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

Choice of the Interim Analysis Timing in Adaptive Enrichment Designs

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum
an der Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von

Laura Benner (geb. Kohlhas)
aus Hachenburg

2019

Dekan: Prof. Dr. med. Andreas Draguhn
Doktorvater: Prof. Dr. sc. hum. Meinhard Kieser

Contents

Abbreviations and Symbols

α	Significance level
ASS	Average sample size
β	Type II error probability
C	Control group
c	Threshold value for selection rule based on estimated effect differences
c_0, c_+	Threshold values for selection rule based on absolute effect estimates
CP	Conditional power
Δ_0	Treatment effect in the total patient population
Δ_+	Treatment effect in the subgroup (with the higher expected benefit)
Δ_-	Treatment effect in the complementary group
$\widehat{\Delta}_0$	Treatment effect estimator in the total population in stage I
$\widehat{\Delta}_+$	Treatment effect estimator in the subgroup in stage I
$\widehat{\Delta}_0^{II}$	Treatment effect estimator in the total population in stage II
$\widehat{\Delta}_+^{II}$	Treatment effect estimator in the subgroup in stage II
$\Delta_{0,A}$	Assumed effect in G_0 under the alternative in the planning stage
$\Delta_{+,A}$	Assumed effect in G_+ under the alternative in the planning stage
$\Delta_{0,\tilde{A}}$	Assumed effect in G_0 under the alternative after stage I
$\Delta_{+,\tilde{A}}$	Assumed effect in G_+ under the alternative after stage I
$Erf(\cdot)$	Error function
$Erfi(\cdot)$	Imaginary error function
EMA	European Medicines Agency

FDA	Food and Drug Administration
G_0	Total patient population
G_+	Subgroup with the higher expected benefit
G_-	Complementary group
$H_0^{(0+)}$	Global null hypothesis
$H_1^{(0+)}$	Global alternative hypothesis
$H_0^{(0)}$	Null hypothesis for G_0
$H_1^{(0)}$	Alternative hypothesis for G_0
$H_0^{(+)}$	Null hypothesis for G_+
$H_1^{(+)}$	Alternative hypothesis for G_+
HER2	Human epidermal growth factor receptor-2
n	Overall sample size per group
n^I	Sample size per group in stage I
n^{II}	Sample size per group in stage II
n_0^I	Sample size per group in G_0 in stage I
n_0^{II}	Sample size per group in G_0 in stage II
n_+^I	Sample size per group in G_+ in stage I
n_+^{II}	Sample size per group in G_+ in stage II
n_{fix}	Sample size in the non-adaptive design
n_{max}^{II}	Upper limit of recalculated sample size in stage II
n_{min}^{II}	Lower limit of recalculated sample size in stage II
$\Phi(\cdot)$	Standard normal distribution function
$\Phi^{-1}(\cdot)$	Inverse of the standard normal distribution function
p	Prevalence of G_+
SD	Standard deviation
T	Treatment group
Th2	Type 2 helper T-cell
t	Timing of the interim analysis
Z_0^I	Stage I test statistic for the assessment of $H_0^{(0)}$
Z_+^I	Stage I test statistic for the assessment of $H_0^{(+)}$
Z_0^{II}	Stage II test statistic for the assessment of $H_0^{(0)}$
Z_+^{II}	Stage II test statistic for the assessment of $H_0^{(+)}$
Z_0	Combined test statistic for the assessment of $H_0^{(0)}$
Z_+	Combined test statistic for the assessment of $H_0^{(+)}$
Z_{0+}^I	Stage I test statistic for the assessment of $H_0^{(0+)}$
$Z_{0+}^{(0)}$	Combined test statistic for the assessment of $H_0^{(0+)}$ if G_0 is selected

$Z_{0+}^{(+)}$	Combined test statistic for the assessment of $H_0^{(0+)}$ if G_+ is selected
Z_{0+}^{II}	Stage II test statistic for the assessment of $H_0^{(0+)}$
Z'_{0+}	Combined test statistic for the assessment of $H_0^{(0+)}$ if both populations are selected
$z_{1-\alpha/2}$	$(1-\alpha/2)$-quantile of the standard normal distribution

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In recent years, interest in targeted therapies has increased greatly. Knowledge of the molecular basis of a disease and the treatment's mode of action often allows to specify a subset of patients based on genetic or molecular biomarkers who are expected to have an increased benefit from the treatment. Especially, in oncology, targeted therapies are developed more and more frequently (see, e.g., Pérez-Herrero and Fernández-Medarde, 2015). An example is the treatment of breast cancer with trastuzumab, for which efficacy was only shown for patients with HER2-positive tumors. An overexpression of the human epidermal growth factor receptor-2 (HER2) is known to play an important role in the development of breast cancer. Trastuzumab binds to the HER2 receptor and inhibits the growth of HER2-overexpressing breast cancer cells (Baselga, 2001). Another example from the field of pneumonology, which is considered later in this thesis, is the treatment of lebrikizumab for patients with asthma that is more efficient in specific subgroups (Corren et al., 2011).

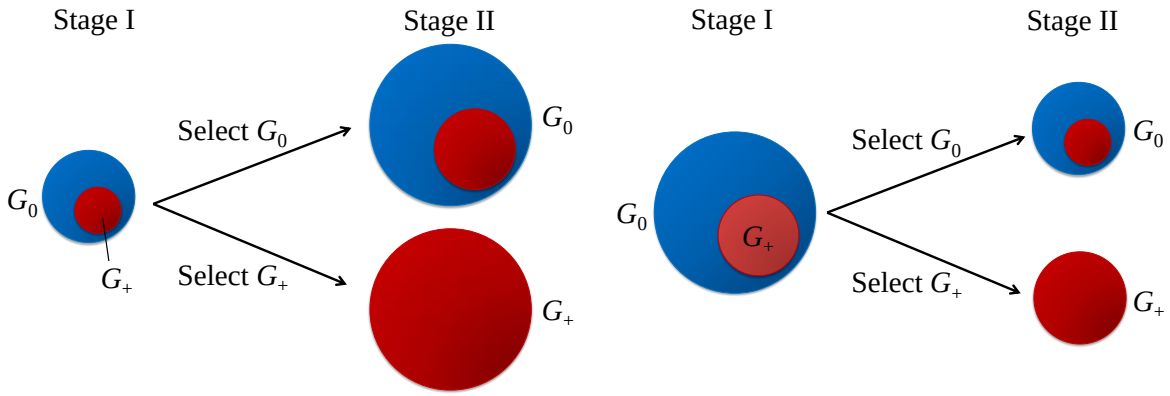
If the benefit of a treatment depends on individual characteristics of a patient, study designs allowing to demonstrate efficacy in particular subgroups of the overall patient population become more important. In case of conjecturing a higher treatment effect in a specific subgroup, the traditional approach consists of two separate clinical trials. In a phase II study, patients from the total population are enrolled, and estimated treatment effects in both the subgroup and the total population are used to select the target population with the most promising benefit. In a subsequent phase III trial, only patients from the selected target population are recruited and its treatment effect is assessed based on the data obtained from the phase III study. However, this approach is very time consuming and resource intensive since selection of the target population and assessment of efficacy is divided into two separate trials.

As a less time consuming and more efficient approach, so-called adaptive enrichment designs have been proposed combining both selection of the patient population and confirmatory assessment of the treatment effect in one trial (see, e.g., Wang et al., 2007; Jenkins et al., 2011). Thereby, patients from the total population are enrolled in the first stage of the study. Then, an interim analysis is conducted to select the target population with the most promising treatment benefit. Depending on the selected target population, patients from the total population are enrolled in the second stage, or recruitment in the second stage is restricted to patients from the subgroup only. In the final analysis, data from both stages are combined for the investigation of efficacy.

1.2 Aim and Structure of the Thesis

In recent years, various statistical methods for adaptive enrichment designs have been developed and proposed in the literature, e.g., different rules for the selection of the target population and several methodologies related to the control of the type I error rate (see, e.g., Wang et al., 2007; Brannath et al., 2009; Jenkins et al., 2011; Friede et al., 2012). However, the choice of the interim analysis timing has not been very well investigated yet. One possible *ad hoc* strategy would be to conduct the interim analysis after half of the patients are enrolled. However, this is rather a rule of thumb and is not based on any statistical considerations. Moreover, the timing of the interim analysis in an adaptive enrichment design has a substantial impact on the composition of the study populations if the subgroup is chosen in the interim analysis. If the subgroup is selected as target population in an early interim analysis, the study contains overall substantially more patients from the subgroup as compared to the case when the subgroup is selected in a late interim analysis. A schematic representation of an early and a late interim analysis (or, in other words, a small and, respectively, a large sample size in the first stage) is shown in Figure 1.1.

Based on heuristic considerations, it is clear that a very early conduct of the interim analysis might be inappropriate since selection of the target population is then based on a small data set and the probability of selecting the wrong population is rather high which may have severe consequences. For example, if the subgroup is selected as target population but there is also a relevant treatment effect in the complementary group, the treatment may be denied to this patient group. Also, if the total population is erroneously selected but there is only a relevant effect in the subgroup, the study might fail to prove efficacy, which is especially likely if the prevalence of the subgroup is small. On the other hand, a very late interim analysis does not seem to be sensible either. In case the subgroup is selected at a late interim analysis, it is no longer possible to substantially



(a) early interim analysis (b) late interim analysis

Figure 1.1: Schematic illustration of an early and a late interim analysis using an adaptive enrichment design with fixed overall sample size. The total patient population is denoted by G_0 and the subgroup by G_+.

enrich the study population with patients from the subgroup. Nevertheless, it is not clear if the interim analysis timing after the enrollment of half the patients is uniformly the best choice.

The aim of this thesis is to investigate the choice of the interim analysis timing in adaptive enrichment designs for the situation of a normally distributed outcome. In the first part, a design with fixed overall sample size specified at the beginning of the trial is considered. Here, the impact of the timing on the power of the study is investigated for various effect sizes, prevalences of the subgroup, and different selection rules. The aim of these investigations is to assess to what extent the timing influences the power of the study in general, and which timings are favorable or unfavorable with regard to the power of the trial.

In the second part of the thesis, scenarios are considered in which the overall sample size is not fixed. Instead, the sample size of the second stage is recalculated based on the treatment effect observed in the interim analysis. In this case, characteristics of sample size distribution are compared for different interim analysis timings. Here, it is investigated to what extent the timing influences the distribution of the sample size and which timing is appropriate regarding the average sample size for scenarios with different effect sizes, prevalences of the subgroup, and selection rules.

Overall, the thesis is structured as follows: In Chapter 2, the theoretical framework on adaptive enrichment designs is given. Chapter 3 gives the basic notation and describes the considered design together with the two different classes of selection rules for selecting the target population. In Chapter 4, the choice of the interim analysis timing is investigated if the overall sample size is fixed. As an alternative, in Chapter 5, the

impact of the interim analysis timing is examined for an adaptive enrichment design with sample size reassessment. Chapter 6 concludes with a discussion, and a summary of the thesis is given in Chapter 7.

Results concerning the adaptive enrichment design with fixed overall sample size were already published in Benner and Kieser (2018). All simulations were performed using R, version 3.5.0 (R Core Team, 2018), and corresponding program code is provided in the Appendix.

Chapter 2

Theoretical Framework

Beside the use of adaptive enrichment designs, adaptive designs in general are very popular as they allow flexible modifications of the design in an ongoing study under control of the type I error rate.
An overview of the development of methodologies for adaptive designs is given by Bauer et al. (2016). After the first seminal publications of Bauer (1989) and Bauer and Köhne (1994), various methods were developed. For example, strategies were proposed to assure control of the type I error rate within an adaptive design. In general, two different approaches exist. The first strategy is based on the combination of p-values or test statistics obtained from the two stages of the study. Commonly used combination methods are Fisher's combination test (Bauer, 1989; Bauer and Köhne, 1994), where the product of both p-values is compared to a critical value, or the weighted inverse normal combination function (Lehmacher and Wassmer, 1999), where statistics from both stages (p-values or test statistics) are weighted and combined. The latter method is used in this work and is described in detail in the subsequent chapter. The second commonly used strategy to assure control of the type I error rate was proposed by Proschan and Hunsberger (1995) and is based on the conditional error function.
One important application of adaptive designs is the reassessment of sample size in the interim analysis. Since assumed treatment effects and standard deviations are often vague in the planning phase for calculating the sample size, the data observed in the first stage can be used to adjust the sample size of the subsequent stage. For example, the sample size can be adjusted upwards if the observed effect size in the interim analysis is smaller than expected, and downwards if the effect size is higher than expected. A commonly used method for sample size reassessment is based on conditional power arguments (Proschan and Hunsberger, 1995), and is applied in this work for sample size recalculation using an adaptive enrichment design.
However, not only the sample size can be modified using an adaptive design. The

flexibility of adaptive designs find a variety of applications. For example, EMA and FDA mention in their guidelines (European Medicines Agency, 2007; Food and Drug Administration, 2018) the adaptation of sample size, the allocation ratio, change of endpoints, selection of the most promising treatments, and selection of subgroups. Also a stop for futility or efficacy is possible if observed effects in the interim analysis provide already sufficient information. In this case, the trial is stopped early after the interim analysis either with acceptance or rejection of the null hypothesis, and the sample size can be reduced.

One useful application in the field of stratified medicine is the adaptive enrichment design, which offers the possibility to select the target population in the interim analysis and accordingly, adapt the population from which patients in the second stage are enrolled. While in general it would be possible to consider multiple subgroups (see, e.g., Wassmer and Dragalin, 2015), this work solely considers the situation of a single subgroup G_+ in which a higher treatment effect is assumed compared to the total population G_0.

The basic process of a study using an adaptive enrichment design is as follows: Patients from the total population G_0 are enrolled in the first stage of the study. After completion of the first stage, based on the data observed so far, an interim analysis is conducted, where the target population (G_+ or G_0) with the most promising treatment effect is selected. Depending on the selected target population in the interim analysis, patients from the total population are enrolled in the second stage, or recruitment is restricted to patients from the subgroup only, which is referred to as enrichment. In the final analysis, data from both stages are combined for the investigation of efficacy. In this way, the selection of the target population and the test for efficacy is combined in a single trial consisting of two stages, which is less resource intensive and less time consuming than conducting two separate trials.

In recent years, several approaches and methodologies for applying adaptive enrichment designs were proposed. To control the type I error rate, most approaches make use of combination tests as, for example, described in Wang et al. (2007), Wang et al. (2009) or Jenkins et al. (2011) but also applying the conditional error function approach is possible as implemented by Friede et al. (2012). The adjustment for multiplicity arising from the two considered populations can be handled using commonly methods such as Bonferroni correction or the closure principle.

Especially, there are various ways to define the rule for selecting the target population. Besides Bayesian decision tools, as described in Brannath et al. (2009), several simple selection rules were proposed based on the difference between effect sizes or comparing effect sizes to a pre-specified threshold value. In this thesis, two different classes are considered, which are both based on the estimated treatment effects in G_0 and G_+

calculated with data observed in the first stage. The first selection rule is based on the difference between standardized effect estimates and is a variant of the ϵ-selection rule proposed by Kelly et al. (2005) and Friede and Stallard (2008) in the context of treatment selection. If the estimated effect in the subgroup is larger than the effect in the total population by a pre-specified amount, only the subgroup is selected; otherwise, patients from the total population are enrolled in the second stage of the trial. The second selection rule, that is considered in this thesis, is based on absolute effect estimates and was originally proposed by Jenkins et al. (2011) for time-to-event endpoints. Using this selection rule, the estimated effect sizes in both populations are compared to pre-specified threshold values. Using this selection rule, either a single population is selected, both populations are selected, or the trial is stopped early for futility.

