technische universität
dortmund

Faculty of Statistics

Subgroup analyses and investigations of treatment effect heterogeneity in clinical dose-finding trials

Marius Thomas

Dissertation

submitted to the Faculty of Statistics at TU Dortmund University in partial fulfilment of the requirements for the degree *Doktor der Naturwissenschaften*

Advisors and referees: Prof. Dr. Katja Ickstadt, Dr. Björn Bornkamp

Referee: Prof. Dr. Jörg Rahnenführer

Chair of the thesis comitee: Prof. Dr. Guido Knapp

Submitted: 30.11.2018

Date of oral examination: 25.01.2019

Acknowledgements

Many people directly or indirectly supported the work on this thesis and made these three years an interesting, educational and fun experience for me.

My greatest thanks go to my advisor Björn Bornkamp, who gave me great freedom to pursue my own ideas and whose door was open at all times, when I needed advice. I am also very grateful to Katja Ickstadt, who was always available for discussions during my visits in Dortmund and gave particularly helpful feedback on the Bayesian aspects of my thesis. Franz König and Martin Posch were great hosts during the three months I spent at the Medical University of Vienna and made my stay a productive and enjoyable experience. My thanks also go to Heidi Seibold for sharing her expertise on the model-based recursive partitioning with me, leading to a fruitful collaboration.

I am very grateful to have been a part of the IDEAS network during my PhD, and I would like to thank the supervisors of the network for creating it and organizing our meetings together. A special shout-out goes to the other IDEAS students, I have many great memories from our trips together to summer schools and conferences. I would also like to thank my colleagues from the Novartis Statistical Methodology Group for creating a very pleasant atmosphere at work, in particular my office mate Johanna, who was always able to answer all my questions about deadlines or PhD formalities. I am also grateful to my friends, the ones in Basel, who were always available for an after work beer, when I needed it, and the ones in Dortmund, who always made me look forward to coming back. Finally I would like to thank my family, my parents and Christina, for encouraging me to take on this challenge and supporting me all the way through.

This dissertation was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 999754557. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government. The project is part of the IDEAS European training network from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567.

List of contributed articles

Thomas, M. and Bornkamp, B. (2017): Comparing approaches to treatment effect estimation for subgroups in early phase clinical trials. *Statistics in Biopharmaceutical Research*, 9 (2), 160–171

Thomas, M., Bornkamp, B. and Seibold, H. (2018c): Subgroup identification in dose-finding trials via model-based recursive partitioning. *Statistics in Medicine*, 37 (10), 1608–1624

Thomas, M., Bornkamp, B. and Ickstadt, K. (2018a): Identifying treatment effect heterogeneity in dose-finding trials using Bayesian hierarchical models. *submitted for publication*. Preprint: arXiv:1811.10488

Thomas, M., Bornkamp, B., Posch, M. and König, F. (2018b): A multiple comparison procedure for dose-finding trials with subpopulations. *submitted for publication*. Preprint: arXiv:1811.09824

Contents

Chapter 1

Introduction

Over the last 20 years there has been a shift in drug development towards stronger taking individual differences between patients into account. Genetic and molecular information for each patient is more easily available than ever and pharmaceutical companies hope to reduce failure rates in clinical trials by identifying the patients, for which a new drug is most effective (Woodcock (2007)). The increased interest in *personalized* or *precision medicine* (see for example Ginsburg and McCarthy (2001); Woodcock (2007); Hamburg and Collins (2010)) further drives these changes.

As a result of these trends in medicine and drug development, exploratory subgroup analyses have become commonplace for many clinical trials. Often such analyses have the aim of identifying a subgroup of patients with high treatment effects, especially when the treatment effect in the overall population is insufficient and development would otherwise be stopped. Alternatively, when the treatment is shown to be effective overall, such analyses can be used to confirm the consistency of the beneficial treatment effect across the whole patient population. For these purposes, exploratory analyses are routinely conducted in addition to the primary analysis of the trial, which is in most cases concerned with assessing overall efficacy of the treatment. Generally these analyses can however not be used to rescue failed confirmatory trials. Instead the results can inform design of future trials, which can for example take the subgroup into account and potentially replicate the finding. The European Medicines Agency (EMA) discusses possible scenarios and consequences of subgroup analyses in detail in their draft guideline on the topic (EMA

(2014)).

The subgroup analyses we consider in this thesis are commonly performed for Phase II and III clinical trials. Phase II clinical trials are usually the first studies in patients. Phase II studies therefore aim to establish *Proof of Concept* (PoC), i.e. that the drug shows the desired effect in patients and is more effective than a placebo. After PoC, dose-finding trials, in which the efficacy and safety of several doses of the treatment are investigated, are conducted in this phase. Trials in Phase II are generally small for financial and ethical reasons and don't recruit more than a few hundred patients. Phase III trials are confirmatory trials, that aim to confirm the results from previous exploratory studies and to show, that the new treatment is efficacious and safe enough to warrant approval and market release. Phase III trials are generally larger than Phase II trials and often recruit thousands of patients. While Phase III studies are primarily confirmatory, additional exploratory analyses, for example investigating treatment effect heterogeneity, might be conducted.

In this cumulative thesis we develop and investigate statistical tools for identifying and confirming subgroups in clinical trials. From a statistical perspective, for most clinical trials we have a clinically relevant response Y and a new treatment T and we want to assess if the treatment has a positive effect on Y. When considering treatment effect heterogeneity and possible subgroups, we have to consider the effect of additional covariates $\boldsymbol{X}$, which might interact with the treatment T and can explain differences in treatment effects between individual patients. Covariates in $\boldsymbol{X}$ can be binary subgroup variables, but in this thesis we often consider the original baseline covariates, which are underlying possible subgroups. Thus $\boldsymbol{X}$ can also include continuous or categorical variables. Broadly speaking, the problems discussed in this thesis boil down to identifying if treatment effect heterogeneity exists and which covariates can be used to explain it. Furthermore we discuss how to define suitable subgroups based on these covariates.

These problems are challenging from a statistical perspective, since the covariate vector $\boldsymbol{X}$ often has a relatively large dimension. In Phase II trials an additional challenge is the small sample size. On the one hand, this leads to a high probability of false positive subgroup identifications, when the resulting multiplicity is not properly taken into account (Pocock et al. (2002)). On the other hand, there is a high probability of a false negative, where an

existing subgroup can not be detected, because of small sample size (Keene and Garrett (2014)). Further complications arise, since we are focusing on identifying interactions (with treatment), which makes detection of existing effects even more difficult. In addition treatment effect heterogeneity might also be caused by covariate-covariate interactions, which further increases the number of possible subgroups.

In recent years several statistical approaches to these problems have been proposed, employing for example tree-based recursive partitioning algorithms, which are well-suited for handling interactions, penalized regression methods, which can be used to prevent overfitting, when explicitly modeling a large number of covariate effects, or Bayesian approaches, which allow incorporating uncertainty and can be used to make optimal decisions with regard to subgroups (Lipkovich et al. (2017)). In confirmatory settings parametric testing approaches can be used to control the type I error rate, while taking the overlap between subgroup and the full population into account (Alosh et al. (2017)).

The available literature focuses however on two-arm clinical trials, where patients are randomized to the experimental treatment or a control (e.g. current standard of care or placebo). A main contribution of this thesis is the development of statistical methodology for identification of subgroups in dose-finding trials with more than two arms. Dose-finding trials play a key role in the drug development process, since they provide valuable information about the effect of the dose on efficacy and safety. In dose-finding trials patients are administered several doses of the new drug. Thus, instead of considering a binary treatment variable T, as is common in the literature on subgroup analyses, we consider a dose variable D and explicitly take the relationship between D and the response Y into account. For identifying subgroups in this setting we consider the treatment effect to be a function of the dose and then try to identify relevant covariate effects $\boldsymbol{X}$ on this treatment effect curve. This allows us to not only identify subgroups with higher treatment effects but also subgroups, which require a different dose of the treatment.

The two chapters following this introduction aim to provide the reader with the necessary background for the contributed articles, which make up the main part of this thesis. Chapter 2 focuses on statistical aspects of subgroup analyses for the routinely considered setting of two-arm clinical trials. The statistical challenges are discussed in detail and an overview over the available statistical methodology is given. Chapter 3 introduces

Phase II dose-finding trials and gives an overview over general principles and commonly used dose-response models. In addition the *MCP-Mod* methodology for the analysis of dose-finding trials, originally proposed in Bretz et al. (2005) is introduced.

In Chapter 4 the four contributed articles are summarized and discussed. The articles deal with different aspects of subgroup analyses. In Thomas and Bornkamp (2017) we consider the problem of unbiased treatment effect estimation after a subgroup has been identified. In Thomas et al. (2018c) and Thomas et al. (2018a) we focus on subgroup and treatment effect heterogeneity identification in dose-finding trials. Thomas et al. (2018b) deals with a more confirmatory setting, where a subgroup is prespecified and tests for a significant dose-response signal are performed under uncertainty about the underlying dose-response model.

Chapter 2

Subgroup analyses in clinical trials

This chapter serves as an introduction to the statistical aspects of subgroup analyses and subgroup identification in clinical trials to provide the necessary background for the contributed articles. The chapter is divided into two main sections. The first deals with exploratory subgroup analyses, which are the main focus of this thesis. In this Section the statistical challenges of these analyses are discussed and some possible approaches are presented, first basic ones in Subsection 2.1.2 and then more advanced approaches in Subsection 2.1.3, which have been proposed in the literature over the recent years. In the second part of this chapter multiplicity adjustments for confirmatory subgroup analyses are briefly discussed to provide the necessary background for Thomas et al. (2018b).

2.1 Exploratory subgroup analyses

2.1.1 Standard models and statistical challenges

In the introduction some intuition was already given on the statistical principles and challenges behind subgroup analyses. In this section these aspects will be discussed in more detail. First we introduce some general notation. Assume we conduct a clinical trial for a new treatment, where the clinical response variable is Y and we observe responses for the n patients in the trial, $y_1, \ldots, y_n$. In general Y could be related to efficacy or

safety of the drug in question, however in the context of subgroup analyses we usually focus on efficacy. Y can be a continuous, binary, time-to-event or count variable. T is the treatment variable and we denote by t_i the treatment indicator for patient i. In this Section we assume, that there are two arms in the clinical trial, one that receives the new treatment (patients with $t_i = 1$) and one that receives a control (patients with $t_i = 0$), for example the current standard of care or a placebo. Such clinical trial designs are common in Phase II Proof of Concept trials and many Phase III trials. Trials with more than two arms and multiple doses of the same treatment will be discussed in Chapter 3.

In the context of subgroup analyses and investigations of treatment effect heterogeneity, baseline covariates $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(k)})'$ play an important role. Such covariates can for example be demographic, clinical, or genetic. These covariates are measured for each patient at baseline, before any treatment has been given. The dimension k of the covariate vector can range from 5-30 for the scenarios we consider in this thesis. For each patient we then have an observed vector of covariates, $\boldsymbol{x_i} = (x_i^{(1)}, \ldots, x_i^{(k)})'$.

Defining $f(\boldsymbol{x}, t) := E(Y|\boldsymbol{X} = \boldsymbol{x}, T = t)$ as the expected response of a patient with covariate vector of $\boldsymbol{x}$ and assigned to treatment t, we can generally write this mean response as (Lipkovich et al. (2017))

$$f(\boldsymbol{X}, T) = g(h(\boldsymbol{X}) + l(z(\boldsymbol{X})T)), \tag{2.1}$$

where g and l are monotone functions, h describes the baseline response without treatment and z describes the treatment effect. This model is very general and encompasses a large range of models for different types of outcomes. For time-to-event data, where generally one is interested in survival times and count data, where rates are of main interest, covariate effects would however be modeled on the hazard function or rates, instead of the expected response. In the case of a continuous response, which we consider in most of this thesis we can simplify the model and write

$$f(\boldsymbol{X}, T) = h(\boldsymbol{X}) + z(\boldsymbol{X})T.$$

If $z(\boldsymbol{X}) = c$, where c is some constant, there is no treatment effect heterogeneity.

Based on model (2.1), a distinction between two types of covariates can be made. Covariates, that contribute to $h(\boldsymbol{X})$ are called *prognostic* and influence the baseline response,

independent of the given treatment. Covariates, that contribute to $z(\boldsymbol{X})$ are called *predictive* and interact with the treatment (Oldenhuis et al. (2008)). Covariates can naturally also be both, prognostic and predictive. These different types of covariate effects are visualized in Figure 2.1. In the context of subgroup analyses predictive covariates are of main interest, since they have an influence on treatment effects. Based on predictive covariates, subgroups of patients with increased treatment effects can be defined as a subset of the covariate space.

