Master's Thesis

Investigations on MCP-Mod Designs

Ludwig-Maximilians-Universität München, Institut für Statistik
in cooperation with
Boehringer Ingelheim Pharma GmbH & Co. KG

Author:
Julia Krzykalla

Supervision:
Prof. Dr. Göran Kauermann
(Ludwig-Maximilians-Universität München, Institut für Statistik)

Dr. Frank Fleischer
(Boehringer Ingelheim Pharma GmbH & Co. KG)

Biberach, April 16, 2015

Abstract

The present master's thesis investigates specific applications of the MCP-Mod approach which is a unification of the two approaches typically applied in the matter of dose-finding, the multiple comparison procedures and the modelling of a parametric dose-response function. By the combination of both, one benefits from the advantages of the continuous modelling, but improves the validity of the results by basing the analyses not only on one pre-specified model but on a set of suitable models.

The MCP-Mod approach by Bretz et al. (2005) has been designed for normally distributed outcomes collected in a basic study design. An enhancement by Pinheiro et al. (2014) makes the approach applicable to a broader range of outcome types, particularly for binary endpoints. As a binary data setting is the underlying scenario for the investigations in the practical part of the thesis, a description of this generalized version is as well included.

Furthermore, a third approach is presented which is based on the same idea: the approach by Klingenberg (2009).

The first aim of this thesis is the comparison of the naive application of the original MCP-Mod approach with its generalized version and the Klingenberg approach for the case of a binary endpoint via simulations. Aspects for the comparison are the achieved power, the preservation of the type-I error and the precision of the target dose estimate. The simulations reveal that the first mentioned approach leads to a loss in power and a potential inflation of the type-I error whereas the other two methods show good performances in both, the testing and the estimation part.

Secondly, the thesis investigates two different approaches for the combination of target dose results of separate trials with the aim of obtaining a common dosage proposal if adequate. The first approach is to pool the data of the separate trials and perform the analysis based on the combined data set. For the second approach, the trials are analyzed separately and the results are combined only afterwards. The two approaches are judged by the same criteria as considered in the first part. Simulations show that for an inconvenient combination of trial-specific design aspects, the pooled analysis approach without adjustments may lead to an inflation of the type-I error while the second approach produces good results for all of the investigated aspects. Evidently, the type-I error inflation of the pooled analysis approach can be avoided by adapting the determination of the p-value.

Acknowledgements

I would like to express my gratitude to Prof. Dr. Göran Kauermann and the company Boehringer Ingelheim Pharma GmbH & Co. KG (BI) for giving me the opportunity to write this thesis in cooperation with the Department of Statistics at the Ludwig-Maximilians-University Munich and BI.

Many thanks to Prof. Dr. Göran Kauermann for taking over the supervision of my thesis. I appreciate your openness in terms of the development of the content and your engagement throughout the whole time.

My supervisor from BI, Dr. Frank Fleischer, deserves special thanks for providing me with this extraordinarily interesting, demanding and up-to-date topic and for giving me the chance to write my thesis at BI. Furthermore, I would like to thank you for your great effort and for always taking time to answer my questions. Your excellent supervision, your tireless rereading of my work and all your suggested improvements helped me a lot.

Furthermore, my gratitude to Dr. Tillmann Krahnke for the extensive discussions and his assistance in the implementation.

Moreover, many thanks to Daniela Fischer and Verena Endriß for the critical rereading of my thesis.

In addition, I want to thank the whole Biostatistics group at BI for the warm welcome and the offered support whenever needed.

Finally, I would like to thank my family and my friends for the encouragement and support during the time of studying, particularly during the work on this thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ANOVA	Analysis Of Variance
AZT	Azidothymidine
BI	Boehringer Ingelheim Pharma GmbH & Co. KG
$\mathbf{E}_0$	Placebo Effect
ED	Effective Dose
$\mathbf{E}_{\mathbf{max}}$	Maximum Effect Over Placebo
FDA	Food And Drug Administration
FDR	False Discovery Rate
FWER	Familywise Error Rate
gAIC	Generalized Akaike Information Criterion
GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
GLS	Generalized Least Squares
HIV	Human Immunodeficiency Virus
ICH	International Conference On Harmonization Of Technical Requirements For Registration Of Pharmaceuticals For Human Use
iid	Independent And Identically Distributed
IUM	Intersection-Union Method
IUT	Intersection-Union Test
IWLS	Iteratively Reweighted Least Squares
MCP	Multiple Comparisons Procedure
MCT	Multiple Contrast Test
MED	Minimum Effective Dose
ML	Maximum Likelihood
Mod	Modelling
MSE	Mean Square Error
MTD	Maximum Tolerated Dose
PAVA	Pool-Adjacent-Violator Algorithm
PCER	Per-Comparison Error Rate
pML	Partial Maximum Likelihood
PoA	Proof-of-Activity
PoC	Proof-of-Concept
TD	Target Dose
UIM	Union-Intersection Method
UIT	Union-Intersection Test

Chapter 1

Introduction

The identification of the appropriate dosage is a decisive step in the development and registration process of a new drug. In the past, there have been several examples of drugs that were marketed with an excessive dosage. The Food and Drug Administration (FDA) reports that for 20% of all new molecular entities approved between 1980 and 1999, the labels were changed belatedly. In 79% of these cases, the dosage was decreased, principally for safety reasons (Cross et al., 2002).

A famous example for a belated dose reduction, which was as well thematized by the Oscar-winning movie "Dallas Buyers Club", is Zidovudine, also known as Azidothymidine (AZT). AZT has been the first government-approved drug for the treatment of Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiency Syndrome (AIDS) patients. In the years after the approval, it became obvious that the applied dosage causes serious adverse events, for example severe anaemia or reduction in white blood cell count (AIDSinfo, 2014). Later studies showed that half the original dose is just as much efficacious and far better tolerated so that the dosage recommendation was revised downwards (The Washington Post, December 10, 2013).

But not only an excessive dose can lead to problems. Contrary to this, choosing a dose that is too low can involve the risk of failing to show the efficacy of the substance in a later confirmatory phase. In fact, selecting an inappropriate dose is regarded as one of the main reasons for a considerable high failure rate of clinical phase III trials. This is not only a loss for the pharmaceutical company in the economic sense, it also means that a potentially beneficial substance will never find its way to the patients.

Generally, if a new drug shall be launched onto the market, the newly developed substance has to pass four different clinical phases of testing in humans.

In *Phase I*, the drug is first administered in men, i.e. to healthy human volunteers (except for substances in oncology), to investigate the safety, pharmacodynamics, pharmacokinetics and digestibility of the new substance. Thereby, pharmacodynamics deal with the effects of a drug on the processes in a living organism whilst pharmacokinetics vice versa describe the influences of a living organism on the drug. Additionally, this phase serves for the determination of the dose range to be set up in *Phase II* with respect to efficacy as well as safety. One step to address this matter can be the identification of the Maximum Tolerated Dose (MTD).

For trials conducted in the context of dose finding in *Phase II*, according to Ruberg (Ruberg, 1995), the following questions should be addressed:

- "Is there any evidence of a drug effect?"
  And beyond that, is the efficacy and safety of the drug sufficient to detect a benefit over existing substances in the subsequent phase III trials? This is often referred to as Proof-of-Concept (PoC) or Proof-of-Activity (PoA). For the investigation of this matter, the drug is administered to a limited number of patients that suffer from the disease or condition to treat.

- "What doses are (relevantly) different from control?"
  ... and are therefore remarkable for potential dosages to be recommended.

- "What is the dose-response relationship?"
  Broader as in the previous question, the objective is to identify a functional model for the dose-response relationship. The response variable can thereby either describe the efficacy or the safety of the drug under investigation.

- "What is the optimal dose?"
  The difficulty of this question is that an unambiguous definition for which dose is considered optimal is missing. A possible solution could be the Minimum Effective Dose (MED) or the dose leading to a certain proportion p of the maximal effect to be achieved, the so-called ED$_p$.

Only if the drug shows promising results in this first administration to patients in Phase II, the drug is transferred to the confirmatory stage. Using the dosages identified as reasonable in the previous phase, the trials in *Phase III* are aimed to statistically proof the benefit and safety of the new drug and therewith support the submission for registration to the responsible authorities.
But even when a drug has already passed the registration process, additional trials might be conducted, for example to detect any possible long-term side effects. These studies are referred to as trials of the clinical *Phase IV* or post-marketing trials.

This thesis will concentrate on the process of establishing the dose-response relationship, i.e. on the design and analysis of dose-finding studies in clinical phase II. The expression of "dose-response" is here generally referring to the population average of the dose-response instead of individual dose-response relationships. For these studies, patients are typically randomized into different prespecified dose groups of the substance under investigation or a placebo group. An additional group with an active comparator can optionally be included.
In some cases, a crossover study design can be utilized. Therefor, patients are administered a sequence of dosages in two or more periods of the study. The simplest crossover design is a 2 × 2 where one arm is treated with dose A in period I and dose B in period II while the patients randomized into the second arm take dose B first and dose A in the second part of the study. The advantage of such a design is that it is possible to account for potential trends in the manifestation of the disease such as progression or seasonal variation. Furthermore, each patient serves as its own control and thus, the unexplained variability in the study population can be reduced. However, the efficacy of a multitude of drugs cannot be observed within a short time period and thus, the usage of a crossover design would be (too) time-consuming in the context of dose-finding (Ting, 2006, Section 7.2.2).
Apart from these commonly used study designs, one can also use designs that include possible dose-escalation steps (administration of prespecified doses in an ascending order) or an eventual up- or down-titration (individual adaptation of the administered dose dependent on the observed response or the occurrence of side effects) as well as adaptive designs.
However, this thesis will focus on the classic case of a parallel fixed dose design. The latter designs will not be part of this thesis.

Historically, the matter of finding an optimal dose in later stages of drug development has been addressed by two different methods which both have their deficiencies.

Either the selected dose is the result of multiple comparisons of all doses under investigation against placebo/an active control dose or it is determined with the help of a modelled functional dose-response relationship. The crucial difference between those methods is that the former treats the dose as a qualitative factor while the modelling approach considers the dose as a quantitative factor with regard to the response variable. These characteristics at the same time represent the pitfalls of both methods. Applying a multiple comparison procedure implies the restriction of the appropriate dose to the doses defined in the planning phase of the trial. Furthermore, this method does not take into account possible dependencies between the responses of different doses. On the other hand, the great advantage of this approach is that no assumptions are to be met for the dose-response relation.

In contrast to this, when modelling the relationship by a parametric function, basically every value in the range of the investigated dosages can be identified as the optimal dose. But the validity of the results strongly depends on the choice of an appropriate model to fit to the data.

The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (1994) points out that "what is most helpful in choosing the starting dose of a drug is knowing the shape and location of the population (group) average dose-response curve for both desirable and undesirable effects".

To address this aim and to overcome the drawbacks of the historical approaches mentioned above, Bretz, Pinheiro, and Branson (2005) proposed to combine these methods. The so-called MCP-Mod approach consists of two major steps and enables the user to simultaneously address both aims of phase II trials by using a seamless design.

In a first step, the null hypothesis of a flat dose-response curve is tested for a predefined set of candidate models. If at least one model has a significant test result given the data, a non-flat dose-response curve is established. While doing so, the Familywise error rate (FWER) is controlled by the use of a **M**ultiple **C**omparisons **P**rocedure. Each model for which the corresponding null hypothesis has been rejected can be considered as satisfactory approximation of the true model and is hence included in the reference set for the second step.

In case more than one appropriate model has been identified for the next step, either a model selection procedure has to be set in place or the reference models have to be combined using model averaging techniques.

The resulting model is then fitted to the actual data in the **Mod**elling step and characteristics for the dose-response relationship are estimated, for example the minimum effective dose.

The original paper introducing the MCP-Mod approach only concentrates on normally distributed, homoscedastic outcome measures collected for a single time point in a parallel group design. Pinheiro et al. (2014) enhanced the MCP-Mod approach so that it is also applicable to binary or survival data, repeated measurements or data from crossover studies. Rather practical issues related to the original approach such as sample size determination and sensitivity analyses are considered in the paper of Pinheiro et al. (2006).

All these methods have been implemented in the R package called DoseFinding (Bornkamp et al., 2014).

Another approach for dose-finding in binary data which in its basic idea is similar to the MCP-Mod approach is the one presented by Klingenberg (2009). It also starts with a set of eligible candidate models for binary data to take into account the uncertainty about the true shape of the dose-response relationship. But contrary to the original approach, the identification of the best model(s) is done using a permutation-based test instead of a multiple contrast test. However, this also ensures the control of the FWER. The final model is then obtained by averaging over all models showing a significant dose effect.

The aim of this thesis is to compare the naive application of the MCP-Mod approach for normal data with the enhancement proposed by Pinheiro et al. (2014) and the Klingenberg approach for a binary data setting. The comparison will be done with respect to power evaluations, the preservation of the type-I error and the precision of the target dose estimator via simulations. Furthermore, it investigates different approaches for the combination of the target dose results of two separate studies. These studies are conducted in two different populations that vary in their expected responses and with that also in their dose-response profiles. However, a common recommendation for the dose to administer to those patients is aimed for.

The thesis is structured as follows. The second chapter serves as an introduction to the basic methods for the analysis of dose-finding studies in later phases (Phase II/III). Hereby, both principles that are usually applied, multiple comparison procedures and modelling approaches, are considered. The third chapter introduces the MCP-Mod approach as a hybridization of the previously mentioned principles and deals with the question how to design such a study optimally, that is to determine the dose groups to be used in the trial, as well as how many patients are needed to achieve a certain target power and how they are allocated to the different dose groups. For every step of the procedure, the commands for the practical implementation by means of the statistical software R are given and important options are pointed out. In a second section, the chapter also includes the extension for data that is not normally distributed as well as the related approach of Klingenberg (2009).

In the fourth chapter, the methods presented for binary data are evaluated in terms of power and the preservation of the type-I error via simulations. Furthermore, the precision of the target dose estimate is investigated. Secondly, two different approaches for the combination of two separate studies are presented and their performance is investigated again via simulations. The criteria for this are the same as for the comparison of the methods in the first part of this chapter: power, preservation of the type-I error and the precision of the target dose estimate. In the last chapter, the thesis is summed up and the methods presented are discussed in basic matters as practicality and statistical inference.

Chapter 2

Basic Principles

Before introducing the MCP-Mod approach itself, this chapter addresses some general characteristics of dose-response relationships and then gives a (non-exhaustive) overview of commonly used procedures for dose-finding studies in Phase II/III. These methods cover both, several approaches for multiple comparisons procedures as well as the main methodological aspects of parametric modelling of dose-response relationships. Methods for dose-finding in early phases (as for example $3+3$ designs, up-and-down designs or continual reassessment methods) will not be considered in this thesis. Information about such methods can be found for example in Ting (2006, Chapter 2 & 3), Chevret (2006) and Krishna (2006).

2.1 Dose-Response Relationships

The analysis of dose-response relations is implicitly based on the assumption that the effect of a drug is in a way dependent on the amount of medicine administered to the patient. Thereby, the "effect" has to be an accurately defined (and observable) event which is appropriate to evaluate the severity of the disease or medical condition to treat. It can either be a quantitative measure as the increase or reduction in some clinical value or a qualitative measure, for example the occurrence of an asthmatic attack.

The intuitive assumption on the dose-response relationship is that a pharmacological effect increases monotonically with an increasing dosage and at some dosage achieves saturation, i.e. a level where the dose-response curve plateaus. In some cases, also a subsequent decrease in the response, resulting in an "inverted U-shaped" dose–response pattern, cannot be ruled out. However, when considering a range of doses that are assumed to be therapeutically beneficial, such types of relationships will be rather rare.

Two important characteristics of the dose-response relationship are the Maximum Tolerated Dose (MTD) referring to the dose "which, if exceeded, would put patients at unacceptable risk for toxicity" (Rosenberger and Haines, 2002) and the Minimum Effective Dose (MED) defined by Ruberg (1995) as "the lowest dose producing a clinically important response that can be declared statistically, significantly different from the placebo response". In case of a monotonous relationship, the range between these two measures is called the therapeutic window and contains the values that could be recommended in the label (cf. Figure 2.1). In the case of a U-shaped dose-response curve, the therapeutic window can be narrower than the range between MED and MTD, for example if the favourable dose effect reaches its peak at a dose lower than the MTD.

Another important family of measures that can be used as optimal doses are the ED_p defined as the smallest dose resulting in $p\%$ of the maximum effect E_{max}. Hereby, E_{max} is the maximum effect attributable to the drug which can be derived as the difference between the absolute maximum response

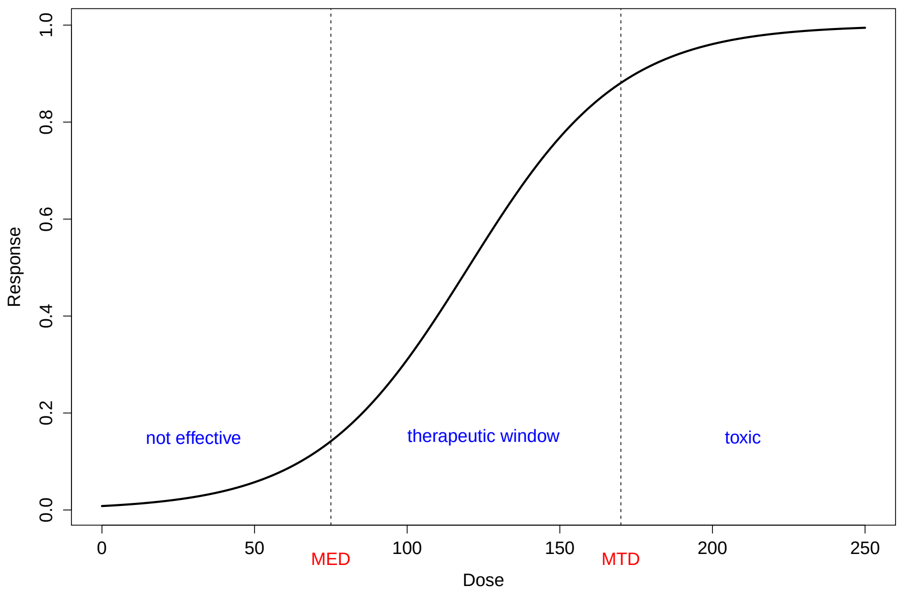


Figure 2.1: Dose-Response Relationship: Dose Ranges

and the placebo effect E_0 (cf. Figure 2.2). A common choice is $p = 50\%$ implying ED_{50} to be the dose that leads to half of the E_{max}.

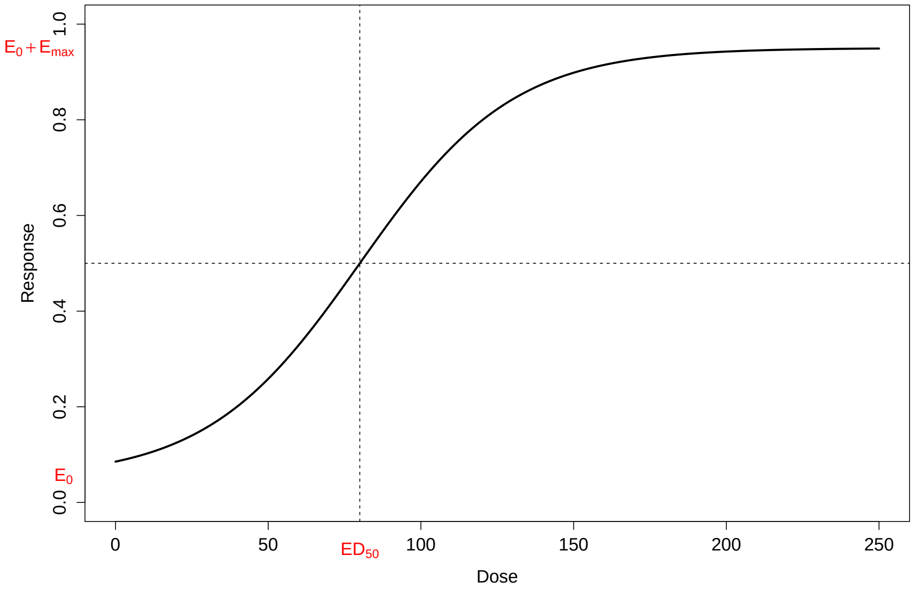


Figure 2.2: Dose-Response Relationship: Characterizing Quantities

For more detailed discussion of dose-response relationships, it is referred to Unkelbach and Wolf (1985, p.4), Ting (2006, Chapter 1) and Senn (1997, Chapter 20).

2.2 Multiple Comparisons

This section contains basic methods for multiple comparison procedures in the application to dose-finding studies. Firstly, possible generalizations of the type-I error for multiple testing procedures are given. The ensuing subsections include the principle of ordered alternatives and other stepwise procedures as well as closed test procedures. Finally, the class of contrast tests is explained in more detail as this is as well the method used in the MCP-Mod approach. The methodology presented in this section can be found in Benjamini and Hochberg (1995), Shaffer (1995), Hsu (1996) and Hochberg and Tamhane (1987).

2.2.1 Generalization of the Type-I Error for Multiple Testing

Generally, two types of errors can occur when conducting a test, either the null hypothesis is rejected although it is true (type-I error) or it cannot be rejected although it is not true (type-II error). Formally, the type-I error (denoted by α) is defined as

$$\alpha = \mathbb{P}(\mathrm{H}_0 \text{ rejected} \mid \mathrm{H}_0 \text{ true}) \ .$$

One of the main issues in testing is to keep the type-I error below a certain designated level which is referred to as the significance level, usually of a value of 2.5%, 5% or 10%.

In the framework of dose-finding, it is usually the case that more than one (pairwise) comparison has to be drawn among the different dose groups. Each of those k comparisons is represented by one null hypothesis $\mathrm{H}_0^i, \ i = 1, \dots, k$. Conducting each of the pairwise comparison tests at the same (local) level α can eventually produce a rate of false positives (meaning erroneously rejected null hypotheses) above this predefined significance level. However, different definitions of error rates for a multiple testing procedure are in place.
The most conservative one is the Familywise error rate (FWER), defined as the probability of committing at least one type-I error, e.g. the probability to erroneously reject any of the k null hypotheses in the whole set of comparisons:

$$\text{FWER} = \mathbb{P}(\# \text{ false positives} > 0) \ .$$

As especially for a huge set of null hypotheses, rejecting one single true hypothesis is more or less unavoidable, it can be more appropriate to consider the False Discovery Rate (FDR), defined as the expected number of true null hypotheses among all that have been rejected

$$\text{FDR} = \mathbb{E}\left(\frac{\# \text{ false positives}}{R} \,\middle|\, R > 0\right) \cdot \mathbb{P}(R > 0)$$

with R being the number of rejected hypotheses.
If all k null hypotheses are correct, the number of false positives equals the number of rejected null hypotheses R. This comes true for two cases: either there are no false positives, then the probability of $\mathbb{P}(R > 0)$ and as a consequence the FDR is zero, or on the other hand, the number of false positives

is not zero implying

$$\mathrm{FDR} = \mathbb{E}\left(\frac{\#\ \text{false positives}}{\#\ \text{false positives}}\right) = \mathbb{E}(1) = 1\ .$$

In this special case, the two error rates are equivalent. Differently spoken, if the FDR is controlled, also the FWER is controlled, but in a weak sense. Strong control is only achieved if the FWER is controlled under all configurations (for the exact definitions, see Hochberg and Tamhane (1987, p.3)). In general, the implication for these two error rates only holds for the opposite direction, i.e. the control of the FWER implies the control of the FDR as the FWER is more conservative.
Another type of definition is the Per-Comparison error rate (PCER) defined as the probability for each hypothesis of committing a type-I error

$$\mathrm{PCER} = \frac{\mathbb{E}(\#\ \text{false positives})}{k}\ .$$

The PCER is the least conservative error rate out of the three presented here. Hence, if one of the others, FWER or FDR is controlled, also the PCER is controlled.

2.2.2 Types of Multiple Comparisons Procedures (MCPs)

Generally, there are four main types of MCPs (Hsu, 1996):

1. All-contrast comparisons: all contrasts (cf. subsection 2.2.6 for the definition of a contrast) are to be tested

2. All-pairwise comparisons: all pairwise differences are to be tested

3. Multiple comparisons with the best: all treatment/dose groups shall be tested against the treatment/dose with the best effect

4. Multiple comparisons with the control: all treatment/dose groups shall be tested against the placebo/active control group

The most common type for dose-finding studies and therefore most thoroughly discussed in this thesis is the last one: multiple comparisons with the control.

2.2.3 Methods Based on Ordered p-Values

The setting for the methods presented in the following is a set of null hypotheses $H_1,\ H_2, \ldots, H_k$ with corresponding p-values $P_1,\ P_2, \ldots, P_k$ that shall be tested at a global significance level α. By sorting them by the size of the p-values, one obtains a list of ordered p-values $P_{(1)} \leq\ P_{(2)} \leq \ldots \leq P_{(k)}$ for the hypotheses $H_{(1)},\ H_{(2)}, \ldots, H_{(k)}$. In the context of dose-finding, each of these hypotheses is representing the comparison of one of the k dose groups with the placebo/active control group, or differently stated, the comparison of the mean responses $\mu_1, \ldots, \mu_k$ in the active dose groups with the mean response μ_0 in the placebo/active control group.

Bonferroni

A simple but effective procedure is to split the global significance level α (equally) between the set of hypotheses such that H_i is rejected if $p_i \le \alpha_i$ with the α_i summing up to 1. This results in a local significance level of $\alpha_i = \frac{\alpha}{k}$ for each of the tests. Alternatively, one could adjust the p-value p_i by multiplying it by the number of hypotheses k. Expressed by the adjusted p-value, H_i is rejected if

$$p_i^* = \min\{1, k \cdot p_i\} \le \alpha \quad \text{for all } i = 1, \ldots, k \ .$$

The order of p-values or tests is not relevant in this case, although if $H_{(i)}$ cannot be rejected, all subsequent hypotheses won't be rejected either.

The so-called (unweighted) Bonferroni method controls the FWER in a strong sense, that is under all configurations.

The main disadvantage of this procedure is that the power for the individual tests decreases with an increasing number of hypotheses to be tested. This is due to the fact that the method ensures the probability of committing at least one type-I error (FWER) to be less or equal to α by augmented critical values. Consequently, higher values for the test statistics are needed to reject the individual null hypotheses which then results in a lower power.

Holm

The Holm procedure (see Holm (1979)) is slightly less conservative than the Bonferroni method and hence more efficient. For this procedure, the ordered p-values $P_{(1)}$, $P_{(2)}, \ldots, P_{(k)}$ are compared sequentially against the adjusted significance levels $\alpha_i = \frac{\alpha}{k-i+1}$. If the p-value is below the corresponding level, $H_{(i)}$ is rejected and the next hypothesis belonging to the next larger p-value is tested. If not, the procedure is stopped and all following hypotheses are considered as not rejectable.

Formulating it by means of adjusted p-values, $H_{(i)}$ is rejected if

$$p_i^* = \min\{1, (k - i + 1) \cdot p_i\} \le \alpha \quad \text{for all } i = 1, \ldots, k \ .$$

The Holm procedure also controls the FWER in a strong sense and therewith also shows a (substantial) loss of power with an increasing number of hypotheses.

Benjamini-Hochberg

A further correction procedure is the method of Benjamini and Hochberg (1995). Contrary to the previous method, the null hypothesis to be tested first is the one with the highest p-value, namely $H_{(k)}$. If the corresponding p-value $p_{(k)} \le \alpha$, all k hypotheses in the set are rejected. If not, $H_{(k)}$ cannot be rejected and the next smaller p-value is taken into consideration. This is repeated until one p-value, $p_{(i)}$ say, stays below the corresponding adjusted α-level of $\alpha_{(i)} = \frac{i}{k}\alpha$. If this is true for one i, all null hypotheses $H_{(j)}$ with $j \le i$ are rejected and the method stops. In other words, the aim is to find the largest i for which the p-value $p_{(i)}$ is smaller than the corresponding $\alpha_{(i)}$.

The method of Benjamini-Hochberg is less conservative than the two presented before. In contrast to them, this method only ensures the control of the FDR instead of the FWER. This strong control holds as long as the test statistics are independent or at least positively dependent (cf. Benjamini and Yekutieli (2001)).

Resampling-Based Step-Down Procedure by Westfall and Young (1993)

More complex methods for the multiplicity adjustment make use of resampling methods. Therefor, the original data set is resampled B times under the global null hypothesis by means of permutation or bootstrap (with replacement). Let the observed p-values for the original data be denoted by $p_1^{\mathrm{obs}},\ p_2^{\mathrm{obs}}, \ldots, p_k^{\mathrm{obs}}$ whereas the p-values for the b-th permutation or bootstrap sample shall be designated by $p_1^{(b)},\ p_2^{(b)}, \ldots, p_k^{(b)},\ b = 1, \ldots, B$.

According to the single-step method presented in Westfall and Young (1993, Section 2.5.2), the adjusted p-value corresponding to the i-th null hypothesis is given by the proportion of permutations for which the minimum p-value is greater than or equal to the observed one:

$$p_i^{\mathrm{adj}} = \frac{1}{B} \sum_{b=1}^{B} I(p_i^{\mathrm{obs}} \leq \min_{j=1,\ldots,k} p_j^{(b)}) \ .$$

Here, $I(\cdot)$ is denoting the indicator function taking the value 1 if $p_i^{\mathrm{obs}} \leq \min\limits_{j=1,\ldots,k} p_j^{(b)}$ is true or 0 else. The use of the minimal p-value for the global test decision is implied by defining the global null hypothesis as the intersection of all single null hypotheses. For detailed explanation it is referred to subsection 2.2.6.

A more powerful approach deduced from the previous method is the step-down procedure by Westfall and Young (1993, Section 2.6)) based on ordered p-values. The procedure starts with the adjustment of the smallest p-value, say p_1^{obs}, by using the minimum p-value distribution like for the single-step adjustment

$$p_1^{\mathrm{adj}} = \frac{1}{B} \sum_{b=1}^{B} I(p_1^{\mathrm{obs}} \leq \min_{j=1,\ldots,k} p_j^{(b)}) \ .$$

But, in contrast to the previous approach, the remaining p-values are no longer adjusted according to the minimum p-value distribution, but according to a reduced set of p-values. This means that all resampling p-values $p_1^{(b)}$ are deleted and the adjustment of the second smallest p-value, say p_2^{obs}, is done based on the minimum p-value distribution of all remaining p-values

$$p_2^{\mathrm{adj}} = \frac{1}{B} \sum_{b=1}^{B} I(p_2^{\mathrm{obs}} \leq \min_{j \neq 1} p_j^{(b)}) \ .$$

The other adjusted p-values are calculated analogously and in an ascending order. After every adjustment step, the corresponding resampling p-values are deleted from the sampled set.
The advantage of using a reduced set of p-values for the adjustment is that also the adjusted p-values are smaller than the ones obtained by the single-step method. Hence, the power is improved.

It was shown that both methods ensure the control of the FWER in a strong sense, i.e. independent of the number of true individual null hypotheses and which ones are true or false. In fact, the strong control is based on the condition of subset pivotality.
Let P be a random vector following a certain distribution and define P_K as an arbitrary subvector of P. The property of subset pivotality is true if the joint distribution of $P_K = \{P_i;\ i \in K\}$ under the global null hypothesis H_0 and the subset of null hypotheses $\cap_{i \in K} H_0^i$ is identical. This must hold for all arbitrary subsets K of true null hypotheses.
Also important to note is that this approach is more efficient than for example the Bonferroni or Holm procedure due to the possibility of taking into account potential correlations between the test statistics.

2.2.4 Partitioning Principle

The basis for this hierarchical testing method (also known as principle of ordered alternatives) presented in the following is a disjoint family of hypotheses. Suppose there are k active doses to be tested versus placebo (dose 0), one could (pre-)define a series of hypotheses as follows:

1. H_{0k}: dose k is ineffective ($H_{0k}:\ \mu_k = \mu_0$),

2. $H_{0(k-1)}$: dose k is effective, but dose $k-1$ is ineffective ($H_{0(k-1)}:\ \mu_k \neq \mu_0 \wedge \mu_{k-1} = \mu_0$),

$\vdots$

i. H_{0i}: doses $i+1, \ldots, k$ are effective, but dose i is ineffective
($H_{0i}:\ \mu_k \neq \mu_0 \wedge \ldots \wedge \mu_{i+1} \neq \mu_0 \wedge \mu_i = \mu_0$),

$\vdots$

k. H_{01}: doses $2, \ldots, k$ are effective, but dose 1 is ineffective
($H_{01}:\ \mu_k \neq \mu_0 \wedge \ldots \wedge \mu_2 \neq \mu_0 \wedge \mu_1 = \mu_0$).

By means of the appropriate tests, a local significance level of α can be applied to all hypotheses ensuring the strong control of the FWER at the same time. Although k hypotheses are tested simultaneously by this procedure, an adjustment for multiplicity is not needed as there is always only one true null hypothesis (at most). However, in some cases the construction of a test for those disjoint hypotheses may be complicated.
A common misconception is that the procedure is based on the assumption of a monotonic dose-response function. But this is not always true as the ordering of the hypotheses is arbitrary. If for example the assumed relationship is that of a quadratic curve, the sequence could be specified as: dose 3, then dose 2, then dose 4, then dose 1 (cf. Figure 2.3).
This approach is very efficient as each of the k hypotheses can be tested with respect to a significance level of α. The disadvantage is that the order of the hypotheses has to be specified in advance. This involves the risk that possibly efficient doses may not be detected due to unfavourable ordering.
For more detailed information of methods using the partitioning principle, see Bretz et al. (2008), Finner and Strassburger (2002) and Ting (2006, Chapter 11).