It should be noted that most adaptive enrichment designs are based on the assumption of an increased treatment benefit in an already pre-defined subgroup. This is usually the case if the biomarker defining the subgroup has already been validated in prior trials. If there is a high uncertainty regarding the actual tailoring of the subgroup, designs which allow to define the subgroup based on data from the current trial might be more advisable. Two examples are the design by Renfro et al. (2014), where a biomarker cutoff is determined at interim, and the design by Chen et al. (2016), where part of the trial data is used to potentially modify design elements such as the investigated subgroup. In this thesis, however, the situation of a pre-defined subgroup which will not be subject to any data-driven alterations is considered.

Chapter 3

Methods

3.1 Design and Notation

In this thesis, a parallel-group clinical trial with an adaptive two-stage enrichment design is considered, where a higher treatment benefit is expected in a prespecified subpopulation. The total population is denoted by G_0 and the subgroup with the higher expected benefit by $G_+ \subset G_0$. The complement is indicated by G_-, and the prevalence of G_+ is given by p. It is assumed that the prevalence in the total patient population and the study population is equal, and that the prevalence in the study population is fixed and not variable. Furthermore, in both populations, equal allocation to both treatment groups is assumed. The overall sample size per treatment group is denoted by n, the sample size in stage I is given by n^I and in stage II by n^{II}. The number of patients per group from the subgroup in the first stage is given by $n_+^I = pn^I$. In the second stage, the number of patients from the subgroup is

$$n_+^{II} = \begin{cases} n^{II} & \text{if } G_+ \text{ is selected} \\ pn^{II} & \text{if } G_0 \text{ is selected.} \end{cases}$$

Throughout this thesis, a normally distributed outcome is considered. Therefore, the independent random samples for treatment and control group

$$X_{T+i} \sim \mathcal{N}(\mu_{T+}, \sigma_+),\ i = 1, ..., n_+^I + n_+^{II}$$
$$X_{C+j} \sim \mathcal{N}(\mu_{C+}, \sigma_+),\ j = 1, ..., n_+^I + n_+^{II}$$

are assumed for the biomarker positive subgroup, where μ_{T+} and μ_{C+} are the means in the treatment and control group in G_+, and σ_+ is the common known standard deviation.

Similarly, the independent random samples

$$X_{T-k} \sim \mathcal{N}(\mu_{T-}, \sigma_-),\ k = 1, ..., n - n_+^I - n_+^{II}$$
$$X_{C-l} \sim \mathcal{N}(\mu_{C-}, \sigma_-),\ l = 1, ..., n - n_+^I - n_+^{II}$$

are assumed for the biomarker negative subgroup, where μ_{T-} and μ_{C-} are the means in the treatment and control group in G_-, and σ_- is the common known standard deviation. The standardized treatment effects are defined as $\Delta_+ = (\mu_{T+} - \mu_{C+})/\sigma_+$ in G_+ and $\Delta_- = (\mu_{T-} - \mu_{C-})/\sigma_-$ in G_-. Thereby, the treatment effect in G_0 is given by $\Delta_0 = p\Delta_+ + (1-p)\Delta_-$. Estimated treatment effects from the first-stage data in population G_0 and G_+ are denoted by $\widehat{\Delta}_0$ and $\widehat{\Delta}_+$, respectively, and are calculated by

$$\widehat{\Delta}_+ = \frac{1}{\sigma_+ n_+^I} \left(\sum_{i=1}^{n_+^I} X_{T+i} - \sum_{j=1}^{n_+^I} X_{C+j} \right)$$
$$\widehat{\Delta}_0 = p\widehat{\Delta}_+ + \frac{1}{\sigma_- n^I} \left(\sum_{k=1}^{(1-p)n^I} X_{T-k} - \sum_{l=1}^{(1-p)n^I} X_{C-l} \right)$$

using the maximum likelihood estimator. In the second stage, estimated treatment effects are denoted by $\widehat{\Delta}_0^{II}$ for G_0 and $\widehat{\Delta}_+^{II}$ for G_+, and are calculated analogously:

$$\widehat{\Delta}_+^{II} = \frac{1}{\sigma_+ n_+^{II}} \left(\sum_{i=n_+^I+1}^{n_+^I+n_+^{II}} X_{T+i} - \sum_{j=n_+^I+1}^{n_+^I+n_+^{II}} X_{C+j} \right)$$
$$\widehat{\Delta}_0^{II} = p\widehat{\Delta}_+^{II} + \frac{1}{\sigma_- n^{II}} \left(\sum_{k=(1-p)n^I+1}^{n-n_+^I-n_+^{II}} X_{T-k} - \sum_{l=(1-p)n^I+1}^{n-n_+^I-n_+^{II}} X_{C-l} \right).$$

However, $\widehat{\Delta}_0^{II}$ only exists if G_0 is selected as target population.
Furthermore, for the two considered populations, different hypotheses are formulated. For convenience, it is assumed that higher values for the outcome are related to favourable results. Hence, the following one-sided hypotheses are considered:

$$H_0^{(0)} : \Delta_0 \leq 0 \quad \text{versus} \quad H_1^{(0)} : \Delta_0 > 0$$

for the effect in the total population, and

$$H_0^{(+)} : \Delta_+ \leq 0 \quad \text{versus} \quad H_1^{(+)} : \Delta_+ > 0$$

for the effect in the subgroup. To cover the aim of rejecting at least one of the hypotheses

$H_0^{(0)}$ or $H_0^{(+)}$, the following hypotheses are considered:

$$H_0^{(0+)} : \Delta_0 \leq 0 \ \cap \ \Delta_+ \leq 0 \quad \text{versus} \quad H_1^{(0+)} : \Delta_0 > 0 \ \cup \ \Delta_+ > 0.$$

The timing of the interim analysis t is defined by the ratio of number of patients in the first stage and the overall sample size, i.e. $t = n^I/n$, yielding possible values for t between 0 and 1.

3.2 Selection Rules

In the following, two different classes of selection rules are described, that are considered in this thesis. For a detailed assessment of the statistical properties of the two classes of rules and their performance in trials with subgroup selection, see, e.g. Krisam and Kieser (2014).

3.2.1 Selection Rule Based on Estimated Effect Differences

The first selection rule is a variant of the ϵ-selection rule proposed by Kelly et al. (2005) and Friede and Stallard (2008) in the context of treatment selection. In the originally proposed approach, the most promising treatments are selected from a pool of several different treatments, and the predefined difference is related to the treatment showing the largest effect in the interim analysis.

This selection rule is transferred to the situation of an adaptive enrichment design which is based on the design described by Wang et al. (2007), where a higher treatment benefit is expected in a prespecified subpopulation. In the interim analysis it is decided if the total population or the subpopulation is considered as the target population defining which patients are recruited in the second stage of the trial and which hypothesis is tested in the final analysis.

The selection rule for selecting the target population considered in this thesis is based on the difference between estimated effect sizes from the interim analysis in the total patient population and the subgroup, i.e. $\widehat{\Delta}_+ - \widehat{\Delta}_0$. This difference is compared to a predefined constant c. If $\widehat{\Delta}_+ - \widehat{\Delta}_0 \leq c$, patients from G_0 are enrolled in the second stage and $H_0^{(0)}$ is tested at the end of the trial. If $\widehat{\Delta}_+ - \widehat{\Delta}_0 > c$, patients from the subgroup only are enrolled in the second stage and the null hypothesis $H_0^{(+)}$ is tested in the final analysis. A schematic illustration of the application of this selection rule is shown in Figure 3.1.

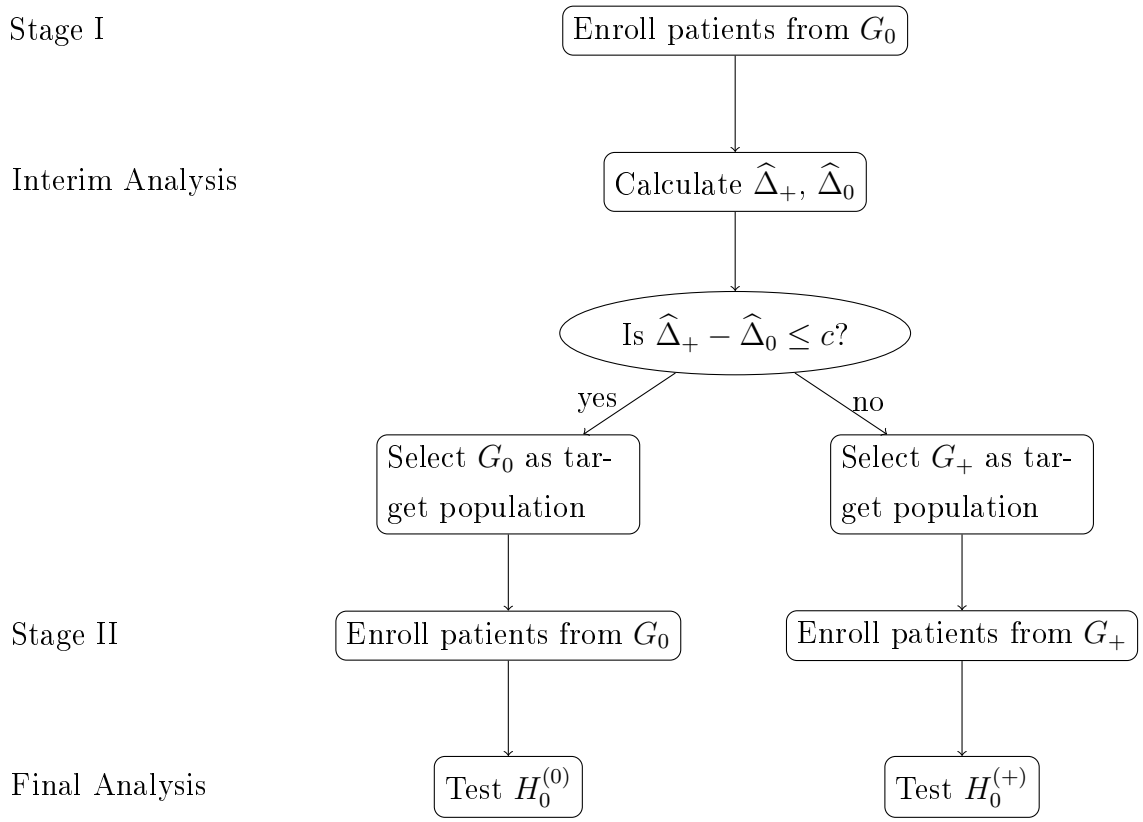


Figure 3.1: Flow chart of an adaptive enrichment design using the selection rule based on estimated effect differences

It should be noted that it also would be possible to test both hypotheses in the final analysis. If G_0 is selected and $H_0^{(0)}$ is tested, data from both stages can be used to test also $H_0^{(+)}$. If G_+ is selected, only data from the first stage can be used to test $H_0^{(0)}$. However, investigations in this thesis are restricted to the case where only the hypothesis related to the selected population is tested since this hypothesis should be of primary interest. Nevertheless, although only one hypothesis is tested in the final analysis, the α-level has to be adjusted.

In general, this simple class of selection rules has the disadvantage that the absolute effect sizes within the two target populations are irrelevant. Exactly one hypothesis is tested in the final analysis, thus ignoring whether the observed effects in the interim analysis are both very high (thus justifying to test both hypotheses in the final analysis) or both very small (thus indicating that there is no treatment effect in either of the populations, justifying an early stop for futility). A selection rule overcoming this drawback is presented in the following subsection where the absolute effect sizes are considered.

3.2.2 Selection Rule Based on Absolute Effect Estimates

The second class of selection rules is based on absolute effect estimates. Using this selection rule, two threshold values c_0 and c_+ have to be specified for the continuation or termination of the total population and the subgroup, respectively. This approach was originally proposed by Jenkins et al. (2011) for time-to-event data, where the hazard ratio estimates were compared to predefined target values. In contrast to the selection rule based on the estimated effect differences where either $H_0^{(0)}$ or $H_0^{(+)}$ is tested, the selection rule proposed by Jenkins et al. (2011) additionally includes the option to stop for futility as well as the option to test both hypotheses $H_0^{(0)}$ and $H_0^{(+)}$ at the end of the trial. The latter is also referred to as the co-primary option.

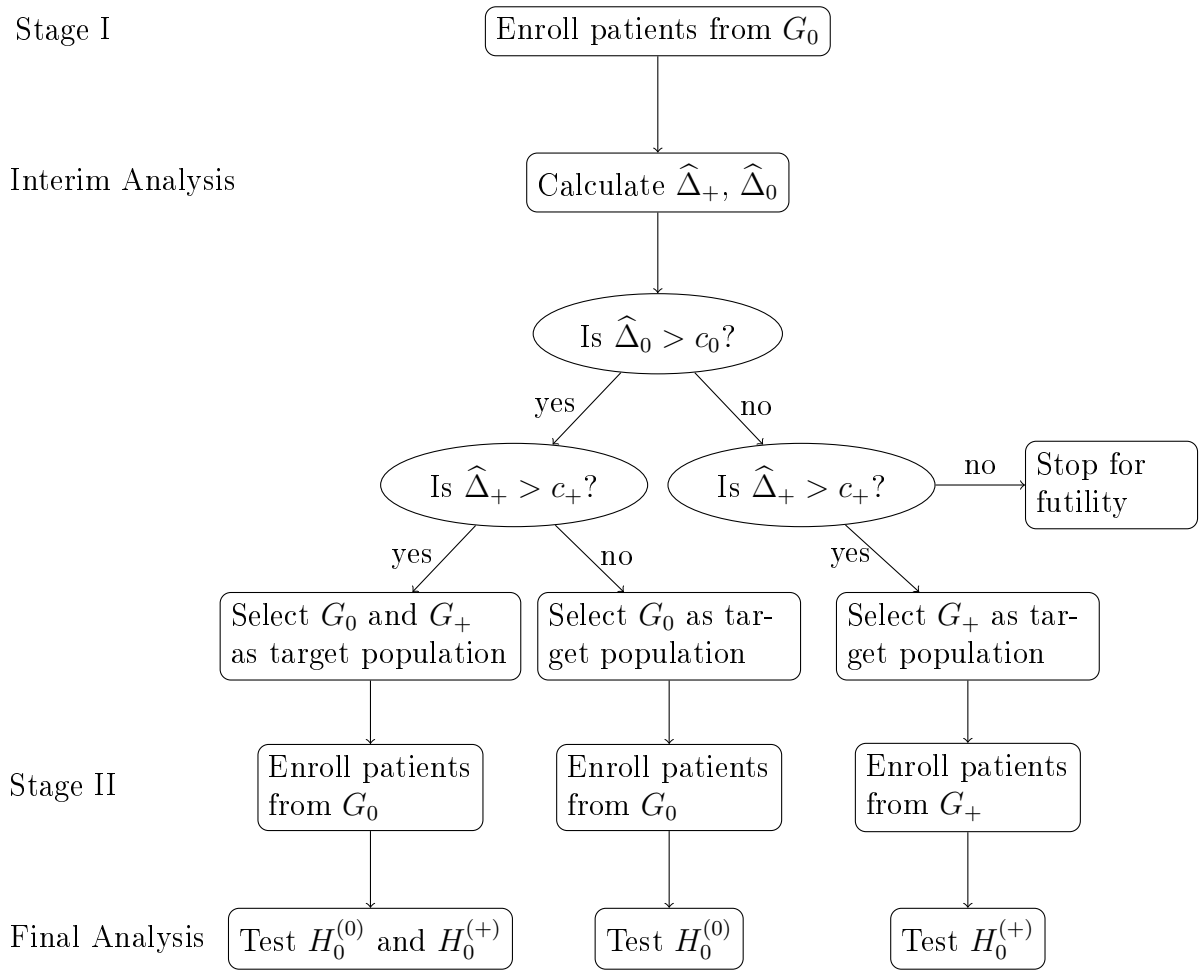


Figure 3.2: Flow chart of an adaptive enrichment design using the selection rule based on absolute effect estimates