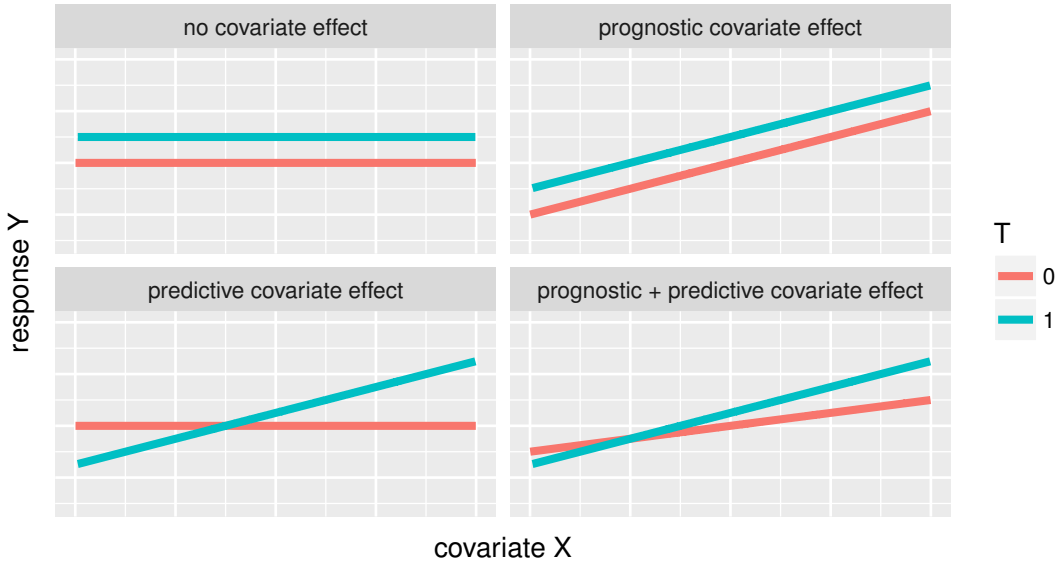

Figure 2.1: Different types of covariate effects. The difference between the two lines is the treatment effect, which is only altered by predictive covariates.

For example, a simple rule for defining a subgroup could be $S = \{\boldsymbol{X} | z(\boldsymbol{X}) > \nu\}$, where ν is a relevant threshold on the treatment effect. When z is a complex function, such a subgroup definition could however depend on a large number covariates. This means, that for a new patient, all covariates have to be measured to determine if the patient belongs to the subgroup, which can often be impractical. Therefore simpler subgroup definitions are often used, for example by finding a suitable cut-off q on a continuous covariate and defining $S = \{X^{(1)} > q\}$, so that the subgroup definition is univariate.

In practice, instead of using the raw baseline covariates for the exploratory analyses, a set of candidate subgroups will often be considered. These candidate subgroups are commonly derived from baseline covariates, for example by using quantiles of the empirical distribution of the covariate as cut-offs. Alternatively cut-offs can be specified based on

clinical considerations. A set of subgroup variables $S^{(1)}, \ldots, S^{(k^*)}$ is then obtained, where generally $k^* \geq k$, since multiple cut-offs could possibly be considered for each covariate. The aim of the analysis would then be to find a promising subgroup among this predefined set. Considering only a set of candidate subgroups can simplify the analysis, since the underlying covariates can then be ignored, however dichotomization of continuous covariates will generally lead to a loss of information and is often criticized (Weinberg (1995); Royston et al. (2006)).

From a statistical perspective subgroup analyses in the settings described above are challenging. Multiplicity is commonly considered to be one of the major challenges of subgroup analyses (Lagakos (2006); Wang et al. (2007); Ruberg and Shen (2015)). Multiplicity arises firstly through the large number of possible predictive covariates and is then amplified by multiple possible combinations of covariates and also multiple cut-offs for continuous covariates. If multiplicity is ignored, there is a high chance for false positive findings, which could then lead to failures in later trials. Obtaining reliable treatment effect estimates is a related issue, since unadjusted treatment effect estimates in selected subgroups will generally suffer from selection bias and confidence intervals will not account for the uncertainty with regard to the selection procedure (Rosenkranz (2016)). In addition sample sizes in clinical trials are primarily chosen based on power considerations for treatment effects in the whole population. Statistical tests to detect possible subgroups focus on interactions (with treatment), and generally have low power, because of the lower sample sizes in subgroups and the often comparably small differences between subgroups (Wang and Ware (2013)).

2.1.2 Basic approaches

Before giving an overview over the methods proposed in the literature, we will go over some basic approaches to subgroup analyses and subgroup identification, which are commonly employed in practice. We will discuss the shortcomings of these approaches to further illustrate the challenges of subgroup analyses and to provide some motivation for the use of more advanced approaches.

In practice, many exploratory subgroup analyses are conducted using univariate subgroup

models, where for each candidate subgroup a separate model is fit to the data. For example, for a continuous response Y and assuming a standard linear model, k^* models can be fit to the data using standard Maximum-Likelihood estimation (MLE), with conditional mean modeled as

$$E(Y|S^{(j)}, T) = \beta_0 + \beta_1 S^{(j)} + (\beta_2 + \beta_3 S^{(j)})T, \;\; j = 1, \ldots, k^*. \tag{2.2}$$

Then for each model an interaction test for the null hypothesis $\beta_3 = 0$ is performed. A naive subgroup identification strategy chooses either all subgroups, where the p-value for the interaction test is below some threshold α (for example $\alpha = 0.1$) or chooses the subgroup with the lowest p-value.

It's easy to see, that such an approach suffers from the issues with subgroup analyses described above. Multiplicity is not properly taken into account and the chance for false positive findings is rapidly increasing with k^*. While standard multiplicity adjustments such as Bonferroni (Dunn (1961)) could be used to adjust for the number of tests, the resulting critical values would require extreme evidence for a subgroup to pass the threshold, if k^* is reasonably large. It is then almost impossible to detect an existing subgroup, because of small sample sizes and the low power of interactions tests. In addition unadjusted treatment effect estimates for the most promising subgroup from the above models will be strongly biased upwards.

One issue with the univariate models above is that only one covariate or subgroup is considered at a time. Thus there is no adjustment for (prognostic or predictive) effects of other covariates and possible interactions between covariates are ignored. As a simple alternative a multivariate linear model including all candidate covariates,

$$E(Y|\boldsymbol{X}, T) = \beta_0 + \sum_{j=1}^{k} \beta_{1.j} X^{(j)} + (\beta_2 + \sum_{j=1}^{k} \beta_{3.j} X^{(j)})T, \tag{2.3}$$

could be considered. However such a model would be highly prone to overfitting (Hawkins (2004)) for the considered settings, where n is relatively small and k can be large. These tendencies to overfit would be further increased, when interactions between covariates would be included. In addition to the overfitting problems, interaction tests for single covariates would have greatly reduced power.

Regression trees (Breiman (2017)) and similar tree-based approaches, like random forests

(Breiman (2001)), intuitively seem like a good approach to the problem, since they partition the overall data-set into subgroups, in which different models are adequate. Compared to the univariate and multivariate models described above, they are good at handling interactions between covariates and also find suitable cut-offs for continuous covariates. Fitting a regression tree with Y as the outcome and $(\boldsymbol{X}, T)$ as predictors returns however subgroups with differential outcomes, while of main interest would be subgroups with differential treatment effects. The regression tree does not distinguish between prognostic and predictive covariates, since both would affect the outcome Y. Here we are however particularly interested in identifying covariates as predictive and the distinction between prognostic and predictive effects is therefore of great importance. While tree-based approaches can find interactions between covariates and treatment, as well as covariate-covariate interactions, a subgroup identified with a simple regression tree does not necessarily indicate heterogeneous treatment effects.

The basic approaches discussed here are therefore not sufficient for the problems at hand and can only be used as a starting point for further analyses with more advanced methods. In the following an overview over such methods, which have been proposed in the literature in recent years will be provided.

2.1.3 Review of the literature

This section provides an overview over the recent literature on subgroup analyses. This overview is split into three parts, one for each of the main approaches used, recursive partitioning, penalized regression and Bayesian approaches. The literature on the topic is extensive, therefore a complete review of all relevant publications in this field would be beyond the scope of this thesis. We will therefore focus on the most important contributions, illustrating the main ideas of each. For extensive reviews of the literature on subgroup analyses see Ondra et al. (2016) or Lipkovich et al. (2017).

Recursive partitioning A large number of approaches proposed in the literature can be classified as recursive partitioning methods. In the context of subgroup analyses recursive partitioning methods use tree-based algorithms to detect optimal splits over the

available baseline covariates, resulting in subgroups with differential treatment effects. The final output of such methods is commonly a tree, in which each terminal node represents a subgroup and the splitting rules of the tree determine the subgroup definitions. An example of such a tree output is also shown in Chapter 4 in Figure 4.2. Recursive partitioning approaches are well suited to the problems discussed in this thesis, since they are good at handling interaction through the tree structure, and are usually flexible with regards to the underlying model (Ruberg and Shen (2015)). In addition the tree outputs are easily interpretable. In comparison to the simple regression tree procedure described in the previous section, the following methods focus on the function determining the treatment effect, $z(\boldsymbol{X})$ in model (2.1), to partition the patients based on predictive covariates and obtain subgroups with differences in treatment effects instead of the overall outcome. To prevent overfitting and to reduce the number of false positives, many of the approaches described below use pruning or stopping criteria to control the size of the final tree or the number of identified subgroups. Many of the approaches also use bootstrapping or similar resampling techniques to obtain adjusted treatment effect estimates or p-values in identified subgroups. In addition variable importance measures are often provided to rank the predictive covariates.

The *interaction tree* (*IT*) method (Su et al. (2009)) was one of the first approaches using recursive partitioning for subgroup identification. The *IT* algorithm is based on the *CART* algorithm (Breiman (2017)), however instead of splitting based on differences in the outcome Y, *IT* uses a splitting criterion, which is based on interaction tests in models similar to 2.2. The optimal split is chosen by maximizing the interaction test statistic over all covariates and over all possible splits for each covariate. This guarantees that only predictive covariates, which contribute to $z(\boldsymbol{X})$, are used for splitting. Similar to *CART*, the method uses pruning to control the complexity of the tree and prevent overfitting. Variable importance measures are provided.

Virtual Twins (Foster et al. (2011)) is a two step procedure, which combines random forests and regression trees. In the first step random forests are used to make a prediction for the missing outcome of each patient, i.e. the potential outcome under control is predicted for patients on the treatment arm and the potential outcome under treatment is predicted for patients on the control arm, using the available covariates as predictors.

With these predictions estimates of the individual treatment effect for each patient can be obtained. These are then used as outcomes in a regression tree, which again uses the covariates as predictors. This way a regression tree can be fit directly on individual treatment effects and can essentially be used to model $z(\boldsymbol{X})$, ignoring the prognostic terms of model (2.1). To provide reliable treatment effect estimates in identified subgroups resampling methods are used.

SIDES (Lipkovich et al. (2011); Lipkovich and Dmitrienko (2014)) is a recursive algorithm, which searches for promising subgroups. Compared to the methods described above, *SIDES* does not try to model $z(\boldsymbol{X})$ over the whole covariate space and instead focuses on regions of the covariate space, where treatment effects are large. To identify these regions the authors provide several possible splitting criteria to find the best covariates and the best splits on each level of the recursion. These splitting criteria for example maximize the differences in treatment effect between a subgroup and a complement or simply maximize the treatment effect in one of the subgroups. Based on these criteria, *SIDES* then retains the M most promising subgroups, where M is a tuning parameter of the algorithm. Only in these subgroups with large treatment effects the algorithm continues with the next step of the recursion and the complements of these subgroups, which generally will have small treatment effects, will be ignored. The algorithm continues until the improvement in the splitting criterion becomes negligible. For each candidate subgroup the algorithm produces an adjusted p-value using a permutation approach. Additionally variable importance measures are provided to identify predictive covariates.

GUIDE (Loh (2002)) is a general regression tree algorithm like *CART*, which was extended to subgroup identification in Loh et al. (2015). In contrast to traditional regression tree approaches like *CART*, which find the optimal split over all possible cut-offs for all covariates, *GUIDE* separates the selection of covariate and cut-off. This avoids a bias for selecting variables with more possible splits, which would, for example, lead to continuous covariates ($n-1$ possible cut-offs) being selected more often than binary covariates (1 possible cut-off). For the purpose of subgroup identification, *GUIDE* first calculates lack-of-fit test statistics for univariate models with only the prognostic effect of each covariate and the treatment effect, without any interactions of treatment and covariate. If a covariate is truly predictive (e.g. interacting with the treatment) it is expected, that the

lack-of-fit statistics will be large for this covariate. Therefore the covariate, which maximizes this lack-of-fit test statistics is chosen for the split. The optimal cut-off for the split is then selected, so that the residual sum of squares in the two resulting daughter nodes are minimized. To provide adjusted treatment effect estimates in subgroups a bootstrap procedure is used.

Model-based recursive partitioning (*MOB*) (Zeileis et al. (2008); Seibold et al. (2016)) is a recursive partitioning algorithm, which tries to detect parameter instabilities to identify predictive (and prognostic) covariates. The idea behind this approach is similar to *GUIDE*, in that essentially lack-of-fit in models without covariate effects is evaluated. As the name suggests, *MOB* is however model-based and allows fitting a wide range of parametric models in each node of the tree. The test statistics used to detect covariate effects are based on these parametric models and tend to be more powerful than the tests used by *GUIDE* (Seibold et al. (2016)). Similar to *GUIDE*, *MOB* also separates selection of covariate and cut-off to avoid the aforementioned bias for selecting covariates with more possible splits. *MOB* fits a parametric model without covariate effects in each node of the tree. Then tests on the score function, the first derivative of the objective function of the model (for example negative log-likelihood or residual sum of squares) are employed to determine, if the scores fluctuate randomly around zero or if the scores depend on covariates. The covariate which maximizes the test statistics is chosen for the split. An optimal cut-off is determined by minimizing the objective function in the two resulting daughter nodes. To control the complexity of the tree, the p-value corresponding to the best covariate, which can be adjusted for multiplicity using a Bonferroni correction, has to be below a prespecified significance level α for a split to occur. An extension of *MOB* for dose-finding trials is presented in Thomas et al. (2018c).