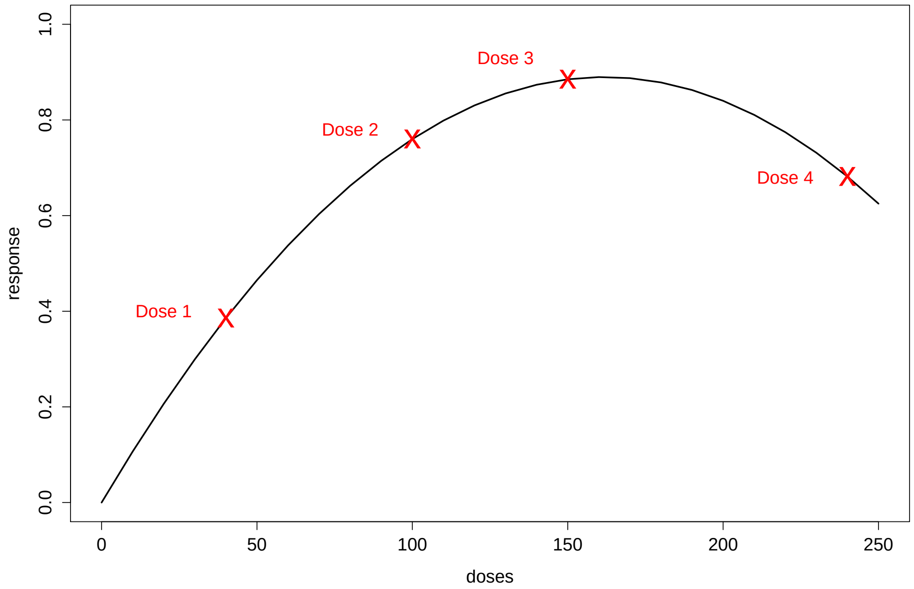


Figure 2.3: Partitioning Principle for a Quadratic Dose-Response Relationship

2.2.5 Closed Testing Procedure

In contrast to the method based on the partitioning principle, the closed testing procedure requires a closed set of hypotheses under investigation. A closed set of hypotheses is a set which contains the hypotheses themselves as well as all their distinct intersections. It is hierarchical as some of the hypotheses are proper components of others. The top of the hierarchy in such a closed set is represented by the intersection of all single hypotheses.

For a closed testing procedure, each hypothesis in the closed set is tested at the (global) significance level α. In order to be able to control the FWER, a null hypothesis of the original set can only be rejected if all hypotheses in the hierarchy that are above the considered one are also rejected. This implies that no hypothesis can be rejected if the hypothesis on the top of the hierarchy does not show a significant test result.

Practical examples and further information on closed testing procedures can be found in Marcus et al. (1976) and Ting (2006, Chapter 11).

2.2.6 Multiple Contrast Tests (MCTs)

The principle of (multiple) contrast tests allows to test a more general set of hypotheses than the methods presented before. Not only pairwise comparisons of an active dose and the control can be addressed, but all types of comparisons listed in subsection 2.2.2. Contrary to the previous methods, the null hypotheses are no longer formulated directly on the basis of the group means themselves but by means of contrasts of these group means. A contrast is a linear combination of the group means $\sum\limits_{i=0}^{k} c_i\mu_i$ with the restriction that all c_i's sum up to 0. One could also express this by the product of $\boldsymbol{c}^\top \boldsymbol{\mu}$ with the vector forms $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_k)^\top$ and $\boldsymbol{c} = (c_0, \ldots, c_k)^\top$ of the group means and contrasts respectively.

Assuming the responses $Y_{ij},\ i = 0, \ldots, k,\ j = 1, \ldots, n_i$ to be normally distributed and independent within and across the different dose groups, the following test statistic can be used for testing the null hypothesis $H_0 : \boldsymbol{c}^\top \boldsymbol{\mu} = 0$

$$T(\boldsymbol{Y}) = \frac{\sum_{i=0}^{k} c_i \bar{Y}_i}{S\sqrt{\sum_{i=0}^{k} \frac{c_i^2}{n_i}}} \ . \tag{2.1}$$

Thereby $\boldsymbol{Y}$ is the matrix of all responses, $\bar{Y}_i$ denotes the mean response in dose group i and

$$S^2 = \frac{1}{\nu} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \tag{2.2}$$

is the estimator of the pooled variance with $\nu = \sum\limits_{i=0}^{k} n_i - k$ degrees of freedom.

As the test statistic consists of the normally distributed contrast $\sum\limits_{i=0}^{k} c_i \bar{Y}_i$ divided by the independent chi-squared distributed estimator of the pooled variance, under the null hypothesis H_0, the test statistic follows a central t-distribution $T(\boldsymbol{Y}) \sim t_\nu$ with ν defined as above. This implies a rejection of the null hypothesis if the value of the test statistic exceeds the $(1-\alpha)$-quantile of the central t-distribution $t_{1-\alpha,\ \nu}$ in the case of a single one-sided contrast test and the $(1 - \frac{\alpha}{2})$-quantile $t_{1-\frac{\alpha}{2},\ \nu}$ in the case of a single two-sided contrast test.

In the case of a Multiple Contrast Test, i.e. when testing several contrasts simultaneously, one of the methods presented in the previous subsections can be applied. So if m contrasts are to be tested, one could use the $(1 - \frac{\alpha}{m})$-quantile of the central t-distribution instead of the $(1 - \alpha)$-quantile for each individual (one-sided) test in order to adjust according to the Bonferroni method. This, however, leads to a rather conservative control of the FWER.

A method less conservative than the Bonferroni correction is the Union-Intersection Method (UIM) (Roy and Bose, 1953). It is applicable if the global null hypothesis can be expressed by an intersection of all individual null hypotheses

$$H_0 = \bigcap_{i=1}^{m} H_0^i \ .$$

In the matter of dose-finding for example, one would express the global null hypothesis of no overall drug effect H_0 by the intersection of the single null hypotheses H_0^i stating that there is no drug effect for dose i.

For each of these individual null hypotheses H_0^i, an appropriate test T_i should be available which rejects H_0^i if $T_i(\boldsymbol{y}) > a$.

Consequently, the global null hypothesis will be rejected if at least one individual null hypothesis can be rejected

$$\bigcup_{i=1}^{m} \{\boldsymbol{y} : T_i(\boldsymbol{y}) > a\} = \{\boldsymbol{y} : \max_{i=1,\ldots,m} T_i(\boldsymbol{y}) > a\}$$

resulting in the maximum of all individual test statistics as a possible test statistic for the global null hypothesis H_0

$$T(\boldsymbol{y}) = \max_{i=1,\ldots,m} T_i(\boldsymbol{y}) \ .$$

The exact opposite of this principle is called the Intersection-Union Method (IUM) and will not be considered further in this thesis as it is not common in the field of dose-finding. This is due to the fact that it is not necessary for all dose groups to show an effect but PoC is established if at least in one of the dose groups, the response is significantly better than in the placebo group.

To derive the distribution of this maximum t-statistic, consider the following

$$\mathbb{P}\left(\max_{i=1,\ldots,m} T_i \leq t\right) = \mathbb{P}(T_1 \leq t,\ \ldots\ ,\ T_m \leq t) = \mathbb{P}(\boldsymbol{T} \leq \boldsymbol{t})$$

with $\boldsymbol{T} = (T_1, \ldots, T_m)^\top$ and $\boldsymbol{t} = (t, \ldots, t)^\top$.
This means that the maximum t-statistic follows a central m-dimensional t-distribution $T^m_{\nu,\boldsymbol{R}}$ with ν degrees of freedom and a correlation matrix $\boldsymbol{R} = (\rho_{ij})$ where

$$\rho_{ij} = \frac{\sum\limits_{\ell=1}^{k} \frac{c_{i\ell} c_{j\ell}}{n_\ell}}{\sqrt{\left(\sum\limits_{\ell=1}^{k} \frac{c_{i\ell}^2}{n_\ell}\right)\left(\sum\limits_{\ell=1}^{k} \frac{c_{j\ell}^2}{n_\ell}\right)}}, \quad 1 \leq i,\ j \leq m\ . \tag{2.3}$$

Here, $c_{i\ell}$ is the contrast coefficient for dose group ℓ within the i-th hypothesis.

Proof.

$$\begin{aligned}
\operatorname{Cov}\left(\sum_{i=0}^{k} c_{1i}\bar{Y}_i, \sum_{j=0}^{k} c_{2j}\bar{Y}_j\right) &= \mathbb{E}\left[\left(\sum_{i=0}^{k} c_{1i}\bar{Y}_i - \mathbb{E}\left(\sum_{i=0}^{k} c_{1i}\bar{Y}_i\right)\right)\left(\sum_{j=0}^{k} c_{2j}\bar{Y}_j - \mathbb{E}\left(\sum_{j=0}^{k} c_{2j}\bar{Y}_j\right)\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=0}^{k} c_{1i}\bar{Y}_i\right)\left(\sum_{j=0}^{k} c_{2j}\bar{Y}_j\right)\right] \\
&= \sum_{i=0}^{k}\sum_{j=0}^{k} c_{1i}c_{2j}\mathbb{E}(\bar{Y}_i\bar{Y}_j) \\
&= \sum_{i=0}^{k} c_{1i}c_{2i}\mathbb{E}(\bar{Y}_i^2) + \sum_{\substack{i,j=0 \\ i\neq j}}^{k} c_{1i}c_{2j}\mathbb{E}(\bar{Y}_i\bar{Y}_j) \\
&= \sum_{i=0}^{k} c_{1i}c_{2i}\operatorname{Var}(\bar{Y}_i) = \sum_{i=0}^{k} c_{1i}c_{2i}\frac{\sigma^2}{n_i}\ .
\end{aligned}$$

Inserting this into the definition of the correlation

$$\rho_{ij} = \frac{\operatorname{Cov}\left(\sum\limits_{i=0}^{k} c_{1i}\bar{Y}_i, \sum\limits_{j=0}^{k} c_{2j}\bar{Y}_j\right)}{\sqrt{\operatorname{Var}\left(\sum\limits_{i=0}^{k} c_{1i}\bar{Y}_i\right)\operatorname{Var}\left(\sum\limits_{j=0}^{k} c_{2j}\bar{Y}_j\right)}}\ ,$$

using $\operatorname{Var}\left(\sum\limits_{i=0}^{k} c_{1i}\bar{Y}_i\right) = \sigma^2 \sum\limits_{i=0}^{k} \frac{c_{1i}^2}{n_i}$ and $\operatorname{Var}\left(\sum\limits_{j=0}^{k} c_{2j}\bar{Y}_j\right) = \sigma^2 \sum\limits_{j=0}^{k} \frac{c_{2j}^2}{n_j}$ for the variance of the contrasts respectively, this implies the above formula for the correlation of two contrasts. □

Hence, the preservation of the FWER is ensured by comparing the maximum t-statistic against the $(1-\alpha)$-quantile of the m-dimensional t-distribution $\boldsymbol{t}^m_{1-\alpha,\ \nu,\ \boldsymbol{R}}$ with correlation matrix $\boldsymbol{R}$. The quantile of a multivariate t-distribution can be defined in a non-unique sense by the following equation

$$\mathbb{P}_m(|\boldsymbol{T}| \leq \boldsymbol{t}^m_{1-\alpha,\ \nu,\ \boldsymbol{R}}) = 1-\alpha \tag{2.4}$$

with $\boldsymbol{T} = (T_1, \dots, T_m)^\top$ and $\boldsymbol{t}^m_{1-\alpha,\ \nu,\ \boldsymbol{R}} = (t_{1,\ 1-\alpha,\ \nu,\ \boldsymbol{R}}, \dots, t_{m,\ 1-\alpha,\ \nu,\ \boldsymbol{R}})^\top$.
The most straightforward version is the equicoordinate quantile $\boldsymbol{t}^m_{1-\alpha,\ \nu,\ \boldsymbol{R}} = (t, \dots, t)^\top \in \mathbb{R}$ resulting in a cubic confidence region. Due to the fact that its computation (via numerical integration or sampling methods) is comparatively easy, it is well suited for multiple contrast tests. Alternatively, one could also define quantiles that lead to spherical or ellipsoid confidence regions.
More information about the UIM can be found in Hochberg and Tamhane (1987, Chapter 2).

By defining the null hypotheses by means of contrasts, e.g. $H_0 : \boldsymbol{c}^\top \boldsymbol{\mu} = 0$ as mentioned above, it is possible to address every hypothesis as long as only linear components of the means are involved. Particularly, by specifying appropriate contrasts, every set of pairwise comparisons can be defined. This also includes the four different types of MCTs presented in subsection 2.2.2.
Depending on the type of MCT, the contrast vectors for all pairwise comparisons that are involved in this particular contrast test are set up in a matrix, called contrast matrix:

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_i^\top \\ \vdots \\ \boldsymbol{c}_m^\top \end{pmatrix} = \begin{pmatrix} c_{10} & \cdots & c_{1k} \\ \vdots & & \vdots \\ c_{i0} & \cdots & c_{ik} \\ \vdots & & \vdots \\ c_{m0} & \cdots & c_{mk} \end{pmatrix}$$

with the single contrasts $\boldsymbol{c}_i = (c_{i0}, \dots, c_{ik})^\top$ as row vectors.
By multiplying this matrix with the vector of mean responses, one obtains a vector of all pairwise differences that are to be tested. Hence, also the corresponding null hypothesis can be expressed by means of this matrix: $H_0 : \boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{0}$. It is rejected if the maximum of the individual test statistics constructed according to the formula (2.1) exceeds the quantile of a m-dimensional t-distribution $t^m_{1-\alpha,\ \nu,\ \boldsymbol{R}}$ where m is the number of (pairwise) hypotheses that are part of the test.
In the following, a choice of popular MCTs will be presented. For the sake of illustration, a trial with 3 active dose groups and 1 placebo group shall serve as an example.

Tukey Test

The largest set of pairwise differences is the one containing all comparisons between the treatment or dose groups under investigation, also referred to as "all-pairs comparison". It is generally addressed by the Tukey test (Tukey, 1953). The contrast matrix of a Tukey test contains the contrasts of all $m = \binom{k}{2}$ pairwise comparisons (for our example $\binom{4}{2} = 6$ comparisons in total).

In case of the example trial, the contrast matrix would be the following:

$$\boldsymbol{C}_{\mathrm{T}} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

and the maximum test statistic would be compared against the $t^{6}_{1-\alpha,\ \nu,\ \boldsymbol{R}}$-quantile.

Dunnett Test

The Dunnett procedure (Dunnett, 1955) is a method for the comparison of multiple treatment groups (or dose groups as in the present case of dose-finding studies) with a control. As the set of hypotheses is a subset of the all-pairwise comparisons set, one could apply the Tukey test and only consider the comparisons of interest for the Dunnett test. However this would be a rather conservative approach. Therefore, a smaller contrast matrix is defined only containing the contrasts corresponding to the comparisons of the active treatment groups with the control group. In the introduced example, the contrast matrix would be:

$$\boldsymbol{C}_{\mathrm{D}} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

and the maximum test statistic would be compared against the $t^{3}_{1-\alpha,\ \nu,\ \boldsymbol{R}}$-quantile.

Williams-Type MCT

The Williams-type MCT is a procedure to test the existence of a treatment effect by comparing all active dose groups with a control in case of an underlying monotonic dose-response relationship. Originally, Williams (1971) presented his procedure as a combination of a test for the PoC and, if this has been successfully shown, a stepwise procedure to identify the lowest dose with a significant change in the response variable.

The method itself uses the Maximum Likelihood (ML) estimates $\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = (\hat{\mu}_{0,\mathrm{ML}}, \ldots, \hat{\mu}_{k,\mathrm{ML}})$ as estimates for the mean responses in the different dose groups. But as a monotonic dose-response relationship is assumed, also the ML estimates shall satisfy

$$\hat{\mu}_{0,\mathrm{ML}} \leq \hat{\mu}_{1,\mathrm{ML}} \leq \ldots \leq \hat{\mu}_{k,\mathrm{ML}} \ . \tag{2.5}$$

One way to ensure this is via the so-called Pool-Adjacent-Violator algorithm (PAVA). If the inequality is fulfilled by all estimates of the mean responses, the ML estimates remain unrevised. Otherwise, if there is one $i > 0$ with $\hat{\mu}_{i,\ \mathrm{ML}} > \hat{\mu}_{i+1,\ \mathrm{ML}}$, both estimates are replaced by the weighted mean of themselves

$$\hat{\mu}_i,\ \hat{\mu}_{i+1} = \frac{w_i \hat{\mu}_{i,\mathrm{ML}} + w_{i+1} \hat{\mu}_{i+1,\mathrm{ML}}}{w_i + w_{i+1}}$$

with equal weights $w_i = w_{i+1} = 1$. If these means are again part of an averaging step, they are weighted by 2 and so forth. The process is repeated until all means satisfy the inequality condition (2.5). Note that the estimate for the placebo group is excluded from this process, that is $\hat{\mu}_0 = \hat{\mu}_{0,\text{ML}}$.

For testing the global null hypothesis of no treatment effect, the mean response in the highest dose group is tested against the placebo response by means of the following test statistic

$$\bar{T}_{k,W} = \frac{\hat{\mu}_k - \hat{\mu}_{0,\text{ML}}}{s\sqrt{\frac{1}{n_k} + \frac{1}{n_0}}}$$

with s^2 being some kind of variance estimator.
Originally, Williams (1971) used the denominator from the usual two-sample t-test

$$\sqrt{\frac{2}{r}\frac{(n_i-1)S_i^2 + (n_j-1)S_j^2}{n_i + n_j - 2}}$$

with sample variances S_i^2 and S_j^2 and replication r for the studentization of the test statistic.
But Bretz (1999) showed in his PhD thesis that the use of a usual (pooled) variance estimator of the whole sample

$$\frac{\sum\limits_{i=0}^{k}\sum\limits_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}{\sum\limits_{i=0}^{k} n_i - (k+1)}$$

shows more favourable results in terms of power.
In any of the two cases, the test statistic will be compared against the same critical value $\bar{t}_{1-\alpha,k,\nu}$. It should be emphasized that if $k > 1$, the critical value cannot be derived from the Student's t-distribution because of the potential averaging of the estimates. Instead it has to be computed numerically or can be derived theoretically. Only in case $k = 1$ the critical value is a quantile of the Student's t-distribution.

If the global null hypothesis has been rejected, the single tests to identify the lowest effective dose can be conducted successively. By means of the analogous test statistics (exchanging $\hat{\mu}_k$ by $\hat{\mu}_{k-1}$ et cetera), the procedure starts with the comparison of the second highest dose group with the placebo group and does continue until no significant difference between the mean responses can be detected. Unlike for the other MCTs, it is not the same critical value to be used for every comparison but the critical values differ depending on the number of mean responses involved in the PAVA process.

Bretz (1999) established a link between the Williams test (in the variation described above) and a MCT for the evidence of an overall treatment effect by defining the following contrast matrix

$$\boldsymbol{C}_{\text{W}} = \begin{pmatrix} -1 & 0 & \cdots & 0 & 1 \\ -1 & 0 & \cdots & \frac{n_{k-1}}{n_{k-1}+n_k} & \frac{n_k}{n_{k-1}+n_k} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ -1 & \frac{n_1}{n_1+\ldots+n_k} & \cdots & \frac{n_{k-1}}{n_1+\ldots+n_k} & \frac{n_k}{n_1+\ldots+n_k} \end{pmatrix}.$$

This implies

$$\max \boldsymbol{C}_W \hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \hat{\mu}_k - \hat{\mu}_{0,\mathrm{ML}} \ .$$

Hence, by comparing the maximum of the single contrast tests with the same critical value that is used for testing the global null hypothesis in the original Williams test, the existence of an overall treatment effect can be tested. The exploration of the lowest effective dose as it is realized by the stepwise procedure presented in the original paper is not included in this MCT procedure.

2.3 Parametric Modelling of the Dose-Response Relationship

Instead of applying multiple comparison procedures, another method to address the identification of the optimal dose is to model the dose-response relationship by a prespecified parametric function.
In the following, a selection of frequently used models will be presented and displayed by plotting the function with varying parameter values. Besides, as is it needed for the construction of appropriate contrast tests in the MCP-Mod approach later on, for each model the derivation of prior estimates will be explained. These prior estimates can be used as initial parameters for iterative procedures of non-linear model fitting as well.
Furthermore, the procedure of testing for the existence of a dose-response and the estimation of an adequate dose on the basis of the fitted model will be outlined.
The theory in this section is mainly based on Ting (2006, Chapter 10), Branson et al. (2003) and Bretz et al. (2008).

2.3.1 General Notation

Generally, a clinical outcome Y (either an efficacy or a safety measure) is observed for a population of patients assigned to one of the active doses $d_1, \ldots, d_k$ or the control d_0. In total, this amounts to $k+1$ dose groups, mostly investigated in a parallel group design. Hence, let Y_{ij} denote the response of patient j in dose group i, $i = 0, \ldots, k$, $j = 1, \ldots, n_i$. In the basic case, the response is assumed to be normally distributed $Y_{ij} \sim N(\mu_i, \sigma^2)$ in consequence of the following model

$$Y_{ij} = f(d, \boldsymbol{\theta}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{2.6}$$

with $f(\cdot)$ being a linear or non-linear function parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^p$.
In practice, it is often sufficient to consider the standardized version f^0 of a dose-response model which can be obtained according to the decomposition

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

Thereby $\boldsymbol{\theta}^0 \in \mathbb{R}^{p-2}$ is the standardized model parameter of f^0.
If such a decomposition is possible, the model is called location-scale model.

2.3.2 Frequently Used Model Shapes

Linear Model

The simplest dependency is a linear dose-response model which is expressed by

$$f(d, \boldsymbol{\theta}) = \mathrm{E}_0 + \delta d, \qquad \boldsymbol{\theta}^\top = (\mathrm{E}_0,\ \delta)$$

and a standardized version is given by

$$f^0(d) = d \ .$$

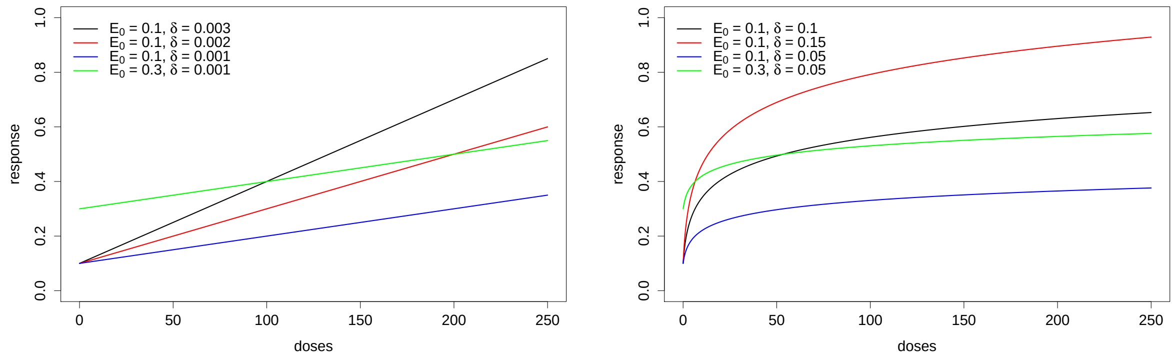
The parameter E_0 will hereinafter be termed as basal or placebo effect, that is to say the value of the response for $d = 0$, and δ can be considered the scale parameter of the model.

Another type of linear model is the linear log-dose model

$$f(d, \boldsymbol{\theta}) = \mathrm{E}_0 + \delta \log(d + c), \qquad f^0(d) = \log(d + c) \ ,$$

where the constant $c > 0$ is only included to avoid issues if $d = 0$. Typically, c is chosen to be 1.

The influence of the model parameters E_0 and δ are illustrated in Figure 2.4(a) for a linear dose-response model, and in Figure 2.4(b) for a log-dose model. In both cases, only positive values for δ are considered resulting in a positive slope. A negative value would lead to a decreasing dose-response curve.



2.4 (a): Linear Dose-Response Relationship

2.4 (b): Linear Log-Dose Model

Figure 2.4: Linear Dose-Response Relationship

For the linear dose-response models presented in this paragraph, no prior estimates for the parameters are necessary as the standardized function is completely independent from their choice. Only the doses are needed.

Quadratic Model

As already mentioned at the beginning of this chapter, a monotonously increasing dose-response relationship is very likely in most of the cases. However, if also a non-monotonous model is worth to consider, there is the option to fit a quadratic model

$$f(d, \boldsymbol{\theta}) = \mathrm{E}_0 + \beta_1 d + \beta_2 d^2, \qquad \boldsymbol{\theta}^\top = (\mathrm{E}_0, \beta_1, \beta_2) \ . \tag{2.8a}$$

Thus also a possible non-monotonic relationship can be captured. Again, the model can be varied by substituting d with $log(d + c)$ in equation (2.8a).
The determination of the standardized version for the quadratic model can be obtained by dividing

the whole term by the absolute value of β_1

$$f^0(d,\delta) = \begin{cases} d+\delta d^2, & \text{if } \beta_2 < 0 \\ -d+\delta d^2, & \text{if } \beta_2 > 0 \end{cases}, \qquad \delta = \frac{\beta_2}{|\beta_1|}\ . \tag{2.8b}$$

The more common inverted-U shape (also called umbrella shape) stems from a quadratic model with $\beta_2 < 0$. It is illustrated in Figure 2.5 for a set of different parameter values.

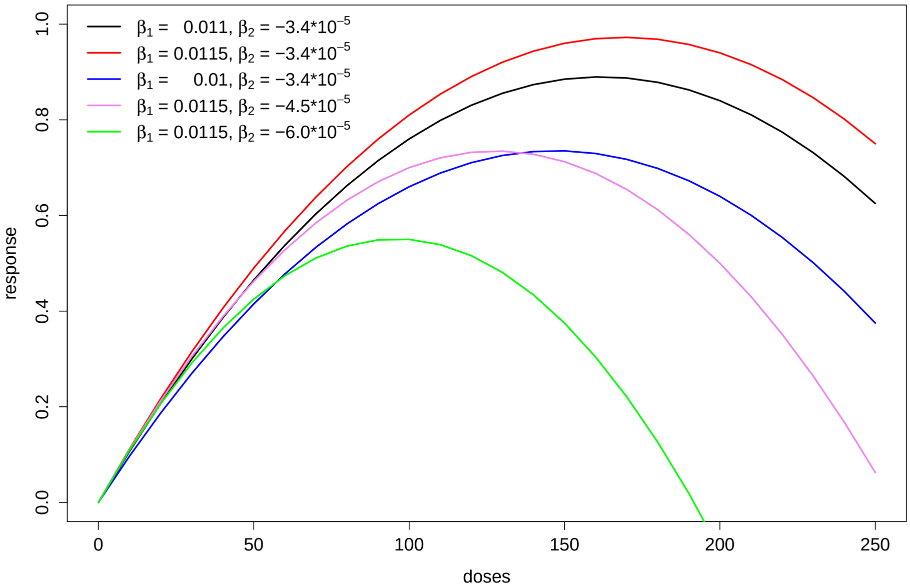


Figure 2.5: Quadratic Dose-Response Relationship

An initial value for the parameter of the standardized model function δ can be obtained by conditioning on the dose which is assumed to produce the maximum (minimum for U-shaped curves) response $d_{\text{opt}} = -\frac{\beta_1}{2\beta_2} = -\frac{1}{2\delta}$. Without loss of generality, let d_{opt} be the dose associated with the maximum response which is synonymous with an underlying umbrella shaped curve. For a pair of values (d^*, p^*) derived from prior knowledge, with p^* being the suspected percentage of the maximum change over placebo for a given dose d^*, an estimate for δ^* can be derived by solving the equation

$$\hat{\delta} = \begin{cases} -\frac{1-\sqrt{1-p^*}}{2d^*}, & \text{if } d^* < d_{\text{opt}} \\ -\frac{1+\sqrt{1-p^*}}{2d^*}, & \text{if } d^* \geq d_{\text{opt}}\ . \end{cases}$$

Proof.
If d_{opt} is the dose associated with the maximum change in the response, the maximum change itself is given by

$$\begin{aligned} f^0(d_{\text{opt}},\delta) &= d_{\text{opt}} + \delta d_{\text{opt}}^2 \\ &= -\frac{1}{2\delta} + \delta\left(-\frac{1}{2\delta}\right)^2 \\ &= -\frac{1}{2\delta} + \frac{1}{4\delta} = -\frac{1}{4\delta}\ . \end{aligned}$$

By inserting this into the standardized model formula (2.8b) and solving the equation for δ, the above formula is obtained:

$$\begin{aligned} & p^*\left(-\frac{1}{4\delta}\right) = d^* + \delta(d^*)^2 \\ \Leftrightarrow \qquad & p^* = -4\delta d^* - 4\delta^2(d^*)^2 \\ \Leftrightarrow \qquad & \hat{\delta} = \begin{cases} -\frac{1-\sqrt{1-p^*}}{2d^*}, & \text{if } d^* < d_{\text{opt}} \\ -\frac{1+\sqrt{1-p^*}}{2d^*}, & \text{if } d^* \geq d_{\text{opt}} \ . \end{cases} \end{aligned}$$

□

Exponential Model

If the relationship between the administered dose and the response can be assumed to be convex, the exponential model is a suitable way to describe this. It is defined as

$$f(d,\boldsymbol{\theta}) = \mathrm{E}_0 + \mathrm{E}_1 \exp\left(\frac{d}{\delta}\right)$$

with model parameter $\boldsymbol{\theta}^\top = (\mathrm{E}_0, \mathrm{E}_1, \delta)$ and a standardized version

$$f^0(d,\delta) = \exp\left(\frac{d}{\delta}\right) \quad . \tag{2.9}$$

The parameter δ can be interpreted as the rate of increase in the response variable (or decrease if $\delta < 0$ respectively).

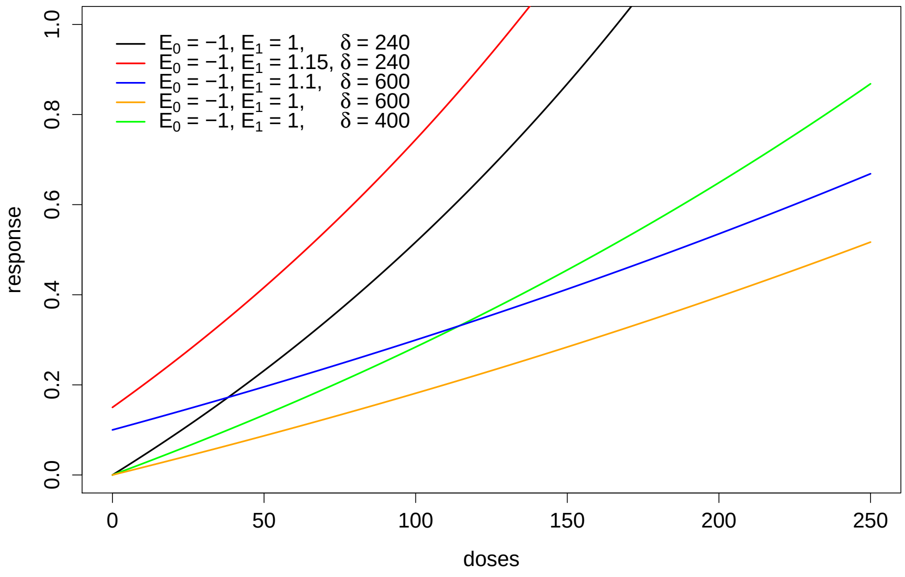


Figure 2.6: Exponential Dose-Response Relationship

Again, the shapes of an exponential model are illustrated for a set of different values for E_1 and δ in Figure 2.6. By increasing the parameter E_1, not only the slope of the dose-response curve changes, but also the intercept increases.

An initial parameter estimate for δ can be obtained analogously to the case of a quadratic model by using prior knowledge of an expected percentage of the maximum change over placebo p^* for a given dose d^*. Note that for the exponential model the percentage increase in the response over placebo can be expressed by $f^0(d, \delta) - 1$. By inserting the prior information into the model formula (2.9), the following estimate can be derived

$$\hat{\delta} = \frac{d^*}{\log(1 + p^*)} \ .$$

$\mathrm{E}_{\mathrm{max}}$ Model

Another very common descriptor of a dose-response relationship is the (hyperbolic) $\mathrm{E}_{\mathrm{max}}$ model

$$f(d, \boldsymbol{\theta}) = \mathrm{E}_0 + \frac{\mathrm{E}_{\mathrm{max}} d}{\mathrm{ED}_{50} + d} \ , \qquad \boldsymbol{\theta}^\top = (\mathrm{E}_0, \mathrm{E}_{\mathrm{max}}, \mathrm{ED}_{50}) \tag{2.10}$$

where E_0 is again the placebo effect as described for the linear model, $\mathrm{E}_{\mathrm{max}}$ is the maximum effect over placebo, i.e. the difference between the maximum response (at an infinite dose) and the response for placebo and ED_{50} is the dose which is expected to induce half of the maximum change (cf. section 2.1). The corresponding standardized version of the model is given by

$$f^0(d, \mathrm{ED}_{50}) = \frac{d}{\mathrm{ED}_{50} + d}$$

and models the percentage of the maximum effect over placebo achieved by dose d (cf. equation (2.10)).

The sign of the maximum effect in the $\mathrm{E}_{\mathrm{max}}$ model is decisive for the monotonous behaviour of the dose-response curve. A positive value $\mathrm{E}_{\mathrm{max}} > 0$ represents an increase in the response with increasing dose level whereas a negative value $\mathrm{E}_{\mathrm{max}} < 0$ indicates a monotonously decreasing dose-response function. A higher (absolute) value of $\mathrm{E}_{\mathrm{max}} > 0$ is accompanied with a broader range of the dose-response curve. Again, this model family is illustrated for a choice of parameter values in Figure 2.7.