The overall procedure that is here merely adapted to normally distributed outcomes is shown in Figure 3.2. Basically, estimated treatment effects $\widehat{\Delta}_0$ and $\widehat{\Delta}_+$ are calculated in the interim analysis and are compared to the predefined threshold values c_0 and c_+, respectively. If only one effect exceeds the threshold, the respective population is selected as target population and the associated hypothesis is tested in the final analysis. If both estimated effects are larger than the respective threshold values, i.e. $\widehat{\Delta}_0 > c_0$ and $\widehat{\Delta}_+ > c_+$ and hence, the treatment is promising for both populations, patients from G_0 are enrolled in the second stage and both hypotheses $H_0^{(0)}$ and $H_0^{(+)}$ are tested in the final analysis. If both effect estimates are below the chosen threshold values, i.e. $\widehat{\Delta}_0 \leq c_0$ and $\widehat{\Delta}_+ \leq c_+$, no second stage is performed and the study is stopped with prematurely accepting both null hypotheses. This option has the advantage that it prevents further patients receiving an ineffective therapy and moreover, resources are saved.
Furthermore, it should be noted that it would also be possible to base the selection of the target population in the interim analysis on a different endpoint than the endpoint used in the final analysis. In the originally proposed decision framework by Jenkins et al. (2011) for time-to-event endpoints, a surrogate endpoint for selecting the target population was used, and the actual primary endpoint was considered in the final analysis. This might especially be useful for time-to-event endpoints where the observation period is long until an event occurs. In this thesis, where a normally distributed endpoint is considered, the endpoint used for subgroup selection is also the endpoint used for hypothesis testing at the end of the trial.

3.3 Testing Procedures

In the following, the testing procedure is presented for the previously described adaptive enrichment design for both selection rules. Different methods exist to handle the multi-stage structure of an adaptive design under control of the type I error rate. In this work, the inverse normal combination method (Lehmacher and Wassmer, 1999) is used. Within this approach, single z-test statistics are calculated for each stage separately and are combined to one statistic in the final analysis. It should be noted that the inverse normal combination method can also be used if z-tests are not appropriate, for example, if the standard deviation is not known, which is usually the case in clinical trials. In this case, the fact is utilized that the transformation $\Phi^{-1}(1-\mathrm{p})$ of any uniformly distributed p-value p is standard normal. Here, Φ^{-1} denotes the inverse of the cumulative standard normal distribution function. However, for sake of simplicity, in this thesis, known standard deviations are assumed and the use of a z-test is considered. In this case, the

single test statistics in the first stage are given by

$$Z_0^I = \widehat{\Delta}_0 \sqrt{\frac{n^I}{2}} \tag{3.3.1}$$

for G_0 and by

$$Z_+^I = \widehat{\Delta}_+ \sqrt{\frac{n_+^I}{2}} \tag{3.3.2}$$

for G_+. In the second stage, the test statistics depend on the selected population. The test statistic for testing $H_0^{(0)}$, which is only available if G_0 is selected, is given by

$$Z_0^{II} = \widehat{\Delta}_0^{II} \sqrt{\frac{n^{II}}{2}}. \tag{3.3.3}$$

The test statistic for G_+ is given by

$$Z_+^{II} = \widehat{\Delta}_+^{II} \sqrt{\frac{n_+^{II}}{2}}. \tag{3.3.4}$$

Using the inverse normal combination method, test statistics from both stages are combined. Weights w_1 and w_2 have to be pre-specified so that the sum of squared weights is equal to 1 ($w_1^2 + w_2^2 = 1$). One reasonable approach is to choose the weights so that the resulting test statistic is equal to the statistic when no interim analysis were performed. This is achieved for weights chosen according to the information time, which leads to the overall test statistic

$$Z_0 = \sqrt{t} Z_0^I + \sqrt{1-t} Z_0^{II} \tag{3.3.5}$$

for testing $H_0^{(0)}$ (only if the total population is selected). The weights w_1 and w_2 in the statistic for testing $H_0^{(+)}$

$$Z_+ = w_1 Z_+^I + w_2 Z_+^{II}, \tag{3.3.6}$$

depend on the selected population. If G_0 is selected, the weights are selected as above using $w_1 = \sqrt{t}$ and $w_2 = \sqrt{1-t}$. When selecting the subgroup only, the sample size in the second stage for G_+ is increased, which is also represented by the weights given by $w_1 = \sqrt{\frac{tp}{tp+1-t}}$ and $w_2 = \sqrt{\frac{1-t}{tp+1-t}}$.
Furthermore, an adjustment for multiplicity has to be conducted since two hypotheses are considered. Even if only one hypothesis is tested in the final analysis, the tested hypothesis was not selected in the planning phase and a multiplicity adjustment is nec-

essary to control the familywise error rate.

Bonferroni correction

A simple method to adjust for multiplicity is the Bonferroni correction, where each hypothesis is tested at level $\alpha/2$ to assure that the familywise error rate does not exceed α. If a one-sided significance level of $\alpha/2$ is used, $H_0^{(+)}$ is rejected if $Z_+ > z_{1-\alpha/4}$ for the case G_+ is selected, where $z_{1-\alpha/4}$ gives the $(1-\alpha/4)$-quantile of the standard normal distribution. In case G_0 is selected, $H_0^{(0)}$ is rejected if $Z_0 > z_{1-\alpha/4}$. If both populations are selected as target population, which is possible using the selection rule based on absolute effect estimates, both hypotheses are independently tested at level $\alpha/4$.
However, this easily applicable adjustment method has the disadvantage of being very conservative, especially for positively correlated test statistics.

Closure Principle

A less conservative approach is to make use of the closure principle (Marcus et al., 1976), which was also applied in the design proposed by Jenkins et al. (2011). Following this method, the single hypothesis ($H_0^{(+)}$ or $H_0^{(0)}$) can be rejected if the intersection hypothesis $H_0^{(0+)}$ and the respective single hypothesis is rejected at level α (or $\alpha/2$ for a one-sided significance level). The intersection hypothesis itself can be tested applying the Simes' procedure (Simes, 1986), which controls the familywise error rate for positively correlated bivariate normally distributed test statistics (Sarkar and Chang, 1997).
If only one hypothesis is tested in the final analysis, which is always the case using the selection rule based on estimated effect differences as described above, Simes' procedure is applied only on the data of the first stage. Thereafter, for testing $H_0^{(0+)}$, the test statistic resulting from the Simes' procedure in the first stage is combined with the test statistic of the selected population from the second stage.
For testing $H_0^{(0+)}$, Simes' procedure controlling the familiy wise error rate $\alpha/2$ for one-sided hypotheses is as follows: P-values related to the test of $H_0^{(0)}$ and $H_0^{(+)}$ are sorted according to size, and $H_0^{(0+)}$ can be rejected if the smaller p-value is less than $\alpha/4$ or the larger p-value is less than $\alpha/2$. Since a test statistic is needed that is combined with the test statistic from the second stage using the inverse normal combination test, Simes' procedure is expressed as

$$Z_{0+}^I = \Phi^{-1}\left(1 - \min[\,2 - 2\Phi(\max(Z_0^I, Z_+^I)), 1 - \Phi(\min(Z_0^I, Z_+^I))\,]\right), \tag{3.3.7}$$

for stage I, where Φ denotes the standard normal distribution function. The combination of stage I and II applied to test the intersection hypothesis using the inverse normal

method is then given by

$$Z_{0+}^{(0)} = \sqrt{t} Z_{0+}^{I} + \sqrt{1-t} Z_{0}^{II}$$

if G_0 is selected and

$$Z_{0+}^{(+)} = \sqrt{t} Z_{0+}^{I} + \sqrt{1-t} Z_{+}^{II}$$

in case G_+ is selected. Finally, $H_0^{(0+)}$ is rejected if $Z_{0+}^{(0)} > z_{1-\alpha/2}$ or $Z_{0+}^{(0)} > z_{1-\alpha/2}$ depending on whether G_0 or G_+ is selected as target population.
In case that both populations are selected in the interim analysis, which is an option using the selection rule based on absolute effect estimates, both hypotheses $H_0^{(+)}$ and $H_0^{(0)}$ are to be tested. The procedure using the closure principle for this case is described in the following: The intersection hypothesis can be tested using

$$Z'_{0+} = \sqrt{t} Z_{0+}^{I} + \sqrt{1-t} Z_{0+}^{II},$$

where Z_{0+}^{I} is defined in formula (3.3.7) and Z_{0+}^{II} is given by

$$Z_{0+}^{II} = \Phi^{-1}\left(1 - \min[\,2 - 2\Phi(\max(Z_0^{II}, Z_+^{II})), 1 - \Phi(\min(Z_0^{II}, Z_+^{II}))\,]\right).$$

Thus, $H_0^{(0+)}$ is rejected if $Z'_{0+} > z_{1-\alpha/2}$.
To summarize, when applying the closure principle with the use of Simes' procedure for testing the intersection hypothesis $H_0^{(0+)}$, the testing procedure is as follows:

Case 1: G_0 is selected as target population:

- Reject $H_0^{(0)}$ if $Z_0 > z_{1-\alpha/2}$ and $Z_{0+}^{(0)} > z_{1-\alpha/2}$

Case 2: G_+ is selected as target population:

- Reject $H_0^{(+)}$ if $Z_+ > z_{1-\alpha/2}$ and $Z_{0+}^{(+)} > z_{1-\alpha/2}$

Case 3: G_0 and G_+ are selected as target population:

- Reject $H_0^{(0)}$ if $Z_0 > z_{1-\alpha/2}$ and $Z'_{0+} > z_{1-\alpha/2}$
- Reject $H_0^{(+)}$ if $Z_+ > z_{1-\alpha/2}$ and $Z'_{0+} > z_{1-\alpha/2}$

It should be noted that the third case is only possible using the selection rule based on absolute effect estimates. Moreover, using this selection rule, a further case is possible, namely the stop for futility, where all null hypotheses are prematurely accepted.

Chapter 4

Design with Fixed Sample Size

In this chapter, the timing of the interim analysis is investigated using an adaptive enrichment design with fixed overall sample size specified in the planning phase of the study. This means that the decision in the interim analysis only determines whether patients enrolled in the second stage originate from the total population or the subgroup, but the overall sample size is not adjusted. For this design, power characteristics are investigated for different interim analysis timings. The power considered here is defined as the probability to reject either $H_0^{(0)}$ or $H_0^{(+)}$.
In Section 4.1, the power function is presented as a mathematical expression for the selection rule based on estimated effect differences. Since this could not be converted into a closed form, simulation studies were used to investigate power characteristics. Results of simulation studies are presented in Section 4.2 for the two different classes of selection rules (based on estimated effect differences, and respectively, based on absolute effect estimates). The main results of the simulation studies presented in this section are taken from Benner and Kieser (2018). In Section 4.3, the impact of the interim analysis on power is investigated for parameters obtained from a real clinical trial examining a therapy for patients with asthma. The chapter closes with a summary in Section 4.4.

4.1 Analytical Derivation of Power Function

In the following, the power function is derived for the selection rule based on estimated effect differences. Due to matters of simplicity, the Bonferroni correction is applied. For solving integrals in the expression of the power function, Mathematica 11.3 (Wolfram Reseach, Inc., 2018) was used.
The aim is to specify the power as a function of the effect sizes Δ_0 and Δ_+, t, p and n. Hence, the sample sizes for different groups and stages are expressed as a function of n, t and p in the following way: $n^I = tn$ and $n_+^I = tnp$ for the sample sizes in stage I, and

$n^{II} = (1-t)n$ for the sample size in stage II. Without loss of generality, it is assumed that $\sigma_+ = \sigma_- = 1$. Theoretically, the derivative with respect to t of this power function could deliver the timing with the maximal power. However, the integral cannot be solved using elementary functions. Some integrals can be expressed with the help of the error function $Erf(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}$ and the imaginary error function $Erfi(x) = -iErf(ix)$, for other integrals, no results could be found in terms of standard mathematical functions. Hence, a power maximum cannot be given analytically. Nevertheless, an analytical derivation of the power function is provided in the following. Results were confirmed comparing the last transformed expression calculated using numerical integration with results from simulation studies.

For calculating the power function, the joint probability density function of $\widehat{\Delta}_0$ and $\widehat{\Delta}_+$ is used (see Krisam and Kieser (2014) for derivation), which is given by

$$(\widehat{\Delta}_0, \widehat{\Delta}_+) \sim \mathcal{N}\left((\Delta_0, \Delta_+), \begin{pmatrix} 2/(tn) & 2/(tn) \\ 2/(tn) & 2/(ptn) \end{pmatrix}\right)$$

as well as the distribution of the effect in the second stage, that is

$$\widehat{\Delta}_0^{II} \sim \mathcal{N}\left(\Delta_0, \frac{2}{(1-t)n}\right)$$

if G_0 is selected, and

$$\widehat{\Delta}_+^{II} \sim \mathcal{N}\left(\Delta_+, \frac{2}{(1-t)n}\right)$$

if G_+ is selected. It can easily be seen that the correlation between $\widehat{\Delta}_0$ and $\widehat{\Delta}_+$ is $Corr(\widehat{\Delta}_0, \widehat{\Delta}_+) = \sqrt{p}$. The joint density function of $\widehat{\Delta}_0$ and $\widehat{\Delta}_+$ is denoted by $f_{\widehat{\Delta}_0, \widehat{\Delta}_+}$, and the densities for the effects in the second stage are denoted by $f_{\widehat{\Delta}_0^{II}}$ and $f_{\widehat{\Delta}_+^{II}}$.