Huang et al. (2017) investigate the use of *bootstrapping and aggregating of thresholds from trees* (*BATTing*) (McKeegan et al. (2015)) for subgroup identification. The proposed approaches are similar to *SIDES*, in that they only consider interesting regions of the covariate space. The algorithm sequentially builds a subgroup definition by adding the covariates and corresponding cut-offs, which maximize the score test statistics for the interaction terms between subgroup and treatment. To find the optimal cut-off for each covariate *BATTing* is used: Bootstrap samples are drawn from the data and for each

bootstrap sample the optimal cut-off (maximizing the score tests statistics) for each covariate is obtained. Then for each covariate the average cut-off over all bootstrap samples is chosen to determine, which split is included in the subgroup definition. Through *BAT-Ting*, variance introduced through the choice of the cut-off is reduced. Similar to *MOB* and *GUIDE* this also removes the bias for covariates with more possible splits, since for each covariate only one split is considered in the final ranking. The authors propose the use of cross-validation to obtain adjusted p-values for the resulting subgroup.

Penalized Regression As discussed in Section 2.1.2, we could assume linear functions of the covariates for h and z in (2.1) and estimate a model as (2.3) using standard maximum-likelihood estimation. When sample sizes are however comparatively low, as in the settings considered in this thesis, and the dimension of the covariate vector k is large, this quickly becomes infeasible, since overfitting will occur for models with too many parameters. Penalized regression methods deal with this problem through regularization of coefficients and add penalty terms to the objective function of the model. A standard penalized regression method is lasso regression (Tibshirani (1996)), which uses the absolute size of coefficients in the penalty terms. In a Bayesian setting penalized regression can be achieved through shrinkage priors, which put more prior mass on values close to zero. Examples for such priors are for example the Bayesian lasso (Park and Casella (2008)) or the horseshoe (Carvalho et al. (2010)). Bayesian penalized regression for subgroup analyses is also considered in Thomas et al. (2018a).

Penalized regression methods are an attractive tool in the context of subgroup analyses, since they generally lead to sparse models, in which only few covariates interact with the treatment. Compared to the recursive partitioning methods discussed in the previous section they generally require stronger parametric assumptions (for example linearity of covariate effects and no interactions between covariates) but provide a model for the outcome and the treatment effect, on which further inferences can be based. While recursive partitioning methods generally provide subgroup definitions directly, penalized regression approaches often define subgroups based on individual treatment effect estimates, using subgroup definitions of the form $\mathrm{S} = \{\boldsymbol{X} | z(\boldsymbol{X}) > \nu\}$, where ν is some threshold. Some examples of penalized regression approaches used in the context of subgroup analyses are

provided below.

Imai et al. (2013) propose an approach for binary outcomes using a Support Vector Machine (SVM) to estimate model (2.3). The authors propose lasso penalization on the prognostic effects $\beta_{1.1}, \ldots, \beta_{1.k}$ and predictive effects $\beta_{3.1}, \ldots, \beta_{3.k}$, with different lasso penalty parameters. The obtained model can be used to obtain individual treatment effect estimates for each patient, based on which subgroups can be identified.

A *modified outcome* approach is presented in Tian et al. (2014) and combined with lasso regression to obtain individual treatment effects. For a continuous outcome Y and assuming equal randomization to treatment and control the authors propose using the modified outcome $Y^* = 2YT^*$, where $T^* = 2T - 1$ is a modified treatment variable. Then, instead of fitting model (2.3) a model, where $E(Y^*|\boldsymbol{X} = \boldsymbol{x}) = \beta_1 + \sum_{j=1}^{k} \beta_{3.k} X^{(k)}$ is used, so that prognostic covariate effects don't have to be estimated. Tian et al. (2014) provide justifications for this approach and extend it to different types of outcomes. As Imai et al. (2013) they use lasso regression to obtain estimates for the coefficients $\beta_{3.1}, \ldots, \beta_{3.k}$.

Ghosh et al. (2015) propose an approach for identifying subgroups with heterogeneous treatment effects, which is similar to *Virtual Twins* by Foster et al. (2011). As for *Virtual Twins*, in the first step random forests are used to predict the missing outcome for each patient, which is then used to obtain individual treatment effect estimates. In the second step lasso regression is employed to model $z(\boldsymbol{X})$ and select relevant covariates. As a further extension the authors argue, that the first step can essentially be considered as imputation of a missing value. Instead of performing a single imputation, the extension uses multiple imputations to take into account uncertainty in the prediction: Imputed values are drawn from a normal distribution with the random forest prediction as the mean and the average mean squared error of the random forest as the variance. Therefore multiple sets of individual treatment effect estimates are obtained. For the second step the authors then propose Lasso regression for the average individual treatment effects over these datasets or fit separate lasso regressions for each imputed dataset.

Bayesian approaches Bayesian approaches are used in the context of subgroup analyses to shrink subgroup effects through hierarchical modeling. This can prevent overfitting

and reduce false positive identifications of subgroups. This makes them conceptually similar to the penalized regression methods discussed above, however the methods discussed in the following do not fit multivariate models for the underlying covariates as in (2.3) but rather consider binary subgroup variables. In addition Bayesian decision-theoretic approaches have been proposed to provide complex decision rules, which apart from the treatment effect size also take the size of the subgroup or the complexity of the subgroup definition into account, when deciding on an optimal subgroup.

Jones et al. (2011) propose a Bayesian hierarchical model for modeling treatment effects in non-overlapping subgroups defined by binary covariates. The approach extends an earlier Bayesian model for subgroup analysis by Dixon and Simon (1991). Dixon and Simon (1991) proposed modeling the treatment effect in each subgroup as a linear function of the binary covariates with only main effects of each covariates. Jones et al. (2011) extend this approach to also include two-way and three-way interactions between covariates. Shrinkage is induced through priors, which assume covariate main effects are drawn from prior distributions with the same variance components and similarly two-way and three-way interactions are also drawn from the same prior. The model can be applied to different types of outcome, however only to binary covariates. In addition this approach is only feasible for a small number of covariates (Jones et al. (2011) consider three), since treatment effects are explicitly modeled in each possible subgroup. For larger number of covariates the number of non-overlapping subgroups quickly becomes very large and as a result the number of patients in each is very small.

Bayesian model averaging is proposed for subgroup identification in Berger et al. (2014). In the context of model averaging each candidate subgroup can be considered to correspond to a model. Berger et al. (2014) consider binary covariates and consider separate models for prognostic and predictive effects of covariates. They use linear models, in which up to one of the covariates is prognostic, and up to one is predictive and include all possible combinations of prognostic and predictive covariates in their model space. In addition null models, where the treatment effect is zero or the treatment effect is the same for all patients are included in the model space. Appropriate choices of prior model probabilities to adjust for multiplicity are discussed. The approach then provides posterior model probabilities for each model, as well as several quantities of interest related to

patients' individual treatment effects and to treatment effects in subgroups.

A Bayesian model averaging approach is also considered in Bornkamp et al. (2017). However the focus for this paper lies on treatment effect estimation in prespecified subgroups. The authors argue, that subgroup selection can be seen as model selection. Following this line of thought, unadjusted treatment effect estimates in selected subgroups do not properly account for model uncertainty and thus suffer from a selection bias. Model averaging is proposed to take model uncertainty into account and obtain adjusted treatment effect estimates in subgroups. For each subgroup a corresponding univariate model of type (2.2) is included in the model space. Additionally a null model with no treatment effect heterogeneity can be included. The model averaging approach is then used to obtain treatment effect estimates and credible intervals for candidate subgroups with reduced selection bias. This model averaging approach is compared to other possible treatment effect estimation methods for subgroups in Thomas and Bornkamp (2017).

A Bayesian decision-theoretic approach to identify the best subgroup from a set of candidate subgroups was proposed by Sivaganesan et al. (2017). The proposed approach is divided into two steps. First MCMC samples from the posterior predictive distribution of the individual treatment effect for each patient are obtained from a Bayesian regression tree (*BART*) (Chipman et al. (2010)). A utility function, which depends on the treatment effect in the subgroup, the size of the subgroup and the number of covariates included in the subgroup definition is proposed. The best subgroup is then found by maximizing posterior expected utility over the set of candidate subgroups, using the MCMC samples of the individual treatment effects. The candidate set for this approach always includes an empty subgroup to allow for not identifying any subgroups.

A very similar decision-theoretic approach, termed *Bayesian population finding* (*BaPoFi*), is presented in Morita and Müller (2017). This approach explicitly considers the situation of a Phase II trial and poses the decision problem at the end of Phase II in a Bayesian framework. The approach tries to optimize utility over a set of possible decisions: the development for the treatment is stopped (no-go decision) or continued (go decision) with Phase III trials in the entire population, only in a subgroup (identified from a set of candidate subgroups) or in the entire population and the subgroup.

2.2 Confirmatory subgroup analyses

The previous sections in this chapter were concerned with exploratory subgroup analyses, where the main aim is commonly the identification of a subgroup with an increased treatment effect. If a potential subgroup is identified as a consequence of such an analysis, it typically needs to be confirmed in a later trial. Depending on the size of the treatment effect in the subgroup and the overall treatment effect, several different trial designs with subgroups are possible. If the treatment is shown to be effective only in the subgroup the trial might only be conducted in this subgroup. Alternatively the trial might investigate the subgroup in addition to the overall population. Enrichment trial designs, in which patients from the subgroup are recruited at a higher rate are possible as well. Possible designs for trials with subgroups are for example discussed in Ondra et al. (2016) and Dmitrienko et al. (2016).

In confirmatory clinical trials with subgroups multiplicity issues arise, when tests for efficacy are performed in the subgroup and the full population. In such trials there are generally two main null hypotheses of interest: $H_0^{(F)}$, the null hypothesis of no treatment effect in the full population and $H_0^{(S)}$, the null hypothesis of no treatment effect in the subpopulation. Let Z_F and Z_S then denote the corresponding test statistics. From the sponsor's perspective the trial will often be successful, if one of these null hypotheses can be rejected. The aim of multiplicity adjustments in confirmatory clinical trials is to control the family-wise error rate (FWER), i.e. the probability to reject at least one true null hypothesis at a nominal level α. The multiplicity adjustments discussed in the following provide FWER control in the strong sense. This means, that they control the aforementioned probability under all combinations of false and true null hypotheses. Weak FWER control on the other hand would only control the probability to reject at least one true null hypothesis under the global null hypothesis $H_0^{(F)} \cap H_0^{(S)}$.

Simple nonparametric multiplicity adjustments, as for example the Bonferroni-Holm procedure (Holm (1979)), control FWER in this setting are however unnecessarily conservative by not taking all available information into account. Semiparametric procedures are a more powerful alternative, for example Hochberg (Hochberg (1988)) or Hommel procedures (Hommel (1988)). These semiparametric methods make some assumptions about

the test statistics' distributions, namely that test statistics are non-negatively correlated. When tests are performed in subgroup and full population, where one population is a subset of the other, this assumption obviously holds. While these approaches will therefore guarantee FWER control, they will generally still be conservative for tests in subgroup and full population, which are positively correlated.

Parametric procedures, which explicitly take the correlation between Z_F and Z_S into account have more power to reject a false null hypothesis. A simple parametric procedure can be constructed by using the joint distribution of Z_F and Z_S under the null hypothesis. A critical value for the testing procedure can be obtained by finding c, for which $P_{H_0}(Z_F > c \text{ or } Z_S > c) = \alpha$, or equivalently $P_{H_0}(Z_F \leq c, Z_S \leq c) = 1 - \alpha$, assuming that large values of the test statistics will lead to rejections of the null hypotheses. Since $P_{H_0}(Z_F > c \text{ or } Z_S > c) = P_{H_0}(max\{Z_F, Z_S\} > c)$ such tests are also known as *maximum tests*. As this procedure uses the exact joint distribution under the null hypothesis, it exactly controls the FWER and is more powerful than the other procedures described above, which make only vague assumptions about the test statistics' distributions. An example of a parametric testing procedure used in clinical trials is the Dunnett test for comparing multiple treatments to the same control (Dunnett (1955)). *MCP-Mod* (Bretz et al. (2005)), which is introduced in Chapter 3 of this thesis, also makes use of parametric procedures. In Thomas et al. (2018b) we derive an extended testing procedure based on *MCP-Mod*, which uses a parametric test for dose-response signals in multiple populations.

Further extensions of parametric tests, which will not be discussed in detail here, make use of sequential testing procedures (Alosh and Huque (2009)) or graphical testing approaches (Bretz et al. (2011)).

Chapter 3

Phase II dose-finding studies

Dose-finding is an essential component of drug development. The term *dose-finding study* is often used to refer to both Phase I and Phase II studies, which investigate several doses of the same treatment. However Phase I and Phase II studies are very different with regard to their design and their objectives. Phase I dose-finding studies are usually the first trials in humans. Commonly, the main aim is the estimation of a *maximum tolerated dose* (MTD), below which no severe side effects are expected to occur. Phase I trials (outside the area of oncology) are typically conducted in healthy volunteers, therefore efficacy cannot be assessed in these trials.