As the standardized version of the $\mathrm{E}_{\mathrm{max}}$ model directly represents the percentage of increase in the response over placebo, an estimate for ED_{50} can be derived as before using a pair of values (d^*, p^*) from prior knowledge. This implies an estimate given by

$$\widehat{\mathrm{ED}}_{50} = \frac{d^*(1 - p^*)}{p^*} \ .$$

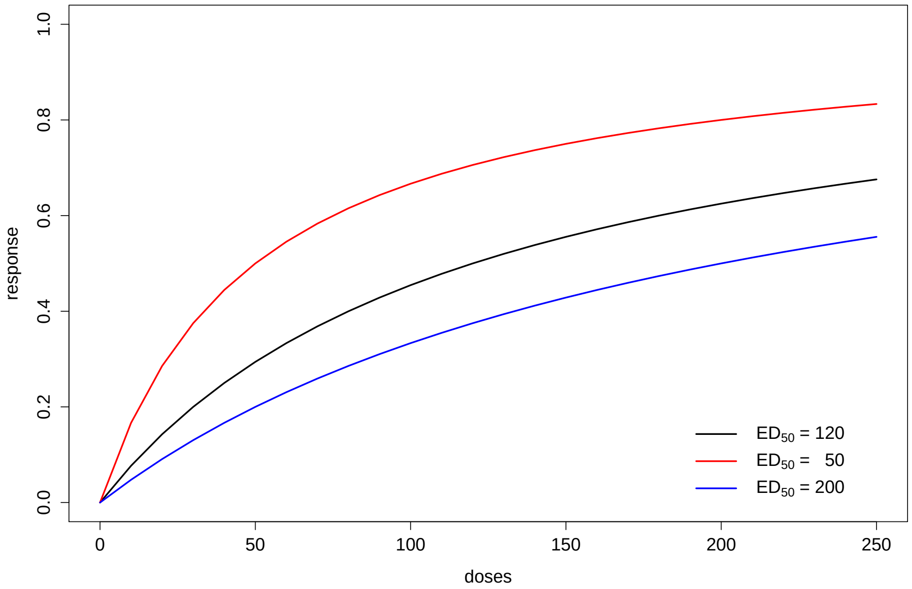


Figure 2.7: $\mathrm{E}_{\max}$ Model as Dose-Response Relationship

Sigmoid $\mathrm{E}_{\max}$ Model

An extension of the (hyperbolic) $\mathrm{E}_{\max}$ model is the sigmoid $\mathrm{E}_{\max}$ model which includes an additional slope factor h, also termed Hill factor

$$f(d,\boldsymbol{\theta}) = \mathrm{E}_0 + \frac{E_{\max} d^h}{\mathrm{ED}_{50}^h + d^h}$$

with parameter $\boldsymbol{\theta}^\top = (\mathrm{E}_0, \mathrm{E}_{\max}, \mathrm{ED}_{50}, h)$. The standardized version is given by

$$f^0(d, \mathrm{ED}_{50}) = \frac{d^h}{\mathrm{ED}_{50}^h + d^h}$$

and, analogously to the hyperbolic $\mathrm{E}_{\max}$ model, represents the percentage of the maximum change over placebo related to a certain dose d.
For the sigmoid $\mathrm{E}_{\max}$ model, the value of ED_{50} determines the inflection point of the dose-response curve, but does not have impact on the slope of the curve so that with a changing value of ED_{50}, the curve is shifted along the x-axis. Besides, the Hill factor can be interpreted as a measure for the sensitivity of the response variable to a change in the administered dose as it is regulating the steepness of the curve. This behaviour can also be observed in Figure 2.8.

Due to the additional slope factor, it is not that straightforward to derive the initial parameter estimates for the sigmoid $\mathrm{E}_{\max}$ model as for the other models presented in this section. One possibility is to fit a smoothing spline function to the observed data and to extract the parameters E_0, $\mathrm{E}_{\max}$ and ED_{50} from the resulting plot.

Additionally, the hill factor h can be estimated using the following rule of thumb

$$h \approx \frac{1.91}{\log_{10}\left(\frac{\mathrm{ED}_{90}}{\mathrm{ED}_{10}}\right)} .$$

The two parameters ED_{90} and ED_{10} can again be obtained from the plot.
This method is described in more detail in Ting (2006, Chapter 9).

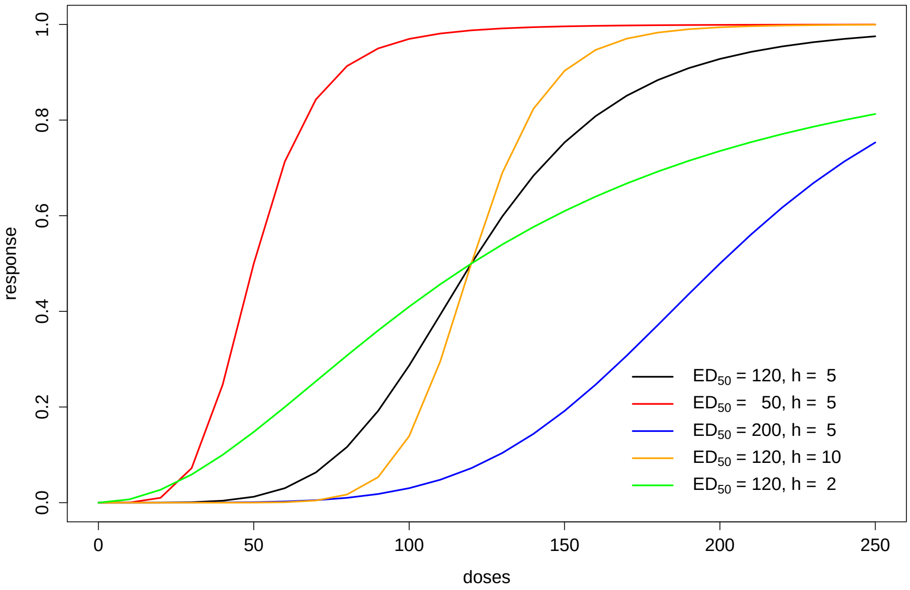


Figure 2.8: Sigmoid $\mathrm{E}_{\max}$ Model as Dose-Response Relationship

Logistic Model

An alternative for the modelling of an S-shaped dose-response curve is the logistic model

$$f(d,\boldsymbol{\theta}) = \mathrm{E}_0 + \frac{\mathrm{E}_{\max}}{1+\exp\left[\frac{\mathrm{ED}_{50}-d}{\delta}\right]}$$

with parameter $\boldsymbol{\theta}^\top = (\mathrm{E}_0, \mathrm{E}_{\max}, \mathrm{ED}_{50}, \delta)$ and a standardized version of

$$f^0(d,\boldsymbol{\theta}) = \frac{1}{1+\exp\left[\frac{\mathrm{ED}_{50}-d}{\delta}\right]} . \tag{2.12}$$

In contrast to the interpretation of the previous models, E_0 can still be seen as some kind of basal effect. However, it is no explicit placebo effect as it is not the response for $d = 0$ but the left limit of the function, i.e. the limit for $d \longrightarrow -\infty$.
The meaning of the other parameter, ED_{50}, remains the same as before. It again determines the inflection point and thus can be regarded as a kind of location parameter. The steepness of the curve is controlled by the parameter δ. Figure 2.9 shows the curves of a logistic model for varying values of the parameters ED_{50} and δ.

As the standardized logistic model is a function of two parameters, one also needs (at least) two pairs of values (d_1^*, p_1^*) and (d_2^*, p_2^*) for the derivation of the initial estimates. Just like in the $\mathrm{E}_{\max}$ model, the standardized version of the logistic model can be interpreted as the maximum effect $\mathrm{E}_{\max}$ related to a certain dose. Therefore, the initial estimates can be directly derived from the inversion of formula (2.12) and are hence given by

$$\hat{\delta} = \frac{d_2^* - d_1^*}{\operatorname{logit}(p_2^*) - \operatorname{logit}(p1^*)}$$

and

$$\widehat{\mathrm{ED}}_{50} = \frac{d_1^* \operatorname{logit}(p_2^*) - d_2^* \operatorname{logit}(p_1^*)}{\operatorname{logit}(p_2^*) - \operatorname{logit}(p_1^*)} \ .$$

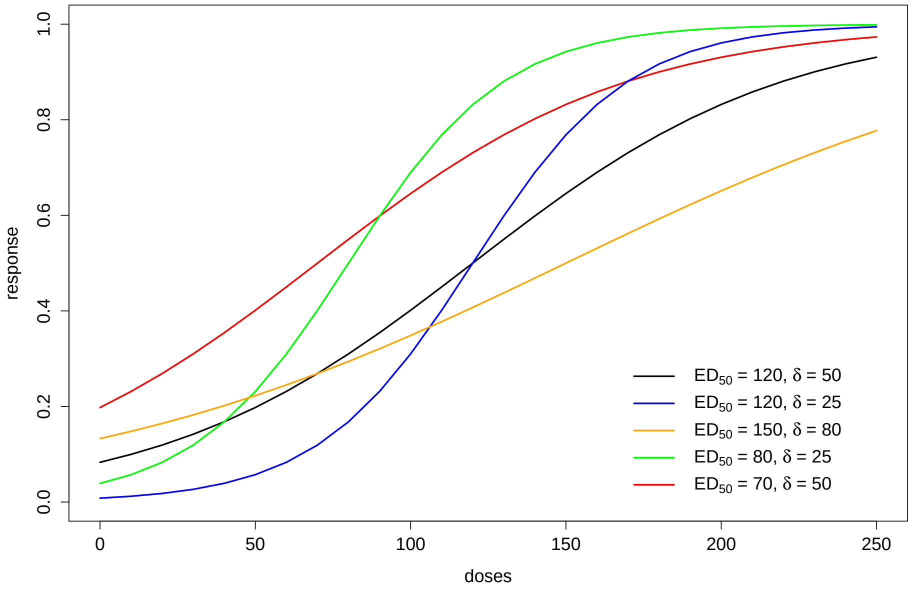


Figure 2.9: Logistic Model as Dose-Response Relationship

In general, if more pairs of values are available than it would be necessary for the derivation of the estimates, one can determine the initial estimates for every pair of values and afterwards use the average of all estimates as the "final" initial estimate.

The fitting of these dose-response models under the assumption of independent and identically distributed (iid) errors ϵ_{ij} can either be conducted by means of least squares estimation in case of a linear model or via Generalized Least Squares (GLS) procedures as for example the iterative Newton's method. For the latter, as mentioned at the beginning of this section, the initial estimates are used as starting values for the algorithm according to the presented formulae.

2.3.3 Estimation of the Minimum Effective Dose (MED)

Having fitted the prespecified dose-response model to the data, one can determine the target dose of interest on the basis of the fitted dose-response curve.

As previously mentioned, there are several criteria for defining a dose as optimal. To be consistent with the MCP-Mod approach, the focus in this thesis is on the MED, i.e. the dose that produces a certain (clinically relevant) difference Δ in the outcome compared to placebo.

Formally, the MED is defined as

$$\text{MED} = \underset{d \in (d_0, d_k]}{\arg\min} \{ f(d, \boldsymbol{\theta}) > f(d_0, \boldsymbol{\theta}) + \Delta \} \ . \tag{2.13}$$

It is restricted to the interval $(d_1, d_k]$ to prevent issues caused by an extrapolation beyond the investigated dose range. The lower limit of the interval, d_1, represents placebo whereas d_k is the highest dose included in the study.

Possible estimates for the MED are given by the following formulae:

$$\widehat{\text{MED}}_1 = \underset{d \in (d_0, d_k]}{\arg\min} \{ U_d > f(d_0, \hat{\boldsymbol{\theta}}) + \Delta, \ L_d > f(d_0, \hat{\boldsymbol{\theta}}) \} \tag{2.14a}$$

$$\widehat{\text{MED}}_2 = \underset{d \in (d_0, d_k]}{\arg\min} \{ f(d, \hat{\boldsymbol{\theta}}) > f(d_0, \hat{\boldsymbol{\theta}}) + \Delta, \ L_d > f(d_0, \hat{\boldsymbol{\theta}}) \} \tag{2.14b}$$

$$\widehat{\text{MED}}_3 = \underset{d \in (d_0, d_k]}{\arg\min} \{ L_d > f(d_0, \hat{\boldsymbol{\theta}}) + \Delta \} \tag{2.14c}$$

with L_d and U_d denoting the lower and upper $(1-2\gamma)$ confidence limits of the expected outcome value $f(d, \hat{\boldsymbol{\theta}})$ associated with dose d. Thereby it is not absolutely necessary to choose γ small enough to produce a statistically significant effect at the significance level of α. But, if chosen too generously, it may happen that the estimate for the MED is smaller than a dose that failed to show any significant effect in the study which leads to interpretation issues.

By construction, the estimates are in an ascending order $\widehat{\text{MED}}_1 \leq \widehat{\text{MED}}_2 \leq \widehat{\text{MED}}_3$ implying that in general, the estimate given by formula (2.14a) tends to determine a dose that is smaller than the true MED while using equation (2.14c) in contrast may lead to an overestimation of the MED. This has been shown by simulations, for example in the paper of Bretz et al. (2005).

2.3.4 Precision of Estimation

There are several ways to determine the precision of the MED estimate or the estimated response at a fixed dose $d = d^*$ for a certain underlying dose-response model.

One option would be to use non-parametric or parametric bootstrap methods.

In the case of a non-parametric bootstrap, the patient data is re-sampled by randomly drawing observations with replacement of the original data set and analyzed analogously to the analysis of the original data for an adequate number of times. The resulting characteristic(s) of interest (estimated MED values and/or expected response at dose $d = d^*$) for each run are collected and represent the bootstrap sample.

Alternatively, the parameter vector $\boldsymbol{\theta}$ of the dose-response model can be directly re-sampled using parametric bootstrap. This means that a sample of parameter vectors is produced by generating random numbers of a normal distribution as $\boldsymbol{\theta}$ is asymptotically normally distributed with mean $\hat{\boldsymbol{\theta}}$

and covariance matrix $\widehat{V(\boldsymbol{\theta})}$ as a result of the assumed dose-response model (cf. equation (2.6)). The bootstrap sample is then obtained by reading out the response values from the dose-response model with the inserted parameter vectors and, if desired, by estimating the MED on the basis of the resulting dose-response model.

The bootstrap samples, no matter whether they are obtained by non-parametric or parametric bootstrap, can then be used in the final step to derive a confidence interval for the MED estimate and/or the response value at $d = d^*$ by means of Monte-Carlo methods, i.e. by using their empirical analogues.

Another option is to make use of the asymptotic behaviour of the least square estimate (that is also used for the parametric bootstrap as previously described) to analytically derive a variance formula for $\hat{\boldsymbol{\theta}}$ in model (2.6).

In a usual non-linear regression setting (each observed value of the response corresponds to a unique value of the independent variable), the least squares estimate $\hat{\boldsymbol{\theta}}$ asymptotically follows a normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix

$$V(\boldsymbol{\theta}) = \sigma^2 \left(G(\boldsymbol{\theta})^\top G(\boldsymbol{\theta})\right)^{-1} = \sigma^2 \left(\sum_{j=0}^{k} g(d_j, \boldsymbol{\theta}) g^\top(d_j, \theta)\right)^{-1}$$

where

$$G(\boldsymbol{\theta}) = \left(\frac{\partial f(d_i, \boldsymbol{\theta})}{\partial \theta_j}\right)_{i=1,\ldots,N;\ j=0,\ldots,p-1}$$

is representing the matrix of partial derivatives of the response function $f(d, \boldsymbol{\theta})$ and

$$g^\top(d, \boldsymbol{\theta}) = \frac{\partial f(d, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(1, f^0(d, \boldsymbol{\theta}^0), \theta_1 \, \frac{\partial f^0(d, \boldsymbol{\theta}^0)}{\partial \theta_2}, \ldots, \theta_1 \, \frac{\partial f^0(d, \boldsymbol{\theta}^0)}{\partial \theta_{p-1}}\right)$$

is denoting the gradient of the response function with respect to $\boldsymbol{\theta}$ accordingly. This results from linearization by means of Taylor's theorem, if the appropriate regularity conditions are fulfilled.

In the underlying case of a dose-finding study where the independent variable is in fact discrete, i.e. there are several observations for a certain dose d_i, the formula has to be adapted such that

$$V(\boldsymbol{\theta}) = \frac{\sigma^2}{N} \left(\sum_{j=0}^{k} w_j g(d_j, \theta) g^\top(d_j, \theta)\right)^{-1} \tag{2.15}$$

with allocation rates $(w_0, \ldots, w_k)$.

By application of the delta method, an approximately normal distribution can be derived also for the transformation of $\hat{\boldsymbol{\theta}}$

$$f(d, \hat{\boldsymbol{\theta}}) \sim \mathcal{N}\left(f(d, \boldsymbol{\theta}),\ g^\top(d, \theta) V^{-1}(\boldsymbol{\theta}) g(d, \theta)\right) \ .$$

Hence, the limits of the point-wise confidence interval for the predicted response at a certain dose $d = d^*$ are given by

$$f(d^*, \hat{\boldsymbol{\theta}}) \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{N}} \left\{ g^\top(d^*, \hat{\boldsymbol{\theta}}) \left(\sum_{i=0}^{k} w_i \ g(d_i, \hat{\boldsymbol{\theta}}) g^\top(d_i, \hat{\boldsymbol{\theta}})\right)^{-1} g(d^*, \hat{\boldsymbol{\theta}}) \right\}^{\frac{1}{2}} + o\left(\frac{1}{\sqrt{N}}\right)$$

with

$$\hat{\sigma}^2 = \frac{1}{N-p} \|\boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\theta}})\|^2 = \frac{1}{N-p} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (y_{ij} - f(d_i,\ \hat{\boldsymbol{\theta}}))^2$$

denoting the common least squares estimate of σ^2 and z_β the β-quantile of the standard normal distribution.
By replacing the (deterministic) σ^2 in the formula by its estimate $\hat{\sigma}^2$, the normal distribution turns into a studentized t-distribution. However, if the sample size is sufficiently large, the quantile of the normal distribution is an adequate approximation of the quantile of a t-distribution.

If it is of further interest to investigate the precision of the MED estimate, it is useful to start from an explicit formula for the estimate itself. When considering the decision rule given in equation (2.14b), a possible approximation would be

$$\widehat{\mathrm{MED}} = a_p(\hat{\boldsymbol{\theta}}) := h^0 \left(f^0(d_0, \hat{\boldsymbol{\theta}}^0) + \frac{\Delta}{\hat{\theta}_1} \right)$$

where h^0 is the inverse of the standardized model function f^0 with respect to d.
By using the delta method again, the variance of this estimate can be derived as

$$\begin{aligned} \mathrm{Var}(\widehat{\mathrm{MED}}) &= \mathrm{Var}(a_p(\hat{\boldsymbol{\theta}})) + o\left(\frac{1}{N}\right) \\ &= b^\top(\boldsymbol{\theta}) V^-(\boldsymbol{\theta}) b(\boldsymbol{\theta}) + o\left(\frac{1}{N}\right) \end{aligned}$$

where

$$\begin{aligned} b(\boldsymbol{\theta}) = b(\theta_0, \ldots, \theta_{p-1}) &= \frac{\partial}{\partial \boldsymbol{\theta}} a_p(\theta_0, \ldots, \theta_{p-1}) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} h^0 \left(f^0(d_0, \boldsymbol{\theta}^0) + \frac{\Delta}{\theta_1} \right) \end{aligned} \tag{2.16}$$

is the gradient of function a_p with respect to $\boldsymbol{\theta}$ and $V^-(\boldsymbol{\theta})$ is the generalized inverse of the matrix $V(\boldsymbol{\theta})$ so that the formula also holds in the case of a singular covariance matrix.
Consequently, the asymptotic confidence interval for the MED estimate is given by

$$\widehat{\mathrm{MED}} \pm z_{1-\frac{\alpha}{2}} \sqrt{b^\top(\hat{\boldsymbol{\theta}}) V^-(\hat{\boldsymbol{\theta}}) b(\hat{\boldsymbol{\theta}})}$$

where z_β is denoting the β-quantile of the standard normal distribution as before.

The theory presented in the last two sections is taken from Branson et al. (2003, Section 4.3) and Dette et al. (2008, Section 3). For more detailed information about the asymptotic behaviour of GLS estimates, it is referred to the books of Seber and Wild (2003, Chapter 5) and Gallant (1987, Chapter 1 & 4).

Chapter 3

The MCP-Mod Approach

3.1 MCP-Mod Approach for Normally Distributed Outcomes

As both of the common procedures for the planning and analysis of dose-finding studies presented in the last chapter have their shortcomings, Bretz et al. (2005) introduced an approach for normally distributed data that combines both principles in one. Thus, it is possible to benefit from the modelling approach such that the choice of the optimal dose is not restricted to those doses included in the trial. At the same time, the validity of the results is improved in comparison to the basic modelling approach by considering not only one but several shapes of dose-response relationships and selecting the one that fits best to the collected data. Furthermore, a study conducted according to the MCP-Mod approach is able to simultaneously address the aim of PoC and estimating the optimal dose in the course of one single study.

The general framework of the MCP-Mod approach is the same as for the parametric modelling of the dose-response curve presented in subsection 2.3.1. The current section will first cover the steps that must be considered in the planning phase of the study, meaning prior to the start of the trial, and will then take into consideration the methods applied to the analysis of the data. For each step, the realization by means of the functions implemented in the R package `DoseFinding` (Bornkamp et al., 2014) will be presented and important options will be cited.
Before starting with the detailed explanation of the single design and analysis steps, the following flow chart (Figure 3.1) shall serve as an overview over the basic idea behind this approach.

After the definition of the main study characteristics (e.g. the primary endpoint and the study population) as it is essential for any study, the study-specific features have to be set up such that the outcome of the study is as promising as possible. Those features include

- candidate models for the dose-response relationship (selected on the basis of available prior knowledge which can be obtained for example from similar compounds),
- the choice of dose groups to be included in the study as well as the corresponding allocation ratios (may be restricted by practicability or technical reasons),
- the optimal contrasts for the selected candidate models to maximize the power of the trend tests conducted in the MCP-step,
- the sample size providing a certain target power for the establishment of PoC or a certain precision for one of the estimates of interest.

Once the study has been designed thoroughly and the data has been collected accordingly, the analysis is carried out in two subsequent steps. First, the models are tested separately for an existing dose

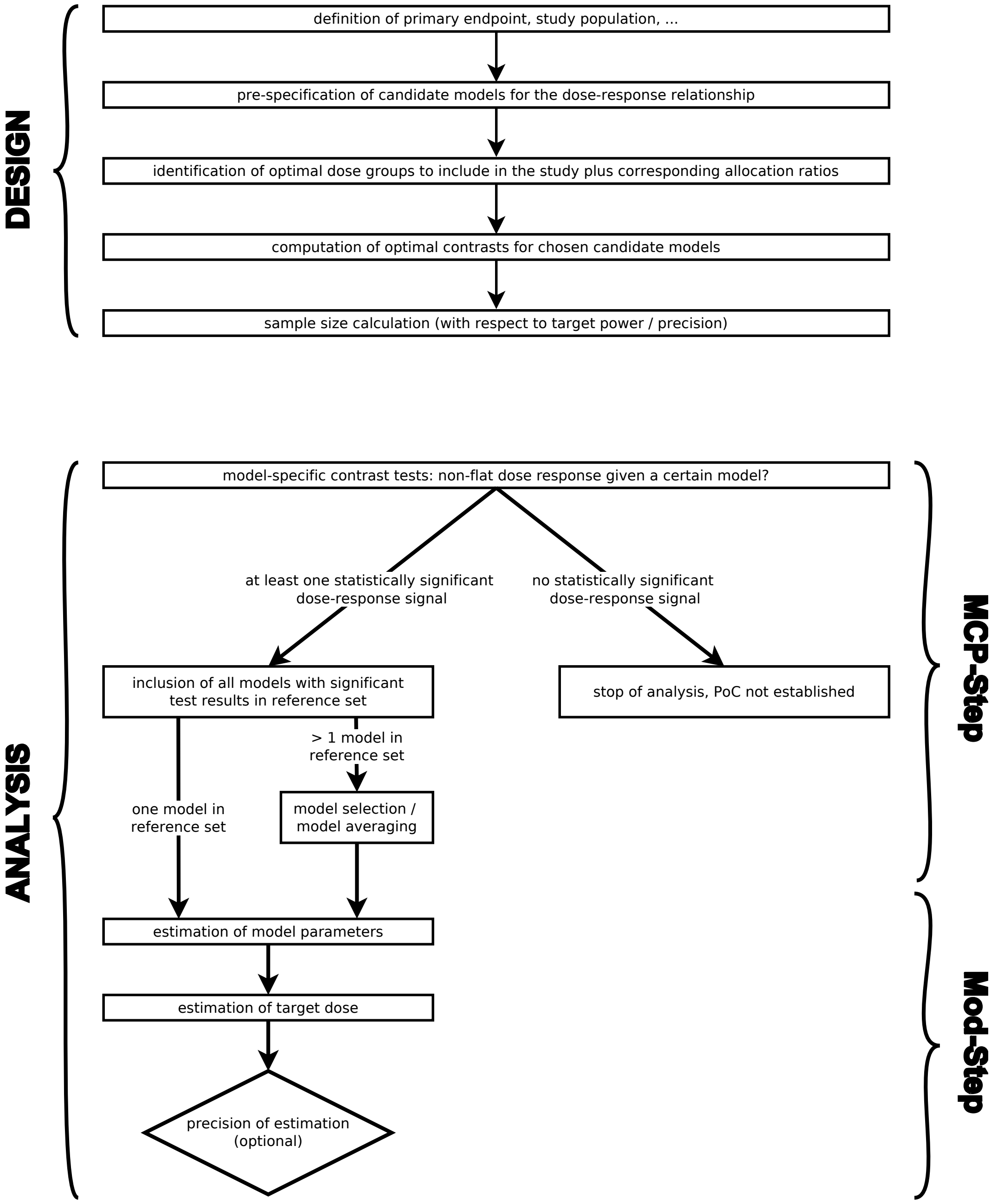

DESIGN
definition of primary endpoint, study population, ...
pre-specification of candidate models for the dose-response relationship
identification of optimal dose groups to include in the study plus corresponding allocation ratios
computation of optimal contrasts for chosen candidate models
sample size calculation (with respect to target power / precision)
ANALYSIS
model-specific contrast tests: non-flat dose response given a certain model?
at least one statistically significant dose-response signal
no statistically significant dose-response signal
inclusion of all models with significant test results in reference set
stop of analysis, PoC not established
> 1 model in reference set
one model in reference set
model selection / model averaging
estimation of model parameters
estimation of target dose
precision of estimation (optional)
MCP-Step
Mod-Step

Figure 3.1: Flow Chart of MCP-Mod Approach

effect while still adhering to the overall type-I error. If that could be proven for at least one of the candidate models, the dose-response relationship is modelled by means of a parametric model that is either the candidate model that fits the data best or an average over those candidate models with a significant test result. On the basis of the fitted model, the target dose can be estimated via inverse regression techniques and precision of the estimates can be assessed if desired.

The explanations in this section are based on the papers of Bretz et al. (2005), Pinheiro et al. (2006) and Branson et al. (2003). Further considerations concerning the planning of the study and the robustness of the chosen design can be found in the paper of Dette et al. (2008).

3.1.1 Definition of the Candidate Models

The starting point of the planning phase is the determination of possible shapes for the dose-response model, i.e. one has to select functions $f(d, \boldsymbol{\theta})$ that fit to the prior suppositions of the functional dose-response relationship. Therefor, the same model shapes as presented for the simple parametric modelling approach in section 2.3.2 can be used. For the selection it is advisable to take into consideration prior knowledge about the true dose-response relationship on the one hand and to enable flexibility within the assumed range of shapes as much as possible on the other hand. It might also be reasonable to include several versions of the same model family, namely with different parameter specifications. But the more models are included in the candidate set, the less powerful the testing procedure is to differentiate between them and the stricter the multiplicity adjustment will be for the PoC part. However the latter fact is moderated by an increasing correlation between the models in the candidate set as the p-value is computed from the multivariate distribution of the test statistics (cf. equation (2.4)).

For each of the model shapes selected in the previous step, initial estimates for the parameters of the standardized model, the so-called "guesstimates", are to be computed on the basis of some prior knowledge about the true underlying dose-response curve. The prior knowledge can be taken for example from pharmacokinetic data as well as from the dose-response relationship itself that was identified for a similar compound. The theoretical derivation of these "guesstimates" for the choice of common models presented in this thesis can be found in section 2.3.2.

In R, the guesstimates can be obtained by means of the function `guesst` for a choice of parametric models including the ones presented in section 2.3.2. Note that for the correct computation, the function demands for the "expected percentages of the maximum effect achieved at [dose] d" (cf. Bornkamp et al., 2012) instead of the absolute values of the outcome variable.

The set of candidate models can be defined by aggregating the prespecified model shapes and corresponding guesstimates with information about the placebo effect and the maximum change from placebo via the `Mods` function.

3.1.2 Determining the Optimal Study Design

Once the set of candidate models $\mathcal{M} = \{M_m,\ m = 1, \dots, M\}$ has been prespecified, one can search for the optimal selection of dose groups to be included in the study and identify how to allocate the patients optimally to the selected dose groups. The identification of optimal design features is implemented in the function `optDesign`, which offers three different kinds of optimality criteria (specified via `designCrit = "Dopt" | "TD" | "Dopt& TD"`). Either the study design is optimized with regard to the estimation of the model parameters (D-optimality), with regard to the Target Dose (TD) estimation (TD-Optimality) or with regard to both. In practice, D-optimality signifies the minimization of a criterion which involves the variance of the model parameters whereas TD-Optimality means minimizing the length of the confidence interval for the TD as proposed by Dette et al. (2008).

Formally, the criterion for the D-Optimality is given by

$$\Psi(\xi, \boldsymbol{\theta}) = -\sum_{m=1}^{M} \frac{p_m}{k_m} \log\left(\det |V_m{}^{-}(\xi, \boldsymbol{\theta})|\right)$$

where $\xi = \{d_i, w_i\}_{i=0}^k$ contains the design information (dose groups $d_0, \dots, d_k$ with corresponding allocation weights $w_0, \dots, w_k$) and $V_m(\xi, \boldsymbol{\theta})$ is the covariance matrix of the parameter estimate belonging to model M_m as defined in equation (2.15).

A penalization for the complexity of the model can be included by choosing the k_m equal to the number of model parameters used in model M_m. The penalization can be suppressed by setting the values k_m equal to one (in R: option `standDopt = FALSE`). Furthermore, the models can be weighted by specifying appropriate model probabilities $p_m,\ m = 1, \dots, M$.

Alternatively, the criterion for the TD-Optimality can be expressed by

$$\Psi(\xi, \boldsymbol{\theta}) = \sum_{m=1}^{M} p_m \log(b_m^\top(\boldsymbol{\theta}) V_m{}^-(\xi, \boldsymbol{\theta}) b_m(\boldsymbol{\theta}))$$

with $b_m(\boldsymbol{\theta})$ as defined in equation (2.16) with respect to a particular model M_m.

The criterion for the joint optimization is a combination of the single criteria, given by

$$\Psi(\xi, \boldsymbol{\theta}) = \sum_{m=1}^{M} p_m \left(\frac{1}{2} \cdot \frac{-\log\left(\det |V_m(\xi, \boldsymbol{\theta})|\right)}{k_m} + \frac{1}{2} \cdot \log(b_m^\top(\boldsymbol{\theta}) V_m{}^-(\xi, \boldsymbol{\theta}) b_m(\boldsymbol{\theta})) \right)$$

again with the possibility to suppress the penalization by fixing $k_m = 1,\ m = 1, \dots, M$.

The optimal design for one of those optimality criteria is the one that minimizes the appropriate criterion $\Psi(\xi, \boldsymbol{\theta})$ with respect to the design vector ξ.

However, it must be mentioned that due to feasibility matters, it might be necessary to deviate from the optimal study design, for example if the manufacturing of the optimal dosages is not possible or due to technical restrictions. In these cases, it is recommended to evaluate the efficiency of the chosen design $\tilde{\xi}$ compared to the optimal design ξ_{opt}. The efficiency can be computed as the ratio of optimality criteria for the two models

$$\text{eff}(\tilde{\xi}, \boldsymbol{\theta})) = \frac{\Psi(\tilde{\xi}, \boldsymbol{\theta})}{\Psi(\xi_{\text{opt}}, \boldsymbol{\theta})} \ .$$

As in R, the `calcCrit` function outputs the criterion on the log-scale, the efficiency is obtained by `exp(calcCrit(design.actual,...)-calcCrit(design.opt,...))`.

3.1.3 Computation of the Optimal Contrasts

For every model that has been included in the candidate set, the optimal contrast coefficients have to be computed separately. A contrast is meant to be optimal with respect to a specific model if the resulting test has maximum power in case the assumed model is correct. This implies that a test for a null hypothesis $H_0 : \boldsymbol{c}^\top \boldsymbol{\mu} = 0$ with a model specific contrast vector $\boldsymbol{c}$ can be interpreted as a test of the assumed model versus a null model with a flat dose-response. The actual testing (including use of maximum statistic as global test statistic, joint distribution of the single contrast test statistics under null and alternative hypothesis, critical values) will be analogous to what has been described in subsection 2.2.6.