The probability to reject either $H_0^{(0)}$ or $H_0^{(+)}$ is given by the sum of the two probabilities

$$Pr(select\ G_0\ \cap\ reject\ H_0^{(0)}) + Pr(select\ G_+\ \cap\ reject\ H_0^{(+)}).$$

The first probability referring to the rejection of $H_0^{(0)}$ at the end of the trial is given by

$$\begin{aligned} &Pr(reject\ H_0^{(0)}\ \cap\ select\ G_0) \qquad (4.1.1) \\ &= Pr(Z_0 > z_{1-\alpha/4}\ \cap\ \widehat{\Delta}_0 \geq \widehat{\Delta}_+ - c) \\ &= \frac{\sqrt{n(1-t)}}{8\sqrt{\pi}} \left(1 + Erf\left\{\frac{1}{2}\sqrt{\frac{ntp}{1-p}}\,(c + \Delta_0 - \Delta_+)\right\}\right) \end{aligned}$$

$$\cdot \int_{\delta_0^{II}=-\infty}^{\infty} exp\left\{\frac{n}{4}(t-1)\left(\Delta_0-\delta_0^{II}\right)^2\right\}$$
$$\cdot\left(1-Erf\left\{\frac{1}{2}\sqrt{nt}\left(\left(z_{1-\alpha/4}-(1-t)\delta_0^{II}\sqrt{\frac{n}{2}}\right)\frac{\sqrt{2}}{t\sqrt{n}}-\Delta_0\right)\right\}\right)\ d\delta_0^{II}$$

For the last integral, no solution could be found with Mathematica. The detailed calculation of the probability given in (4.1.1) can be found in Appendix A.
The probability for rejecting the hypothesis $H_0^{(+)}$ can be derived in a similar way (see Appendix A for detailed derivation):

$$Pr(Z_+ > z_{1-\alpha/4} \cap \widehat{\Delta}_0 < \widehat{\Delta}_+ - c) \tag{4.1.2}$$
$$= \int_{\delta_+=-\infty}^{\infty} \frac{1}{8}\sqrt{\frac{npt}{\pi}}\cdot exp\left\{-\frac{npt}{4}(\Delta_+-\delta_+)^2\right\}$$
$$\cdot\left(1+Erf\left\{\frac{\sqrt{nt}}{2\sqrt{1-p}}(-\Delta_0+p(\Delta_+-\delta_+)+\delta_+-c)\right\}\right)$$
$$\cdot\left(1+Erf\left\{-\frac{\sqrt{n(1-t)}}{2}\left(-\Delta_+ + \left(z_{1-\alpha/4}\sqrt{\frac{2(tp+1-t)}{n}}-tp\delta_+\right)\frac{1}{1-t}\right)\right\}\right)\ d\delta_+$$

As already seen for the calculation of the first probability, no solution for the last integral could be found. In addition to the presented approach, other strategies were explored, namely using the density of $\widehat{\Delta}_-$ and $\widehat{\Delta}_+$ instead of $f_{\widehat{\Delta}_0,\widehat{\Delta}_+}$ or using a different integration order. In any case, the term could not be transformed into a closed form.
For the selection rule based on absolute effect estimates, the power function can be formulated in a similar way. Here, the overall power is the sum of the probabilities

$$\begin{aligned} &Pr(Z_+ > z_{1-\alpha/4} \cap \widehat{\Delta}_0 \le c_0 \cap \widehat{\Delta}_+ > c_+) \\ +&Pr(Z_0 > z_{1-\alpha/4} \cap \widehat{\Delta}_0 > c_0 \cap \widehat{\Delta}_+ \le c_+) \\ +&Pr\left((Z_+ > z_{1-\alpha/4} \cup Z_0 > z_{1-\alpha/4}) \cap \widehat{\Delta}_0 > c_0 \cap \widehat{\Delta}_+ > c_+\right). \end{aligned}$$

However, also in this case, no solution for the integral can be found and calculations are not presented.

4.2 Simulation Study

4.2.1 Simulation Setup

The power is simulated for different interim analysis timings between 0.05 and 0.95 in increments of 0.025. Furthermore, different scenarios with varying effect sizes, prevalences and selection rules are considered. The standardized effect in the subgroup is fixed to 0.5 and the power for various Δ_- is determined. Values between 0 and 0.5 for Δ_- are considered in steps of 0.05. The influence of the prevalence is investigated for $p = 0.2, 0.4$ and 0.7, and the one-sided significance level $\alpha/2 = 0.025$ is used. For both classes of selection rules, different choices of the threshold values are considered. Applying the selection rule based on estimated effect differences, scenarios are investigated for $c = 0$ and $c = 0.2$, and for the selection rule based on absolute effect estimates, scenarios with $c_+ = 0.1, 0.3$ are considered while c_0 remains constant at 0.1. To adjust for multiplicity, Bonferroni method as well as the closure principle with Simes' procedure for testing the intersection hypothesis was applied. Since results are very similar for both adjustment methods, only output for the latter method is shown. In every scenario, the total sample size is determined such that a power of 80% is achieved at $t = 0.5$. Hence, it is investigated whether the power for $t \neq 0.5$ is higher or smaller than 80% using the same sample size. For every scenario, $1,000,000$ study results are simulated (standard error for a power of 50% equals $5 \cdot 10^{-4}$).
In practice, it is not sensible to do the interim analysis extremely early or extremely late. When conducting the interim analysis very early, the interim decision is based on few data resulting in a high probability to select the wrong population. If the interim analysis is performed towards the end of the trial, the probability to select the correct population is increased but it is not possible anymore to relevantly affect the composition of the study population. For this reason, the focus is on timings between 0.3 and 0.7 and the power range in this interval is considered to specify the variability in power for different timings.
In addition to power considerations, the type I error rate is investigated for three different types of null distributions: the global null hypothesis is true ($\Delta_0 = 0$ and $\Delta_+ = 0$), only $H_0^{(0)}$ is true and only $H_0^{(+)}$ is true. For each null scenario, a sample size of $n = 200$ is chosen.

4.2.2 Selection Rule Based on Estimated Effect Differences

In this section, results are presented for the selection rule based on estimated effect differences, where G_+ is selected if $\widehat{\Delta}_+ - \widehat{\Delta}_0 > c$, and patients from G_0 are enrolled in

stage II otherwise.

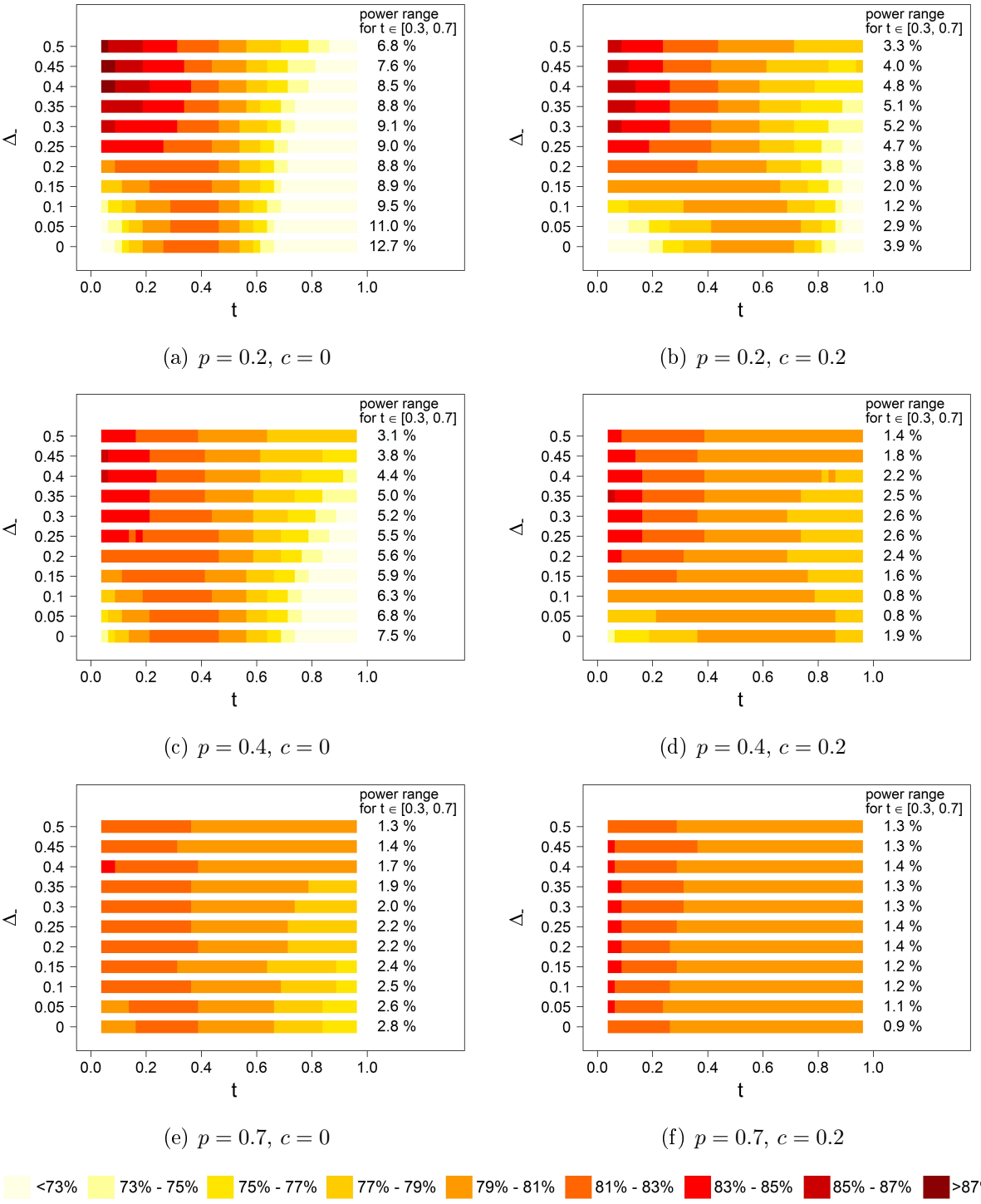
Power

Figure 4.1 shows the power, defined as the probability to reject $H_0^{(0)}$ or $H_0^{(+)}$, for different interim analysis timings. Shades of color represent the amount of power; dark red indicates a power higher than 87% and the brightest yellow characterizes a power smaller than 73%. In Figure 4.2, the individual power curves are depicted for the rejection of $H_0^{(0)}$ and $H_0^{(+)}$ as well as the probability to select G_0 or G_+ in the interim analysis as a function of the interim analysis timing. In addition, Table 4.1 shows timings yielding the minimal and maximal power within the interval $t \in [0.3, 0.7]$.
In every scenario, the power is 80% for $t = 0.5$ since the sample size is determined for this particular timing. For most of the considered simulation scenarios, the general tendency is a power advantage for early timings. Table 4.1 reveals that the maximal power is in most scenarios at $t = 0.3$ when considering timings between 0.3 and 0.7. However, the power gain is not substantially higher compared to a timing of 0.5 as the power only increases approximately between 1% and 3% for the considered scenarios. In contrast, the power can get considerably smaller than 80% for late interim analysis timings in some scenarios. This is also represented by the range of power (shown in Figure 4.1 and Table 4.1), which is especially high for scenarios with a small prevalence ($p = 0.2$) and $c = 0$ (see Figure 4.1a). In this case, for $\Delta_- > 0.25$, the power is highest for an interim analysis performed at the beginning of the study and decreases for later timings. This was to be expected since in case of high effects in both populations it is rather irrelevant in terms of power which one is selected at the beginning of the trial. However, the later the conduct of an enrichment, the smaller sample size of the subgroup is resulting in a smaller power. This is also apparent in Figure 4.2a for the scenario with $p = 0.2$ and $\Delta_- = 0.5$. The probability to select the subgroup or the total population is 0.5 and constant for varying t. Also the probability to reject $H_0^{(0)}$ is relatively constant since this selection has no impact on the sample size of the total population. However, if G_+ is selected, the power decreases with increasing t due to the decreasing sample size of the subgroup. For smaller Δ_-, the power maximum is approximately achieved between $t = 0.3$ and 0.4. In this case, selection of the subgroup is crucial to achieve a high power as selection of the total population would decrease the power of the study due to the small effect size. Obviously, a correct interim decision (selection of the subgroup) with a high probability is not possible at the beginning of the study. Instead, a sufficient amount of data has to be available for the interim analysis. However, when conducting the interim analysis much later where the probability for selecting the subgroup is higher, it is no longer possible to increase the sample size of the subgroup to a substantial extent

what is in contrast feasible when starting enrichment early. For this reason, the power decreases clearly for later timings. For $t \in [0.3, 0.7]$ the power range is given for the considered scenarios in the column right to the graphs in Figure 4.1. For the different scenarios with $p = 0.2$ and $c = 0$, this power range lies between 6.8% for $\Delta_- = 0.5$ and 12.7% for $\Delta_- = 0$.

In contrast, corresponding scenarios with $p = 0.4$ (Figure 4.1c) and $p = 0.7$ (Figure 4.1e) show smaller power ranges in the time interval $[0.3, 0.7]$: For $p = 0.4$ power ranges lie between 3.1% and 7.5% and between 1.3% and 2.8% for $p = 0.7$. Thus, one main difference to scenarios with $p = 0.2$ is that power only slightly decreases for late interim analysis timings. The relatively small power for late interim analysis timings in case $p = 0.2$ can be explained as follows: If the subgroup is selected relatively late, the sample size in the second stage is rather small and only 20% of the data from the first stage can be used. In case $p = 0.7$, a late selection of the subgroup does not lead to a considerably smaller sample size since 70% of the data from the first stage can be included in the analysis. This characteristic is also highlighted in Figure 4.2, where the deviation between the probability to reject $H_0^{(+)}$ and the probability to select G_+ is higher, the smaller the prevalence when considering late interim analysis timings.

In case $c = 0.2$ (Figure 4.1b, d and f), power ranges are smaller than for the corresponding scenarios with $c = 0$. In general, use of a higher c leads to a smaller probability to select G_+, which can be seen in Figure 4.2. Hence, the power related to $H_0^{(0)}$ has a greater contribution to the overall power which is more stable for different interim analysis timings. The smaller power range for higher c applies both for small and high prevalences. For example, for $p = 0.7$, power is relatively constant for different interim analysis timings as the maximum power range is only 1.4% in the considered scenarios. For $p = 0.2$, the power range lies between 1.2% and 5.2%. For rather high Δ_-, the power is highest for an interim analysis at the beginning of the study and decreases with increasing t as displayed in the case of $c = 0$. In contrast, the power maximum is shifted to later interim analysis timings for small Δ_-. For example, in case $\Delta_- = 0.1$, the power is maximal for $t \approx 0.575$ using $c = 0.2$, and in case using $c = 0$ the power is maximal for $t \approx 0.35$ (see Table 4.1). One reason for the later power maximum in comparison to the corresponding scenario with $c = 0$ is the higher sample size in this situation ($n = 232$ for $c = 0.2$ compared to $n = 157$ for $c = 0.2$). As a consequence, a later selection of the subgroup still leads to a sufficiently large sample size to reject $H_0^{(+)}$ with a high probability.