Phase II dose-finding studies are usually run after a proof-of-concept has been achieved in patients, i.e. an early phase efficacy trial was positive. Phase II trials usually have two main objectives (Bornkamp (2017)): Firstly, to investigate general efficacy of the drug to determine if development should continue into Phase III. Secondly, to decide on a dose for a potential Phase III trial, which is both efficacious and safe. Phase II dose-finding trials commonly focus on efficacy instead of safety, since safety-related events are rare for most treatments and hard to model reliably. Instead Phase II dose-finding trials are designed to estimate the efficacy dose-response curve and assume, that toxicity of the drug and related safety-events will increase monotonically with the dose. From a safety standpoint the dose is therefore best chosen as low as possible. The optimal dose for Phase III is thus often the lowest dose, which provides close to maximum efficacy or efficacy beyond a clinically relevant threshold.

The articles related to dose-finding trials, which are part of this thesis, are considering Phase II dose-finding trials. Thus, when we talk about dose-finding trials from here on, we are referring to Phase II trials. For more information on Phase I trials, see for example the relevant chapters in Ting (2006) or O'Quigley et al. (2017). To provide the necessary background for the articles on dose-finding a brief overview over statistical aspects of dose-finding trials and commonly used models will be provided in this chapter, following Bornkamp (2017). Additionally the *MCP-Mod* methodology for the analysis of dose-finding trials is introduced, which was originally proposed in Bretz et al. (2005) and an extension of which is presented in Thomas et al. (2018b).

3.1 General principles

In Phase II dose-finding trials patients are randomized to different doses $d_1^*, \ldots, d_t^*$ of an investigational treatment, where the first dose is commonly a placebo, so that $d_1^* = 0$. Typically the number of doses is between 4-7. To illustrate common methods for the statistical analysis of dose-finding trials a continuous normally distributed clinical response of interest Y is assumed in the following. These types of analyses are however not limited to continuous response variables and can be applied to other outcomes as well. A general model for a normally distributed response of each patient can be specified as

$$Y_{ij} = \mu(d_i^*, \boldsymbol{\theta}) + \epsilon_{ij}, \;\; \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \;\; i = 1, \ldots, t, j = 1, \ldots, n_i, \tag{3.1}$$

where n_i is the number of patients on dose level d_i^* and $\mu(d_i^*, \boldsymbol{\theta})$ describes the mean response at dose d_i^* for some parameter vector $\boldsymbol{\theta}$. Possible models for μ will be discussed in Section 3.2. Most dose-response profiles follow monotonic, plateauing shapes (Thomas et al. (2014)), as for example depicted in Figure 3.2, and the treatment effect $\mu(d) - \mu(0)$ will generally approach a maximum, commonly denoted as E_{max}, as the dose increases.

In dose-finding trials a main quantity of interest is often the dose, for which a specific treatment effect is achieved. The dose for which a treatment effect of $p\%$ of E_{max} is achieved for $p \in [0, 100]$ is denoted by ED_p. ED_{50} and ED_{90}, which are frequently used in practice, are depicted in Figure 3.1. An alternative dose of interest is the *target dose* (TD), which is the minimum dose for which a target treatment effect is achieved. If the

target is the minimum clinically relevant treatment effect, the term *minimum effective dose* (MED) is commonly used. Which of these doses is used as the dose for a possible Phase III trial, depends on the investigated treatment and the related efficacy and safety considerations.

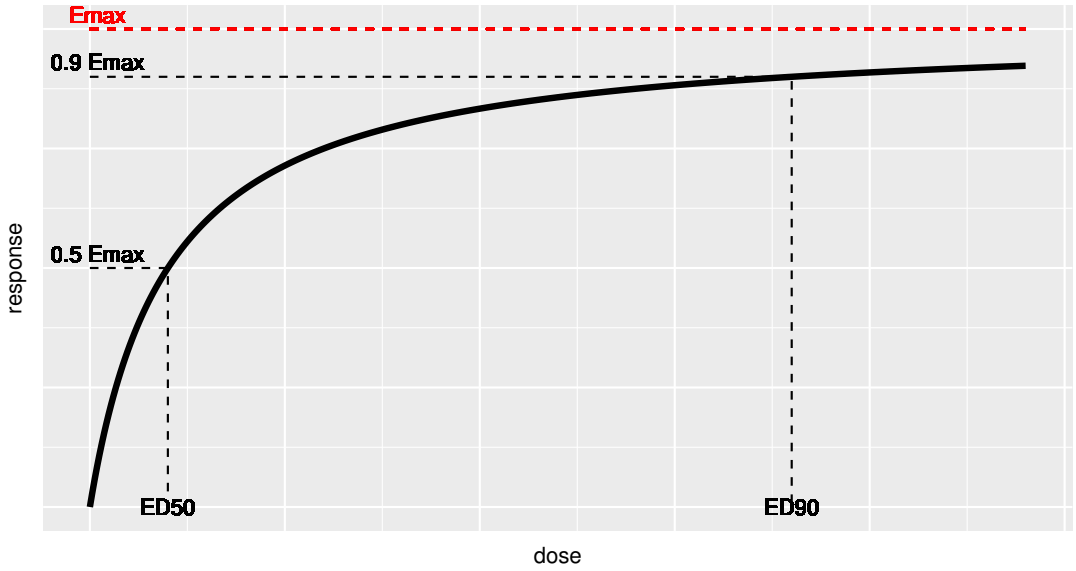

Figure 3.1: ED_{50} and ED_{90} for an E_{max} model.

While the design of dose-finding trials is not a focus of this thesis, it is worth noting, that the choice of the active dose levels $d_2^*, \ldots, d_t^*$ is crucial to guarantee, that the dose-response curve can be adequately estimated. Ideally the investigated doses should include doses, which lie on the ascending part of the dose-response curves, as well as doses which lie on the plateau of the curve. If, for example, the first dose is already on the plateau, then it is unclear if a lower dose could maybe provide the same treatment effect at a lower safety risk. On the other hand, if the plateau is not reached at the highest dose in the trial, E_{max} can not be estimated. For an extensive discussion on optimal designs of Phase II dose-finding trials see Pinheiro and Bornkamp (2017).

3.2 Dose-response models

One of the most common models for dose-response relationships is the E_{max} model, for which the mean dose response from (3.1) is

$$\mu(d, \boldsymbol{\theta}) = E_0 + E_{max} \frac{d^h}{ED_{50}^h + d^h}.$$

Here E_0 represents the response in the placebo group and h is the so called hill parameter, which controls the steepness of the curve. ED_{50} and E_{max} were already introduced in the previous section. The special case of $h = 1$ is sometimes called the *hyperbolic* E_{max} model. Possible E_{max} shapes are depicted in Figure 3.2. The E_{max} model is frequently used, since it can be derived from underlying pharmacological principles (Källén (2007)), is easily interpretable and has been shown to be an appropriate model for the dose-response relationship of a large number of drugs (Thomas et al. (2014)).

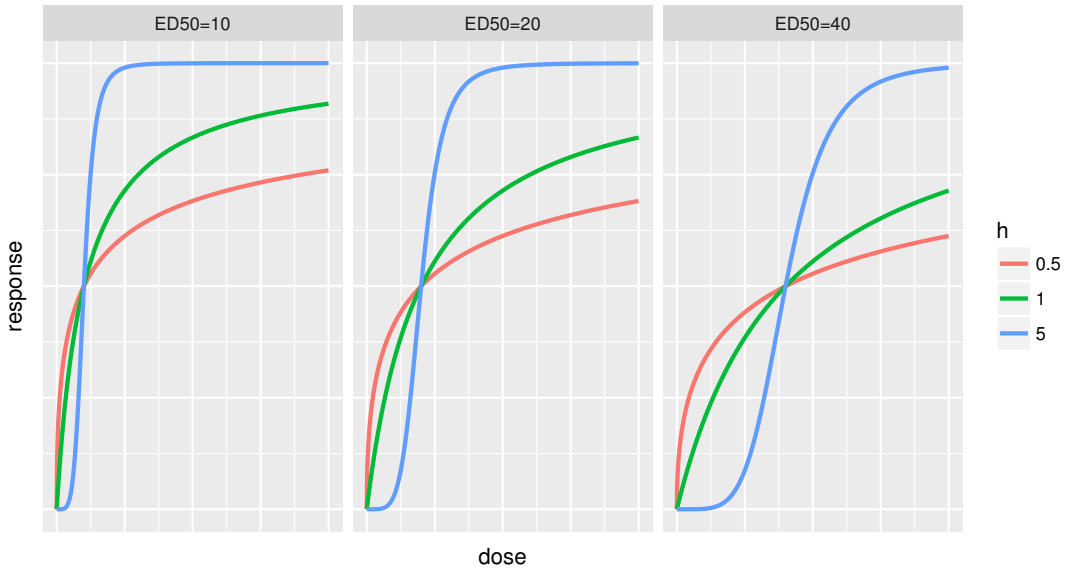

Figure 3.2: E_{max} shapes for different ED_{50} and hill parameters h. The E_{max} parameter is fixed for all depicted shapes.

Still, there might be situations in which other models are more appropriate. For example, when the MTD, which determines the upper limit of the possible dose range, lies below the plateau of the dose-response curve, there is no way to estimate the whole dose-response curve. In such situations simpler models can be adequate, for example exponential or linear models. In addition E_{max} models only allow for monotonous dose-response relationships. Non-monotonous (e.g. quadratic) dose-response could occur, when large doses

lead to increased safety issues, which could then have a negative effect on efficacy as well. Table 3.1 gives an overview over common dose-response models. Possible non E_{max}-models are depicted in Figure 3.3.

The above parametric models can be fit in a frequentist framework using least-squares methods, as for example in the *DoseFinding* package for R (Bornkamp et al. (2013)). Alternatively, Bayesian estimation is possible and has for example been discussed in Thomas (2006) and Bornkamp (2014). Non-parametric dose-response models have been considered in Bornkamp and Ickstadt (2009). Complete nonparametric estimation is however considered difficult, because of usually low number of doses combined with high variability of the response (Bornkamp (2017)). *MCP-Mod*, which is introduced in the next section, relaxes the assumptions of parametric models, by taking model uncertainty into account and allowing different parametric dose-response models.

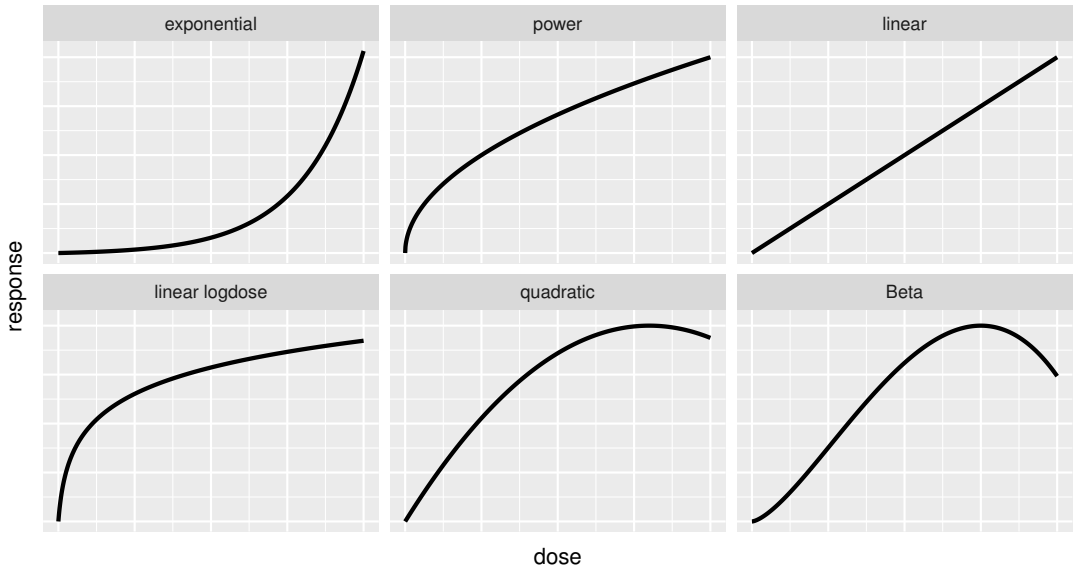

Figure 3.3: Non-E_{max} dose-response models (see also Table 3.1).

3.3 MCP-Mod

MCP-Mod is a method for the analysis of Phase II dose-finding studies. It was first proposed by Bretz et al. (2005) for normally distributed endpoints and has since been extended to general parametric models in Pinheiro et al. (2014). Further details regarding design and analysis were also discussed in Pinheiro et al. (2006). Xun and Bretz (2017)

is a recent book chapter on the method, which discusses many practical aspects.

MCP-Mod combines multiple comparison techniques with dose-response modeling in a two-step procedure. The main objective of the method is to establish a significant dose-response signal and provide an estimate for the recommended dose for a possible Phase III trial, while accounting for possible uncertainty with regard to the underlying dose-response model. Often there is not a lot of information about the shape of the dose-response before the trial, which makes it sensible to allow for some flexibility with regard to the dose-response model, which will be fit to the data.