As the optimality criterion for the contrasts is related to the power, the distribution of the contrast test statistic for the respective model under the alternative hypothesis is decisive for the derivation of the optimal contrast. To recall what has been discussed in general for (multiple) contrast tests in subsection 2.2.6, note that under the null hypothesis, the test statistic for one single model (cf. formula (2.1)) is centrally t-distributed. Under the alternative hypothesis $H_1 : \boldsymbol{c}^\top \boldsymbol{\mu} \neq 0$, where $\boldsymbol{\mu} = (\mu_0, \dots, \mu_k)^\top = (f(d_0, \boldsymbol{\theta}), \dots, f(d_k, \boldsymbol{\theta}))^\top$ is denoting the vector of unknown treatment means under the assumed model, the test statistic follows a non-central t-distribution with non-centrality parameter

$$\tau = \tau(\boldsymbol{c}) = \frac{\boldsymbol{c}^\top \boldsymbol{\mu}}{\left(\sigma^2 \sum\limits_{i=0}^{k} \frac{c_i^2}{n_i}\right)^{\frac{1}{2}}} \ . \tag{3.1}$$

As shown by Abelson and Tukey (1963), the maximization of the power can be achieved by maximizing the non-centrality parameter which is again equivalent to the maximization of the correlation between the model specific contrast $\boldsymbol{c}_m$ and the standardized mean response $\boldsymbol{\mu}_m^0$ according to model m.
Hence the optimal contrast for a specific model (for a two-sided test) can be defined as

$$c_{\text{opt}}(f) = \arg\max_c \tau(\boldsymbol{c}) = \arg\max_c \frac{(\boldsymbol{c}^\top \boldsymbol{\mu})^2}{\sum\limits_{i=0}^{k} \frac{c_i^2}{n_i}}$$

with the additional condition that the coefficients of the contrast sum up to 0 (cf. first paragraph of subsection 2.2.6). But, as this only defines c_{opt} up to a multiplicative factor, it is further required that $\|c_{\text{opt}}\| = 1$ with $\|\cdot\|$ being the L_2-norm to make the optimal contrast unique (except for the sign as $\pm c_{\text{opt}}$ both are optimal).
For the one-sided test, $-c_{\text{opt}}$ is the optimal contrast for the alternative hypothesis $H_1^- : \boldsymbol{c}^\top \boldsymbol{\mu} < 0$ and c_{opt} is the optimal contrast for the alternative hypothesis $H_1^+ : \boldsymbol{c}^\top \boldsymbol{\mu} > 0$.
If a standardized version of the model exists, it suffices to find the optimal contrast for the standardized model as

$$\arg\max_c \frac{(\boldsymbol{c}^\top \boldsymbol{\mu})^2}{\sum\limits_{i=0}^{k} \frac{c_i^2}{n_i}} = \arg\max_c \theta_1^2 \frac{(\boldsymbol{c}^\top \boldsymbol{\mu}^0)^2}{\sum\limits_{i=0}^{k} \frac{c_i^2}{n_i}}$$

with $\boldsymbol{\mu}^0$ representing the vector of unknown standardized means $(f(d_0, \boldsymbol{\theta}^0), \dots, f(d_k, \boldsymbol{\theta}^0))$.

In the case of equal patient allocation $n_0, \dots, n_k = n$, the calculation of c_{opt} simplifies to

$$c_{\text{opt}}(f) = \arg\max_c \frac{n(\boldsymbol{c}^\top \boldsymbol{\mu})^2}{\sum\limits_{i=0}^{k} c_i^2} = \arg\max_c \ (\boldsymbol{c}^\top \boldsymbol{\mu})^2$$

due to the restriction $\|c_{\text{opt}}\| = 1$. By application of the Cauchy-Schwarz inequality (and the assumption that the coefficients of c sum up to 0) it follows that

$$(\boldsymbol{c}^\top \boldsymbol{\mu})^2 = \left(\boldsymbol{c}^\top (\boldsymbol{\mu} - \bar{\mu}\mathbf{1})\right)^2 \leq \|\boldsymbol{\mu} - \bar{\mu}\mathbf{1}\|^2$$

where $\bar{\mu}$ is denoting the overall mean across all treatment groups. This implies that in case of equal allocation a closed-form solution for the optimal contrast is given by

$$c_{\text{opt}} = \frac{\boldsymbol{\mu} - \bar{\mu}\mathbf{1}}{\|\boldsymbol{\mu} - \bar{\mu}\mathbf{1}\|} = \frac{\boldsymbol{\mu}^0 - \bar{\mu}^0\mathbf{1}}{\|\boldsymbol{\mu}^0 - \bar{\mu}^0\mathbf{1}\|} \ .$$

In general, if the group sample sizes are not equal, the optimal contrast has to be determined by numerical optimization. Therefor it is preferable to express the contrast vector by means of a parametrization function $c = h(\gamma)$ where the vector γ contains all of the $k-1$ free parameters of c. The other two parameters of c can be expressed as a function of the elements in γ because of the two restricting assumptions $\sum c_i = 0$ and $\sum c_i^2 = 1$.
In R, the computation of optimal contrasts can be handled with the function `optContr` using a quadratic programming algorithm. By choosing the option `type = "constrained"`, the optimization algorithm allows for a further restriction of the contrast coefficients, namely that the coefficients for the control groups need to have a different sign than the ones for the active dose groups.

As the contrast coefficients have to be fixed prior to the collection of any study data and only on the basis of some prior estimates for the model parameters, this introduces a possible risk for a loss of power due to weak prior knowledge. However, the MCP-Mod approach was shown to be robust against moderate misspecification of the prior estimates (e.g. see Pinheiro et al. (2006)). In case of unreliable information about the true dose-response curve in the planning phase of the study, it is advisable to include a choice of different parameter specifications for one model in the candidate set.

3.1.4 Sample Size Calculation

The last step in the planning phase of the study is the determination of the required sample size. For the MCP-Mod approach, the sample size can be chosen with respect to different criteria, dependent on what is the main focus of the trial. If the establishment of the PoC is considered of prime importance, the sample size can be calculated to meet a certain target power for the PoC test whereas for the target dose estimation, the sample size should be chosen in a way that provides a prespecified precision of the resulting estimate. A combination of these criteria is possible as well.

The starting point for the derivation of the required sample size N^* in order to achieve a prespecified target power π^* is to derive a formula for the power under the assumption of a single true model and afterwards generalizing it for a multiple-model scenario.
Suppose model M_m is the true underlying dose-response model with corresponding mean vector $\boldsymbol{\mu}_m = (f_m(d_0, \boldsymbol{\theta}_m), \ldots, f_m(d_k, \boldsymbol{\theta}_m))$. The power to detect a non-flat dose-response curve (i.e. to show the existence of dose-response) is the probability that the maximum test statistic exceeds the critical value $q_{1-\alpha}$ given $\boldsymbol{\mu}_m$ is the true mean vector

$$\pi_m(N) = \mathbb{P}(\max_{i=1,\ldots,M} T_i \geq q_{1-\alpha} \mid \boldsymbol{\mu} = \boldsymbol{\mu}_m) = 1 - \mathbb{P}(T_1 < q_{1-\alpha}, \ldots, T_M < q_{1-\alpha} \mid \boldsymbol{\mu} = \boldsymbol{\mu}_m) \ . \tag{3.2}$$

Analogously to the contrast test for one specific model, the joint distribution of all test statistics $T_1, \ldots, T_M$ under model M_m is a non-central (multivariate) t-distribution with $N-k$ degrees of freedom, correlation matrix $\boldsymbol{R} = (\rho_{ij}),\ i,j = 1, \ldots, M$ and non-centrality parameter $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$ where the single parameters τ_j are defined as in equation (3.1) and reduce to

$$\tau_j = \sqrt{n}\frac{\boldsymbol{c}_j^\top \boldsymbol{\mu}_j}{\sigma}, \quad j = 1, \ldots, M$$

in case of equal allocation to the dose groups. Note that $\boldsymbol{c}_m$ is the contrast vector for model M_m.
In practice, the power under one single true model as in equation (3.2) can be calculated via numerical integration (cf. Genz and Bretz (2000)).

As one of the main advantages of the MCP-Mod approach is the possibility to use different candidate models instead of one prespecified model, it is preferable to consider the vector of power values for all models in the candidate set $\boldsymbol{\pi}(N) = (\pi_1(N), \ldots, \pi_M(N))^\top$ and to define a monotonically increasing summary function $s : [0,1]^M \to [0,1]$ that combines the single power values to a measure for the overall power. That could be for example the minimum/maximum of all values, but also the (weighted) mean of all power values or any quantile. In any case, the chosen summary function should map into the range of individual power values, or in other words, they should not exceed or fall below the minimum and maximum power respectively.

This generalized power definition is also the basis for the sample size considerations. The required number of patients N^* is the smallest integer value that results in an overall power equal to or greater than the prespecified target power. Practically, this N^* is calculated using an iterative algorithm which starts with a given upper bound N^u and reduces the number of patients by 1 (or k in the case of equal allocation) until the corresponding overall power falls below the target power. The required sample size is then chosen as the smallest integer resulting in an overall power not less than the target power. For the algorithm to work, the upper bound N^u has to be chosen as the maximum of the sample sizes required to achieve the target power of π^* for the single contrast tests using the multiplicity adjusted critical value of the global test statistic.
An alternative approach uses a root-finding algorithm to solve the equation $s(\boldsymbol{\pi}(N^*)) - \pi^* = 0$ for N^* to find the required sample size. The latter method is also the one that is implemented in R in the function `sampSizeMCT`. It uses a bisection search algorithm and within each step, it calls the `powMCT` function for the calculation of the single power values. The summary function can be specified by means of the option `sumFct= "min" | "mean" | "max"`.

Another option, as already mentioned at the beginning of this subsection, is to aim at a certain level of precision for one or more estimates of interest (for example the estimates for the parameters in the dose-response model, the expected response for a given dose or the estimate for the MED). To address this matter, one would use the formulae derived in subsection 2.3.4 (either the variance formulae themselves or the length of the confidence intervals, calculated as the difference between the upper and lower limits) and express them as a function of the total number of patients N. By setting this function equal to the target value and transforming it adequately, the required sample size N^* can be derived with the help of a root finding algorithm.
In R, one can realize this by means of the `sampSize` function with a user-defined target function in `targFunc` which is supposed to achieve the target value defined in `target`.

3.1.5 Analysis of Study Data

Test for PoC via MCP

The final analysis of the collected data consists of two main steps.

The first step is the establishment of PoC via a multiple testing procedure. Therefor, the mean responses in the individual dose groups $\bar{Y}_0, \ldots, \bar{Y}_k$ are calculated and the covariance matrix $\boldsymbol{S}^2$ is estimated on the basis of the data according to the formula given in equation (2.2). These measures are then entered into the single test statistics (cf. equation (2.1)) and the resulting values of the test statistics are one by one compared against the (equicoordinate) $(1-\alpha)$-quantile of the M-dimensional t-distribution with correlation matrix $\boldsymbol{R}$ as defined in equation (2.3). Thus, it is possible to test if each of the models in the candidate set (seen individually) is significantly different from a flat dose-response curve given the observed data whilst controlling the FWER at a certain level α. All models which are shown to be statistically significant by means of the single contrast tests are included in the reference set for the subsequent modelling phase.

For the global null hypothesis of no overall dose-response, the maximum of all the single test statistics is compared against the same quantile as above. If the test statistic exceeds the quantile, the existence of a significant dose-response signal is proven and hence, PoC is established. If not, this means that no model is significantly different from a flat dose-response and hence the procedure is stopped after the first stage without being able to establish PoC.

Note that failing to show a dose-response effect might also be due to an insufficient sample size or a high variance in the collected data. Another reason might be that the models included in the candidate set don't describe the true dose-response shape appropriately.

Modelling of the Dose-Response Relationship & Estimation of the MED

The second step of the final analysis is modelling the dose-response relationship. Therefor, all the models in the reference set are taken into consideration. If more than one model was included in the reference set, one has to decide which of those shall be used for the dose estimation. This can either be the model with the smallest p-value as it is most likely to be (closest to) the true model or the best model with respect to some goodness-of-fit criterion (e.g. the Akaike Information Criterion (AIC)). The latter could be preferable in case the candidate models are more complex. Then the AIC identifies the model that shows the best trade-off between goodness of fit and complexity of the model. Another option to develop a final model is to use model averaging over all significant models, as for example in Verrier et al. (2014). As Bornkamp (2015) states in his paper, the usage of model averaging techniques is theoretically superior to model selection for the following two reasons. Firstly, a model selection process, whether based on an information criterion or some other indicator, is not necessarily robust. This means that small changes in the underlying data set can lead to substantially different models and by that to a different conclusion at the extreme. Furthermore, deriving confidence intervals without taking into account the model selection process may lead to overoptimistic confidence intervals (too narrow), i.e. to an incorrect coverage probability and an inflation of the type-I error. On the contrary, model averaging techniques offer the possibility to take into account the uncertainty in the model selection process and prevent potential bias. Concrete examples for model averaging are given

in a later section for the Klingenberg approach (subsection 3.2.3).
The selected model(s) is/are fitted to the data via GLS estimation and the MED for each model is estimated as described in subsection 2.3.3. In case the final model is obtained by model averaging, the final MED estimate is calculated as a weighted mean of the model-specific MEDs, using the same weights as in the model averaging process itself. The precision for the estimated MED as well as for the expected response at a certain dose can be assessed via bootstrap methods or using the asymptotic behaviour of the GLS estimates (see subsection 2.3.4).

In R, the contrast tests can be conducted via the function `MCTtest`; fitting one of the built-in models to the data can be realized via the function `fitMod`. The whole analysis procedure is also implemented in one single function called `MCPMod` which offers three different options for the selection of the final model (`selModel = "AIC"` for the model with the smallest AIC, `selModel = "maxT"` for the model with the greatest value of the test statistic or smallest p-value respectively or `selModel = "aveAIC"` for a weighted average of the significant models with model weights

$$w_i = \frac{\exp(-0.5 * AIC_i)}{\sum\limits_{j=1}^{k} \exp(-0.5 * AIC_j)} \ . \tag{3.3}$$

Additionally, it includes the estimation of the optimal dose, either the MED for a certain effect of `Delta` over placebo or the Effective Dose (ED) that produces a certain percentage `p` of the maximum effect over placebo.
The target dose can also be estimated separately using the R functions `TD` and `ED` respectively.

3.2 MCP-Mod Approach for Binary Distributed Outcomes

The original MCP-Mod approach as outlined in the previous section is constructed for homoscedastic normally distributed outcomes collected in a simple study setting, i.e. in a parallel group design with non-repeated measurements. However, the practical part of this thesis focuses on a binary data setting, i.e. on studies which have a responder rate as their primary outcome variable. Naively using the same methods for binary data, especially in the case of small sample sizes, may lead to questionable results. In this section, three main approaches to deal with binary outcome data in the framework of a unified dose-finding procedure are presented and their advantages and disadvantages are briefly discussed.
The first is the above mentioned naive application of the original MCP-Mod approach to binary data. The second approach was presented by Pinheiro et al. (2014) and enhances the original MCP-Mod approach in order to make it more generally applicable, e.g. particularly for binary distributed data, but also for count data, longitudinal data and even for time-to-event settings. It is based on the idea of transforming the data via an appropriate parametric model and using the essentially unmodified MCP-Mod methods on this parameter level. This can be justified by the fact that for most of the common estimation problems, the model parameters asymptotically follow a normal distribution.
The third approach by Klingenberg (2009) is in its basic idea similar to the original MCP-Mod approach, but is constructed for binary data. The candidate models are basically Generalized Linear Models (GLMs), but allow non-linear influencing variables such as the logarithm of the given dose. Furthermore, the dose-response signal is tested by means of the deviance difference between the assumed dose-response model and a model only including the intercept instead of a contrast test.

3.2.1 "Naive" Approach on Outcome Level

As mentioned in the introductory paragraph of this section, the simplest idea is to apply the unmodified MCP-Mod methods to the data in spite of the fact that the data is not normal but follows a binomial distribution. According to the de Moivre–Laplace theorem, the binomial distribution can be satisfactorily approximated by a normal distribution for a sufficiently large sample size n and a success probability p that is not too extreme (Krengel, 2002, Chapter 5). A rule of thumb says that an approximation can be seen as valid if $np(1-p) > 9$.
However, the following simulations show that in some cases, the type-I error for the PoC test may be inflated and the power may not reach the target level although the sample size was calculated in view of that. Simulations were done for the following four scenarios:

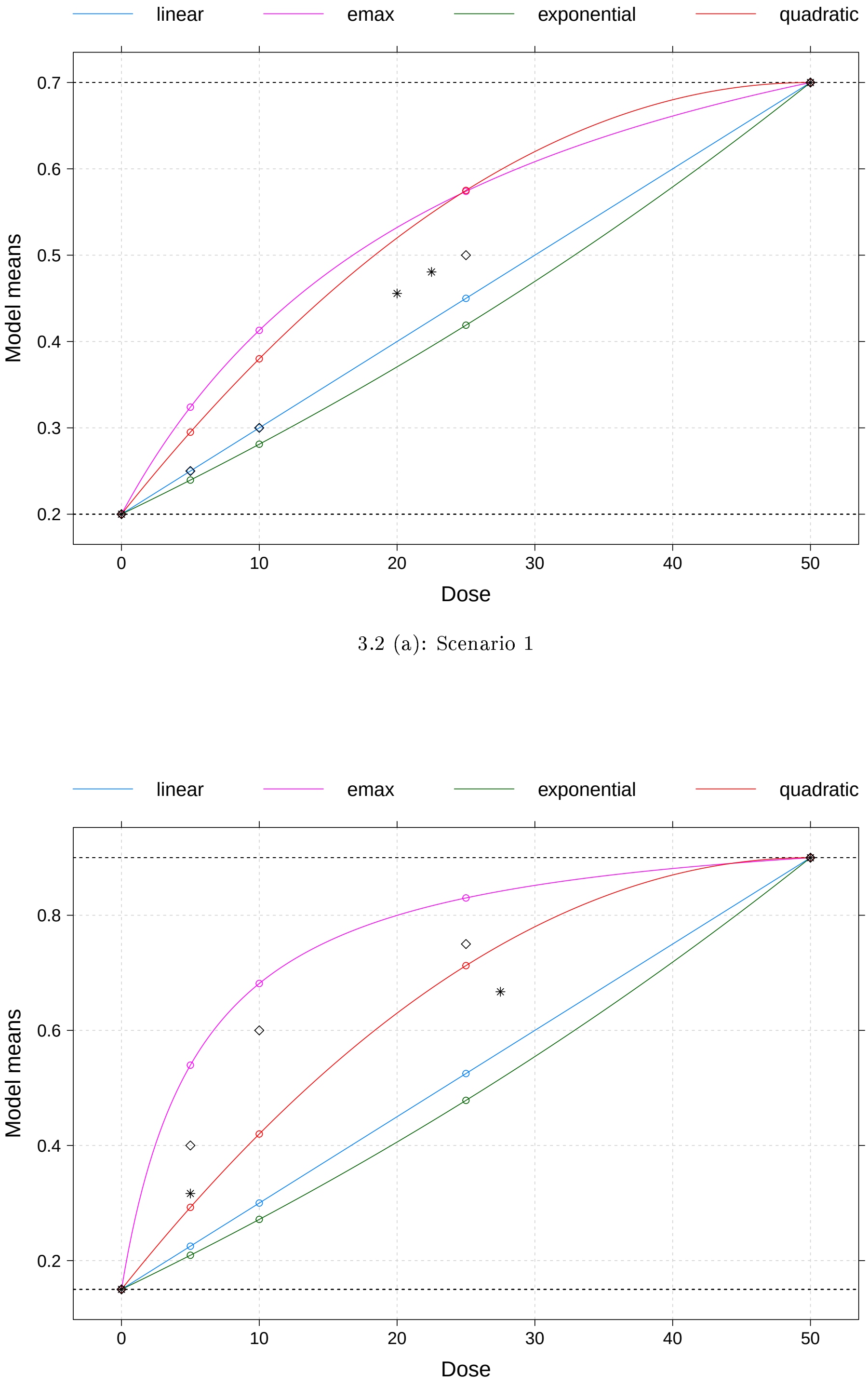
- **Scenario** 1: moderate response rates
  0 mg: 0.2, 5 mg: 0.25, 10 mg: 0.3, 25 mg: 0.5, 50 mg: 0.7

- **Scenario** 2: large maximum effect over placebo
  0 mg: 0.15, 5 mg: 0.4, 10 mg: 0.6, 25 mg: 0.75, 50 mg: 0.9

- **Scenario** 3: small response rates
  0 mg: 0.05, 5 mg: 0.08, 10 mg: 0.1, 25 mg: 0.15, 50 mg: 0.2

- **Scenario** 4: high response rates
  0 mg: 0.6, 5 mg: 0.7, 10 mg: 0.8, 25 mg: 0.85, 50 mg: 0.9

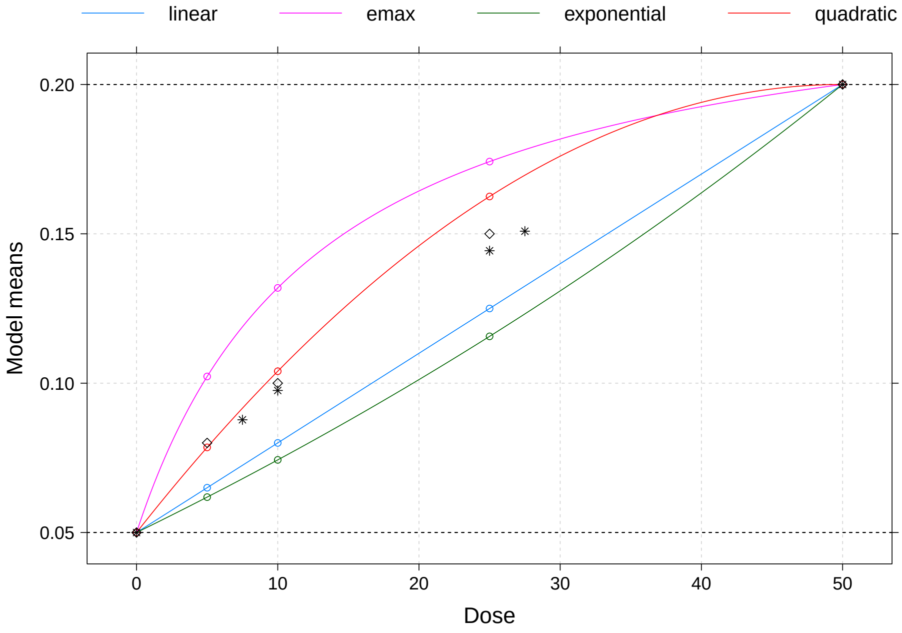
For each scenario, i.e. for each set of expected response rates, the set of candidate models consists of the linear model, the E_{max} model, the exponential model and the quadratic model to cover a wide range of possible dose-response shapes. The guesstimates for these models are derived on the basis of the above listed response rates resulting in the dose-response profiles visualized in Figures 3.2(a)-(d).



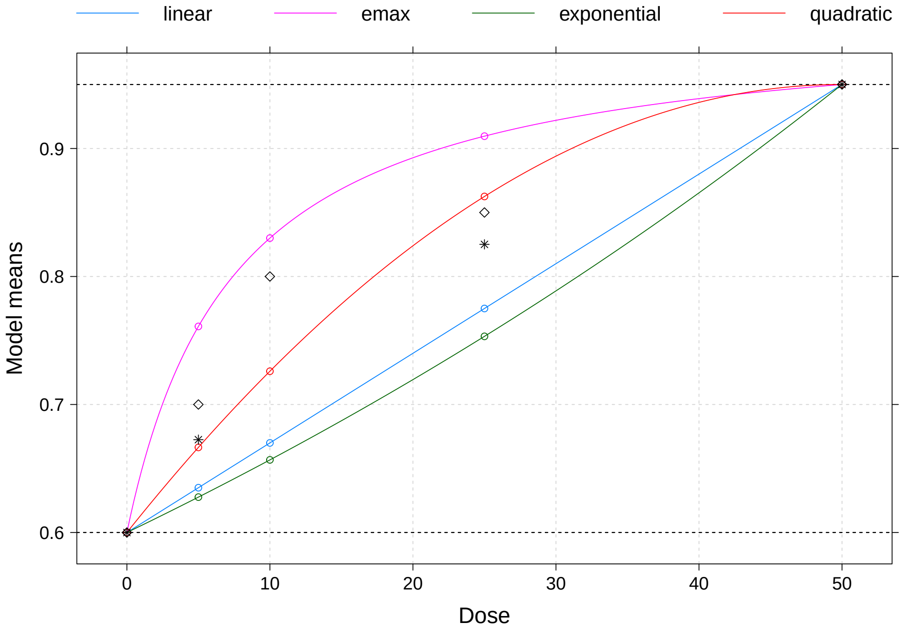
3.2 (a): Scenario 1

3.2 (b): Scenario 2

Figure 3.2: Candidate Models for the Naive Approach



3.2 (c): Scenario 3



3.2 (d): Scenario 4

The rhombi in the four plots mark the expected response rates that serve as prior information for the prespecification of the candidate models. Based on these candidate models, the optimal dose groups and the corresponding allocation ratios for the simulation process are derived and the total sample size needed to reach a mean power (averaged over all candidate models) of at least 80% is computed. After all design matters have been determined, the "study" data is simulated and analyzed according to the MCP-Mod approach for normal data (as described in section 3.1) multiple times, that is for 10 000 simulation runs. The significance level for the contrast test was defined as $\alpha = 5\%$. As generally, the optimal dose groups are not equal to the doses for which prior information is available, the response rates for the power simulation are chosen to be the means of the response values predicted by the four candidate models for the corresponding doses. These mean values are marked as asterisks in the plots of the candidate models 3.2(a)-(d). For the simulation of the actual type-I error, the response rates are set to be equal to the placebo response for all dose groups involved.

The power and type-I error values are estimated from the simulations as the percentage of simulation runs for which the null hypothesis has been rejected. As several models are tested in each run, different power definitions can be applied. In the following Table 3.1, two of those definitions are listed. The "average power" (the "average type-I error") represents the mean power (type-I error) over all four models. The definitions in columns 4 and 6 (i.e. the power to reject at least one of the model-specific null hypotheses and therewith achieving the PoC or the type-I error of erroneously rejecting at least one of the model-specific null hypotheses and hence also falsely confirming PoC) match the decision over PoC as proposed for the MCP-Mod approach. The results for other versions of the power (type-I error) term such as minimum/maximum power (type-I error) as well as the model-specific characteristics can be found in Tables A.1 and A.2 in appendix A. Additional to the simulations with an optimized sample size, simulations have been conducted with a remarkably increased sample size in order to investigate if this can compensate the inflation of the type-I error and is hence a problem of insufficient approximation and not a consequence of an inadequate testing procedure.

Table 3.1: Power and Type-I Error for the Naive Application of the MCP-Mod Approach to Binary Data

Scenario	**Sample Size**	**Average Power**	**Power to Reject at least one H_0**	**Average Type-I Error**	**Type-I Error of Rejecting at least one H_0**
Scenario 1	9 5 3 9	0.6818	0.7657	0.0425	0.0493
	20 each	0.9501	0.9646	0.0540	0.0659
	40 each	0.9992	0.9997	0.0444	0.0579
Scenario 2	2 2 2 3	0.8212	0.8308	0.0352	0.0519
Scenario 3	42 11 10 8 27 54	0.4385	0.5096	0.0298	0.0480
	60 each	0.8577	0.9017	0.0980	0.1254
Scenario 4	12 6 10 11	0.6164	0.6709	0.0638	0.0891
	70 each	0.9999	0.9999	0.0353	0.0548

The results in Table 3.1 allow the conclusion that for the scenario with moderate response rates (Scenario 1), the power stays below the target power of 80% for which it was originally powered (cf. first row in the table). Also the type-I error shows increased values, especially for the simulation with 20 patients per group. The inflation of the type-I error reduces with an increasing number of patients.

In contrast to this, the simulations for Scenario 2 (big change over placebo in the highest dose group) show that in some cases, the naive application of the original methods is acceptable and does not lead to a loss in power or an inflation of the type-I error respectively.
For the scenario with small response rates (Scenario 3), the power values are low despite of a relatively high number of simulated observations. The type-I error shows acceptable results for the simulation with the optimal sample size but is clearly inflated when increasing the sample size. The reason for this counter-intuitive behaviour is unclear.
The simulations for Scenario 4 again come up with low power values if the optimal sample size is used but this improves with an increasing number of observations. Also the type-I error shows better results for a higher sample size.
The simulations reveal that the naive usage of the original methods may lead to a worsening in terms of power. Furthermore, it seems to be impossible to control the type-I error, even with sample sizes which would theoretically allow the approximation of the binary distribution with a normal distribution. Hence, Pinheiro et al. recently published a proceeding paper that proposes to use an adequate transformation of the non-normal data up-front and to analyze the data on the parameter level of the transformation.

3.2.2 Pinheiro et al. (2014): Transformation to Parameter Level

As already mentioned, the approach developed by Pinheiro et al. is generally applicable to a broad range of endpoint types as binary or count data, survival data and longitudinal data (also if resulting from crossover studies). As the practical part of this thesis focuses on binary data, the following explanations of the generalized MCP-Mod approach will be illustrated for the case of binary data. However, the procedure for other data situations only differs in the transformation step at the beginning and hence, is very similar.

The basic idea of the extension to non-normal data is a transformation of the original data via a parametric model in a way that one of the parameters still captures the dose-response relationship (which was formerly the role of the expected response value in the original formulation). Formally, this means that the random variable Y describing the response follows a certain distribution with distribution function F

$$Y \sim F(\mu(x), \eta, z) \tag{3.4}$$

where $\mu(x)$ is the dose-response parameter, η the nuisance parameter and the vector of possible covariates is denoted by z.
As soon as the data has been transformed, everything is formulated with respect to the dose-response parameter $\mu(x)$, meaning that also the candidate models and the target effect are specified on this parameter level. This can sometimes be challenging and hence it is important to keep in mind that the dose-response parameter should be well interpretable.
Further demands on the parametrization are as follows. Firstly, it has to be an Analysis of Variance (ANOVA) parametrization ensuring that the dose-response for every single dose level is represented by a separate parameter. Furthermore, the estimate $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, \ldots, \hat{\mu}_k)^\top$, obtained for example via maximum likelihood estimation or GLS estimation, follows a normal distribution $N(\boldsymbol{\mu}, \boldsymbol{S})$ where $\boldsymbol{S}$ is the covariance matrix of $\hat{\boldsymbol{\mu}}$. The latter assumption is known to hold for most of the common paramet-

ric estimation problems including, inter alia, all generalized linear models and mixed-effect models. In the case of binary data, the transformation commonly used to obtain ANOVA-type parameters would be a logistic regression without intercept. Hence, the parameters the subsequent MCP-Mod methods are based on are the means of responses on the logit scale. This implies that also the candidate models have to be specified on a logit scale.

Apart from the prespecification of the candidate models and the computation of the optimal contrasts (will be explained later on in this section) that are necessary for the conduction of the contrast test in the analysis part, none of the design aspects of the study has been explicitly discussed by Pinheiro et al. (2014).

Problems arise when trying to plan the study with respect to which dose groups shall be included and how many patients are needed to reach the target power. In the case of homoscedastic normal data, the variance is equal to σ^2 for all dose groups and therefore, can be estimated by means of the common pooled variance estimate. But this is not generally true, particularly not in the case of binary data. Here, the variance in each dose group is directly dependent on the corresponding response rate which makes optimization more complicated because prior information about the nuisance parameters denoted by η is needed already at planning stage.

In the following, the computation of the optimal contrasts will be addressed and the two analysis steps of the extended MCP-Mod approach will be explained in more detail.

Computation of the Optimal Contrasts

When having prespecified a set of candidate models on the parameter level (for binary data that is on the logit scale), the next step is the computation of optimal contrasts. Analogously to the approach for normal data, a contrast for a specific model is considered optimal if the power of the corresponding univariate contrast test is maximal. Again, this can be obtained by maximizing the non-centrality parameter

$$\boldsymbol{\tau} = \boldsymbol{\tau}(\boldsymbol{c}) = \frac{\boldsymbol{c}^\top \boldsymbol{\mu}}{\sqrt{\boldsymbol{c}^\top \boldsymbol{S} \boldsymbol{c}}} \tag{3.5}$$

with respect to $\boldsymbol{c}$ where $\boldsymbol{\mu}$ is the mean vector on parameter level and $\boldsymbol{S}$ the covariance matrix of $\boldsymbol{\mu}$. Furthermore, the optimal contrast $\boldsymbol{c}_{\text{opt}}$ has to meet the condition $\boldsymbol{c}_{\text{opt}}^\top \mathbf{1} = 0$.

To directly include the condition on the contrast coefficients in the maximization problem, one of the coefficients has to be expressed by means of all other coefficients, e.g. $c_0 = -\sum_{i=1}^{k} c_i$ such that the reformulated contrast vector is given by

$$\tilde{\boldsymbol{c}} = \begin{pmatrix} -\sum\limits_{i=1}^{k} c_i \\ c_1 \\ \vdots \\ c_k \end{pmatrix} = (c_0, c_1, \dots, c_k)^\top \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{pmatrix} = \boldsymbol{c}^\top \boldsymbol{C}_0 \ .$$

Consequently, the maximization of equation (3.5) is equivalent to the maximization of

$$\frac{\left(\boldsymbol{c}^\top \boldsymbol{C}_0 \boldsymbol{\mu}\right)^2}{\boldsymbol{c}^\top \boldsymbol{C}_0 \boldsymbol{S} \boldsymbol{C}_0^\top \boldsymbol{c}}$$

which is the (only) solution of a generalized eigenvalue problem (see Ahrens and Läuter, 1981, formula (2.66)) given by

$$\boldsymbol{C}_0 \boldsymbol{\mu} \boldsymbol{\mu}^\top \boldsymbol{C}_0^\top \boldsymbol{x} = \lambda \boldsymbol{C}_0 \boldsymbol{S} \boldsymbol{C}_0^\top \boldsymbol{x} \ .$$

This implies that a closed form of the optimal contrast for model m is proportional to

$$\boldsymbol{c}_{\text{opt}} \propto \boldsymbol{S}^{-1} \left(\boldsymbol{\mu}_m^0 - \frac{\boldsymbol{\mu}_m^{0\ \top} \boldsymbol{S}^{-1} \boldsymbol{1}}{\boldsymbol{1}^\top \boldsymbol{S}^{-1} \boldsymbol{1}} \right) \ .$$

Again the condition $\sum c_i = 0$ has to be fulfil.
The derivation of this formula can be found in more detail in the appendix of Pinheiro et al. (2014).