(a) $p = 0.2$, $c = 0$

(b) $p = 0.2$, $c = 0.2$

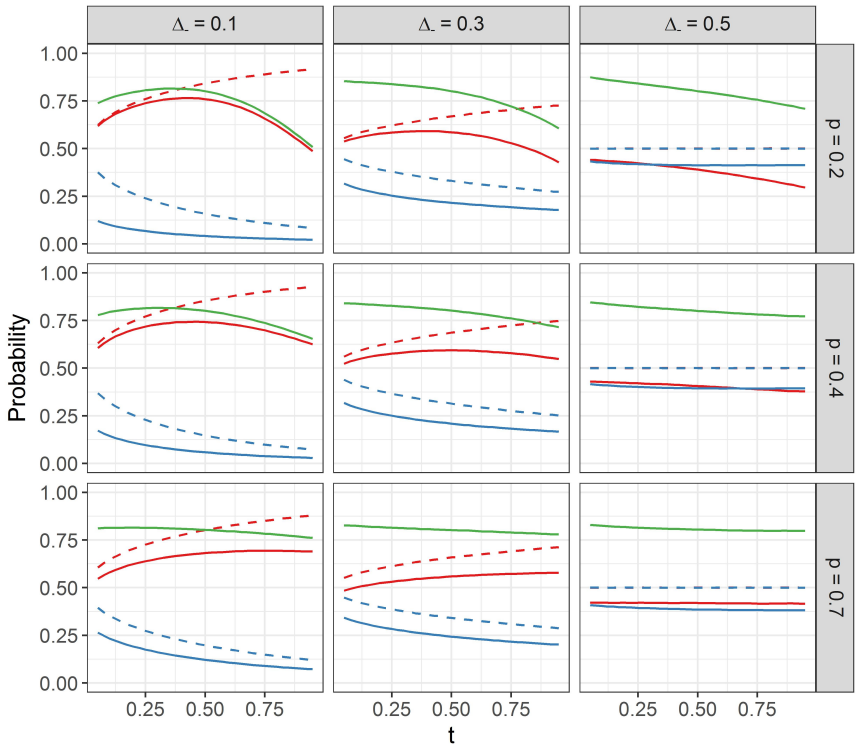
(c) $p = 0.4$, $c = 0$

(d) $p = 0.4$, $c = 0.2$

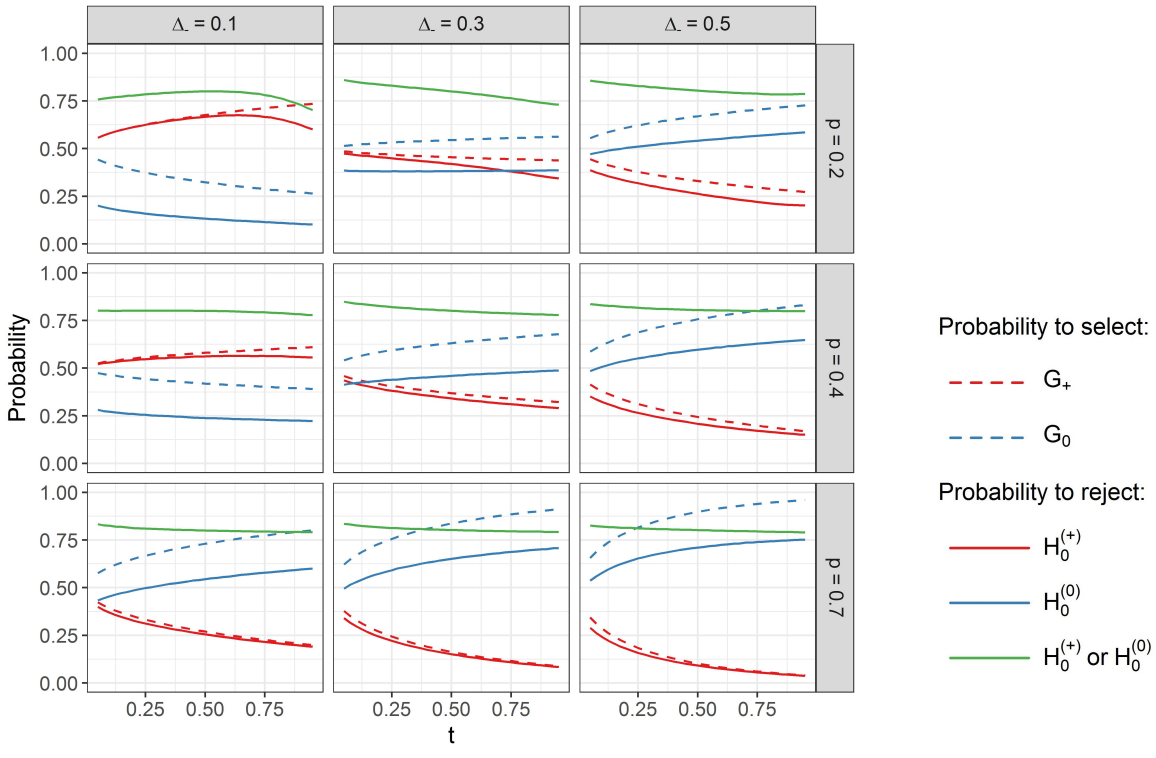
(e) $p = 0.7$, $c = 0$

(f) $p = 0.7$, $c = 0.2$

Figure 4.1: Probability to reject $H_0^{(0)}$ or $H_0^{(+)}$ using the selection rule based on estimated effect differences; $\Delta_+ = 0.5$; total sample size determined to assure a power of 80% at $t = 0.5$.



(a) $c = 0$



(b) $c = 0.2$

Figure 4.2: Probability to select G_+ or G_0, respectively, and probability to reject different hypotheses using the selection rule based on estimated effect differences; $\Delta_+ = 0.5$; total sample size determined to assure a power of 80% at $t = 0.5$.

Table 4.1: Minimal, maximal and range of power (probability to reject $H_0^{(0)}$ or $H_0^{(+)}$) for $t \in [0.3, 0.7]$ applying selection rule based on estimated effect differences; $\Delta_+ = 0.5$; total sample size (n) determined to assure a power of 80% at $t = 0.5$.

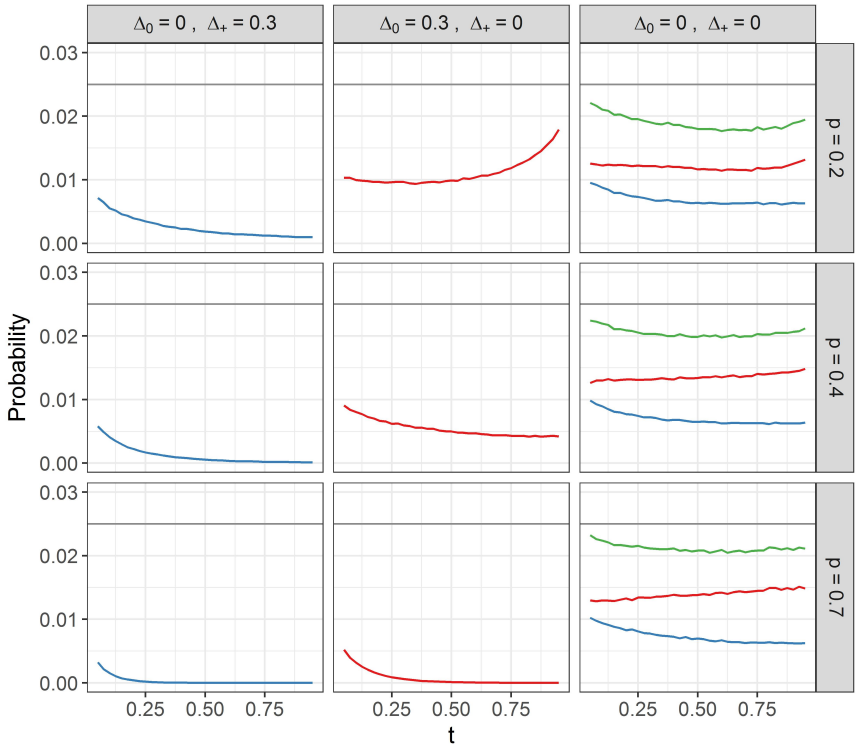
					minimal power		maximal power		
c	p	Δ_-	Δ_0	n	t	power	t	power	power range
0	0.2	0.1	0.18	157	0.7	0.72	0.35	0.815	0.095
		0.3	0.34	120	0.7	0.742	0.3	0.833	0.091
		0.5	0.50	81	0.7	0.766	0.3	0.834	0.068
	0.4	0.1	0.26	115	0.7	0.753	0.3	0.816	0.063
		0.3	0.38	98	0.7	0.771	0.3	0.822	0.052
		0.5	0.50	74	0.7	0.785	0.3	0.817	0.031
	0.7	0.1	0.38	86	0.7	0.788	0.3	0.813	0.025
		0.3	0.44	79	0.7	0.793	0.3	0.813	0.02
		0.5	0.50	70	0.7	0.799	0.3	0.812	0.013
0.2	0.2	0.1	0.18	232	0.3	0.788	0.575	0.8	0.012
		0.3	0.34	129	0.7	0.772	0.3	0.824	0.052
		0.5	0.50	77	0.7	0.79	0.3	0.823	0.033
	0.4	0.1	0.26	153	0.7	0.794	0.35	0.802	0.008
		0.3	0.38	106	0.7	0.789	0.3	0.815	0.026
		0.5	0.50	72	0.7	0.8	0.3	0.814	0.014
	0.7	0.1	0.38	101	0.7	0.795	0.3	0.808	0.012
		0.3	0.44	84	0.7	0.797	0.3	0.81	0.013
		0.5	0.50	69	0.675	0.796	0.3	0.809	0.013

Type I Error Rate

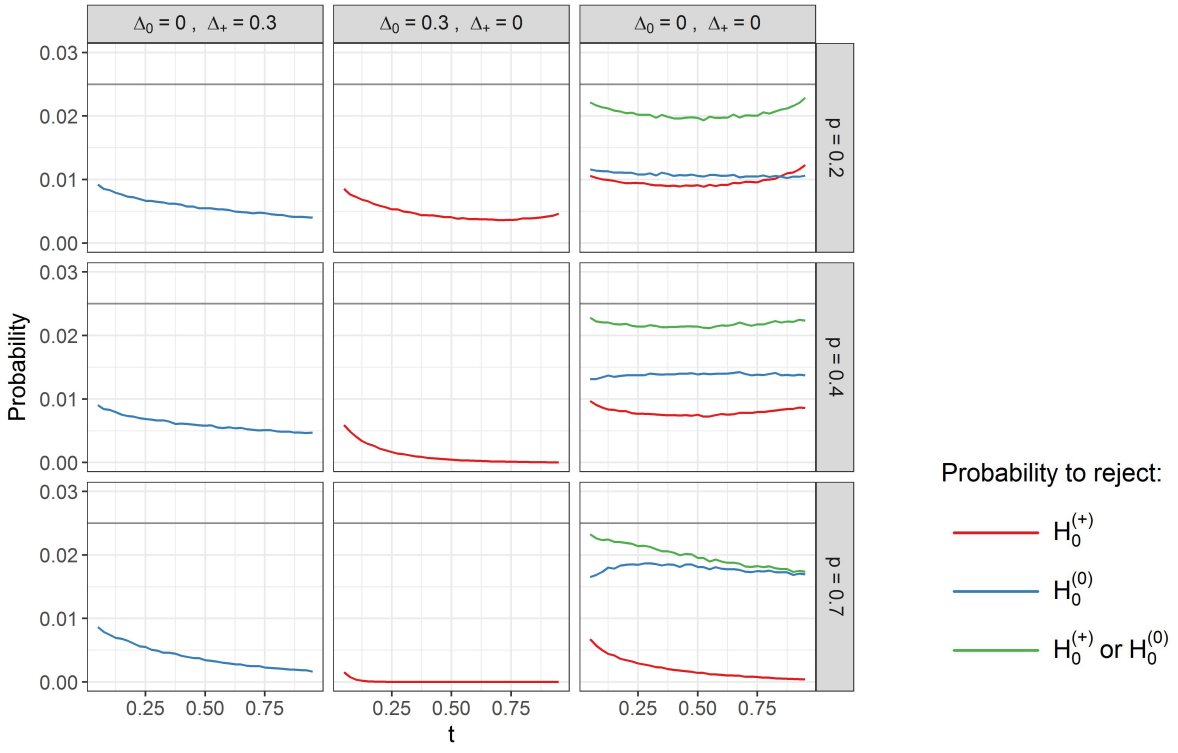
The type I error rate depending on different interim analysis timings is investigated for three different types of null distributions:

- $\Delta_0 = 0$, $\Delta_+ = 0.3$ (only $H_0^{(0)}$ is true)
- $\Delta_0 = 0.3$, $\Delta_+ = 0$ (only $H_0^{(+)}$ is true)
- $\Delta_0 = 0$, $\Delta_+ = 0$ ($H_0^{(0)}$ and $H_0^{(+)}$ are true).

The size of Δ_- depends on the prevalence of the subgroup in the specific scenario. For the first case with $\Delta_0 = 0$ and $\Delta_+ = 0.3$, Δ_- is equal to -0.075 for $p = 0.2$, $\Delta_- = -0.2$ for $p = 0.4$ and $\Delta_- = -0.7$ for $p = 0.7$. For the second null scenario with $\Delta_0 = 0.3$ and $\Delta_+ = 0$, $\Delta_- = 0.375$ for $p = 0.2$, $\Delta_- = 0.5$ for $p = 0.4$ and $\Delta_- = 1$ for $p = 0.7$. If both null hypotheses are true, $\Delta_- = 0$ for each prevalence. A sample size of $n = 200$ is used in each scenario. The global one-sided significance level was set to 0.025. To control the familywise error rate in the strong sense, the closure principle including Simes' correction to test the intersection hypothesis is applied. Figure 4.3 shows the type I error rate for the different null scenarios. In agreement with theory, the probability to reject one null hypothesis is smaller than the chosen α-level of 0.025. Obviously, the probability to reject the null hypothesis which is not true is larger than 0.025 and is not shown in the diagrams.
If only $H_0^{(0)}$ is true and $\Delta_+ = 0.3$, the probability to reject $H_0^{(0)}$ is smaller than 0.01 in every considered scenario and decreases with increasing interim analysis timing. Furthermore, the type I error rate is slightly higher for higher c since in this case the total population is selected with a higher probability. In case $H_0^{(+)}$ is true and $\Delta_0 = 0.3$, the type I error rate decreases with increasing t for most of the scenarios. However, for $c = 0$ and $p = 0.2$ the probability to reject $H_0^{(+)}$ increases for late interim analysis timings. In case both null hypotheses are true, the probability to reject either $H_0^{(+)}$ or $H_0^{(0)}$ has kind of a slightly u-shaped form for most scenarios. Since only one hypothesis is tested at the end of the trial when using the selection rule based on estimated effect differences, this probability is the sum of the probability to reject $H_0^{(+)}$ and the probability to reject $H_0^{(0)}$. While the familywise error rate lies slightly below 0.025 for very early and very late timings, an interim analysis in between leads to rather conservative decisions. Only for the scenario with $c = 0.2$ and $p = 0.7$, the familywise error rate decreases with increasing t.



(a) $c = 0$



(b) $c = 0.2$

Figure 4.3: Type I error rate for different interim analysis timings in case $H_0^{(0)}$, $H_0^{(+)}$ and both hypotheses are true using the selection rule based on estimated effect differences; $n = 200$.

4.2.3 Selection Rule Based on Absolute Effect Estimates

In this section, results are presented using the selection rule based on absolute treatment effect estimates. Threshold values are set to $c_0 = 0.1$ and $c_+ = 0.1, 0.3$.