MCP-Mod derives its name from the two steps, which make up the method. The first part of the method is the multiple comparison procedure (MCP) step, in which a set of possible candidate dose-response models is considered and tests for a significant dose-response signal are performed for each of those candidate models. In the second part of the method, the Mod step, the best candidate model from the MCP step is used to estimate the dose-response curve. Alternatively all significant models can be combined via model averaging.

To explain the method in detail we will again use the basic model (3.1), which assumes a normally distributed clinical response. As mentioned above, *MCP-Mod* has also been extended to other types of outcomes in Pinheiro et al. (2014). For the MCP step a set of Z candidate dose response models for the mean response $\mu(d, \boldsymbol{\theta})$ is considered, which can generally be written as

$$f(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(d, \boldsymbol{\theta^0}),$$

where $f^0(d, \boldsymbol{\theta^0})$ is the *standardized model* for $f(d, \boldsymbol{\theta})$. Commonly considered dose-response models with their standardized forms are summarized in Table 3.1.

Linear contrast tests are used to test for a dose-response signal. A single contrast test tests the null hypothesis of no dose-response signal $H_0 : \boldsymbol{c}'\boldsymbol{\mu} = 0$ versus the alternative of a positive dose-response trend, $H_1 : \boldsymbol{c}'\boldsymbol{\mu} > 0$, where $\boldsymbol{c} = (c_1, \ldots, c_t)'$ is a vector of contrast coefficients with $\sum_{i=1}^{t} c_i = 0$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_t)'$ is the vector of mean responses at the dose levels in the trial. Contrast tests are preferable over simple pairwise comparisons between each dose and placebo, since they take the underlying dose-response relationship into account and combine information across all doses. For a mean response vector $\boldsymbol{\mu}$,

Model	$f(d, \boldsymbol{\theta})$	$f^0(d, \boldsymbol{\theta^0})$
Hyperbolic E_{max}	$E_0 + E_{max}d/(ED_{50} + d)$	$d/(ED_{50} + d)$
Sigmoid E_{max}	$E_0 + E_{max}d^h/(ED_{50}^h + d^h)$	$d^h/(ED_{50}^h + d^h)$
Exponential	$E_0 + E_1 \cdot exp(d/\delta) - 1$	$exp(d/\delta) - 1$
Power	$E_0 + E_1 d^\alpha$	d^α
Linear	$E_0 + \delta d$	d
Linear logdose	$E_0 + \delta log(d + c)$	$log(d + c)$
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	$d + (\beta_2/\lvert\beta_1\rvert)d^2$
Beta	$E_0 + E_{max}B(\delta_1, \delta_2)(\delta/D)^{\delta_1}(1 - d/D)^{\delta_2}$	$(\delta/D)^{\delta_1}(1 - d/D)^{\delta_2}$

Table 3.1: Common dose response-models and their standardized form (from Xun and Bretz (2017)). Here $B(\delta_1, \delta_2) = (\delta_1 + \delta_2)^{\delta_1+\delta_2}/(\delta_1^{\delta_1}\delta_2^{\delta_2})$.

resulting from an assumed underlying dose-response shape, Bretz et al. (2005) show, that contrast coefficients with optimal power for the single contrast tests are proportional to

$$c_i = n_i(\mu_i - \bar{\mu}), \text{ for } i = 1, \ldots, t,$$

where $\bar{\mu} = \sum_{i=1}^{t} n_i\mu_i / \sum_{i=1}^{t} n_i$. A unique solution is given after normalization as $\boldsymbol{c}/\|\boldsymbol{c}\|$.

Since the main idea of *MCP-Mod* is to take model uncertainty into account, instead of performing just a single contrast test, contrast tests are performed for all Z candidate models simultaneously. From each candidate model $m = 1, \ldots, Z$ a resulting mean response vector $\boldsymbol{\mu_m} = (\mu_{m1}, \ldots, \mu_{mt})$ can be obtained, which can then be used to derive optimal contrast coefficients under this dose-response shape. The optimal contrast coefficients provided above are invariant to any change in shift or scale of the mean response vectors, so that the standardized models f^0 are sufficient for deriving optimal coefficients. For the parameter vector $\boldsymbol{\theta^0}$ guesses have to be made, based on prior knowledge. Using these *guesstimates*, optimal contrast vectors $\boldsymbol{c_m} = (c_{m1}, \ldots, c_{mk})'$ for each of the candidate models can be derived.

For each candidate model m a test of $H_0^{(m)} : \boldsymbol{c_m'}\boldsymbol{\mu} = 0$ versus $H_1^{(m)} : \boldsymbol{c_m'}\boldsymbol{\mu} > 0$ is performed using test statistics of the form

$$T_m = \frac{(\boldsymbol{c_m})'\bar{\boldsymbol{y}}}{\sqrt{S^2 \sum_{i=1}^{t} c_{mi}^2/n_i}} \quad \text{for } m = 1, \ldots, Z,$$

where $\bar{\boldsymbol{y}} = (\bar{y}_1, \ldots, \bar{y}_t)'$ is the vector of observed means at the dose levels and $S^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(\sum_{i=1}^{t} n_i - t)$ is the pooled variance estimator. Univariately, each of these test statistics follows a t-distribution with $\sum_{i=1}^{t} n_i - t$ degrees of freedom under the null hypothesis.

When performing multiple contrast tests, multiplicity has to be taken into account. *MCP-Mod* uses the joint distribution of the Z test statistics and takes into account correlations between the tests. Under the global null hypothesis $\boldsymbol{H_0}$: $H_0^{(1)} \cap \cdots \cap H_0^{(Z)}$ the test statistics $T_1, \ldots, T_Z$ jointly follow a multivariate t-distribution with $\sum_{i=1}^{t} n_i - t$ degrees of freedom and a correlation matrix, which depends on the sample sizes on each dose level and the contrast coefficients.

The joint distribution can then be used to obtain a multiplicity-adjusted critical value $q_{1-\alpha}$ for a given α, so that $P_{\boldsymbol{H_0}}(T_{max} \leq q_{1-\alpha}) = P_{\boldsymbol{H_0}}(T_1 \leq q_{1-\alpha}, \ldots, T_Z \leq q_{1-\alpha}) = 1 - \alpha$, where $T_{max} = max_m T_m$. If $T_{max} > q_{1-\alpha}$, a significant dose-response signal has been established and the null hypothesis can be rejected. Generally, all candidate shapes for which $T_m > q_{1-\alpha}$ can then be declared as significant. This multiple comparison procedure guarantees control of the family-wise error rate (FWER), e.g. the probability to falsely reject at least one true null hypothesis is controlled at significance level α.

If a significant dose-response signal could be established in the MCP step the method continues to the Mod step. Several possible options for modeling dose-response are available. A single dose-response model out of the set of significant candidate models could be used. This can either be the model with the largest test statistic or the model with the best fit according to a model selection criterion like AIC. Model averaging over all selected models is another alternative. Based on the resulting dose-response curve, a dose for a possible Phase III trial can then be chosen, for example by estimating one of the doses described in Section 3.1.

Chapter 4

Overview and discussion of contributed articles

4.1 Overview

The four contributed articles, that form the main part of this cumulative thesis, deal with different aspects of subgroup analyses.

In Thomas and Bornkamp (2017) we consider the problem of treatment effect estimation in subgroups in the two-arm trial setting. As discussed in Chapter 2, treatment effect estimates in selected subgroups often suffer from selection bias. We discuss several ways to obtain adjusted treatment effect estimates with reduced bias and compare them in an extensive simulation study. Our results show, that several of the considered methods provide treatment effect estimates with strongly reduced selection bias and provide confidence intervals, which give close to nominal coverage.

In Thomas et al. (2018c) and Thomas et al. (2018a) we propose approaches for identification of subgroups and treatment effect heterogeneity in dose-finding trials. While a large number of methods have been proposed for subgroup identification in the literature (see Chapter 2), they generally target two-arm trials. The setting of dose-finding trials is quite different, since underlying dose-response relationships have to be considered and the used models are often non-linear (see Chapter 3), so that methods for two-arm trials can not

be easily transferred to dose-finding settings. The first approach, proposed in Thomas et al. (2018c), is based on the model- based recursive partitioning algorithm by Zeileis et al. (2008), which was briefly discussed in Section 2.1.3 in Chapter 2. The algorithm tries to detect covariate effects on the parameters of dose-response models. The proposed approach produces a tree of subgroups, in each of which a separate dose-response model is fit. As an alternative approach, in Thomas et al. (2018a) we consider Bayesian hierarchical models using shrinkage priors to model covariate effects on dose-response parameters. This approach is conceptually similar to penalized regression approaches proposed for two-arm trials, which were introduced in Chapter 2.

Thomas et al. (2018b) again considers subgroup analyses in a dose-finding setting, but considers a more confirmatory situation. Here we assume, that a subgroup has been identified before the start of the trial and we want to test for a significant dose-response signal in multiple populations, for example in the subgroup and the full population. In a confirmatory setting the aim is usually to control the family-wise error rate, the probability to reject at least one true null hypothesis, at the desired α-level. For this purpose we extend the *MCP-Mod* Methodology method, which was introduced in Chapter 3, to trials with multiple populations.

In the following we provide a short synopsis for each article, in which we highlight the main contributions.

4.1.1 Treatment effect estimation for subgroups in clinical trials

Contributed material

Thomas, M. and Bornkamp, B. (2017): Comparing approaches to treatment effect estimation for subgroups in early phase clinical trials. *Statistics in Biopharmaceutical Research*, 9 (2), 160–171

As discussed in Section 2.1.2 in Chapter 2 subgroups are in practice often analyzed using univariate models of type (2.2) and by testing for significant interaction terms in these models. A common identification strategy is to fit such models for all subgroups and check for significant p-values among all interactions without properly taking multiplicity

into account. Subgroups with significant p-values are then investigated further. When subgroups are identified with this procedure, treatment effect estimates in the identified subgroups will suffer from selection bias, especially if the total number of considered subgroups was large. Such naive estimates can therefore lead to over-optimism with regard to the subgroup finding. Since the size of the treatment effect is essential to determine, if the subgroup is clinically relevant, it is important to obtain an adjusted estimate, that does not suffer from selection bias.

In Thomas and Bornkamp (2017) we specifically focus on the issue of obtaining an adjusted treatment effect estimate after subgroup selection. We consider several proposed estimators for treatment effects in subgroups from the literature and perform an extensive simulation study to investigate their ability to reduce the bias, their mean squared error (MSE) and also coverage of confidence intervals. We consider a large number of scenarios with varying sample size, number of subgroups, effect sizes in the subgroup and functional forms for the underlying covariate effects.

In total, we compare five adjusted estimators, which can be divided into three main approaches, model averaging, resampling and penalized regression. The main ideas behind these approaches are provided in the following.

1. **Model averaging (*ma*)**: Bayesian model averaging for treatment effect estimation in subgroups was proposed in Bornkamp et al. (2017), motivated by an earlier paper by Berger et al. (2014). When univariate subgroups model are used, each subgroup has its own model and subgroup selection can essentially be considered as a model selection procedure. The *ma*-estimator aims to provide an adjusted treatment effect estimate in the selected subgroup, which takes model uncertainty of this selection procedure into account. Instead of using a naive estimate for the identified subgroup, which is obtained using just the model for that subgroup, treatment effect estimates are obtained under all considered models. The *ma*-estimator is then given by a weighted average of estimates under all considered models, with weights determined by the posterior model probabilities. The resulting estimate reflects how strong the evidence for the identified subgroup is. If the model for the selected subgroup has a high posterior model probability the final estimate will be close to the naive

estimate. On the other hand, if posterior model probabilities are spread across several models, the final estimate will be shrunken towards the overall treatment effect.

2. **Resampling**: A second class of estimators uses resampling methods, motivated by Sun and Bull (2005), which used resampling approaches to adjust for selection bias in genome-wide association studies. The main idea of these approaches is to perform subgroup identification and treatment effect estimation on separate datasets. The approaches use bootstrapping to create these two datasets. Bootstrap samples are obtained by drawing a sample of size n with replacement from the original dataset. In total B bootstrap samples are generated and for each bootstrap sample the subgroup identification procedure is repeated. Treatment effect estimates for these subgroups are then obtained in the corresponding out-of-bag sample, which consists of all patients, which did not end up in the bootstrap sample. The difference between the size of the treatment effect in the bootstrap sample and in the out-of-bag sample selection can be used to gain information about the amount of selection bias introduced through the identification procedure. Three adjusted estimators based on resampling are compared in the simulation study:

   (a) ***rs632***, the *0.632-estimator* first proposed in Efron (1983).

   (b) ***rsbias***, which subtracts the selection bias averaged over the bootstrap samples from the original naive estimate.

   (c) ***rsma***, which is similar to the model averaging approach and in each bootstrap sample estimates the treatment effect for the selected subgroup using the best model for that bootstrap sample (selected via BIC).

3. **Penalized regression**: An estimator based on lasso regression (Tibshirani (1996)) is included as well. For this estimator a multivariate model, as model (2.3) in Section 2.1.2 is fit to the data, using lasso penalties on covariate effects. A treatment effect estimate for the selected subgroup can be obtained by averaging all individual treatment effect predictions for patients in the subgroup obtained from this model. In contrast to the other treatment effect estimators described above, the lasso estimator explicitly models the effect of the underlying baseline covariates (instead of dichotomizing them into subgroups).

