Aus dem Institut für Medizinische Biometrie und Informatik
Universitätsklinik Heidelberg
Abteilung Medizinische Biometrie
Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Flexible Designs for Single-Arm Phase II Trials in Oncology

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Stefan Englert

aus
Schweinfurt, Bayern

2013

Dekan:	Prof. Dr. med. Claus R. Bartram
Doktorvater:	Prof. Dr. sc. hum. Meinhard Kieser

Contents

List of Tables

List of Figures

List of Source Codes

List of Abbreviations and Symbols

Statistics may be defined as "a body of methods for making wise decisions in the face of uncertainty."

(Wilson Allen Wallis)

1

Introduction

Motivated by the recent developments in adaptive and flexible designs methodology, we propose new methods for oncological phase II designs with the option to redesign aspects of the trial in a flexible manner. Before we describe our proposed design and investigate its characteristics, we introduce the context.

1.1. Development of new therapies and role of phase II trials

Clinical trials in the development of new therapies are divided into four different phases (ICH Topic E 8, 1998). Following pre-clinical research, phase I trials are usually the first studies in which a new drug or therapy is tested in human subjects. Their aim is to assess the toxicity (pharmacovigilance), tolerability, pharmacokinetics and pharmacodynamics of the new drug. In contrast to phase I studies, where healthy people are normally recruited, the primary objective of phase II studies is to explore the therapeutic efficacy in the targeted patient group. Once safety and anti-disease activity have been demonstrated, large-scale phase III studies are conducted to confirm how effective the treatment is. Phase III studies are usually designed as randomized controlled trials (RCT) comparing the new treatment with the current standard treatment or a placebo. Once effectiveness has been demonstrated in phase III studies, it is possible to obtain approval for market release from the appropriate regulatory agencies, such as the US Food and Drug Administration (FDA) or the European Medicines Agency (EMA). Phase IV trials are post-approval studies and are sometimes referred to as post-marketing surveillance trials. They are designed to

detect drug-drug interactions, rare adverse events and side-effects of the approved drug.

Phase II designs play a key role within the clinical development process. Their main objective is to provide the information required for the decision to progress to a phase III trial or to halt the development of the therapy. The consequences of a wrong decision may be far-reaching: Ending the clinical evaluation of an effective therapy implies that future patients will be deprived of a valuable therapeutic option. On the other hand, continuing the development of an ineffective drug leads to a significant binding of resources. The costs of phase III studies usually amount to several million euros and hundreds of patients are treated. Additionally, the use of an ineffective therapy would expose the patient collective to unnecessary risks during the course of the phase III study. According to the FDA, only between 5% and 8% of all products investigated in phase I go on to receive a license for commercial use (FDA, 2004; Adjei et al., 2009). Due to the key role of phase II designs, there is an urgent need for adequate and well-designed phase II trials.

This thesis focuses on oncology trials in early phase II with a binary primary endpoint, e.g., tumor response. In this setting, there is an ongoing debate on the relative merits of single-arm versus randomized phase II trials (see, for example, Gan et al., 2010; Stewart, 2010). Single-arm trials are recommended if limited numbers of patients are available for recruitment or if a single-agent therapy is to be tested in pretreated patients. Randomized trials should be used in the case of inadequate historical data or if evaluation of clinical response is difficult or highly sensitive to clinical, pathologic or molecular parameters. Randomized phase II oncology studies are, however, still the exception (Stone et al., 2007). Therefore, single-arm phase II trials remain an essential tool in cancer research (Gan et al., 2010).

1.2. Aims and structure of the thesis

The designs for single-arm phase II oncology trials presented in the literature do not permit the flexibility desired in conducting clinical trials. Let us consider the clinical evaluation of a single-agent therapy, conducted according to some single-arm phase II design. The complete layout of the trial has to be specified in detail *a priori* and adhered to strictly during the course of the study. Especially in early drug development, there is usually a considerable extent of uncertainty in the planning stage of a clinical trial. It may become apparent during the trial that the initial assumptions do not hold true. For example, data of a parallel trial may indicate that the efficacy of the new therapy is higher than anticipated (or lower but still clinically relevant). In this situation, the initially planned design is inadequate and modification would be appropriate. In phase II designs to date, however, the study still has to be conducted and evaluated according to the initially

specified rules. As another example, data on patients treated may continue to accumulate for some time after the criterion for ending the study is fulfilled. This can occur for various reasons. In multicenter clinical trials, for example, it may be difficult to exactly time the end of recruitment. This phenomenon is generally known as overrunning. On the other hand, safety considerations or insufficient recruitment may force the trial to end before the stopping criterion has been fulfilled. This is called underrunning. In both situations, the procedure for proper inference is unclear.

The aim of the present work is to develop and evaluate methods that allow flexibility in the setting of single-arm phase II study designs in oncology. Before proceeding to present our proposed method and evaluate our findings, we give in Chapter 2 some background information on the complex of themes that will be covered. We start with an overview of phase II trials in oncology in Section 2.1 and present standard designs that are frequently used in this setting. In Section 2.2, we introduce existing design methods that allow flexibility in the course of a trial. The methods developed to date are oriented towards controlled trials and continuous outcomes; application to discrete test statistics, as in oncological phase II trials, has not been investigated so far.

In Chapter 3, we show that direct application of flexible methods to oncological phase II designs will lead to inflation of the type I error rate or to conservative methods. Strict control of the specified type I error rate is a desirable property of study design and is especially important when these trials play a major role in the approval process of new therapies. Approval of a new oncological drug is sometimes based solely on phase II results (Gan et al., 2010; Tsimberidou et al., 2009). In a regulated environment strict control of the type I error rate is mandatory (Committee for Medicinal Products for Human Use, 2007; ICH Topic E 9, 1998). In fact, from 1973 through 2006, 46% (31/68) of investigational anti-cancer drugs were approved without randomized trials that used a comparator (Tsimberidou et al., 2009). Furthermore, approvals by the FDA for solid tumors from 1998 through 2008 were based solely on single-arm phase II data in 13% (4/34) of indications for cytotoxic drugs and in 15% (4/26) of indications for targeted drugs (Gan et al., 2010).

In Chapters 4 and 5 we present in detail the approach we developed to achieve the goal of flexible phase II oncology trials with control of the type I error rate. Designs are developed for the setting of a fixed second-stage sample size (Chapter 4) and for the situation that already in the planning phase the second-stage sample size may depend on the interim outcome (Chapter 5). For each case, an algorithm is provided to construct the proposed phase II designs.

The performance of the resulting designs is investigated in Chapter 6, where we also analyze the profile of different recalculation strategies and describe their properties. Ap-

plication of the proposed method is demonstrated in Chapter 7 with the example of a clinical trial conducted by Combs et al. (2012) that uses the methodology described in this thesis. We close with a discussion of the findings in Chapter 8.

This work is based on the articles Englert and Kieser (2012a), Englert and Kieser (2012b), Englert and Kieser (2013a), and Englert and Kieser (2013b).

To understand God's thoughts we must study statistics, for these are the measure of His purpose.

(Florence Nightingale)

2

Background

2.1. Phase II trials in oncology

It is the aim of clinical phase II trials in oncology to determine whether a new agent or combination of agents has sufficient anti-tumor activity to merit further investigation in larger patient groups. In contrast with phase II designs in other medical fields, these trials are usually not performed in a controlled design but as single-arm studies. Treatment performance is evaluated according to the RECIST (Response Evaluation Criteria In Solid Tumors) guidelines (Therasse et al., 2000; Eisenhauer et al., 2009). The primary endpoint is a binary response variable indicating treatment success. In these trials, the null hypothesis that the response rate π is lower than a pre-specified uninteresting response rate π_0 is tested. Since a new therapy will only be investigated further if the response rate is higher than π_0, a one-sided test is performed. To formalize the statistical approach, the following null and alternative hypothesis can be defined:

$$H_0 : \pi \leq \pi_0 \text{ vs. } H_1 : \pi > \pi_0.$$

The study is usually powered to reject the null hypothesis for a response rate π_1 ($\pi_1 > \pi_0$) that indicates sufficient efficacy.

In the following, we only consider designs for the simple null hypothesis $H_0 : \pi = \pi_0$ and powered for the simple alternative $H_1 : \pi = \pi_1$. Such tests are appropriate for testing the composite null hypothesis $H_0 : \pi \leq \pi_0$ versus the composite alternative hypothesis $H_1 : \pi \geq \pi_1$, as the power function is monotone in π (Chang et al., 1987). For convenience, and adhering to the usual formulation in oncological phase II trials, the hypotheses tested

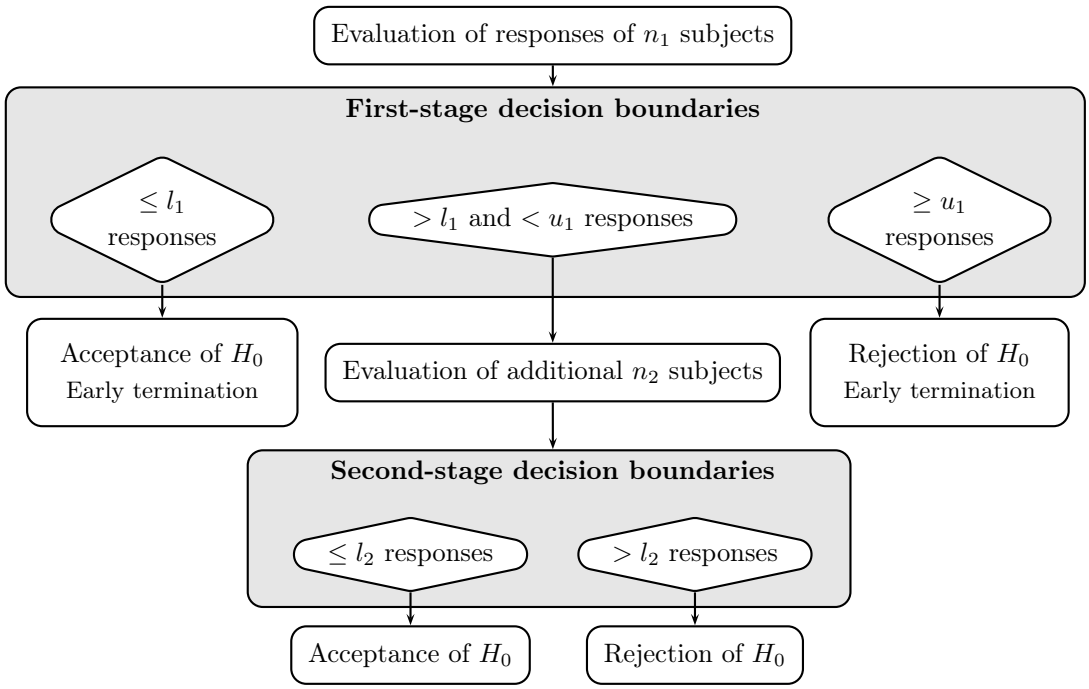


Figure 2.1.: *Layout of classical phase II designs in oncology*

are written as

$$H_0 : \pi = \pi_0 \text{ vs. } H_1 : \pi = \pi_1.$$

Note that rejection of the null hypothesis leads to the conclusion that $\pi > \pi_0$, i.e., the drug is promising enough to move to the next step in drug development, not that $\pi > \pi_1$. If the treatment has not shown sufficient activity to reject the null hypothesis, it is convenient to say that the null hypothesis is accepted.

Typically, phase II trials in oncology are performed due to ethical considerations with planned interim analyses to allow early termination. Usually, only one interim analysis is implemented according to logistical and efficiency considerations (Mariani and Marubini, 1996; McPherson, 1982). In the first stage, n_1 patients are recruited and treated. If the number of observed responses out of the n_1 initial patients is less or equal than l_1 or at least u_1, the trial is terminated after the interim analysis for futility or efficacy, respectively. Otherwise, the trial continues to the second stage with inclusion of a further n_2 patients. If the total number of observed responses after stage two exceeds l_2, the treatment is proven to be promising for further investigation. Otherwise, the treatment is rejected. The general layout of the design is illustrated in Figure 2.1.

If the true response rate is π, these trials are terminated after the first stage with proba-

bility

$$\mathrm{PET}(\pi) = \sum_{k=0}^{l_1} b(k; \pi, n_1) + \sum_{k=u_1}^{n_1} b(k; \pi, n_1), \tag{2.1}$$

where b denotes the binomial probability mass function. This leads to an expected sample size of

$$\mathrm{EN}(\pi) = n_1 \cdot \mathrm{PET}(\pi) + (n_1 + n_2) \cdot \big(1 - \mathrm{PET}(\pi)\big).$$

The type I and II error rates are given by

$$\alpha' = 1 - \left[B(l_1; \pi_0, n_1) + \sum_{k=l_1+1}^{\min(n_1, u_1-1)} b(k; \pi_0, n_1) \cdot B(l_2 - k; \pi_0, n_2) \right] \tag{2.2}$$

and

$$\beta' = B(l_1; \pi_1, n_1) + \sum_{k=l_1+1}^{\min(n_1, u_1-1)} b(k; \pi_1, n_1) \cdot B(l_2 - k; \pi_1, n_2) \tag{2.3}$$

with the cumulative binomial distribution function B.

The sample sizes n_1 and n_2 and the decision boundaries u_1, l_1 and l_2 are determined such that the type I error rate is at most α and the type II error rate is at most β under the null hypothesis $H_0 : \pi = \pi_0$ and the alternative $H_1 : \pi = \pi_1$, respectively, where π_0 and π_1, $\pi_0 < \pi_1$, define insufficient and promising anti-tumor activity. One class of designs stops after the first stage for futility only, i.e., if the initial proportion of observed responses is too low. Not stopping for reasons of efficacy after the first stage might be in the best interest of the patients. The patient collective would be treated with an effective therapy. Methodologically, this can be achieved by additionally requiring $u_1 > n_1$. The designs by Simon (1989) are the most popular representatives of this class of designs. However, there are also situations where it is desirable to terminate a phase II trial early if the initial response rate is high enough to give evidence of activity (Fleming, 1982; Shuster, 2002). For example, ending a trial early because efficacy has been demonstrated speeds up the development process. The new therapy can move on to phase III faster than would have been the case otherwise.

Among all parameter constellations $(l_1, u_1, n_1, l_2, n_2)$ fulfilling the type I and type II error constraints, a specific one is usually selected to satisfy an optimality criterion that is appropriate for the objectives of the current trial. Up to now, a multitude of optimality criteria and corresponding designs have been developed, with minimization of, for example, (a) the expected sample size under the null hypothesis (Simon, 1989), (b) the maximum sample size $n = n_1 + n_2$ (Simon, 1989), (c) the average of the expected sample size under the null and alternative hypotheses (Chang et al., 1987), (d) the average or maximum sample size under the alternative hypothesis (Mander and Thompson, 2010),

Table 2.1.: *Simon's optimal designs ($\pi_1 - \pi_0 = 0.2$)*

π_0	π_1	α	β	l_1	n_1	l_2	n_2	n	EN(π_0)	α'	β'
0.05	0.25	0.05	0.2	0	9	2	8	17	12.0	0.047	0.188
		0.05	0.1	0	9	3	21	30	16.8	0.049	0.098
0.1	0.3	0.05	0.2	1	10	5	19	29	15.0	0.047	0.195
		0.05	0.1	2	18	6	17	35	22.5	0.047	0.098
0.2	0.4	0.05	0.2	3	13	12	30	43	20.6	0.050	0.200
		0.05	0.1	4	19	15	35	54	30.4	0.048	0.096
0.3	0.5	0.05	0.2	5	15	18	31	46	23.6	0.050	0.197
		0.05	0.1	8	24	24	39	63	34.7	0.050	0.097
0.4	0.6	0.05	0.2	7	16	23	30	46	24.5	0.049	0.199
		0.05	0.1	11	25	32	41	66	36.0	0.049	0.098
0.5	0.7	0.05	0.2	8	15	26	28	43	23.5	0.050	0.196
		0.05	0.1	13	24	36	37	61	34.0	0.049	0.099
0.6	0.8	0.05	0.2	7	11	30	32	43	20.5	0.049	0.198
		0.05	0.1	12	19	37	34	53	29.5	0.043	0.099
0.7	0.9	0.05	0.2	4	6	22	21	27	14.8	0.049	0.196
		0.05	0.1	11	15	29	21	36	21.2	0.046	0.095

(e) the median sample size (Hanfelt et al., 1999), or (f) the globally maximized expected sample size for all $\pi \in [0, 1]$ (Shuster, 2002). Table 2.1 (for $\pi_1 - \pi_0 = 0.2$) and Table B.1 in the Appendix (for $\pi_1 - \pi_0 = 0.15$) list for a variety of parameter constellations the layout of Simon's phase II designs minimizing the expected sample size under the null hypothesis (optimal designs). Tables 2.2 (for $\pi_1 - \pi_0 = 0.2$) and B.2 (for $\pi_1 - \pi_0 = 0.15$) list the corresponding designs minimizing the maximum sample size (minimax designs). As Simon (1989) does not allow for early stopping for efficacy, the parameter $u_1 > n_1$ is omitted in the presentation.

In recent years, a multitude of variants and refinements of these designs have been suggested. Jung et al. (2004), for example, considered admissible designs that are a compromise between designs minimizing the average sample size under the null hypothesis and the total sample size. Further generalizations include the implementation of a third stage (Chang et al., 1987; Chen, 1997; Ensign et al., 1994), the use of stratification (London and Chang, 2005; Sargent et al., 2001; Tournoux-Facon et al., 2011; Chang et al., 2012), consideration of a combined endpoint or two or more endpoints (Chen and Chi, 2011; Lin and Chen, 2000; Lin et al., 2008; Kunz and Kieser, 2011a), use of stochastic and non-stochastic curtailment (Kunz and Kieser, 2011b, 2012; Chen and Chi, 2011), and the application of full sequential plans (Tan and Xiong, 2010).

Additionally, designs have been presented where the second-stage sample size and the

Table 2.2.: *Simon's minimax designs* ($\pi_1 - \pi_0 = 0.2$)

π_0	π_1	α	β	l_1	n_1	l_2	n_2	n	EN(π_0)	α'	β'
0.05	0.25	0.05	0.2	0	12	2	4	16	13.8	0.043	0.199
		0.05	0.1	0	15	3	10	25	20.4	0.034	0.099
0.1	0.3	0.05	0.2	1	15	5	10	25	19.5	0.033	0.198
		0.05	0.1	2	22	6	11	33	26.2	0.041	0.098
0.2	0.4	0.05	0.2	4	18	10	15	33	22.3	0.046	0.199
		0.05	0.1	5	24	13	21	45	31.2	0.048	0.100
0.3	0.5	0.05	0.2	6	19	16	20	39	25.7	0.045	0.196
		0.05	0.1	7	24	21	29	53	36.6	0.047	0.098
0.4	0.6	0.05	0.2	17	34	20	5	39	34.4	0.049	0.198
		0.05	0.1	12	29	27	25	54	38.1	0.049	0.099
0.5	0.7	0.05	0.2	12	23	23	14	37	27.7	0.048	0.199
		0.05	0.1	14	27	32	26	53	36.1	0.046	0.100
0.6	0.8	0.05	0.2	8	13	25	22	35	20.8	0.050	0.192
		0.05	0.1	15	26	32	19	45	35.9	0.044	0.100
0.7	0.9	0.05	0.2	19	23	21	3	26	23.2	0.045	0.199
		0.05	0.1	13	18	26	14	32	22.7	0.050	0.099

corresponding decision boundary depend in a pre-specified way on the responses observed in the interim analysis. These methods can result in more effective phase II designs, but require that the rule for the design modifications is already specified in the protocol. First attempts to construct such designs were made by Lin and Shih (2004) and Banerjee and Tsiatis (2006). Lin and Shih presented a design where, based on the results of the interim analysis, the study is powered for either a skeptical or an optimistic target response rate. Therefore, depending on the number of responses in the first stage, two different sample sizes are possible. Banerjee and Tsiatis used a Bayesian decision-theoretic construct to develop optimal adaptive two-stage designs. In the setting of two parallel Simon designs, Jones and Holmgren (2007) presented how one design can influence the other.

As a common feature of all these approaches, the sample sizes of the two stages and the decision rules for the interim and the final analysis have to be specified *a priori* and adhered to strictly during the course of the study in order to assure control of the type I error rate. However, there is usually a considerable degree of uncertainty in the planning stage of a clinical trial, especially in early drug development. For example, the activity of the new agent may be higher than anticipated when specifying the target response rate (or lower but still of clinical importance). Even if the interim results suggest that the initial assumptions would not hold true and modification of the design would therefore be appropriate, the study has to be continued according to the specified rules. As another example, stopping recruitment exactly after accrual of a predefined number of patients

may be difficult, especially in multicenter studies, possibly leading to violation of the predetermined sample size.

Some attempts have been made to ease the strong restrictions imposed by the common two-stage designs. Green and Dahlberg (1992) described approaches to deal with the situation that the attained sample size is not equal to the planned one. However, this method only allows for reaction to unintentional over- or underrunning and does not guarantee that the significance level is kept. Following up Green and Dahlberg, Chen and Ng (1998) developed flexible designs with control of the average type I error rate under the assumption that each of the possible scenarios of over- or underrunning has the same probability of occurrence. For three specific scenarios, Wu and Shih (2008) investigated how to handle the data of a phase II trial when Simon's two-stage design is pre-specified but the trials deviates from it. Koyama and Chen (2008) presented a method for proper inference from Simon's two-stage design when the actual sample size in the second stage differs from the planned one. Their procedures, however, only cover unintentional or non-informative sample size changes, i.e., when the decision to change the sample size of the second stage was made blinded to any information gained throughout the trial. Therefore, unforeseen events or situations where a variety of aspects have to be taken into account when re-shaping the design of an ongoing trial cannot be adequately handled by these procedures. For example, data from the current or a parallel trial may indicate that the response rate of the treatment is higher than expected (or lower but still clinically relevant), making a decrease (or increase) of the sample size desirable.

Due to these limitations, the universe of potential situations arising in practice cannot be covered. Therefore, two-stage designs that allow greater flexibility while maintaining the type I error rate would be desirable. Adaptive design methodology that has been developed to perform arbitrary design modifications while controlling the type I error rate is presented in the next section.

2.2. Adaptive and flexible designs

In confirmatory clinical trials with fixed sample size and a given significance level, the total sample size of the design is chosen to assure sufficient statistical power for a pre-specified treatment difference. In practice, the prior assumption with regard to the treatment effect is derived from earlier studies or literature. Through data from a parallel trial it may become apparent during the conduct of the trial that the assumption does not reflect the actual treatment difference for the considered population. When this happens, the power of the trial will be different from what is needed and the trial is underpowered or oversized. In these situations, it may be desirable to adjust the design of the ongoing trial accordingly.

Adaptive designs have been proposed that enable design modifications during an ongoing clinical trial under control of the overall significance level. Special attention was given to adaptive designs in 2006, when the FDA released a Critical Path Opportunities List that calls for advancing innovative trial designs by using accumulated information in designing trials.

The term *adaptive design* is nowadays so frequently used in clinical trials methodology that a Google search will yield more than half a million direct hits (accessed January 24, 2013). Unfortunately, no clear definition of adaptive design seems to exist. Proposed definitions include the following:

> "[An] adaptive design [is] a design that allows modifications to some aspects (e.g., trial procedures and/or statistical procedures) of an on-going clinical trial after its initiation, without undermining the validity and integrity of the trial."
> Chow et al. (2005)

> "By adaptive design we [the Pharmaceutical Research Manufacturer Association (PhRMA) working group] refer to a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial."
> Gallo et al. (2006)

> "An adaptive design clinical study is defined as a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study."
> FDA (2010)

Other authors further categorize adaptive designs into classes depending on what aspects of the trial are changed. The categories used by Pong and Chow (2010) or Chow and Chang (2008) include, for example, classical and adaptive group-sequential designs, flexible sample size re-estimation methods and the drop-the-losers, adaptive dose finding, seamless phase I/II and phase II/III, adaptive randomization, hypothesis-adaptive and biomarker-adaptive designs.

In this thesis, the major focus lies on the degree of flexibility allowed by adaptive designs. Therefore, adaptive designs are classified not based on *what* adaptations are performed, but according to *how* these adaptations are carried out. The terms *adaptive design* and *flexible design*, used throughout this thesis, are defined as follows:

Definition 2.1. (Adaptive design). A *(per-design) adaptive design* is a design allowing modifications of an ongoing trial that follow strict predefined rules for adaptation.

Definition 2.2. (Flexible design). A *flexible design* is a design where the rules for design changes do not have to be pre-specified.

These definitions of *adaptive* and *flexible* correspond to the terms *planned flexible* and *fully flexible* used by Bauer (2008) and the terms *adaptivity* and *flexibility* used by Brannath et al. (2007).

Several versions of adaptive and flexible designs controlling the type I error rate are available and it is not possible to give a comprehensive overview of these designs here. In the following sections, we point out the most fundamental concepts and methods of these designs.

2.2.1. Adaptive designs

One of the first types of adaptive designs proposed in the literature were group-sequential designs. In these designs one or more interim analyses are implemented and at each interim analysis it is decided whether the study is to be stopped early for efficacy or futility or whether the study should be continued to the next stage. A special case of group-sequential designs are full sequential designs, where an interim analysis is performed after each observational unit. If significance tests at a predetermined level are performed repeatedly at each stage during data collection, it becomes more likely that a significant result will be obtained under the assumption of no effect (Armitage et al., 1969). Group-sequential methods therefore use adjusted levels at each stage to control the nominal significance level for the complete trial. These adjusted levels can be selected in a variety of possible ways. The two most common layouts, widely used in clinical research, were presented by Pocock (1977) and O'Brian and Fleming (1979). These authors proposed group-sequential plans for normal responses with known variance and a fixed number of looks into the data. In these sequential sampling schemes the final sample size is not fixed but random and depends on the interim results. In standard group-sequential tests, the sample sizes for each stage and the rules at each interim analysis are fixed. The α-spending function or use function approach proposed by Lan and DeMets (1983) and DeMets and Lan (1994) eases these restrictive requirements and allows the sample sizes of the different stages to vary. However, all following stages must be planned independently of the observed data.

Adaptive designs allow the layout of the following stages and corresponding rejection regions to depend on accumulated data of an interim analysis. This idea is closely related to

internal piloting, first explored by Wittes and Brittain (1990). For the necessary adjustment of the design parameters, several authors have suggested sample size re-estimation based on blinded data to provide an updated estimate of a nuisance parameter (Gould, 1995; Kieser and Friede, 2000, 2003; Friede and Kieser, 2004). Also, methods using the unblinded interim data have been proposed. For phase II oncology trials, such designs have been presented, for example, by Lin and Shih (2004) and Banerjee and Tsiatis (2006) (see Section 2.1). As the rule for design modifications is already a priori specified, it cannot be changed during the trial without undermining trial integrity. This is not the case for flexible designs, where changes to the design are allowed in a flexible manner based on the sequentially computed observed treatment differences. These designs may not only account for the accumulated information in the trial but may also make use of external information. Note that every flexible design with a pre-fixed adaptation rule is *per definitionem* a per-design adaptive design. This is also the reason why most per-design adaptive designs are planned using flexible design methodology. Flexible designs are described in more detail in the following section. Tsiatis and Mehta (2003) showed that per-design adaptive designs can be uniformly improved by using standard group-sequential tests based on the sequentially computed likelihood ratio test statistic. However, pre-specification of recalculation rules counteracts the flexibility desired in clinical research (Timmesfeld et al., 2007). In practice, it is impossible to specify a suitable recalculation rule that adequately reacts to every possible eventuality in the course of the trial.

2.2.2. Flexible designs

Bauer (1989a,b) introduced in a hotly debated article the idea of incorporating the unblinded data from the current trial as well as from parallel studies for mid-trial adaptations. Flexible designs methodology includes both terminating the trial early and redesigning the trial, with no necessity for pre-specification of a recalculation rule, thus achieving far-reaching flexibility. Historically, two concepts were introduced to allow for these changes under control of the nominal type I error rate. A good overview is given in a tutorial on flexible designs by Bretz et al. (2009).

We present both concepts in detail for studies with two stages, i.e., one interim analysis, continuous test statistics and a single one-sided null hypothesis. Under this setting, both approaches can be defined by the one-sided p-values p_1 and p_2 obtained from the separate stages of the trial. This ensures that the distributions of p_1 and p_2 are independent and that, per construction of continuous p-values, both p_1 and p_2 are uniformly distributed under H_0.

The two approaches, denoted as *combination test method* and *conditional error function*

method are outlined in the following sections. The former concept, introduced by Bauer and Köhne (1994), involves combining p-values of the different stages through an appropriate combination function $C(p_1, p_2)$. The latter strategy, developed by Proschan and Hunsberger (1995) is based on a function that defines the conditional type I error rate of the second stage through a function $A(p_1)$ depending on the first-stage p-value p_1.

Combination test method

A two-stage combination test procedure for testing a null hypothesis H_0 can be defined in terms of a combination function $C(p_1, p_2)$ which combines the one-sided p-values of the two stages. Usually, early stopping boundaries α_0 and α_1 with $0 \leq \alpha_1 < \alpha_0 \leq 1$ are defined. The value α_1 is the local significance level to reject the null hypothesis after the first stage and α_0 defines a futility boundary.

After the first stage of the trial, the p-value obtained from the first-stage data p_1 is computed. If p_1 falls below α_1, the trial is stopped after the first stage with rejection of the null hypothesis. If p_1 exceeds α_0, the trial is stopped for futility. Otherwise, based on the interim results and the observed test statistic p_1, adaptations may be carried out and the trial continues to the second stage. After the end of the study, the p-value of the second stage p_2 is computed and a combination function $C(p_1, p_2)$ is applied. The null hypothesis is rejected after the second stage if the combined value falls below a boundary c_α. The general layout of the combination test method is illustrated in Figure 2.2.

Several versions of combination functions have been proposed in the literature (Bauer and Köhne, 1994; Chang, 2007; Cui et al., 1999; Lehmacher and Wassmer, 1999). The three most common ones are listed below:

- Fisher's combination criterion, product of p-values (Bauer and Köhne, 1994):

$$C(p_1, p_2) = p_1 \cdot p_2$$

- Sum of p-values (Chang, 2007):

$$C(p_1, p_2) = p_1 + p_2$$

- Inverse normal (Lehmacher and Wassmer, 1999):

$$C(p_1, p_2) = \big(\Phi^{-1}(1 - p_1) + \Phi^{-1}(1 - p_2)\big)/\sqrt{2}.$$

Given a specific combination method, the values of α_0, α_1 and c_α are chosen such that the type I error rate is controlled. It is given by

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \Pr_{H_0}\big(C(p_1, P_2) \leq c_\alpha\big)\, dp_1, \tag{2.4}$$

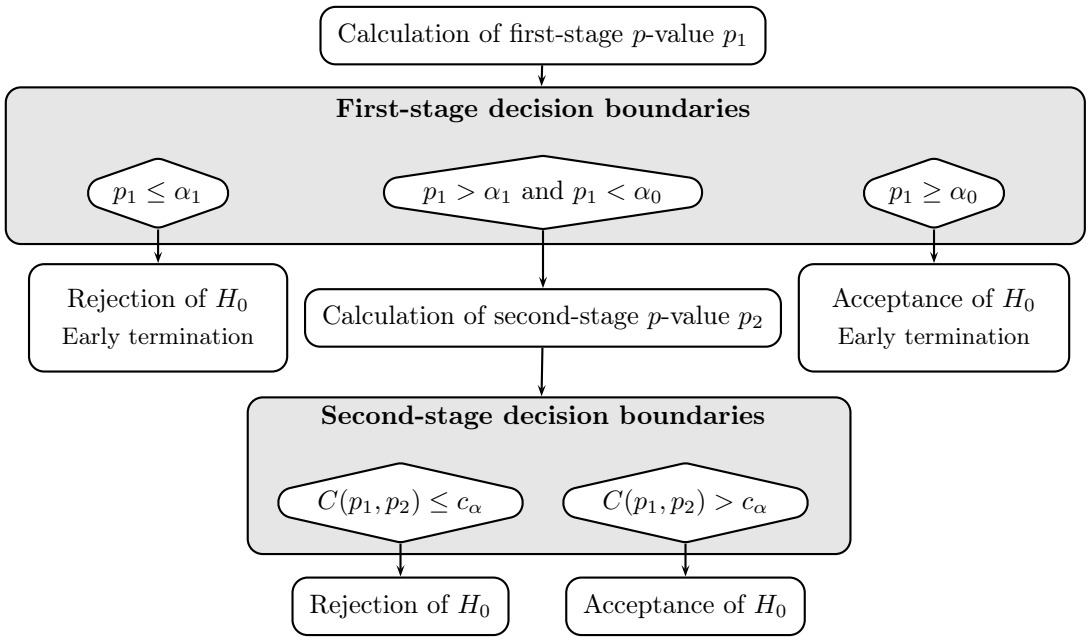


Figure 2.2.: *Layout of classical combination test procedures*

where P_2 denotes the random variable of the second-stage p-value p_2. As p_2 is uniformly distributed on $[0, 1]$ under H_0 regardless of the adaptations made after the interim analysis, $\Pr_{H_0}\big(C(p_1, P_2) \leq c_\alpha\big)$ is independent with respect to the adaptations performed. Appropriate selection of the values α_0, α_1 and c_α therefore guarantees that the type I error rate is not inflated within a flexible combination test procedure. For a formal proof see Bauer (1989a) or Bauer and Kieser (1999).

Bauer and Köhne (1994) used Fisher's combination criterion where H_0 is rejected at a specified nominal level α after the second stage if the product of the p-values from the two stages is small enough, that is if $p_1 p_2 \leq c_\alpha$, where $c_\alpha = \exp[-\frac{1}{2}\chi^2_4(1-\alpha)]$ and $\chi^2_4(1-\alpha)$ is the $(1-\alpha)$-quantile of the central χ^2 distribution with 4 degrees of freedom. Following Bauer and Köhne, the values α_0 and α_1 are determined such that the overall type I error rate (2.4) is exhausted, i.e., by solving the equation

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \Pr_{H_0}\big(p_1 P_2 \leq c_\alpha\big)\, dp_1 = \alpha_1 + \int_{\alpha_1}^{\alpha_0} c_\alpha / p_1 \, dp_1 = \alpha_1 + c_\alpha(\ln \alpha_0 - \ln \alpha_1) \stackrel{!}{=} \alpha. \quad (2.5)$$

Table 2.3 lists for different significance levels α suitable choices of c_α, α_0 and α_1. Here, c_α is calculated as above and α_1 is calculated by inverting (2.5) given the value of c_α and an early futility boundary α_0. Note that for $\alpha_0 = 1$ no early stopping for futility is considered.

Table 2.3.: *Critical boundaries for the Bauer and Köhne combination test*

α	0.10	0.05	0.025	0.010
c_α	0.0205	0.0087	0.0038	0.0013
α_0		α_1		
0.3	0.0703	0.0299	0.0131	0.0045
0.4	0.0618	0.0263	0.0115	0.0040
0.5	0.0548	0.0233	0.0102	0.0035
0.6	0.0486	0.0207	0.0090	0.0031
0.7	0.0429	0.0183	0.0080	0.0027
1	0.0205	0.0087	0.0038	0.0013

Conditional error function method

Proschan and Hunsberger (1995) introduced the concept of conditional error functions as a method for testing a null hypothesis within a two-stage design while allowing data-dependent modifications of the sample size after the first stage. In these designs, the conditional significance level of the second part of the trial depends on the outcome of the first stage in a pre-specified way while the unconditional type I error rate is still controlled.

For a given level α, the conditional error function A is defined as a non-increasing function $A(p) : [0, 1] \to [0, 1]$ with

$$\int_0^1 A(p)\, dp \leq \alpha. \tag{2.6}$$

The value $A(p_1)$ specifies the conditional type I error rate used for the second stage, given the first-stage p-value p_1. The additional requirement that $A(p)$ is non-increasing is a logical one and ensures that outcomes of the first stage that are less likely under the null hypothesis, i.e., smaller p-values, are associated with higher local significance levels of stage two.