Test for PoC via MCP

The actual contrast tests can be performed analogously to the procedure described in subsection 3.1.5 for the original MCP-Mod approach, i.e. the (model-specific) test statistics for testing the null hypotheses $H_0^m : \boldsymbol{c}_m^\top \boldsymbol{\mu} = 0$ versus the alternative hypotheses $H_1^m : \boldsymbol{c}_m^\top \boldsymbol{\mu} > 0$ are given by

$$T_m = \frac{\boldsymbol{c}_m^\top \hat{\boldsymbol{\mu}}}{[\boldsymbol{C}^\top \boldsymbol{S} \boldsymbol{C}]_{m,m}^{\frac{1}{2}}} \ , \qquad m = 1, \ldots, M$$

with $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M]$ being the matrix of all optimal contrast vectors and $[\boldsymbol{A}]_{m,m}$ referring to the m-th element on the diagonal of a matrix $\boldsymbol{A}$.
For testing the global null hypothesis, that is for establishing PoC, the maximum of these (model-specific) test statistics

$$T_{\text{global}} = \max_m T_m$$

is compared to a critical value derived from the asymptotic joint distribution of all single test statistics, a multivariate normal distribution.
The same critical value is also used for the individual contrasts tests when computing multiplicity adjusted p-values.

The only difference to the basic homoscedastic case is concerning the covariance matrix $\boldsymbol{S}$. In this basic setting, $\boldsymbol{S}$ was proportional to a diagonal matrix with elements equal to the reciprocal of the number of observations in the dose groups. More generally however, $\boldsymbol{S}$ may additionally depend on the nuisance parameter η and, as for example in the binary case, also on the expected response rates in the different dose groups. Therefore, in the planning phase of the study, guesstimates are needed also for all nuisance parameters that are contained in the covariance matrix. In most cases, prior information about those nuisance parameters is quite unreliable. The solution is that once the actual study data is available, the nuisance parameters can/should be re-estimated and used for the revaluation of the contrasts and the critical value involved in the contrast tests. Important to stress is that the re-estimation is stringently restricted to nuisance parameters as a re-calculation of the model parameters $\boldsymbol{\theta}$ or $\boldsymbol{\theta}^0$ respectively, would result in a serious inflation of the type-I error.

Modelling of the Dose-Response Relationship & Estimation of the MED

Also the fitting of the final dose-response model and the estimation of the target dose is generally similar to the original methods. Only the transformation of the observed data to the parameter level has to be run up-front so that the actual modelling can be done on the basis of the dose-response parameters. Hence, the Mod-part consists of two consecutive stages.

The ANOVA estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{S}}$, as already described in the introductory paragraph of this section, can be obtained using the common methods for the respective general parametric model (3.4), for example ML estimation, Partial Maximum Likelihood (pML) estimation or Generalized Estimating Equations (GEE).

The actual model can then be fitted to the resulting ANOVA estimates via GLS as in the original MCP-Mod approach by minimizing the following equation with respect to $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{\Psi}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} (\hat{\boldsymbol{\mu}} - \boldsymbol{f}(\boldsymbol{d},\boldsymbol{\theta}))^\top \hat{\boldsymbol{S}}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{f}(\boldsymbol{d},\boldsymbol{\theta})) \tag{3.6}$$

where $\boldsymbol{f}(\boldsymbol{d},\boldsymbol{\theta}) = (f(d_0,\boldsymbol{\theta}),\dots,f(d_k,\boldsymbol{\theta}))^\top$.

The reason for choosing this two-stage fitting approach instead of a standard ML estimation is that the optimization has to be conducted with respect to only $k+1$ different parameters (as many as there are different dose levels included) whereas the ML estimation is based on the full likelihood depending on the complete data. For the same reason, it is preferred to use a generalized model selection criterion as the Generalized Akaike Information Criterion (gAIC) defined as

$$\hat{\Psi}(\hat{\boldsymbol{\theta}}) + 2\dim(\boldsymbol{\theta})$$

for the selection of the final dose-response model.

The motivation of preferring the GLS estimation to other estimation methods is that it produces similar results as the ML estimation; in the case of homoscedastic normal data, the results are even identical. The same yields for the model selection criterion: for homoscedastic normal data, the gAIC is equal to the AIC.

Once a final model has been worked out and fitted to the data, the MED is estimated as described in subsection 2.3.3. Note that the clinically relevant improvement over placebo has to be defined on parameter level, i.e. as a difference in the dose-response parameter compared to placebo.

Precision of Estimation

As for the basic setting, the precision for the extended version of the MCP-Mod approach can be assessed on the basis of the asymptotic normality of the estimator given in equation (3.6)

$$\sqrt{a_n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, (\boldsymbol{F}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}^{-1}\boldsymbol{F}(\boldsymbol{\theta}_0)^\top)^{-1})$$

with a_n being a non-decreasing sequence fulfilling $a_n \xrightarrow{n\to\infty} \infty$ and $a_n\boldsymbol{S} \xrightarrow{P} \boldsymbol{\Sigma}$.

Alternatively, a parametric bootstrap method can be used. Herein, one element of the bootstrap sample of model parameters is generated by sampling from the multivariate normal distribution of the ANOVA estimates and estimating the model parameters on their basis via GLS methods. Confidence

intervals can then be obtained by determining the appropriate empirical quantile of the bootstrap sample.
Simulations show that for small sample sizes, the bootstrap method is preferable over the asymptotic procedure even though it is computationally more complex (cf. Pinheiro et al. (2014, page 16)).

For the practical implementation of the generalized MCP-Mod approach, the same R functions of the DoseFinding package can be applied. In the function calls, the results of the parametrization procedure, i.e. the estimate for the dose-response parameter and its covariance matrix, need to be specified instead of the original data. For the analysis functions (`MCTtest`, `fitMod` and `MCPMod`), the option `type = "general"` has to be selected. This implies that on the one hand, the model fitting is conducted by means of GLS instead of ordinary least squares estimation and on the other hand, the functions for the testing procedure skip the fitting of an ANOVA model and interpret the entered responses as the results of the transformation conducted beforehand.

3.2.3 Klingenberg (2009)

The third approach that combines a test for PoC based on a set of candidate models with the subsequent fitting of the best model is the approach presented by Klingenberg (2009). Contrary to the previous ones, the approach as presented in the original paper is specific for binary data collected under a parallel group design.
The methods of this approach have already been implemented in R by the author of the paper and made available at `http://sites.williams.edu/bklingen/research/poc/rcode/` including additional explanations and example code.

Candidate Models

As already mentioned, the starting point is a binary outcome variable, for example a responder variable indicating if the patient experienced a certain improvement in a specific (laboratory/score/...) value or not. For patient j receiving dose d_i, the response is denoted by Y_{ij} with $i = 0, \dots, k$ and $j = 1, \dots, n_i$. The responses are assumed to be independent within and across the different dose groups. The candidate models for the dose-response curve are defined directly on the response level, that is they model the success probabilities $\pi(d_i) = \mathbb{P}(Y_{ij} = 1)$, $i = 0, \dots, k$. The structure of the models is given by a link function (log-link, logit-link, identity-link, ...) and a predictor describing the influence of the dose, i.e. it is structurally similar to a GLM. If the predictor is linear, the model is a GLM by construction.
The number of parameters in the predictor is restricted by the number of different dose groups $k + 1$. When defining a complex predictor with more than $k + 1$ parameters, this can lead to problems of overfitting. Concerning the decision which models to include in the candidate set, it is advisable to cover a broad range of different dose-response profiles, but matching the anticipations of the clinical team. A list of possible candidate models (Table 3.2) is extracted from Klingenberg (2009, page 277).

When plotting those models prior to study start, initial estimates for the model parameters are needed. They can be derived from "educated guesses" of the placebo and maximum effect in the case of a model with only two parameters. For a model with three parameters, information about the dose which is expected to result in the maximum effect is required in addition. However, contrary to the MCP-Mod approach, these estimates are only needed for the premature visualization of the models, but are not involved in the analysis of the data. The estimation of the prior guesses and the plotting of the resulting models is implemented in the function `plotModels` for a set of different link functions and structures of the predictor. The plots for the candidate models listed in Table 3.2 are presented in Figure 3.3.

Table 3.2: Examples of Candidate Models for the Klingenberg Approach

Model	**Link Function**	**Predictor**
Model 1	Logit	$\beta_0 + \beta_1 d$
Model 2	Logit	$\beta_0 + \beta_1 \sqrt{d}$
Model 3	Logit	$\beta_0 + \beta_1 \log(d+1)$
Model 4	Logit	$\beta_0 + \frac{\beta_1}{\sqrt{d+1}}$
Model 5	Logit	$\beta_0 + \frac{\beta_1}{(d+1)}$
Model 6	Log	$\beta_0 + \beta_1 d$
Model 7	Identity	$\beta_0 + \beta_1 \exp(\exp(\frac{d}{\max d}))$
Model 8	Logit	$\beta_0 + \beta_1 d + \beta_2 d^2$
Model 9	Logit	$\beta_0 + \beta_1 \log(d+1) + \frac{\beta_2}{(d+1)}$
Model 10	Logit	$\beta_0 + \beta_1 \log(d+1) + \beta_2 d$

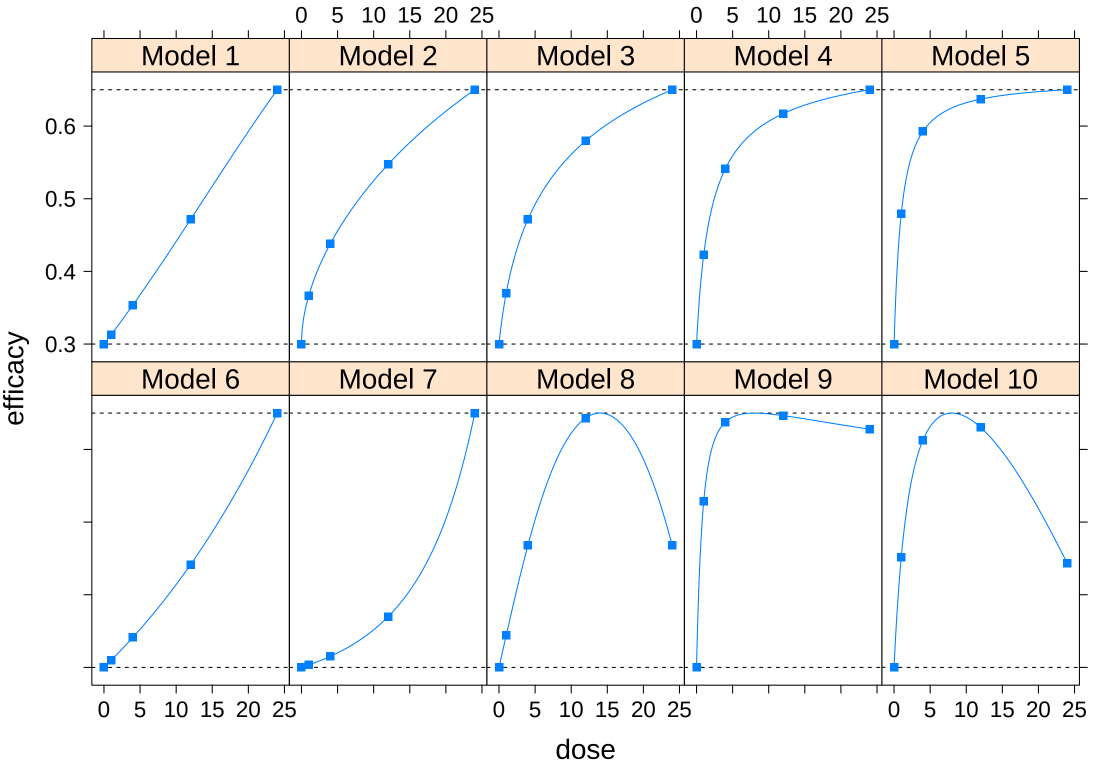


Figure 3.3: Examples of Candidate Models for the Klingenberg Approach

The filled squares in the plots mark the responses for the set of doses to be included in the study (specified in `dose`). As apparent from the plots, this set of candidate models already represents a wide range of dose-response profiles, including some non-monotonous scenarios.

Test for PoC via a Permutation Test

Analogously to the MCP-Mod approach, the first analysis step is to conduct separate tests for each of the candidate models to investigate the existence of a potential dose-response signal. But instead of a contrast test, one uses the following signed and penalized likelihood ratio statistic to test the model-specific null hypotheses $H_0^m : \pi_m(d_i) = \beta_0,\ m = 1, \dots, M$

$$T_m = (-1)^{I(\hat{\pi}_m(d_{\max}) \leq \hat{\pi}_m(d_0))} \{-2[\log L(\boldsymbol{y}, \boldsymbol{n}, M_0) - \log L(\boldsymbol{y}, \boldsymbol{n}, M_m)]\} - 2\mathrm{df}_m \tag{3.7}$$

with $I(\cdot)$ denoting the indicator function taking the value 1 if $\hat{\pi}_m(d_{\max}) \leq \hat{\pi}_m(d_0)$ is true or 0 else and $\hat{\pi}_m(d)$ is the ML estimate for the success probability $\pi(d)$ for dose d assuming model M_m. Furthermore, $L(\boldsymbol{y}, \boldsymbol{n}, M_m)$ is the maximum of the binomial likelihood under the assumed model M_m if a number of $\boldsymbol{y} = (y_0, \dots, y_k)^\top$ successes have been observed for the $k+1$ dose groups in $\boldsymbol{n} = (n_0, \dots, n_k)^\top$ patients respectively. This test statistic is constructed to compare a specific model M_m with the null model $M_0 : \pi_m(d_i) = \beta_0$ via the deviance difference between these two models (part in curly brackets). As the response is assumed to be binomially distributed, the deviance difference is explicitly given by the following formula

$$2 \sum_{i=0}^{k} y_i \log \left(\frac{\hat{\pi}_m(d_i)}{\hat{\pi}_0(d_i)} \right) + 2 \sum_{i=0}^{k} (n_i - y_i) \log \left(\frac{1 - \hat{\pi}_m(d_i)}{1 - \hat{\pi}_0(d_i)} \right)$$

with $\hat{\pi}_0(d) = \frac{1}{N} \sum_{i=0}^{k} y_i$ being the mean number of successes across all doses; $N = \sum_{i=0}^{k} n_i$.
The sign of the test statistic is intended to restrict a positive test decision to the existence of a positive dose effect, implying that the outcome variable Y_{ij} has to be coded such that high probabilities $\pi(d)$ are desirable. It is achieved by considering the dose effect under model M_m positive when $\hat{\pi}_m(d_{\max}) > \hat{\pi}_m(d_0)$ is met. Here, $d_{\max}$ represents the dose that maximizes the absolute difference of the associated effect over placebo $\arg\max_d |\hat{\pi}_m(d) - \hat{\pi}_m(d_0)|$. Note that this definition still covers dose-response profiles starting with a relatively small negative effect (called "J-shaped" profiles) but excludes those where the extend of the negative effect compared to placebo is too large (which is the case for some quadratic models).
Furthermore, subtracting two times the degrees of freedom of T_m (which are equal to the difference in the number of parameters involved in the two models) from the signed deviance difference signifies a penalization for complex models.

As the exact distribution of the test statistic in equation (3.7) is not known, the derivation of an exact p-value can only be attained via permutation. Therefor, one repeatedly arranges a random permutation of the patients' assignments to the different dose groups and subsequently calculates the test statistics for those permutations, denoted as $(T_1^{(b)}, \dots, T_M^{(b)})$ for the b-th of B permutations. The (estimated) p-value for the observed test statistic T_m^{obs} can then be computed as follows

$$p_m^{\text{obs}} = \frac{1}{B} \sum_{b=1}^{B} I(T_m^{(b)} \geq T_m^{\text{obs}})$$

where again, $I(\cdot)$ denotes the indicator function. In the following, it is referred to as "raw p-value". The described approach can be justified by the fact that, under the null hypothesis, the response values are independent of the given dose and hence interchangeable.

Another option would be to use the asymptotic distribution of T_m. It is well-known that the likelihood ratio statistic (part in curly brackets) is asymptotically chi-square distributed with df_m degrees of freedom. Consequently, the asymptotic p-value for T_m is given by

$$p_m = \begin{cases} \frac{1}{2} + \frac{1}{2}\mathbb{P}\left(\chi^2_{\mathrm{df}_m} \leq -(T_m + 2\mathrm{df}_m)\right) & \text{if } T_m + 2\mathrm{df}_m \leq 0 \\ \frac{1}{2}\mathbb{P}\left(\chi^2_{\mathrm{df}_m} \geq T_m + 2\mathrm{df}_m\right) & \text{if } T_m + 2\mathrm{df}_m > 0 \end{cases}$$

with $\chi^2_{\mathrm{df}_m}$ being a chi-square distributed random variable with df_m degrees of freedom.

The last paragraph dealt with the procedure to derive p-values for the model-specific test statistics. However, the main aim of the testing step is the establishment of PoC. It is done by comparing the minimum of the individual p-values with an appropriate critical value c, i.e. PoC is established if $\min\limits_m p_m \leq c$. This is equivalent to the usage of the maximum statistic in the MCP-Mod approach. Herein, the right choice for c ensures the preservation of the overall type-I error.

As for the individual p-values, the distribution of this minimum p-value can be estimated by means of its permutational distribution. For each permutation $b \in \{1, \dots, B\}$, the minimum p-value is denoted by $\min\limits_m p_m^{(b)}$ where

$$p_m^{(b)} = \frac{1}{B}\sum_{l=1}^{B} I(T_m^{(l)} \geq T_m^{(b)})$$

is the p-value that corresponds to the test statistic $T_m^{(b)}$ for model M_m under the b-th permutation. To ensure that the type-I error of falsely declaring PoC is kept below the global significance level of α, c has to be equal to the α-percentile of the distribution of $\min\limits_m p_m$.

Alternatively to the adjustment of the critical value, one can use the step-down procedure proposed by Westfall and Young (1993) for the direct adjustment of the p-values as presented in subsection 2.2.3. The procedure adjusts the p-values in an ordered fashion, starting with the one corresponding to the most significant model. The adjustment of this minimum p-value is based on the permutational minimum p-value distribution; all subsequent adjustments are carried out on the basis of stepwise reduced sets. This implies that the adjusted version of the i-th smallest p-value is equal to "the proportion of permutations for which the minimum p-value over the $[M - i + 1]$ remaining models is smaller than the observed one" (Klingenberg, 2009).

As already stated in subsection 2.2.3, the stepwise adjustment procedure ensures the preservation of the FWER at a specified level α when conducting all model-specific tests simultaneously. At the same time, the FWER for the PoC test is controlled by rejecting the global hypothesis if one of the adjusted p-values does not exceed the global significance level α.

Modelling of the Dose-Response Relationship & Estimation of the MED

Analogously to the previously described approaches, the final model can either be the best of all candidate models or a model obtained by averaging over all models that are significantly different from a model with intercept only. Methods for the fitting of the final model are not mentioned in the paper. However, the function that is implemented to fit the dose-response model in R calls the `glm` function which uses Iteratively Reweighted Least Squares (IWLS) estimation.

For the determination of the MED under a specific model M_m, one can use the same definition and estimator as described in subsection 2.3.3, formulae (2.13) and (2.14b) respectively. If the assumed model is relatively simple with respect to the predictor, as it is the case for models M_1 or M_6 in Table 3.2, the MED estimate can be obtained by solving a polynomial. For more complex predictors, numerical optimization methods such as the gradient descent, Newton's method or the Quasi-Newton method are necessary.
In case the final model is not the best of all candidate models but the result of model averaging, the true MED is also a (weighted) average of the MED estimates from each model (if existent)

$$\widehat{\text{MED}} = \frac{\sum\limits_m w_m \widehat{\text{MED}}_m}{\sum\limits_m w_m} \ .$$

Herein, the weights w_m have to be chosen suitably. Common examples are

- $w_m = \exp(\frac{1}{2}T_m)$,
  such that a ratio of weights is equal to the likelihood ratio of two models provided that they have the same number of parameters,

- $w_m \propto \exp(-\frac{1}{2}IC_m)$
  with IC_m being an information criterion for model M_m (e.g. AIC or BIC; the `MCPMod` function for the original MCP-Mod approach uses AIC); see Bornkamp (2015),

- $w_m = \frac{\mathbb{P}(M_m)\exp(-\frac{1}{2}IC_m)}{\sum_{l=1}^{K}\mathbb{P}(M_l)\exp(-\frac{1}{2}IC_l)}$
  with prior model weights $\mathbb{P}(M_m)$ which is a generalized version of the previous weights and has been proposed in Bornkamp et al. (2009),

- $w_m = \mathbb{P}(M_m|D)$,
  the posterior distribution of the model given the observed data; a Bayesian approach described for example in Hoeting et al. (1999).

Although only presented for binary data, this approach can also be applied for other data situations such as count data, other non-normal continuous data and repeated measurements, as long as the model fitting process is not too complex. Some of the mentioned data situations demand the usage of non-likelihood-based estimation methods for the model parameters, for example via GEEs. If this is the case, the test statistic proposed in the original paper can be replaced by a penalized generalized score statistic (see Boos, 1992).

Realization in R

As already stated at the beginning of this section, the R implementation of the Klingenberg method is available on the author's homepage. The candidate models can be specified by uniting the required information (distribution family, link function, structure of the predictor) in a `list` object. The specifications are analogous to those of the `glm` function to fit a Generalized Linear Model. If no predictor is specified, a linear relationship is assumed. Models of the form $a + b(dose + off)^p$ can be defined by the command `model=pow(dose,p, <off, dmax>)` where `off` is optional. If `p=0`, the model includes the logarithm of the dose instead of the original dose. In addition, entering a value for `dmax` implies that the maximum response is assumed at the specified dose instead of the highest one. Finally, all models are stored together in a `list` object. Optionally, one can provide a label for each candidate model in the same statement. The models can be plotted by means of the `plotModels` function for a given placebo effect and a maximum effect (to be specified in the arguments `low` and `high` respectively).

For the analysis of the data, the response data has to be handed over in the form of a matrix. This matrix has to contain the number of responders for every dose group in the first column and the number of non-responders in the second column. Hence, the row sums are equal to the sample sizes in the individual dose groups. The permutation test is implemented in the `permT` function. It first fits the models in the candidate set to the data and then computes the test statistics and the corresponding adjusted p-values. Besides, it is able to determine the MED for a given target effect entered in the `clinRel` argument. A plot of the most significant model can be obtained by using the `plot` with the object resulting of the `permT` function as a parameter.

For more detailed information, see the example code on the homepage.

Chapter 4

Simulations

The present chapter contains the descriptions and results of simulations conducted to investigate certain aspects of the approaches presented in this thesis. All investigations are made in a setting of binary outcome data collected in a parallel group design.
The aim of the first set of simulations is to compare the methods for binary data presented in section 3.2 with regard to their power and type-I error results as well as the precision of the target dose estimates. As already seen in subsection 3.2.1, the naive application of the original MCP-Mod methods to binary distributed outcome data leads to a loss of power and in some cases also to a substantial inflation of the type-I error. Hence, this approach is not further investigated. The simulations are restricted to the enhanced MCP-Mod method for binary data developed by Pinheiro et al. and the Klingenberg approach.
The second section of this chapter is addressing two different approaches for combining the target dose results of two separate studies to ideally come to a common target dose.

The underlying scenario for the simulations is the following:
The aim is to find the optimal dose for a new drug in patients suffering from a chronic disease. From other substances in this indication, it can be suspected that the patients' reactions will be different depending on whether or not they were unsuccessfully pretreated with a certain agent. Therefore, dose-finding will be initially done independently for the two subpopulations. Patients who failed to respond to the previous treatment are called "Failures", those that have not been pretreated are referred to as "Naives". The Naives will be randomized into five dose groups: 0 mg, 90 mg, 120 mg, 180 mg and 240 mg; the Failures will be assigned to only four different dose groups: 0 mg, 90 mg, 150 mg and 240 mg. For each of the subpopulations, a high and a low response scenario will be investigated. The assumed response rates characterizing these scenarios are presented in Table 4.1.

Table 4.1: Assumed Response Rates for the Simulations

Population	**Scenario**	**0 mg**	**90 mg**	**120 mg**	**150 mg**	**180 mg**	**240 mg**
Naives	High Scenario	0.3	(0.3)	0.3		0.5	0.7
	Low Scenario	0.3	(0.3)	0.3		0.5	0.6
Failures	High Scenario	0.25	(0.25)	0.25	(0.35)	0.45	0.65
	Low Scenario	0.25	(0.25)	0.25	(0.35)	0.45	0.55

The numbers not written in parentheses (response rates for dose groups 0 mg, 120 mg, 180 mg and 240 mg) are used for the (pre-)specification of the candidate models, and hence represent the prior knowledge for the estimation of the initial model parameters. The additional ones in parentheses are only used for the generation of the data.
Not taking into account the latter information for the prespecification of the candidate models reflects

the situation when prior knowledge is available, for example from existing study data for similar compounds, but the doses that are selected for the actual trial are not the same as the doses for which the prior information is given.

Although the subpopulations potentially react differently on the drug under investigation, the ideal case would be to identify a common optimal dose for the whole population. Therefore, two different possibilities for the development of a universal dose recommendation are investigated by means of simulations. Roughly spoken, on the one hand, one could aim for a pooled analysis, that is to fit an overall model to the pooled data from both subpopulations and to derive the optimal dose from the resulting dose-response curve. On the other hand, separate dose-response models could be fitted and given they are not "too different", the optimal doses resulting from these models are in a way combined to find one common optimal dose.

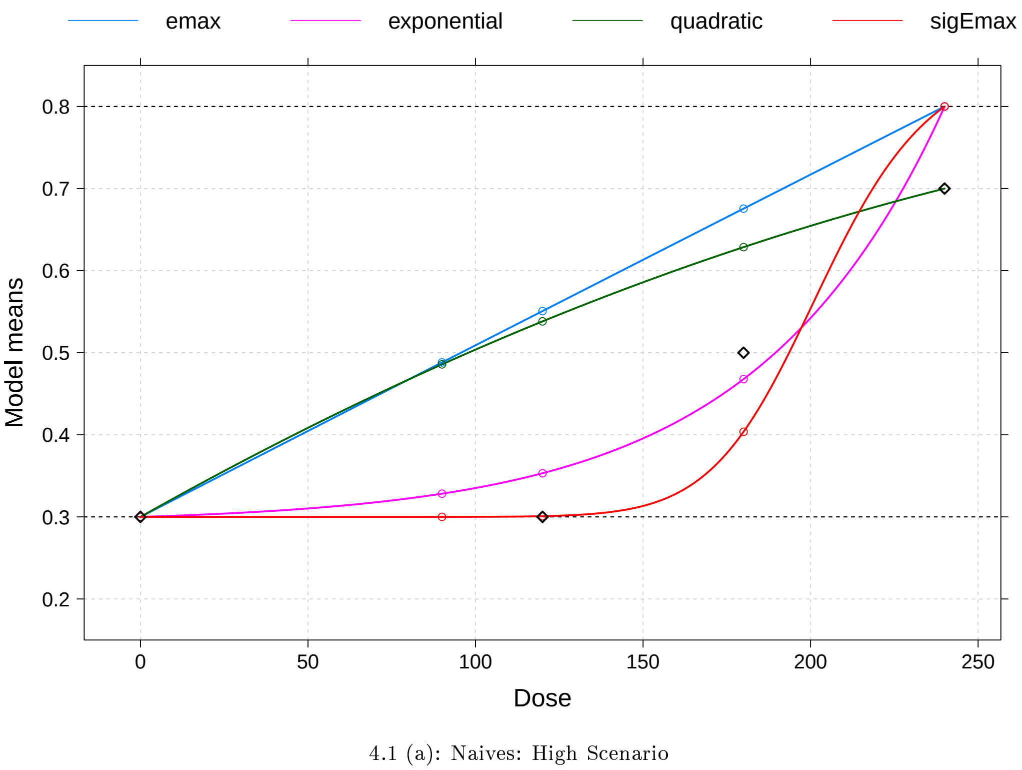
In the following, the candidate models for the two approaches that will be compared for the use in binary data are presented and illustrated by means of plots. The candidate models for the generalized MCP-Mod approach are also used in the second part for the investigations concerning the combination of target dose results.
In the first section of this chapter, the simulation macros to assess the performance of the methods are explained and the comparison between the two methods is drawn on the basis of the obtained results. The second section deals with the combination of study results. The principle of the two approaches is delineated and their performance with respect to power, type-I error and precision of the MED estimates is investigated via similar simulations as used for comparison part.

Candidate Models for the Generalized MCP-Mod Approach

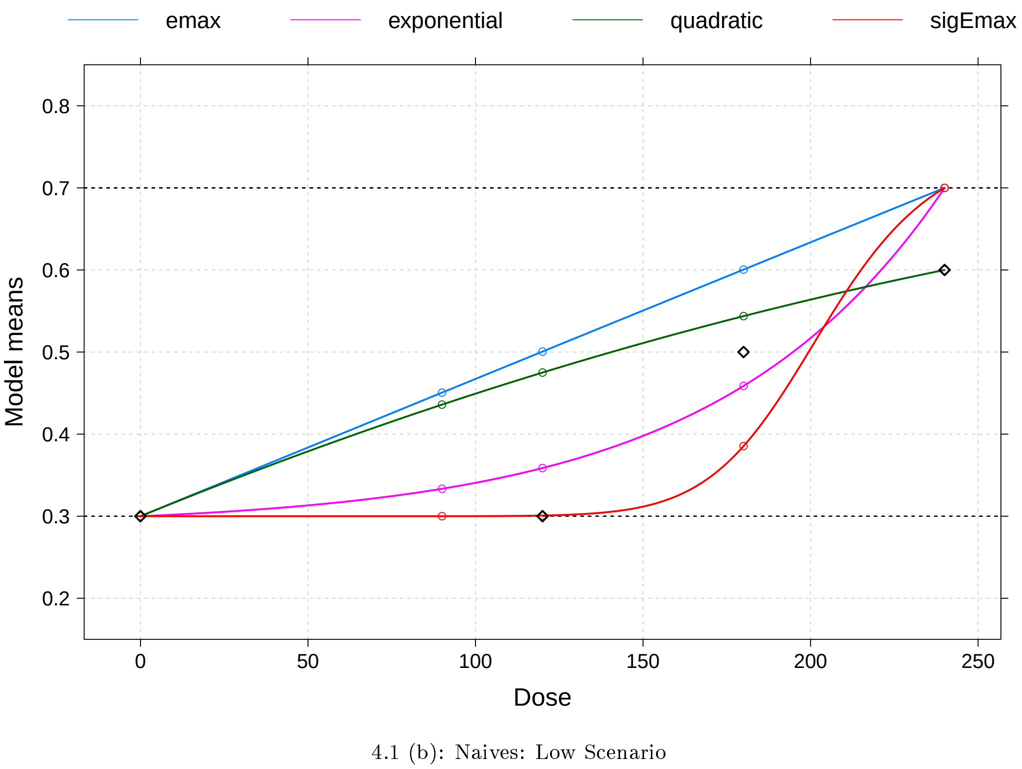
The candidate models for the generalized MCP-Mod approach are defined on response level instead of on parameter level to achieve a better comparability of the models over the two approaches. As a consequence, the contrasts that result from these candidate models are also defined on response level. Nevertheless, they can be used for the contrast test on parameter level as applying the logit function to the models on response level does not essentially change their shapes. The same proceedings can be found in the example for binary data given in the paper by Pinheiro et al. (2014, Section 4.2.1).

The candidate set for the generalized MCP-Mod approach comprises the $\mathrm{E_{max}}$ model, the sigmoid $\mathrm{E_{max}}$ model, the exponential model and the quadratic model with initial parameters derived from the assumptions given in Table 4.1 (numbers that are not in parentheses).
Important to note is that the `guesst` function in R that is used for the estimation of the initial parameters can only take into account the percentages of the maximum effect over placebo associated with a certain set of doses if they are not zero, that is if the responses in the active dose groups are different from the placebo response. Therefore, to ensure that the information for the 120 mg dose group is considered for the estimation of the initial model parameters, the percentage of the maximum effect for this dose group is minimally increased by 0.1 percentage points.
Additionally, in order to determine the location and scale parameters, the maximum effect over placebo is defined as the difference between placebo and maximum response plus 0.1 for the Naives and plus

0.05 for the Failures. The resulting candidate shapes on response level are depicted in Figure 4.1(a)-(d).