For all considered estimators above confidence intervals are provided in the paper.

Based on our simulation study the model averaging estimator, and two of the resampling estimators (*rs632* and *rsma*) show good performance. These three estimators all have close to zero biases over all considered scenarios. Similarly, MSE is strongly reduced compared to naive estimates and confidence intervals generally have close to nominal coverage. On the other hand, the lasso estimator and rsbias are both overly conservative and generally over-adjust in scenarios with existing subgroups, which leads to negatively biased estimates.

4.1.2 Subgroup identification in dose-finding trials via model-based recursive partitioning

Contributed material

Thomas, M., Bornkamp, B. and Seibold, H. (2018c): Subgroup identification in dose-finding trials via model-based recursive partitioning. *Statistics in Medicine*, 37 (10), 1608–1624

In Section 2.1.3 in Chapter 2 an overview over the recent literature on subgroup identification was provided. While the literature on the topic is quite extensive, almost all proposed approaches focus on clinical trials with two arms, where patients are administered either one dose of the new treatment or a control. The exploratory subgroup analyses considered in this thesis are commonly performed in Phase II, where also dose-finding studies take place. However subgroup identification for dose-finding trials has, to our knowledge, previously not been considered in the literature. Subgroup identification approaches designed for two-arm trials can generally not be directly transferred to settings with multiple doses, since they do not account for the underlying dose-response relationship. In addition many dose-response models are non-linear (see Section 3.2), which provides an additional challenge.

In Thomas et al. (2018c) we transfer the subgroup identification problem to dose-finding trials and propose a possible approach to identify subgroups in this setting. We combine the subgroup models described in Section 2.1.1 in Chapter 2 with the dose-response models

presented in Chapter 3 and use dose-response models, in which the parameters of the model can depend on covariates. We consider E_{max} models of the form

$$\mu(d, \boldsymbol{\theta}) = E_0(\boldsymbol{X}) + E_{max}(\boldsymbol{X}) \frac{d^h}{ED_{50}(\boldsymbol{X})^h + d^h}, \tag{4.1}$$

focusing however on the special case for $h = 1$ (the hyperbolic E_{max} model), since sigmoid E_{max} model proved to be hard to fit with the tree-based approach we consider in this paper. Using the terminology introduced in Section 2.1, we can consider covariates with effects on E_0 to be prognostic, while covariates with effects on E_{max} or ED_{50} can be considered to be predictive, since they lead to differential treatment effect curves $E_{max}(\boldsymbol{X}) \frac{d^h}{ED_{50}(\boldsymbol{X})^h + d^h}$. As for subgroup identification in two-arm trials, we are mostly interested in identifying predictive covariates.

While subgroup analyses for two-arm trials are generally only concerned with efficacy of the treatment and aim to identify subgroups with large treatment effects, in dose-finding trials we are additionally interested in dose estimation. Therefore, instead of only considering subgroups with increased treatment effects (e.g. with higher E_{max}) we could also aim at identifying subgroups of patients, which require different doses of the treatment. Figure 4.1 visualizes the two scenarios, where subgroups have different E_{max} or different ED_{50}, leading to subgroups with differences in maximum treatment effects or with differences in the steepness of the dose-response curve. If we estimate target doses in the latter scenario, these could possibly differ greatly between the three subgroups.

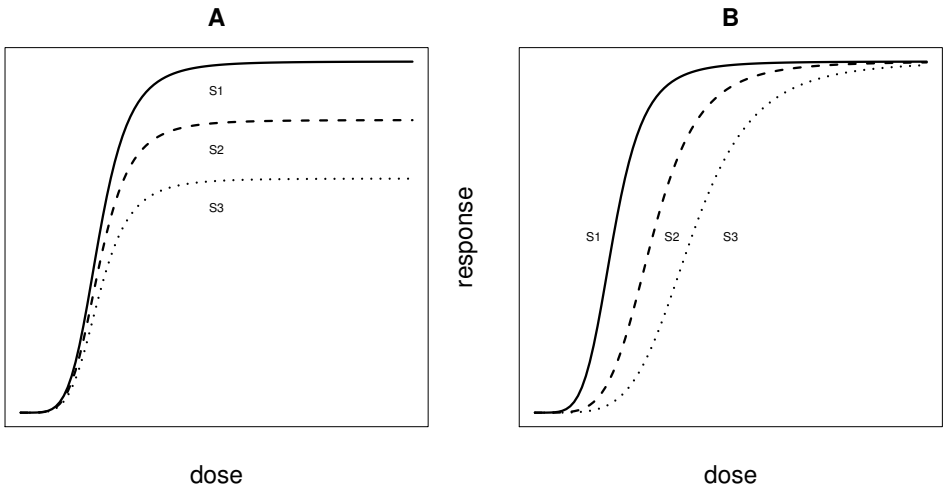

Figure 4.1: Exemplary dose-response shapes in three subgroups S1-S3. **A** shows subgroups with different E_{max}, **B** shows subgroups of with different ED_{50}.

In this paper we propose to use *model-based recursive partitioning* (*MOB*) for subgroup identification in dose-finding trials. *Model-based recursive partitioning* was first introduced in Zeileis et al. (2008) and was applied to subgroup identification in two-arm trials in Seibold et al. (2016). The method belongs to the class of recursive partitioning algorithms discussed in Section 2.1.3 and shares the advantages of these approaches, namely ability to deal with interactions, model non-linearities and additionally good interpretability. In contrast to other recursive partitioning methods, which are mostly non-parametric, *MOB* is in addition, as the name suggests, model-based. This makes it a particular attractive approach in the dose-finding situations, because it allows us to use dose-response models in the recursive partitioning algorithm, so that we can model covariate effects on the dose-response parameters as in model (4.1). *MOB* has previously only been applied to generalized non-linear models and investigating the feasibility of this approach for non-linear dose-response models is a key contribution of the article.

MOB's algorithm, in the configuration that we propose for subgroup identification, tries to identify covariate effects on the model parameters by testing for parameter instabilities of the dose-response parameters E_{max} or ED_{50}. Formally, the hypotheses of independence between the partial scores and the covariates are tested. The partial score function is the first partial derivative of the objective function, for example $\psi_{E_{max}}((y, d, \boldsymbol{x}), \hat{\boldsymbol{\varphi}}) = \frac{\partial \Psi((y,d,\boldsymbol{x}),\hat{\boldsymbol{\varphi}})}{\partial \boldsymbol{E_{max}}}$ is the partial score function for E_{max}, where $\boldsymbol{\varphi}$ is the parameter vector of the dose-response model and Ψ is the model's objective function (for example the residual sum-of-squares for normally distributed data). The algorithm tests the null hypotheses

$$H_0^{(j,p)} : \psi_p((y, d, \boldsymbol{x}), \hat{\boldsymbol{\varphi}}) \quad \perp \quad x^{(j)}, \quad j = 1, \ldots, k, \quad p \in \{E_{max}, ED_{50}\}.$$

The intersection of these null hypotheses is the global null of no parameter instability. The hypotheses are tested using M-fluctuation tests, which aim to detect structural changes in the scores (see Zeileis et al. (2008) and Zeileis and Hornik (2007) for details). To control for multiplicity, Bonferroni corrections are used. If covariate effects are detected using this approach the data is partitioned into two subgroups. This procedure is recursively repeated, until a stopping criterion is reached. The final output is a tree, with separate dose-response parameter estimates in each node of the tree, an example is shown in Figure 4.2.

In simulation studies the performance of the proposed approach is assessed. In these

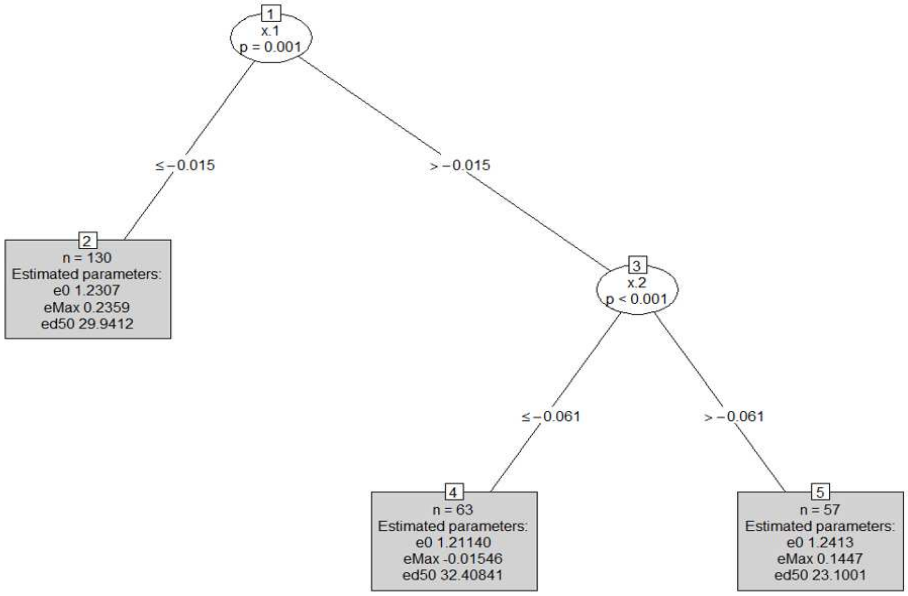

Figure 4.2: Example output for the *MOB* approach with E_{max} models.

studies *MOB* shows good performance with regard to subgroup identification and selection of predictive covariates. We also consider linear models (for example splines) as an alternative to the non-linear E_{max} models, however E_{max} models show good performance in all scenarios and there seems to be no benefit, when using linear models. In addition to the simulation studies we also apply the proposed method to a real Phase II dose-finding trial to illustrate its use in practice.

4.1.3 Identifying treatment effect heterogeneity in dose-finding trials using Bayesian hierarchical models

Contributed material

Thomas, M., Bornkamp, B. and Ickstadt, K. (2018a): Identifying treatment effect heterogeneity in dose-finding trials using Bayesian hierarchical models. *submitted for publication*. Preprint: arXiv:1811.10488

In Thomas et al. (2018c), which was discussed in the previous section, we introduced subgroup analyses in the context of dose-finding trials and presented model-based recursive

partitioning as a possible approach to identify subgroups in dose-finding settings. In this article we propose an alternative approach, which is partially motivated by similar penalized regression approaches proposed in the subgroup analyses literature for two-arm trials. As in the previous section, one of the key contributions of this article is the extension of such methods to the dose-finding setting.

The recursive partitioning approach we proposed for subgroup identification in dose-finding trials in Thomas et al. (2018c), which was described in Section 4.1.2 is easily interpretable, straightforward to use and is generally a good approach to identify relevant predictive covariates as well as subgroups. However, the dose-response models in the trees are fit separately in each terminal node, without borrowing information across subgroups or taking the uncertainty over the whole identification procedure into account. Thus dose-response models obtained from the resulting trees often have small precision, since only the patients in the subgroup are used. While recursive partitioning is suitable, when the main aim of the analysis is to identify subgroups, it might not be optimal, when one is also interested in estimating dose-response curves in the subgroups, for example to estimate target doses (see Chapter 3).

In Thomas et al. (2018a) we propose an alternative approach using Bayesian hierarchical dose-response models, which is able to identify relevant predictive covariates and subgroups as *MOB* but with improved dose-response estimation. We achieve this by fitting a single hierarchical dose-response model, which estimates underlying covariate effects and allows borrowing of information across possible subgroups. We propose shrinkage priors on covariate effects to prevent issues with overfitting, which are likely to occur, when multivariate models with many covariates are considered (see also Section 2.1.2 in Chapter 2). In addition we also incorporate dependencies between prognostic and predictive effects in our model.

For this article we consider sigmoid E_{max} models (see Table 3.1 in Chapter 3) with linear covariate effects of the form

$$\begin{aligned} E_0 &= \alpha_{E_0} + \beta_1 x^{(1)} + \cdots + \beta_k x^{(k)} \\ E_{max} &= \alpha_{E_{max}} + \gamma_1 x^{(1)} + \cdots + \gamma_k x^{(k)} \\ log(ED_{50}) &= \alpha_{ED_{50}} + \delta_1 x^{(1)} + \cdots + \delta_k x^{(k)}. \end{aligned} \tag{4.2}$$

We choose flat priors for α_{E_0}, $\alpha_{E_{max}}$ and σ and functional uniform priors (see Bornkamp (2014) for details) for the non-linear parameters $\alpha_{ED_{50}}$ and h. If flat priors would also be chosen for the covariate effects $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, overfitting would be likely to occur. To avoid this we propose using shrinkage priors, such as the horseshoe (Carvalho et al. (2010)). In the proposed model, covariate effects on each of the dose-response models have horsehoe priors

$$\begin{aligned} \beta_j &\sim N(0, \tau_\beta^2 {\lambda_j^{(prog)}}^2), \quad j = 1, \ldots, k \\ \gamma_j &\sim N(0, \tau_\gamma^2 {\lambda_j^{(pred)}}^2), \quad j = 1, \ldots, k \\ \delta_j &\sim N(0, \tau_\delta^2 {\lambda_j^{(pred)}}^2), \quad j = 1, \ldots, k. \end{aligned} \tag{4.3}$$

The prior variance for each of the coefficients is a mixture of a local component λ and a global component τ. A priori $\lambda_j^{(prog)}$ and $\lambda_j^{(pred)}$ follow positively bounded Cauchy ($C^+(0,1)$)-distributions and determine how strongly the prognostic and predictive effect of a covariate is shrunken to zero. The global components $\tau_\beta, \tau_\gamma, \tau_\delta$ determine the overall amount of shrinkage. The priors for these global components determine how many covariates can be expected to have non-zero effects a priori. The choice of priors for these parameters is specific to each analysis and should take the total number of considered covariates into account. In the paper we give some guidance on the choice of these priors.