Consider a one-sided null hypothesis H_0 which is tested in a two-stage design. Let p_1 and p_2 again denote the p-values for H_0, such that p_1 and p_2 are based only on the first- and second-stage data, respectively. If in the first stage a p-value p_1 of the corresponding random variable P_1 is realized, any stochastically independent test statistic can be used to test H_0 in the second stage at a nominal level less than or equal to $A(p_1)$. The null hypothesis is rejected if the second-stage p-value p_2 satisfies $p_2 \leq A(p_1)$. Of course, with $A(p_1) = 0$ or $A(p_1) = 1$, respectively, early stopping after the first stage with acceptance or rejection of the null hypothesis is possible. As the first- and second-stage p-values are uniformly distributed under the null hypothesis, (2.6) guarantees that the overall type I error rate is controlled even if the design characteristics of the second stage have been modified based on any information available at the end of the first stage. The concept of

conditional error function was generalized by Müller and Schäfer (2001, 2004) to implement flexible design changes any time during the course of a trial. They used a special *natural conditional error function* calculated from the planned design, the conditional rejection region. This function is then used for the further design of interim analyses or for redesign of the trial.

In both the combination test method and the conditional error function method, the rejection region for the final decision rule is invariant with respect to mid-trial design adaptations conditional on the results of the interim analysis. This fundamental principle of flexible designs is called the *conditional invariance principle* by Brannath et al. (2007). The two approaches are basically two different ways of specifying a level α rejection region in the two-dimensional (p_1, p_2)-plane (Posch and Bauer, 1999; Schäfer et al., 2006). For all combination test procedures presented above, equivalent representations in term of conditional error functions are possible. For example, the choice

$$A(p_1) = \begin{cases} 1 & \text{if } p_1 \leq \alpha_1 \\ c_\alpha / p_1 & \text{if } \alpha_1 < p_1 < \alpha_0 \\ 0 & \text{if } p_1 \geq \alpha_0 \end{cases}$$

defines the Bauer and Köhne combination test in terms of a conditional error function. For illustration, Figure 2.3 displays the corresponding level α rejection region in the (p_1, p_2)-

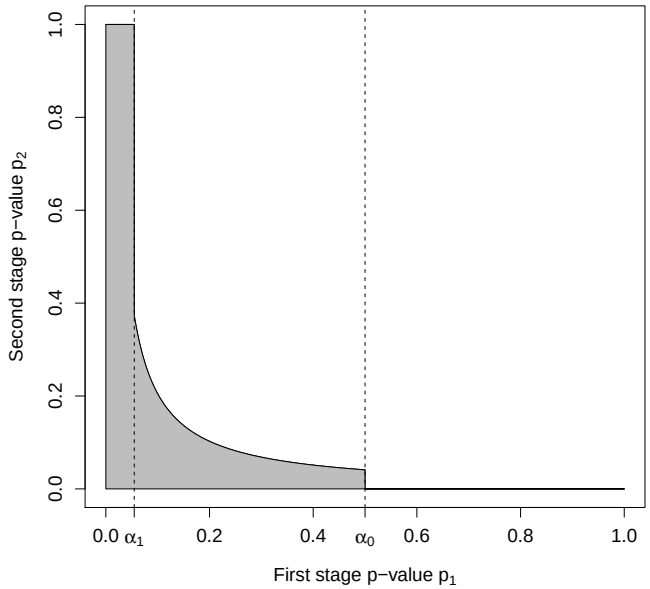


Figure 2.3.: *Rejection region of the Bauer and Köhne design*

plane as a gray area for the choices $\alpha = 0.1$ and $\alpha_0 = 0.5$ and therefore $c_\alpha = 0.0205$ and $\alpha_1 = 0.0548$ (see Table 2.3).

A formal proof that the two approaches are essentially equivalent was provided by Vandemeulebroecke (2006). Therefore, all general findings and advances in flexible designs apply to both approaches. The methodology for flexible designs has been considerably extended in recent years and now covers, for instance, implementation of further interim analyses (Brannath et al., 2002), adjustment for covariates (Ayanlowo and Redden, 2008), calculation of the overall study p-value (Brannath et al., 2002), estimation of the treatment effect and calculation of confidence intervals (Coburger and Wassmer, 2001; Posch et al., 2005; Brannath et al., 2006), and multiple test procedures (Bauer and Kieser, 1999; Kieser et al., 1999; Brannath et al., 2007). However, all these methods are tailored to comparative studies with continuous outcomes. In the next chapter, we will investigate the characteristics of flexible design methods applied to discrete test statistics.

You can't fix by analysis what you bungled by design.

(Light, Singer and Willett 1990*)*

3

Drawbacks with Adaptive Designs Applied to Discrete Test Statistics

Brannath et al. (2002) showed that the p-values in flexible designs do not necessarily need to be uniformly distributed. A sufficient requirement for type I error rate control with the combination test approach is the p clud condition.

Definition 3.1. (p clud). The distribution of p values p_t $(t = 1, 2)$ is called p *clud*, if they satisfy

$$\Pr_{H_0}(p_1 \leq \alpha) \leq \alpha \text{ and } \Pr_{H_0}(p_2 \leq \alpha | p_1) \leq \alpha \text{ for all } 0 \leq \alpha \leq 1.$$

Remark. The property p *clud* means that the distribution of p_1 and the conditional distribution of p_2 given p_1 are stochastically at least as large as the uniform distribution on $[0, 1]$.

Let us assume, as in Section 2.1, that we are testing, in an oncological phase II design, a binary endpoint with the two outcomes *success* and *failure*. The corresponding null hypothesis is $H_0 : \pi = \pi_0$, which should be tested at level α using a design with two stages. Further, as in Section 2.2 on adaptive and flexible designs, let p_1 and p_2 denote the stage-wise p-values, such that p_1 is based on the n_1 observations of the first stage and p_2 on the n_2 observations of the second stage. At each stage a binomial test is performed and the p-values of the two stages with k and l successes in the first and second stage, respectively, are therefore given by

$$p_1(k) = \Pr_{H_0}(X_1 \geq k) = 1 - B(k - 1; \pi_0, n_1) \tag{3.1}$$

and

$$p_2(l) = \Pr_{H_0}(X_2 \geq l) = 1 - B(l-1; \pi_0, n_2), \tag{3.2}$$

where X_i denotes the random variable of the number of successes in stage $i, i = 1, 2$, and B the cumulative binomial distribution. Note that the random variables X_1 and X_2 are binomially distributed, and therefore the distributions of P_1 and P_2 are stochastically larger than the uniform distribution on $[0, 1]$. Consequently, the distribution of p_1 and p_2 satisfy the p clud condition given in Definition 3.1. Therefore, the flexible designs methodology presented in the preceding chapter can directly be applied to binary endpoints. In the following sections we will see, however, that the resulting designs will generally be conservative and that apparently self-evident solutions will lead to inflation of the type I error rate.

3.1. Combination test method

As presented in Section 2.2.2, combining stochastically independent uniformly distributed p-values of the two stages is a common approach to construct flexible designs for continuous test statistics. Bauer and Köhne (1994) used Fisher's combination criterion where H_0 is rejected after the second stage if the product of the p-values from the two stages is smaller than a boundary c_α. For continuous test statistics and uniformly distributed p-values, the overall level α is then maintained regardless of the adaptations made. Due to the fact that for binary endpoints the distributions of P_1 and P_2 are stochastically larger than the uniform distribution, the designs and decision boundaries developed for uniformly distributed p-values can be applied and the level will be maintained. However, the level α will not be exhausted as for continuous test statistics, associated with a loss in power or an increased sample size. Without stopping for futility, the actual type I error rate of the Bauer and Köhne (1994) two-stage design for binary endpoints and sample sizes for both stages fixed is given by

$$\alpha' = \Pr_{H_0}(P_1 P_2 \leq c_\alpha) = \sum_{k,l \text{ with } p_1(k)p_2(l) \leq c_\alpha} b(k; \pi_0, n_1) b(l; \pi_0, n_2),$$

where b denotes the probability mass function of the binomial distribution.

As an example we consider a two-stage trial with $n_1 = 19$, $n_2 = 26$, $\pi_0 = 0.3$ and $\alpha = 0.05$ with corresponding $c_\alpha = 0.0087$ (see Table 2.3). This results in $\alpha' = 0.0282$, which is considerably below the nominal level α. Applying boundaries for early stopping for futility or efficacy $\alpha_0 < 1$ and $\alpha_1 > c_\alpha$, respectively, as introduced in Section 2.2.2, may result in an even more extreme undershooting of the desired level and consequently in a

further loss of power. For example, the use of $\alpha_0 = 0.5$ with a resulting $\alpha_1 = 0.0233$ leads to an actual level of

$$\begin{aligned}\alpha' &= \Pr_{H_0}(P_1 \leq \alpha_1) + \Pr_{H_0}(P_1 P_2 \leq c_\alpha, \alpha_1 < P_1 < \alpha_0) \\ &= \sum_{\substack{k,l \text{ with } p_1(k) \leq \alpha_1 \vee \\ \{p_1(k)p_2(l) \leq c_\alpha \wedge \alpha_1 < p_1(k) < \alpha_0\}}} b(k; \pi_0, n_1) b(l; \pi_0, n_2) \\ &= 0.0251\end{aligned}$$

and thus to a spending of only about half of the nominal level. With increasing n_1 and n_2 the approximation of the distribution of the discrete p-values by the uniform distribution will improve, and hence the actual level will increase towards α. However, the approximation remains poor for sample sizes typically used in single-arm phase II trials in oncology. For example, even for the rather extreme scenario of $n_1 = 50$ and $n_2 = 100$ the actual level still amounts to only $\alpha' = 0.0317$ for the above mentioned two-stage study ($\pi_0 = 0.3$, $\alpha = 0.05$, $\alpha_0 = 0.5$, $\alpha_1 = 0.0233$).

On first sight, a straightforward solution of the above described conservativeness may be obtained by enlarging α_1 or c_α and thus achieving a better exhaustion of the level. However, this results in a dependence of c_α and α_1 from the sample size of the second stage which can lead to conflicting decisions and counterintuitive events if the sample size is changed after the interim analysis. For example, with α_1 depending on n_2 it is possible to change the sample size after the interim analysis such that an early stopping with rejection of H_0 is possible although the trial would have been continued according to the initially specified decision rules. If the second-stage sample size is changed from n_2 to n_2^* this occurs for $\alpha_1(n_2) < p_1 \leq \alpha_1(n_2^*)$.

The dependence of the decision rule for the second stage on n_2 is also problematic. Consider again a two-stage design with $n_1 = 19$, $n_2 = 26$, $\pi_0 = 0.3$ and $\alpha = 0.05$. With $\alpha_0 = 0.5$ and $\alpha_1 = 0.0233$ fixed, an increase of c_α to $c'_\alpha = 0.0202$ will maintain the level with

$$\alpha' = \Pr_{H_0}(P_1 \leq \alpha_1) + \Pr_{H_0}(P_1 P_2 \leq c'_\alpha, \alpha_1 < P_1 < \alpha_0) = 0.0493.$$

Likewise, $c^*_\alpha = 0.0233$ is possible for the choices of $n_1 = 19$ and $n_2 = 19$. Here we have

$$\alpha' = \Pr_{H_0}(P_1 \leq \alpha_1) + \Pr_{H_0}(P_1 P_2 \leq c^*_\alpha, \alpha_1 < P_1 < \alpha_0) = 0.0489.$$

Even though both designs separately control the nominal level for fixed sample sizes, an adaptive design combined of both will not share this property. Consider the following adaptive scenario: With $p_1 = 0.033$ resulting from $k = 10$ successes out of $n_1 = 19$ observations, the sample size of the second stage is reduced from $n_2 = 26$ to $n_2^* = 19$

and the corresponding boundary is used; otherwise the originally planned design and corresponding decision rule is used. The resulting adaptive test has a type I error rate of

$$\begin{aligned}\alpha' &= \Pr_{H_0}(P_1 \leq \alpha_1) + \Pr(P_1 P_2 \leq c'_\alpha, \alpha_1 < P_1 < \alpha_0, P_1 \neq 0.033) \\ &\quad + \Pr(P_1 P_2^* \leq c^*_\alpha, P_1 = 0.033) \\ &= 0.053,\end{aligned}$$

where P_2^* denotes the random p-value of the modified second stage. Surprisingly, this design does not control the level any more.

From a more general viewpoint, every dependence of the decision rules from the second-stage data will violate a common principle of adaptive designs which is called the conditional invariance principle (Brannath et al., 2007). It states that the final decision rule must be invariant with respect to mid-trial design changes conditional on the results from the interim analysis.

3.2. Conditional error function method

As noted in Section 2.2.2, for each combination test procedure there exists an equivalent representation in terms of the conditional error function. Therefore, a direct application of (continuous) conditional error functions will result in the same findings as in the preceding section. However, it is possible to adapt the conditional error function to the setting of discrete test statistics.

In analogy to (2.6), we define for two-stage designs with discrete outcomes the discrete conditional error function D.

Definition 3.2. (Discrete conditional error function). A *discrete conditional error function* is defined as a function $D(p) : [0,1] \to [0,1]$ with support $\boldsymbol{P}_1$ and

$$\sum_{p \in \boldsymbol{P}_1} D(p) \cdot \Pr_{H_0}(P_1 = p) \leq \alpha, \tag{3.3}$$

where $\boldsymbol{P}_1$ denotes the finite set of possible outcomes p_1 of the random variable P_1.

We further require that D is non-increasing on its support $\boldsymbol{P}_1$. As for the continuous counterpart, this is a logical restriction and ensures that smaller p-values observed in stage one are associated with higher conditional error levels for the second stage. Note that the discrete conditional error function only depends on the design characteristics of the first stage.

As in the continuous case, the null hypothesis is rejected if the second-stage p-value p_2 satisfies $p_2 \leq D(p_1)$. With $D(p_1) = 0$ or $D(p_1) = 1$, respectively, early stopping after the first stage with acceptance or rejection of the null hypothesis is possible. If the first- and second-stage p-values are p clud, (3.3) guarantees that the overall type I error rate is controlled even if the design characteristics of the second stage have been modified based on any information available at the end of the first stage.

We now show that classical phase II designs presented in Section 2.1 can alternatively be formulated in terms of our novel concept of discrete conditional error functions. As the first- and second-stage p-values are p clud, flexible design changes are then possible. Therefore, this is the first fundamental step to achieve the goal of flexible phase II oncology trials. Later, in Chapters 4 and 5, we will refine and extend the concept of discrete conditional error functions leading to designs that are both flexible and more efficient as compared to classical phase II designs.

Theorem 3.3. *In phase II designs with binary endpoint and p-values given by* (3.1) *and* (3.2)*, the type I error rate α' of such a design can alternatively to* (2.2) *be calculated by*

$$\alpha' = \sum_{k=0}^{n_1} CE(k) \cdot Pr_{H_0}\big\{P_1 = p_1(k)\big\}, \tag{3.4}$$

where

$$CE(k) = \begin{cases} 0 & \text{if } k \leq l_1 \\ 1 - B\left(l_2 - k, \pi_0, n_2\right) & \text{if } l_1 < k < u_1 \\ 1 & \text{if } k \geq u_1 \end{cases} \tag{3.5}$$

defines the conditional type I error rate when k responses are observed in the first stage.

Proof. We have

$$\begin{aligned} \alpha' &\overset{(2.2)}{=} 1 - \left[B(l_1; \pi_0, n_1) + \sum_{k=l_1+1}^{\min(n_1, u_1-1)} b(k; \pi_0, n_1) \cdot B(l_2 - k; \pi_0, n_2)\right] \\ &= \sum_{k=0}^{n_1} b(k; \pi_0, n_1) - \sum_{k=0}^{l_1} b(k; \pi_0, n_1) - \sum_{k=l_1+1}^{\min(n_1, u_1-1)} b(k; \pi_0, n_1) \cdot B(l_2 - k; \pi_0, n_2) \\ &= \sum_{k=l_1+1}^{u_1-1} \left[1 - B\left(l_2 - k, \pi_0, n_2\right)\right] \cdot b(k; \pi_0, n_1) + \sum_{k=u_1}^{n_1} b(k; \pi_0, n_1) \\ &= \sum_{k=0}^{n_1} \mathrm{CE}(k) \cdot \mathrm{Pr}_{H_0}\big\{P_1 = p_1(k)\big\}. \end{aligned}$$

□

With (3.3) and (3.4), for any two-stage design with actual level $\alpha' \leq \alpha$ a discrete conditional error function can be defined by $D(p_1(k)) = \text{CE}(k)$, $k \in \{0, \ldots, n_1\}$. Note that this corresponds to the application of the method proposed by Müller and Schäfer (2001, 2004) to the current setting of phase II trials in oncology. By using this 'natural' conditional error function, arbitrary design modifications after the first stage can be performed while still controlling the type I error rate by α. If the design is not changed after the interim analysis, the applied decision rules and with it the performance characteristics are identical to those of the original design.

Application of the natural conditional error function makes classical phase II trials flexible but results in conservative procedures. This conservativeness lies in the fact that the original two-stage designs are *per se* conservative due to the discreteness of the used test statistics. For example, Simon's design minimizing the total sample size for testing $H_0 : \pi \leq 0.1$ against the alternative $H_1 : \pi = 0.3$ with $\alpha = 0.05$ and $\beta = 0.1$ is defined by $(l_1, n_1, l_2, n_2) = (2, 22, 6, 11)$ and $u_1 > n_1$ (see Table 2.2). The actual type I error rate of this design amounts to $\alpha' = 0.0409$. Choosing $D(p_1(k)) = \text{CE}(k)$, $k \in \{0, \ldots, n_1\}$, the discrete conditional error function exactly matches the decision rules of the classical design if no design changes are performed and therefore the level then also equals $\alpha' < \alpha$.

The three most important areas where statisticians can contribute, and be influential are design, design and design.

(Andy Grieve 2002*)*

4

Flexible Design Methods for Discrete Test Statistics

Our aim is to derive two-stage designs for testing the one-sided null hypothesis $H_0 : \pi \leq \pi_0$ at a prefixed level α and with power $1 - \beta$ at $H_1 : \pi = \pi_1$, $\pi_1 > \pi_0$, where the sample size for the second stage can be adapted after the interim analysis in a flexible way under control of the significance level. For this, we first propose in Section 4.1 a new fixed two-stage combination test design that is inspired by the Bauer and Köhne (1994) design. The suggested method combines the independent p-values of the two stages under a nonadaptive framework, i.e., n_1 and n_2 fixed, such that the predefined level α is exhausted to a greater extent than described in the examples above. By calculating the power for a specified alternative, feasible designs that satisfy the α and β constraints can be selected. We then show in Section 4.2 that the proposed fixed two-stage design can be directly transferred into a flexible design by applying adaptive conditional tests. Selecting a combination test based design as initial start design, the sample size of the second stage can be changed mid-course while at the same time controlling the type I error rate. If no adaptations are performed, this design results in identical decision rules and operational characteristics as the corresponding original design. With no adaptations, the conditional adaptive test is also identical to the Müller and Schäfer method, which lacks of the same conservativeness as the original design when the latter does not exhaust the nominal type I error rate. We show below how this deficiency can be resolved within our flexible two-stage design based on the combination test approach.

An alternative way to introduce flexibility to phase II designs in oncology is by application of the conditional error function approach, which allows any change of the design as long

as the conditional type I error rate of the modified design is kept equal or lower to the corresponding conditional error function value. We present in Section 4.3 how flexible two-stage designs can be constructed based on the conditional error function method. This will lead to a refinement of the concept of discrete conditional error functions presented in Section 3.2. We will see that application of the conditional error function approach is similar to the adaptive conditional test. Both are, however, methodologically different, as the conditional error function starts with an arbitrary fixed function, whereas the second enables adaptations within fixed designs.

Finally, in Section 4.4 we show how combination of both ideas leads to flexible designs for phase II oncology trials that are more efficient than those designs presented in the literature so far.

4.1. Fixed two-stage design based on combination test approach

In the following, a fixed two-stage design is presented that uses the same type of decision rules as the Bauer and Köhne (1994) design, i.e., the trial is stopped early for futility if the first-stage p-value p_1 (see (3.1) on page 19) satisfies $p_1 \geq \alpha_0$ or for efficacy if $p_1 \leq \alpha_1$. Otherwise, the trial is continued and the null hypothesis is rejected in the final stage if the product of the p-values of both stages, p_1 and p_2 (see (3.2) on page 20), falls below a boundary c_α. To derive such a design, we need to specify in the planning stage a lower bound α_0 for early acceptance of H_0 and an upper bound $\alpha_1 < \alpha$ for the amount of type I error rate that should be spent for the interim analysis.

At first we calculate for given values of α_0, α_1, n_1 and n_2 the decision boundary c_α such that the level α is maintained. Due to the discreteness of the test statistics the type I error rate spent in the first stage $\gamma_1 = \Pr_{H_0}(P_1 \leq \alpha_1)$ will usually not fully exhaust α_1. The actual level γ_1 of the first stage can be calculated as

$$\gamma_1 = \Pr_{H_0}(P_1 \leq \alpha_1) = \sum_{k \text{ with } p_1(k) \leq \alpha_1} b(k; \pi_0, n_1) \leq \alpha_1.$$

The remaining level $\gamma_2 = \alpha - \gamma_1 \geq \alpha - \alpha_1$ can now be spent in the second stage to maintain the overall level α. Note that γ_1 and γ_2 do not depend on n_2.

In the final analysis, Fisher's combination criterion is used to combine the independent p-values to the final test statistic. For $\alpha_1 < p_1 < \alpha_0$ the null hypothesis is rejected after the second stage if $p_1 p_2 \leq c_\alpha$. If we define c_α by

$$c_\alpha = \max\left\{x \in \boldsymbol{P}_{12} \middle| \sum_{k,l \text{ with } p_1(k)p_2(l) \leq x \,\wedge\, \alpha_1 < p_1(k) < \alpha_0} b(k; \pi_0, n_1) b(l; \pi_0, n_2) \leq \gamma_2\right\},$$

where $\boldsymbol{P}_{12}$ denotes the finite range of P_1P_2, we assure

$$\Pr_{H_0}(P_1P_2 \leq c_\alpha, \alpha_1 < P_1 < \alpha_0) \leq \gamma_2$$

and thus control of the type I error rate by α. It should be mentioned that the dependence of c_α on the choice of n_2 is new in our approach, since in the original design proposed by Bauer and Köhne for uniformly distributed p-values c_α depends on α only. Note that this dependence allows a better exhaustion of the nominal level, but is also the reason why the conditional invariance principle is violated and further refinement of the procedure is required to assure control of the type I error rate in an adaptive framework (see Section 3.1).

The actual level of the procedure can be calculated as

$$\begin{aligned}
&\alpha'_{\text{combination test}} \\
&= \Pr_{H_0}(P_1 \leq \alpha_1) + \Pr_{H_0}(P_1P_2 \leq c_\alpha, \alpha_1 < P_1 < \alpha_0) \\
&= \sum_{\substack{k,l \text{ with } p_1(k)\leq\alpha_1 \vee \\ \{p_1(k)p_2(l)\leq c_\alpha \wedge \alpha_1<p_1(k)<\alpha_0\}}} b(k;\pi_0,n_1)b(l;\pi_0,n_2).
\end{aligned} \tag{4.1}$$

Due to the discreteness, the nominal level α will usually not be fully exhausted, but the extent of conservativeness is small. For the example considered in Section 3.1, $n_1 = 19$, $n_2 = 26$, $\pi_0 = 0.3$ and $\alpha = 0.05$ with $\alpha_0 = 0.5$ and $\alpha_1 = 0.02$, the resulting level is $\alpha' = 0.0493$. This is very close to the nominal level and considerably higher than $\alpha' = 0.0251$, which resulted from direct application of the Bauer and Köhne method developed for continuous test statistics.

The power of the procedure for the alternative $H_1 : \pi = \pi_1$ can be calculated as

$$\begin{aligned}
&1 - \beta'_{\text{combination test}} \\
&= \Pr_{H_1}(P_1 \leq \alpha_1) + \Pr_{H_1}(P_1P_2 \leq c_\alpha, \alpha_1 < P_1 < \alpha_0) \\
&= \sum_{\substack{k,l \text{ with } p_1(k)\leq\alpha_1 \vee \\ \{p_1(k)p_2(l)\leq c_\alpha \wedge \alpha_1<p_1(k)<\alpha_0\}}} b(k;\pi_1,n_1)b(l;\pi_1,n_2).
\end{aligned}$$

In order to determine designs that achieve the desired power $1 - \beta$, we need to calculate the decision boundary c_α as described above for fixed α_0 and α_1 and a wide range of values for n_1 and n_2. Among this set of designs, we select those with an actual power of at least $1 - \beta$. By construction, these feasible designs satisfy the α and β constraints, and among them designs with specific characteristics can be selected as presented in Section 2.1. The average sample number under the null and alternative hypothesis, $\text{EN}(\pi_0)$ and $\text{EN}(\pi_1)$,

respectively, can be calculated via

$$\mathrm{EN}(\pi_i) = \big(1 - \mathrm{Pr}_{H_i}(\alpha_1 < P_1 < \alpha_0)\big) \cdot n_1 + \mathrm{Pr}_{H_i}(\alpha_1 < P_1 < \alpha_0) \cdot (n_1 + n_2) \text{ for } i = 0, 1. \tag{4.2}$$

We now demonstrate that the combination test based designs have very similar properties to classical phase II designs as given in Section 2.1. To allow for a fair comparison the design by Chang et al. (1987) was selected, as it also allows for early stopping for futility. Chang et al. selected the parameter combinations $(n_1, n_2, l_1, u_1, l_2)$ with $l_1 < u_1$ by a computer search algorithm to control the type I and II error rates and to minimize $\mathrm{EN}(\pi_{01}) := \frac{1}{2}\big(\mathrm{EN}(\pi_0) + \mathrm{EN}(\pi_1)\big)$. In the following, the optimal designs of Chang et al. are compared with the corresponding optimal fixed two-stage designs based on the combination test that can be found by selecting the feasible design minimizing $\mathrm{EN}(\pi_{01})$. Chang et al. restricted the sample sizes of the stages to multiples of five. As such a restriction seems not to be necessary in clinical studies we calculated the corresponding oncological phase II designs of Chang et al. without this restriction according to the algorithm given in the original article. For the proposed method we determined for given $(\pi_0, \pi_1, \alpha, \beta)$ the decision boundary c_α for a wide range of the design parameters α_0 (0.2 to 0.9 by 0.1), α_1 (0.010 to 0.045 by 0.005), n_1 (1 to 50) and n_2 (1 to 50). Among all feasible designs, the optimal design minimizing $\mathrm{EN}(\pi_{01})$ was selected.

The results are given in Table 4.1. For the design of Chang et al. the early futility boundary l_1 and early efficacy boundary u_1 are given together with the final futility boundary l_2. As the final decision within the combination test design does not depend on the absolute number of events but on the product of the p-values, the critical boundary c_α is given instead. For ease of comparison, l_1 and u_1 were calculated for the proposed method by selecting the greatest or smallest number of responses in the first stage with p-value greater than α_0 or smaller than α_1, respectively.

The proposed method shows very similar characteristics as compared to the optimal design by Chang et al. In the majority of considered scenarios, the critical boundaries for the interim analysis are the same for both designs and a closer look on the decision rules after stage two showed that these are identical, too, for these cases. However, there are also situations where the proposed design shows a slightly smaller $\mathrm{EN}(\pi_{01})$ than the design of Chang et al. that was optimized with respect to this criterion. In the combination test design, the product of the p-values takes the interim results into account and thus allows for a better exhaustion of the level. This corresponds to a higher power or a lower average sample size to achieve a given power.

It should be mentioned that the proposed method provides a huge number of feasible designs and that the designs minimizing $\mathrm{EN}(\pi_{01})$ given in Table 4.1 are just special ones.

Table 4.1.: *Comparison of the proposed combination test design with the optimal design of Chang et al. (1987)*

Design parameters					Boundaries						Size	Power	Average sample number		
π_0	π_1	α	β	Design	n_1	n_2	l_1	u_1	l_2	c_α	α'	$1-\beta'$	EN(π_0)	EN(π_1)	EN(π_{01})
0.1	0.3	0.05	0.2	Chang	11	16	1	4	5		0.048	0.812	15.5	18.3	16.9
				Combination	11	16	1	4		0.021	0.048	0.812	15.5	18.3	16.9
			0.1	Chang	17	24	2	5	7		0.048	0.901	22.2	24.5	23.3
				Combination	17	24	2	5		0.020	0.048	0.901	22.2	24.5	23.3
0.2	0.4	0.05	0.2	Chang	14	23	3	7	11		0.047	0.802	20.7	27.1	23.9
				Combination	17	26	4	7		0.014	0.050	0.801	22.3	25.4	23.8
			0.1	Chang	23	31	5	9	16		0.045	0.901	31.6	33.4	32.5
				Combination	20	35	4	8		0.013	0.050	0.901	31.8	32.8	32.3
0.3	0.5	0.05	0.2	Chang	20	22	7	11	17		0.050	0.802	24.6	30.0	27.3
				Combination	20	22	7	11		0.021	0.050	0.802	24.6	30.0	27.3
			0.1	Chang	29	34	10	14	25		0.049	0.901	35.8	38.8	37.3
				Combination	29	34	10	14		0.014	0.049	0.901	35.8	38.8	37.3
0.4	0.6	0.05	0.2	Chang	20	27	9	13	24		0.049	0.807	26.0	32.3	29.2
				Combination	20	27	9	13		0.018	0.049	0.807	26.0	32.3	29.2
			0.1	Chang	26	36	11	16	31		0.050	0.905	37.0	41.4	39.2
				Combination	26	36	11	16		0.016	0.050	0.905	37.0	41.4	39.2
0.5	0.7	0.05	0.2	Chang	17	28	9	13	29		0.048	0.802	25.1	31.2	28.2
				Combination	20	22	11	15		0.019	0.049	0.801	25.1	30.3	27.7
			0.1	Chang	22	37	11	16	37		0.049	0.900	36.4	39.3	37.8
				Combination	27	30	14	19		0.017	0.050	0.900	36.7	38.6	37.7

For each value of $(\pi_0, \pi_1, \alpha, \beta)$ the first row corresponds to the design of Chang et al. and the second row correspond to the fixed two-stage design based on combination test. The designs were chosen to minimize EN(π_{01}), the mean of the average sample number under the null and alternative hypothesis denoted by EN(π_0) and EN(π_1), respectively.

The choice of the design can also be tailored to any other specifications made by the study team, for example, with respect to a minimization of the total sample size.

4.2. Flexible two-stage design based on combination test approach

We now assume that for given $(\pi_0, \pi_1, \alpha, \beta)$ a fixed two-stage design based on the combination test approach is selected according to Section 4.1. In the following, we derive an adaptive conditional test that allows within this design a flexible mid-course modification of the sample size of the second stage under control of the type I error rate.

A general justification of adaptive conditional tests can be found in Liu et al. (2002). The idea is as follows. The sample size n_1 of the first stage is fixed in the planning stage. The observed p-value at the first stage, P_1, is random, but the distribution of P_1 under the null hypothesis is known in the planning stage. Conditional on a realization of P_1, say p_1, a conditional type I error rate is used in the second stage. Let $C(p_1)$ be the function defining this conditional type I error rate depending on the p-value p_1 obtained in the first stage. If the p-value of the second stage is independent from the first-stage results and stochastically at least as large as the uniform distribution on $[0, 1]$, then the null hypothesis is rejected after the second stage if $p_2 \leq C(p_1)$. Obviously, for $C(p_1) = 0$ or $C(p_1) = 1$ early acceptance or rejection of H_0 after the first stage, respectively, occurs. The function C defining the conditional type I error rates is calculated by the experimenter before the start of the trial based on a fixed design. Therefore, as the latter, it controls the unconditional type I error, i.e.,

$$\alpha'_{\text{flexible design}} = \sum_{p_1 \in \boldsymbol{P}_1} \Pr_{H_0}(P_1 = p_1) \cdot C(p_1) \leq \alpha, \tag{4.3}$$

where $\boldsymbol{P}_1$ denotes the finite range of P_1. This is conceptionally similar to the approach by Müller and Schäfer (2001, 2004).

For the fixed two-stage design based on the combination test developed in the previous section, the null hypothesis is rejected in the final stage whenever $p_2 \leq \frac{c_\alpha}{p_1}$ with $\alpha_1 < p_1 < \alpha_0$. Therefore, $C(p_1) = \Pr_{H_0}\big(P_2 \leq c_\alpha/p_1 \;\big|\; p_1\big)$ for $\alpha_1 < p_1 < \alpha_0$, $p_1 \in \boldsymbol{P}_1$, is a natural choice for mapping the decision rule of the fixed design to a function C. For a p-value smaller than or equal to α_1, the fixed two-stage design is stopped early for efficacy, and for a p-value of at least α_0 it is stopped for futility, i.e., $C(p_1) = 1$ for $p_1 \leq \alpha_1$ and $C(p_1) = 0$ for $p_1 \geq \alpha_0$. With this choice of C, the adaptive conditional test has the same type I error rate as the fixed combination test design if the sample size is not changed. This follows

from equations (4.1) and (4.3):

$$
\begin{aligned}
&\alpha'_{\text{flexible design}} \\
&= \sum_{p_1 \in \boldsymbol{P}_1} \Pr_{H_0}(P_1 = p_1) \cdot C(p_1) \\
&= \sum_{\substack{p_1 \leq \alpha_1 \\ p_1 \in \boldsymbol{P}_1}} \Pr_{H_0}(P_1 = p_1) \cdot C(p_1) + \sum_{\substack{\alpha_1 < p_1 < \alpha_0 \\ p_1 \in \boldsymbol{P}_1}} \Pr_{H_0}(P_1 = p_1) \cdot C(p_1) \\
&\quad + \sum_{\substack{p_1 \geq \alpha_0 \\ p_1 \in \boldsymbol{P}_1}} \Pr_{H_0}(P_1 = p_1) \cdot C(p_1) \\
&= \sum_{\substack{p_1 \leq \alpha_1 \\ p_1 \in \boldsymbol{P}_1}} \Pr_{H_0}(P_1 = p_1) + \sum_{\substack{\alpha_1 < p_1 < \alpha_0 \\ p_1 \in \boldsymbol{P}_1}} \Pr_{H_0}(P_1 = p_1) \cdot \Pr_{H_0}(P_2 \leq c_\alpha / p_1 | p_1) \\
&= \Pr_{H_0}(P_1 \leq \alpha_1) + \Pr_{H_0}(P_1 P_2 \leq c_\alpha, \alpha_1 < P_1 < \alpha_0) = \alpha'_{\text{combination test}}.
\end{aligned}
$$

By construction, the type I error rate of the flexible design is also controlled if the sample size of the second stage is changed (cf. (4.3)). However, the conditional type I error rate of stage two $C(p_1)$ will then usually not be fully exhausted due to the discreteness of the distribution. Thus, the maximum achievable level in a flexible setting is $\alpha'_{\text{combination test}}$. If for the chosen start design $\alpha - \alpha'_{\text{combination test}} > 0$, the remaining portion of the level can be implemented in the flexible design by increasing the function C for specific values of p_1 and assuring at the same time control of the type I error rate. There exists a variety of different ways how this can be achieved. In the following, the values of C were increased equally for all p_1 with $\mathrm{C}(p_1) \neq 0$ and $\mathrm{C}(p_1) \neq 1$ such that equation (4.3) equals to the nominal level.

Table 4.2 tabulates the function C, i.e., the adaptive conditional test boundaries of the second stage, for the fixed two-stage designs presented in Table 4.1. For each considered combination of $(\pi_0, \pi_1, \alpha, \beta)$, the critical values for the second-stage p-value $C(p_1)$ are given for each possible realization of p_1. To illustrate how to read Table 4.2, consider, for example, a clinical trial planned for $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.20)$. According to Table 4.1, $n_1 = 11$ patients need to be enrolled in the first stage. With at least four responses, i.e., $p_1 \leq \alpha_1 = 0.020$, the trial is stopped after the interim analysis with the rejection of H_0 ($C(p_1) = 1$). If two or three responses are observed in the first stage leading to $p_1 = 0.3026$ or $p_1 = 0.0896$, the second stage is planned with a critical value of $C(0.3026) = 0.073$ or $C(0.0896) = 0.224$, respectively. With $p_1 \geq \alpha_0 = 0.400$, i.e., less than two responses observed in the first stage, the trial is stopped for futility ($C(p_1) = 0$).