Power

Using the selection rule based on absolute effect estimates, selection of a co-primary analysis where $H_0^{(+)}$ and $H_0^{(0)}$ is tested in the final analysis is also possible. Thus, the power defined as the probability to reject either $H_0^{(+)}$, $H_0^{(0)}$ or both is considered. Again, the sample size in each scenario is calculated to assure a power of 80% for an interim analysis timing of 0.5. Power characteristics are presented in Table 4.2 and Figure 4.4 for different scenarios. In addition, probabilities for the different interim decisions and the respective rejection probabilities are displayed in Figure 4.5 for $c_+ = 0.1$ and in Figure 4.6 for $c_+ = 0.3$.

In contrast to the selection rule based on estimated effect differences, the power is small for early interim analysis timings. Considering the power only for timings in the interval of $t \in [0.3, 0.7]$, the power is smallest for $t = 0.3$ (see Table 4.2) in every scenario. The small power for early timings using the selection rule based on absolute effect estimates results from the additional possibility to stop for futility. This is more likely at the beginning of the trial due to the lower precision of the effect estimates for small sample sizes, which is depicted by the orange dotted line in Figure 4.5 and Figure 4.6.

In Figure 4.4a, showing results for $p = 0.2$ and $c_+ = 0.1$, the power is more or less constant for $t > 0.4$ and $\Delta_- > 0.25$. Furthermore, the power ranges for $t \in [0.3, 0.7]$ amount to about 1% only, implying that the timing of the interim analysis in the considered interval has no substantial effect on the power of the study. However, the timing of the interim analysis has an impact on the selection probabilities (see Figure 4.5, $p = 0.2$, $\Delta_- = 0.5$): While the probability to select both populations increases with increasing t, the probability to select only a single population decreases with increasing interim analysis timing. In contrast, for $\Delta_- = 0.1$ (and $p = 0.2$), the overall power decreases for later timings. In this case, the probability to select only the subgroup is higher compared to scenarios with higher Δ_-. If only G_+ is selected, the power decreases with increasing interim analysis timings. While for early interim analysis timings, $H_0^{(+)}$ is rejected with a high probability if G_+ is selected, the probability to reject $H_0^{(+)}$ in case the subgroup is selected declines for timings after around $t = 0.5$. This results from a rather small sample size of G_+ if the subgroup is selected as target population towards the end of the study. This small conditional probability of rejecting $H_0^{(+)}$ given that the subgroup is selected for late timings is especially striking for small prevalence.

In case of a higher prevalence (see Figure 4.4 c and e), the decline of the power for small Δ_- is less pronounced or not present since the probability to select only the subgroup is smaller for scenarios with a higher prevalence due to the higher effect in the overall population.

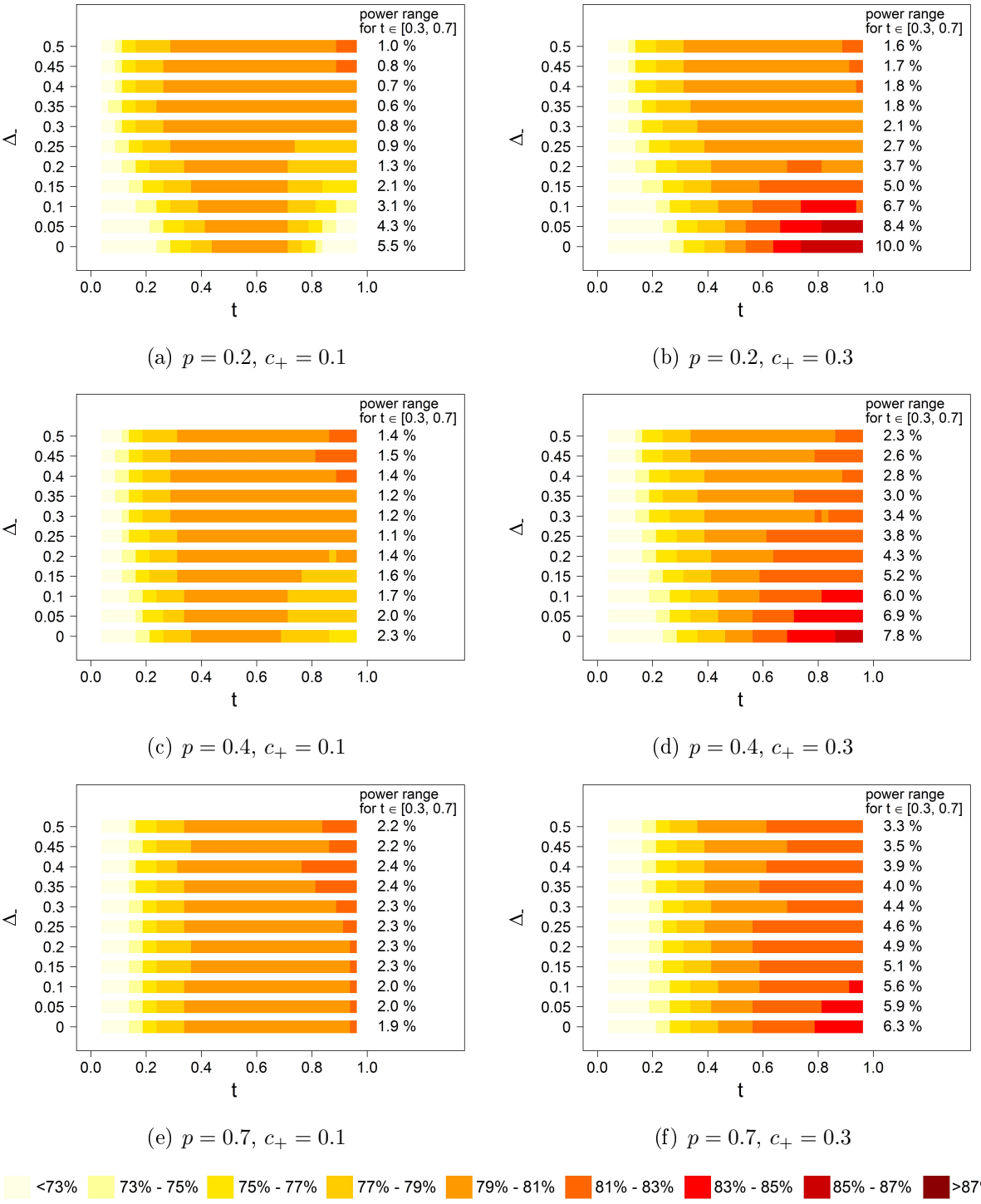
In case the threshold for selecting the subgroup is higher, the probability to select the subgroup is obviously smaller. Figure 4.4 b, d and f as well as Figure 4.6 show the power characteristics for $c_+ = 0.3$. In comparison to the scenarios with $c_+ = 0.1$, the power increases also for later timings in many scenarios. For example, the maximal power difference is 10.0% for $\Delta_- = 0$ and $p = 0.2$.

This can be explained by the fact that the power loss for later timings in case only the subgroup is selected is not present in this case. Firstly, the probability to select G_+ only is much smaller, and secondly, if G_+ is selected, the estimated effect size from the first stage must be larger than 0.3 and thus, a rejection of $H_0^{(+)}$ is more likely as compared to the selection rule using $c_+ = 0.1$. Moreover, the overall sample size needed to achieve a power of 80% at $t = 0.5$ is higher using the stricter selection rule especially for the scenario $p = 0.2$ and $\Delta_- = 0.1$ where sample size is around 25% higher (see Table 4.2).

Overall, the power is smallest for $t = 0.3$ in each considered scenario and maximal for timings between 0.5 and 0.7. However, the power range within the interval $[0.3, 0.7]$ is not very high in most situations, and the power of 80% at a timing of $t = 0.5$ is not much improved for later timings. The highest power gain is achieved for $c_+ = 0.3$, $p = 0.2$ and a small Δ_-. For example, for $\Delta_- = 0.1$ the overall power of 82.7% is reached for $t = 0.7$, which cannot be regarded as a considerably high gain though.

Table 4.2: Minimal, maximal and range of power (probability to reject $H_0^{(0)}$, $H_0^{(+)}$ or both) for $t \in [0.3, 0.7]$ applying selection rule based on absolute effect estimates; $\Delta_+ = 0.5$; $c_0 = 0.1$; total sample size (n) determined to assure a power of 80% at $t = 0.5$.

					minimal power		maximal power		
c_+	p	Δ_-	Δ_0	n	t	power	t	power	power range
0.1	0.2	0.1	0.18	262	0.3	0.772	0.575	0.804	0.032
		0.3	0.34	149	0.3	0.795	0.5	0.803	0.008
		0.5	0.50	78	0.3	0.793	0.7	0.803	0.010
	0.4	0.1	0.26	157	0.3	0.784	0.525	0.801	0.017
		0.3	0.38	116	0.3	0.791	0.525	0.803	0.012
		0.5	0.50	77	0.3	0.79	0.7	0.804	0.014
	0.7	0.1	0.38	100	0.3	0.783	0.7	0.804	0.02
		0.3	0.44	88	0.3	0.783	0.675	0.806	0.023
		0.5	0.50	75	0.3	0.783	0.675	0.806	0.022
0.3	0.2	0.1	0.18	329	0.3	0.76	0.7	0.827	0.067
		0.3	0.34	151	0.3	0.783	0.675	0.804	0.021
		0.5	0.50	77	0.3	0.787	0.7	0.803	0.016
	0.4	0.1	0.26	182	0.3	0.763	0.7	0.823	0.06
		0.3	0.38	118	0.3	0.776	0.675	0.809	0.033
		0.5	0.50	76	0.3	0.783	0.675	0.806	0.023
	0.7	0.1	0.38	106	0.3	0.762	0.7	0.818	0.056
		0.3	0.44	89	0.3	0.768	0.7	0.812	0.044
		0.5	0.50	75	0.3	0.778	0.7	0.812	0.033



(a) $p = 0.2$, $c_+ = 0.1$

(b) $p = 0.2$, $c_+ = 0.3$

(c) $p = 0.4$, $c_+ = 0.1$

(d) $p = 0.4$, $c_+ = 0.3$

(e) $p = 0.7$, $c_+ = 0.1$

(f) $p = 0.7$, $c_+ = 0.3$

Figure 4.4: Probability to reject $H_0^{(0)}$, $H_0^{(+)}$ or both using the selection rules based on absolute effect estimates; $\Delta_+ = 0.5$; $c_0 = 0.1$; total sample size determined to assure a power of 80% at $t = 0.5$.

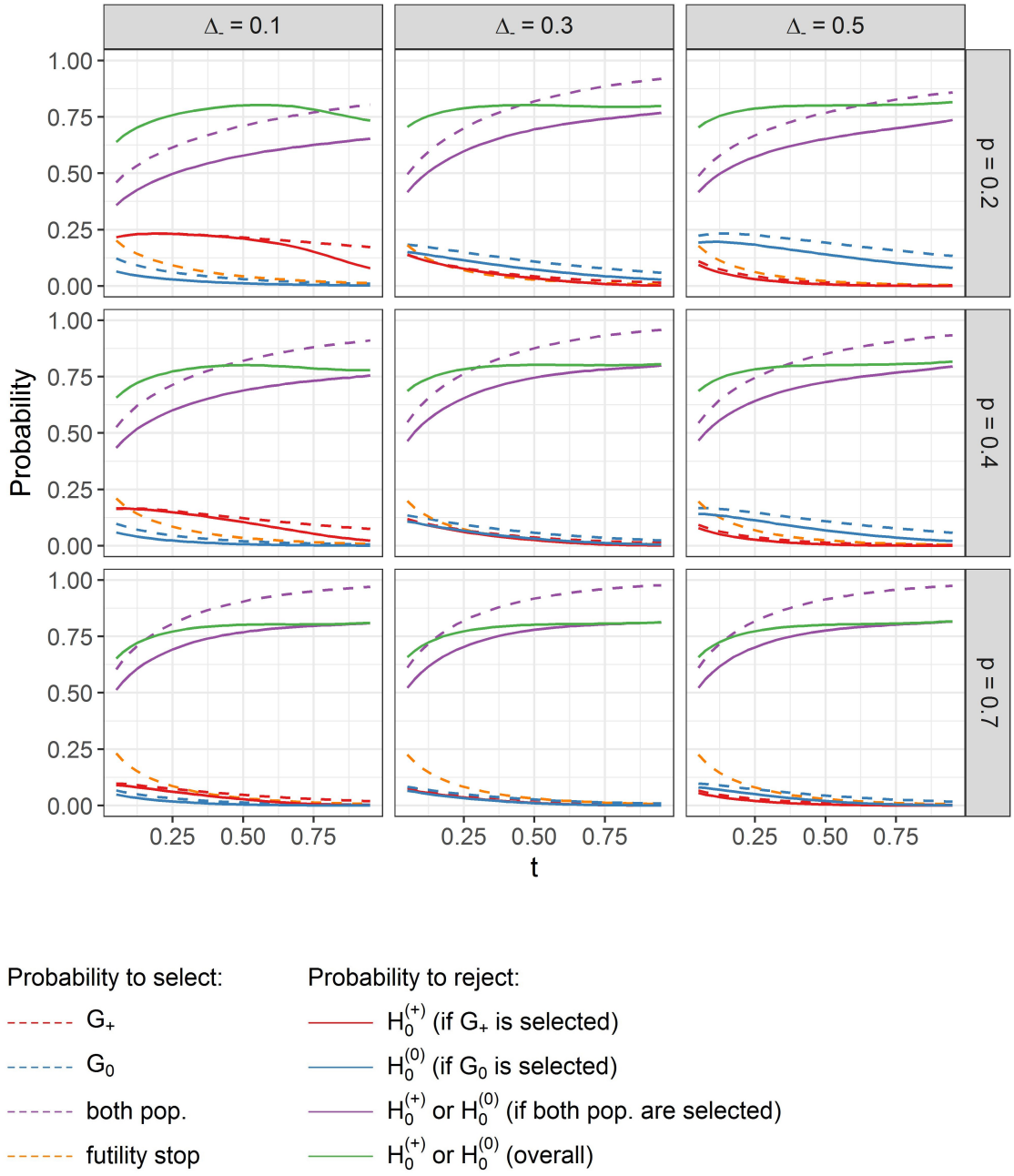


Figure 4.5: Probability for different interim decisions, and probability to reject different hypotheses using the selection rule based on absolute effect estimates; $\Delta_+ = 0.5$; $c_0 = 0.1$; $c_+ = 0.1$; total sample size determined to assure a power of 80% at $t = 0.5$.