The mixture of local and global components allows the horseshoe to clearly separate spurious covariate effects, which are strongly shrunken to zero, from large effects, which can escape the shrinkage through the local components. As a result, horseshoe priors have shown good performance in a number of scenarios and often outperform other common shrinkage priors, like the Bayesian lasso or the spike-and-slab (Carvalho et al. (2010); Polson and Scott (2010)). For our model we compare horseshoe priors to spike-and-slab priors in a simulation study. The results suggest, that horseshoe priors give a more consistent performance over the scenarios we consider.

An additional key aspect of our model is the incorporation of dependencies between prognostic and predictive effects. There are some arguments, why this model specification might be preferable to completely independent priors. First, from a modeling perspective, increasing the probability of prognostic and predictive effects occuring together, helps to clearly distinguish prognostic and predictive effects and avoids possible bias in the size of the predictive effect. Second, from a biological perspective, a covariate, which has already

been identified as predictive, could be considered more likely to be prognostic as well.

For this purpose, as an alternative to independent $C^+(0,1)$-priors for the local shrinkage components proposed previously, we also consider alternative dependent priors, which make it more likely to include a covariate as prognostic, when it is already included as predictive. We choose priors such that marginal probabilities to include predictive coefficients are unaffected. We achieve this by using modified priors for the local shrinkage components of the horseshoe in (4.3),

$$\begin{aligned} \lambda_j^{(*)} &\sim C^+(0,1), \quad j = 1, \ldots, k \\ \lambda_j^{(pred)} &\sim C^+(0,1), \quad j = 1, \ldots, k \\ \lambda_j^{(prog)} &= max(\lambda_j^{(*)}, \lambda_j^{(pred)}), \quad j = 1, \ldots, k. \end{aligned}$$

With these priors prognostic effects are not shrunken more than predictive effects. In our simulation studies the model with dependent priors outperforms the independent prior model in most scenarios.

We compare the proposed Bayesian hierachical modeling approach to the recursive partitioning approach from Thomas et al. (2018c) in simulation studies. In most considered scenarios the Bayesian approach shows similar performance with regard to subgroup identification and identification of predictive covariates as *MOB*. However, the estimation of individual dose-response curves is much improved compared to mob. In Section 4.2 the differences between the two methods are discussed in greater detail.

4.1.4 A multiple comparison procedure for dose-finding trials with subpopulations

Contributed material

Thomas, M., Bornkamp, B., Posch, M. and König, F. (2018b): A multiple comparison procedure for dose-finding trials with subpopulations. *submitted for publication*. Preprint: arXiv:1811.09824

The methods presented in the previous sections deal with exploratory subgroup analyses, where the focus generally lies on identifying a promising subgroup from a large set of candidate subgroups or based on a large number of baseline covariates. In Thomas et al.

(2018b) we consider a different situation, where knowledge about a possible subgroup exists ahead of a clinical trial. If this trial is a dose-finding trial it might be of interest to perform tests for a significant dose-response signal in both the subgroup and the full population. Such a scenario leads to issues of multiplicity, which have to be taken into account.

In Chapter 3 *MCP-Mod* (Bretz et al. (2005); Pinheiro et al. (2014)) was introduced as a methodology for the analyses of dose-finding trials. *MCP-Mod* consists of two steps, the MCP step, which focuses on establishing a significant dose-response signal and the Mod step, which fits a dose-response model for dose-estimation if a significant signal has been established in the MCP step. As discussed in Section 3.3, *MCP-Mod* performs contrast tests for several possible candidate dose-response models and aims to establish a dose-response signal for at least one of the candidate models.

In Thomas et al. (2018b) we extend the MCP part of *MCP-Mod* to settings, where dose-response signals are tested in multiple populations, for example in the full population and a subgroup. With the presented extension it is then possible to test for a dose-response signal in multiple populations, while controlling family-wise error rate over the whole testing procedure. As the standard *MCP-Mod*, the proposed extension uses the joint distribution of test statistics to take the correlation between tests in subgroup and full population into account. The approach therefore falls under the class of parametric approaches to control multiplicity, which were introduced in Section 2.2 in Chapter 2.

In the following we briefly summarize the proposed extension using a similar notation as in Section 3.3, however we introduce population indices to refer to the different populations. We consider a situation, where a subgroup S is prespecified and denote the complement of this subgroup as C. In addition we consider the full population $F = S \cup C$.

For each candidate model m and each population P a contrast test of $H_0^{(P,m)} : \boldsymbol{c_m}' \boldsymbol{\mu^{(P)}} = 0$ against the alternative $H_1^{(P,m)} : \boldsymbol{c_m}' \boldsymbol{\mu^{(P)}} > 0$ is performed, where $\boldsymbol{\mu^{(P)}}$ is the vector of mean responses in population P. We define the *population null hypothesis (in P)* as

$$H_0^{(P)} : H_0^{(P,1)} \cap ... \cap H_0^{(P,Z)},$$

the intersection of the null hypotheses for all Z candidate models in population P. The *global null hypothesis* is then the intersection between the population null hypotheses,

in all tested populations. For example for the full testing strategy, which includes both subgroup and complement, as well as the full population, the global null hypothesis is

$$H_0^{(global)} : H_0^{(F)} \cap H_0^{(S)} \cap H_0^{(C)}.$$

While rejecting the global null is of main interest, the population hypotheses have to be considered to determine if a dose-response trend has been established in all populations or, for example, only in a subgroup.

The test statistics for $H_0^{(P,m)}$ vs $H_1^{(P,m)}$ are contrast tests of the form

$$T_m^{(P)} = \frac{(\boldsymbol{c_m})' \bar{\boldsymbol{Y}}^{(P)}}{\sqrt{(\hat{\sigma}^{(P)})^2 \sum_{i=1}^{t} c_{mi}^2 / n_i^{(P)}}}, \quad m = 1, \ldots, Z; \quad P = F, S, C.$$

Here $\bar{\boldsymbol{Y}}^{(P)} = (\bar{Y_1}^{(P)}, \ldots, \bar{Y_t}^{(P)})'$ is the vector of estimated dose means and $\hat{\sigma}^{(P)}$ is an estimator of the standard deviation in population P.

With our extension we are not only testing for multiple possible candidate shapes, but also in multiple populations. We still want to control the family-wise error rate (FWER), i.e. the probability to falsely reject $H_0^{(global)}$, at the nominal level of α. Similar to the standard *MCP-Mod* we use a parametric approach to control for multiplicity. We consider here a full testing strategy, for which tests are performed in all populations, e.g. in F, S and C, as the most complete case. Simpler testing strategies with tests only in F and S can easily be derived from the formulas below. For the full testing strategy and assuming normally distributed data and homoscedasticity across populations the vector of test statistics
$\boldsymbol{T'} = (T_{M_1}^{(F)}, \ldots, T_{M_Z}^{(F)}, T_{M_1}^{(S)}, \ldots, T_{M_Z}^{(S)}, T_{M_1}^{(C)}, \ldots, T_{M_Z}^{(C)})$ follows a multivariate t-distribution with $\sum_{P \in V} (\sum_{i=1}^{k} n_i^{(P)} - k)$ degrees of freedom, mean vector $\mathbf{0}$ and correlation matrix $\boldsymbol{R}$. The correlation matrix has the form

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R_{FF}} & \boldsymbol{R_{FS}} & \boldsymbol{R_{FC}} \\ \boldsymbol{R'_{FS}} & \boldsymbol{R_{SS}} & \boldsymbol{R_{SC}} \\ \boldsymbol{R'_{FC}} & \boldsymbol{R'_{SC}} & \boldsymbol{R_{CC}} \end{bmatrix}.$$

Each of the 9 submatrices is of dimension $Z \times Z$ and includes the correlations between the contrast tests in the corresponding populations. The correlations depend on the considered candidate models and the overlap between the populations. Based on this

exact joint distribution critical values can be provided, using the approach outlined in Section 3.3, which guarantees FWER-control for this testing procedure.

In heteroscedastic scenarios, where variances in subgroup and complement differ, the exact joint distribution cannot be obtained, since the degrees of freedom for each single test statistic depend on the population and thus the tests statistic no longer jointly follow a multivariate t-distribution. For large sample sizes multivariate normal distributions can be used as an approximation, however for small sample sizes this will lead to a liberal testing procedure with inflation of the FWER. We investigate an alternative approximation to the joint distributions, originally proposed by Hasler (2014). We show, that this approximation controls the FWER for our testing procedure at the nominal level for very small sample sizes, where the multivariate normal approximation leads to large FWER-inflation.

In a simulation study we investigate the power of our multi-population testing procedure compared to the standard *MCP-Mod*, which only considers a single population (the full population F). Depending on the prevalence of the subgroup and the difference in treatment effect sizes, as well as variances, between subgroup and complement, the proposed multi-population test can greatly increase power over the standard *MCP-Mod*.

4.2 Discussion and outlook

The four articles summarized above focus on different aspects, that are related to subgroup analyses in clinical trials.

In Thomas and Bornkamp (2017) (Section 4.1.1) we focus on treatment effect estimation in subgroups after a subgroup has been identified. In addition we consider two-arm trials and not dose-finding trials as in the other articles in this thesis. However, there is some overlap with the methods used in Thomas et al. (2018a) (Section 4.1.3), for example we consider treatment effect estimation based on penalized regression in the form of lasso regression, while in Thomas et al. (2018a) a Bayesian penalized regression approach is proposed. The concept of shrinkage is also common to both articles: The model averaging approach we consider in Section Thomas and Bornkamp (2017), which shows great performance

in the simulation studies, shrinks effects in subgroups to the overall treatment effect by averaging over univariate models for different possible subgroups. In similar fashion, the horseshoe priors we use in Thomas et al. (2018a) shrink the effects of covariates on dose-response parameters to zero. This then leads to shrinkage towards the overall average for the specific dose-response parameter. In both cases, there needs to be a strong signal to prevent shrinkage, either in the form of a very large treatment effect (relative to the overall effect) in a subgroup or a very large covariate effect on the dose-response curve.

Several extensions of the work described in Thomas and Bornkamp (2017) are possible. We specifically focused on simple subgroup identification strategies based on interaction p-values, however it could be interesting to also consider more complex identification strategies, for example using one of the approaches described in Section 2.1.3 in Chapter 2. While several resampling approaches were already considered in the simulation study, the recent method by Rosenkranz (2016) is based on a similar idea, but uses a different estimator for treatment effects in subgroups. This new method could be included as an additional competitor in the simulation study.
As an alternative to the lasso estimator, which did not perform well in the simulation study, one could also consider a Bayesian penalized regression approach, using for example horseshoe priors as in Thomas et al. (2018a). The horseshoe shows good performance with regard to treatment effect estimation in the dose-response scenario, it could therefore also be a promising alternative for two-arm trials.

The main contribution of Thomas et al. (2018c) (Section 4.1.2) and Thomas et al. (2018a) (Section 4.1.3) is the development of methods for subgroup identification for dose-finding trials. While many methods have been proposed in the context of two-arm clinical trials (see Chapter 2), to our knowledge no other articles currently exist, which specifically target the dose-finding setting. The methods we propose in these papers take the underlying dose-response relationship into account and can be used with non-linear models, which are commonly used to model dose-response. In addition they can handle the challenges of subgroup analyses, and prevent high false positive rates and overfitting, either through explicit multiplicity adjustments or by using shrinkage or variable selection priors.

The two approaches are conceptually quite different. *MOB*, starting off from a model with no covariate effects, recursively detects covariate effects through an algorithm, resulting

in a tree, while the Bayesian hierarchical model (abbreviated as *BHM* from here on) explicitly models effects of all covariates, but shrinks most of them to zero. Apart from differences in the methodology used, there are also further dissimilarities in the produced outputs and the amount of input required from users. A detailed discussion of these differences, as well as resulting advantages and disadvantages of the two approaches, is provided below.

For *BHM* we use a sigmoid E_{max} model, *MOB* however uses a hyperbolic E_{max} model, i.e. the special case of a sigmoid E_{max} model with $h = 1$ (see Table 3.1 in Chapter 3). *BHM* can therefore fit a larger range of dose-response shapes. Extensions to sigmoid E_{max} models were considered for *MOB*, however estimating the additional parameter proved difficult, when sample sizes in subgroups became too small. In addition we noticed increased rates of false positive identifications. However it is worth noting that the spline models investigated in Thomas et al. (2018c) should be a reasonable alternative, when the hyperbolic E_{max} model does not provide a good fit.

The *MOB*-algorithm is available in the *partykit* package for *R* (Hothorn and Zeileis (2015)), however only for generalized linear models. Custom fitting functions can be defined for non-standard models. We provide an E_{max} fitting function for *MOB* in Thomas et al. (2018c). *MOB*'s output is a tree of subgroups with separate dose-response models in each, which is easily interpretable. However as mentioned previously, the dose-response modeling is performed separately in each of the found subgroups. This leads to reduced precision due to lower sample sizes. In addition confidence intervals will generally not take the uncertainty over the whole search procedure into account.