It should be noted that while the decision rule of the fixed two-stage design depends on the sample size of the second stage n_2 through the choice of c_α, this does not hold true

Table 4.2.: *Adaptive conditional test boundaries calculated for the fixed combination test design corresponding to the optimal design of Chang et al. (1987)*

π_0	π_1	α	β		Adaptive conditional test boundaries					
0.1	0.3	0.05	0.2	p_1	$\leq$0.020	0.0896	0.3026			$\geq$0.400
				$C(p_1)$	1	0.224	0.073			0
			0.1	p_1	$\leq$0.025	0.0826	0.2382			$\geq$0.300
				$C(p_1)$	1	0.228	0.090			0
0.2	0.4	0.05	0.2	p_1	$\leq$0.040	0.1057	0.2418			$\geq$0.300
				$C(p_1)$	1	0.133	0.024			0
			0.1	p_1	$\leq$0.035	0.0867	0.1958	0.3704		$\geq$0.400
				$C(p_1)$	1	0.147	0.035	0.035		0
0.3	0.5	0.05	0.2	p_1	$\leq$0.020	0.0480	0.1133	0.2277		$\geq$0.300
				$C(p_1)$	1	0.329	0.187	0.092		0
			0.1	p_1	$\leq$0.030	0.0652	0.1294	0.2292		$\geq$0.300
				$C(p_1)$	1	0.202	0.116	0.061		0
0.4	0.6	0.05	0.2	p_1	$\leq$0.025	0.0565	0.1275	0.2447		$\geq$0.300
				$C(p_1)$	1	0.261	0.150	0.078		0
			0.10	p_1	$\leq$0.025	0.0518	0.1082	0.1993	0.3263	$\geq$0.400
				$C(p_1)$	1	0.236	0.146	0.083	0.043	0
0.5	0.7	0.05	0.2	p_1	$\leq$0.025	0.0577	0.1316	0.2517		$\geq$0.300
				$C(p_1)$	1	0.271	0.148	0.070		0
			0.10	p_1	$\leq$0.030	0.0610	0.1239	0.2210	0.3506	$\geq$0.400
				$C(p_1)$	1	0.181	0.101	0.050	0.050	0

for the adaptive conditional test boundaries. The function C and with it the conditional type I error rate used in the second stage are determined before the start of the trial and are therefore independent of changes made throughout the conduct of the trial. For this reason, the conditional invariance principle mentioned in Section 3.1 is adhered to.

The flexible two-stage design based on combination test is based on the fixed two-stage design and, by construction, the type I error rate and power are identical if the initially planned n_2 is not changed. Therefore, it is possible to plan a study quite similar to the design of Chang et al. by selecting an appropriate fixed two-stage design as shown in Section 4.1 and using it as *start design*. Even though this selected design may be modified throughout the study, it is a key element of this proposed method because it mainly influences the extent of exhaustion of the type I error rate and with it the achieved power in case that the sample size is not changed. Consequently, when applying the flexible two-stage design based on combination tests the choice of the start design is essential and should therefore be based on some optimality criterion.

4.3. Flexible two-stage design based on conditional error functions

As a second approach we want to investigate the potential of conditional error functions as a method for construction of flexible two-stage phase II design under control of the type I error rate. We have seen in Section 3.2 that a discrete conditional error function, which was constructed in analogy to the continuous counterpart, allows for arbitrary design modifications in phase II designs while still controlling the nominal significance level. Thus it is directly possible to allow flexibility in all standard phase II designs presented in Section 2.1. However, we also noted that this method suffers from conservativeness for every two-stage design with $\alpha' < \alpha$, which is true for most phase II designs (see Section 3.2 or Table 2.1, 2.2, B.1 and B.2). Our aim is to reduce this conservativeness by refining the application of discrete conditional error function, possibly leading to more efficient flexible phase II designs.

A possible solution of this conservativeness is to use a discrete conditional error function for which equality holds true in (3.3). However, this would lead to an exhaustion of the level only if every conditional error rate was attainable in the second stage. In designs with binary endpoint, only discrete p-values can be achieved and, consequently, the actual conditional level of the second stage rarely reaches the planned level $D(p_1)$. Consider the situation that a study testing $H_0 : \pi = 0.1$ against the alternative $H_1 : \pi = 0.3$ with $\alpha = 0.05$ and $\beta = 0.1$ is planned with the discrete conditional error function

$$D(p_1) = \begin{cases} 0.1108 & \text{if } \boldsymbol{P}_1 \ni p_1 < 0.5 \\ 0 & \text{else.} \end{cases}$$

According to (3.3), the type I error rate of the resulting fixed design is

$$\alpha' = 0.1108 \cdot \Pr_{H_0}(P_1 < 0.5) = 0.05 = \alpha.$$

However, choosing $n_2 = 10$ as the sample size for stage two, only a conditional error rate of 0.0555 instead of the available conditional level of 0.1108 is attainable if $p_1 < 0.5$, and the actual level amounts to

$$\alpha'_{n_2=10} = 0.0555 \cdot \Pr_{H_0}(P_1 < 0.5) = 0.0251 < \alpha.$$

Therefore, only approximately half of the planned level is used for $n_2 = 10$. More insight in this general issue is given by the following theorem.

Theorem 4.1. *The conservativeness in application of discrete conditional error functions cannot be eliminated by constructing a discrete conditional error function that accounts for the discreteness of the second stage if at the same time the option for an arbitrary sample size adaptation is to be maintained.*

Proof. We prove for the application of discrete conditional error functions:
Option for an arbitrary sample size adaptation $\Leftrightarrow$ The second stage is planned for uniformly distributed p-values.

"$\Rightarrow$": A flexible design must adhere to the conditional invariance principle mentioned above, i.e., the rejection region must be invariant with respect to mid-trial design adaptations, especially with respect to a change in second-stage sample size. If n_2 can be modified completely freely, (nearly) every conditional level is attainable even though for each selected sample size only a finite number of levels can be reached. In case that the available conditional level is actually attained, the 'spent' conditional level is identical to the one in the continuous case. Therefore, the invariance requires that the second stage must be planned for uniformly distributed p-values.

"$\Leftarrow$": Even if changes are performed in the interim analysis, the second-stage p-value is stochastically at least as large as the uniform distribution. Therefore, a rejection region calculated for uniformly distributed p-values will maintain the level. Thus, to allow adaptive planning with discrete test statistics it is sufficient to plan the second stage for uniformly distributed p-values. □

According to Theorem 4.1, the discreteness of the second stage must be addressed in another way. For this, we use the natural conditional error function defined in Section 3.2 as a starting point. In case of no adaptations, the related design has the same level $\alpha' \leq \alpha$ as the original design as then the conditional error rates are exactly met. If the original design is conservative, i.e., $\alpha' < \alpha$, the remaining level $\alpha - \alpha'$ can be implemented to overcome the conservativeness in a flexible setting. This is done by increasing $D(p)$ such that equality results in (3.3). With this modified discrete conditional error function, the level is maintained even for uniformly distributed p-values and thus the requirement described above is fulfilled. The increase of the boundaries of the conditional error function results in a design that is at least as powerful as the original design if no adaptations are performed. However, situations exist where the null hypothesis cannot be rejected in the original design but in the new design. This occurs if the p-value of the second stage falls between the conditional levels of the original and the design with increased boundaries.

The boundaries of the conditional error function can be increased in a multitude of ways. This flexibility enables the study team to fine-tune the design for the desired adaptation strategy. If, for example, the sample size will only be changed if few responses are observed in the first stage, it is consequent to increase the discrete conditional error function just for the corresponding values of p_1.

4.3.1. Flexible version of Simon's two-stage design

Simon's designs (Simon, 1989, see Table 2.1, 2.2, B.1 and B.2) are the most popular phase II designs for oncology trials that allow stopping for futility only. We present in the following how the proposed discrete conditional error function approach can be applied to these designs thus allowing a completely free sample size modification after the interim analysis while controlling the overall type I error rate.

We construct the discrete conditional error function starting with the conditional type I error rates of the corresponding Simon design. As the principle for the construction is the same for all parameter constellations, we confine ourselves to demonstrate the procedure in detail for the situation $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$. Simon's minimax design under this setting is defined by $(l_1, n_1, l_2, n_2) = (2, 22, 6, 11)$ and $u_1 > n_1$. The conditional type I error rates $\mathrm{CE}(k)$ for a given number of k responses in the first stage (see (3.5) on page 23) are summarized in Table 4.3 together with the corresponding p-values $p_1(k)$. If $\mathrm{CE}(k)$ is used as discrete conditional error function, the resulting test has an actual level of

$$\alpha' = \sum_{k=0}^{n_1} \mathrm{CE}(k) \cdot \mathrm{Pr}_{H_0}\big\{P_1 = p_1(k)\big\} = 0.0409.$$

To overcome the conservativeness of the classical design, the boundaries $\mathrm{CE}(k)$ with $\mathrm{CE}(k) \neq 0$ and $\mathrm{CE}(k) \neq 1$ can be increased while assuring control of the nominal level. We considered three ways of increasing the boundaries until equality is reached in (3.3): (1) increase proportionally to the probability of observing p_1 ($D_1(p_1)$), (2) distribute the remaining level $\alpha - \alpha'$ equally among the conditional error function values ($D_2(p_1)$) and (3) increase of only the smallest conditional error function value unequal to zero ($D_3(p_1)$).

$$D_1\big(p_1(k)\big) = \begin{cases} \mathrm{CE}(k) & \text{if } \mathrm{CE}(k) = 0 \text{ or } \mathrm{CE}(k) = 1 \\ \mathrm{CE}(k) + \dfrac{(\alpha-\alpha')\frac{\mathrm{Pr}_{H_0}\{P_1=p_1(k)\}}{\sum \mathrm{Pr}_{H_0}\{P_1=p_1(j)\}}}{\mathrm{Pr}_{H_0}\{P_1=p_1(k)\}} & \text{else,} \end{cases} \tag{4.4}$$

where the sum is over all j with $\mathrm{CE}(j) \neq 0$ and $\mathrm{CE}(j) \neq 1$.

$$D_2\big(p_1(k)\big) = \begin{cases} \mathrm{CE}(k) & \text{if } \mathrm{CE}(k) = 0 \text{ or } \mathrm{CE}(k) = 1 \\ \mathrm{CE}(k) + \dfrac{\frac{\alpha-\alpha'}{\#\{\mathrm{CE}(k)\neq 0 \text{ and } \mathrm{CE}(k)\neq 1\}}}{\mathrm{Pr}_{H_0}\{P_1=p_1(k)\}} & \text{else,} \end{cases} \tag{4.5}$$

where # denotes the count function.

$$D_3\big(p_1(k)\big) = \begin{cases} \mathrm{CE}(k) & \text{if } \mathrm{CE}(k) \text{ is not the smallest conditional} \\ & \text{error function value unequal to zero} \\ \mathrm{CE}(k) + \dfrac{\alpha-\alpha'}{\mathrm{Pr}_{H_0}\{P_1=p_1(k)\}} & \text{else.} \end{cases} \tag{4.6}$$

The resulting discrete conditional error functions are also given in Table 4.3.

Table 4.3.: *Flexible version of Simon's minimax design for the parameter constellation* $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$

k	$\Pr_{H_0}\{P_1 = p_1(k)\}$	$p_1(k)$	CE(k)	$D_1(p_1)$	$D_2(p_1)$	$D_3(p_1)$
0	0.0985	1	0	0	0	0
1	0.2407	0.90152	0	0	0	0
2	0.2808	0.66080	0	0	0	0
3	0.2080	0.37996	0.0185	0.0429	0.0295	0.0625
4	0.1098	0.17193	0.0896	0.1139	0.1104	0.0896
5	0.0439	0.06213	0.3026	0.3270	0.3547	0.3026
6	0.0138	0.01822	0.6862	0.7105	0.8514	0.6862
7	0.0035	0.00439	1	1	1	1
8	0.0007	0.00088	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	<.0001	<.0001	1	1	1	1

Note that if the remaining level is spent proportionally to the probability of observing p_1, this results in an equal increase of the CE(k) values. For the example we have $\text{CE}(k) - D_1\big(p_1(k)\big) = \frac{\alpha - \alpha'}{\sum \Pr_{H_0}\{P_1 = p_1(j)\}} = 0.0243$. In contrast, a discrete conditional error function as in $D_2(p_1)$ will result in higher increases for higher CE(k) values, whereas $D_3(p_1)$ per construction increases only the smallest conditional error function value unequal to zero. We have developed an R function for the statistical software package R (R Development Core Team, 2011) that performs all necessary computations. The full source code is given in Chapter A.1 and execution of the function is demonstrated in Source code 4.1.

Planning a trial with any of these discrete conditional error functions starts with recruiting $n_1 = 22$ patients in the first stage. If the sample size of the second stage is kept at $n_2 = 11$, any of the conditional error functions CE(k), $D_1(p_1)$, $D_2(p_1)$ and $D_3(p_2)$ lead to the same decision rules as for Simon's minimax design. Note, however, that this is not generally true because $D_1(p_1)$, $D_2(p_1)$ and $D_3(p_1)$ may result in designs that are different from the original one for other parameter settings. Now consider the situation that the sample size of the second stage is adapted and that a total of $k = 3$ responses were observed in stage one. When basing the design on CE(k), the conditional significance level for the second stage is 0.0185. If $D_1(p_1)$, $D_2(p_1)$ or $D_3(p_1)$ was specified instead, a conditional level of 0.0429, 0.0295 or 0.0625 can be used. If, for example, the sample size is doubled to $n_2 = 22$ and if $l = 5$ responses were observed in the second stage ($p_2 = 0.0621$), the null hypothesis can be rejected if the study was planned for $D_3(p_1)$ but not with CE(k), $D_1(p_1)$ or $D_2(p_1)$.

These considerations show that it is advisable to evaluate in the planning phase how the remaining level $\alpha - \alpha'$ can be best implemented to lead to favorable design characteristics.

Source code 4.1: *Increase in conditional error function values for Simon's minimax design* $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$.

```
> updatedcef(0.1, ce, nominalalpha=0.05, how="proportionally")

dCEF:
 0 0 0 0.01853476 0.08956185 0.3026431 0.6861894 1 1 1 1 1 1 1 1
     1 1 1 1 1 1 1 1

Updated dCEF:
 0 0 0 0.04287696 0.1139041 0.3269853 0.7105316 1 1 1 1 1 1 1 1
    1 1 1 1 1 1 1 1

Alpha:   0.05

> updatedcef(0.1,ce, nominalalpha=0.05, how="equally")

dCEF:
 0 0 0 0.01853476 0.08956185 0.3026431 0.6861894 1 1 1 1 1 1 1 1
     1 1 1 1 1 1 1 1

Updated dCEF:
 0 0 0 0.02952131 0.1103785 0.3546847 0.8514978 1 1 1 1 1 1 1 1
    1 1 1 1 1 1 1 1

Alpha:   0.05

> updatedcef(0.1,ce, nominalalpha=0.05, how="border")

dCEF:
 0 0 0 0.01853476 0.08956185 0.3026431 0.6861894 1 1 1 1 1 1 1 1
     1 1 1 1 1 1 1 1

Updated dCEF:
 0 0 0 0.06248095 0.08956185 0.3026431 0.6861894 1 1 1 1 1 1 1 1
     1 1 1 1 1 1 1 1

Alpha:   0.05
```

Table 4.4.: *Sample size needed in the second stage to achieve a conditional power of* 0.90 *for* $\pi_1^* = 0.25$ *given a flexible version of Simon's minimax design for the parameter constellation* $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$

k	$\Pr_{H_0}\{P_1 = p_1(k)\}$	$\Pr_{H_1^*}\{P_1 = p_1(k)\}$	CE(k)	$D_1(p_1)$	$D_2(p_1)$	$D_3(p_1)$
3	0.2080	0.1017	69	59	64	50
4	0.1098	0.1611	45	40	40	45
5	0.0439	0.1933	25	20	20	25
6	0.0138	0.1826	9	9	9	9

A discrete conditional error function similar to the shape of $D_3(p_1)$ may be chosen if sample size recalculation should only be performed when few responses occur, while $D_1(p_1)$ or $D_2(p_1)$ may serve as a good choice when the sample size is recalculated irrespective of the first stage outcome. This can be illustrated by assuming that the minimax design for $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$ has been chosen in the planning phase but that it becomes evident after the first stage that the alternative $\pi_1^* = 0.25$ is still of clinical interest. For this alternative $H_1^* : \pi = \pi_1^*$, the fixed designs with a discrete conditional error function $D(p_1)$ based on CE(k), $D_1(p_1)$, $D_2(p_1)$ and $D_3(p_1)$ have a power of only 0.75. If in the second stage the sample size is recalculated to achieve a conditional power $\Pr_{H_1^*}(\text{Reject } H_0|k)$ of at least 0.90, an overall power of 0.88 is achieved for all these approaches. Note that as the study can also be stopped early, the desired conditional power is only achieved if the study continues to the second stage. Therefore, recalculation does not guarantee that the overall power will be equal to the conditional power.

Recalculation based on conditional power is performed by choosing n_2 as the minimum integer where the conditional power is greater than a specified boundary. An R function capable to perform the necessary calculations is given in the Appendix in Section A.2. For the considered flexible version of Simon's minimax design and $k = 3$ responses in stage one, application of this function is demonstrated in Source code 4.2. Table 4.4 lists the resulting sample sizes for CE(k), $D_1\big(p_1(k)\big)$, $D_2\big(p_1(k)\big)$ and $D_3\big(p_1(k)\big)$, $k \in \{0, \ldots, n_1\}$. Note that with less than three or more than seven responses in the first stage, the trial is stopped for futility or efficacy, respectively, and no additional sample size is needed. Due to the different discrete conditional error functions used, the sample size needed to achieve a conditional power of 0.90 for $\pi_1^* = 0.25$ varies. The expected sample size $\text{EN}(\pi_1^*) = 39.95$ for the design based on the conditional error function $D_1(p_1)$ is smaller as compared to using CE(k) ($\text{EN}(\pi_1^*) = 42.75$), $D_2(p_1)$ ($\text{EN}(\pi_1^*) = 40.46$) or $D_3(p_1)$ ($\text{EN}(\pi_1^*) = 40.81$). Of particular interest is the comparison between $D_1(p_1)$ and $D_2(p_1)$ in Table 4.4. While $D_2(p_1)$ allows for a higher local significance level for more than $k = 5$ responses in the first stage, this does not translate to a lower sample size in order to achieve a conditional

Source code 4.2: *Sample size recalculation to achieve a conditional power of* 0.90 *for* $\pi_1^* = 0.25$ *given a flexible version of Simon's minimax design for the parameter constellation* $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$ *and* $k = 3$ *responses in stage one*

```
> #Design parameters
> pi0 <- 0.1; n1 <- 22; pi1star <- 0.25; CP <- 0.9
>
> #Discrete conditional error function based on CE (k=3)
> recalcn2(pi0,n1,pi1star,0.0185,CP)
Sample size needed in the second stage:
69
>
> #Discrete conditional error function based on D1 (k=3)
> recalcn2(pi0,n1,pi1star,0.0429,CP)
Sample size needed in the second stage:
59
>
> #Discrete conditional error function based on D2 (k=3)
> recalcn2(pi0,n1,pi1star,0.0295,CP)
Sample size needed in the second stage:
64
>
> #Discrete conditional error function based on D3 (k=3)
> recalcn2(pi0,n1,pi1star,0.0625,CP)
Sample size needed in the second stage:
50
```

power of 0.90 for $\pi_1^* = 0.25$. In case of three responses, the design based on $D_2(p_1)$ needs a higher second-stage sample size. Therefore, for this particular recalculation scenario $D_1(p_1)$ is uniformly better than $D_2(p_1)$ with respect to second-stage sample size. This raises the question whether there exists an optimal discrete conditional error function that is uniformly more effective than every other discrete conditional error function. We address this question in Section 4.4 and we demonstrate how to construct these more efficient phase II designs.

The values of $\mathrm{CE}(k)$ and $D(p_1)$ can be calculated in a similar way for all designs presented by Simon (1989). The results are given in Englert and Kieser (2012b). Despite the fact that Simon's designs do not explicitly allow early stopping for efficacy after the first stage,

some of the resulting designs can be stopped early with rejection of H_0 after the first stage due to $D(p_1) = 1$ which occurs whenever $k > l_2$.

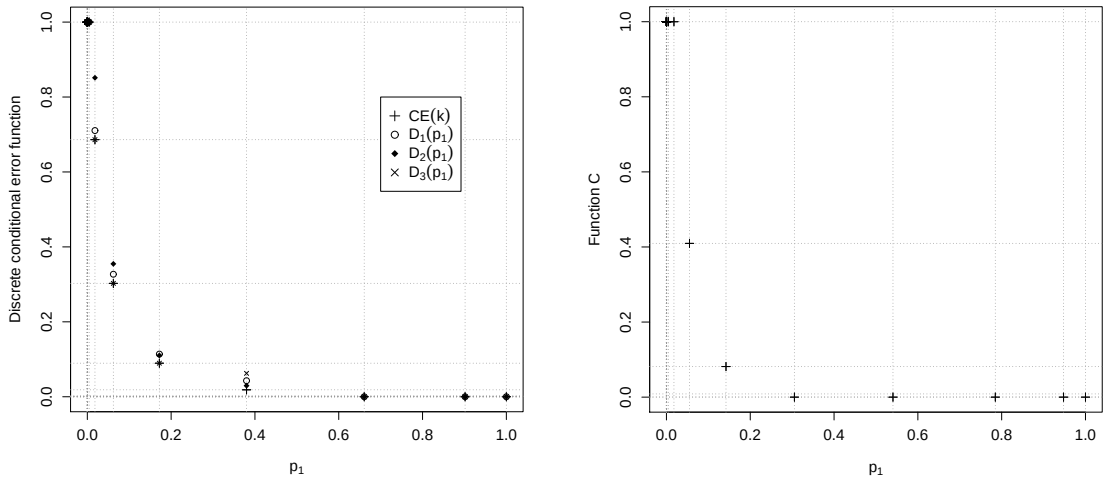
4.3.2. Flexible two-stage designs with early stopping for efficacy

There are also situations where it is desirable to terminate a phase II trial early if the initial response rate is high enough to give evidence of activity (see, e.g. Shuster, 2002; Mander and Thompson, 2010). Designs that allow stopping also for efficacy have been proposed by Fleming (1982), Chang et al. (1987) and Mander and Thompson (2010). The designs by Mander and Thompson (2010) where optimized according to the same criteria as Simon's design, i.e., by selecting that combination of $(l_1, u_1, n_1, l_2, n_2)$ that minimizes the expected or total sample size. The procedure to construct flexible two-stage designs with early stopping for efficacy is identical to the preceding section. Discrete conditional error functions for all designs presented in the article by Mander and Thompson (2010) are given in Englert and Kieser (2012b).

4.4. Construction of flexible and more efficient phase II designs

We have shown how by means of discrete conditional error functions that every phase II trial in oncology can be directly transferred into a flexible design. Additionally, we showed how the conservativeness of the designs can be overcome by increasing the conditional error bounds for the second stage. The left panel of Figure 4.1 illustrates for Simon's minimax design for $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$ the different discrete conditional error functions considered in Table 4.3. Vertical and horizontal gray dotted lines represent the attainable first- and second-stage p-values, respectively, for the given sample sizes $n_1 = 22$ and $n_2 = 11$. It can be seen that this approach usually results in values of the conditional error functions that cannot be attained if the originally planned sample size n_2 of Simon's design is not changed. In this case, the flexible designs based on these conditional error functions will lead to the same decision rules as for Simon's design and, consequently, shows the same characteristics.

In Section 4.1 and 4.2 on the other hand, we were in some cases able to improve on standard phase II designs by applying combination test methodology to discrete test statistics. The right panel of Figure 4.1 illustrates for the same parameter constellation the layout of a flexible two-stage design based on combination test approach minimizing the total sample size. These designs apparently make better use of the overall significance level and likewise allowed flexibility. Via the product of the p-values they made direct use of the second-stage



(a) *Discrete conditional error functions for Simon's minimax two-stage design as given in Table 4.3 ($n_1 = 22, n_2 = 11$)*

(b) *Function C of the flexible two-stage design based on combination test approach minimizing the total sample size ($n_1 = 28, n_2 = 5$)*

Figure 4.1.: *Comparison of conditional error function approach and combination test approach $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$. Vertical and horizontal gray dotted lines represent the attainable first- and second-stage p-values, respectively, for the given sample sizes n_1 and n_2.*

p-value if the originally planned sample size n_2 is not changed. Therefore, the rejection region defined by the function C matches to attainable second-stage p-values in this case and guarantees a good exhaustion of the level.

The general framework of discrete conditional error functions will now be combined with these findings. We construct conditional error functions that match to attainable second-stage p-values in case the planned second-stage sample size is not changed. This leads to new and more efficient phase II designs (denoted in the following as proposed method) that allow flexible design modifications and show at least as good characteristics as standard designs if no adaptations are performed.

4.4.1. Methodology and search strategy

Given n_1 and n_2, each p-value attainable in the first and second stage is determined by (3.1) and (3.2), respectively. Let $\boldsymbol{P}_i$ be the set of attainable p-values in stage i, $i = 1, 2$. At $c \in \boldsymbol{P}_2$, the distribution function of the uniform distribution equals the distribution function of the p-value p_2. Therefore, a rejection region with critical value c will show the

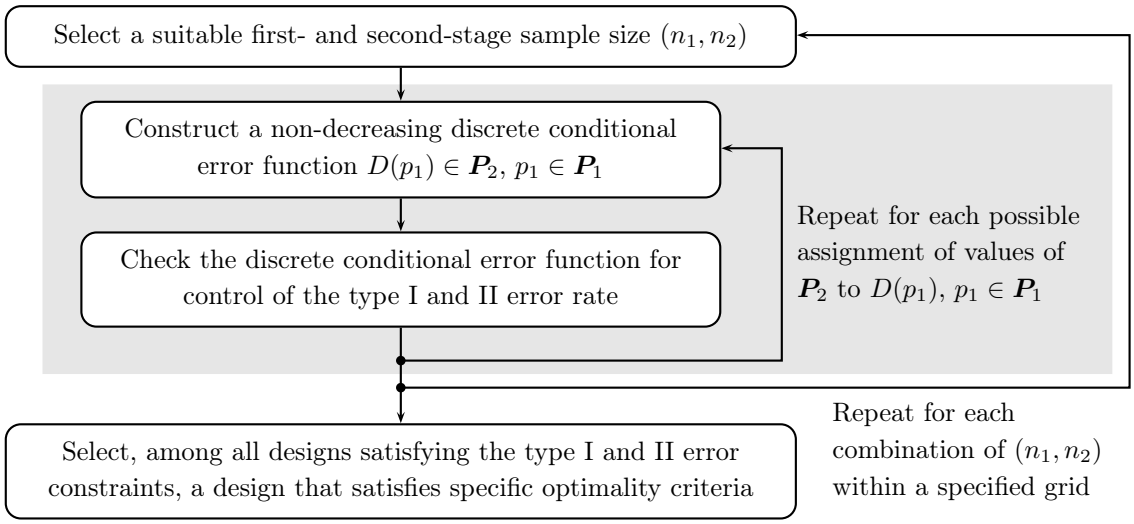


Figure 4.2.: *Algorithm for identifying flexible and more efficient phase II designs*

same characteristics in both cases. As shown in Theorem 4.1 it is necessary and sufficient to plan the second stage for uniformly distributed p-values to allow flexibility. Restricting the values of $D(p_1)$ to $\boldsymbol{P}_2$ and exhausting the level α to a maximal extent by selecting for each $p_1 \in \boldsymbol{P}_1$ the optimal $D(p_1) \in \boldsymbol{P}_2$ results in a discrete conditional error function that fulfills (3.3) and that simultaneously accounts for the discreteness of the second stage. Thus, the related design will show good exhaustion of the level in a flexible setting while at the same time exhibit favorable characteristics if no design changes are performed.

According to this idea, the search algorithm proposed to determine the conditional error function D for a two-stage design testing $H_0 : \pi = \pi_0$ against $H_1 : \pi = \pi_1$ and fulfilling the type I and II error rate constraints α and β is described in Figure 4.2. Note that the type I error rate is given by (3.3) and the type II error rate by

$$\beta' = 1 - \sum_{p \in \boldsymbol{P}_1} \Pr_{H_1}\{P_2 \leq D(p)\} \cdot \Pr_{H_1}(P_1 = p).$$

In a comprehensive computer-aided search it is now possible to search for given π_0, π_1, α and β for suitable flexible designs that satisfy certain optimality criteria. As demonstrated in the previous Section 4.3, the phase II designs proposed by Simon (1989), Chang et al. (1987), Fleming (1982), Shuster (2002) and others can alternatively be represented by a discrete conditional error function. Therefore, all these designs are contained within the set of designs supplied by the above algorithm. Consequently, the optimal design chosen among all feasible designs obtained by the algorithm will show at least as good characteristics as the classical design optimized with respect to the same criterion.

The main problem of this search strategy is that the number of possible combinations of

$D(k)$ (gray box in Figure 4.2) rapidly increases with the size of the set of possible discrete conditional error function values $|\boldsymbol{P}_2| = n_2 + 1$.

Theorem 4.2. *For given first- and second-stage sample sizes, n_1 and n_2, there exist*

$$\binom{n_1 + n_2 + 2}{n_1 + 1}$$

different non-decreasing discrete conditional error functions.

Proof. The set of possible first-stage p-values $\boldsymbol{P}_1$ is of size $n_1 + 1$ and the set of possible second-stage p-values $\boldsymbol{P}_2$ of size $n_2 + 2$, where $0 \in \boldsymbol{P}_2$ is added to represent early stopping for futility. Each discrete conditional error function is defined by the $n_1 + 1$ values of $\boldsymbol{P}_2$ assigned to each element of $\boldsymbol{P}_1$. Due to its monotony, it can also be represented as the set of different p-values used out of $\boldsymbol{P}_2$ plus position markers that indicate that the number at the next position is not smaller than the previous one. The number of discrete conditional error functions is therefore identical to the number of sets of $n_1 + 1$ items out of $n_2 + 2$ numbers and $n_1 - 1$ position markers (from position 1 to $n_1 - 1$), which is

$$\binom{(n_2 + 2) + (n_1 - 1)}{n_1 + 1} = \binom{n_1 + n_2 + 2}{n_1 + 1}.$$

□

Simon's minimax design for $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$ requires $n_1 = 22$ patients in the first stage and $n_2 = 11$ patients in the second stage (see Table 2.2). Under this constellation and with Theorem 4.2 there exist

$$\binom{35}{23} = 8.3 \cdot 10^8$$

different discrete conditional error functions for these values of n_1 and n_2. With a standard personal computer (Pentium Dual-Core @ 2.60GHz, 3.46 GB Ram) the construction and evaluation of one discrete conditional error function takes about $3.5 \cdot 10^{-5} s$. The evaluation of all possible solutions would take about 8 hours. Consequently, it is very time-consuming to perform a naïve search among a grid of possible combination of n_1 and n_2 to obtain the optimal design.

We do not want to restrict the set of valid designs, e.g., by imposing constraints on the discrete conditional error function values for specific number of responses. This may leave room for improvement of the resulting designs. Instead, we implement an intelligent algorithm that uses the branch-and-bound method, which is described in the next section and which allows an exhaustive and non-restricted search for the optimal design.

4.4.2. Branch-and-bound algorithm for identifying optimal designs

Branch-and-bound is a general algorithm for finding optimal solutions of one-dimensional discrete optimization problems. The branch-and-bound approach consists of two steps: a branching step that splits the problem into similar sub-problems, and a bounding step that discards branches that cannot lead to optimal solutions of the test problem. A general description of branch-and-bound can be found in most books on integer optimization (see, for example, Wolsey, 1998; Nemhauser and Wolsey, 1999).

In our application, the branching recursively defines the layout of the discrete conditional error function. In the first branching step, the optimization problem is split up into $|\boldsymbol{P}_2|$ sub-problems, where for each sub-problem the discrete conditional error level used for $k = 0$ responses in the first stage is set equal to a corresponding value of $\boldsymbol{P}_2$. The next branching step splits these sub-problems into further sub-problems, where in each of them the discrete conditional error level used for $k = 1$ responses is defined. Here, the restriction $D(1) \geq D(0)$ is used to ensure monotonicity. This procedure is iterated until the complete discrete conditional error function is defined.

Within the recursion, the bounding step in the branch-and-bound algorithm discards all sub-problems that cannot lead to the optimal design. In our application, all sub-problems are dropped that either cannot control the type I or II error rate or that cannot lead to a smaller average sample sizes than the designs found so far. Due to the monotonicity of the discrete conditional error function, the minimal type I error rate, minimal type II error rate and minimal expected sample size of all following sub-problems can be determined. After $m + 1$ branching steps, i.e., when the conditional error function is defined for 0 to m responses, the minimal type I error rate of all following sub-problems is given by

$$\alpha_{\min} = \sum_{k=0}^{m} D(k) \cdot \Pr_{H_0}(K = k) + D(m) \cdot \sum_{k=m+1}^{n_1} \Pr_{H_0}(K = k), \tag{4.7}$$

the minimal type II error rate by

$$\begin{aligned} \beta_{\min} = 1 - \Bigg[& \sum_{k=0}^{m} \Pr_{H_1}\big\{P_2 \leq D(k)\big\} \cdot \Pr_{H_1}(K = k) \\ & + \Pr_{H_1}\big\{P_2 \leq D(m)\big\} \cdot \sum_{k=m+1}^{n_1} \Pr_{H_1}(K = k) \Bigg] \end{aligned} \tag{4.8}$$

and the minimal average sample size is calculated as

$$\mathrm{EN}_{\min} = \sum_{k=0}^{m} \big\{n_1 + n_2 \cdot \mathbb{I}_{D(k)\neq 0 \,\wedge\, D(k)\neq 1}\big\} \cdot \Pr_{H_0}(K = k) + n_1 \cdot \sum_{k=m+1}^{n_1} \Pr_{H_0}(K = k), \tag{4.9}$$

where $\mathbb{I}$ denotes the indicator function

$$\mathbb{I}_{D(k)\neq 0 \,\wedge\, D(k)\neq 1} = \begin{cases} 1 & \text{if } D(k) \neq 0 \text{ and } D(k) \neq 1 \\ 0 & \text{else,} \end{cases}$$

and K denotes the random number of first-stage responses.

Every sub-problem is discarded that shows a value $\alpha_{\min}$ greater than the pre-specified type I error rate α, a value $\beta_{\min}$ greater than the type II error rate β, or a value $\mathrm{EN}_{\min}$ that is greater than the smallest average sample size found for all designs considered so far. Thus, the branch-and-bound algorithm identifies the optimal design minimizing the average sample size in an efficient way by avoiding to evaluate the characteristics of each potential solution. This enables a complete search over all possible discrete conditional error functions (gray box in Figure 4.2).

We developed an `R`-program that performs all necessary calculations presented here. The scheme of the layout of the program is given in Source code 4.3. It consists of three functions, the above mentioned `branch`- and `bound`-functions and a `launch`-routine, which defines all design parameters, calculates needed variables as, for example, $\boldsymbol{P}_2$ and initializes the first branching-step. The full code is given in the Appendix in Section A.3.