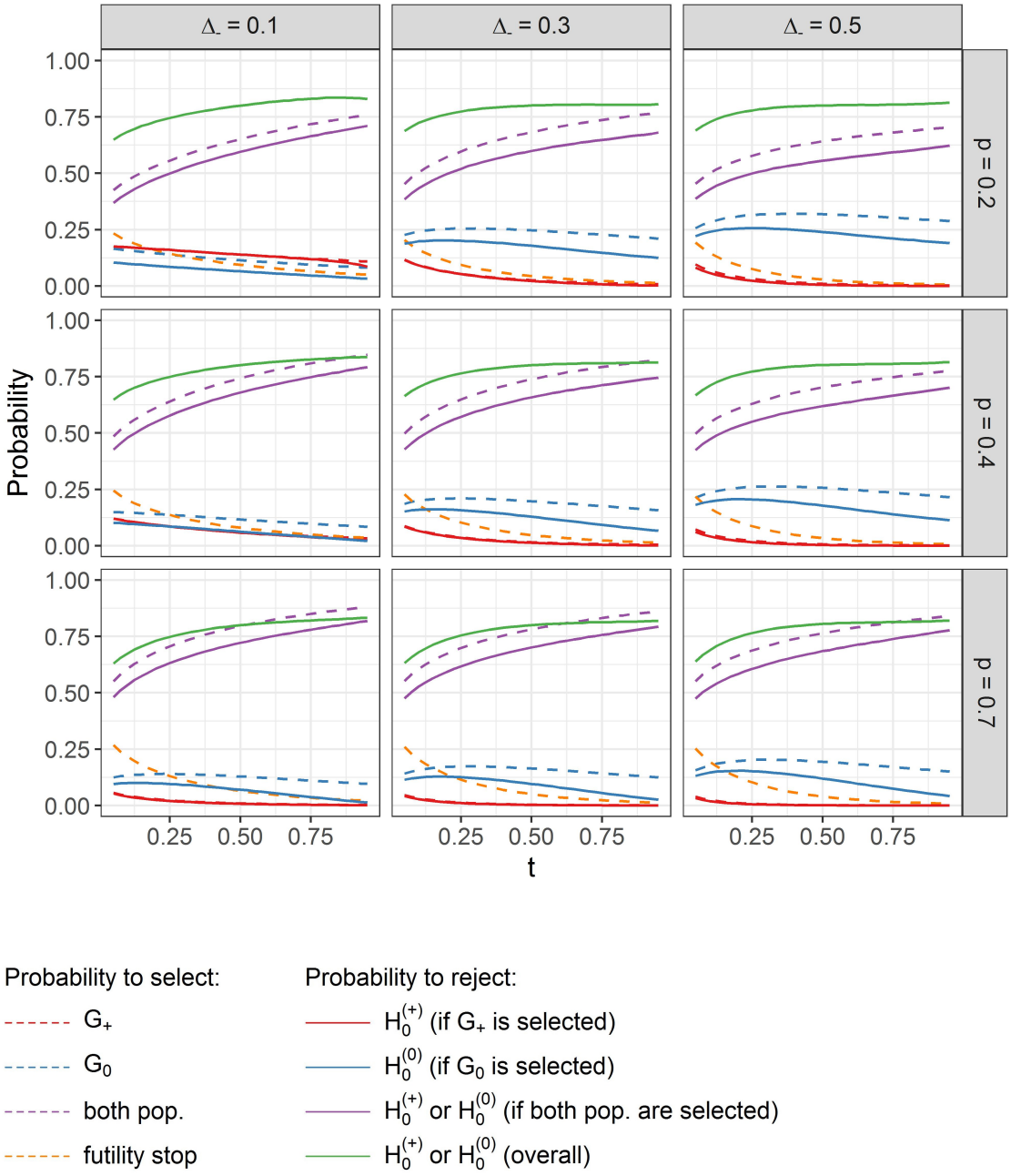
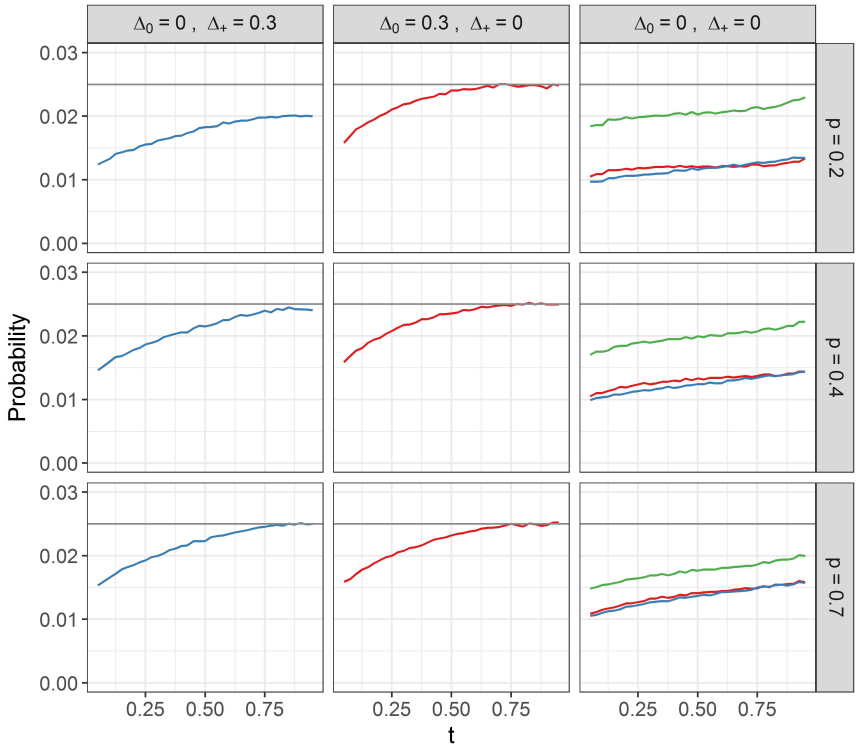


Figure 4.6: Probability for different interim decisions, and probability to reject different hypotheses using the selection rule based on absolute effect estimates; $\Delta_+ = 0.5$; $c_0 = 0.1$; $c_+ = 0.3$; total sample size determined to assure a power of 80% at $t = 0.5$.

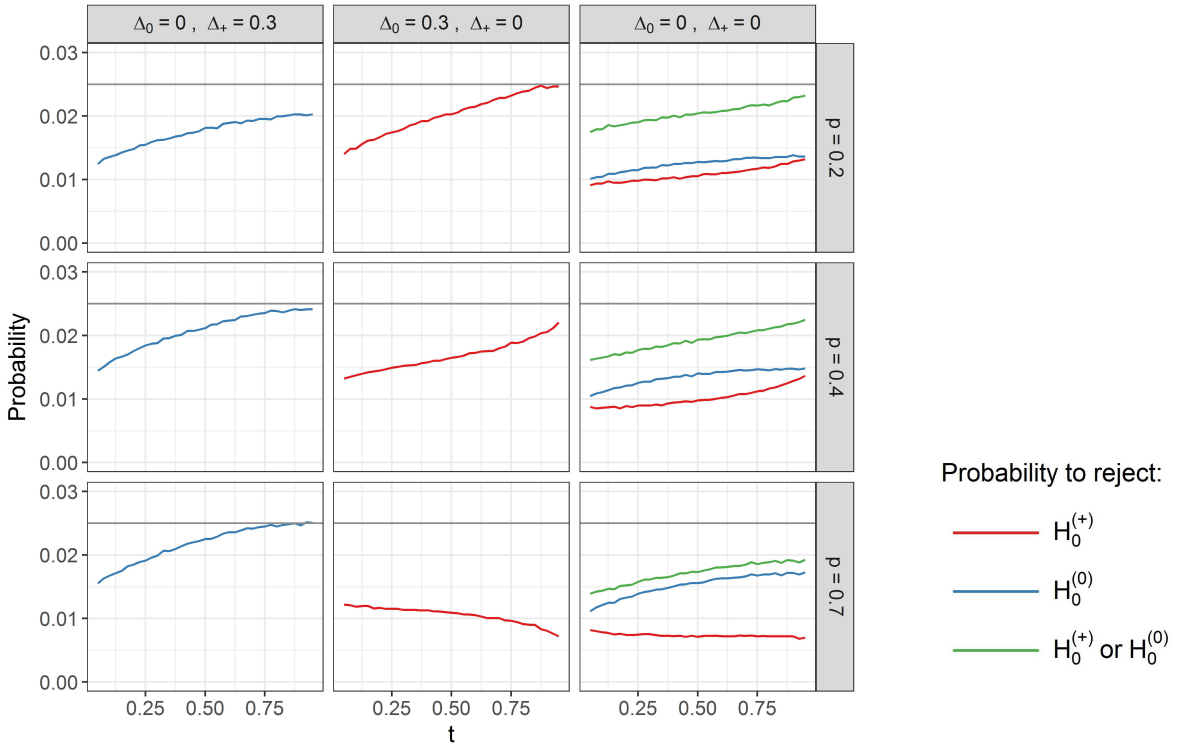
Type I Error Rate

Figure 4.7 depicts the type I error rate for the three null scenarios considered previously: only $H_0^{(0)}$ is true ($\Delta_0 = 0, \Delta_+ = 0.3$), only $H_0^{(+)}$ is true ($\Delta_0 = 0.3, \Delta_+ = 0$) and both null hypotheses are true ($\Delta_0 = 0, \Delta_+ = 0$). In contrast to the type I error rate investigations for the selection rule based on estimated effect differences, where the rejection probabilities are presented separately for each interim decision, in this section the probabilities to reject $H_0^{(+)}$ (in case $\Delta_0 = 0, \Delta_+ = 0.3$), $H_0^{(0)}$ (in case $\Delta_0 = 0.3, \Delta_+ = 0$) and at least one of the hypotheses (in case $\Delta_0 = 0, \Delta_+ = 0$) are considered irrespective of the interim decision.

In the considered scenarios, the type I error rate never exceeds the alpha level of 0.025. In most cases, early interim timings are rather conservative, and the probability to reject a true null hypothesis increases with increasing t. For late interim analysis timings, the type I error reaches almost 0.025 for most of the scenarios. One exception is the scenario $\Delta_0 = 0.3, \Delta_+ = 0, p = 0.7$ and $c_+ = 0.3$ shown in 4.7b, where the type I error rate decreases with increasing t.



(a) $c_+ = 0.1$



(b) $c_+ = 0.3$

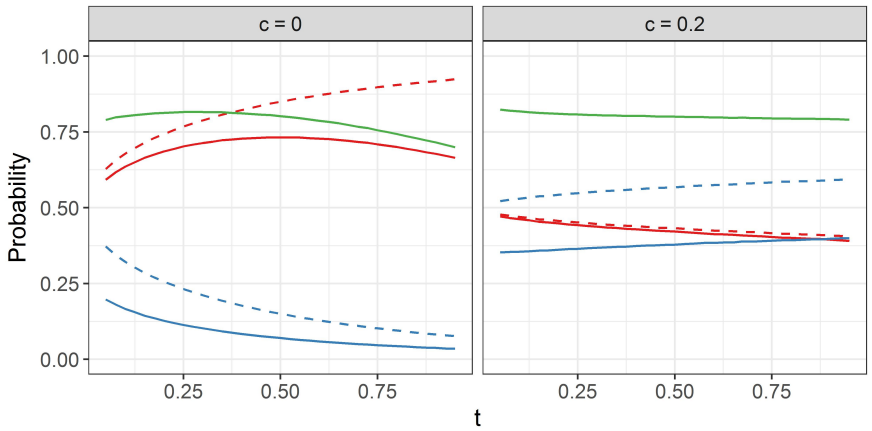
Figure 4.7: Type I error rate for different interim analysis timings in case $H_0^{(0)}$, $H_0^{(+)}$ and both hypotheses are true using the selection rule based on absolute effect estimates; $n = 200$.

4.3 Clinical Trial Example

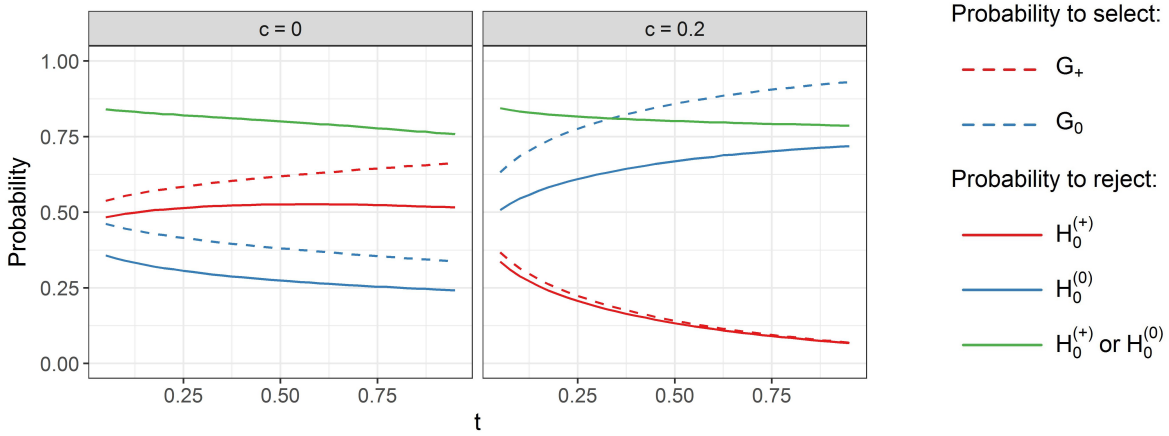
In this section, the impact of the interim analysis timing on the power of the study is investigated for a clinical trial example from the field of pneumonology. The MILLY trial is a randomized, double-blind, placebo-controlled clinical trial investigating the efficacy of lebrikizumab for patients with uncontrolled asthma (Corren et al., 2011). The primary endpoint was the relative change in prebronchodilator forced expiratory volume in 1 second from baseline to week 12. It was supposed that the treatment effect was higher for patients with a high serum periostin level as well as for patients with high type 2 helper T-cell (Th2) status. The MILLY trial was conducted as a single-stage study. Subgroup analyses were performed for prespecified subgroups defined by high and low periostin levels as well as high and low Th2 levels. For illustrative purposes, it is assumed that the trial was planned as a two-stage adaptive enrichment design. It is assumed that the observed treatment effects in the MILLY trial are the true effects and the standard deviation equals 19%, which was used for sample size calculation. This leads to the effect sizes $\Delta_{per_+} = 0.43$ and $\Delta_{per_-} = 0.08$ for subgroups defined by high and low periostin levels, and $\Delta_{Th2_+} = 0.34$ and $\Delta_{Th2_-} = 0.25$ for subgroups defined by high and low Th2 levels. A prevalence of 0.5 is assumed for each biomarker, respectively. As for the selection rules described in the previous sections, thresholds of $c = 0$ and $c = 0.2$ are considered when using the selection rule based on estimated effect differences, and $c_0 = 0.1$ together with $c_+ = 0.1$ and $c_+ = 0.3$ when using the selection rule based on absolute effect estimates. The sample size is calculated for each biomarker and the used threshold values to assure a power of 80% for an interim analysis timing at $t = 0.5$ in the specific scenario.

Figure 4.8 shows the power and selection probabilities for the selection rule based on estimated effect differences. When considering periostin level to define the subgroup (see Figure 4.8a), effect sizes differ considerably between both populations. For $c = 0$, this leads to a power maximum at around $t = 0.3$. With an overall sample size of 140, the timing of 0.3 corresponds to a sample size of 42 in the first stage and 98 in the second stage. If the subgroup with high periostin levels is selected, the number of patients to test the effect in the subgroup is 119. In case the total population is selected, all enrolled patients are included in the final analysis testing the effect in the total population. For late interim analyses, the probability is high to select the population with high periostin values but the probability to reject the null hypothesis related to this subgroup decreases as the sample size in this subgroup decreases with increasing interim analysis timing. For example, an interim analysis performed at $t = 0.7$ leads to a sample size of only 91 in case the subgroup is selected.