BHM on the other hand requires more input from the user. The use of Bayesian models makes it necessary to use MCMC sampling tools, as for example *Stan* (Carpenter et al. (2017)) or *JAGS* (Plummer (2017)). Additional input is required with regard to hyper-parameter distributions for the shrinkage priors, for which we provide some guidance in Thomas et al. (2018a). There is however a reward for the additional work, since the output of *BHM* is much richer than for *MOB* and provides posterior distributions for individual parameters of the dose-response model and individual treatment effect curves, which allow for much more detailed statements about uncertainty. Relevant predictive (and prognostic) covariates can be identified and while the method does not directly

suggest promising subgroups as *MOB*, some possible strategies to identify subgroups are outlined in Thomas et al. (2018a).

Summarizing, *MOB*'s main advantages are, that it is possibly easier to use and allows for non-linearities and interactions. It is therefore particularly suitable for a quick search for possible subgroups. *BHM* requires more time for setting up and is more complex, however it provides a much richer output, which provides more information for decision-making. It is therefore more suitable for a detailed analysis, possibly after a first analysis with *MOB* suggested, that there might be subgroups with differential dose-response.

As an avenue for further research on *MOB*, one could consider using a *forest* instead of just a single tree. Similar to random forests (Breiman (2001)), which average over multiple regression trees on bootstrapped datasets, one could consider bagging multiple *MOB* trees to reduce variance in treatment effect estimates in subgroups. While such an approach could possibly be used to close the gap to *BHM* in terms of estimation of individual treatment curves, it would reduce interpretability.

As a possible extension for *BHM* non-linear functions of the covariates could be considered. For example basis function expansions, like splines, could be used to allow for non-linearities. Incorporating such non-linearities and possibly also (selected) covariate-covariate interactions in Bayesian models with shrinkage priors is however not straightforward and would require further investigation. In addition it might be of interest to further investigate possible subgroup identification strategies based on *BHM*. Some strategies are suggested in Thomas et al. (2018a), but it might be useful to compare operating characteristics in a simulation study.

While the other articles in this thesis focus on exploratory subgroup analyses, Thomas et al. (2018b) (Section 4.1.4) considers a more confirmatory setting. An approach like the one we outline in this paper would often be the next step after a possible subgroup was identified in a previous trial, for example using one of the tools we discuss in Thomas et al. (2018c) and Thomas et al. (2018a).

In the contributed article we focus on the MCP step of *MCP-Mod* to develop a testing strategy to control FWER over tests for all candidate shapes and in all populations. We do not extend the Mod step, in which one or multiple dose-response models are fit

and used for dose estimation. A natural extension would therefore be an incorporation of the Mod step, which is however not straightforward. First of all different candidate models could be significant in different populations. In addition it is unclear how the subgroup should be incorporated in the dose-response model. For example dose-response models could be fit only in the subgroup or subgroup effects could be modeled on the dose-response parameters. Model averaging could be considered to take several different dose-response models with and without subgroups into account, using a similar approach as the one proposed in Bornkamp et al. (2017) and utilized in Thomas and Bornkamp (2017).

Bibliography

Alosh, M. and Huque, M. F. (2009): A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*, 28 (1), 3–23.

Alosh, M., Huque, M. F., Bretz, F. and D'Agostino Sr, R. B. (2017): Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Statistics in Medicine*, 36 (8), 1334–1360.

Berger, J. O., Wang, X. and Shen, L. (2014): A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics*, 24, 110–129.

Bornkamp, B. (2014): Practical considerations for using functional uniform prior distributions for dose-response estimation in clinical trials. *Biometrical Journal*, 56, 947–962.

Bornkamp, B. (2017): Dose-finding studies in phase II: Introduction and overview. In: J. O'Quigley, A. Iasonos and B. Bornkamp (Hrsg.) *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-finding Trials*, Kapitel 11, 189–204. CRC Press, Boca Raton, 1st edition.

Bornkamp, B. and Ickstadt, K. (2009): Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, 65, 198–205.

Bornkamp, B., Ohlssen, D., Magnusson, B. P. and Schmidli, H. (2017): Model averaging for treatment effect estimation in subgroups. *Pharmaceutical Statistics*, 16 (2), 133–142.

Bornkamp, B., Pinheiro, J. and Bretz, F. (2013): *DoseFinding: Planning and Analyzing Dose Finding experiments*. R package version 0.9-6.

Breiman, L. (2001): Random forests. *Machine learning*, 45 (1), 5–32.

Breiman, L. (2017): *Classification and regression trees*. Routledge, New York, 1st edition.

Bretz, F., Pinheiro, J. C. and Branson, M. (2005): Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61, 738–748.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W. and Rohmeyer, K. (2011): Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal*, 53 (6), 894–913.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017): Stan: A probabilistic programming language. *Journal of statistical software*, 76 (1).

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010): The horseshoe estimator for sparse signals. *Biometrika*, 97 (2), 465–480.

Chipman, H. A., George, E. I., McCulloch, R. E. et al. (2010): Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4 (1), 266–298.

Dixon, D. O. and Simon, R. (1991): Bayesian subset analysis. *Biometrics*, 47, 871–881.

Dmitrienko, A., Muysers, C., Fritsch, A. and Lipkovich, I. (2016): General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *Journal of Biopharmaceutical Statistics*, 26 (1), 71–98.

Dunn, O. J. (1961): Multiple comparisons among means. *Journal of the American statistical association*, 56 (293), 52–64.

Dunnett, C. W. (1955): A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50, 1096–1121.

Efron, B. (1983): Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78 (382), 316–331.

EMA (2014): *Guideline on the investigation of subgroups in confirmatory clinical trials*. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf` (accessed 17 July 2018).

Foster, J. C., Taylor, J. M. and Ruberg, S. J. (2011): Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30 (24), 2867–2880.

Ghosh, D., Zhu, Y. and Coffman, D. L. (2015): Penalized regression procedures for variable selection in the potential outcomes framework. *Statistics in Medicine*, 34 (10), 1645–1658.

Ginsburg, G. S. and McCarthy, J. J. (2001): Personalized medicine: revolutionizing drug discovery and patient care. *TRENDS in Biotechnology*, 19 (12), 491–496.

Hamburg, M. A. and Collins, F. S. (2010): The path to personalized medicine. *New England Journal of Medicine*, 363 (4), 301–304.

Hasler, M. (2014): Multiple contrast tests for multiple endpoints in the presence of heteroscedasticity. *The International Journal of Biostatistics*, 10 (1), 17–28.

Hawkins, D. M. (2004): The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44 (1), 1–12.

Hochberg, Y. (1988): A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75 (4), 800–802.

Holm, S. (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Hommel, G. (1988): A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75 (2), 383–386.

Hothorn, T. and Zeileis, A. (2015): partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16, 3905–3909.

Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L. and Devanarayan, V. (2017): Patient subgroup identification for clinical drug development. *Statistics in Medicine*, 36 (9), 1414–1428.

Imai, K., Ratkovic, M. et al. (2013): Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7 (1), 443–470.

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A. and Branson, M. (2011): Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 8, 129–143.

Källén, A. (2007): *Computational Pharmacokinetics*. Chapman and Hall, Boca Raton.

Keene, O. N. and Garrett, A. D. (2014): Subgroups: time to go back to basic statistical principles? *Journal of Biopharmaceutical Statistics*, 24 (1), 58–71.

Lagakos, S. W. (2006): The challenge of subgroup analyses - reporting without distorting. *New England Journal of Medicine*, 354 (16), 1667–1669.

Lipkovich, I. and Dmitrienko, A. (2014): Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of Biopharmaceutical Statistics*, 24, 130–153.

Lipkovich, I., Dmitrienko, A. and D'Agostini, R. B. (2017): Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36 (1), 136–196. ISSN 1097-0258.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011): Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30 (21), 2601–2621.

Loh, W.-Y. (2002): Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, 361–386.

Loh, W.-Y., He, X. and Man, M. (2015): A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34 (11), 1818–1833.

McKeegan, E. M., Ansell, P. J., Davis, G., Chan, S., Chandran, R. K., Gawel, S. H., Dowell, B. L., Bhathena, A., Chakravartty, A., McKee, M. D. et al. (2015): Plasma biomarker signature associated with improved survival in advanced non-small cell lung cancer patients on linifanib. *Lung Cancer*, 90 (2), 296–301.

Morita, S. and Müller, P. (2017): Bayesian population finding with biomarkers in a randomized clinical trial. *Biometrics*, 73 (4), 1355–1365.

Oldenhuis, C., Oosting, S., Gietema, J. and De Vries, E. (2008): Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*, 44 (7), 946–953.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N. and Posch, M. (2016): Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of Biopharmaceutical Statistics*, 26 (1), 99–119.

O'Quigley, J., Iasonos, A. and Bornkamp, B. (2017): *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-Finding Trials*. CRC Press, Boca Raton, 1st edition.

Park, T. and Casella, G. (2008): The Bayesian lasso. *Journal of the American Statistical Association*, 103 (482), 681–686.

Pinheiro, J. and Bornkamp, B., Bornkamp (2017): Designing phase II dose-finding studies: Sample size, doses, and dose allocation weights. In: J. O'Quigley, A. Iasonos and B. Bornkamp (Hrsg.) *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-finding Trials*, Kapitel 13, 229–246. CRC Press, Boca Raton, 1st edition.

Pinheiro, J. C., Bornkamp, B. and Bretz, F. (2006): Design and analysis of dose finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics*, 16, 639–656.

Pinheiro, J. C., Bornkamp, B., Glimm, E. and Bretz, F. (2014): Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, 33, 1646–1661.

Plummer, M. (2017): Jags version 4.3. 0 user manual [computer software manual]. *Retrieved from sourceforge. net/projects/mcmc-jags/files/Manuals/4. x*.

Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002): Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine*, 21 (19), 2917–2930.

Polson, N. G. and Scott, J. G. (2010): Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9, 501–538.

Rosenkranz, G. (2016): Exploratory subgroup analysis in clinical trials by model selection. *Biometrical Journal*, 58, 1217–1228.

Royston, P., Altman, D. G. and Sauerbrei, W. (2006): Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25 (1), 127–141.

Ruberg, S. J. and Shen, L. (2015): Personalized medicine: four perspectives of tailored medicine. *Statistics in Biopharmaceutical Research*, 7 (3), 214–229.

Seibold, H., Zeileis, A. and Hothorn, T. (2016): Model-based recursive partitioning for subgroup analyses. *International Journal of Biostatistics*, 12 (1), 45–63.

Sivaganesan, S., Müller, P. and Huang, B. (2017): Subgroup finding via Bayesian additive regression trees. *Statistics in Medicine*, 36 (15), 2391–2403.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M. and Li, B. (2009): Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10 (Feb), 141–158.

Sun, L. and Bull, S. B. (2005): Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, 28, 352–367.

Thomas, M. and Bornkamp, B. (2017): Comparing approaches to treatment effect estimation for subgroups in early phase clinical trials. *Statistics in Biopharmaceutical Research*, 9 (2), 160–171.

Thomas, M., Bornkamp, B. and Ickstadt, K. (2018a): Identifying treatment effect heterogeneity in dose-finding trials using Bayesian hierarchical models. *submitted for publication*. Preprint: arXiv:1811.10488.

Thomas, M., Bornkamp, B., Posch, M. and König, F. (2018b): A multiple comparison procedure for dose-finding trials with subpopulations. *submitted for publication*. Preprint: arXiv:1811.09824.

Thomas, M., Bornkamp, B. and Seibold, H. (2018c): Subgroup identification in dose-finding trials via model-based recursive partitioning. *Statistics in Medicine*, 37 (10), 1608–1624.

Thomas, N. (2006): Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose designs. *Journal of Biopharmaceutical Statistics*, 16, 657–677.

Thomas, N., Sweeney, K. and Somayaji, V. (2014): Meta-analysis of clinical dose-response in a large drug development portfolio. *Statistics in Biopharmaceutical Research*, 6, 302–317.

Tian, L., Alizadeh, A. A., Gentles, A. J. and Tibshirani, R. (2014): A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109 (508), 1517–1532.

Tibshirani, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Ting, N. (2006): *Dose finding in drug development*. Springer, New York, 1st edition.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. and Drazen, J. M. (2007): Statistics in medicine - reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357 (21), 2189–2194.

Wang, R. and Ware, J. H. (2013): Detecting moderator effects using subgroup analyses. *Prevention Science*, 14 (2), 111–120.

Weinberg, C. R. (1995): How bad is categorization? *Epidemiology*, 345–347.

Woodcock, J. (2007): The prospects for personalized medicine in drug development and drug therapy. *Clinical Pharmacology & Therapeutics*, 81 (2), 164–169.

Xun, X. and Bretz, F. (2017): The MCP-Mod methodology: Practical considerations and the DoseFinding R package. In: J. O'Quigley, A. Iasonos and B. Bornkamp (Hrsg.) *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-finding Trials*, Kapitel 12, 205–228. CRC Press, Boca Raton, 1st edition.

Zeileis, A. and Hornik, K. (2007): Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61 (4), 488–508.

Zeileis, A., Hothorn, T. and Hornik, K. (2008): Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17 (2), 492–514.