4.4.3. Resulting optimal designs

The branch-and-bound algorithm presented in the preceding section was used to identify the discrete conditional error function of flexible and more efficient optimal and minimax designs. We chose the same parameter constellations as in Simon (1989). For given $(\pi_0, \pi_1, \alpha, \beta)$ the search algorithm described in Figure 4.2 on page 42 was applied. The parts of the search algorithm marked with the gray box were performed within the branch-and-bound algorithm to speed up the optimization process. As an example, Source code 4.4 invokes the program and displays the output for $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$ with $n_1 = 22$ and $n_2 = 11$. Here, the program fully evaluates only 2 of the $8.3 \cdot 10^8$ possible discrete conditional error functions within the search for the optimal flexible phase II design. Therefore, only a tiny fraction of the total number of possible designs needs to be checked and the results are displayed immediately. This procedure was repeated to every combination of first- and second-stage sample size (n_1, n_2) with $n_1 + n_2 \leq 120$ and every combination of design parameters considered. The allowed maximum sample size of 120 thereby lies way above the range of total sample sizes in the designs by Simon making it very unlikely that this restriction influenced the identification of optimal designs. Among all discrete conditional error functions found for each combination of (n_1, n_2) the ones minimizing the expected (optimal design) or total sample size (minimax design) were selected. If different minimax designs had the same total sample size, the one minimizing

Source code 4.3: *Scheme of the implemented branch-and-bound approach*

```
#LAUNCH
launch <- function(pi0,pi1,nominalalpha,nominalbeta,n1,n2min,
   n2max){
  #Calculation of variables
  P_2
  #Initialize first branching-step
  branch(0,1)
  #Print final results
  }

#BRANCH
branch <- function(k,j){
  #Until here D is recursively defined for 1,...,k with
  #D(k) equal to the j-th element of P_2, i.e., P_2[j]
  if (k < n1){
    #Deeper into the tree
    for (P_2[i] >= P_2[j]){
      D(k+1) = P_2[i]           #Define D(k+1)
      if(bound(k+1,i) = TRUE){  #Bounding
        branch(k+1,i)           #Branching
        }
      }
    }
  else{
    #Output solution if D(k) is defined completely
    }
  }

#BOUND
bound <- function(k,j)){
  if(alpha_min > nominalalpha) {FALSE}
  if(beta_min > nominalbeta) {FALSE}
  if(en_min > en) {FALSE}
  else {TRUE}  #Further branching only if no restraints are
     fulfilled
  }
```

Source code 4.4: *Invoking and output of the branch-and-bound approach*

```
> launch(0.1,0.3,0.05,0.10,22,11)

Searching optimal solutions:

[1] 26.13126
 [1]  1  1  1  9 10 11 12 13 13 13 13 13 13 13 ... 13
[1] 25.97917
 [1]  1  1  1  9 10 11 13 13 13 13 13 13 13 13 ... 13

Search completed. In total, 2  of  834451800 combinations were
   evaluated.

Optimal D(k):
 0 0 0 0.01853476 0.08956185 0.3026431 1 1 1 1 1 1 1 ... 1

Optimal n_2(k):
 22 22 22 33 33 33 22 22 22 22 22 22 22 22 22 ... 22


Design characteristics:
Alpha:   0.04519655
Beta:    0.09465674
EN_p0:   25.97917
EN_p1:   25.22021
```

the expected sample size was selected. In case of optimal designs with identical expected sample sizes, the one with the smaller total sample size was chosen.

Figure 4.3 shows the resulting discrete conditional error function for the parameter choice of the example $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$. It can be seen how, by construction, the conditional error function matches to attainable second-stage p-values and therefore will guarantee an optimal exhaustion of the level if the originally planned sample size n_2 is not changed. With $n_1 = 21$ and $n_2 = 11$ the total sample size of the identified minimax design is also smaller compared to Simon's minimax design by one patient.

Further design characteristics for a variety of parameter settings are given in Table 4.5, 4.6, B.3 and B.4. For the purpose of comparison with classical phase II designs, the total and

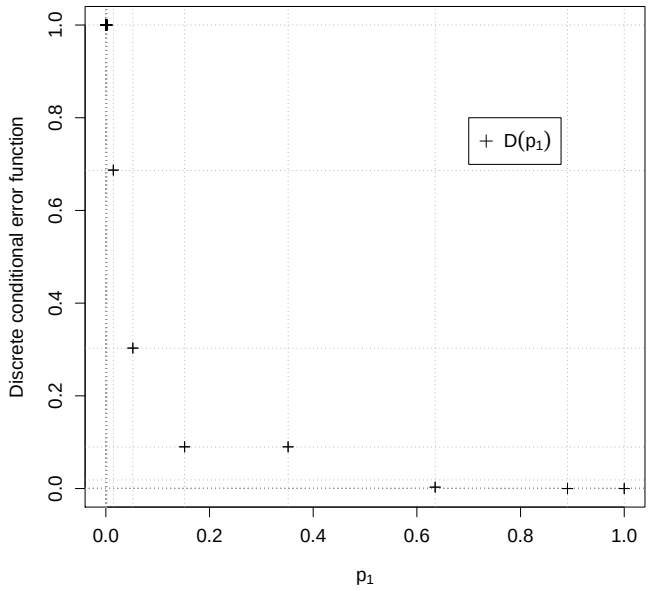


Figure 4.3.: *Discrete conditional error function resulting from the proposed flexible design ($n_1 = 21, n_2 = 11$). Vertical and horizontal gray dotted lines represent the attainable first- and second-stage p-values, respectively, for the given sample sizes n_1 and n_2.*

expected sample size for designs without (Simon, 1989) and with (Mander and Thompson, 2010) early stopping for efficacy, which are optimized with respect to the same criteria, are given additionally. The corresponding discrete conditional error functions are presented in Table 4.7, 4.8, B.5 and B.6. All computations were done using R.

To illustrate how to read these Tables, consider, for example, the optimal design identified for $(\pi_0, \pi_1, \alpha, \beta) = (0.3, 0.5, 0.05, 0.10)$. According to Table 4.7, $n_1 = 22$ patients need to be enrolled in the first stage. Depending on the first-stage p-value, the following conditional significance levels are used for the second part of the trial:

$$D(p_1) = \begin{cases} 0 & \text{if } \boldsymbol{P}_1 \ni p_1 > 0.3287 \\ 0.061 & \text{if } p_1 = 0.3287 \\ 0.113 & \text{if } p_1 = 0.1865 \\ 0.302 & \text{if } p_1 = 0.0916 \\ 0.302 & \text{if } p_1 = 0.0387 \\ 0.436 & \text{if } p_1 = 0.0140 \\ 0.581 & \text{if } p_1 = 0.0043 \\ 1 & \text{if } \boldsymbol{P}_1 \ni p_1 < 0.0043 \end{cases}$$

With a p-value p_1 greater than 0.3287 or smaller than 0.0043 the trial is stopped for futility or efficacy, respectively. In all other cases the trial continues with the second stage for

Table 4.5.: *Design characteristics of optimal flexible designs ($\pi_1 - \pi_0 = 0.2$)*

				Proposed method		Simon (1989)		Mander and Thompson (2010)	
π_0	π_1	α	β	n	EN(π_0)	n	EN(π_0)	n	EN(π_0)
0.05	0.25	0.05	0.20	21	11.17	17	11.95	17	11.89
			0.10	30	16.75	30	16.76	30	16.75
0.1	0.3	0.05	0.20	29	14.98	29	15.01	29	14.98
			0.10	41	22.19	35	22.53	41	22.19
0.2	0.4	0.05	0.20	43	20.54	43	20.58	43	20.54
			0.10	53	30.05	54	30.43	54	30.36
0.3	0.5	0.05	0.20	46	23.52	46	23.63	46	23.63
			0.10	59	34.12	63	34.72	63	34.69
0.4	0.6	0.05	0.20	46	24.49	46	24.52	46	24.49
			0.10	66	35.80	66	35.98	66	35.93
0.5	0.7	0.05	0.20	43	23.40	43	23.50	43	23.50
			0.10	59	33.47	61	34.01	59	33.47
0.6	0.8	0.05	0.20	37	20.42	43	20.48	43	20.48
			0.10	52	28.99	53	29.47	53	29.29
0.7	0.9	0.05	0.20	27	14.82	27	14.82	27	14.82
			0.10	36	20.92	36	21.23	36	21.13

Table 4.6.: *Design characteristics of minimax flexible designs ($\pi_1 - \pi_0 = 0.2$)*

				Proposed method		Simon (1989)		Mander and Thompson (2010)	
π_0	π_1	α	β	n	EN(π_0)	n	EN(π_0)	n	EN(π_0)
0.05	0.25	0.05	0.20	16	13.76	16	13.83	16	13.76
			0.10	24	20.42	25	20.36	25	18.55
0.1	0.3	0.05	0.20	23	19.20	25	19.51	24	20.30
			0.10	32	27.95	33	26.18	33	23.96
0.2	0.4	0.05	0.20	32	23.24	33	22.25	32	24.93
			0.10	44	33.43	45	31.23	44	35.68
0.3	0.5	0.05	0.20	36	29.32	39	25.69	36	30.68
			0.10	50	41.06	53	36.62	50	42.47
0.4	0.6	0.05	0.20	39	27.14	39	34.44	39	34.33
			0.10	53	42.68	54	38.06	54	38.03
0.5	0.7	0.05	0.20	37	26.90	37	27.74	37	26.90
			0.10	51	37.74	53	36.11	51	41.14
0.6	0.8	0.05	0.20	33	23.22	35	20.77	33	23.97
			0.10	45	31.52	45	35.90	45	33.30
0.7	0.9	0.05	0.20	25	18.05	27	14.82	26	23.11
			0.10	32	22.66	32	22.66	32	22.66

Table 4.7.: *Discrete conditional error function for optimal two-stage designs. Only values different from zero or one are presented. The discrete conditional error function equals zero for p-values greater than the ones given here and equals one for smaller ones.*

π_0	π_1	α	β	n_1	n_2	Discrete conditional error function						
0.05	0.25	0.05	0.2	7	14	p_1	0.3017	0.0444				
						$D(p_1)$	0.154	0.161				
		0.05	0.1	9	21	p_1	0.3698	0.0712	0.0084			
						$D(p_1)$	0.086	0.289	0.708			
0.1	0.3	0.05	0.2	10	19	p_1	0.2639	0.0702	0.0128			
						$D(p_1)$	0.120	0.310	0.661			
		0.05	0.1	17	24	p_1	0.2382	0.0826				
						$D(p_1)$	0.090	0.228				
0.2	0.4	0.05	0.2	13	30	p_1	0.2527	0.0991	0.0300	0.0070		
						$D(p_1)$	0.129	0.240	0.394	0.578		
		0.05	0.1	19	34	p_1	0.3267	0.1631	0.0676	0.0233	0.0067	
						$D(p_1)$	0.063	0.128	0.232	0.713	0.741	
0.3	0.5	0.05	0.2	15	31	p_1	0.2784	0.1311	0.0500	0.0152		
						$D(p_1)$	0.108	0.194	0.316	0.323		
		0.05	0.1	22	37	p_1	0.3287	0.1865	0.0916	0.0387	0.014	0.0043
						$D(p_1)$	0.061	0.113	0.302	0.302	0.436	0.581
0.4	0.6	0.05	0.2	16	30	p_1	0.2839	0.1423	0.0583	0.0191	0.0049	
						$D(p_1)$	0.099	0.178	0.291	0.438	0.627	
		0.05	0.1	25	41	p_1	0.2677	0.1538	0.0778	0.0344	0.0132	
						$D(p_1)$	0.098	0.164	0.254	0.258	0.629	
0.5	0.7	0.05	0.2	15	28	p_1	0.3036	0.1509	0.0592	0.0176		
						$D(p_1)$	0.093	0.174	0.289	0.295		
		0.05	0.1	21	38	p_1	0.3318	0.1917	0.0946	0.0392	0.0133	
						$D(p_1)$	0.072	0.128	0.209	0.314	0.436	
0.6	0.8	0.05	0.2	14	23	p_1	0.2793	0.1243	0.0398	0.0081	0.0004	
						$D(p_1)$	0.124	0.237	0.238	0.390	0.407	
		0.05	0.1	19	33	p_1	0.3081	0.1629	0.0696	0.023		
						$D(p_1)$	0.093	0.170	0.171	0.412		
0.7	0.9	0.05	0.2	6	21	p_1	0.4202	0.1176				
						$D(p_1)$	0.087	0.202				
		0.05	0.1	16	20	p_1	0.2459	0.0994	0.0261	0.0033		
						$D(p_1)$	0.107	0.238	0.609	0.895		

Table 4.8.: *Discrete conditional error function for minimax flexible designs. Only values different from zero or one are presented. The discrete conditional error function equals zero for p-values greater than the ones given here and equals one for smaller ones.*

π_0	π_1	α	β	n_1	n_2	Discrete conditional error function								
0.05	0.25	0.05	0.2	12	4	p_1	0.4596	0.1184						
						$D(p_1)$	0.025	0.223						
		0.05	0.1	16	8	p_1	0.5599	0.1892	0.0429					
						$D(p_1)$	0.058	0.06	0.349					
0.1	0.3	0.05	0.2	11	12	p_1	0.6862	0.3026	0.0896	0.0185				
						$D(p_1)$	0.026	0.111	0.113	0.349				
		0.05	0.1	21	11	p_1	0.6353	0.3516	0.152	0.0522	0.0144			
						$D(p_1)$	0.003	0.090	0.090	0.303	0.687			
0.2	0.4	0.05	0.2	19	13	p_1	0.3267	0.1631	0.0676	0.0233	0.0067	0.0016		
						$D(p_1)$	0.030	0.099	0.499	0.499	0.769	0.776		
		0.05	0.1	23	21	p_1	0.4993	0.3053	0.1598	0.0715	0.0273	0.0089		
						$D(p_1)$	0.043	0.043	0.109	0.231	0.416	0.827		
0.3	0.5	0.05	0.2	20	16	p_1	0.5836	0.3920	0.2277	0.1133	0.048	0.0171	0.0051	
						$D(p_1)$	0.007	0.026	0.075	0.176	0.341	0.756	0.906	
		0.05	0.1	32	18	p_1	0.5049	0.3560	0.2283	0.1326	0.0694	0.0327	0.0138	0.0052
						$D(p_1)$	0.021	0.021	0.060	0.141	0.279	0.467	0.67	0.842
0.4	0.6	0.05	0.2	18	21	p_1	0.4366	0.2632	0.1347	0.0576	0.0203	0.0058		
						$D(p_1)$	0.012	0.085	0.175	0.31	0.479	0.808		
		0.05	0.1	35	18	p_1	0.4272	0.2997	0.1935	0.1143	0.0615	0.03	0.0133	0.0053	0.0019
						$D(p_1)$	0.006	0.020	0.058	0.135	0.437	0.626	0.792	0.907	0.909
0.5	0.7	0.05	0.2	20	17	p_1	0.4119	0.2517	0.1316	0.0577	0.0207			
						$D(p_1)$	0.025	0.072	0.167	0.316	0.503			
		0.05	0.1	28	23	p_1	0.4253	0.2858	0.1725	0.0925	0.0436	0.0178	0.0063	
						$D(p_1)$	0.047	0.047	0.105	0.203	0.339	0.501	0.799	
0.6	0.8	0.05	0.2	14	19	p_1	0.4859	0.2793	0.1243	0.0398	0.0081			
						$D(p_1)$	0.024	0.071	0.165	0.493	0.512			
		0.05	0.1	23	22	p_1	0.3884	0.2373	0.1240	0.0540	0.0190	0.0052		
						$D(p_1)$	0.027	0.073	0.159	0.456	0.458	0.785		
0.7	0.9	0.05	0.2	13	12	p_1	0.4206	0.2025	0.0637	0.0097				
						$D(p_1)$	0.086	0.087	0.259	0.524				
		0.05	0.1	18	14	p_1	0.3327	0.1646	0.06	0.0142	0.0016			
						$D(p_1)$	0.048	0.161	0.355	0.585	0.972			

which the corresponding conditional significance level can be used. According to Table 4.5, if $n_2 = 37$ patients are enrolled in the second stage ($n = 59$), i.e., if no adaptations are performed, the expected sample size amounts to 34.12. For this example, Simon's optimal design is given by $(l_1, n_1, l_2, n_2) = (8, 24, 24, 39)$, $u_1 > n_1$, with an expected sample size of 34.72. An optimal design with early stopping for efficacy (Mander and Thompson, 2010) is defined by $(l_1, u_1, n_1, l_2, n_2) = (8, 15, 24, 24, 39)$ and shows an average sample size of 34.69 patients. Thus, the maximum and the expected sample size of the new proposed design is by 4 patients and 0.60 patients less, respectively, as compared to the corresponding Simon's optimal design, and by 4 and 0.57 less if an optimal design with early stopping for efficacy is used.

In summary, the proposed optimal designs outperform the designs by Simon and by Mander and Thompson in 21 of the considered 34 cases with regard to expected sample size and the minimax designs with respect to the total or expected sample size in 28 of 34 cases.

This benefit has to be paid with the side-effect, that the order of responses occurring throughout the trial may influence the test decision. In the given example, 10 responses ($p_1 = 0.0916$) in the first stage and 13 ($p_2 = 0.302$) responses in the second stage, i.e., a total of 23, lead to rejection of the null hypothesis, whereas with 9 ($p_1 = 0.1865$) responses observed in the first stage a total of 24 responses are needed to reject H_0.

By a small sample, we may judge of the whole piece.

(Miguel de Cervantes from Don Quixote)

5

Optimal Adaptive Designs for Phase II Trials in Oncology

In all designs developed so far, the second-stage sample size of the initially planned design does not depend on the number of responses observed in the interim analysis. It can, however, be changed throughout the trial in a flexible way. Per-design adaptive designs, which allow the sample size at the second stage to depend on the results from the first stage, may result in even more effective phase II designs than presented in the preceding Chapter 4, if there are no derivations from the planned scheme.

First attempts to construct adaptive designs for phase II oncology trials were made by Lin and Shih (2004) and Banerjee and Tsiatis (2006). To cope with uncertainty concerning the choice of an adequate response rate π_1, Lin and Shih presented a design where, based on the results of the interim analysis, the second stage of the study is powered for either a skeptic or an optimistic target response rate. Therefore, depending on the number of responses in the first stage, two different choices for the sample size of the second stage are possible. Banerjee and Tsiatis used a Bayesian decision-theoretic framework to construct optimal adaptive two-stage designs. They minimized an expected loss function by backwards induction to find adaptive two-stage designs with specified type I and type II error rates for the original test problem that minimize the expected sample size under the null hypothesis. Using this Bayesian decision-theoretic construct, the approach provides for each number of responses K observed in the n_1 patients of the first stage the optimally chosen second-stage sample size n_2 and the corresponding critical boundary. However, it remains unclear what restrictions the applied Bayesian framework constitutes on the resulting designs. The authors state:

> The designs we present may not necessarily be the optimal two-stage sequential design; i.e. the design which minimizes the expected sample size at $\pi = \pi_0$ subject to the type I and type II error constraints. However, we would expect such designs to be very close to optimal ... (Banerjee and Tsiatis, 2006)

An obvious way to identify optimal adaptive designs in the sense of Banerjee and Tsiatis is to apply an exhaustive search over all combinations of second-stage sample sizes and associated decision boundaries for each possible number of responses observed in the first stage. However, if the interplay between critical boundaries and sample sizes is not taken into account, this approach may provide contra-intuitive designs: For two study results with identical outcome in the second stage, only the result with fewer responses in the first stage, i.e., the less favorable outcome, may lead to a rejection of the null hypothesis. Additionally, due to the huge number of possible designs, a naïve exhaustive search over all combinations is only feasible for very small sample sizes that are even considerably below those usually applied in phase II designs.

We now show how to modify the discrete conditional error function representation of phase II designs proposed in Chapter 4 to allow for an exhaustive search for identifying optimal adaptive phase II designs. To search for the optimal design we apply the branch-and-bound algorithm introduced in Section 4.4.2. The adapted discrete conditional error function representation and the search algorithm are described in the following Section 5.1. Section 5.2 gives the results and compares our method to the approach proposed by Banerjee and Tsiatis (2006).

5.1. Modified discrete conditional error function methodology and search strategy

Recall that we consider the test problem $H_0 : \pi = \pi_0$ versus $H_1 : \pi = \pi_1$, $\pi_1 > \pi_0$, where the response rate π_0 represents insufficient activity while π_1 indicates sufficient efficacy to move the therapy to phase III. To present the proposed method, we use the same notation as in Banerjee and Tsiatis (2006). An adaptive two-stage design is represented by $\{n_1, n_2(K), l_2(K)\}$, where n_1 denotes the number of patients in the first stage, $n_2(K)$ the number of patients in the second stage, which may depend on the random number or responses K observed in the first stage, and $l_2(K)$ the decision boundary to reject the null hypothesis, which may also depend on K and therefore on $n_2(K)$. The null hypothesis is rejected if the total number of responses exceeds $l_2(K)$. Note that for all K with $n_2(K) = 0$ the trial is stopped after the first stage.

In the following, we use the conditional error function representation of phase II designs to

construct suitable designs. We will apply an exhaustive search over all suitable designs to find the corresponding optimal one, i.e., the one minimizing the average sample size under the null hypothesis. Additionally, we demonstrate that the designs considered by Banerjee and Tsiatis form a subgroup of the designs considered in our approach. Therefore, our approach will show at least equal but potentially more favorable design characteristics.

The concept of discrete conditional error functions was introduced in Chapter 4. Given a representation $\{n_1, n_2(K), l_2(K)\}$ of a per-design adaptive two-stage design, the corresponding discrete conditional error function representation can be obtained similar to fixed designs by setting $D(k) = 1 - B\{l_2(k) - k; \pi_0, n_2(k)\}$ for $k \in \{0, \ldots, n_1\}$ (see (3.5) on page 23). In particular, the designs by Simon (1989), Lin and Shih (2004) and Banerjee and Tsiatis (2006) can be characterized in such a way. The overall type I error rate of the design is given by

$$\sum_{k=0}^{n_1} D(k) \cdot \Pr_{H_0}(K = k). \tag{5.1}$$

Within the conditional error function framework, the null hypothesis is rejected after the second stage if the second-stage p-value of the applied binomial test $p_{2,n_2(k)}(l) := 1 - B\{l - 1; \pi_0, n_2(k)\}$ is equal to or falls below $D(k)$. Here, l denotes the number of responses observed in the second stage with sample size $n_2(k)$. In terms of the number of responses, the null hypothesis is rejected if the total number of observed responses after stage two exceeds the boundary $l_2(k)$. Hence, the null hypothesis is rejected if one of the following inequalities holds true:

$$\begin{aligned}
&l + k > l_2(k) \\
&\Leftrightarrow l - 1 \geq l_2(k) - k \\
&\Leftrightarrow 1 - B\{l - 1; \pi_0, n_2(k)\} \leq 1 - B\{l_2(k) - k; \pi_0, n_2(k)\} \\
&\Leftrightarrow p_{2,n_2(k)}(l) \leq D(k)
\end{aligned}$$

The equivalence of the first and last inequality shows that decision making based on the discrete conditional error function is identical to the classical evaluation of phase II designs based on boundaries formulated in terms of the observed number of responses. To construct optimal adaptive phase II designs, it is therefore sufficient to consider the discrete conditional error rate representation.

For any specified n_2, the attainable values of $D(k)$, i.e., the possible conditional type I error rates, are given by

$$\boldsymbol{P}_{2,n_2} := \{1 - B(x - 1; \pi_0, n_2) \mid x \in \{0, \ldots, n_2\}\}. \tag{5.2}$$

This set contains for each possible critical boundary the corresponding value of $D(k)$. For a given range of second-stage sample sizes $n_2 \in \mathcal{N}_2$ let $\boldsymbol{P}_2 = \bigcup_{n_2 \in \mathcal{N}_2} \boldsymbol{P}_{2,n_2} \cup \{0, 1\}$ denote the

possible values of the discrete conditional error function. The values 0 and 1 were added representing stopping after the first stage for futility or efficacy, respectively. The key idea is now that an exhaustive search for optimal adaptive designs over all combinations of second-stage sample sizes $n_2(K)$ and associated decision boundaries $l_2(K)$ is equivalent to a search over the corresponding discrete conditional error functions, i.e., over $D(K) \in \boldsymbol{P}_2$. As guaranteed by the construction above, the set $\boldsymbol{P}_2$ simultaneously accounts for the second-stage sample size and the decision boundary. Therefore, optimization must be done only in a one-dimensional set $\boldsymbol{P}_2$ and not in a two-dimensional array of all combinations of $n_2(K)$ and $l_2(K)$. This advantage of discrete conditional error functions will be utilized later in the search strategy as then the branch-and-bound algorithm can be applied.

It is a natural restriction to consider only sequences $D(k)$ that are non-decreasing in $k \in \{0, \ldots, n_1\}$. This restriction ensures that a higher conditional type I error level is associated with more responses observed in the first stage. In the special case where n_2 does not depend on the first-stage outcome, this restriction implies that with fewer responses in the first stage fewer responses in the second stage cannot lead to a rejection of the null hypothesis. This condition is satisfied for all phase II designs presented in the literature and excludes contra-intuitive designs as those mentioned at the beginning of this chapter.

The proposed algorithm for identifying optimal adaptive phase II designs with fixed maximum type I error rate α and maximum type II error rate β is described in Figure 5.1. Note that the type II error rate is given by

$$\beta' = 1 - \sum_{k=0}^{n_1} \Pr_{H_1}\left\{P_{2,n_2(k)} \leq D(k)\right\} \cdot \Pr_{H_1}(K = k), \tag{5.3}$$

with $P_{2,n_2(k)}$ being the random second-stage p-value, and that the average sample size under the null hypothesis is given by

$$\mathrm{EN}(\pi_0) = \sum_{k=0}^{n_1} \left\{n_1 + n_2(k)\right\} \cdot \Pr_{H_0}(K = k). \tag{5.4}$$

As the designs by Simon (1989), Lin and Shih (2004) and Banerjee and Tsiatis (2006) have a discrete conditional error function representation, they are included in the set of designs considered by the above-presented search strategy. The optimal design identified by the proposed algorithm will therefore necessarily show at least as good characteristics as these designs.

Similar to the non-adaptive case, the main problem of this search strategy is that the number of possible combinations of $D(k)$ (gray box in Figure 5.1) rapidly increases with

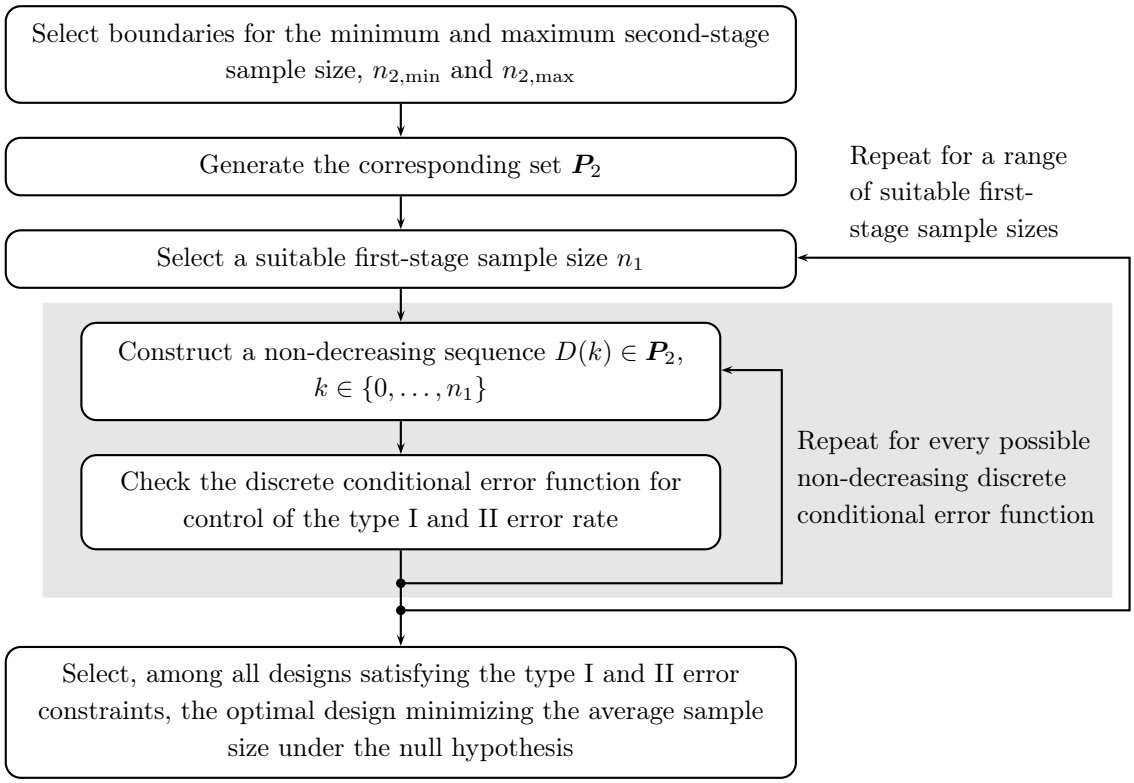


Figure 5.1.: *Algorithm for identifying optimal adaptive phase II designs*

the size of the set of possible discrete conditional error function values $|\boldsymbol{P}_2|$. Similar to Theorem 4.2 on page 43, it can be shown that for a fixed value of n_1 there exist

$$\binom{n_1 + |\boldsymbol{P}_2|}{n_1 + 1}$$

different non-decreasing discrete conditional error functions. As this design allows a dependency of the second-stage sample size on the number of responses in the first stage and therefore $|\boldsymbol{P}_2| \gg 1$, it is not feasible to perform a naïve search among the entire set of designs. As a numerical example, we consider the test problem $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.10)$ with $n_1 = 22$ and the specification that the second-stage sample size might vary in the range of $n_{2,\min} = 1$ to $n_{2,\max} = 15$. These parameter choices result in $|\boldsymbol{P}_2| = 137$ and therefore in $3.12 \cdot 10^{27}$ different discrete conditional error functions. With this multitude of design variants, it is not feasible to simply check each of these designs for type I and II error rate control and to select the optimal one.

In contrast to preceding work, we do not want to restrict the set of valid designs, e.g., by imposing constraints on the second-stage sample size using frequentist or Bayesian methods, which still leave a room for improvement of the resulting designs. Instead, we

implemented as in Section 4.4 the branch-and-bound method, which allows an exhaustive and non-restricted search for the optimal design. By this, our algorithm inevitably identifies the optimal adaptive two-stage design. Note that, due to the discreteness of the binomial distribution, the optimal design may not show a complete exhaustion of the type I and II error rate, but is optimal in the sense that every other design that can be represented by $\{n_1, n_2(K), l_2(K)\}$ has an equal or higher average sample size under the null hypothesis. We will see later that the optimal designs found here have type I and II error rates very close to nominal level.

From a methodological point of view, the only difference between the setting for the fixed size scenario in Section 4.4 and the application to adaptive designs here is the layout of the set $|\boldsymbol{P}_2|$. In the first case, this set defines the decision boundaries for the second-stage sample size. In the latter scenario, the set $\boldsymbol{P}_2$ simultaneously accounts for the second-stage sample size and the decision boundary. In both cases, however, optimization must be done only in this one-dimensional set $\boldsymbol{P}_2$. Therefore, the branch-and-bound algorithm can be directly used. The conceptional layout and implementation of the algorithm is the same as in Section 4.4: the branching recursively defines the layout of the discrete conditional error function and the bounding step algorithm discards all sub-problems that cannot lead to the optimal design. For the bounding algorithm, calculation of the minimal type I error rate is given in (4.7). Note, however, that by (5.2) with a value chosen for $D(k) \in \boldsymbol{P}_2$ the second-stage sample size $n_2(k)$ corresponding to this value is also determined. The minimal type II error rate and average sample size are therefore now calculated as

$$\beta_{\min} = 1 - \left[\sum_{k=0}^{m} \Pr_{H_1}\left\{P_{2,n_2(k)} \leq D(k)\right\} \cdot \Pr_{H_1}(K = k) + \sum_{k=m+1}^{n_1} \Pr_{H_1}(K = k) \right]$$

and

$$\mathrm{EN}_{\min} = \sum_{k=0}^{m} \left\{n_1 + n_2(k)\right\} \cdot \Pr_{H_0}(K = k) + n_1 \cdot \sum_{k=m+1}^{n_1} \Pr_{H_0}(K = k).$$

The R-program given in the Appendix in Section A.3 is capable to perform all necessary calculations. For the example above, Source code 5.1 invokes the program and displays the output. The program fully evaluates only 3889 of the $3.12 \cdot 10^{27}$ possible discrete conditional error functions within the search for the optimal adaptive phase II design. Therefore, only a tiny fraction of the total number of possible designs needs to be checked resulting in a significant gain in performance. For this parameter setting this resulted in a computation time of only a few seconds. For other parameter settings, with a higher range of second-stage sample sizes and greater $|\boldsymbol{P}_2|$, the computational effort still becomes impractically large. Restricting the set of discrete conditional error function values $\boldsymbol{P}_2$ solves this problem elegantly. This is due to the fact that from a practical point of view it is logical to consider besides zero and one only discrete conditional error function values that

Source code 5.1: *Invoking and output of the branch-and-bound approach for adaptive designs*

```
> launch(0.1,0.3,0.05,0.10,22,1,15)

Searching optimal solutions:

[1] 23.95059
 [1]   1   1   1   1  94 110 122 122 122 122 122 122 122 ... 122
 ...
[1] 23.95059
 [1]   1   1   1   1  94 110 122 122 122 122 122 137 137 ... 137
[1] 23.95057
 [1]   1   1   1   1  94 110 122 122 122 122 137 137 137 ... 137
[1] 23.95045
 [1]   1   1   1   1  94 110 122 122 122 137 137 137 137 ... 137
[1] 23.94972
 [1]   1   1   1   1  94 110 122 122 137 137 137 137 137 ... 137
[1] 23.94620
 [1]   1   1   1   1  94 110 122 137 137 137 137 137 137 ... 137
[1] 23.93238
 [1]   1   1   1   1  94 110 137 137 137 137 137 137 137 ... 137

Search completed. In total,  3889  of  3.115117e+27 combinations
    were evaluated.

Optimal D(k):
 0 0 0 0 0.1108700 0.4153709 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Optimal n_2(k):
 22 22 22 22 34 36 22 22 22 22 22 22 22 22 22 22 22 22 ... 22


Design characteristics:
Alpha:   0.04863099
Beta:    0.09952334
EN_p0:   23.93238
EN_p1:   25.24193
```

are away from zero or one by a certain small amount. Values smaller than α correspond, for example, to situations where it is, with respect to patient resources, better to start a new trial than to continue with the current one with a possibly modified second stage. On the other hand, values close to one constitute situations, where only a few patients suffice to reach sufficient statistical power for the next stage. In our applications a cutoff value of $\alpha/2$ and $1-\alpha/2$, respectively, has proven to lead to situations that are feasible from the computational aspect and logical with respect to these considerations. As we will see in the next section all resulting optimal adaptive designs show discrete conditional error function values far away from these cutoff values. Therefore, this restriction did not influence the optimization process.

5.2. Resulting optimal adaptive designs

We compare the adaptive two-stage designs developed by Banerjee and Tsiatis (2006) with our optimal two-stage adaptive designs found by the branch-and-bound algorithm. Additionally, the characteristics of Simon's designs (1989) are given for comparison. Banerjee and Tsiatis developed both unrestricted and restricted designs (where the total maximum sample size $n_1 + n_{2,max}$ does not exceed Simon's maximum sample size by more than 10%). As in practical applications it is reasonable to restrict the maximum total sample size and to allow for a fair comparison, we only consider designs with the same restraints on the maximum sample size as in Banerjee and Tsiatis' restricted designs. Otherwise, no restrictions on the second-stage sample size were made, i.e., every $n_2(K)$ between 1 and $n_{2,\max}$ was allowed. For the first-stage sample size n_1, we searched for the optimal solution within a range of ± 4 centered at the corresponding first-stage sample size of the Banerjee and Tsiatis design. All designs aim at minimizing the expected sample size under the null hypothesis. The resulting design characteristics are summarized in Table 5.1.