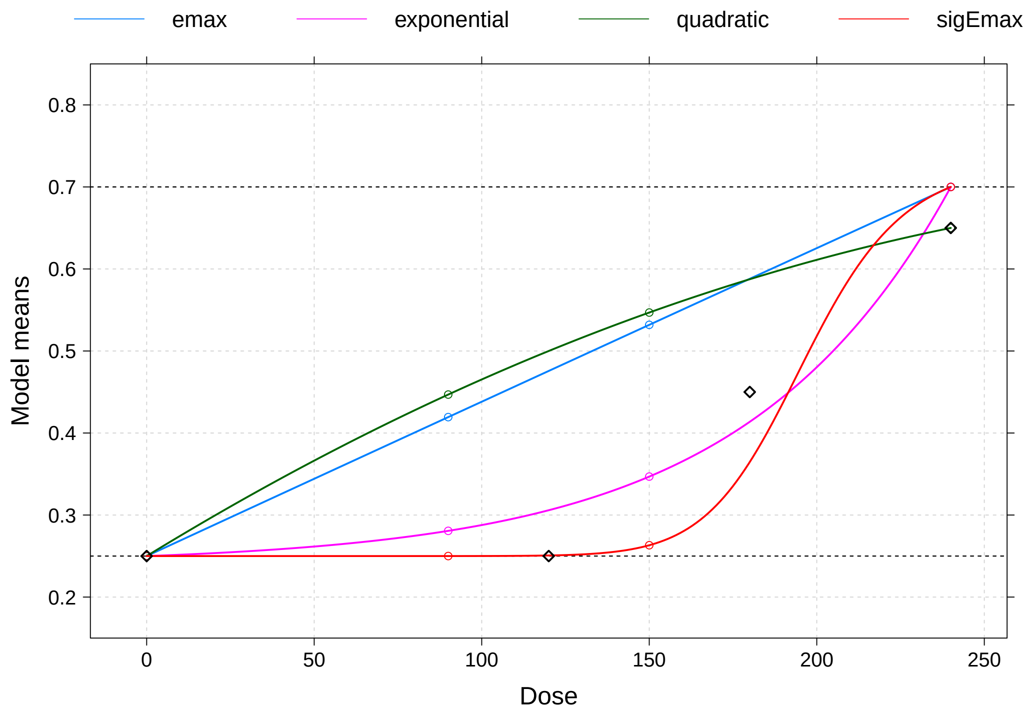


4.1 (a): Naives: High Scenario

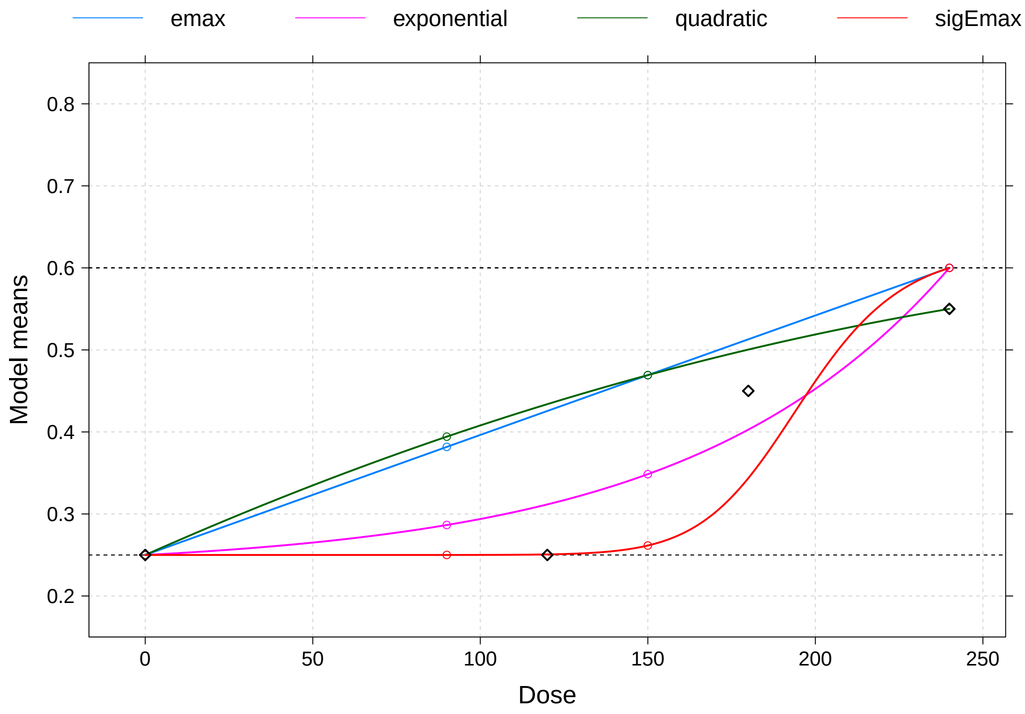


4.1 (b): Naives: Low Scenario

Figure 4.1: Candidate Models for the Simulations: Generalized MCP-Mod Approach



4.1 (c): Failures: High Scenario



4.1 (d): Failures: Low Scenario

As visible from those plots, the set of candidate models covers a broad range of dose-response profiles; the $\mathrm{E}_{\mathrm{max}}$ model (blue curve) describes a relationship that is almost linear, the quadratic model (green curve) represents a concave dose-response shape and the exponential (pink curve) and the sigmoid $\mathrm{E}_{\mathrm{max}}$ model (red curve) a convex one. The prior knowledge used for the computation of the guesstimates is added to the plots as black rhombi.

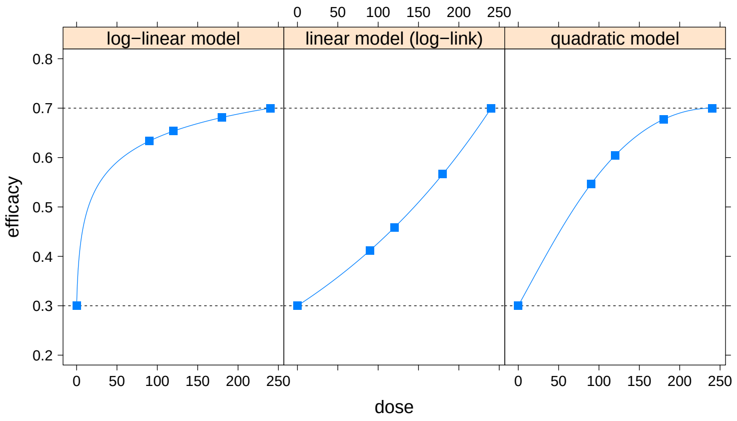
Candidate Models for the Klingenberg Approach

As already stated in subsection 3.2.3, the candidate models for the Klingenberg approach are basically GLMs and hence differ from the models used for the generalized MCP-Mod approach. However, the following candidate models for the Klingenberg approach are chosen to match those for the MCP-Mod approach as good as possible to make the results comparable:

Table 4.2: Candidate Models for the Simulations of the Klingenberg Approach

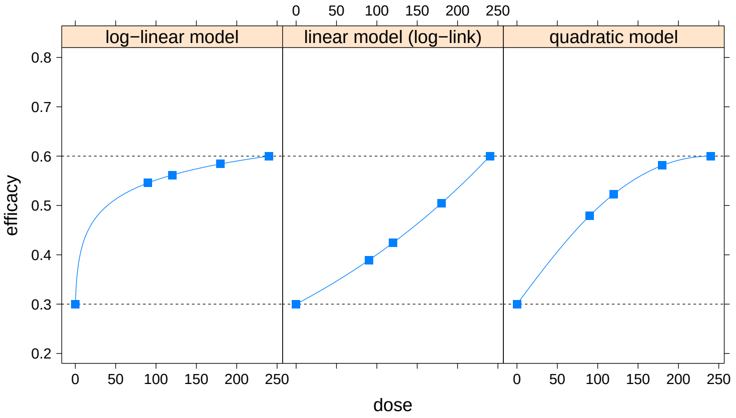
Model	**Link Function**	**Predictor**
Log-Linear Model	Logit	$\beta_0 + \beta_1 \log(d+1)$
Linear Model (Log-Link)	Log	$\beta_0 + \beta_1 d$
Quadratic Model	Logit	$\beta_0 + \beta_1 d + \beta_2 d^2$

The corresponding plots of those candidate models are shown in Figure 4.2(a)-(d) for all scenarios. Note that for the displayed plots, the models are only fitted to the placebo and the maximum effect. Information about the expected responses related to other dosages is not taken into account. Besides, no information will be extracted from this fitting process for the analysis of the data.
As before, the set of candidate models covers a broad range of dose-response shapes, i.e. convex shapes are represented by the linear model with log-link, and concave shapes by the log-linear and the quadratic model respectively. The filled rhombi mark the model-predicted response rates associated with the dose groups for which data shall be simulated.

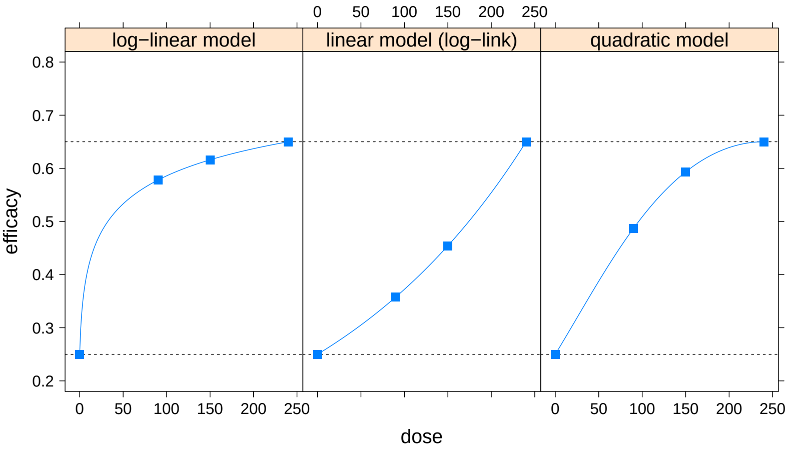


4.2 (a): Naives: High Scenario

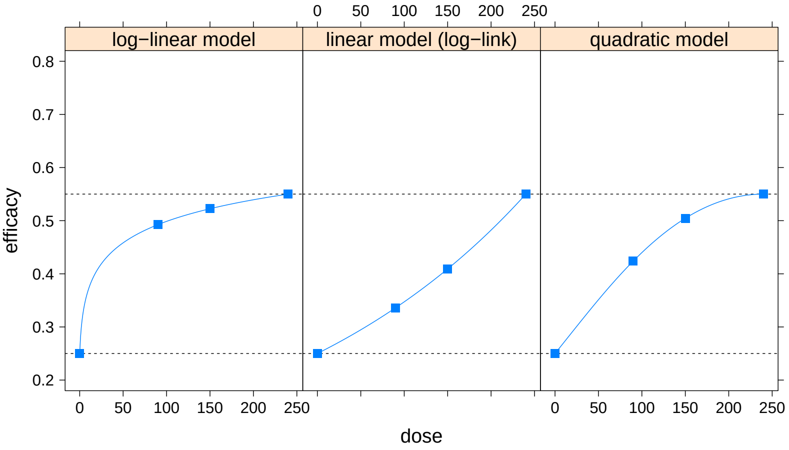
Figure 4.2: Candidate Models for the Simulations: Klingenberg Approach



4.2 (b): Naives: Low Scenario



4.2 (c): Failures: High Scenario



4.2 (d): Failures: Low Scenario

4.1 Comparison of the Methods for Binary Outcomes

The data sets that are generated for the comparison of the two methods shall contain 40 patients per dose group for the Naives and 50 patients per dose group for the Failures such that the total number of patients for the separate analyses of both subpopulations amounts to 200. This is true for the power and type-I error simulations as well as for the simulation to assess the precision of the MED.

Simulation Macro for the Estimation of Power and Type-I Error

The data is generated according to the response rates that are assumed under the alternative hypothesis (cf. Table 4.1) for the simulation of power, or according to the null hypothesis for the type-I error simulation (i.e. response rates in all dose groups equal to placebo, namely 0.3 for the Naives and 0.25 for the Failures). The simulated data is then used for the test of PoC as described in subsection 3.2.2 and subsection 3.2.3 respectively.

The simulation and analysis steps are repeated 10 000 times for the generalized MCP-Mod approach and 3 000 times with 3 000 permutations for the Klingenberg approach. The reduced number of simulation runs for the Klingenberg results is due to long run times caused by the permutation testing.

The test decisions of all simulation runs are stored together in a matrix. A positive test decision for model m in simulation run r is coded as $\text{success}_{rm} = 1$ if the adjusted p-value for model m is less than or equal to $\alpha = 0.05$; a negative test decision as $\text{success}_{rm} = 0$ if the condition wasn't met. The probability of rejecting a model-specific null hypothesis given the null hypothesis is true (type-I error) and under the alternative hypothesis (power) is estimated by calculating the frequencies of positive test decisions ($\text{success}_{rm} = 1$) per column, that is separately for each of the models. The overall power/type-I error is then obtained by condensing the model-specific values with the help of a summary function such as the minimum, maximum and mean. Beside these, another definition is derived from the PoC decision rule: the frequency with which at least one model-specific test decision was positive within a simulation run. This can be taken as an estimator for the power of rejecting the null hypothesis for at least one model or for the type-I error of erroneously rejecting one model-specific null hypothesis.

Simulation Macro for the Precision Assessment

For the precision assessment, each of the candidate models is assumed to be the true dose-response profile in one of the simulations. Consequently, data is generated according to the responses that are predicted by this specific model. In the analysis step, all the models in the candidate set are fitted to the simulated data, independent of which one was the true model chosen for the data generating process. Based on the fitted dose-response functions, different versions of the MED are estimated (cf. equation (2.14b) and subsection 3.2.3). The clinically significant effect that shall be induced by the MED is defined to be a change of $\delta = 0.3$ in the response rate. The set of estimates comprises all model-specific MEDs as well as the mean MED over all models, the MED of the model with the smallest AIC value and a weighted mean MED. For the latter, the weights are defined as in equation (3.3). Important to mention here is that the model-specific MED estimate was set to "not applicable" in case the corresponding model was not significantly different from a flat dose-response curve. Furthermore, this estimate was not included in the determination of the mean MEDs. The estimates

of all simulation runs are again collected in a matrix.
In a final step, the means, variances, Mean Square Errors (MSEs) and bias of the true MED for each of the seven variants are estimated. To get a better impression of the precision of all the different MED estimates, their distributions are illustrated by means of box plots.
The number of simulation runs for the precision assessment is set to 10 000 for both approaches. This is possible as only the models need to be fitted, the permutation test is not conducted here. For the decision whether the MED estimate will be excluded due to a negative test decision, the asymptotic p-value is used for the Klingenberg approach.

The macros for the generalized MCP-Mod approach are programmed in R, Version 3.0.3, the macros for the Klingenberg approach in Version 2.14.2 as some functions of the `stats` package that are used in the implemented functions were updated for newer R versions and hence, lead to errors.

4.1.1 Power and Type-I Error Results

The first characteristic for the comparison of the two approaches is the power. The simulation results for the generalized MCP-Mod approach are displayed in Table 4.3, for the Klingenberg approach they are listed in Table 4.4.

Table 4.3: Power Results for the Generalized MCP-Mod Approach

Scenario	**E_{max} Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid E_{max} Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
Naives High	0.9803	0.9963	0.9589	0.9944	0.9975	0.9824	0.9589	0.9963
Naives Low	0.9004	0.9312	0.8685	0.9087	0.9502	0.9022	0.8685	0.9312
Failures High	0.9908	0.9967	0.9778	0.9959	0.9982	0.9903	0.9778	0.9967
Failures Low	0.9227	0.9470	0.8859	0.9243	0.9618	0.9200	0.8859	0.9470

Table 4.4: Power Results for the Klingenberg Approach

Scenario	**Log-Linear Model**	**Linear Model (Log-Link)**	**Quadratic Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
Naives High	0.7947	0.9870	0.9837	0.9897	0.9218	0.7947	0.9870
Naives Low	0.6493	0.9203	0.9047	0.9290	0.8248	0.6493	0.9203
Failures High	0.8883	0.9930	0.9927	0.9940	0.9580	0.8883	0.9930
Failures Low	0.7167	0.9327	0.9103	0.9397	0.8532	0.7167	0.9327

It becomes obvious that for each scenario, all models except the log-linear model used in the Klingenberg approach show power estimates exceeding the 80% power margin, most of them are even greater than 90%. The power that matches the decision rule for the global PoC test for both approaches, namely the power to detect a statistically significant dose-response signal for at least one model, is the most liberal of all overall power terms. That means it results in the highest number of rejected global null hypotheses throughout all simulation runs and hence in the highest overall power.
Concerning the results for this power term, as well as for all other summarizing functions, the values for the generalized MCP-Mod approach are slightly higher than the ones for the Klingenberg approach.

When comparing the power values for the concave (quadratic) models only, the power to detect a non-flat dose-response is slightly higher for the Klingenberg approach than for the generalized MCP-Mod approach. Comparable power values are obtained for the convex models (sigmoid E_{max} model and exponential model in the generalized MCP-Mod approach versus linear model with log-link in the Klingenberg approach). In conclusion, none of the two approaches can be identified as clearly superior to the other in terms of power.

The results of the type-I error simulations are again displayed in two separate tables, the ones for the generalized MCP-Mod approach in Table 4.5, the ones for the Klingenberg approach in Table 4.6. For all scenarios and all definitions of the type-I error, the estimates stay below the prespecified significance level of $\alpha = 0.05$ or only show negligible exceedances. This is true for both the generalized MCP-Mod approach as well as for the Klingenberg approach. In general, the type-I error for the Klingenberg approach is slightly higher than that for the generalized MCP-Mod approach throughout all simulations, i.e. the significance level seems to be better exploited by the Klingenberg approach. This is desirable as a test that is too conservative (does not make use of the maximum error probability) is usually less powerful. However, a loss in power could not be observed in the investigated scenarios.

Table 4.5: Type-I Error Results for the Generalized MCP-Mod Approach

Scenario	**E_{max} Model**	**Expo-nential Model**	**Quadratic Model**	**Sigmoid E_{max} Model**	**Type-I Error of Rejecting at least one H_0**	**Mean Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**
Naives High	0.0282	0.0267	0.0269	0.0270	0.0449	0.0272	0.0267	0.0282
Naives Low	0.0254	0.0288	0.0246	0.0278	0.0440	0.0267	0.0246	0.0288
Failures High	0.0273	0.0277	0.0264	0.0289	0.0460	0.0276	0.0264	0.0289
Failures Low	0.0247	0.0274	0.0241	0.0291	0.0431	0.0263	0.0241	0.0291

Table 4.6: Type-I Error Results for the Klingenberg Approach

Scenario	**Log-Linear Model**	**Linear Model (Log-Link)**	**Quadratic Model**	**Type-I Error of Rejecting at least one H_0**	**Mean Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**
Naives High	0.0293	0.0317	0.0323	0.0500	0.0311	0.0293	0.0323
Naives Low	0.0293	0.0317	0.0323	0.0500	0.0311	0.0293	0.0323
Failures High	0.0270	0.0287	0.0353	0.0513	0.0303	0.0270	0.0353
Failures Low	0.0270	0.0287	0.0353	0.0513	0.0303	0.0270	0.0353

To be able to assess more distinctively which of the two approaches is preferable, another set of simulations has been conducted with dose-response signals that are not that highly significant. The dose groups as well as the sample size of 40 per group have been adopted from the first set of simulations for the Naives.

Again, several scenarios for the response rates have been considered (cf. Table 4.7): a scenario with a weak dose-response signal, a moderate scenario and a scenario with a strong dose-response signal. Note that even for the best of those scenarios, the corresponding dose-response signal is not as strong as in the previous settings.

Table 4.7: Assumed Response Rates for the Second Set of Simulations

Scenario	**0 mg**	**90 mg**	**120 mg**	**180 mg**	**240 mg**
Low Dose-Response Signal	0.35	0.35	0.4	0.5	0.5
Moderate Dose-Response Signal	0.35	0.35	0.35	0.5	0.55
High Dose-Response Signal	0.3	0.3	0.4	0.5	0.6

As intended, the estimated power for the second set of scenarios (cf. Tables 4.8 and 4.9) is in general (substantially) lower than in Tables 4.3 and 4.4.
However, the comparison between the two approaches reveals that also for the second set of scenarios, no clear preference for one of the approaches can be inferred. Both result in similar power values, independent of the underlying scenario. The modest inferiority of the Klingenberg approach could be explained by the fact that the corresponding candidate set only comprises three different candidate models instead of four models as for the generalized MCP-Mod approach.

Table 4.8: Power Results for the Generalized MCP-Mod Approach - Part 2

Scenario	**E_{max} Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid E_{max} Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
Low	0.3900	0.3246	0.3892	0.2457	0.4465	0.3374	0.2457	0.3900
Moderate	0.6018	0.6351	0.5695	0.5813	0.7038	0.5967	0.5695	0.6351
High	0.8328	0.7985	0.8190	0.7285	0.8805	0.7947	0.7285	0.8328

Table 4.9: Power Results for the Klingenberg Approach - Part 2

Scenario	**Log-Linear Model**	**Linear Model (Log-Link)**	**Quadratic Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
Low	0.3267	0.4133	0.3567	0.4473	0.3656	0.3267	0.4133
Moderate	0.3727	0.6300	0.5833	0.6520	0.5287	0.3727	0.6300
High	0.7393	0.8453	0.8007	0.8603	0.7951	0.7393	0.8453

Apart from the power, the same scenarios have also been investigated in terms of the type-I error. However, the results are similar to those already presented in Tables 4.5 and 4.6 for the first set of simulations and hence, are not shown here.

4.1.2 Precision Results

The second aspect for the comparison of the two approaches is the precision of the target dose estimate, that is of the MED. Separate simulations are conducted for every of the four scenarios and every assumed true dose-response model within these scenarios. As already described in a previous section, seven different estimates of the MED have been collected for the generalized MCP-Mod approach and six different estimates for the Klingenberg approach respectively. For each simulation, the precision of these estimates is illustrated by means of a plot. Each plot comprises a single box plot for one version of the MED estimate. The red horizontal line marks the MED according to the true underlying model. Important to state is that an estimate for the MED is only available for those simulation runs in which the prespecified effect over placebo of $\delta = 0.3$ is achieved at a dose between 0 and 240 mg (i.e. within the dose range of the study), else it is not applicable. This means that the box plots display the conditional distribution of the MED estimates, given the estimated target dose does not exceed 240 mg.

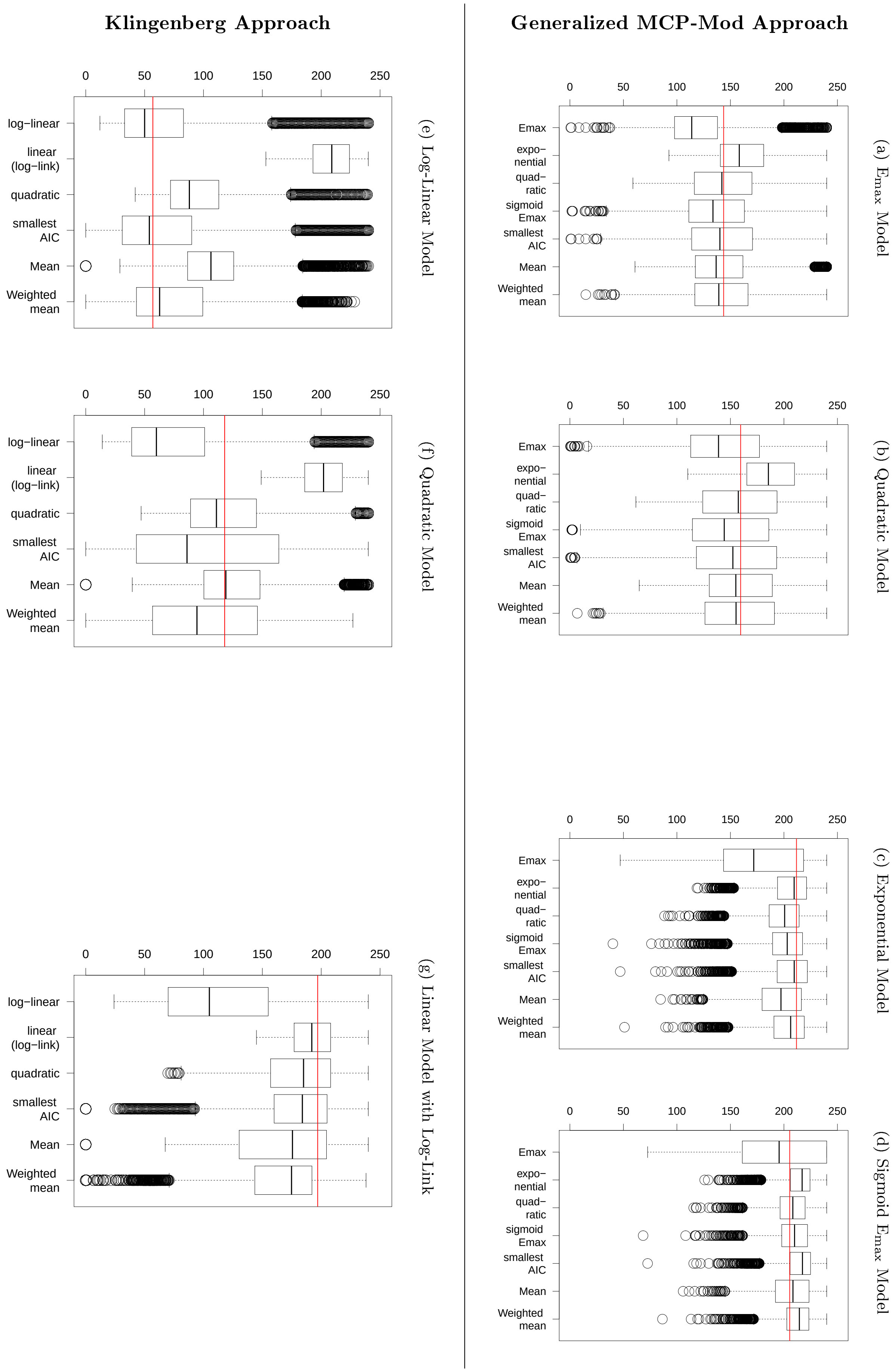
Furthermore, as stated in subsection 4.1, the MED estimate for a specific model was only included for the precision assessment if the corresponding model was shown to be significantly different from a non-flat dose-response curve. However, the box plots do not change remarkably when generated without this restriction (cf. plots in the Appendix, subsection B.2).

The plots of the results for the two scenarios of the Naives are presented in Tables 4.10 and 4.11. For both scenarios, all MED estimates for the generalized MCP-Mod approach show good performances. None of them can be identified as severely biased or highly variable. If the true underlying model is convex (exponential or sigmoid $\mathrm{E}_{\mathrm{max}}$ model), the variance of the estimates is generally smaller. On the contrary, if the true model is either the $\mathrm{E}_{\mathrm{max}}$ model or the quadratic model, the estimates are more variable. This can be explained by the fact that in the latter cases, the MED is located in a flatter part of the dose-response curve. This means that although the given dose varies, the response rate is almost constant. However, if the MED is located in a steep part, as it is the case for convex model shapes, a small change in the dose implies a large change in the response rate.

For the Klingenberg approach, the estimates for the high scenario also show acceptable precision characteristics. However, the performances of the estimates are more heterogeneous. The linear model with log-link results in a highly biased MED estimate if one of the concave models (log-linear model or quadratic model) is the true underlying dose-response profile. On the other hand, if the linear model with log-link is the true dose-response relationship, the MED estimate resulting from the log-linear model proposes a value that is much too small. In general, the estimates for the Klingenberg approach are less precise than the ones for the first approach and also tend to have a higher variance.

When considering the plots for the low scenario, it becomes obvious that it is indeed the conditional distribution of the estimates which is shown. The estimates are clearly biased towards smaller doses and vary clearly. From these plots one can deduce that the dose range of 0 to 240 mg is not appropriate for the low response scenario.

As the results for the Failures are very similar to the ones shown for the Naives, the plots are not included in the main part of this thesis. They can be found in the Appendix (subsection B.1).



(e) Log-Linear Model

(f) Quadratic Model

(g) Linear Model with Log-Link

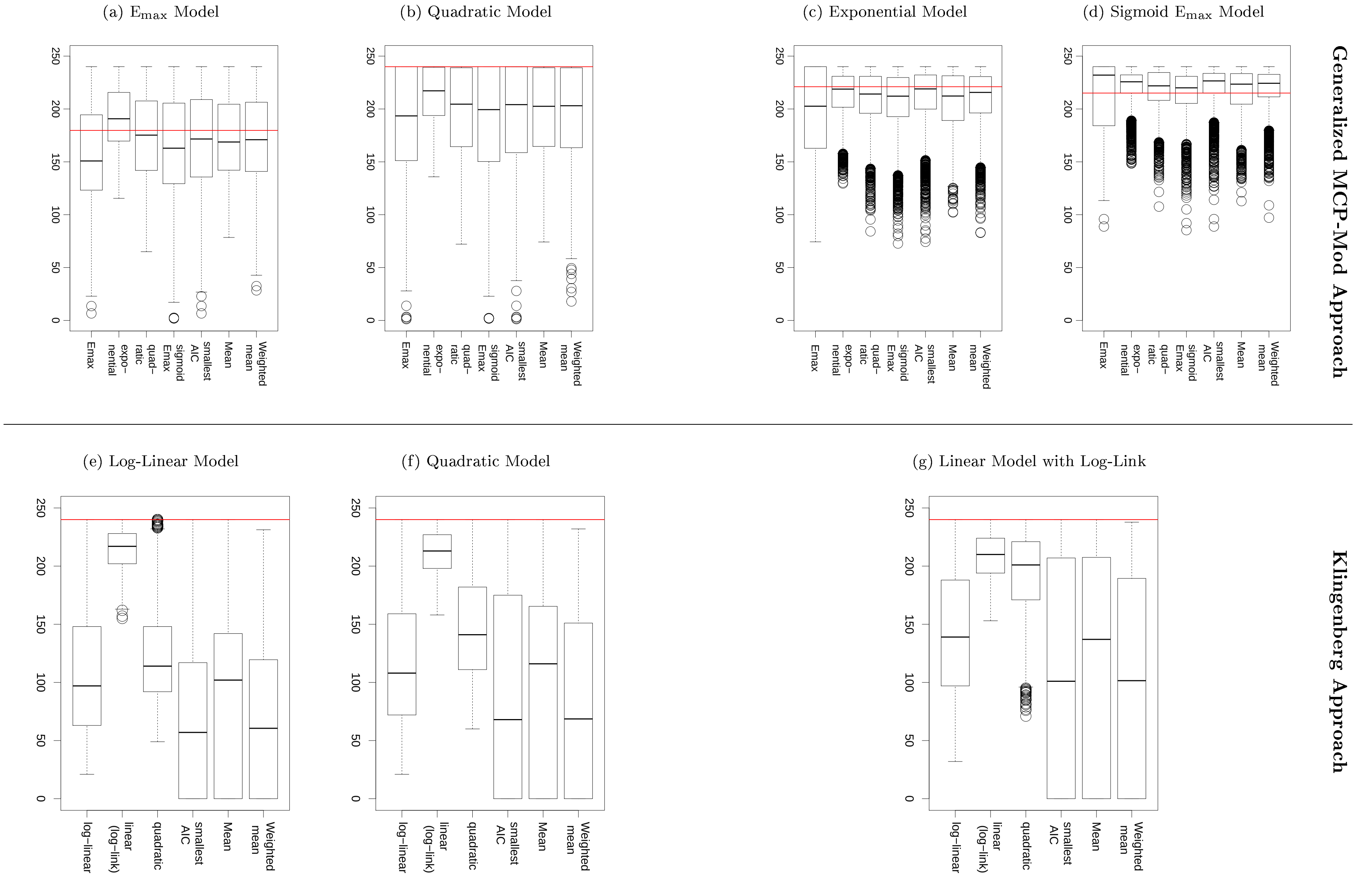
(a) E_{max} Model

(b) Quadratic Model

(c) Exponential Model

(d) Sigmoid E_{max} Model

Table 4.10: Precision of Target Dose Estimates - Naives, High Response Scenario



(a) $\mathrm{E}_{\max}$ Model

(b) Quadratic Model

(c) Exponential Model

(d) Sigmoid $\mathrm{E}_{\max}$ Model

(e) Log-Linear Model

(f) Quadratic Model

(g) Linear Model with Log-Link

Table 4.11: Precision of Target Dose Estimates - Naives, Low Response Scenario

4.2 Combined Analysis of Study Data

As stated in the introductory paragraph, the aim of the second section of this chapter is to investigate two approaches for the combination of the target dose results for the two subpopulations. Ideally, one would like to see results leading to a common optimal dose for the whole population which is independent from the specific population or line of treatment. The criteria for the comparison of the approaches are the same as in the last section: the power, the preservation of the type-I error and the precision of the common target dose estimate.

The two approaches to be considered in this section are the following:

- pooled analysis:
  A contrast test is conducted for H_0 : *flat dose-response curve for the overall population*. The contrasts stem from the candidate models with initial parameters calculated on the basis of the assumed response rates for both subpopulations. If the null hypothesis can be rejected, the best model is fitted to the pooled data and the MED is estimated from this overall dose-response model.