The improvements in terms of average sample sizes that can be achieved by application of the branch-and-bound algorithm as compared to the method by Banerjee and Tsiatis are moderate with mean savings in average sample size of 0.32 (range $0-1.15$). However, taking into account that Simon's designs were deemed to be optimal for decades and that Banerjee and Tsiatis could further improve them by the adaptive approach, the additional decrease in sample size that can be achieved by applying the branch-and-bound algorithm may be regarded as remarkable. For example, for the parameter constellation $(\pi_0, \pi_1, \alpha, \beta) = (0.20, 0.35, 0.05, 0.20)$ the expected sample sizes for the Simon, the Banerjee and Tsiatis and the proposed design are 35.37, 34.77 and 34.05, respectively. The mean reduction in expected sample size achieved for the designs presented in Table 5.1 by application of the proposed method as compared to Simon's designs amounts to 1.01

Table 5.1.: *Comparison of average sample size under the null hypothesis EN(π_0) of the optimal designs by Simon (S), the designs by Banerjee and Tsiatis (BT), and the proposed optimal adaptive designs found by the branch-and-bound algorithm (BB)*

π_0	π_1	α	β	$\mathrm{EN}(\pi_0)_{\mathrm{S}}$	$\mathrm{EN}(\pi_0)_{\mathrm{BT}}$	$\mathrm{EN}(\pi_0)_{\mathrm{BB}}$
0.05	0.25	0.05	0.2	11.89	11.03	11.03
			0.1	16.75	16.67	16.40
	0.20	0.05	0.2	17.60	17.45	17.45
			0.1	26.60	25.83	25.83
0.10	0.30	0.05	0.2	15.01	15.01	14.72
			0.1	22.50	22.29	21.70
	0.25	0.05	0.2	24.65	24.65	24.41
			0.1	36.82	36.24	35.13
0.20	0.40	0.05	0.2	20.58	20.18	19.80
			0.1	30.43	29.21	29.02
	0.35	0.05	0.2	35.37	34.77	34.05
			0.1	51.45	50.23	50.07
0.30	0.50	0.05	0.2	23.63	23.21	23.02
			0.1	34.72	33.78	33.31
	0.45	0.05	0.2	41.71	40.80	40.61
			0.1	60.77	58.60	58.56
0.40	0.60	0.05	0.2	24.52	24.43	24.09
			0.1	35.98	34.95	34.48
	0.55	0.05	0.2	44.93	43.32	43.20
			0.1	63.96	62.88	62.85
0.50	0.70	0.05	0.2	23.50	23.08	23.03
			0.1	34.01	33.45	32.95
	0.65	0.05	0.2	43.72	42.20	42.15
			0.1	62.28	60.90	60.77
0.60	0.80	0.05	0.2	20.48	20.08	19.72
			0.1	29.47	29.08	28.15
	0.75	0.05	0.2	39.35	39.01	37.86
			0.1	55.60	54.35	54.30
0.70	0.90	0.05	0.2	14.82	14.82	14.82
			0.1	21.23	20.89	20.42
	0.85	0.05	0.2	30.29	29.62	29.16
			0.1	43.40	41.59	41.57
0.80	0.95	0.05	0.2	17.72	17.56	17.56
			0.1	24.45	24.38	23.75

(range $0 - 2.21$). In view of the high number of phase II trials performed worldwide, application of the new method may thus lead to a considerable saving of patients. This gain has not to be paid with an increased loss in power or conservativeness. In contrast, the type I and type II error rates of the new optimal adaptive designs are on average closer to the nominal values than for the other two design variants. For all designs depicted in Table 5.1, the mean undershooting of the actual type I error rate to the nominal level $\alpha = 0.05$ amounts to 0.0016, 0.0012 and 0.0004 for Simon's designs, Banerjee and Tsiatis' designs, and the proposed per-design adaptive designs, respectively, while the actual type II error rate falls below the nominal level by 0.0020, 0.0009 and 0.0002 for $\beta = 0.1$ and by 0.0028, 0.0023 and 0.0005 for $\beta = 0.2$. Note that there is a typo in Table I of the publication of Banerjee and Tsiatis concerning the expected sample size for the design parameter constellation $(\pi_0, \pi_1, \alpha, \beta) = (0.70, 0.90, 0.05, 0.20)$.

The layouts of the proposed designs are presented in Table 5.2. For each design specification and number of responses in the first stage, the corresponding optimal second-stage sample size and the discrete conditional error function value are given. For convenience, we added the corresponding critical boundary $l_2(k)$.

As could be expected, the new designs are generally similar to those given by Banerjee and Tsiatis, but there exist also marked differences for some parameter constellations.

A possibly counter-intuitive feature of the proposed optimal adaptive design is that initially the second-stage sample size increases with the number of responses observed in the first stage. Only when the number of observed responses exceeds a certain value, the second-stage sample size slightly decreases in most cases. A similar pattern occurred for the designs given in the work by Banerjee and Tsiatis, where here the sample size for the second stage does not decrease with increasing number of observed responses for any of the reported designs. The authors stated that this observed sample size pattern is dictated by the Bayesian decision rule applied to obtain the optimal adaptive design. As, however, our approach does not use this rule, we can conclude from our calculations that this pattern is not caused by the applied loss function, but is due to the optimization process and the discreteness of the applied test statistic. The total sample size associated with more responses observed in the interim analysis, i.e., with situations that are rather unlikely under the null hypothesis, has a less pronounced influence on the average sample size under H_0. It appears that reducing the second-stage sample size for fewer responses (which are more likely under H_0) is more favorable, even at the price of a higher stage two sample size in case of a higher number of responses observed in stage one.

All results given so far aim at minimization of the average sample size under the null hypothesis with the restriction that the total maximum sample size $n_1 + n_{2,max}$ does not exceed Simon's maximum sample size by more than 10%. This restriction is not necessary

Table 5.2.: *Layout of proposed optimal adaptive designs for (a) $\pi_1 - \pi_0 = 0.2$ and $\beta = 0.2$; (b) $\pi_1 - \pi_0 = 0.2$ and $\beta = 0.1$; (c) $\pi_1 - \pi_0 = 0.15$ and $\beta = 0.2$; (d) $\pi_1 - \pi_0 = 0.15$ and $\beta = 0.1$. For all designs $\alpha = 0.05$ was specified.*

(a)

$\pi_0 = 0.05$

$n_1 = 8$, $n_{2,\max} = 10$

k	n_2	$D(k)$	$l_2(k)$
0	0	0	0
1	9	.071	2
2	10	.401	2
≥3	0	1	0

$\pi_0 = 0.1$

$n_1 = 11$, $n_{2,\max} = 21$

k	n_2	$D(k)$	$l_2(k)$
≤1	0	0	0
2	11	.090	4
3	17	.238	5
4	11	.686	4
≥5	0	1	0

$\pi_0 = 0.2$

$n_1 = 11$, $n_{2,\max} = 36$

k	n_2	$D(k)$	$l_2(k)$
≤2	0	0	0
3	17	.106	8
4	30	.129	12
5	34	.227	13
6	35	.255	14
7	32	.465	13
≥8	0	1	0

$\pi_0 = 0.3$

$n_1 = 13$, $n_{2,\max} = 37$

k	n_2	$D(k)$	$l_2(k)$
≤4	0	0	0
5	23	.120	14
6	35	.135	19
7	37	.193	20
8	36	.263	20
9	37	.434	20
≥10	0	1	0

(b)

$\pi_0 = 0.05$

$n_1 = 11$, $n_{2,\max} = 22$

k	n_2	$D(k)$	$l_2(k)$
0	0	0	0
1	10	.086	2
2	22	.095	4
3	15	.537	3
≥4	0	1	0

$\pi_0 = 0.1$

$n_1 = 14$, $n_{2,\max} = 25$

k	n_2	$D(k)$	$l_2(k)$
≤1	0	0	0
2	16	.068	5
3	22	.171	6
4	25	.236	7
5	23	.408	7
6	15	.794	6
≥7	0	1	0

$\pi_0 = 0.2$

$n_1 = 20$, $n_{2,\max} = 39$

k	n_2	$D(k)$	$l_2(k)$
≤4	0	0	0
5	16	.082	10
6	30	.129	14
7	33	.200	15
8	39	.241	17
9	39	.376	17
≥10	0	1	0

$\pi_0 = 0.3$

$n_1 = 22$, $n_{2,\max} = 47$

k	n_2	$D(k)$	$l_2(k)$
≤7	0	0	0
8	25	.098	18
9	38	.137	23
10	46	.191	26
11	45	.254	26
12	46	.403	26
13	46	.530	26
14	10	.972	14
≥15	0	1	0

(c)

$\pi_0 = 0.05$

$n_1 = 10$, $n_{2,\max} = 21$

k	n_2	$D(k)$	$l_2(k)$
0	0	0	0
1	18	.058	3
2	21	.283	3
3	20	.642	3
≥4	0	1	0

$\pi_0 = 0.1$

$n_1 = 14$, $n_{2,\max} = 33$

k	n_2	$D(k)$	$l_2(k)$
≤1	0	0	0
2	22	.062	6
3	30	.175	7
4	33	.230	8
5	20	.608	6
6	16	.815	6
7	24	.920	7
≥8	0	1	0

$\pi_0 = 0.2$

$n_1 = 20$, $n_{2,\max} = 58$

k	n_2	$D(k)$	$l_2(k)$
≤4	0	0	0
5	28	.090	13
6	39	.141	16
7	57	.152	21
8	58	.260	21
9	55	.420	20
10	50	.556	19
≥11	0	1	0

$\pi_0 = 0.3$

$n_1 = 25$, $n_{2,\max} = 63$

k	n_2	$D(k)$	$l_2(k)$
≤8	0	0	0
9	41	.079	25
10	48	.164	27
11	58	.186	31
12	61	.266	32
13	63	.325	33
14	63	.428	33
≥15	0	1	0

(d)

$\pi_0 = 0.05$

$n_1 = 20$, $n_{2,\max} = 23$

k	n_2	$D(k)$	$l_2(k)$
≤1	0	0	0
2	22	.095	4
3	23	.321	4
4	23	.693	4
≥5	0	1	0

$\pi_0 = 0.1$

$n_1 = 23$, $n_{2,\max} = 50$

k	n_2	$D(k)$	$l_2(k)$
≤2	0	0	0
3	17	.083	6
4	42	.121	10
5	47	.186	11
6	47	.329	11
7	50	.384	12
8	50	.750	11
≥9	0	1	0

$\pi_0 = 0.2$

$n_1 = 33$, $n_{2,\max} = 58$

k	n_2	$D(k)$	$l_2(k)$
≤7	0	0	0
8	41	.102	19
9	58	.103	24
10	58	.169	24
11	58	.260	24
12	58	.373	24
13	57	.473	24
14	58	.630	24
15	35	.939	18
≥16	0	1	0

$\pi_0 = 0.3$

$n_1 = 39$, $n_{2,\max} = 81$

k	n_2	$D(k)$	$l_2(k)$
≤13	0	0	0
14	27	.080	25
15	48	.100	33
16	64	.122	39
17	75	.157	43
18	81	.217	45
19	81	.293	45
20	75	.394	43
21	81	.571	43
≥22	0	1	0

Table 5.2.: *continued*

(a)

$\pi_0 = 0.4$

$n_1 = 16$, $n_{2,\max} = 35$

k	n_2	$D(k)$	$l_2(k)$
≤ 7	0	0	0
8	24	.114	20
9	32	.165	24
10	35	.300	25
11	34	.373	25
12	34	.509	25
13	11	.970	14
14	15	.973	16
≥ 15	0	1	0

$\pi_0 = 0.5$

$n_1 = 15$, $n_{2,\max} = 31$

k	n_2	$D(k)$	$l_2(k)$
≤ 8	0	0	0
9	23	.105	23
10	30	.181	27
11	31	.237	28
12	29	.356	27
13	19	.676	21
14	18	.881	20
15	4	.938	15

$\pi_0 = 0.6$

$n_1 = 10$, $n_{2,\max} = 37$

k	n_2	$D(k)$	$l_2(k)$
≤ 6	0	0	0
7	21	.096	22
8	31	.143	29
9	31	.245	29
10	34	.354	31

(b)

$n_1 = 22$, $n_{2,\max} = 51$

k	n_2	$D(k)$	$l_2(k)$
≤ 9	0	0	0
10	21	.085	21
11	35	.114	28
12	45	.144	33
13	50	.234	35
14	51	.272	36
15	49	.393	35
≥ 16	0	1	0

$n_1 = 20$, $n_{2,\max} = 47$

k	n_2	$D(k)$	$l_2(k)$
≤ 10	0	0	0
11	19	.084	23
12	32	.108	31
13	47	.121	40
14	44	.226	38
15	47	.280	40
16	46	.329	40
17	43	.500	38
18	35	.750	33
≥ 19	0	1	0

$n_1 = 18$, $n_{2,\max} = 39$

k	n_2	$D(k)$	$l_2(k)$
≤ 11	0	0	0
12	18	.094	25
13	32	.116	35
14	38	.186	39
15	36	.262	38
16	36	.384	38
17	35	.436	38
18	0	1	0

(c)

$n_1 = 28$, $n_{2,\max} = 65$

k	n_2	$D(k)$	$l_2(k)$
≤ 12	0	0	0
13	37	.108	31
14	52	.148	38
15	62	.169	43
16	65	.262	44
17	64	.312	44
18	65	.446	44
19	63	.568	43
≥ 20	0	1	0

$n_1 = 25$, $n_{2,\max} = 66$

k	n_2	$D(k)$	$l_2(k)$
≤ 13	0	0	0
14	37	.094	36
15	51	.131	44
16	63	.157	51
17	65	.229	52
18	63	.307	51
19	63	.401	51
20	60	.551	49
21	14	.971	24
≥ 22	0	1	0

$n_1 = 24$, $n_{2,\max} = 49$

k	n_2	$D(k)$	$l_2(k)$
≤ 15	0	0	0
16	36	.090	41
17	45	.143	47
18	49	.184	50
19	49	.272	50
20	48	.422	49
21	45	.564	47
≥ 22	0	1	0

(d)

$n_1 = 45$, $n_{2,\max} = 69$

k	n_2	$D(k)$	$l_2(k)$
≤ 19	0	0	0
20	38	.078	39
21	57	.103	48
22	68	.144	53
23	68	.206	53
24	68	.283	53
25	68	.371	53
26	69	.410	54
27	60	.549	50
≥ 28	0	1	0

$n_1 = 39$, $n_{2,\max} = 77$

k	n_2	$D(k)$	$l_2(k)$
≤ 20	0	0	0
21	38	.072	44
22	54	.11	53
23	73	.121	64
24	73	.175	64
25	76	.211	66
26	76	.283	66
27	77	.410	66
28	76	.454	66
29	77	.590	66
30	73	.680	64
≥ 31	0	1	0

$n_1 = 31$, $n_{2,\max} = 73$

k	n_2	$D(k)$	$l_2(k)$
≤ 19	0	0	0
20	48	.081	53
21	66	.108	65
22	72	.150	69
23	73	.189	70
24	73	.261	70
25	68	.340	67
26	73	.437	70
27	66	.612	65
≥ 28	0	1	0

Table 5.2.: *continued*

(a) $\pi_0 = 0.7$

$n_1 = 6$, $n_{2,\max} = 22$

k	n_2	$D(k)$	$l_2(k)$
≤ 4	0	0	0
5	21	.086	22
6	21	.198	22

(b) $\pi_0 = 0.7$

$n_1 = 15$, $n_{2,\max} = 25$

k	n_2	$D(k)$	$l_2(k)$
≤ 11	0	0	0
12	15	.127	24
13	22	.165	30
14	25	.341	32
15	20	.608	28

(c) $\pi_0 = 0.7$

$n_1 = 18$, $n_{2,\max} = 47$

k	n_2	$D(k)$	$l_2(k)$
≤ 13	0	0	0
14	24	.111	33
15	42	.148	47
16	46	.233	50
17	47	.311	51
18	20	.772	30

(d) $\pi_0 = 0.7$

$n_1 = 24$, $n_{2,\max} = 62$

k	n_2	$D(k)$	$l_2(k)$
≤ 17	0	0	0
18	33	.094	44
19	47	.124	55
20	61	.143	66
21	61	.219	66
22	62	.285	67
23	54	.425	61
24	30	.730	43

(c) $\pi_0 = 0.8$

$n_1 = 9$, $n_{2,\max} = 21$

k	n_2	$D(k)$	$l_2(k)$
≤ 7	0	0	0
8	19	.083	25
9	21	.179	27

(d) $\pi_0 = 0.8$

$n_1 = 16$, $n_{2,\max} = 27$

k	n_2	$D(k)$	$l_2(k)$
≤ 13	0	0	0
14	19	.083	31
15	27	.182	38
16	25	.421	36

and, as pointed out in Section 2.1, a variety of different optimization rules are used in the construction of phase II trials in oncology. Therefore, we further investigated unrestricted adaptive designs for different minimization strategies. For convenience of tabulation, we present results only for three specific parameter situations, namely $(\pi_0, \pi_1, \alpha, \beta) = (0.1, 0.3, 0.05, 0.2)$, denoted as example 1, $(\pi_0, \pi_1, \alpha, \beta) = (0.2, 0.4, 0.05, 0.2)$, denoted as example 2 and $(\pi_0, \pi_1, \alpha, \beta) = (0.3, 0.5, 0.05, 0.2)$, denoted as example 3. We considered unrestricted adaptive designs that minimize for these design parameters (a) the expected sample size under the null hypothesis $\mathrm{EN}(\pi_0)$, (b) the expected sample size under the alternative hypothesis $\mathrm{EN}(\pi_1)$, (c) the maximum sample size $n_1 + \max\big(n_2(k)\big)$ and (d) the sum of the expected sample size under the null and alternative hypothesis $\mathrm{EN}(\pi_0) + \mathrm{EN}(\pi_1)$. The latter optimization rule was also used by Levin et al. (2012) to investigate efficient types of adaptation for continuous test statistics.

It is straightforward to modify the code of the branch-and-bound approach to obtain unrestricted adaptive designs that are optimal with respect to these optimization criteria thus extending the work by Simon (1989), Mander and Thompson (2010) and Jung et al. (2004) (for the practical implementation, see Appendix A.3.4). For computational reasons, we investigated designs with a total sample size smaller than $1'000$ only. All results are given in Table 5.3.

First we have a closer look on the resulting optimal designs with respect to the null hy-

Table 5.3.: *Layout of proposed optimal adaptive designs for different minimization strategies*

(a) $\mathrm{EN}(\pi_0)$ (b) $\mathrm{EN}(\pi_1)$ (c) $n_1 + \max\big(n_2(k)\big)$ (d) $\mathrm{EN}(\pi_0) + \mathrm{EN}(\pi_1)$

Example 1

$n_1 = 10$				$n_1 = 12$				$n_1 = 11$				$n_1 = 11$			
k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$
≤ 1	0	0	0	≤ 1	0	0	0	0	0	0	0	≤ 1	0	0	0
2	14	.158	4	2	17	.022	6	1	11	.019	4	2	11	.083	4
3	24	.214	6	3	9	.225	4	2	12	.026	5	3	21	.152	6
4	25	.463	6	≥ 4	0	1	0	3	12	.341	4	≥ 4	0	1	0
≥ 5	10	1	0					4	10	.651	4				
								≥ 5	0	1	0				

Example 2

$n_1 = 11$				$n_1 = 16$				$n_1 = 19$				$n_1 = 15$			
k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$
≤ 2	0	0	0	≤ 3	0	0	0	≤ 4	0	0	0	≤ 4	0	0	0
3	17	.106	8	4	28	.015	14	5	12	.073	9	5	18	.051	11
4	27	.156	11	5	18	.051	11	6	13	.100	10	6	24	.089	13
5	36	.168	14	6	13	.252	9	7	13	.253	10	7	14	.302	10
6	48	.241	17	≥ 7	0	1	0	8	11	.678	9	≥ 8	0	1	0
7	54	.274	19					9	8	.832	9				
8	68	.276	23					≥ 10	0	1	0				
9	48	.361	19												
≥ 10	0	1	0												

Example 3

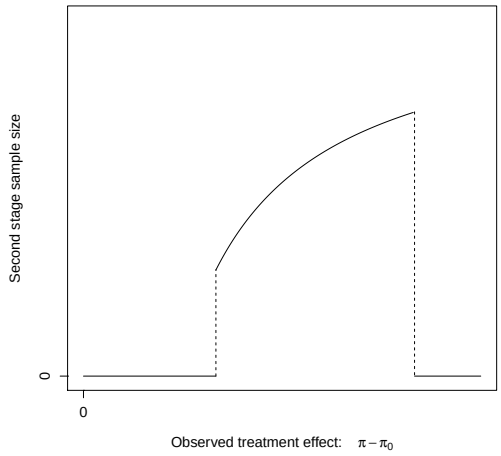
$n_1 = 13$				$n_1 = 21$				$n_1 = 20$				$n_1 = 18$			
k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$	k	n_2	$D(k)$	$l_2(k)$
≤ 4	0	0	0	≤ 7	0	0	0	≤ 5	0	0	0	≤ 6	0	0	0
5	20	.113	13	8	23	.055	18	6	15	.015	15	7	23	.055	17
6	33	.161	18	9	22	.092	18	7	16	.026	15	8	29	.129	19
7	43	.192	22	10	14	.219	15	8	16	.074	15	9	21	.277	16
8	50	.218	25	≥ 11	0	1	0	9	16	.175	15	≥ 10	0	1	0
9	56	.305	27					10	16	.340	15				
10	61	.470	28					11	16	.550	15				
≥ 11	0	1	0					≥ 11	0	1	0				

pothesis. As observed in the restricted case, the second-stage sample size increases with the number of responses observed in the first stage. This effect becomes more prominent for unrestricted designs. In example 2, the sample size of the second stage increases up to $n_2 = 68$ with 8 observed responses in the first stage and decreases only slightly for 9 observed responses. The greater search parameter space of unrestricted designs results in average sample sizes that are somewhat lower compared to the restricted case. For the considered examples, the average sample size of the unrestricted designs are $\mathrm{EN}(\pi_0) = 14.37$, 19.72 and 22.84 compared to the values given for the restricted case

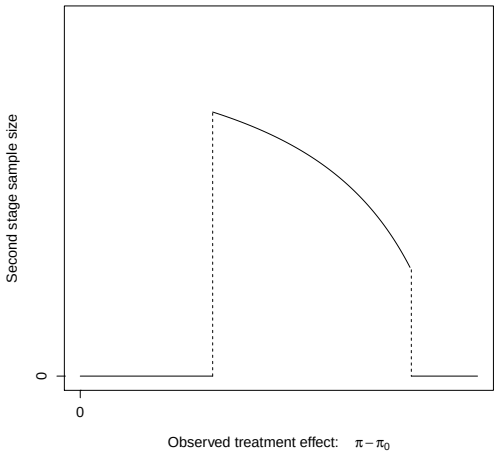
in Table 5.1 of $\mathrm{EN}(\pi_0) = 14.72$, 19.80 and 23.02, respectively. Similar findings were made by Banerjee and Tsiatis (2006), when they compared restricted to unrestricted designs. Again, the Bayesian decision-theoretic construct applied by Banerjee and Tsiatis left room for improvement and our unrestricted designs show smaller average sample sizes then the designs by Banerjee and Tsiatis, which have average sample sizes of $\mathrm{EN}(\pi_0) = 14.48$, 20.07 and 23.21, respectively. For designs that are optimal with respect to the alternative hypothesis, the sample size pattern is reversed and the second-stage sample size decreases with the number of responses observed in the first stage. Here, lower sample sizes for higher number of responses are favorable, as these outcomes are more likely under the alternative hypothesis and contribute more to the average sample size under H_1. The optimal design shows regions, where the trial is stopped after the interim analysis for efficacy and futility. In minimax designs the second-stage sample sizes lie close together and the complete design shows a layout that is similar to a fixed group-sequential case. In all examples we considered, the first-stage sample size and the maximum sample size coincide with those of designs with a fixed second stage sample size (see Table 4.6). The maximum sample size of these designs lies below the sample size needed for a one-stage fixed sample design. For example 1, 2 and 3, the maximum sample size amounts to $n = 23$, $n = 32$ and $n = 36$, whereas an exact one-sample binomial test requires a sample size of $n = 25$, $n = 35$ and $n = 39$, respectively. For continuous test statistics the one-sample design is the optimal design with respect to maximum sample size (Wang and Tsiatis, 1987). With discrete test statistics, the fixed sample size designs are, however, conservative. The designs presented here have, with the timing of the interim analysis and sample size scheme, more free parameters and thus allow for a more stringent exhaustion of the type I and II error rate making a decrease in total sample size possible. The designs that minimize the sum of $\mathrm{EN}(\pi_0)$ and $\mathrm{EN}(\pi_1)$ show greater symmetry compared to the situation when the design is optimized only with respect to one criteria. It uses the highest sample size when the number of responses is in the middle, away from both the boundaries to stop for futility or efficacy. This shape is the sample size layout that should be desired for adaptive tests according to Pong and Chow (2010, Chapter 5) and is similar to the layout found by Levin et al. (2012) for continuous test statistics.

For the investigated optimality criteria, we obtained similar patterns of second-stage sample sizes for all other parameter constellations we considered. Figure 5.2 summarizes qualitatively the layout of optimal designs with respect to (a) the null hypothesis, (b) the alternative hypothesis, (c) the maximum sample size and (d) the sum of null and alternative hypothesis.

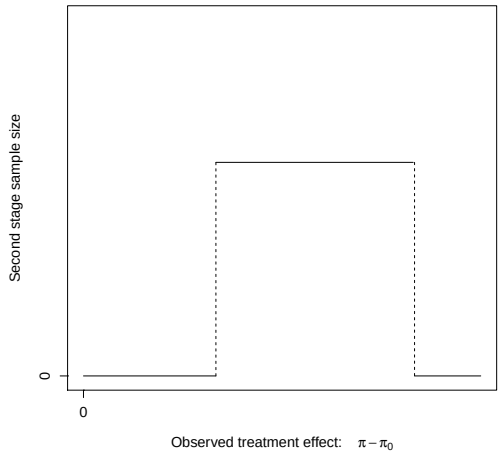
In this chapter, we considered sample size recalculation rules that lead to an overall performance optimization. It should be mentioned that the achieved gains in average sample size are solely due to the new representation of adaptive phase II designs by means of



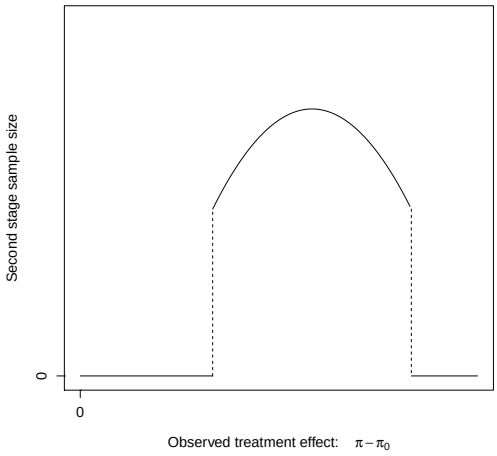
(a) *Optimal choice of sample size for stage two under the null hypothesis*



(b) *Optimal choice of sample size for stage two under the alternative hypothesis*



(c) *[Optimal choice of sample size for stage two for minimizing the maximum sample size*



(d) *Optimal choice of sample size for stage two under the mean of the null and alternative hypothesis*

Figure 5.2.: *Sample size scheme of optimal adaptive designs for different optimization criteria*

the discrete conditional error function and due to the improved search strategy. All these designs keep the type I and II error rates at the prefixed levels.

Recently, Jin and Wei (2012) proposed an adaptive design based on Simon's two-stage optimal design that determines the sample size of stage two by conditional power arguments. Here, the gains observed in average sample size come at the price of a reduced overall power. In fact, if we allow in our design the same reduction in overall power, designs with a smaller average sample size as compared to the designs by Jin and Wei (2012) can be found. By construction, every other adaptive phase II design for the same test problem will show an equal or higher average sample size as compared to our proposed approach. Therefore, our work finally solves the open problem of finding the ultimate optimal design.

We would like to note that the discrete conditional error function representation of phase II designs allows much more flexible sample size recalculation strategies than these *predefined* recalculation rules considered here. Alternative data-driven sample size adjustment rules may be advisable if, for example, there is high uncertainty in specifying the treatment effect in the planning phase. In the next chapter, the characteristics of various sample size recalculation strategies are investigated.

Uncertainty is the rule in medicine.

(William Osler 1849 – 1919*)*

6

Evaluating the Performance of Flexible Phase II Designs

A major requirement that led to the development of group-sequential designs is their ability to reduce the average number of patients needed per trail. Group-sequential designs examine the accumulating data at defined time points and thus allow early termination of a clinical trial. Later, flexible designs have been proposed that allow more far-reaching changes to the study layout based on the results of an interim analysis. All these methods can improve the efficiency of clinical trials, either by stopping a hopeless or an effective trial early or by adjusting the sample size of an ongoing trial depending on the interim results. Until today, several proposals have been made for efficient adaptive and flexible designs applying a continuous endpoint (Jennison and Turnbull, 2006; Shih, 2006; Liu et al., 2008; Bauer and König, 2006; Levin et al., 2012). Jennison and Turnbull (2006) proposed a design where the experimenter starts with a group-sequential design powered for a conservative effect with the option for early termination in order to generate satisfactory design characteristics and average sample sizes for greater effect sizes. Liu et al. (2008) considered designs with sufficient power over a wide interval of possible values of the true response rate and a sample size close to the *ideal* sample size, i.e., the one needed in a single stage design. Bauer and König (2006) among others investigated the efficiency of a reassessment of trial perspective from interim data.

In the setting of single arm phase II designs in oncology, all designs presented so far aim at optimizing the design with respect to a certain criterion. We especially considered designs that minimize either the average sample size (optimal designs) or the total sample size (minimax designs). Applying fixed rules, these designs neglect an important option of

flexible designs: Flexible designs offer the opportunity to make changes *ad hoc* in response to interim data. This wide-ranging room for design changes makes it difficult to evaluate the performance of flexible designs.

Consider a clinical trial, where after the interim analysis an adaptation was performed. At least conceptually, one may ask then for the recalculation rule that would have been applied under all other possible outcomes k in the first stage. Then the performance of the specific rule used, which is only a special one out of the universe of all flexible monitoring schemes, is identical to the one of the pre-specified adaptive design defined by these rules. Thus, in evaluating some pre-specified adaptation rules, it is possible to learn what flexible designs, which are not fully pre-specified, may offer. For pre-specified adaptation rules, we can write out exactly the sampling scheme and we can numerically calculate and compare operating characteristics for group-sequential and adaptive designs.

In this way, we now investigate how the gained flexibility developed in Chapters 4 and 5 can improve the efficiency of single-arm phase II trials under different scenario settings. Moreover, we will see whether a recalculation based on the interim results leads to favorable characteristics, or whether a fixed group-sequential design should be used, i.e., whether the possibility for recalculation should better be neglected. We first present different approaches to evaluate the efficiency of adaptive phase II designs in general, followed by a comprehensive comparison of different design variants and recalculation strategies.

6.1. Methodology for evaluating the efficiency of group-sequential and adaptive designs

Efficiency of designs may be evaluated in a number of different ways. The statistical power, i.e., the probability that the test will reject the null hypothesis when the null hypothesis is false, is a general efficiency indicator for all statistical tests. As group-sequential designs allow early stopping of the trial before all patients are recruited, the expected sample size, i.e., the average sample size needed for each trial in a series of experiments, is a frequently applied method for assessing the efficiency of group-sequential designs. For phase II designs, the overall power $1 - \beta'(\pi)$ and the average sample size $\mathrm{EN}(\pi)$ at a particular alternative π can be written as a weighted average of the conditional powers or total sample sizes of the trial for all different interim outcomes, respectively.

$$1 - \beta'(\pi) = \sum_{k=0}^{n_1} \Pr_\pi(\text{Reject } H_0 \mid k) \cdot b(k; n_1, \pi)$$

and

$$\mathrm{EN}(\pi) = \sum_{k=0}^{n_1} \big\{ n_1 + n_2(k) \big\} \cdot b(k; n_1, \pi),$$

where b denotes the binomial distribution function. In phase II designs in oncology, patient resources are usually limited and thus the maximum total sample size needed, $\max\big(n(k)\big)$, is an additional indicator of efficiency.

The practioner may refuse to accept long-term arguments on overall power and average sample size which summarize the possible interim outcomes. Instead, he may be interested in the impact of a specific sample size rule that – given the observed interim data – he wants to apply in his trial. If it turns out that the true treatment effect was overestimated but is still clinically relevant, the original study is underpowered. In this situation, an increase in sample size may be justifiable and efficient, as planning the study with a smaller treatment effect would likewise have required a greater sample size. If, however, the true treatment effect was overestimated, a smaller sample size would have been sufficient and too many patient resources are spent. To evaluate the performance of flexible sample size designs we must reasonably account for their recalculation possibilities.

Liu et al. (2008) proposed to evaluate the performance of designs over an interval and not only at an isolated point. Given the interval $[\pi_l, \pi_u]$, $\pi_l < \pi_u$, of parameters of interest, they proposed an average performance score (APS), which is defined as

$$\mathrm{APS} = \int_{\pi_l}^{\pi_u} R(\pi)\, \omega(\pi)\, d\pi, \tag{6.1}$$

where $\omega(\pi)$ is a weight function on $[\pi_l, \pi_u]$ and $R(\pi)$ a performance indicator.

The performance marker developed by Liu et al. (2008) for continuous endpoints compares the random sample size of adaptive designs with the sample size needed in a single stage design for the same setting. It allows for an objective judgment of whether a sample size increase/decrease is justified. However, continuous data are required to apply this measure. Therefore, we now develop a similar performance indicator that fits to binary outcomes and adequately addresses the recalculation possibilities. For a given type I error rate α, null hypothesis $H_0 : \pi = \pi_0$ and true response rate π, the one-stage design to achieve a nominal power of $1 - \beta$ is the design with approximate sample size (see, for example, Chow et al., 2008, p. 88):

$$n_{1-\beta}(\pi) = \left(\frac{z_{1-\beta}\sqrt{\pi(1-\pi)} + z_{1-\alpha}\sqrt{\pi_0(1-\pi_0)}}{\pi - \pi_0} \right)^2, \tag{6.2}$$

where z_γ denotes the γ-quantile of the standard normal distribution.

Before we can propose a performance score $R(\pi)$ for binomial data, some further terms need to be specified. We define the ratio of the study sample size N to the sample size of the one-stage design as

$$\mathrm{SR}(N \mid \pi) = \frac{\text{Study sample size}}{\text{Sample size of the one-stage design}} = \frac{N}{n_{1-\beta}(\pi)},$$

the relative oversize (ROS) as

$$\mathrm{ROS}(\pi) = E[\mathrm{SR}(N \mid \pi) - 1]_+,$$

where $[x]_+$ is the unit step function with $[x]_+ = x$ for $x \geq 0$ and 0 otherwise, and the relative underpower (RUP) as

$$\mathrm{RUP}(\pi) = \frac{[n_{1-\beta}(\pi) - n_{pow}(\pi)]_+}{n_{1-\beta}(\pi) - n_{0.8\cdot(1-\beta)}(\pi)},$$

where $n_{pow}(\pi)$ denotes the ideal one-stage sample size with power set equal to the power value of the given adaptive design at the true treatment difference π.

ROS measures by how much the final study sample size exceeds the ideal sample size, i.e., the sample size needed for a one-stage study. The expectation is taken because the final sample size N is a random variable in group-sequential and flexible designs and depends on the interim data. RUP compares the sample size needed for a one-stage design to achieve a power equal to the overall power of the applied adaptive design with the ideal sample size. Therefore, ROS and RUP measure if the sample size ratio stays close to unity and the power stays close to the targeted power. As in Liu et al. (2008) for continuous endpoints, the study is 100% oversized if the final sample size is twice the ideal sample size and 100% underpowered if the power of the procedure is 80% of the targeted power. Finally, the performance function $R(\pi)$ is defined as the sum of the relative oversize and relative underpower

$$R(\pi) = \mathrm{ROS}(\pi) + \mathrm{RUP}(\pi). \tag{6.3}$$

With (6.1) and (6.3) we are now able to adequately evaluate the performance of phase II designs over the interval $[\pi_l, \pi_u]$.

6.2. Framework for the comparison

We are particularly interested in the question, how different sample size recalculation rules based on the interim results perform in comparison to fixed group-sequential designs, where irrespective of the first-stage outcome the second stage is performed with a fixed

(pre-)specified sample size n_2. For recalculation, we consider rules based on conditional power and recalculation scenarios that minimize certain characteristics of the design, for example, the average sample number under the null hypothesis or the total sample size.

The conditional power is defined as the probability of rejection of the null hypothesis for the alternative $H_1' : \pi = \pi'$ given the data accumulated so far:

$$\mathrm{CP}(\pi' \mid \text{current data}) = \Pr_{H_1'}(\text{Reject } H_0 \mid \text{current data}).$$

If the study is stopped early for futility or efficacy, the conditional power equals zero or one, respectively. Note that with $\pi' = \pi_0$ the conditional power equals the conditional significance level. If the sample size is recalculated based on conditional power at an interim analysis, n_2 is chosen as the minimum integer where $\mathrm{CP}(\pi')$ is greater than a specified boundary. One possible way for recalculations based on condition power is to require a conditional power for the next stage that is equal to the planned power $1-\beta$. The unconditional power of the test is then the weighted average over all conditional powers

$$1 - \beta'(\pi') = \sum_{k=0}^{n_1} \mathrm{CP}(\pi' \mid k) \cdot b(k; n_1, \pi'), \tag{6.4}$$

where the sufficient statistic k, the number of responses observed in the first stage, summaries the data at the interim analysis.