- combination of separate analyses:
  The subpopulations are separately tested for a non-flat dose-response. The global test is then characterized by the null hypothesis of H_0 : *flat dose-response curve for the Naives* ***and*** *the Failures* and the corresponding alternative hypothesis of H_A : *non-flat dose-response curve for the Naives* ***and/or*** *the Failures*. The global null hypothesis can be rejected if one of the separate p-values is smaller than $\frac{\alpha}{2}$, i.e. Bonferroni correction is used to account for multiplicity (cf. subsection 2.2.3). An alternative multiplicity adjustment is the Benjamini-Hochberg procedure (cf. subsection 2.2.3) which means that the smaller one of the two separate p-values is compared against $\frac{\alpha}{2}$ whilst the maximum p-value is compared against α. Again, the global null hypothesis can be rejected if one of the p-values is smaller than its comparative value.
  If the global null hypothesis can be rejected, a separate dose-response model is fitted for each subpopulation and the corresponding MEDs are estimated. The common optimal dose is then chosen as the MED that produces response rates that are higher than the placebo response by at least δ for both subpopulations. In case of monotonously increasing dose-response functions, this is equal to the maximum of both MED estimates. The combination of the MED results can be additionally restricted by demanding that the MED estimates for the Naives and the Failures should not be "too different", for example, they should not differ by more than 30 mg (referred to as the restricted version of the combination approach in the results section).

Both approaches will be implemented in macros similar to the ones that have been described in section 4.1. For the power and type-I error simulations, all combinations of scenarios for the Naives and the Failures will be considered. For the simulations to investigate the precision of the MED estimate, a selection of four different scenarios will be considered. The scenarios are chosen such that the extent of similarity between the true MEDs of the Naives and the Failures varies. In any case, it is assumed that the true underlying dose-response model is the same for the Naives and the Failures. Otherwise, it would be doubtful if a combination of the results makes sense at all. The results of such a scenario with different dose-response models is subject to further research.

As the generalized MCP-Mod method showed better results in terms of the precision of the MED estimate than the Klingenberg approach, the analysis of the data is done according to this approach only.

4.2.1 Power and Type-I Error Results

The results of the power and type-I error simulations for the pooled analysis approach are presented in Tables 4.12 and 4.13. Study data has been simulated according to the response rates given in Table 4.1 with 40 patients per dose group for the Naives and 50 patients per dose group for the Failures.
The power estimates are very similar across the scenarios and all models provide impressively good results. All of the power values are above 99%.

Table 4.12: Power Results for the Pooled Analysis

Scenario Naives	**Scenario Failures**	**$\mathbf{E}_{\max}$ Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid $\mathbf{E}_{\max}$ Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
High	High	1	1	0.9999	1	1	1	0.9999	1
High	Low	0.9999	1	0.9986	0.9999	1	0.9996	0.9986	1
Low	High	1	1	0.999	1	1	1	0.9990	1
Low	Low	0.9977	0.9990	0.9945	0.9976	0.9996	0.9972	0.9945	0.9990

When looking at the estimates for the type-I error of rejecting at least one individual null hypothesis (seventh column), the estimate for the second scenario (high response scenario for the Naives and low response scenario for the Failures) is slightly inflated.

Table 4.13: Type-I Error Results for the Pooled Analysis

Scenario Naives	**Scenario Failures**	**$\mathbf{E}_{\max}$ Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid $\mathbf{E}_{\max}$ Model**	**Type-I Error of Rejecting at least one H_0**	**Mean Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**
High	High	0.0311	0.0233	0.0324	0.0227	0.0452	0.0274	0.0227	0.0324
High	Low	0.0350	0.0285	0.0358	0.0268	0.0506	0.0315	0.0268	0.0358
Low	High	0.0301	0.0245	0.0319	0.0247	0.0465	0.0278	0.0245	0.0319
Low	Low	0.0319	0.0277	0.0321	0.0246	0.0471	0.0291	0.0246	0.0321

This becomes even more obvious if the difference in the sample sizes per dose group between the subpopulations is higher, for example 30 patients for the Naives and 60 for the Failures. The type-I error results for this scenario are shown in Table 4.14.

Table 4.14: Type-I Error Results for the Pooled Analysis -
Sample Sizes of 30 (Naives) and 60 (Failures) per Group

Scenario Naives	**Scenario Failures**	**$\mathbf{E}_{\max}$ Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid $\mathbf{E}_{\max}$ Model**	**Type-I Error of Rejecting at least one H_0**	**Mean Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**
High	High	0.0317	0.0253	0.0329	0.0240	0.0469	0.0285	0.0240	0.0329
High	Low	0.0343	0.0294	0.0349	0.0272	0.0501	0.0315	0.0272	0.0349
Low	High	0.0335	0.0269	0.0365	0.0268	0.0512	0.0309	0.0268	0.0365
Low	Low	0.0313	0.0289	0.0332	0.0277	0.0502	0.0303	0.0277	0.0332

The inflation of the type-I error for the pooled analysis approach results from the fact that the placebo responses for the Naives and the Failures are not identical and at the same time, the dose groups that were chosen for the single trials do not coincide. Hence, as presented in Table 4.15, the mean overall response rates are no longer constant over all dose groups and thus, fitting a model to the pooled data does not necessarily result in a constant function. Therefore, the null hypothesis of a flat overall dose-response curve tends to be rejected too often.

Table 4.15: Mean Response Rates under H_0 for the Pooled Analysis

Population	**0 mg**	**90 mg**	**120 mg**	**150 mg**	**180 mg**	**240 mg**
Naives	0.3	0.3	0.3		0.3	0.3
Failures	0.25	0.25		0.25		0.25
Overall	0.2722	0.2722	0.3	0.25	0.3	0.2722

Apart from this visual explanation, the inappropriateness of a pooled analysis approach in this setting also becomes obvious when considering the distribution of the test statistic under H_0. For the single trials, the mean responses under the null hypothesis are constant over the dose groups. Hence, as the elements of the contrast vector have to sum up to 0, the nominator of the test statistic (product of contrast vector and mean vector of responses) is expected to be zero. Consequently, the contrast test statistic follows a central t-distribution.

However, with varying mean responses across the dose groups as it is the case for the present pooled analysis approach, the centrality of the t-distribution of the contrast test statistic is no longer valid. As a results of that, choosing the critical value as a quantile of that distribution does not ensure the preservation of the type-I error at the given significance level α.

For further support of this explanation, the same simulations have been repeated but response data for the Failures has been generated for the same dose groups that were used for the Naives, namely 0 mg, 90 mg, 120 mg, 180 mg and 240 mg, together with a balanced number of 40 patients per dose group over both subpopulations. The results of these simulations are presented in Tables 4.16 and 4.17.

The power estimates are very similar to the ones from the first set of simulations such that the power clearly exceeds the 90% margin for all cases. Again, the values do not vary much across the four scenarios.

Table 4.16: Power Results for the Pooled Analysis - Equivalent Dose Groups and Equal Sample Sizes

Scenario Naives	**Scenario Failures**	**$\mathbf{E_{max}}$ Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid $\mathbf{E_{max}}$ Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
High	High	0.9998	1	0.9995	1	1	0.9998	0.9995	1
High	Low	0.9995	0.9998	0.9978	0.9997	0.9999	0.9992	0.9978	0.9998
Low	High	0.9991	1	0.9983	0.9998	1	0.9993	0.9983	1
Low	Low	0.9964	0.9987	0.9911	0.9970	0.9990	0.9958	0.9911	0.9987

Again endorsing the above stated explanation, the estimates for the type-I error of rejecting at least one model-specific H_0 are now preserving the significance level of $\alpha = 5\%$ for all scenarios.

Equivalent results can be observed if choosing the same dose groups for both subpopulations but with different sample sizes per dose group (cf. Appendix, section C). In this case, the overall mean responses are weighted means of the responses in the single trials, but still, they are constant over all dosages.

Table 4.17: Type-I Error Results for the Pooled Analysis - Equivalent Dose Groups and Equal Sample Sizes

Scenario Naives	**Scenario Failures**	**E$_{\text{max}}$ Model**	**Expo-nential Model**	**Quadratic Model**	**Sigmoid E$_{\text{max}}$ Model**	**Type-I Error of Reject-ing at least one H_0**	**Mean Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**
High	High	0.0295	0.0296	0.0287	0.0283	0.0463	0.0290	0.0283	0.0296
High	Low	0.0283	0.0281	0.0274	0.0293	0.0476	0.0283	0.0274	0.0293
Low	High	0.0279	0.0283	0.0268	0.0279	0.0455	0.0277	0.0268	0.0283
Low	Low	0.0287	0.0285	0.0302	0.0276	0.0482	0.0288	0.0276	0.0302

As a consequence, the approach of conducting an (unmodified) overall contrast test for the pooled data set of Naives and Failures leads to an inflation of the type-I error if the setting for the single studies do not coincide, i.e. the studies use different dose groups and/or sample sizes. Therefore, this approach cannot be recommended in a general setting.

However, if the global p-value is derived from the appropriate noncentral t-distribution with noncentrality parameter

$$\boldsymbol{\tau} = \frac{\boldsymbol{C}^\top \boldsymbol{\mu}^0}{(\boldsymbol{C}^\top \boldsymbol{S}\boldsymbol{C})^{\frac{1}{2}}}$$

where $\boldsymbol{\mu}^0$ is the overall mean vector of responses under H_0, the inflation of the type-I error can be prevented. The corresponding simulation results are presented in Tables 4.18 and 4.19.
The power estimates are only slightly smaller than those that result from the use of an unmodified p-value and the type-I error estimates stay below the significance level of $\alpha = 5\%$ for all scenarios.

Table 4.18: Power Results for the Pooled Analysis - p-Value from Noncentral t-Distribution

Scenario Naives	**Scenario Failures**	**E$_{\text{max}}$ Model**	**Expo-nential Model**	**Quadratic Model**	**Sigmoid E$_{\text{max}}$ Model**	**Power to Reject at least one H_0**	**Mean Power**	**Minimum Power**	**Maximum Power**
High	High	1	1	0.9999	1	1	1	0.9999	1
High	Low	0.9997	1	0.9984	0.9998	1	0.9995	0.9984	1
Low	High	0.9993	0.9998	0.9984	0.9995	0.9999	0.9993	0.9984	0.9998
Low	Low	0.9980	0.9991	0.9944	0.9981	0.9995	0.9974	0.9944	0.9991

Table 4.19: Type-I Error Results for the Pooled Analysis - p-Value from Noncentral t-Distribution

Scenario Naives	**Scenario Failures**	**E$_{\text{max}}$ model**	**Expo-nential model**	**Quadratic model**	**Sigmoid E$_{\text{max}}$ model**	**Type-I error of rejecting at least one H_0**	**Mean Type-I error**	**Minimum Type-I error**	**Maximum Type-I error**
high	high	0.0279	0.0275	0.0281	0.0267	0.0450	0.0276	0.0267	0.0281
high	low	0.0293	0.0271	0.0305	0.0273	0.0475	0.0286	0.0271	0.0305
low	high	0.0299	0.0269	0.0313	0.0263	0.0464	0.0286	0.0263	0.0313
low	low	0.0302	0.0269	0.0311	0.0271	0.0474	0.0288	0.0269	0.0311

In practice, the computation of the true noncentrality parameter under H_0 may not be perfect as an estimate of the overall mean vector of responses under H_0 is needed. This may involve the risk of not preserving the type-I error due to a misspecification.

The second approach investigated in this thesis is the combination of results stemming from separate analyses of the single trials. The power and type-I error estimates presented first (in Tables 4.20 and 4.21) are a result of the combination of p-values using the Bonferroni correction for multiple testing.
As already observable for the pooled approach, the estimated power is extremely high for all of the considered scenarios. All values are close to 100% power.

Table 4.20: Power Results for the Combination of Separate Analyses (Bonferroni Correction)

Scenario Naives	Scenario Failures	$\mathrm{E}_{\max}$ Model	Exponential Model	Quadratic Model	Sigmoid $\mathrm{E}_{\max}$ Model	Power to Reject at least one H_0	Mean Power	Minimum Power	Maximum Power
High	High	0.9987	0.9999	0.9961	0.9999	0.9999	0.9987	0.9961	0.9999
High	Low	0.9954	0.9992	0.9845	0.9987	0.9997	0.9945	0.9845	0.9992
Low	High	0.9969	0.9997	0.9912	0.9991	0.9999	0.9967	0.9912	0.9997
Low	Low	0.9751	0.9868	0.9584	0.9789	0.9915	0.9748	0.9584	0.9868

Also the results of the type-I error simulations show very good results. All values stay below the critical margin of $\alpha = 5\%$ and the estimated type-I error of rejecting at least one of the null hypotheses is close to 5%. Hence, the significance level is adequately exploited.

Table 4.21: Type-I Error Results for the Combination of Separate Analyses (Bonferroni Correction)

Scenario Naives	Scenario Failures	$\mathrm{E}_{\max}$ Model	Exponential Model	Quadratic Model	Sigmoid $\mathrm{E}_{\max}$ Model	Type-I Error of Rejecting at least one H_0	Mean Type-I Error	Minimum Type-I Error	Maximum Type-I Error
High	High	0.0227	0.0238	0.0225	0.0253	0.0408	0.0236	0.0225	0.0253
High	Low	0.0243	0.0258	0.0235	0.0272	0.0431	0.0252	0.0235	0.0272
Low	High	0.0249	0.0250	0.0253	0.0257	0.0425	0.0252	0.0249	0.0257
Low	Low	0.0248	0.0251	0.0240	0.0257	0.0427	0.0249	0.0240	0.0257

Alternatively to the Bonferroni correction, the same simulations are conducted using the Benjamini-Hochberg correction for the final test decision. The Benjamini-Hochberg method is slightly less conservative than the Bonferroni correction. The power and type-I error results are presented in Tables 4.22 and 4.23.

Table 4.22: Power Results for the Combination of Separate Analyses (Benjamini-Hochberg Correction)

Scenario Naives	Scenario Failures	$\mathrm{E}_{\max}$ Model	Exponential Model	Quadratic Model	Sigmoid $\mathrm{E}_{\max}$ Model	Power to Reject at least one H_0	Mean Power	Minimum Power	Maximum Power
High	High	0.9988	0.9999	0.9972	0.9999	0.9999	0.9990	0.9972	0.9999
High	Low	0.9969	0.9994	0.9876	0.9990	0.9998	0.9957	0.9876	0.9994
Low	High	0.9976	0.9999	0.9926	0.9991	0.9999	0.9973	0.9926	0.9999
Low	Low	0.9792	0.9884	0.9640	0.9816	0.9926	0.9783	0.9640	0.9884

As for the previous simulations, the power estimates are almost at 100% for all scenarios.
The estimates for the type-I error also look very good. They do not exceed the significance level of 5% for any of the scenarios but at the same time, the significance level is well exploited.
When comparing the results for the Benjamini-Hochberg correction against those for the Bonferroni correction, the first is superior. The power estimates show higher values and the significance level is better exploited. However, differences are small such that both of the approaches can be recommended.

Table 4.23: Type-I Error Results for the Combination of Separate Analyses (Benjamini-Hochberg Correction)

Scenario Naives	Scenario Failures	E_{max} Model	Exponential Model	Quadratic Model	Sigmoid E_{max} Model	Type-I Error of Rejecting at least one H_0	Mean Type-I Error	Minimum Type-I Error	Maximum Type-I Error
High	High	0.0230	0.0239	0.0227	0.0253	0.0410	0.0237	0.0227	0.0253
High	Low	0.0246	0.0259	0.0240	0.0273	0.0435	0.0255	0.0240	0.0273
Low	High	0.0252	0.0252	0.0255	0.0261	0.0432	0.0255	0.0252	0.0261
Low	Low	0.0252	0.0253	0.0242	0.0257	0.0430	0.0251	0.0242	0.0257

Altogether, the approach of the subsequent combination of the results for the single subpopulations is preferable to a pooled analysis. It ensures the preservation of the type-I error for arbitrary settings of the single trials and induces good power results at the same time.

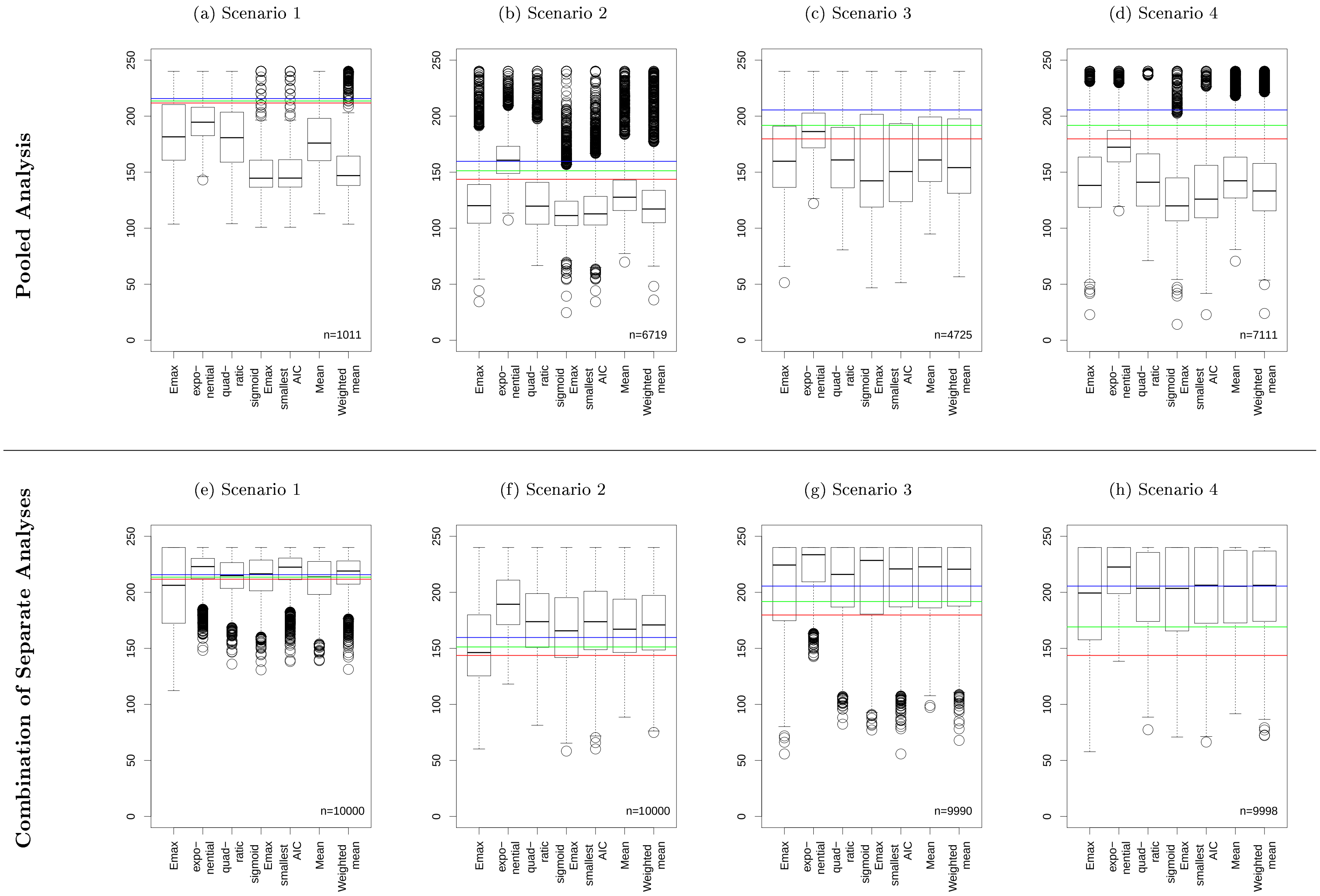
4.2.2 Precision Results

As mentioned at the beginning of this chapter, only a choice of scenarios is considered for the precision assessment. The intention of the selection is to cover different extents of similarity between the true MEDs of the Naives and the Failures, with differences in the true MEDs ranging from 3.99 mg up to 61.88 mg. In any case, the true underlying dose-response profile according to which the data is generated is assumed to be of the same family for both subpopulations. The four chosen scenarios together with the corresponding differences of the true MEDs are listed in Table 4.24.

Table 4.24: Scenarios for the Precision Simulations for the Combined Analysis of Study Data

Scenario	Naives	Failures	True Model	Difference of True MEDs
Scenario 1	High Scenario	High Scenario	Exponential	3.99 mg
Scenario 2	High Scenario	High Scenario	E_{max}	16.02 mg
Scenario 3	Low Scenario	Low Scenario	E_{max}	25.81 mg
Scenario 4	High Scenario	Low Scenario	E_{max}	61.88 mg

For each of the scenarios, simulations have been performed using three different types of analysis: the pooled analysis, the combination of the separate study results and the restricted combination of the separate study results for which the MED estimates must not differ by more than 30 mg.
Just like for the comparison of the methods for binary data, the precision of the MED estimates is again illustrated by means of box plots (Table 4.25). The red horizontal lines in the plots mark the true MEDs for the Naives, the blue horizontal lines represent the true MEDs for the Failures.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 1

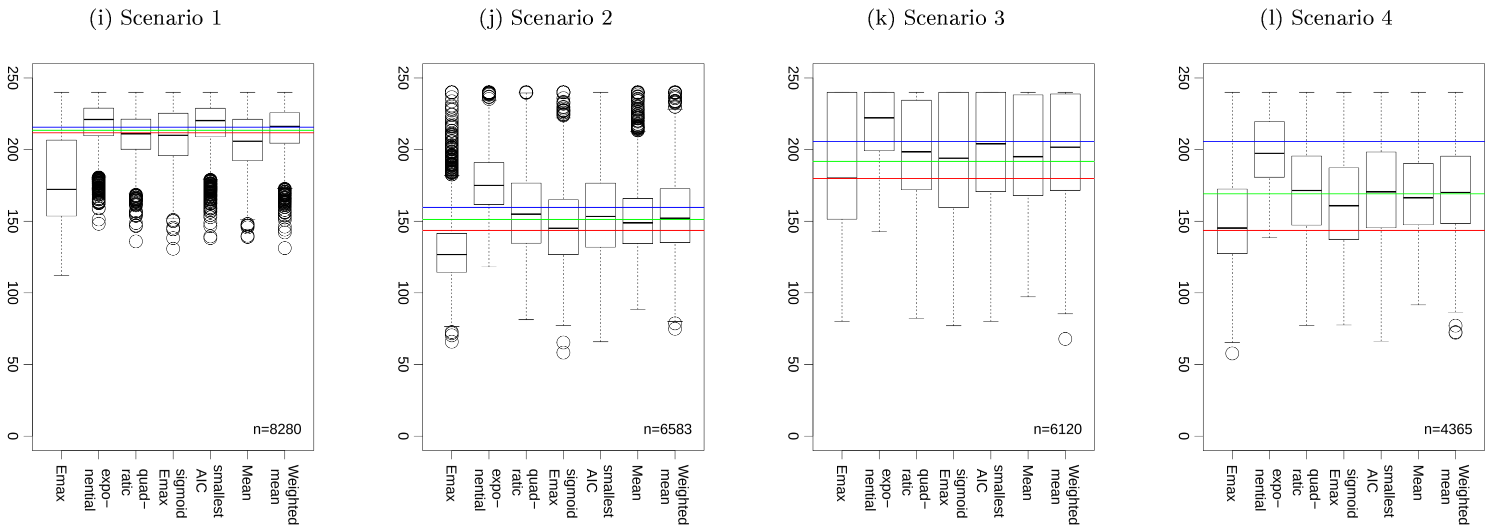
(f) Scenario 2

(g) Scenario 3

(h) Scenario 4

Table 4.25: Precision of Target Dose Estimates - Combined Analysis of Study Data

Combination of Separate Analyses - Restricted



Precision of Target Dose Estimates - Combined Analysis of Study Data (Cont.)

Additionally, a true overall MED has been determined on the basis of the candidate model which is of the same family as the data generating model. The initial parameter estimates of this candidate model have been calculated taking into account the prior knowledge for both subpopulations. The true overall MED is depicted by the green horizontal lines.
Furthermore, the number of MED estimates that were included for the plot is added as a note in the low right hand corner of the plot. The reason for exclusion can either be a non-significant test result in the first analysis step or, in the restricted combination approach, a difference in the MED estimates that is too extreme and exceeds the predefined margin of 30 mg.

The plots show that the estimates of the pooled analysis stay below the true MEDs for the Naives which is the lowest of the three types of true MEDs. This is true for almost all types of MED estimates in all scenarios. That means using the pooled analysis approach, the MED is systematically underestimated.
On the contrary, the estimates resulting from the combination of the subpopulation-specific estimates mostly exceed the true MEDs for the Failures which is the highest of the true MEDs. However, this is a consequence of the applied combination rule for the estimates as the overall MED is chosen to be the maximum of the single MEDs for the Naives and the Failures.
The third approach, the restricted version of the combination of results, shows the best performance of all three investigated approaches. The medians of the estimates coincide quite well with the true overall MEDs for all different versions of the MED and for all scenarios.
Concerning the variance of the estimates, no clear difference can be detected between the three approaches, i.e. none of them can be identified as preferable over the others in terms of variability.

In conclusion, it is apparent that the combination of study results after the analyses of the separate trials have been finalized is a better approach than the pooled analysis, in terms of testing as well as for the estimation of the target dose. Among the two combination approaches, there is no difference in power and the preservation of the type-I error as the testing procedure is not affected by the restriction for the combination of the MED estimates. However, the estimates for the restricted combination approach fit better to the true overall MEDs than the ones for the unrestricted combination approach. Furthermore, conditioning the combination of the single estimates on a certain extent of similarity is well reasonable, either to the margin of 30 mg or another arbitrary value. Because if the single estimates are too different, a common dose may cause that patients from one subpopulation are overdosed whereas others are also not adequately treated and could be medicated more appropriately. So in this case, the conclusion is to recommend different dosages depending on the success of a potential first treatment.

Chapter 5

Discussion and Outlook

This thesis introduced the MCP-Mod approach as a hybridization of the methods that are commonly used for the matter of dose-finding, namely multiple comparisons and the parametric modelling of the dose-response curve. In comparison to the original approach which is restricted to a very basic case of normal data, an enhancement has been presented which enables the application of the approach also to non-normal or heteroscedastic data as well as survival data, repeated measurements and others.
Both methods, the original approach as well as its enhancement, consist of two steps: a contrast test is conducted to investigate if any of the models from a predefined candidate set is significantly different from a flat dose-response curve. Models to be used for this approach are parametric models, e.g. linear models, quadratic models or the so-called $\mathrm{E}_{\mathrm{max}}$ model. If at least one of those models achieves a positive test result, the one that describes the study data best is fitted and the target dose is estimated from the resulting parametric function.
Furthermore, the related Klingenberg approach has been examined. Just like the two previous approaches, it also unifies the PoC with a parametric modelling of the dose-response curve, but is primarily developed for binary data. Contrary to the MCP-Mod approaches, the PoC is tried to be established using penalized deviance difference statistics. The corresponding test decisions are made on the basis of the permutational distribution of the test statistic.

After a theoretical introduction to these methods, the first aim of the practical part was to compare the naive application of the original method to binary data with the other two methods. Criteria for the comparison were the achieved power, the preservation of the type-I error and the precision of the target dose estimate (here the MED). All these qualities have been estimated by means of simulations. The naive application of the original MCP-Mod approach to binary data led to a loss in power and a potential inflation of the type-I error, hence it cannot be recommended for other situations than normal data resulting from a simple study design.
Apart from that, a clear preference for one method could not be established, neither for the generalized version of the MCP-Mod approach nor for the Klingenberg approach. The power and type-I error results were very similar for both approaches. Only in terms of precision, the generalized MCP-Mod approach is slightly preferable. The corresponding target dose estimates come out to be less variant and tend to be more precise.
On the other hand, the Klingenberg approach is clearly in favour in terms of the interpretability of the results as the dose-response models are directly defined on response level. Another advantage is the use of GLMs which are common knowledge also for users without an in-depth understanding of statistics. In contrast to this, the dose-response models for the generalized MCP-Mod approach are defined on a parameter level. Hence, their interpretation is not straightforward, the results are not self-explanatory, especially for non-statisticians.

What is missing for both approaches is a recommendation for the appropriate designing of such a study. Neither the paper by Klingenberg, nor the one by Pinheiro et al. proposes methods for the identification of the optimal dose groups to include in the study or for the sample size assessment. Using the R function `sampSize` as implemented for the original approach leads to power values smaller than the target power aimed for.
Consequently, this would be a topic for the further development of these approaches.

Beside the comparison of the approaches for binary data, the additional objective of the thesis was the combination of target dose results of separate trials with the aim of obtaining a common dosage proposal if adequate. Therefor, two different procedures have been investigated with respect to the same aspects as considered for the comparison of approaches in the first part, namely power, preservation of the type-I error and the precision of the target dose estimate.
The first approach was a pooled analysis of the combined data set of both single trials according to the generalized MCP-Mod approach. Simulations showed that for the considered scenarios, i.e. in case the dose groups from the single trials do not coincide and the placebo responses are assumed to be different for the study populations, the pooled analysis approach leads to an inflation of the type-I error. This results from deriving the global p-value from a central t-distribution whereas the true distribution of the maximum statistic under the global null hypothesis in this setting is a noncentral t-distribution. One possible correction is the derivation of the p-value from the appropriate noncentral t-distribution. Simulations confirmed the preservation of the type-I error in this case. One could also try to modify the computation of the optimal contrast vector such that the matrix multiplication of this vector with the vector of mean overall responses under H_0 equals zero. However, the success of this method is subject to further investigation.
Alternatively to analyzing the pooled data, one could pool the results from the separate trials after having finalized the analyses of both, herein referred to as the combination of separate analyses approach. Therefor, the p-values resulting from the application of the MCP-Mod methods for non-normal data are combined using either the Bonferroni correction for multiplicity or the Benjamini-Hochberg procedure. The results for both correction methods are very similar. The achieved power for the considered scenarios is close to 100% and the type-I error has been preserved for all simulations. Also the precision assessment was contenting. Even better results in terms of the precision of the target dose estimate could be achieved by adding a restriction to the combination of population-specific MEDs. This means that a common dosage is only recommended if the population-specific MEDs do not differ relevantly.
An additional aspect which hasn't been investigated is the precision of the combination approaches in case the response data of the separate trials stem from different model families. For the precision assessment presented in this thesis, the responses for the two trials have been generated on the basis of one family of dose-response models. Hence, the behaviour of the target dose estimate should be examined also for this more inconvenient setting.

Furthermore, one question that arises for these approaches is how to identify the best matching model from the set of candidate models most appropriately. Is it preferable to select the model based on information criteria or to use model averaging techniques? This issue is addressed in a more general context by Bornkamp (2015) and could be investigated more specifically for the background of these unifying dose-finding approaches.

Other topics that go beyond this thesis and are worth future research are combinations of these unified approaches with adaptive designs and/or Bayesian methodology.

For example one could use the first part of an adaptive trial to learn about the rough shape of the dose-response profile and use this information to adapt certain design features in the course of the interim analyses, for example the sample size or the allocation of patients. Besides, one could add/change the dose groups of the ongoing trial to aggregate the data around the suspected range for the target dose. A study putting this into practice can be found in Selmaj et al. (2013).

Another approach is to define a priori distributions for the model parameters as well as prior model probabilities and update the design features according to the posterior means of the model parameters and the posterior model probabilities computed at the interim analyses. The methodology of such an adaptive design involving Bayesian statistics is presented in Bornkamp et al. (2011).

Bibliography

Abelson, R. P. and Tukey, J. W. (1963). “Efficient Utilization of Non-Numerical Information in Quantitative Analysis General Theory and the Case of Simple Order”. *The Annals of Mathematical Statistics*, **34** (4), pp. 1347–1369.

International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (1994). “Topic E4: Dose-response information to support drug registration. (ICH Harmonized Tripartite Guideline)”. URL: `http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E4/Step4/E4_Guideline.pdf`.

Ahrens, H. and Läuter, J. (1981). Mehrdimensionale Varianzanalyse. Akademie-Verlag.

AIDSinfo (2014). AIDSinfo Drug Database - Zidovudine. Visited on 04.11.2014. URL: `http://aidsinfo.nih.gov/drugs/4/zidovudine/0/patient`.

Benjamini, Y. and Hochberg, Y. (1995). “Controlling the False Discovery Rate: A practical and powerful approach to Multiple Testing”. *Journal of the Royal Statistical Society*, **57** (1), pp. 289–300.

Benjamini, Y. and Yekutieli, D. (2001). “The control of the false discovery rate in multiple testing under dependency”. *The Annals of Statistics*, **29** (4), pp. 1165–1188.

Boos, D. D. (1992). “On Generalized Score Tests”. *The American Statistician*, **46** (4), pp. 327–333.

Bornkamp, B. (2015). “Viewpoint: model selection uncertainty, pre-specification, and model averaging”. *Pharmaceutical Statistics*, pp. 79–81.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2009). “MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies”. *Journal of Statistical Software*, **29** (7), pp. 1–23.

Bornkamp, B., Bretz, F., Dette, H., and Pinheiro, J. (2011). “Response-adaptive dose-finding under model uncertainty”. *The Annals of Applied Statistics*, **5** (2B), pp. 1611–1631.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2012). Package ’DoseFinding’. URL: `http://cran.r-project.org/web/packages/DoseFinding/DoseFinding.pdf`.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2014). DoseFinding: Planning and Analyzing Dose Finding experiments. R package version 0.9-11. URL: `http://CRAN.R-project.org/package=DoseFinding`.

Branson, M., Pinheiro, J., and Bretz, F. (2003). Searching for an adequate dose: Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. Tech. rep. No. 2003-08-20. Available at `http://www.bioinf.uni-hannover.de/~bretz/paper/TR_MCPMod.pdf`. Novartis Pharmaceuticals.

Braun, H. I., ed. (1994). The collected works of John W. Tukey. Vol. 8. Multiple comparisons: 1948–1983. New York: Chapman and Hall, Ltd.

Bretz, F. (1999). “Powerful modifications of Williams’ test on trend”. Available at `http://www.biostat.uni-hannover.de/fileadmin/institut/pdf/thesis_bretz.pdf`. PhD thesis. University of Hannover.

Bretz, F., Pinheiro, J. C., and Branson, M. (2005). “Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies”. *Biometrics*, **61** (3), pp. 738–748.