Note that as the study can also be stopped early, recalculation based on conditional power does not necessarily guarantee that the overall power will be achieved. When investigating recalculation based on conditional power, we consider the scenarios of recalculation performed with the originally assumed effect ($\pi' = \pi_1$), recalculation with the observed effect ($\pi' = \pi_{\mathrm{obs}} = k/n_1$) and recalculation with a given fixed (external) effect ($\pi' = \pi_{\mathrm{ext}}$). Recalculation rules can, of course, be combined or restricted to certain scenarios that might happen after the first stage. It might, for example, be realistic to recalculate the sample size only if the boundary for early stop for efficacy was missed by only one response and to continue otherwise with the preplanned design.

Sample size recalculation will be performed for the flexible and more efficient phase II design presented in Section 4.4. Different recalculation rules applied for this design will be compared with the classical group-sequential Simon's designs presented in Section 2.1, the response adaptive phase II designs of Chapter 5 (with restricted maximum total sample size) and the per-design adaptive designs developed for phase II cancer trials by Lin and Shih (2004). In Lin and Shih's design, the layout depends on the results of the interim analysis. Depending on the number of responses in the first stage, the study is powered for a skeptic alternative $H_{11} : \pi = \pi_{11}$, $\pi_{11} > \pi_0$ or an optimistic target response rate $H_{12} : \pi = \pi_{12}$, $\pi_{12} \geq \pi_{11}$. The explicit study layout is as follows: With less or equal

to s_1 responses observed in the first stage, the trial is stopped early for futility. With $s_1 < k \leq r_1$ the study continues with $m - n_1$ additional patients and the null hypothesis is rejected when the total number of responses is greater than s, where k denotes the number of observed responses in the first stage out of n_1 patients. With $k > r_1$ the study continues with $n - n_1$ additional patients and the null hypothesis is rejected when the total number of responses is greater than r. The design is therefore determined by seven parameters $(s_1, r_1, n_1, s, m, r, n)$. Lin and Shih suggested selecting $(s_1, r_1, n_1, s, m, r, n)$ such that the study is powered for $1 - \beta_1$ for the skeptic target response rate π_{11} and powered for $1 - \beta_2$ for the optimistic target response rate π_{12}, $\pi_{12} \geq \pi_{11}$. Specifying $\beta_1 = \beta_2 = \beta$ and $\pi_{11} = \pi_{12} = \pi_1$ the designs by Lin and Shih results in a generalization of the designs by Simon, i.e., the study is powered for $1 - \beta$ for the alternative $H_1 : \pi = \pi_1$. The difference to the classical Simon's design is that generally two different second-stage sample sizes are allowed depending on the interim results. Lin and Shih (2004) describe four different optimization rules to choose $(s_1, r_1, n_1, s, m, r, n)$. They proposed to minimize $\mathrm{EN}(\pi_0)$ (Optimal Type 1), $\max\big(\mathrm{EN}(\pi_0), \mathrm{EN}(\pi_{11}), \mathrm{EN}(\pi_{12})\big)$ (Optimal Type 2), $\max(n, m)$ and $\mathrm{EN}(\pi_0)$ (Optimal Type 3), or $\max(n, m)$ and $\max\big(\mathrm{EN}(\pi_0), \mathrm{EN}(\pi_{11}), \mathrm{EN}(\pi_{12})\big)$ (Optimal Type 4). Designs of Optimal Type 1 are thereby an extension of Simon's optimal designs and designs of Optimal Type 3 are an extension of the minimax designs. We used these two types of designs to allow for a fair comparison with other optimal and minimax designs.

6.3. Performance comparison

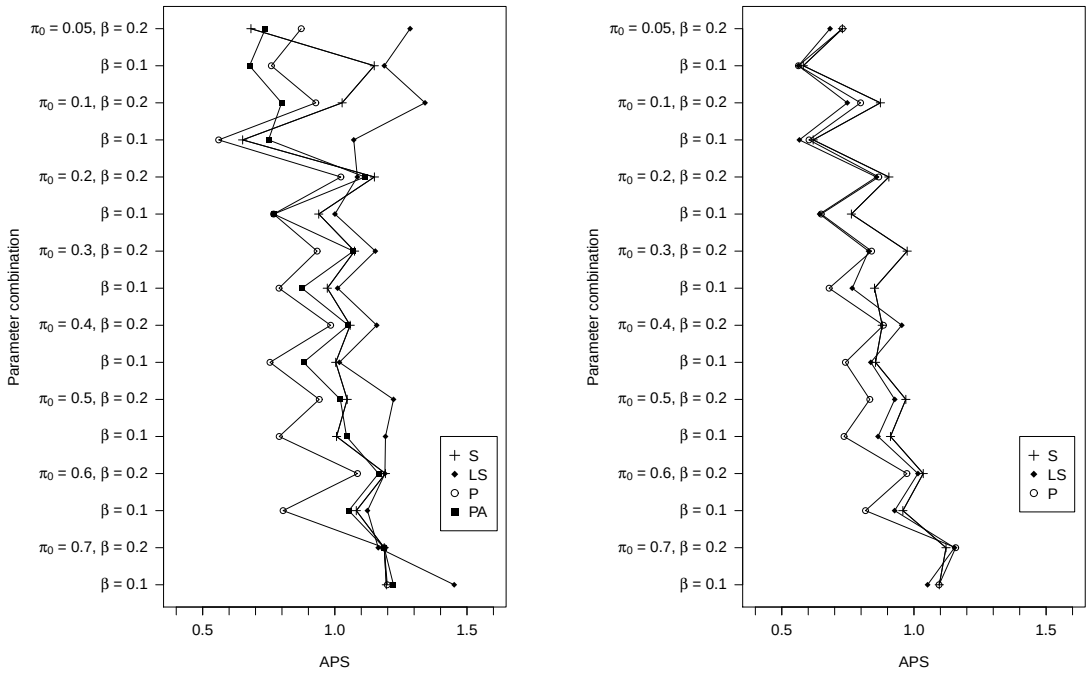
The significance level was fixed to $\alpha = 0.05$ for all comparisons. All fixed designs were determined to achieve a power of $1 - \beta$ for a treatment effect of $\pi_1 - \pi_0 = 0.2$. Therefore, in the comparison of different phase II designs all designs satisfy the same type I and II error rate constraints ensuring that these parameters won't affect the performance comparison. We have previously shown that all designs can be expressed equivalently in the discrete conditional error function framework presented in Chapter 4. We used this framework both for the recalculation and for the evaluation of the designs. Adaptive designs are often criticized for the use of insufficient test statistics associated with a loss of efficiency (Dette et al., 2012; Tsiatis and Mehta, 2003; Jennison and Turnbull, 2003). Evaluation of fixed designs, where no recalculation is performed, based on the discrete conditional error function assures that the evaluation strategy does not affect the performance comparison.

We evaluate the average performance score (6.1) of the considered designs in the interval $[\pi_1 - 0.1, \pi_1 + 0.1]$, i.e., in the case that the true treatment effect was under- or overestimated by at most 10%. The true treatment effect is assumed to be equally likely in the given interval and consequently a uniform weight function ($w \equiv 1$) is used. The performance

score $R(\pi)$ was calculated at 20 equally spaced response rates π within the interval. With the uniform weight function, the APS value can then be approximated by the mean of those performance scores.

6.3.1. Performance comparison of designs applying fixed rules

Figure 6.1 gives the results of the performance score comparison between Simon's design (S), Lin and Shih's design (LS), the design proposed in Section 4.4 (P) and the response adaptive design proposed in Chapter 5 (PA). The left table displays for different parameter settings the characteristics of optimal designs, whereas the right presents minimax designs. The proposed response adaptive designs were included only for optimal designs. Investigating adaptive designs in Chapter 5, we have seen that a fixed second-stage sample size is the optimal choice for minimizing the maximum sample size. Therefore, the average performance score values for minimax group-sequential designs (P) also apply to minimax adaptive designs (PA).



(a) *Average performance scores (APS) of optimal phase II designs*

(b) *Average performance scores (APS) of minimax phase II designs*

Figure 6.1.: *Performance comparison of designs with fixed rules (S = Simon, LS = Lin and Shih, P = Proposed, PA = Proposed adaptive)*

A direct comparison between minimax and optimal designs shows that minimax designs tend to have smaller average performance scores. This can be explained as follows: In Simon's design, the study sample size is a random variable with two outcomes

$$N = \begin{cases} n_1 & \text{if the study is stopped early} \\ n & \text{else.} \end{cases}$$

Since in general $n_1 \leq n_{1-\beta}(\pi)$ the ROS is calculated as

$$\text{ROS}(\pi) = \big(1 - \text{PET}(\pi)\big)\Big[\frac{n}{n_{1-\beta}(\pi)} - 1\Big]_+,$$

where PET denotes the probability for early termination (2.1). Designs with a greater n_1 and a smaller n will lead to smaller performance score values, since ROS penalizes only sample sizes higher than the ideal sample size $n_{1-\beta}(\pi)$. Therefore, minimax designs that usually use a greater n_1 and smaller n tend to lead to smaller ROS values than optimal designs. Similar considerations can be made for Lin and Shih's design and the proposed designs.

Per construction in Section 4.4 and Chapter 5, the proposed optimal designs outperform Simon's optimal designs with respect to average sample size under the null hypothesis and the minimax designs with respect to maximum sample size. From Figure 6.1(a) and (b) we see that the proposed designs are in most cases also superior with respect to the average performance score, i.e., they are superior over an interval of response rates. This is not true for the optimal design by Lin and Shih. As this design includes Simon's design as a special case, the average sample sizes under the null hypothesis are smaller. However, with respect to the average performance score, Simon's optimal design outperforms the design by Lin and Shih (2004). A direct comparison (see Table 6.1) of the proposed design with the design by Lin and Shih may explain these differences. The decrease in average sample size under the null hypothesis $\text{EN}(\pi_0)$ achieved by the Lin and Shih (2004) design is paid by a marked increase in maximum and average sample size under the alternative hypothesis $\text{EN}(\pi_1)$ and in the average performance score. The same findings hold true for the optimal adaptive designs presented in Chapter 5. By allowing the second-stage sample size to depend on the interim outcome a reduction in average sample size under the null hypothesis is possible. This, however, leads to an increase in average sample size under the alternative hypothesis and, in case of a high treatment effect, to an increase of the average performance score.

For minimax designs, Lin and Shih's design is superior in average performance score to Simon's design in most cases, but still less efficient than the proposed method in case of higher response rates. Comparison of standard design characteristics as in Table 6.2 show very similar characteristics for Lin and Shih's design and the proposed method. Maximum sample sizes of the minimax designs are equal, except for one case where the

Table 6.1.: *Performance comparison of optimal phase II designs with fixed rules (LS = Lin and Shih, P = Proposed)*

π_0	β	n_{LS}	$\text{EN}(\pi_0)_{LS}$	$\text{EN}(\pi_1)_{LS}$	n_P	$\text{EN}(\pi_0)_P$	$\text{EN}(\pi_1)_P$
0.05	0.2	29	10.80	21.09	21	11.17	15.72
	0.1	39	16.59	33.79	30	16.75	24.94
0.1	0.2	37	14.80	30.12	29	14.98	23.31
	0.1	46	21.82	41.68	41	22.19	24.47
0.2	0.2	42	20.20	35.00	43	20.54	35.01
	0.1	57	29.29	53.14	53	30.05	44.31
0.3	0.2	50	23.27	42.86	46	23.52	36.65
	0.1	65	33.57	60.75	59	34.12	51.23
0.4	0.2	53	24.26	45.41	46	24.49	39.78
	0.1	67	34.83	61.75	66	35.80	51.60
0.5	0.2	55	22.99	43.75	43	23.40	35.78
	0.1	74	33.13	65.60	59	33.47	48.89
0.6	0.2	42	19.96	36.30	37	20.42	34.01
	0.1	56	28.37	52.16	52	28.99	41.95
0.7	0.2	32	14.66	24.91	27	14.82	24.60
	0.1	46	20.75	39.02	36	20.92	34.63

Table 6.2.: *Performance comparison of minimax phase II designs with fixed rules (LS = Lin and Shih, P = Proposed)*

π_0	β	n_{LS}	$\text{EN}(\pi_0)_{LS}$	$\text{EN}(\pi_1)_{LS}$	n_P	$\text{EN}(\pi_0)_P$	$\text{EN}(\pi_1)_P$
0.05	0.2	17	11.89	13.20	16	13.76	13.44
	0.1	24	20.42	19.16	24	20.42	19.16
0.1	0.2	23	19.92	19.03	23	19.20	20.24
	0.1	32	27.13	24.73	32	27.95	26.99
0.2	0.2	32	23.72	29.46	32	23.24	29.94
	0.1	44	34.23	37.37	44	33.43	37.57
0.3	0.2	36	29.04	33.08	36	29.32	33.56
	0.1	50	41.09	48.83	50	41.06	44.45
0.4	0.2	39	27.17	37.79	39	27.14	35.81
	0.1	53	43.21	52.51	53	42.68	49.25
0.5	0.2	37	26.97	35.18	37	26.90	32.15
	0.1	51	38.93	49.91	51	37.74	45.46
0.6	0.2	33	22.87	31.91	33	23.22	31.33
	0.1	45	31.42	43.56	45	31.52	41.47
0.7	0.2	25	17.71	24.05	25	18.05	24.59
	0.1	32	22.65	31.16	32	22.66	31.61

proposed minimax design needs one patient less. In half of the cases, the proposed method is superior with respect to maximum or average sample size under the null hypothesis, even though the proposed design does not allow to choose between two but only a single second-stage sample size.

6.3.2. Performance comparison of different recalculation rules based on conditional power

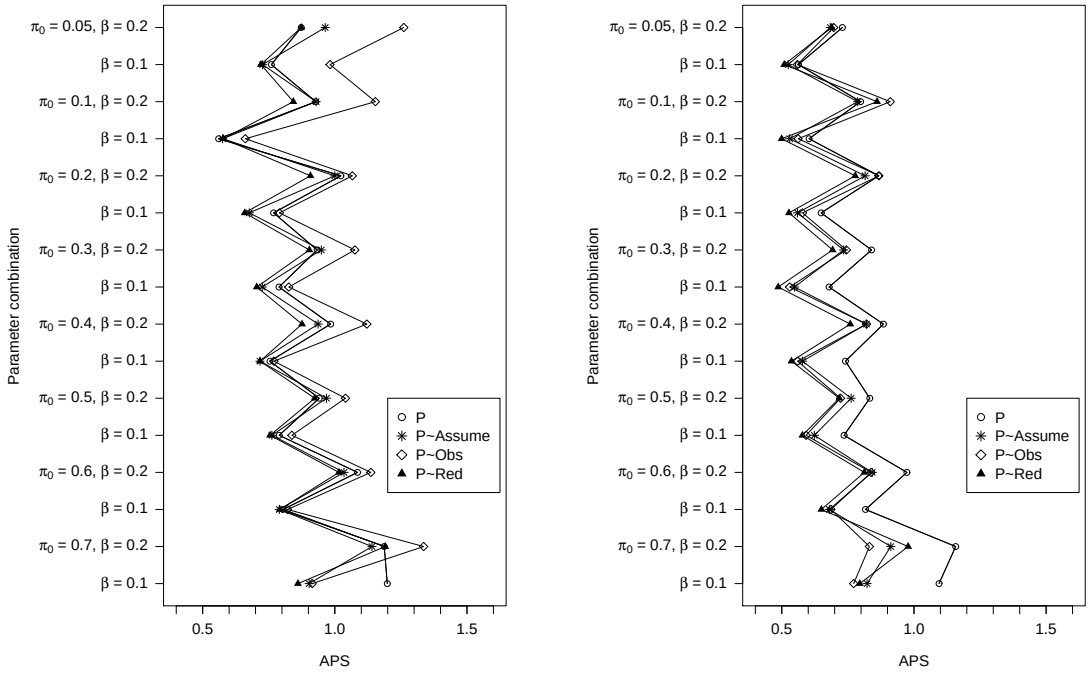
We now investigate *whether* sample size adjustment is advisable at all and if this is true, *what* data-driven sample size adjustment rules may be recommended if, for example, there is high uncertainty in specifying the treatment effect in the planning phase. In Chapter 4, we demonstrated how each classical phase II oncology design can directly be transferred into a flexible design. In the preceding section, we have seen that the proposed flexible design is superior to Simon's and Lin and Shih's design with respect to both the optimization criterion and average performance score, when no recalculation is performed. For optimal designs, it was in most cases also superior to the adaptive design. Therefore, we use the proposed flexible design (P) both as benchmark and as start design in the comparison of different recalculation rules. Given the interim results, the sample size of this start design is recalculated. Details on how recalculation is carried out within the proposed flexible design are given in Chapter 4.

The considered recalculation scenarios aim at achieving a conditional power of $1 - \beta$ (a) based on the assumed effect (P∼Assume), (b) based on the observed effect (P∼Obs) and (c) based on a reduced effect of $\pi' = \pi_1 - 0.05$ (P∼Red). Recalculated sample sizes were truncated at $2n$. The fixed scenario with no recalculation, i.e., the proposed design P, is included for comparison. For each scenario, the average performance score was calculated. Results are given in Figure 6.2.

In optimal designs, recalculation based on the observed effect shows least favorable characteristics with respect to APS. Recalculation based on the assumed or a reduced (external) effect lead to similar average performance values. The APS of the fixed design without flexible sample size adjustment lies in-between the considered recalculation strategies in most cases.

For minimax designs, recalculation based on the assumed, the observed and a reduced effect lead to similar performance score values. Here the fixed design is outperformed by all these recalculation rules for high null response rates π_0.

More insight in the distinct nature of these three recalculation strategies can be gained by considering the performance score for different assumed response rates. Among our



(a) *Average performance scores (APS) of proposed optimal phase II designs*

(b) *Average performance scores (APS) of proposed minimax phase II designs*

Figure 6.2.: *Performance comparison of different recalculation rules (P = Proposed without recalculation, P∼Assume = recalculation based on conditional power on the assumed effect, P∼Obs = recalculation based on conditional power on the observed effect, P∼Red = recalculation based on a reduced effect of* $\pi' = \pi_1 - 0.05$*)*

parameter choices, the minimax design for $(\pi_0, \beta) = (0.3, 0.1)$ leads to the smallest average performance scores. Figure 6.3 plots the performance score $R(\pi) = \text{ROS}(\pi) + \text{RUP}(\pi)$ for the range of true treatment effects $[\pi_1 - 0.1, \pi_1 + 0.1]$. For each curve, the area under the curve represents the average performance score. The gray area highlights the $\text{ROS}(\pi)$ part. The fixed design has the smallest performance score for the planned treatment effect of $\pi_1 = 0.5$. If the true treatment effect differs from the planned one, the fixed design is oversized for greater effects and underpowered for smaller effects. In both cases, the performance score increases. This is characterized by the wedge-shaped curve of the performance score in Figure 6.3. It can be seen that for all re-estimation methods the performance functions are more flat and have relatively low values over a broader interval. This corresponds to designs with a good balance in used sample size and achieved statistical power for these effect sizes. Designs with recalculated sample size reduce the efficiency of the design for the planned response rate but are more efficient if there is uncertainty in the treatment effect.

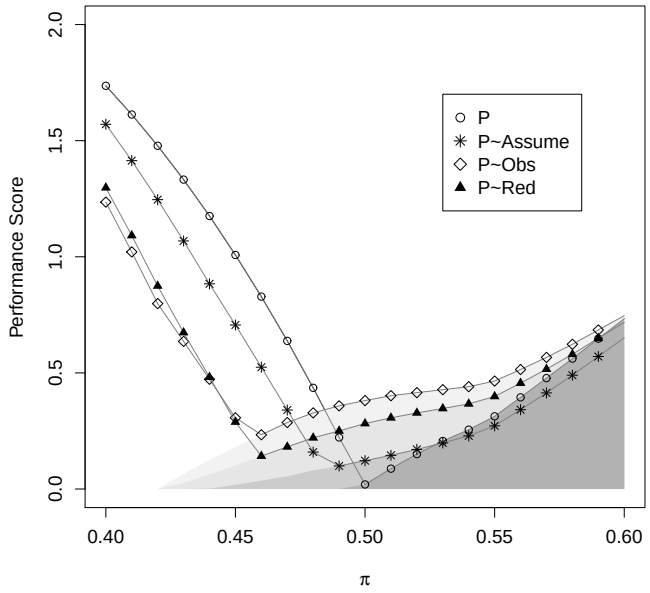


Figure 6.3.: *Performance score for proposed minimax design* $(\pi_0, \beta) = (0.3, 0.1)$ *(P = Proposed without recalculation, P∼Assume = recalculation based on conditional power on the assumed effect, P∼Obs = recalculation based on conditional power on the observed effect, P∼Red = recalculation based on a reduced effect of* $\pi' = \pi_1 - 0.05$*). Shaded areas are explained in the text.*

For recalculation with the assumed effect, low values of the performance score are located around the assumed effect of $\pi_1 = 0.5$. If the sample size is recalculated based on the observed effect, the performance function curve is more flat. However, in absolute values they are lower only for values $\pi \leq \pi_1 - 0.05$. The complete performance score curve is shifted into the direction towards lower treatment effects if recalculation is not based on condition power for the assumed effect but for a reduced effect. In this design, more efficiency for lower treatment effects is achieved. This resulted, however, in less efficiency for the originally assumed treatment effect. This explains why in all investigated combinations recalculation based on the assumed effect and the reduced effect lead to similar average performance scores (see Figure 6.2). Both strategies have different focus and, apparently, do a good job. While there is no difference in the average performance score, recalculation based on a reduced effect increases the overall power of the design if the true treatment effect is lower than assumed. The overall power for a reduced effect of $\pi' = \pi_1 - 0.05$ can be calculated by (6.4). In Table 6.3 and 6.4 we give the overall power in case the sample size is recalculated based on the assumed effect $(1 - \beta(\pi')_{\text{P}\sim\text{Assume}})$

Table 6.3.: *Power comparison of proposed optimal phase II designs for different recalculation rules based on conditional power*

π_0	β	$1-\beta(\pi')_{\text{P}\sim\text{Assume}}$	$1-\beta(\pi')_{\text{P}\sim\text{Red}}$	Δ_{Power}
0.05	0.2	0.58	0.66	0.08
	0.1	0.73	0.79	0.06
0.1	0.2	0.56	0.63	0.07
	0.1	0.75	0.80	0.05
0.2	0.2	0.52	0.60	0.08
	0.1	0.71	0.78	0.07
0.3	0.2	0.53	0.61	0.08
	0.1	0.68	0.78	0.09
0.4	0.2	0.53	0.62	0.08
	0.1	0.68	0.76	0.08
0.5	0.2	0.52	0.62	0.10
	0.1	0.67	0.76	0.10
0.6	0.2	0.51	0.60	0.09
	0.1	0.65	0.76	0.11
0.7	0.2	0.49	0.63	0.15
	0.1	0.61	0.72	0.11

and based on a reduced effect $(1-\beta(\pi')_{\text{P}\sim\text{Red}})$. We also included the difference in powers $\Delta_{\text{Power}} := 1-\beta(\pi')_{\text{P}\sim\text{Red}} - \big(1-\beta(\pi')_{\text{P}\sim\text{Assume}}\big)$. If recalculation is performed with the reduced effect, the overall power $1-\beta(\pi')$ increases on average by 9% for optimal designs and by 10% for minimax designs.

In most cases, recalculation does, however, not guarantee that the overall power equals the conditional power applied for sample size recalculation. The overall power is the weighted average of the conditional powers for all interim results. If the number of responses at the interim analysis is low, the study is stopped early for futility. In this case, the conditional power equals zero. Necessarily, the overall power is smaller than

$$1-\text{Pr}_{H_1'}\big(\text{Stop for futility}\big), \tag{6.5}$$

with $H_1' : \pi = \pi'$. Usually, phase II designs in oncology have a high probability to stop early for futility ($\geq 50\%$), if the null hypothesis is true or if treatment effects are assumed that lie close to the null hypothesis. According to (6.5), no recalculation strategy can achieve sufficient overall power in these cases.

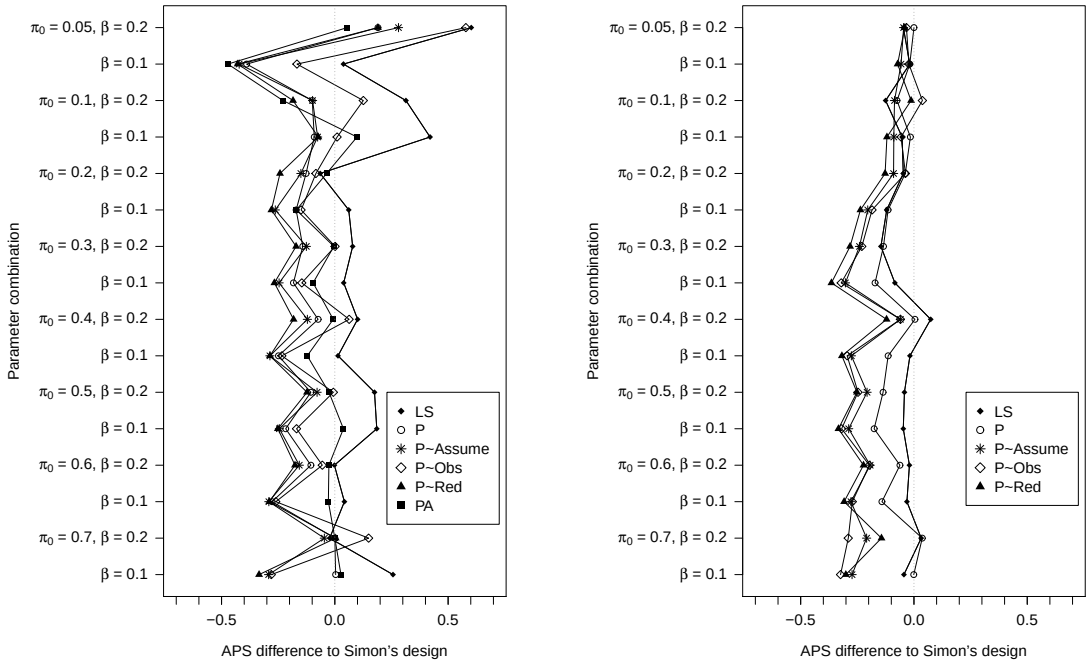
Table 6.4.: *Power comparison of proposed minimax phase II designs for different recalculation rules based on conditional power*

π_0	β	$1-\beta(\pi')_{\text{P}\sim\text{Assume}}$	$1-\beta(\pi')_{\text{P}\sim\text{Red}}$	Δ_{Power}
0.05	0.2	0.76	0.83	0.07
	0.1	0.85	0.92	0.06
0.1	0.2	0.69	0.79	0.10
	0.1	0.83	0.90	0.07
0.2	0.2	0.62	0.70	0.08
	0.1	0.78	0.87	0.09
0.3	0.2	0.66	0.76	0.10
	0.1	0.79	0.88	0.08
0.4	0.2	0.60	0.70	0.11
	0.1	0.78	0.85	0.07
0.5	0.2	0.62	0.72	0.11
	0.1	0.74	0.85	0.10
0.6	0.2	0.58	0.72	0.14
	0.1	0.72	0.82	0.11
0.7	0.2	0.57	0.74	0.17
	0.1	0.68	0.79	0.11

6.4. Properties compared and discussed

Figure 6.4 combines the results of Section 6.3. It illustrates the performance of different designs and different recalculation rules. Simon's design was used as reference (gray dotted line). Changes in average performance scores are given for Lin and Shih's design (LS) and for the proposed designs with and without flexible recalculation of the sample size (PA, P, P∼Assume, P∼Obs and P∼Red). For the same reasons as in Figure 6.1, the proposed adaptive design (PA) was included only for optimal designs.

For optimal designs, Figure 6.4 demonstrates that recalculation with the observed effect performs better than Lin and Shih's design, similar to the proposed adaptive design and Simon's design, but worst among all reassessment strategies considered. When an interim analysis is performed in the conduct of a flexible clinical trial, it is, however, tempting to update the effect size for which the study has been powered for in the planning phase by the interim estimate. This recalculation rule will increase/decrease the sample size if the treatment effect observed in the interim analysis is smaller/higher than assumed. This is due to two aspects: The power calculation was based on a different treatment effect and the conditional error function is small/high for the number of responses observed in the interim analysis, i.e., a small p-value is required/moderately high p-value is sufficient in the second stage to reject the null hypothesis. On first sight, this recalculation rule perfectly



(a) *Average performance score (APS) difference to Simon's design of optimal phase II designs*

(b) *Average performance score (APS) difference to Simon's design of minimax phase II designs*

Figure 6.4.: *Performance comparison of different designs/different recalculation rules. Difference of APS scores to Simon's design are shown (reference indicated by gray dotted line, LS = Lin and Shih, PA = Proposed adaptive, P = Proposed without recalculation, P∼Assume = recalculation based on conditional power on the assumed effect, P∼Obs = recalculation based on conditional power on the observed effect, P∼Red = recalculation based on a reduced effect of* $\pi' = \pi_1 - 0.05$*)*

matches to the current trial results. The treatment effect observed in the interim analysis is, however, a random estimator of the true effect size. As the first-stage sample sizes are small in phase II designs, the variation in the estimated treatment effect is high. This leads to many cases where the observed treatment effect does not well approximate the true treatment effect and where an inadequate recalculation rule is performed (Fleming, 2006). This explains the poor performance of this recalculation rule with respect to APS. This performance marker counterbalances the gain in overall power and the used sample size for a given true effect size to allow for a reasonable judgment.

In 2006, Bauer and König investigated the impact of using conditional power to reassess the sample size. For flexible two-stage combination tests applying continuous test statistics, they determined the density of the conditional power for different reassessment methods.

Although the setting and the principle approach for evaluation differs, similar conclusions were made. Bauer and König (2006) concluded that "mid-trial sample size recalculation based on an interim estimate may lead to an overly large price to be paid in average sample size in relation to the gain in overall power."

From Figure 6.4 we see that recalculation based on the originally assumed effect and recalculation based on a reduced effect show favorable characteristics. Here, the correct recalculation rule is applied if the predicted treatment effect is true. Sample size is increased or reduced if (by chance) a smaller or higher number of responses were observed in the interim analysis, respectively. However, the sample size is not increased enough to guarantee sufficient power for a smaller treatment effect and the recalculated sample size is too large for a greater treatment effect. The average performance score accounts for both of these aspects. For optimal designs, both rules are generally preferable to the design without recalculation and with recalculation based on the observed effect. Additionally, from Figure 6.4 it can be seen that they lead to APS values smaller than Simon's and Lin and Shih's optimal or minimax design and are superior with respect to this criteria. An exception is the first parameter combination for optimal designs $(\pi_0, \beta) = (0.05, 0.2)$, where Simon's design performs best. For this parameter choice the proposed flexible design, which was used as the initial design for recalculation, differs markedly from Simon's design. According to Table 4.5, the proposed flexible design uses for this parameter constellation 4 patients (24%) more. This might explain the differences observed in the average performance score.

In summary, we conclude from Figure 6.4 that recalculation of the sample size can improve the efficiency of the designs, if the there is uncertainty with respect to the true treatment effect. A flexible recalculation strategy can guarantee sufficient power if it becomes apparent throughout the trial that the assumed treatment difference was too optimistic. The price to be paid in terms of additional sample size has to be weighed against the gain in power.

A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.

(Michael Joseph Moroney)

7

Clinical Trial Example

In this chapter, we demonstrate some of the rich possibilities of the proposed methods by re-examining a single-center one-armed phase I/II cancer trial conducted by Combs et al. (2012) (PANDORA-01 trial). The aim of the PANDORA-01 trial was to evaluate the maximum tolerable dose for carbon ion radiotherapy in patients with recurrent rectal cancer that had been previously treated with radiation. In phase I, the safety of a recommended dose is determined by a dose escalation scheme. Subsequently, the effectiveness of this dose level was investigated in patients with recurrent rectal cancer. Phase II was directly included in the study to streamline the development process. The primary endpoint of this second part of the study was the 12-month progression-free survival rate π after re-irradiation. Evaluation was recorded according to the RECIST criteria (Eisenhauer et al., 2009). The one-sided null hypothesis $H_0 : \pi \leq 0.6$ was assessed at type I error rate $\alpha = 0.05$, and a type II error rate $\beta = 0.2$ at $H_1 : \pi = 0.8$ was desired.

Combs et al. calculated the fixed two-stage design based on a combination test that fulfilled the restrictions on type I and II error rate according to the algorithm given in Section 4.1. The authors selected the optimal design minimizing $\mathrm{EN}(\pi_0)$. This resulted in the following design parameters: $(n_1, n_2, \alpha_0, \alpha_1, c_\alpha) = (14, 25, 0.3, 0.03, 0.021)$. Accordingly, if the first-stage p-value p_1 (see (3.1) on page 19) is greater than $\alpha_0 = 0.3$, the study is terminated early for futility. This requirement is met if equal to or less than $l_1 = 9$ of the $n_1 = 14$ patients in the first stage show a response. With $u_1 = 13$ or more responses and, equivalently, $p_1 \leq \alpha_1$, the study is discontinued after the first stage due to the proof of efficacy. Otherwise, the study proceeds to the second stage and enrollment continues until $n_2 = 25$ additional patients are recruited. In the final analysis, the null hypothesis is rejected if the product of the one-sided p-values of the two stages, p_1 and p_2 (see (3.2) on

page 20), is equal to or lower than $c_\alpha = 0.021$; otherwise, the null hypothesis is accepted. This fixed two-stage design can be transferred to a flexible design according to Section 4.2 by the following function C:

$$C(p_1) = \begin{cases} 0 & \text{if } p_1 \geq 0.3 = \alpha_0 \\ 0.082 & \text{if } p_1 = 0.2793 \\ 0.170 & \text{if } p_1 = 0.1243 \\ 0.467 & \text{if } p_1 = 0.0398 \\ 1 & \text{if } p_1 \leq 0.03 = \alpha_1. \end{cases}$$

The PANDORA-01 trial was conducted with these adaptive conditional test boundaries. The null hypothesis is rejected after the second stage if $p_2 \leq C(p_1)$. As noted in Section 3.1, it would have been possible to directly apply flexible designs methodology developed for continuous test statistics. In Figure 7.1, the adaptive conditional test boundaries together with the rejection region of the Bauer and Köhne design for continuous test statistics are plotted for the study example. It can be seen how the proposed method assures better exhaustion of the overall level by increasing the conditional type I error rate for the second stage for each attainable first-stage p-value. If the sample size is not modified, the design assures the desired power of $1 - \beta$ for $H_1 : \pi = 0.8$ and has an average sample size (see (4.2) on page 28) of $\mathrm{EN}(\pi_0) = 20.8$. This design shows characteristics very similar to those of Simon's optimal design, whose expected sample size is 20.5 according to Table 2.1, but requires a maximum total of 43 patients.

The PANDORA-01 trial is the first trial evaluating the efficiency of carbon ion therapy for patients with recurrent rectal cancer. Hence, there is considerable uncertainty with respect to the assumed improvement in 12-month progression-free survival rate. It may therefore not be anticipated to use a fixed group-sequential design such as Simon's optimal design that does not allow for changes in the course of the trial. In Section 4.3, we demonstrated how the phase II study part of the PANDORA-01 trial could have been constructed directly as a flexible version of Simon's optimal design. Then, the conditional type I error rates $\mathrm{CE}(k)$ (see (3.5) on page 23) are used as discrete conditional error function.

$$\mathrm{CE}(k) = \begin{cases} 0 & \text{if } k \leq 7 \\ 0.116 & \text{if } k = 8 \\ 0.205 & \text{if } k = 9 \\ 0.323 & \text{if } k = 10 \\ 0.461 & \text{if } k = 11 = n_1. \end{cases}$$

The null hypothesis is rejected if the second-stage p-value p_2 satisfies $p_2 \leq \mathrm{CE}(k)$.

As the original design is conservative, i.e., $\alpha' = 0.049 < \alpha = 0.05$, the related flexible design is also conservative. We have shown in Section 4.3 how the remaining level $\alpha - \alpha'$ can be implemented to overcome the conservativeness in a flexible setting by increasing the

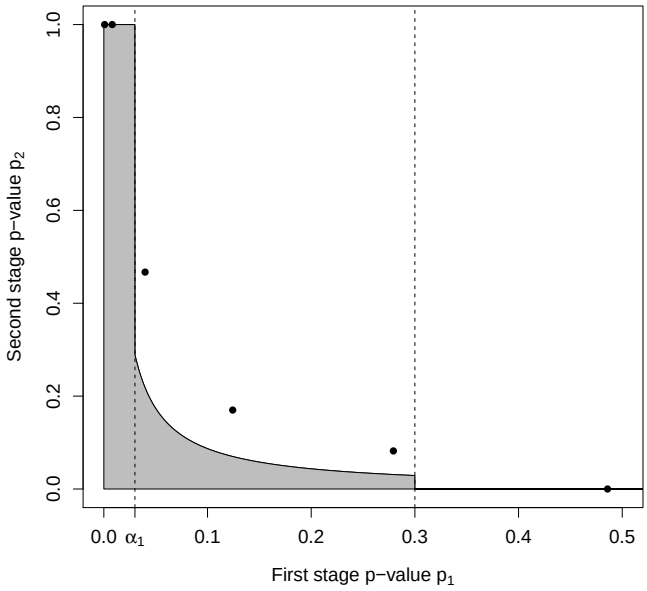


Figure 7.1.: *Rejection regions of the proposed flexible design based on combination test in terms of the observed p-values p_1 and p_2. The rejection boundaries of the proposed method are printed in bold dots. The rejection region of the Bauer and Köhne design for continuous test statistics is given for comparison (line). The figure displays only the region where the rejection boundaries are different from zero.*

boundaries CE(k) with CE(k) $\neq 0$ and CE(k) $\neq 1$. We now (1) increase the conditional type I error rates proportionally to the probability of observing p_1 (4.4), (2) distribute the remaining level $\alpha - \alpha'$ equally among the conditional error function values (4.5) and (3) increase only the smallest conditional error function value unequal to zero (4.6). The resulting discrete conditional error functions exhausting the nominal type I error rate are given below.

$$D_1\big(p_1(k)\big) = \begin{cases} 0 \\ 0.119 \\ 0.208 \\ 0.326 \\ 0.464 \end{cases} \quad D_2\big(p_1(k)\big) = \begin{cases} 0 \\ 0.117 \\ 0.208 \\ 0.332 \\ 0.528 \end{cases} \quad D_3\big(p_1(k)\big) = \begin{cases} 0 & \text{if } k \leq 7 \\ 0.121 & \text{if } k = 8 \\ 0.205 & \text{if } k = 9 \\ 0.323 & \text{if } k = 10 \\ 0.461 & \text{if } k = 11 = n_1. \end{cases}$$

If no design modifications are performed, the same decision rules are applied as for Simon's design. In addition, the resulting designs allow for flexible design changes without undermining the nominal significance level. If, for example, the sample size is changed after an interim analysis, situations exist where, due to the increased boundaries, rejection

of the null hypothesis is possible if the study was planned with $D_1\big(p_1(k),\, D_2\big(p_1(k)\big)$ or $D_3\big(p_1(k)\big)$ but not if the flexible version of Simon's design was applied. This occurs for $\mathrm{CE}(k) < p_2 \leq D\big(p_1(k)\big)$.

The proposed flexible and more efficient phase II designs of Section 4.4 allow for the same degree of flexibility but additionally show better characteristics than standard designs if no adaptations are performed. If the trial of Combs et al. (2012) with parameters $(\pi_0, \pi_1, \alpha, \beta) = (0.6, 0.8, 0.05, 0.2)$ had been constructed within this framework, $n_1 = 14$ patients would have been required in the first stage according to the optimal two-stage design given in Table 4.7. For the second stage only $n_2 = 23$ patients would have sufficed to fulfill the α and β constraints. Evaluation of the design would then have to be performed with the following discrete conditional error function D (see Table 4.7):

$$D(p_1) = \begin{cases} 0 & \text{if } p_1 > 0.2793 \\ 0.124 & \text{if } p_1 = 0.2793 \\ 0.237 & \text{if } p_1 = 0.1243 \\ 0.238 & \text{if } p_1 = 0.0398 \\ 0.390 & \text{if } p_1 = 0.0081 \\ 0.407 & \text{if } p_1 = 0.0004 \\ 1 & \text{if } p_1 < 0.0004. \end{cases} \tag{7.1}$$

This design requires a maximum of $n = 37$ patients and on average $\mathrm{EN}(\pi_0) = 20.42$ per trial. In Chapter 5, we derived these new flexible and more efficient phase II designs also for the situation that the planned second-stage sample size depends on the interim outcome. The discrete conditional error function together with the second stage sample sizes of this optimal adaptive phase II design are given in Table 7.1 (see also Table 5.2 on page 63). Here, a maximum of 44 patients and on average 19.72 per trial are needed.

A summary of the design characteristics of all flexible design variants developed for the considered clinical trial example is given in Table 7.2. Here, it is assumed that the trials were conducted as planned without flexible modification throughout the trial. All these

Table 7.1.: *Layout of the optimal adaptive design (Chapter 5) for the clinical trial example* $(\pi_0, \pi_1, \alpha, \beta) = (0.6, 0.8, 0.05, 0.2)$

		$n_1 = 10$		
k	$p_1(k)$	$n_2(k)$	$n(k)$	$D(k)$
≤ 6	0.6331	0	10	0
7	0.3823	21	31	0.096
8	0.1673	31	41	0.143
9	0.0464	31	41	0.245
10	0.0060	34	44	0.354

Table 7.2.: *Design characteristics of the different approaches for the clinical trial example* $(\pi_0, \pi_1, \alpha, \beta) = (0.6, 0.8, 0.05, 0.2)$

Design	EN(π_0)	n	α'	β'
Flexible design based on combination test (Section 4.2)	20.78	39	0.046	0.199
Flexible version of Simon's design (Section 4.3)	20.48	43	0.049	0.198
Flexible and more efficient design (Section 4.4)	20.42	37	0.050	0.199
Optimal adaptive phase II design (Chapter 5)	19.72	31-44	0.050	0.199

designs aim at minimizing the average sample size under the null hypothesis. We note that all designs control and exhaust the nominal type I error rate in a flexible setting.

The expected and maximum sample size of the flexible and more efficient design is lower by 0.06 and 6 patients, respectively, than the corresponding Simon's optimal design, and lower by 0.36 and 4 if the flexible design based on combination test is used. According to Table 7.2, the gain in performance is not associated with a loss in power but with a better use of the nominal significance level in case that no design changes are performed. The optimal adaptive phase II design that allows the second-stage sample size to depend on the interim outcome, shows an average sample size that is lower by 0.7 compared to the design of Section 4.4 and lower by 0.76 compared to Simon's design. However, the maximum possible sample size also increases by 7 patients or 1 patient, respectively. Taking into account that Simon's design has been deemed to be optimal for decades, the additional decrease in sample size which can be achieved by applying our proposed methods must be regarded as a significant contribution to the field of clinical trial design. There are still hundreds of phase II trials performed each year worldwide that use Simon's design. Application of our proposed design will prevent a considerable number of patients from being included in such trials in case of inefficient therapies under investigation (as we minimize the sample size under the null hypothesis). Noteworthy, the reduction in average sample size has not to be paid with a loss in power or conservativeness but is due to the new methodology and the efficient search algorithm.

Let us now assume the hypothetical situation that the trial by Combs et al. was planned with the proposed flexible and more efficient design of Section 4.4. Further, we assume that 10 responses were observed in the first stage, but external evidence suggests that the response rate is slightly lower than $\pi_1 = 0.8$ while the improvement may still be clinically relevant. The number of responses in the first stage translates to a p-value of

$$p_1 = \Pr_{H_0}(X_1 \geq 10) = 1 - B(9 - 1; 0.6, 14) = 0.2793,$$

with the cumulative distribution function of the binomial distribution B. According to (7.1), $D(0.2793) = 0.124$ and thus a p-value of $p_2 \leq 0.124$ is necessary to reject the

null hypothesis after the second stage. When the planned number of $n_2 = 23$ patients is included in the second stage, the conditional power, i.e., the probability of rejection of the null hypothesis given a reduced effect of $H_1' : \pi = 0.75$, amounts to

$$\begin{aligned}
\Pr_{H_1'}(P_2 \leq 0.124) &= \sum_{l=0}^{23} \Pr_{H_1'}\left\{P_2 \leq 0.124 \;\middle|\; P_2 = p_2(l)\right\} \cdot \Pr_{H_1'}\left\{P_2 = p_2(l)\right\} \\
&= \sum_{l=0}^{23} \mathbb{I}_{\{p_2(l) \leq 0.124\}} \cdot b(l; 0.75, 23) \\
&= 0.654,
\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function and b denotes the binomial probability mass function. It may be desirable to increase the probability for rejection of the null hypothesis at the end of the trial to at least 0.8 under the assumption that the true response rate is $\pi' = 0.75$. Continuing the trial with $n_2 = 43$ instead of the initially planned 23 patients will satisfy this requirement. Such a change in sample size is not possible in standard phase II designs without potential inflation of the type I error rate, as demonstrated in Chapter 3. The proposed flexible two-stage design allows this change without compromising the overall type I error rate. In the performance evaluation in Chapter 6, we saw, based on the average performance scores, that such recalculations are also efficient from a methodological point of view. Figure 6.2(a) shows that for the parameter setting $\pi_0 = 0.6$ and $\beta = 0.2$, a recalculation based on a reduced effect of π' is more efficient than the fixed design without adaptations.

Note that the determination of the sample size in stage two is not restricted to conditional power arguments or other pre-specified rules but can even be set *ad hoc* during the interim analysis and may additionally take economic or safety considerations into account.

Within the proposed flexible designs, it is also possible to react to unintentional sample size changes, e.g., due to overrun. Assume that the trial by Combs et al. continued after the interim analysis and that recruitment was not stopped exactly after attainment of the sample size specified for the second stage but that one additional patient was included. Consequently, $n_2 = 24$ patients instead of 23 were evaluable in the final analysis. Within the flexible design, the test decision depends only on the p-value of the second stage and therefore the evaluation is straightforward.

Statistics are no substitute for judgment.

(Henry Clay)

8

Discussion

This chapter summarizes the advances described in this thesis and their limitations.

8.1. Contributions to research

In this thesis, we have proposed a general method for two-stage one-armed clinical trials with discrete test statistics that allows arbitrary modifications of the second-stage sample size based on the results of the interim analysis or on information from outside the trial.

In a first step towards this goal, we directly applied flexible design methodology developed for continuous outcomes to binary response variables. This resulted in conservative procedures, and apparently self-evident solutions led to inflation of the type I error rate. Therefore, special methods are needed for discrete test statistics. It may be generally questioned whether strict control of the type I error rate is a substantial design aspect of single-arm trials as such trials are performed without including a control treatment. Designs with unknown or inflated type I error rate make things not better but even worse. If phase II designs are part of a regulatory submission, strict type I error rate control is mandatory.

Our first approach to guarantee a better exhaustion of the nominal level applies adaptive conditional tests based on a new fixed two-stage design that uses the combination test approach. In practical applications, a clinical trial can be planned for this fixed design whereby only the first-stage sample size and the adaptive conditional test boundaries need to be specified in the protocol. In the resulting flexible design, the sample size for the second stage can be changed while still controlling the type I error rate. If the sample

size is not changed, the power as well as the total and expected sample size coincide with those of the original combination test design. As these characteristics can be chosen to be very similar to the optimal design of Chang et al. (1987), the huge advantage of the new design with respect to its flexibility does not come at the price of an increase in sample size.

We have proven that a key characteristic of all flexible designs is that the second stage is planned for uniformly distributed p-values. We have developed a second discrete conditional error function approach that defines the conditional significance level of the second stage given the interim results. This approach allows flexible design changes as a free gift for all classical phase II designs presented in the literature (e.g., Simon, 1989; Banerjee and Tsiatis, 2006; Mander et al., 2012). The main contribution of this thesis is that in combination with the first approach, we were able to construct new and more efficient phase II designs that allow flexible design modifications and show better characteristics than standard designs if no adaptations are performed. The discreteness of the second stage is taken into account when planning the trial to guarantee satisfactory properties in practical applications. Applying the discrete conditional error function methodology, we derived these optimal flexible phase II designs both for a planned fixed second-stage sample size and in the situation that the second-stage sample size may already depend on the interim outcome in the planning phase. To calculate the designs efficiently, we showed how the branch-and-bound algorithm can be applied in combination with the discrete conditional error function methodology.

To evaluate the performance of the developed flexible sample size designs, adequate measures are needed to account for the option of sample size recalculation. We adapted a performance score by Liu et al. (2008) to discrete test statistics. With this tool, we evaluated and compared our designs with other phase II designs and analyzed different recalculation scenarios. Our findings may serve to guide researchers seeking recalculation rules that – given the observed interim data – are suitable for their trials. For a clinical trial conducted by Combs et al. (2012) that was planned with our flexible design method, we presented in detail which aspects should be considered in the planning and analysis stage.

8.2. Limitations and directions for further research

Similar to most designs in the area of phase II oncological trials, we treated the unfavorable response rate π_0 as a constant. In practice, π_0 may be set equal to a historical response level or some reference response level and is affected with some uncertainty. Ignoring this uncertainty, our frequentist method depends as little as possible on subjective input from

the judgment of physicians or previous studies. Bayesian methods, on the other hand, use this information by establishing a prior distribution for the response rate. To our knowledge, all Bayesian methods are restricted to per-design adaptive designs and do not allow the flexibility desired in our thesis. Further research will show whether and how uncertainty with respect to π_0 can be implemented into our flexible designs.

In all flexible designs presented, the rules for sample size recalculation need not be defined in advance, but it is possible to change the second-stage sample size freely after the interim analysis. Depending on the situation at hand, recalculation may be done according to conditional power arguments or may include economic or safety arguments. Likewise, the sample size can be chosen to obtain confidence intervals of specified width, and it is also possible to take into account information gained from parallel studies. It should be noted that the approach of using the conditional power based on the observed interim results showed undesirable properties. As a consequence, recalculations using the estimated interim effect should be considered with caution. A study team should investigate which adaptation strategy is to be preferred in a given situation. Evaluation of different rules for interim modifications leads, however, to a more complicated protocol and thus to an extended design stage of the trial. It must then be assessed whether the gains in efficiency during the trial justify this prolongation of the design stage.

The proposed flexible designs can deal with the situation that the sample size of the second stage is not definitely fixed after the interim analysis or that the planned sample size is not met. On the one hand, this ability enables investigators to cope with unintentional departures from protocol definitions such as over- or underrunning. On the other hand, there is potential for misuse of the method. Our approaches are valid only, when the data are examined exactly twice (at the interim and final analyses). All available data should be used at each analysis without arbitrarily adding or removing outcomes to fish for a desirable result. Therefore, adequate measures have to be taken to assure that the integrity of the trial is maintained when applying the proposed flexible designs, e.g., standard operation procedures specifically tailored to this type of design. Furthermore, detailed definitions should be given in the protocol and in the statistical analysis plan and must be adhered to strictly (Gallo, 2006a,b; Hung et al., 2006).

In our methodology, the first-stage sample size is fixed and needs to be attained at the interim analysis. Green and Dahlberg (1992), Chen and Ng (1998) and Li et al. (2012) developed methodology for classical phase II oncology designs to allow for unplanned changes in the sample size of the first stage. They imposed, either by frequentist or Bayesian procedures, assumptions on the distribution of the possible scenarios of over- or underrunning and developed designs with appropriate characteristics. From a methodological point of view, these designs control the type I error rate *averaged* over a range of first-stage sample

sizes. Therefore, these methods allow reaction to unintentional sample size changes. It is straightforward to apply these methods to our proposed flexible design to allow a certain degree of flexibility. However, these approaches do not control the type I error rate for each specific sample size that may occur in the course of the trial. Adequate measures for intentional sample size changes are still lacking. In this work, we have shown that our discrete conditional error function methodology allows for an alternative equivalent representation and evaluation of phase II designs. Further research will show whether this new presentation can be utilized to deal adequately with both planned and unplanned deviations from the study protocol, even in the first stage. Promising early results were presented by Englert and Kieser (2013a) at the 3rd Joint Statistical Meeting DAGStat 2013.

Furthermore, a more general usage and benefit of the branch-and-bound algorithm and its adaptation are possible. This method is a far more efficient means of identifying classical fixed two-stage phase II designs with specified optimality criteria than investigating all possible combinations of sample sizes and decision rules. This is, however, still the method applied by most authors (see, for example, Simon, 1989; Lin and Shih, 2004; Mander et al., 2012). A naïve search for optimal designs becomes infeasible for high total sample sizes or small treatment effects. In these situations, most authors impose restrictions on the total sample size in their search procedure thus, however, potentially leaving out the optimal solution (Dong et al., 2012; Mander et al., 2012). The algorithms can also be applied to modern methodological developments in phase II cancer trials, which account for new aspects and require more free parameters. For example, Chang et al. (2012) developed an improved two-stage phase II design that stratifies patients into subgroups to account for a different prognosis. The complexity of the design is increased as the different strata result in more free parameters. The test statistic used by Chang et al. is a linear combination of the observed number of responders. This allows a direct application of the branch-and-bound algorithm to ease the high complexity of the computations. In a recently accepted paper, Hou et al. (2013) present randomized phase II clinical trials with two treatment arms that are compared to a common historical control. The search algorithm for the design accounts for the number of responses in both treatment arms, doubling the number of free parameters. The authors state that "the computing time is intractably long if an exhaustive search is attempted" and restricted the parameter space considerably. Again, the optimization problem is linear and the branch-and-bound algorithm can be applied potentially thus allowing for an exhaustive search. More advanced phase II designs can in some instances only be derived when more efficient computational methods are applied. Availability of the algorithms and corresponding R programs (see Appendix A) may stimulate other authors to calculate their phase II designs efficiently possibly leading to further improvements in phase II designs.

8.3. Conclusions

Flexible designs for single-arm phase II trials in oncology are an important addition to the currently available methodological spectrum. They allow investigators to react to every eventuality that may occur in the course of a clinical trial without undermining the nominal type I error rate. Moreover, although they allow more flexibility, these designs do not feature a higher average sample size or a lower statistical power. In fact, the converse is true: it is possible to construct flexible designs that outperform existing designs.

Flexible interim analyses are the methodological realization of item 36 in the Critical Path Opportunities List released by the FDA, namely use of accumulated information in trial design. According to the FDA, items in the Critical Path Opportunities List provide "opportunities that, if implemented, can help speed the development and approval of medical products" (FDA, 2006).

We acknowledge that statistical considerations should never be the only reason for selecting a particular study design. A multitude of ethical, scientific, practical and economic issues must also be taken into account at the design stage.

Statistics may be defined as "a body of methods for making wise decisions in the face of uncertainty."

(Wilson Allen Wallis)

9

Summary

Clinical phase II trials in oncology are conducted to determine whether the activity of a new anticancer treatment is promising enough to merit further investigation. Two-stage designs are commonly used for this situation to allow for early termination. Although there is an ongoing debate on the relative merits of single-arm versus randomized phase II trials, the standard tool in cancer research remain single-arm trials.

Designs proposed in the literature so far have the common drawback that the sample sizes for the two stages have to be specified in the protocol and have to be adhered to strictly during the course of the trial. As a consequence, designs that allow a higher extent of flexibility are desirable. Currently available flexible design methods are tailored to comparative studies with continuous test statistics. We have shown that direct transfer of these methods to discrete test statistics results in conservative procedures and, likewise, in a loss in power. Therefore, special methods are needed for discrete test statistics.

In this thesis, we propose flexible methods that allow an arbitrary modification of the sample size of the second stage using the results of the interim analysis or external information while controlling the type I error rate. We constructed new designs based on a combination test and based on the conditional error function principle that directly account for the discreteness of the outcome. It is shown, further, how both approaches can be combined to construct new phase II designs that are more efficient as compared to currently applied designs and that allow flexible mid-course design modifications. We derived these new flexible and more efficient phase II designs for both a planned fixed second-stage sample size and for the situation that the planned second-stage sample size depends on the interim outcome. Results are tabulated for a wide range of frequently used design parameters.

Taking into account that classical phase II designs in oncology have been deemed to be optimal for decades, the additional decrease in sample size which can be achieved by applying our proposed methods must be regarded as a significant contribution to the field of clinical trial design. As hundreds of phase II oncology trials are performed each year worldwide, the application of our new designs will not only allow for flexibility in the conduct of these trials, but also preclude a considerable number of patients from being included in such trials.

The search algorithms used to identify these designs are computationally intensive. Therefore, ways to improve the search strategy were developed and the implementation of these methods was described in detail. All computer programs are provided and illustrated with examples.

Emphasis was placed on evaluation of the adaptive performance of the developed flexible phase II designs. When adjustments are made, the consequences in terms of increasing/decreasing the sample size have to be weighed against the gain/loss in power. We developed a performance indicator that fits to binary outcomes and satisfactorily addresses recalculation possibilities. Thus, we identified recalculation rules which improved the performance of the designs, if there is uncertainty with respect to the treatment effect.

Application and the rich possibilities of the proposed methods were illustrated with a clinical trial that was planned with methodology described in this thesis.

In summary, the new designs we developed allow the use of mid-course information for planning the second stage of the study, thus meeting practical requirements when performing phase II clinical trials in oncology. The observed reduction in average sample size when applying our new flexible study designs, with no resultant loss in power, is due to the new methodology and the efficient search algorithm.

A

Source Codes and Technical Notes for Programmers

A.1. Modified discrete conditional error function

Source code A.1: *Updatedcef-function*

```
updatedcef <- function(p0, inputdcef, nominalalpha = 0.05, how =
    "proportionally"){
  #Calculate parameters
  n1 <- length(inputdcef)-1
  propp0 <- dbinom(0:n1,n1,p0)
  alpha <- drop(inputdcef %*% propp0) #Calculation of type I
     error rate

  #Extract middle part of dcef
  inputdcefmiddle <- inputdcef[inputdcef!=1 & inputdcef!=0]
  propp0middle <- propp0[inputdcef!=1 & inputdcef!=0]
  n1middle <- length(inputdcefmiddle)
  restalpha <- nominalalpha - alpha

  if(restalpha < 0){
    cat("\n")
    cat("\n Nominal significance level lower than type I error
       rate. \n")
```

```
    cat("\n")
    updatetodcef <-rep(0,length(inputdcef))
        }
  else
    {
    switch(how,
      proportionally =
      updatetodcef <- c(rep(0,length(inputdcef[inputdcef==0]))
         ,((restalpha * (propp0middle/sum(propp0middle)))) /
         propp0middle,rep(0,length(inputdcef[inputdcef==1]))),
      equally =
      updatetodcef <- c(rep(0,length(inputdcef[inputdcef==0])),(
         restalpha / n1middle) / propp0middle,rep(0,length(
         inputdcef[inputdcef==1]))),
      border =
      updatetodcef <- c(rep(0,length(inputdcef[inputdcef==0])),c
         (restalpha, rep(0,n1middle-1)) / propp0middle,rep(0,
         length(inputdcef[inputdcef==1]))),
      updatetodcef <- rep(0,length(inputdcef))
      )
    }

  #Output results
  minimum_vector <- function(x){min(x,1)}
  outputdcef <- sapply(inputdcef + updatetodcef, minimum_vector)
  cat("\n")
  cat("dCEF: \n", inputdcef, "\n")
  cat("\n")
  cat("Updated dCEF: \n", outputdcef, "\n")
  cat("\n")
  cat("Alpha:  ",drop(outputdcef %*% propp0),"\n")
  cat("\n")
  }
```

The `updatedcef`-function has two mandatory parameters: `p0` $:= \pi_0$ and `inputdcef`, the discrete conditional error function that does not exhaust the nominal level. The two optional parameters `nominalalpha` $:= \alpha$ and `how` define the significance level used and how the discrete conditional error function should be increased to exhaust the nominal level. As standard, the significance level is set equal to 0.05 and the remaining level is

spent *proportionally*. Note that this results in terms of absolute values in an equal increase of all discrete conditional error function values unequal from zero or one. Other options are *equally* or *border*, where the remaining level is distributed equally among these values or the complete remaining level is spent on the smallest conditional error function value unequal to zero, respectively.

Within the `updatedcef`-function, the first step is to identify the "middle" region of discrete conditional error function with values unequal from zero or one. In a next step, these values are increased according to the specification how the remaining significance level $\alpha - \alpha'$ should be used (see (4.4), (4.5) and (4.6) on page 35). Finally, the original and updated (with increased values) discrete conditional error function are printed out.

A.2. Sample size recalculation

Source code A.2: *Recalculation of the second-stage sample size based on conditional power*

```
condpower <- function(p0,n1,p1,n2,dCEFvalue){
  condpoweriter <- 0
  for(l in 0:n2){
    if(1-pbinom(l-1, n2, p0) <= dCEFvalue){
      condpoweriter <- condpoweriter + dbinom(l,n2,p1)
    }
  }
  #Output result
  condpoweriter
  }

recalcn2 <- function(p0,n1,p1,dCEFvalue,boundary,n2max=Inf){
  n2 <- 0
  condpoweriter <- 0
  if (dCEFvalue == 0 | dCEFvalue == 1){
      n2 <- 0}
  else{
    while(condpoweriter < boundary & n2 <= n2max){
      n2 <- n2 + 1
      condpoweriter <- condpower(p0,n1,p1,n2,dCEFvalue)
    }
  }
  #Output result
```

```
  cat("Sample size needed in the second stage:\n")
  cat(n2)
  cat("\n")
  }
```

The source code for recalculation of the sample size based on conditional power is spitted into two functions: A `condpower`-function that is capable of calculating conditional powers and a `recalcn2`-function that determines the recalculated sample size.

The `condpower`-function has five mandatory parameters: `p0` $:= \pi_0$, `n1` $:= n_1$, `p1` $:= \pi_1$, `n2` $:= n_2$ and the value of the discrete conditional error function used for recalculation `dCEFvalue` $:= D$. Given D, the conditional power for $H_1 : \pi = \pi_1$ is calculated according to the law of total probability:

$$\begin{aligned} \Pr_{H_1}\{P_2 \leq D\} &= \sum_{l=0}^{n_2} \Pr_{H_1}\{P_2 \leq D \mid P_2 = p_2(l)\} \cdot \Pr_{H_1}\{P_2 = p_2(l)\} \\ &= \sum_{l=0}^{n_2} \mathbb{I}_{\{p_2(l) \leq D\}} \cdot b(l; \pi_1, n_2), \end{aligned}$$

where $\mathbb{I}$ denotes the indicator function and b denotes the binomial probability mass function.

The `recalcn2`-function also has `p0` $:= \pi_0$, `n1` $:= n_1$, `p1` $:= \pi_1$ and `dCEFvalue` $:= D$ as mandatory parameters. In addition, the value of the conditional power that should be achieved with the recalculated sample size is required. An additional optional parameter `n2max` defines if the recalculated sample sizes should be truncated at a certain value. As standard, no restrictions are made, i.e., the maximum second-stage sample size is set equal to infinity. The recalculated sample size n_2 is determined by choosing n_2 as the minimum integer where the conditional power is greater than the specified boundary. With $D = 0$ or $D = 1$ early stopping after the first stage is possible and the second-stage sample size is set equal to $n_2 = 0$.

A.3. Branch-and-bound

The algorithms to determine the sample sizes of the proposed flexible designs are computationally intensive. We developed an intelligent search algorithm that uses the branch-and-bound method and thus allows an exhaustive and non-restricted search for the optimal design. It consists of three routines, (a) a `launch`-function that defines all design parameters, calculates needed variables, initializes the branching algorithm and afterwards displays the results, (b) a `branch`-function that splits the problem into similar sub-problems

and (c) a bound-function that discards sub-problems that cannot lead to optimal solutions of the test problem. The following sections give the complete source code together with technical notes on the programming.

A.3.1. Launch-function

Source code A.3: *Branch-and-bound – Launch-function*

```
launch <- function(p0,p1,nominalalpha,nominalbeta,n1,n2min,n2max
    = n2min,minpnext = 0,en = n1+n2max){
  #Define all design parameters
  p0 <<- p0
  p1 <<- p1
  nominalalpha <<- nominalalpha
  nominalbeta  <<- nominalbeta
  nominalpower <<- 1-nominalbeta
  n1 <<- n1
  n2min <<- n2min
  n2max <<- n2max
  minpnext <<- minpnext

  #Calculation of variables
  propp0 <<- dbinom(0:n1,n1,p0)
  propp1 <<- dbinom(0:n1,n1,p1)
  count_n1 <<- length(propp0);
  dcpf <<- c()
  dcef <<- c()
  dcss <<- c()
  for (n2iter in n2min:n2max) {
    dcef <<- c(dcef,pbinom(0:n2iter-1,n2iter,p0,lower.tail =
       FALSE))
    dcpf <<- c(dcpf,pbinom(0:n2iter-1,n2iter,p1,lower.tail =
       FALSE))
    dcss <<- c(dcss,rep(n1+n2iter,length(pbinom(0:n2iter-1,
       n2iter,p0,lower.tail = FALSE))))
  }
  possiblecef <<- cbind(dcef,dcpf,dcss)
  possiblecef <<- subset(possiblecef, dcef >= minpnext & dcef <=
      1-minpnext)
```

```
possiblecef <<- possiblecef[order(possiblecef[,1]),]
dcef <<- c(0,possiblecef[,1],1)
dcpf <<- c(0,possiblecef[,2],1)
dcss <<- c(n1,possiblecef[,3],n1)

set_P_2 <<- length(dcef);
niter <<- rep(1,count_n1)

#Preparation for start
i<<-0
en<<-en
combination<<-c()

cat("\n")
cat("Searching optimal solutions:\n")
cat("\n")
#Initialize first branching-step
branch(0,1)
#Print final results
cat("Search completed. In total,",i,"of",choose(count_n1+set_P
   _2-1,count_n1),"combinations were evaluated.\n")
cat("\n")
if(i==0){
  cat("No solution possible.\n")
  }
else {
  cat("Optimal D(k): \n", dcef[combination], "\n \n")
  cat("Optimal n_2(k): \n", dcss[combination], "\n \n")
  cat("\n")
  cat("Alpha:  ",drop(dcef[combination] %*% propp0),"\n")
  cat("Beta:   ",1-drop(dcpf[combination] %*% propp1),"\n")
  cat("EN_p0:  ",drop(dcss[combination] %*% propp0),"\n")
  }
cat("\n")
}
```

The `launch`-function initializes the branch-and-bound algorithm. The function has six mandatory parameters: `p0` $:= \pi_0$, `p1` $:= \pi_1$, `nominalalpha` $:= \alpha$, `nominalbeta` $:= \beta$, `n1` $:= n_1$ and `n2min` $:= n_{2,\min}$. The first optional parameter is $n_{2,max}$. If $n_{2,max}$ is

not specified, it is set equal to $n_{2,max} = n_{2,min} = n_2$. With this parameter setting the branch-and-bound algorithm searches for an optimal discrete conditional error function as presented in Section 4.4. If $n_{2,\max}$ is specified, the second-stage sample size is allowed to vary between $n_{2,\min}$ and $n_{2,\max}$ and the program searches for an optimal adaptive design as introduced in Chapter 5. The remaining two optional parameters `minpnext` and `en` can be used to further customize the optimization process.

Within the `launch`-function, first all design parameters are defined globally to allow their use also within the `branch` and `bound` routines. Additionally, some other variables that will be frequently used are calculated and defined globally. Later only the evaluated expressions are used in order to save runtime. These variables include the probability under the null and alternative hypothesis to observe exactly k responses out of n_1 patients, $k \in 0, \ldots, n_1$, which are stored in `propp0` and `propp1`, respectively, and the possible values of the discrete conditional error functions, which are calculated for each second-stage sample size n_2 between $n_{2,\min}$ and $n_{2,\max}$ and each number of responses, see (5.2). The discrete conditional error function values are later used to calculate the (minimal) type I error rate, see (4.7) and (5.1). The (minimal) type II error rate and the (minimal) average sample sizes are calculated similarly, see (4.8), (4.9), (5.3) and (5.4). Therefore, for each discrete conditional error function value the corresponding value of $\Pr_{H_1}\{P_{2,n_2(k)} \leq D(k)\}$ and the corresponding sample size are stored all together in the matrix `possiblecef`. As mentioned in Chapter 5, it may be desirable to consider besides zero and one only discrete conditional error function values that are away from zero or one by a certain amount. This threshold is defined by the second optional parameter `minpnext`. As standard it is set equal to zero.

Before the actual branch-and-bound algorithm is started with the first branching step, the counter for the number of fully evaluated designs is set equal to zero and a first guess for the average sample size is handed over by the third optional parameter `en`. If this parameter is not specified, the very conservative value of $n_1 + n_{2,\max}$ is used. As the algorithm fully evaluates only designs that can lead to smaller average sample sizes, see (4.9), a reasonable first choice can significantly speed up the optimization process. After completion of the searching procedure, the optimal combination of discrete conditional error function values and second-stage sample sizes is printed out together with the design characteristics.

A.3.2. Branch-function

Source code A.4: *Branch-and-bound – Branch-function*

```
branch <- function(k,j){
  if (k < count_n1){
    #Deeper into the tree
    for (jiter in j:set_P_2){

      niter[k+1]<<-jiter      #Built up index-vector

      if(bound(k+1,jiter)){   #Bounding
        branch(k+1,jiter)     #Branching
        }

      }
    }

  else{
    #Output solution if index-vector is defined completely
      i<<-i+1

      en<<-drop(dcss[niter] %*% propp0)
      combination<<-niter
      print(en)
      print(niter)
    }
  }
```

The `branch`-function recursively defines the layout of the discrete conditional error function. The recursion starts with defining the discrete conditional error function for $k = 0$ responses and ends when the layout is defined for 0 to n_1 responses. However, instead of directly defining recursively the real-valued discrete conditional error function, we built up an integer based index-vector `niter`. The ith-element of the index-vector defines which value of the increasingly ordered set $\mathbf{P}_2$ is used for $D(i-1)$. Therefore, the index-vector is of length $n_1 + 1$ and each element can range between one and the number of different discrete conditional error function values $|\mathbf{P}_2|$.

The `branch`-function checks first, if the index-vector and therefore the layout of the discrete conditional error function is already defined completely. If the index vector is

defined only up to $k < n_1$ responses, the `branch`-function splits the optimization problem into similar sub-problems. In each sub-problem, the index level used for $k + 1$ responses is set equal to a specific value out of all possible index values, that are at least as large as the index value for k responses. This restriction ensures monotonicity of the underlying discrete conditional error function, i.e., $D(k + 1) \geq D(k)$. The `bound`-subroutine then checks for each sub-problem, if it can lead to the optimal design. If not, the recursion is stopped. Otherwise, the `branch`-function is recursively invoked again. Note that in case that the current branching step has fully defined the discrete conditional error function, the `bound`-function checks if a new solution to the optimization process has been found. In that case the next `branch`-function stops the recursion, increases the counter for fully evaluated designs by one and prints the expected sample size together with the index vector of the corresponding design.

A.3.3. Bound-function

Source code A.5: *Branch-and-bound – Bound-function*

```
bound <- function(k,j){
  mindcef <- if (k < count_n1){
    drop(dcef[niter[1:k]] %*% propp0[1:k]) + dcef[j] * sum(
       propp0[(k+1):count_n1])
    }
    else
    {
    drop(dcef[niter] %*% propp0)
    }

  if(mindcef > nominalalpha) {
    return(FALSE)
    }

  else {
      maxdcpf <- if (k < count_n1){
        drop(dcpf[niter[1:k]] %*% propp1[1:k]) + sum(propp1[(k
           +1):count_n1])
        }
        else
        {
        drop(dcpf[niter] %*% propp1)
```

```
        }

      if(maxdcpf < nominalpower) {
      return(FALSE)
      }
      else
      {
        minen <- if (k < count_n1){
          drop(dcss[niter[1:k]] %*% propp0[1:k]) + n1 * sum(
             propp0[(k+1):count_n1])
          }
          else
          {
          drop(dcss[niter] %*% propp0)
          }

        if(minen > en) {
        return(FALSE)
        }
        else
        {
          if(k==1 & j>1) {
          return(FALSE)
          }
          else
          {
          return(TRUE) #Further branching only if no restraints
             are fulfilled
          }
        }
      }
    }
  }
```

The `bound`-function checks within the `branch`-recursion, if the current sub-problem can lead to an optimal solution. Therefore, it checks if the considered sub-problem shows a minimal type I error rate (4.7) smaller or equal to the nominal significance level, a minimal type II error rate (4.8) smaller or equal to the nominal value and a minimal average sample size at least as large as the smallest average sample size found for all designs considered so

far. If all these conditions are fulfilled, the bound-function returns TRUE and otherwise FALSE.