Bretz, F., Hsu, J., Pinheiro, J., and Liu, Y. (2008). “Dose Finding – A Challenge in Statistics”. *Biometrical Journal*, **50** (4), pp. 480–504.

Chevret, S., ed. (2006). Statistical Methods for Dose-Finding Experiments. John Wiley & Sons, Ltd.

Cross, J., Lee, H., Westelinck, A., Nelson, J., Grudzinskas, C., and Peck, C. (2002). "Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980–1999". *Pharmacoepidemiology and Drug Safety*, **11** (6), pp. 439–446.

Dette, H., Bretz, F., Pepelyshev, A., and Pinheiro, J. (2008). "Optimal Designs for Dose-Finding Studies". *Journal of the American Statistical Association*, **103** (483), pp. 1225–1237.

Dunnett, C. W. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control". *Journal of the American Statistical Association*, **50** (272), pp. 1096–1121.

Finner, H. and Strassburger, K. (2002). "The partitioning principle: a powerful tool in multiple decision theory". *The Annals of Statistics*, **30** (4), pp. 1194–1213.

Gallant, A. (1987). Nonlinear Statistical Models. Wiley Series in Probability and Statistics. Wiley.

Genz, A. and Bretz, F. (2000). "Methods for the Computation of Multivariate t-Probabilities". *Computing Sciences and Statistics*, **25**, pp. 400–405.

Hochberg, Y. and Tamhane, A. C. (1987). Multiple Comparison Procedures. John Wiley & Sons, Inc.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian Model Averaging: A Tutorial". *STATISTICAL SCIENCE*, **14** (4), pp. 382–417.

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure". *Scandinavian Journal of Statistics*, **6** (2), pp. 65–70.

Hsu, J. (1996). Multiple Comparisons: Theory and Methods. Guilford School Practitioner. Taylor & Francis.

Klingenberg, B. (2009). "Proof of concept and dose estimation with binary responses under model uncertainty". *Statistics in Medicine*, **28** (2), pp. 274–292.

Krengel, U. (2002). "Approximationen der Binomialverteilung". In: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Vol. 59. vieweg studium; Aufbaukurs Mathematik. Vieweg+Teubner Verlag.

Krishna, R. (2006). Dose Optimization in Drug Development. Drugs and the Pharmaceutical Sciences. Taylor & Francis.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). "On closed testing procedures with special reference to ordered analysis of variance". *Biometrika*, **63** (3), pp. 655–660.

Pinheiro, J., Bornkamp, B., and Bretz, F. (2006). "Design and Analysis of Dose-Finding Studies Combining Multiple Comparisons and Modeling Procedures". *Journal of Biopharmaceutical Statistics*, **16** (5), pp. 639–656.

Pinheiro, J., Bornkamp, B., Glimm, E., and Bretz, F. (2014). "Model-based dose finding under model uncertainty using general parametric models". *Statistics in Medicine*, **33** (10), pp. 1646–1661.

Rosenberger, W. F. and Haines, L. M. (2002). "Competing designs for phase I clinical trials: a review". *Statistics in Medicine*, **21** (18), pp. 2757–2770.

Roy, S. N. and Bose, R. C. (1953). "Simultaneous Confidence Interval Estimation". *The Annals of Mathematical Statistics*, **24** (4), pp. 513–536.

Ruberg, S. J. (1995). "Dose response studies I. some design considerations". *Journal of Biopharmaceutical Statistics*, **5** (1), pp. 1–14.

Seber, G. and Wild, C. (2003). Nonlinear Regression. Wiley Series in Probability and Statistics. Wiley.

Selmaj, K. et al. (2013). "Siponimod for patients with relapsing-remitting multiple sclerosis (BOLD): an adaptive, dose-ranging, randomised, phase 2 study". *The Lancet Neurology*, **12** (8), pp. 756–767.

Senn, S. (1997). Statistical Issues in Drug Development. Statistics in Practice. John Wiley & Sons.

Shaffer, J. P. (1995). "Multiple Hypothesis Testing". *Annual Review of Psychology*, (46), pp. 561–584.

The Washington Post (December 10, 2013). "What 'Dallas Buyers Club' got wrong about the AIDS crisis". Dylan Matthews. Visited on 04.11.2014. URL: `http://www.washingtonpost.com/blogs/wonkblog/wp/2013/12/10/what-dallas-buyers-club-got-wrong-about-the-aids-crisis/`.

Ting, N. (2006). Dose Finding in Drug Development. Statistics for Biology and Health. Springer.

Tukey, J. W. (1953). "The problem of multiple comparisons". Unpublished manuscript in: H. I. Braun, ed. (1994). The collected works of John W. Tukey. Vol. 8. Multiple comparisons: 1948–1983. New York: Chapman and Hall, Ltd.

Unkelbach, H. and Wolf, T. (1985). Qualitative Dosis-Wirkungs-Analysen: Einzelsubstanzen und Kombinationen. Fischer.

Verrier, D., Sivapregassam, S., and Solente, A.-C. (2014). "Dose-finding studies, MCP-Mod, model selection, and model averaging: Two applications in the real world". *Clinical Trials*, **11** (4), pp. 476–484.

Westfall, P. and Young, S. (1993). Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. A Wiley-Interscience publication. Wiley.

Williams, D. A. (1971). "A Test for Differences between Treatment Means When Several Dose Levels are Compared with a Zero Dose Control". *Biometrics*, **27** (1), pp. 103–117.

Appendix

A MCP-Mod for Binary Data: Simulation Results for the Naive Approach (Relates to subsection 3.2.1)

Table A.1: Power for the Naive Application of the MCP-Mod Approach to Binary Data (Full Table)

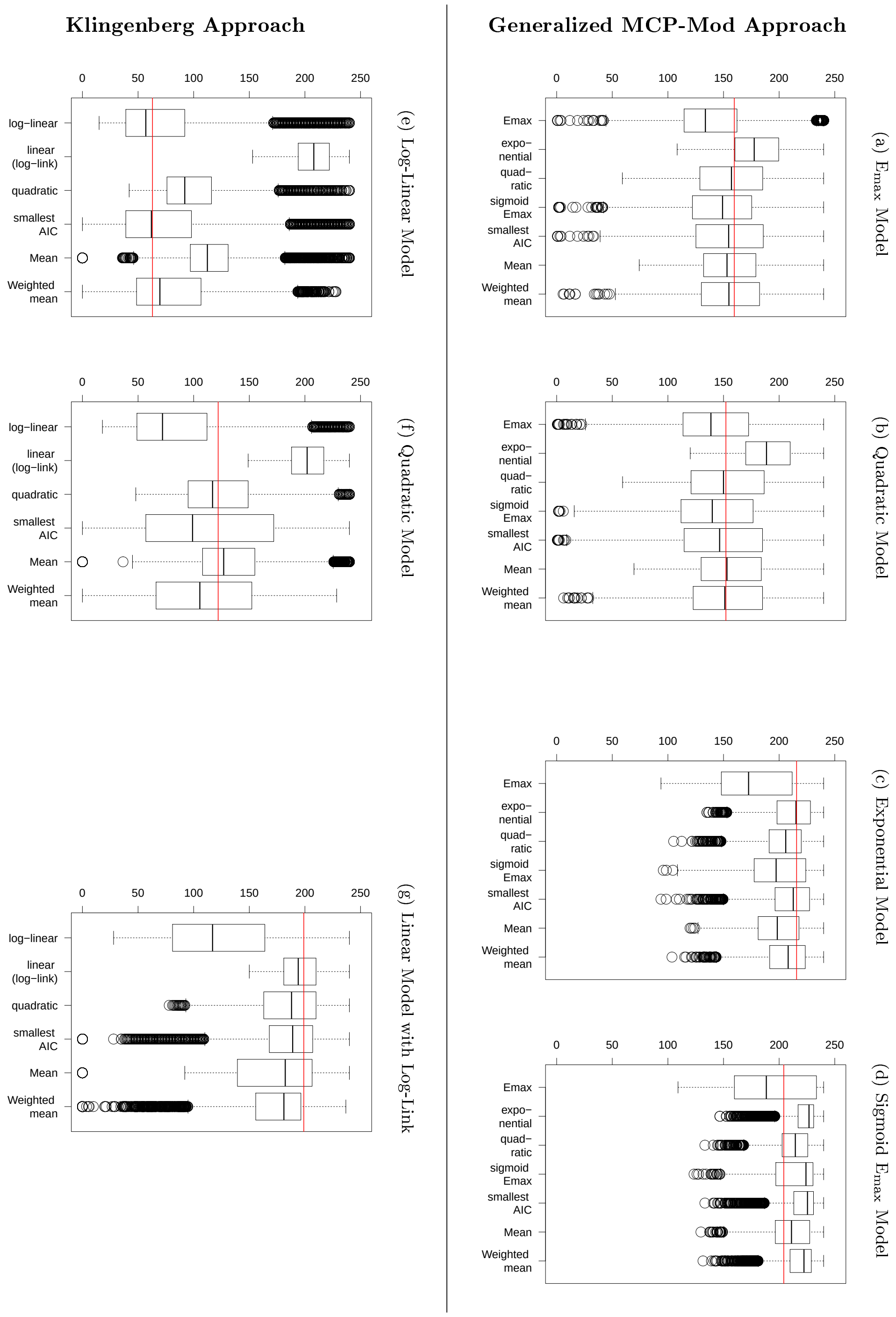
Scenario	**Optimal Doses**	**Sample Size**	**Average Power**	**Minimum Power**	**Maximum Power**	**Power to Reject at least one H_0**	**Power Linear Model**	**Power E_{max} Model**	**Power Exponential Model**	**Power Quadratic Model**
Scenario 1	0-20-22.5-50	9 5 3 9	0.6818	0.6739	0.6904	0.7657	0.6904	0.674	0.689	0.6739
		20 20 20 20	0.9501	0.9485	0.9527	0.9646	0.9527	0.9485	0.9492	0.9498
		40 40 40 40	0.9992	0.999	0.9994	0.9997	0.9994	0.999	0.9991	0.9991
Scenario 2	0-5-27.5-50	2 2 2 3	0.8212	0.8113	0.8247	0.8308	0.8247	0.8113	0.8245	0.8243
Scenario 3	0-7.5-10-25-27.5-50	42 11 10 8 27 54	0.4385	0.2691	0.4991	0.5096	0.4991	0.2691	0.4988	0.4869
		60 60 60 60 60 60	0.8577	0.8217	0.8742	0.9017	0.8742	0.8217	0.8719	0.863
Scenario 4	0-5-25-50	12 6 10 11	0.6164	0.5814	0.6471	0.6709	0.5938	0.6431	0.5814	0.6471
		70 70 70 70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
		150 150 150 150	1	1	1	1	1	1	1	1

Table A.2: Type-I Error for the Naive Application of the MCP-Mod Approach to Binary Data (Full Table)

Scenario	**Optimal Doses**	**Sample Size**	**Average Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**	**Type-I Error of Rejecting at least one H_0**	**Type-I Error Linear Model**	**Type-I Error E_{max} Model**	**Type-I Error Exponential Model**	**Type-I Error Quadratic Model**
Scenario 1	0-20-22.5-50	9 5 3 9	0.0425	0.0393	0.0456	0.0493	0.0394	0.0456	0.0393	0.0456
		20 20 20 20	0.054	0.0518	0.0553	0.0659	0.0538	0.055	0.0518	0.0553
		40 40 40 40	0.0444	0.0434	0.0456	0.0579	0.0434	0.0456	0.0434	0.0452
Scenario 2	0-5-27.5-50	2 2 2 3	0.0352	0.0192	0.0487	0.0519	0.0265	0.0487	0.0192	0.0465
Scenario 3	0-7.5-10-25-27.5-50	42 11 10 8 27 54	0.0298	0.0051	0.047	0.048	0.0439	0.0051	0.047	0.0231
		60 60 60 60 60 60	0.098	0.0636	0.1127	0.1254	0.1126	0.0636	0.1127	0.1031
Scenario 4	0-5-25-50	12 6 10 11	0.0638	0.0609	0.066	0.0891	0.0625	0.066	0.0609	0.0659
		70 70 70 70	0.0353	0.0346	0.0362	0.0548	0.035	0.0362	0.0346	0.0354
		150 150 150 150	0.033	0.0327	0.0334	0.0481	0.0327	0.0327	0.0334	0.0333

B Comparison of the Methods for Binary Data - Precision (Relates to subsection 4.1.2)

B.1 Box Plots of MED Estimates from Significant Models



Klingenberg Approach

Generalized MCP-Mod Approach

(a) E$_{\max}$ Model

(b) Quadratic Model

(c) Exponential Model

(d) Sigmoid E$_{\max}$ Model

(e) Log-Linear Model

(f) Quadratic Model

(g) Linear Model with Log-Link

Table B.3: Precision of Target Dose Estimates - Failures, High Response Scenario

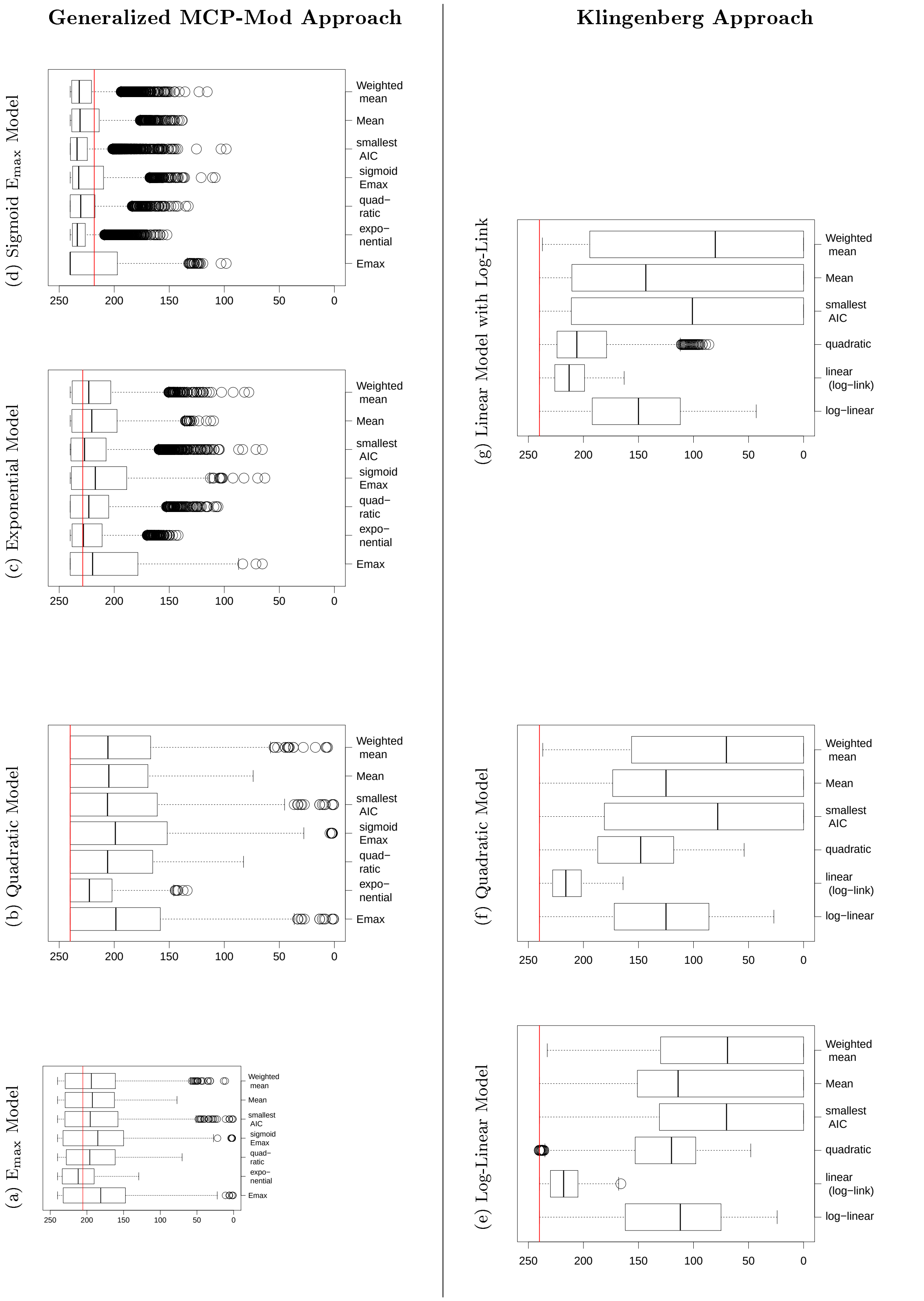
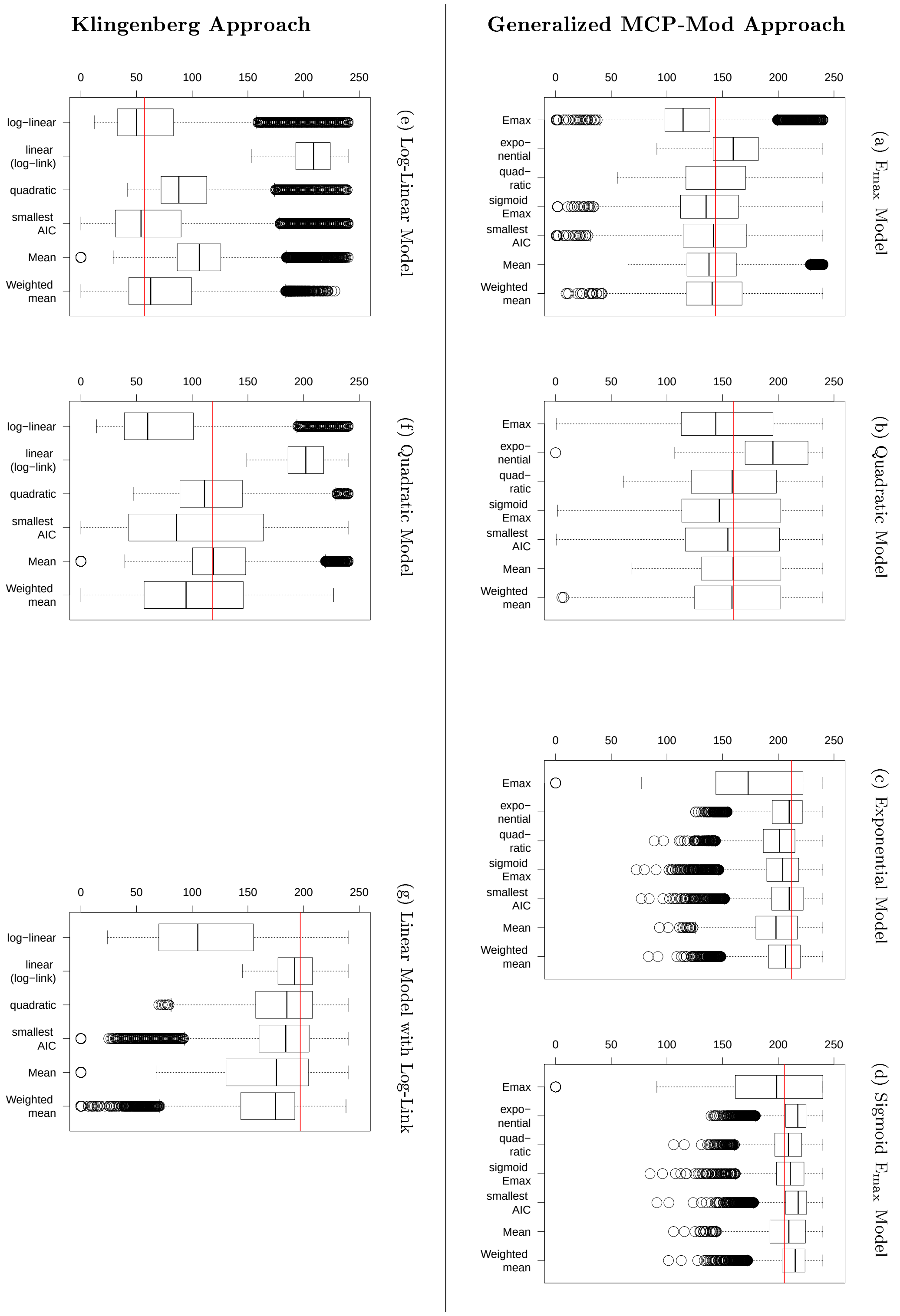


Table B.4: Precision of Target Dose Estimates - Failures, Low Response Scenario

B.2 Box Plots of MED Estimates from Significant & Non-Significant Models



Klingenberg Approach

Generalized MCP-Mod Approach

(e) Log-Linear Model

(f) Quadratic Model

(g) Linear Model with Log-Link

(a) E_{max} Model

(b) Quadratic Model

(c) Exponential Model

(d) Sigmoid E_{max} Model

Table B.5: Precision of Target Dose Estimates - Naives, High Response Scenario - Unconditional Distributions

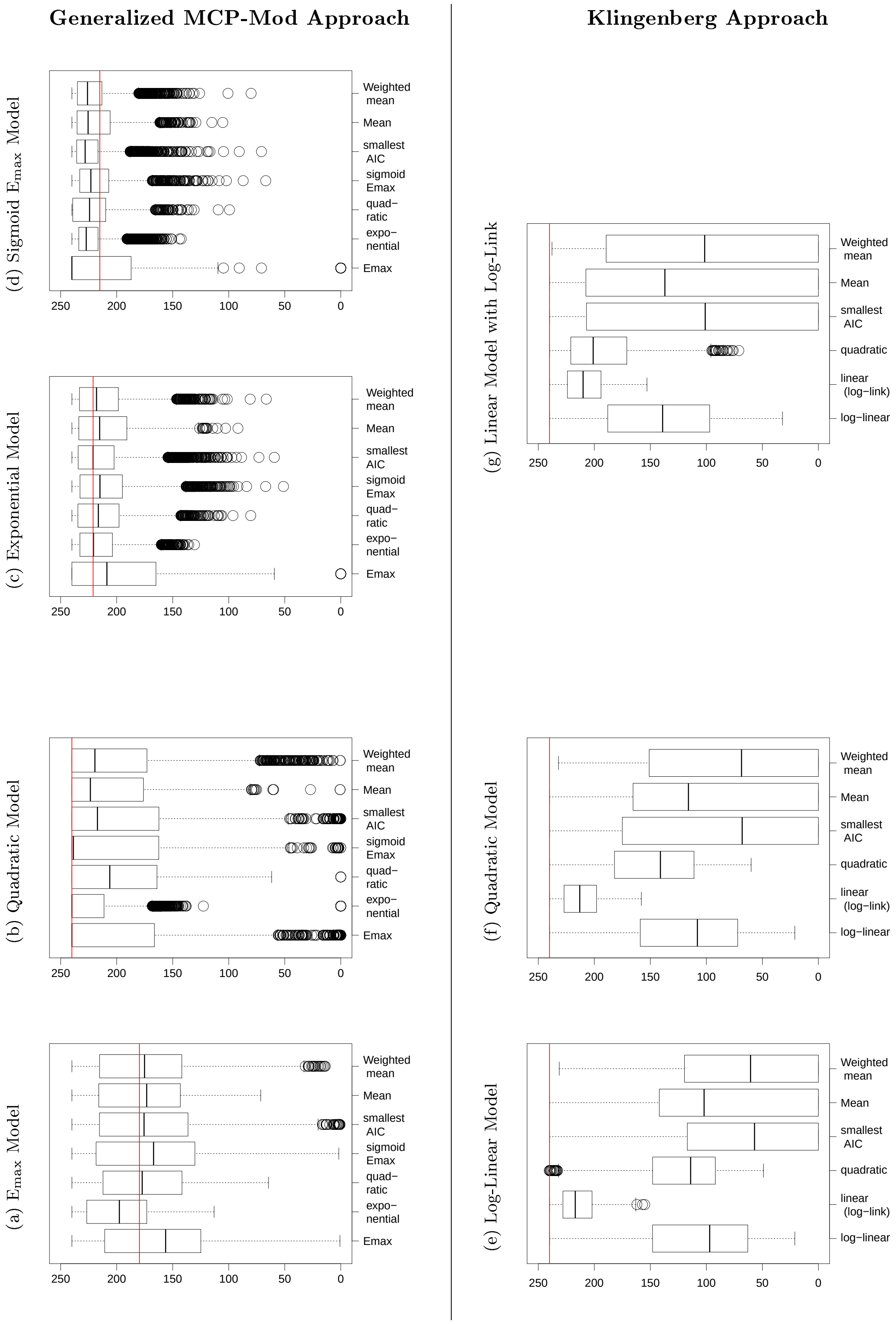
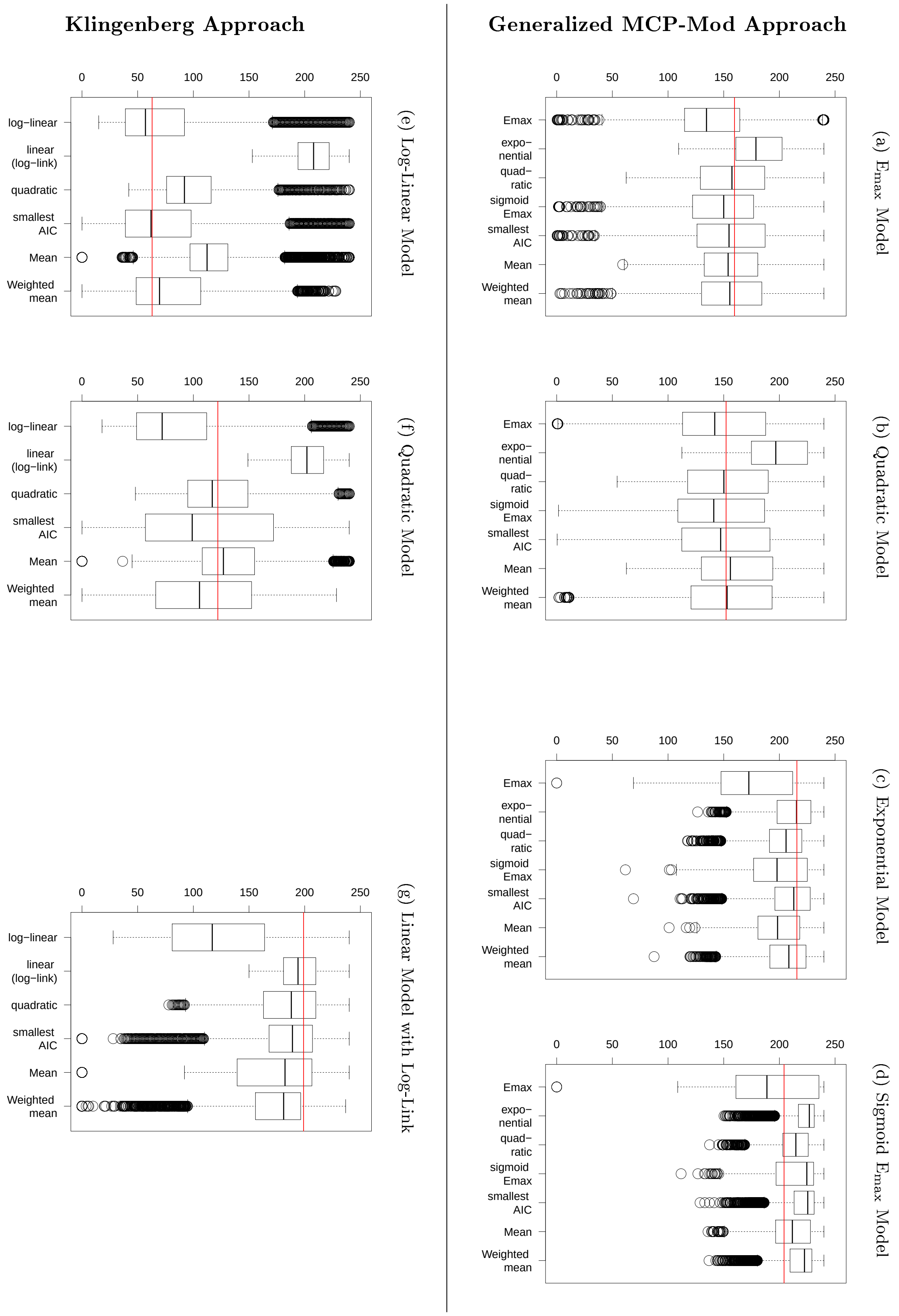


Table B.6: Precision of Target Dose Estimates - Naives, Low Response Scenario - Unconditional Distributions



Klingenberg Approach

Generalized MCP-Mod Approach

(a) Emax Model

(b) Quadratic Model

(c) Exponential Model

(d) Sigmoid Emax Model

(e) Log-Linear Model

(f) Quadratic Model

(g) Linear Model with Log-Link

Table B.7: Precision of Target Dose Estimates - Failures, High Response Scenario - Unconditional Distributions

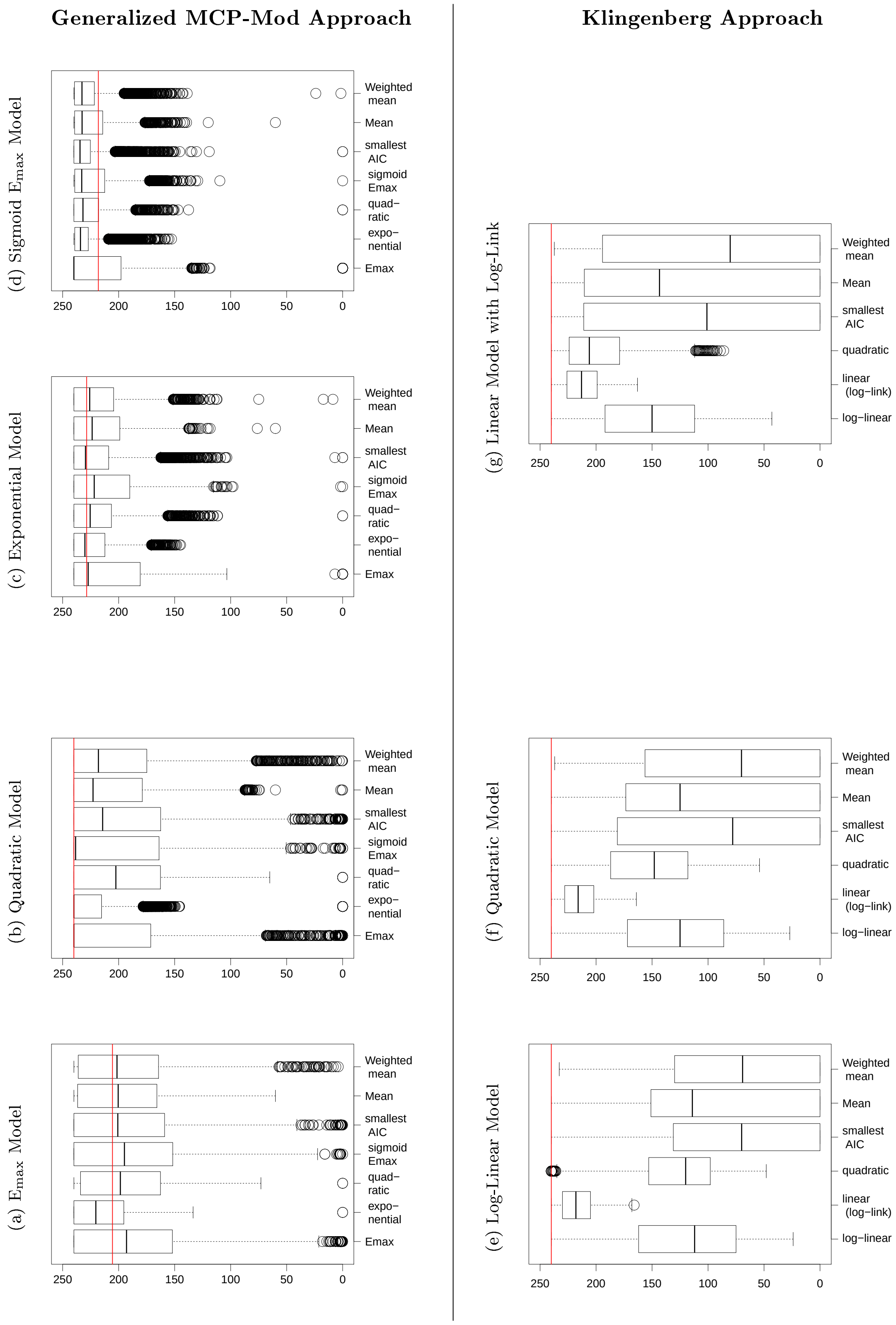


Table B.8: Precision of Target Dose Estimates - Failures, Low Response Scenario - Unconditional Distributions

C Combination of Study Results - Type-I Error for Pooled Analysis Approach (Relates to subsection 4.2.1)

Table C.9: Type-I Error Results for the Pooled Analysis - Equivalent Dose Groups and Different Sample Sizes

Scenario Naives	**Scenario Failures**	**E_{max} Model**	**Exponential Model**	**Quadratic Model**	**Sigmoid E_{max} Model**	**Type-I Error of Rejecting at least one H_0**	**Mean Type-I Error**	**Minimum Type-I Error**	**Maximum Type-I Error**
high	high	0.0288	0.0303	0.0286	0.0302	0.0487	0.0295	0.0286	0.0303
high	low	0.0265	0.0303	0.0256	0.0302	0.0470	0.0282	0.0256	0.0303
low	high	0.0276	0.0285	0.0272	0.0288	0.0460	0.0280	0.0272	0.0288
low	low	0.0304	0.0298	0.0305	0.0302	0.0503	0.0302	0.0298	0.0305

Name: Julia Krzykalla

Statutory Declaration

I declare that I have developed and written this master's thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked.
This master's thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

München, April 16, 2015 ....

Julia Krzykalla