Note that, if the bound-function is invoked when the discrete conditional error function is defined completely, it checks if the design satisfies the type I and II error rate constraints (5.1, 5.3) and leads to a smaller average sample size (5.4). In this case, a new optimal solution to the test problem has been found.

A.3.4. Modifications

The branch-and-bound approach as given so far identifies the designs that minimize the average sample size under the null hypothesis. It can easily be modified to optimize the design with respect to other optimization criteria. Mander and Thompson (2010) and Mander et al. (2012) constructed, for example, fixed designs that are optimal with respect to the alternative hypothesis. It is straightforward to identify with the branch-and-bound approach flexible designs as in Chapter 4 or adaptive designs as in Chapter 5 that are optimal with respect to the same criteria.

Note that the average sample size under the alternative hypothesis is calculated as

$$\mathrm{EN}(\pi_1) = \sum_{k=0}^{n_1} \big\{ n_1 + n_2(k) \big\} \cdot \Pr_{H_1}(K = k).$$

The minimal average sample size of all following sub-problems after $m+1$ branching steps, i.e., when the conditional error function is defined for 0 to m responses, is given by

$$\mathrm{EN}(\pi_1)_{\min} = \sum_{k=0}^{m} \big\{ n_1 + n_2(k) \big\} \cdot \Pr_{H_1}(K = k) + n_1 \cdot \sum_{k=m+1}^{n_1} \Pr_{H_1}(K = k).$$

Therefore, only the corresponding statement in the Bound-function source code A.5 needs to be replaced by the following code snippet.

Source code A.6: *Branch-and-bound – Modification*

```
minen <- if (k < count_n1){
  drop(dcss[niter[1:k]] %*% propp1[1:k]) + n1 * sum(propp1[(k+1)
     :count_n1])
  }
  else
  {
  drop(dcss[niter] %*% propp1)
  }
```

B

Additional Tables

B.1. Simon's design

Table B.1.: *Simon's optimal designs ($\pi_1 - \pi_0 = 0.15$)*

π_0	π_1	α	β	l_1	n_1	l_2	n_2	n	EN(π_0)	α'	β'
0.05	0.20	0.05	0.2	0	10	3	19	29	17.6	0.047	0.199
		0.05	0.1	1	21	4	20	41	26.7	0.046	0.098
0.10	0.25	0.05	0.2	2	18	7	25	43	24.7	0.048	0.200
		0.05	0.1	2	21	10	45	66	36.8	0.050	0.098
0.20	0.35	0.05	0.2	5	22	19	50	72	35.4	0.049	0.200
		0.05	0.1	8	37	22	46	83	51.4	0.049	0.099
0.30	0.45	0.05	0.2	9	27	30	54	81	41.7	0.050	0.198
		0.05	0.1	13	40	40	70	110	60.8	0.048	0.099
0.40	0.55	0.05	0.2	11	26	40	58	84	44.9	0.049	0.195
		0.05	0.1	19	45	49	59	104	64.0	0.050	0.100
0.50	0.65	0.05	0.2	15	28	48	55	83	43.7	0.047	0.198
		0.05	0.1	22	42	60	63	105	62.3	0.050	0.099
0.60	0.75	0.05	0.2	17	27	46	40	67	39.3	0.047	0.200
		0.05	0.1	21	34	64	61	95	55.6	0.048	0.099
0.70	0.85	0.05	0.2	14	19	46	40	59	30.3	0.049	0.193
		0.05	0.1	18	25	61	54	79	43.4	0.049	0.096
0.80	0.95	0.05	0.2	7	9	26	20	29	17.7	0.049	0.198
		0.05	0.1	16	19	37	23	42	24.4	0.048	0.097

Table B.2.: *Simon's minimax designs* ($\pi_1 - \pi_0 = 0.15$)

π_0	π_1	α	β	l_1	n_1	l_2	n_2	n	EN(π_0)	α'	β'
0.05	0.20	0.05	0.2	0	13	3	14	27	19.8	0.042	0.199
		0.05	0.1	1	29	4	9	38	32.9	0.039	0.100
0.10	0.25	0.05	0.2	2	22	7	18	40	28.8	0.040	0.197
		0.05	0.1	3	31	9	24	55	40.0	0.042	0.099
0.20	0.35	0.05	0.2	6	31	15	22	53	40.4	0.050	0.198
		0.05	0.1	8	42	21	35	77	58.4	0.044	0.100
0.30	0.45	0.05	0.2	16	46	25	19	65	49.6	0.050	0.197
		0.05	0.1	27	77	33	11	88	78.5	0.050	0.099
0.40	0.55	0.05	0.2	28	59	34	11	70	60.1	0.050	0.198
		0.05	0.1	24	62	45	32	94	78.9	0.049	0.100
0.50	0.65	0.05	0.2	39	66	40	2	68	66.1	0.049	0.199
		0.05	0.1	28	57	54	36	93	75.0	0.048	0.100
0.60	0.75	0.05	0.2	18	30	43	32	62	43.8	0.047	0.198
		0.05	0.1	48	72	57	12	84	73.2	0.050	0.100
0.70	0.85	0.05	0.2	16	23	39	26	49	34.4	0.047	0.199
		0.05	0.1	33	44	53	24	68	48.5	0.049	0.098
0.80	0.95	0.05	0.2	7	9	26	20	29	17.7	0.049	0.198
		0.05	0.1	31	35	35	5	40	35.3	0.049	0.100

B.2. Proposed design

Table B.3.: *Design characteristics of optimal flexible designs ($\pi_1 - \pi_0 = 0.15$)*

				Proposed method		Simon (1989)		Mander and Thompson (2010)	
π_0	π_1	α	β	n	EN(π_0)	n	EN(π_0)	n	EN(π_0)
0.05	0.20	0.05	0.20	29	17.60	29	17.62	29	17.60
			0.10	43	26.07	41	26.66	41	26.60
0.1	0.25	0.05	0.20	43	24.49	43	24.66	43	24.49
			0.10	66	36.24	66	36.82	63	36.63
0.2	0.35	0.05	0.20	63	35.11	72	35.37	72	35.29
			0.10	87	50.90	83	51.45	83	51.29
0.3	0.45	0.05	0.20	77	41.49	81	41.71	81	41.69
			0.10	103	60.08	110	60.77	100	60.22
0.4	0.55	0.05	0.20	81	44.12	84	44.93	84	44.78
			0.10	104	63.91	104	63.96	104	63.91
0.5	0.65	0.05	0.20	81	43.12	83	43.72	83	43.37
			0.10	109	61.91	105	62.29	105	62.26
0.6	0.75	0.05	0.20	75	38.56	67	39.35	78	39.00
			0.10	97	55.02	95	55.60	95	55.52
0.7	0.85	0.05	0.20	60	30.14	59	30.29	60	30.14
			0.10	75	42.57	79	43.40	80	43.24
0.8	0.95	0.05	0.20	29	17.72	29	17.72	29	17.72
			0.10	42	24.45	42	24.45	42	24.45

Table B.4.: *Design characteristics of minimax flexible designs ($\pi_1 - \pi_0 = 0.15$)*

				Proposed method		Simon (1989)		Mander and Thompson (2010)	
π_0	π_1	α	β	n	EN(π_0)	n	EN(π_0)	n	EN(π_0)
0.05	0.20	0.05	0.20	26	24.20	27	19.81	27	18.60
			0.10	38	28.33	38	32.86	38	28.33
0.1	0.25	0.05	0.20	38	29.85	40	28.84	38	33.94
			0.10	53	41.39	55	40.03	53	47.87
0.2	0.35	0.05	0.20	53	40.41	53	40.44	53	40.41
			0.10	74	59.68	77	58.42	76	66.51
0.3	0.45	0.05	0.20	64	48.10	65	49.63	64	51.32
			0.10	88	68.30	88	78.51	88	78.45
0.4	0.55	0.05	0.20	69	49.88	70	60.07	69	54.17
			0.10	94	74.24	94	78.88	94	76.30
0.5	0.65	0.05	0.20	67	56.40	68	66.11	68	66.04
			0.10	93	69.84	93	75.00	93	72.20
0.6	0.75	0.05	0.20	61	45.39	62	43.79	62	42.88
			0.10	84	60.12	84	73.20	84	73.13
0.7	0.85	0.05	0.20	49	33.17	49	34.44	49	34.36
			0.10	65	48.84	68	48.52	65	50.46
0.8	0.95	0.05	0.20	29	17.72	29	17.72	29	17.72
			0.10	40	27.98	40	35.30	40	35.21

Table B.5.: *Discrete conditional error function for optimal flexible designs*

π_0	π_1	α	β	n_1	n_2	Discrete conditional error function								
0.05	0.2	0.05	0.2	10	19	p_1	0.4013	0.0861	0.0115					
						$D(p_1)$	0.070	0.259	0.724					
		0.05	0.1	20	23	p_1	0.2642	0.0755	0.0159	0.0026	3e-04			
						$D(p_1)$	0.105	0.321	0.693	0.695	0.713			
0.1	0.25	0.05	0.2	18	25	p_1	0.2662	0.0982	0.0282					
						$D(p_1)$	0.099	0.239	0.47					
		0.05	0.1	26	40	p_1	0.2591	0.1118	0.0399	0.0119				
						$D(p_1)$	0.100	0.207	0.373	0.784				
0.2	0.35	0.05	0.2	23	40	p_1	0.3053	0.1598	0.0715	0.0273	0.0089			
						$D(p_1)$	0.088	0.161	0.269	0.27	0.568			
		0.05	0.1	33	54	p_1	0.3343	0.200	0.1068	0.0508	0.0216	0.0082		
						$D(p_1)$	0.060	0.107	0.178	0.274	0.527	0.78		
0.3	0.45	0.05	0.2	25	52	p_1	0.3231	0.1894	0.0978	0.0442	0.0175			
						$D(p_1)$	0.072	0.120	0.189	0.279	0.505			
		0.05	0.1	43	60	p_1	0.2919	0.1919	0.1169	0.0658	0.0342	0.0164		
						$D(p_1)$	0.063	0.104	0.162	0.238	0.437	0.549		
0.4	0.55	0.05	0.2	28	53	p_1	0.3050	0.1868	0.1025	0.0499	0.0215	0.0081	0.0027	
						$D(p_1)$	0.070	0.115	0.258	0.258	0.463	0.575	0.577	
		0.05	0.1	45	59	p_1	0.3214	0.2223	0.1436	0.0865	0.0483	0.025	0.012	0.0053	0.0021
						$D(p_1)$	0.060	0.098	0.151	0.22	0.306	0.404	0.511	0.619	0.63
0.5	0.65	0.05	0.2	28	53	p_1	0.2858	0.1725	0.0925	0.0436	0.0178	0.0063	0.0019	
						$D(p_1)$	0.084	0.136	0.205	0.392	0.5	0.5	0.709	
		0.05	0.1	40	69	p_1	0.3179	0.2148	0.1341	0.0769	0.0403	0.0192	0.0083	0.0032	0.0011
						$D(p_1)$	0.074	0.114	0.168	0.235	0.315	0.405	0.406	0.502	0.505
0.6	0.75	0.05	0.2	25	50	p_1	0.2735	0.1536	0.0736	0.0294	0.0095			
						$D(p_1)$	0.097	0.159	0.242	0.458	0.479			
		0.05	0.1	38	59	p_1	0.2897	0.1864	0.1089	0.0572	0.0268	0.0110	0.0040	
						$D(p_1)$	0.086	0.138	0.206	0.291	0.291	0.494	0.600	
0.7	0.85	0.05	0.2	19	41	p_1	0.2822	0.1332	0.0462					
						$D(p_1)$	0.096	0.174	0.283					
		0.05	0.1	30	45	p_1	0.2814	0.1595	0.0766	0.0302	0.0093			
						$D(p_1)$	0.093	0.165	0.262	0.380	0.381			
0.8	0.95	0.05	0.2	9	20	p_1	0.4362	0.1342						
						$D(p_1)$	0.072	0.211						
		0.05	0.1	19	23	p_1	0.2369	0.0829	0.0144					
						$D(p_1)$	0.137	0.306	0.546					

Table B.6.: *Discrete conditional error function for minimax flexible designs*

π_0	π_1	α	β	n_1	n_2	Discrete conditional error function											
0.05	0.2	0.05	0.2	21	5	p_1	0.6594	0.283	0.0849								
						$D(p_1)$	0.025	0.028	0.242								
		0.05	0.1	24	14	p_1	0.3392	0.1159									
						$D(p_1)$	0.031	0.155									
0.1	0.25	0.05	0.2	19	19	p_1	0.5797	0.2946	0.1150	0.0352							
						$D(p_1)$	0.036	0.036	0.116	0.584							
		0.05	0.1	28	25	p_1	0.5406	0.3054	0.1421	0.055	0.0179						
						$D(p_1)$	0.010	0.098	0.098	0.237	0.73						
0.2	0.35	0.05	0.2	31	22	p_1	0.4289	0.2700	0.1508	0.0746	0.0327	0.0127	0.0044				
						$D(p_1)$	0.020	0.056	0.133	0.268	0.458	0.671	0.853				
		0.05	0.1	47	27	p_1	0.4708	0.3331	0.2174	0.1306	0.0721	0.0366	0.0171	0.0074	0.0029		
						$D(p_1)$	0.011	0.030	0.074	0.156	0.287	0.461	0.462	0.653	0.983		
0.3	0.45	0.05	0.2	32	32	p_1	0.5049	0.3560	0.2283	0.1326	0.0694	0.0327	0.0138	0.0052			
						$D(p_1)$	0.014	0.033	0.070	0.133	0.356	0.357	0.506	0.792			
		0.05	0.1	51	37	p_1	0.4675	0.3505	0.2473	0.1637	0.1015	0.0589	0.0319	0.0161	0.0076	0.0033	0.0014	5e-04	2e-04
						$D(p_1)$	0.013	0.029	0.060	0.113	0.193	0.302	0.434	0.576	0.576	0.824	0.824	0.824	0.983
0.4	0.55	0.05	0.2	37	32	p_1	0.4032	0.2819	0.1820	0.1080	0.0586	0.0290	0.0131	0.0053	0.002		
						$D(p_1)$	0.046	0.046	0.092	0.165	0.268	0.539	0.539	0.679	0.889		
		0.05	0.1	54	40	p_1	0.5073	0.3981	0.2967	0.2094	0.1396	0.0877	0.0518	0.0287	0.0149	0.0072	0.0033
						$D(p_1)$	0.008	0.019	0.039	0.074	0.13	0.209	0.312	0.432	0.683	0.683	0.966
0.5	0.65	0.05	0.2	48	19	p_1	0.4427	0.3327	0.2354	0.1562	0.0967	0.0557	0.0297	0.0147	0.0066	0.0028	0.001
						$D(p_1)$	0.010	0.010	0.032	0.084	0.324	0.324	0.500	0.676	0.821	0.969	0.97
		0.05	0.1	55	38	p_1	0.3939	0.2950	0.2094	0.1403	0.0885	0.0524	0.029	0.015	0.0072		
						$D(p_1)$	0.017	0.037	0.072	0.128	0.209	0.314	0.565	0.565	0.793		
0.6	0.75	0.05	0.2	32	29	p_1	0.4618	0.3233	0.2046	0.1156	0.0575	0.0248	0.0091	0.0028	7e-04		
						$D(p_1)$	0.023	0.057	0.119	0.119	0.343	0.343	0.638	0.639	0.771		
		0.05	0.1	43	41	p_1	0.4178	0.3013	0.2013	0.1238	0.0695	0.0354	0.0162	0.0066	0.0024	7e-04	
						$D(p_1)$	0.012	0.057	0.106	0.178	0.275	0.391	0.392	0.642	0.644	0.85	
0.7	0.85	0.05	0.2	25	24	p_1	0.3407	0.1935	0.0905	0.0332	0.0090	0.0016					
						$D(p_1)$	0.043	0.111	0.229	0.566	0.568	0.744					
		0.05	0.1	37	28	p_1	0.4241	0.2889	0.1763	0.0947	0.044	0.0172	0.0055				
						$D(p_1)$	0.016	0.048	0.113	0.221	0.366	0.685	0.69				
0.8	0.95	0.05	0.2	9	20	p_1	0.4362	0.1342									
						$D(p_1)$	0.072	0.211									
		0.05	0.1	22	18	p_1	0.3320	0.1545	0.048	0.0074							
						$D(p_1)$	0.019	0.272	0.274	0.884							

Bibliography

Adjei, A. A., Christian, M., and Ivy, P. (2009). Novel designs and end points for phase II clinical trials. *Clinical Cancer Research*, 15:1866–1872.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132:235–244.

Ayanlowo, A. O. and Redden, D. (2008). A two stage conditional power adaptive design adjusting for treatment by covariate interaction. *Contemporary Clinical Trials*, 29:428–438.

Banerjee, A. and Tsiatis, A. A. (2006). Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine*, 25:3382–3395.

Bauer, P. (1989a). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie*, 20:130–148.

Bauer, P. (1989b). Sequential tests of hypotheses in consecutive trials. *Biometrical Journal*, 6:663–676.

Bauer, P. (2008). Adaptive designs: Looking for a needle in the haystack—a new challenge in medical research. *Statistics in Medicine*, 27:1565–1580.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50:1029–1041.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, 18:1833–1848.

Bauer, P. and König, F. (2006). The reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine*, 25:23–36.

Brannath, W., König, F., and Bauer, P. (2006). Estimation in flexible two stage designs. *Statistics in Medicine*, 25:3366–3381.

Brannath, W., Koenig, F., and Bauer, P. (2007). Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics*, 6:205–216.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, 97:236–244.

Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28:1181–1217.

Chang, M. (2007). Adaptive design method based on sum of p-values. *Statistics in Medicine*, 26:2772–2784.

Chang, M. N., Shuster, J. J., and Hou, W. (2012). Improved two-stage tests for stratified phase II cancer clinical trials. *Statistics in Medicine*, 31:1688–1698.

Chang, M. N., Therneau, T. M., Wieand, H. S., and Cha, S. S. (1987). Designs for group sequential phase II clinical trials. *Biometrics*, 43:865–874.

Chen, C.-M. and Chi, Y. (2011). Curtailed two-stage designs with two dependent binary endpoints. *Pharmaceutical Statistics*, 11:57–62.

Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 16:2701–2711.

Chen, T. T. and Ng, T.-H. (1998). Optimal flexible designs in phase II clinical trials. *Statistics in Medicine*, 17:2301–2312.

Chow, S.-C. and Chang, M. (2008). Adaptive design methods in clinical trials – a review. *Orphanet Journal of Rare Diseases*, 3:11.

Chow, S.-C., Chang, M., and Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, 15:575–591.

Chow, S.-C., Shao, J., and Wang, H. (2008). *Sample Size Calculations in Clinical Research*. Chapman & Hall / CRC, Boca Raton, second edition. ISBN 978-1584889823.

Coburger, S. and Wassmer, G. (2001). Conditional point estimation in adaptive group sequential test designs. *Biometrical Journal*, 43:821–833.

Combs, S. E., Kieser, M., Habermehl, D., Weitz, J., Jäger, D., Fossati, P., Orrechia, R., Engenhart-Cabillic, R., Pötter, R., Dosanjh, M., Jäkel, O., Büchler, M. W., and Debus, J. (2012). Phase I/II trial evaluating carbon ion radiotherapy for the treatment of recurrent rectal cancer: the PANDORA-01 trial. *BMC Cancer*, 12:137.

Committee for Medicinal Products for Human Use (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. CHMP/EWP/2459/02.

Cui, L., Hung, H. M. J., and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 55:853–857.

DeMets, D. L. and Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13:1341–1352.

Dette, H., Bornkamp, B., and Bretz, F. (2012). On the efficiency of two-stage response-adaptive designs. *Statistics in Medicine*, DOI 10.1002/sim.5555.

Dong, G., Shih, W. J., Moore, D., Quan, H., and Marcella, S. (2012). A bayesian-frequentist two-stage single-arm phase II clinical trial design. *Statistics in Medicine*, DOI 10.1002/sim.5330.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., and Verweij, J. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European Journal of Cancer*, 45:228–247.

Englert, S. and Kieser, M. (2012a). Adaptive designs for single-arm phase II trials in oncology. *Pharmaceutical Statistics*, 11:241–249.

Englert, S. and Kieser, M. (2012b). Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics*, 68:886–892.

Englert, S. and Kieser, M. (2013a). An approach for unplanned sample size changes in one-armed phase II cancer clinical trials. In *3rd Joint Statistical Meeting DAGStat 2013*.

Englert, S. and Kieser, M. (2013b). Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biometrical Journal*, under review.

Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*, 13:1727–1736.

FDA (2004). Challenge and opportunity on the critical path to new medical products. Technical report, U.S. Department of Health and Human Services.

FDA (2006). Critical path opportunities list. Technical report, U.S. Department of Health and Human Services.

FDA (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics. Technical report, U.S. Department of Health and Human Services.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, 38:143–151.

Fleming, T. R. (2006). Standard versus adaptive monitoring procedures: A commentary. *Statistics in Medicine*, 25:3305–3312.

Friede, T. and Kieser, M. (2004). Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics*, 3:269–279.

Gallo, P. (2006a). Confidentiality and trial integrity issues for adaptive designs. *Drug Information Journal*, 40:445–450.

Gallo, P. (2006b). Operational challenges in adaptive design implementation. *Pharmaceutical Statistics*, 5:119–124.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development – an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics*, 16:275–283.

Gan, H. K., Grothey, A., Pond, G. R., Moore, M. J., Siu, L. L., and Sargent, D. (2010). Randomized phase II trials: Inevitable or inadvisable? *Journal of Clinical Oncology*, 28:2641–2647.

Gould, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine*, 14:1039–51; discussion 1053–5.

Green, S. J. and Dahlberg, S. (1992). Planned versus attained design in phase II clinical trials. *Statistics in Medicine*, 11:853–862.

Hanfelt, J. J., Slack, R. S., and Gehan, E. A. (1999). A modification of Simon's optimal design for phase II trials when the criterion is median sample size. *Controlled Clinical Trials*, 20:555–566.

Hou, W., Chang, M. N., Jung, S.-H., and Li, Y. (2013). Designs for randomized phase II clinical trials with two treatment arms. *Statistics in Medicine*, accepted.

Hung, H. M. J., O'Neill, R., Wang, S.-J., and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal*, 48:565–573.

ICH Topic E 8 (1998). General considerations for clinical trials. CPMP/ICH/291/95.

ICH Topic E 9 (1998). Statistical principles for clinical trials. CPMP/ICH/363/96.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22:971–933.

Jennison, C. and Turnbull, B. W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, 25:917–932.

Jin, H. and Wei, Z. (2012). A new adaptive design based on Simon's two-stage optimal design for phase II clinical trials. *Contemporary Clinical Trials*, 33:1255–1260.

Jones, C. L. and Holmgren, E. (2007). An adaptive Simon two-stage deign for phase 2 studies of targeted therapies. *Contemporary Clinical Trials*, 28:654–661.

Jung, S.-H., Lee, T., Kim, K., and George, S. L. (2004). Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 23:561–569.

Kieser, M., Bauer, P., and Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analysis. *Biometrical Journal*, 41:261–277.

Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, 19:901–911.

Kieser, M. and Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, 22:3571–3581.

Koyama, T. and Chen, H. (2008). Proper inference from Simon's two-stage designs. *Statistics in Medicine*, 27:3145–3154.

Kunz, C. U. and Kieser, M. (2011a). Optimal two-stage designs for single-arm phase II oncology trials with two binary endpoints. *Methods of Information in Medicine*, 50:372–377.

Kunz, C. U. and Kieser, M. (2011b). Simon's minimax and optimal and Jung's admissible two-stage designs with or without curtailment. *The Stata Journal*, 11:240–254.

Kunz, C. U. and Kieser, M. (2012). Curtailment in singe-arm two-stage phase II oncology trials. *Biometrical Journal*, 54:445–456.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55:1286–1290.

Levin, G. P., Emerson, S. C., and Emerson, S. S. (2012). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Statistics in Medicine*, DOI 10.1002/sim.5662.

Li, Y., Mick, R., and Heitjan, D. F. (2012). A Bayesian approach for unplanned sample sizes in phase II cancer clinical trials. *Clinical Trials*, 9:293–302.

Lin, S. P. and Chen, T. T. (2000). Optimal two-stage designs for phase II clinical trials

with differentiation of complete and partial responses. *Communications in Statistics – Theory and Methods*, 29:923–940.

Lin, X., Allred, R., and Andrews, G. (2008). A two-stage phase II trial design utilizing both primary and secondary endpoints. *Pharmaceutical Statistics*, 7:88–92.

Lin, Y. and Shih, W. J. (2004). Adaptive two-stage designs for single-arm phase IIa cancer clinical trials. *Biometrics*, 60:482–490.

Liu, G. F., Zhu, G. R., and Cui, L. (2008). Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Statistics in Medicine*, 27:584–596.

Liu, Q., Proschan, M. A., and Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97:1034–1041.

London, W. B. and Chang, M. N. (2005). One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine*, 24:2597–2611.

Mander, A. and Thompson, S. (2010). Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary Clinical Trials*, 31:572–578.

Mander, A. P., Wason, J. M. S., Sweeting, M. J., and Thompson, S. G. (2012). Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics*, 11:91–96.

Mariani, L. and Marubini, E. (1996). Design and analysis of phase II cancer trials: A review of statistical methods and guidelines for medical researchers. *International Statistical Review*, 64:61–88.

McPherson, K. (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine*, 1:25–36.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57:886–891.

Müller, H.-H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*, 23:2497–2508.

Nemhauser, G. L. and Wolsey, L. A. (1999). *Integer and Combinatorial Optimization*. Wiley, Hoboken. ISBN 978-0471359432.

O'Brian, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199.

Pong, A. and Chow, S.-C., editors (2010). *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. Chapman & Hall / CRC, Boca Raton. ISBN 978-1439810163.

Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal*, 41:689–696.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., and Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*, 24:3697–3714.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51:1315–1324.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org`.

Sargent, D., Chang, V., and Goldberg, R. M. (2001). A three-outcome design for phase II clinical trials. *Controlled Clinical Trials*, 22:117–125.

Schäfer, H., Timmesfeld, N., and Müller, H.-H. (2006). An overview of statistical approaches for adaptive designs and design modifications. *Biometrical Journal*, 48:507–520.

Shih, W. J. (2006). Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison. *Statistics in Medicine*, 25:933–941.

Shuster, J. (2002). Optimal two-stage designs for single arm phase II cancer trials. *Journal of Biopharmaceutical Statistics*, 12:39–51.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10:1–10.

Stewart, D. J. (2010). Randomized phase II trials: Misleading and unreliable. *Journal of Clinical Oncology*, 28:e649–e650.

Stone, A., Wheeler, C., and Barge, A. (2007). Improving the design of phase II trials of cytostatic anticancer agents. *Contemporary Clinical Trials*, 28:138–145.

Tan, M. T. and Xiong, X. (2010). A flexible multi-stage design for phase II oncology trials. *Pharmaceutical Statistics*, 10:369–383.

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L.,

Verweij, J., Van Glabbeke, M., van Oosterom, A. T., Christian, M. C., and Gwyther, S. G. (2000). New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92:205–216.

Timmesfeld, N., Schäfer, H., and Müller, H.-H. (2007). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine*, 26:2449–2464.

Tournoux-Facon, C., Rycke, Y. D., and Tubert-Bitter, P. (2011). How a new stratified adaptive phase II design could improve targeting population. *Statistics in Medicine*, 30:1555–1562.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90:367–378.

Tsimberidou, A.-M., Braiteh, F., Stewart, D. J., and Kurzrock, R. (2009). Ultimate fate of oncology drugs approved by the US Food and Drug Administration without a randomized trial. *Journal of Clinical Oncology*, 27:6243–6250.

Vandemeulebroecke, M. (2006). An investigation of two-stage tests. *Statistica Sinica*, 16:933–951.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–199.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9:65–71; discussion 71–2.

Wolsey, L. A. (1998). *Integer Programming*. Wiley, Hoboken. ISBN 978-0471283669.

Wu, Y. and Shih, W. J. (2008). Approaches to handling data when a phase II trial deviates from the pre-specified Simon's two-stage design. *Statistics in Medicine*, 27:6190–6208.

Index

Curriculum Vitae

Stefan Englert
born 3rd May 1986 in Schweinfurt, Germany

Education

University of Heidelberg Doctoral student (Dr. sc. hum.)	Since 02/2011
University of Würzburg Diploma in mathematics (Dipl.-Math.) Thesis: "Species richness estimation"	09/2005 – 02/2010
Alexander-von-Humboldt-Gymnasium Schweinfurt University-entrance Diploma (Abitur)	09/1996 – 06/2005
Grundschule Euerbach	09/1992 – 08/1996

Professional experience

University of Heidelberg Research assistant at the Institute of Medical Biometry and Informatics	Since 03/2010
University of Würzburg Member of a student initiative on statistical consulting	07/2008 – 02/2010
University of Würzburg Teaching assistant at the Chair of Statistics at the Faculty of Mathematics and Computer Science	04/2008 – 02/2010
University of Würzburg Student assistant at the Faculty of Mathematics and Computer Science	04/2007 – 03/2009

Publications

Methodological articles

Englert, S. and Kieser, M. (2012a). Adaptive designs for single-arm phase II trials in oncology. *Pharmaceutical Statistics*, 11:241–249.

Englert, S. and Kieser, M. (2012b). Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics*, 68:886–892.

Englert, S. and Lorenzo Bermejo, J. (2011). Book review – handbook of adaptive deigns in pharmaceutical and clinical development. *Biometrical Journal*, 53:708–709.

Englert, S. and Kieser, M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biometrical Journal* (under review).

Articles

Domschke, C., Diel, I., **Englert, S.**, Kalteisen, S., Mayer, L., Rom, J., Sohn, C., and Schuetz, F. (2013). Prognostic value of disseminated tumor cells in the bone marrow of patients with operable primary breast cancer: A long-term follow up study. *Annals of Surgical Oncology*, DOI 10.1245/s10434-012-2814-4.

Dobner, B. C., Riechardt, A. I., Joussen, A. M., **Englert, S.**, and Bechrakis, N. E. (2012). Expression of haematogenous and lymphogenous chemokine receptors and their ligands on uveal melanoma in association with liver metastasis. *Acta Ophthalmologica*, 90:e638–e644.

Rahbari, N. N., Lordick, F., Fink, C., Bork, U., Stange, A., Jager, D., Luntz, S. P., **Englert, S.**, Rossion, I., Koch, M., Büchler, M. W., Kieser, M., and Weitz, J. (2012). Resection of the primary tumour versus no resection prior to systemic therapy in patients with colon cancer and synchronous unresectable metastases (UICC stage IV): SYNCHRONOUS – A randomised controlled multicentre trial. *BMC Cancer*, 12:142.

Schwenk, M., Gogulla, S., **Englert, S.**, Czempik, A., and Hauer, K. (2012). Test-retest reliability and minimal detectable change of repeated sit-to-stand analysis using one body fixed sensor in geriatric patients. *Physiological Measurement*, 33:1931–1946.

Krämer, N., **Englert, S.**, Michel, R., Petschelt, A., and Frankenberger, R. (2011a). Zahngesundheit bayerischer schulkinder. Bayerisches Zahnärzteblatt.

Krämer, N., Michel, R., **Englert, S.**, Petschelt, A., and Frankenberger, R. (2011b). Zahngesundheit bayerischer schulkinder 2009. *Oralprophylaxe & Kinderzahnheilkunde*, 34:74–82.

Monographs

Falk, M., Marohn, F., Michel, R., Hofmann, D., Macke, M., Tewes, B., Dinges, P., Spachmann, C., and **Englert, S.** (2011). *A First Course on Time Series Analysis : Examples with SAS*. Epubli GmbH, Berlin. 2012.august.01 edition. ISBN 978-3-8442-2845-8.

Ortseifen, C., Ramroth, H., Weires, M., and Minkenberg, R., editors (2011). *Proceedings der 15. Konferenz der SAS-Anwender in Forschung und Entwicklung (KSFE)*, KSFE 2011 – Voneinander Lernen. Shaker Verlag, Aachen. (book contribution) ISBN 978-3-8440-0379-6.

Englert, S. (2009). Species richness estimation. Diplomarbeit, University of Würzburg. URN urn:nbn:de:bvb:20-opus-71362.

Abstracts

Englert, S. and Kieser, M. An approach for unplanned sample size changes in one-armed phase II cancer clinical trials. *3rd Joint Statistical Meeting DAGStat 2013*. Freiburg

Englert, S. and Kieser, M. Evaluation of sample size adaptation rules in clinical studies aiming at an overall performance optimization. *Adaptive Designs And Multiple Testing Procedures Workshop 2012*. Heidelberg

Englert, S. Adaptive designs. *6. Herbsttagung der Deutschen Gesellschaft für Allgemein- und Viszeralchirurgie 2012*. Hamburg (Invited speaker)

Wirths, M., **Englert, S.** and Kieser, M. R Paket zur Planung und Auswertung einarmiger onkologischer Phase-II-Studien (poster) *57. GMDS-Jahrestagung 2012*. Braunschweig

Bruckner, T., Rochon, J. and **Englert, S.** Analysis of safety data using SAS (poster). *33rd Annual Meething of the Society for Clinical Trials, SCT 2012*. Miami

Englert, S. and Kieser, M. Evaluation of sample size adaptation rules in clinical studies aiming at an overall performance optimization. *58. Biometrisches Kolloquium 2012*. Berlin

Kieser, M., **Englert, S.** and Stucke, K. Innovative Methoden und Anwendungen zur Fallzahlplanung für klinische Studien. *58. Biometrisches Kolloquium 2012*. Berlin

Englert, S. "A First Course on Time Series Analysis with SAS" – an Open-Source Book Project (poster). *Time Series Workshop 2012*. Karlsruhe

Englert, S. and Kieser, M. Verbesserung der Effizienz adaptiver Designs bei diskreten Teststatistiken. *56. GMDS-Jahrestagung 2011*. Mainz

Englert, S. and Kieser, M. Improving adaptive group sequential designs with discrete outcomes. *2nd Conference of the Central European Network, CEN 2011.* Zürich

Englert, S. and Kieser, M. Evaluating the efficiency of adaptive two-stage designs with discrete test statistics. *Adaptive Designs And Multiple Testing Procedures Workshop 2011.* Lancaster

Englert, S. Empirische Poweranalysen. *15. Konferenz der SAS® Anwender in Forschung und Entwicklung, KSFE 2011.* Heidelberg

Englert, S. and Kieser, M. Adaptive two-stage designs for single-arm trials with discrete test statistics. *Adaptive Designs And Multiple Testing Procedures Workshop 2010.* Wien

Englert, S. and Kieser, M. Adaptive Designs für Phase-II-Studien in der Onkologie. *55. GMDS-Jahrestagung 2010.* Mannheim

Kieser, M. and **Englert, S.**. Adaptive designs for phase II studies in oncology. *XXVth International Biometric Conference (IBC) 2010.* Florianápolis, Brasilien

Memberships

German Region of the International Biometric Society	Since 01/2012
Society for Clinical Trials	05/2011 – 12/2012

Peer-Reviewer

Journal of Biopharmaceutical Statistics	Since 07/2012
Pharmaceutical Statistics	Since 05/2011
Statistics in Medicine	Since 06/2010

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Dr. Kieser for his support throughout my PhD study and research. His guidance, motivation, enthusiasm, and expertise in adaptive and flexible designs helped me all the time during research and writing of this thesis.

I thank the Deutsche Forschungsgemeinschaft (DFG) for supporting my research on flexible designs for single-arm phase II trials in oncology by grant KI 708/1-1.

Last but not the least, I would like to thank my colleagues at the Institute of Medical Biometry and Informatics, University of Heidelberg.