For $c = 0.2$, the power decreases very slightly with increasing t. In this case, the probability is higher to select the total population. The effect size of the total population is 0.255 and hence, the difference to Δ_{per_+} is with 0.175 slightly smaller than $c = 0.2$. Furthermore, the conditional rejection probabilities for both interim decisions seems to be relatively constant. The higher sample size for $c = 0.2$ and the higher effect estimate from the first stage in G_+ are probably the main reasons that power does not decline considerably if the subgroup is selected. Thus, the impact of the choice of the interim analysis timing on the overall power of the study is negligible in this scenario.
In the setting using Th2 level as biomarker, the effect sizes are similar between both populations, and therefore, it is advantageous in terms of power to conduct an early interim analysis.
Figure 4.9 shows the selection and rejection probabilities for the different populations and hypotheses when using the selection rule based on absolute effect estimates. In each scenario, both populations are selected in the with a relatively high probability, which is the correct decision in every considered setting. Due to the option to stop for futility, the overall probability to reject at least one hypothesis is smaller for early interim analysis timings as already described in Subsection 4.2.3. For $c_+ = 0.1$, the power is more or less constant after around $t = 0.4$ for both considered biomarker settings. For $c_+ = 0.3$, the power also slightly increases for later timings, which suggests that later timings are more favourable when using the stricter c_+. However, the power gain is not substantial.

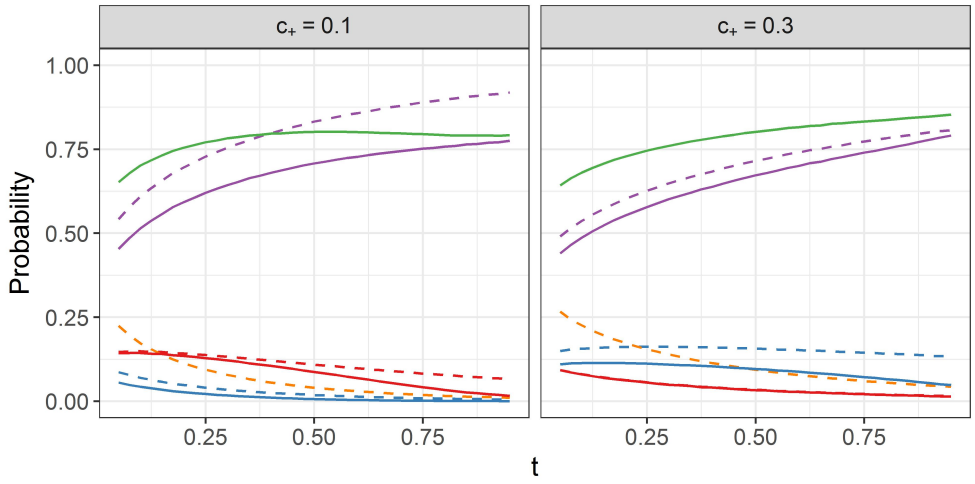


(a) Subgroups defined by periostin level, $\Delta_{per_+} = 0.43$ and $\Delta_{per_-} = 0.08$, $n = 140$ for $c = 0$, $n = 190$ for $c = 0.2$

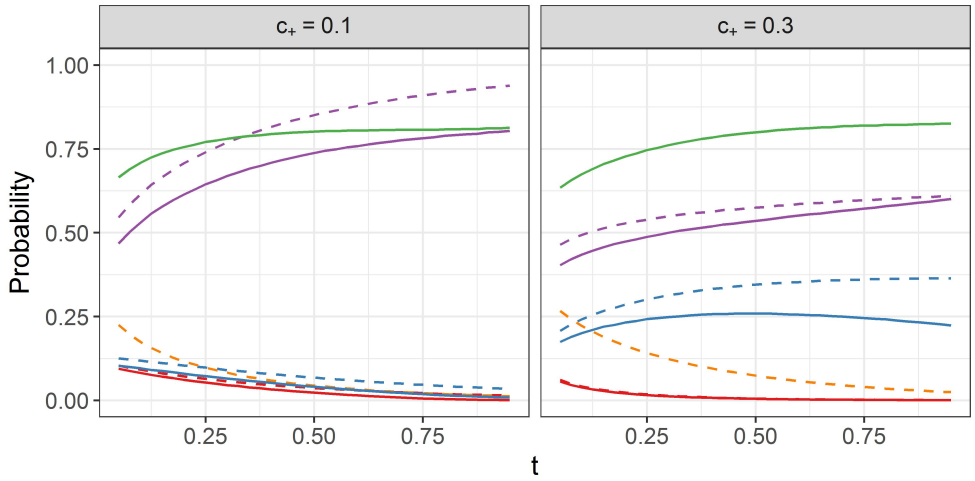


(b) Subgroups defined by Th2 level, $\Delta_{Th2_+} = 0.34$ and $\Delta_{Th2_-} = 0.25$, $n = 182$ for $c = 0$, $n = 192$ for $c = 0.2$

Figure 4.8: Probability to select G_+ and G_0, and probability to reject different hypotheses using the selection rule based on estimated effect differences for effects observed in the MILLY trial.



(a) Subgroups defined by periostin level, $\Delta_{per_+} = 0.43$ and $\Delta_{per_-} = 0.08$, $n = 179$ for $c_+ = 0.1$, $n = 214$ for $c_+ = 0.3$



(b) Subgroups defined by Th2 level, $\Delta_{Th2_+} = 0.34$ and $\Delta_{Th2_-} = 0.25$, $n = 205$ for $c_+ = 0.1$, $n = 209$ for $c_+ = 0.3$

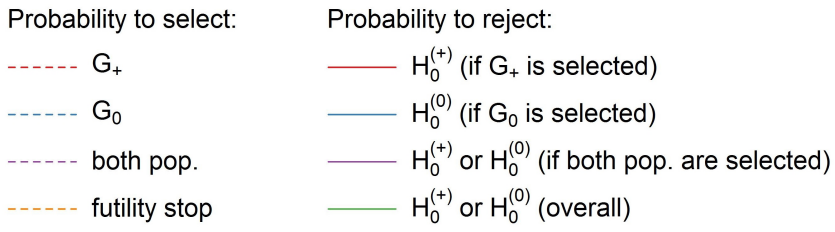


Figure 4.9: Probability for different interim decisions, and probability to reject different hypotheses using the selection rule based on absolute effect estimates for effects observed in the MILLY trial; $c_0 = 0.1$.

4.4 Chapter Summary

In this chapter, the impact of different interim analysis timings was investigated using an adaptive enrichment design with a prespecified fixed overall sample size. Results were obtained using simulation studies since an analytical derivation of the power function was not possible. In addition to the investigation of simulation studies for various scenarios, findings were demonstrated by a clinical trial example.

Results of simulation studies show that there are indeed situations in which the timing of the interim analysis has an impact on power. In particular, different selection rules lead to different power characteristics for varying interim analysis timings.

Using the selection rule based on differences of effect estimates, where a test for efficacy is conducted either in the subgroup or the total population at the end of the trial, early timings are in general more favorable in terms of power. Power ranges are especially high and thus, the timing should be selected with special care if the prevalence and the threshold value c is small. In this case, the power is small for late timings since if the subgroup is selected, only a small fraction of the first-stage data can be used, which leads to a small overall sample size of the subgroup. Therefore, late interim analyses should be avoided in case the prevalence is small and c is not too strict. Furthermore, power is maximal at extremely early timings and decreases for increasing interim analysis time if both the effect in the subgroup and the total population are high. In this case, it is irrelevant in terms of power which population is selected in an early interim analysis, but the sample size in the subgroup is smaller for later interim analysis timings. On the contrary, the probability to select the correct population is not very high for early interim analyses. When using a higher threshold value c it is less likely to select only the subgroup, which is more useful in practice to prevent such a restrictive selection. Furthermore, the power is more stable over different interim analysis timings for a higher c. Moreover, power is relatively constant for different interim analysis timings if the prevalence is high irrespective of the choice of c.

Using the selection rule based on absolute effect estimates that additionally includes the option to stop the trial for futility as well as selection of both populations, the power is small for early interim analysis timings. Whether power decreases, increases, or stays relatively constant after $t = 0.5$ depends on the chosen threshold values, the prevalence, and the treatment effects. A reason for the small power when conducting early interim analyses is the probability to stop for futility that decreases with increasing interim timing. Therefore, applying this selection rule at the end of the trial is pointless. However, simulation results illustrate that in many scenarios, power is approximately constant or increases only slightly after $t \approx 0.5$ which makes an interim analysis around

$t = 0.5$ meaningful in these cases.

Besides, the type I error rate was investigated for three different null scenarios. In each considered scenario, type I error rate or rather familywise error rate was controlled. The size of the actual type I error rate depends on the considered parameters.

In summary, using the selection rule based on estimated effect differences, in general, early interim analysis timings lead to a power advantage. In particular, scenarios with small prevalence and small c showed a considerable small power for later interim analysis timings. In these settings, late interim analysis timings should be avoided. In contrast, for large prevalence and large c, the timing of the interim analysis does not have a major impact on the power of the study. On the contrary, using the selection rule based on absolute effect estimates, early timings should be avoided since the option to stop for futility leads to a small power for interim analyses conducted at the beginning of the study.

To conclude, the power of a study can differ considerably for different interim analysis timings in many scenarios. Power characteristics depend on the selection rule, the prevalence, and effect sizes. However, in many situations, there were no large power gains for timings before or after a timing of 0.5. Nevertheless, no general rules could be established and no specific timing of the interim analysis can be recommended that uniformly fits to all scenarios.

Eventually, when choosing the interim analysis timing in the planning phase of a study, not only the power should be taken into account but also the probability to stop for futility and other selection probabilities. This is especially true if the highest power occurs for an interim analysis at the beginning or at the end of a study.

Chapter 5

Design with Sample Size Reassessment

In this chapter, an adaptive enrichment design with sample size reassessment is considered where not only the population is selected in the interim analysis but also the sample size of the second stage is recalculated. For this design, different interim analysis timings are compared regarding the distribution of the recalculated sample size. For this issue, the definition of the timing of the interim analysis has to be modified. If the overall sample size is not fixed, the definition of the interim analysis timing as a ratio of the number of patients in the first stage (n^I) and the total sample size (n) as used in the previous chapter is not appropriate. Instead, the ratio of n^I and the sample size required in a fixed design without interim analysis (n_{fix}) is considered, that is $t = n^I/n_{fix}$, where the sample size for the fixed design is given by

$$n_{fix} = \frac{2(z_{1-\alpha/4} + z_{1-\beta})^2}{\Delta_{0,A}^2}, \tag{5.0.1}$$

and $\Delta_{0,A}$ is the assumed effect in G_0 under the alternative hypothesis. Since the population to be tested is not specified in the planning phase, the sample size formula incorporates the Bonferroni adjustment using $z_{1-\alpha/4}$.

This chapter has the following structure. In Section 5.1, the general methodology of sample size reassessment is described. Simulation studies were conducted to investigate characteristics of different interim analysis timings. Results are presented in Section 5.2 for the two selection rules: based on estimated effect differences and based on absolute effect estimates. Thereafter, in Section 5.3, the sample size distribution for different interim analysis timings is investigated for the clinical trial example that was already described in the previous chapter. At the end of the chapter, results are summarized in Section 5.4.

5.1 Methods for Sample Size Reassessment

In the previous chapter, the overall sample size was fixed in advance, and only the composition of the population in the second stage (G_0 or G_+) was adapted. However, an adaptive design offers also the possibility to recalculate the sample size based on the effects observed in the interim analysis. This option is especially useful if there is a considerable uncertainty with regard to the treatment effects assumed in the planning stage. The basic idea is to adjust the sample size upwards if the observed effect in the interim analysis is smaller than expected, and downwards if the observed effect is higher than expected. In the setting of adaptive enrichment designs, the sample size for the selected population is reassessed to assure the sample size is sufficient to reject $H_0^{(+)}$ or $H_0^{(0)}$, depending on which of these two hypotheses is selected, with a certain probability. For the design incorporating sample size reassessment, test statistics for the single stages (Z_0^I, Z_+^I, Z_0^{II}, Z_+^{II}) given in formulas (3.3.1), (3.3.2), (3.3.3), (3.3.4) remain valid. However, the weights to combine the single test statistics from the first and the second stage using the inverse normal combination method have to be modified. In the fixed sample size setting, the weights were chosen so that they reflect the sample sizes in both stages. However, if sample size recalculation is applied, the ratio of the first and second-stage sample size is not known in the planning phase where weights have to be specified. Therefore, weights are chosen as $\sqrt{t}$ and $\sqrt{1-t}$ for the first- and second-stage test statistics, respectively, with $t = n^I/n_{fix}$ irrespective of whether G_0 or G_+ is selected. Thus, in each formula addressing the combination using the inverse normal method, the redefined t is used, and formula (3.3.6) changes to

$$Z_+ = \sqrt{t}Z_+^I + \sqrt{1-t}Z_+^{II}. \tag{5.1.1}$$

Even if these weights do not correspond to the actual sample size allocation in both stages, the loss in power is relatively small (Lehmacher and Wassmer, 1999).
The most commonly used method to recalculate the sample size in the interim analysis is based on conditional power arguments (Proschan and Hunsberger, 1995). The idea of this approach is to calculate the sample size in the second stage so that a specific probability to reject a given null hypothesis conditional on the observed test statistic in the first stage is assured for a certain assumed effect size. When using the described adaptive enrichment design, the second-stage sample size n_+^{II} or n_0^{II} is recalculated depending on the interim decision. The assumed effects under the alternative hypotheses used in the interim analysis for sample size recalculation are denoted by $\Delta_{0,\tilde{A}}$ for G_0 and $\Delta_{+,\tilde{A}}$ for the effect in G_+. There are different approaches for specifying the assumed effects. One possible method is to adhere to the assumed effect from the planning phase ($\Delta_{0,A}$ for

the effect in G_0 and $\Delta_{+,A}$ for the effect in G_+). Another approach is to use the effect estimate observed in the interim analysis for sample size recalculation, i.e. $\Delta_{0,\tilde{A}} = \hat{\Delta}_0$ and $\Delta_{+,\tilde{A}} = \hat{\Delta}_+$. A combination of both approaches can be achieved using the Bayesian posterior mean, that is a weighted sum of the prior density mean and the observed effect size in the interim analysis (Wassmer and Brannath, 2016, page 180). As a special case, for example, the mean value $\Delta_{+,\tilde{A}} = (\Delta_{+,A} + \hat{\Delta}_+)/2$ or $\Delta_{0,\tilde{A}} = (\Delta_{0,A} + \hat{\Delta}_0)/2$ for the effect in G_+ and G_0, respectively, can be used.
In the situation that G_+ is selected at interim, the conditional power for the rejection of $H_0^{(+)}$ is given by

$$\begin{aligned}
CP &= Pr_{\Delta_{+,\tilde{A}}}\left(Z_+ \geq z_{1-\alpha/4} \mid Z_+^I\right) \\
&= Pr_{\Delta_{+,\tilde{A}}}\left(\sqrt{t}Z_+^I + \sqrt{1-t}Z_+^{II} \geq z_{1-\alpha/4} \mid Z_+^I\right) \\
&= Pr_{\Delta_{+,\tilde{A}}}\left(Z_+^{II} \geq \frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_+^I \,\middle|\, Z_+^I\right) \\
&= Pr_{\Delta_{+,\tilde{A}}}\left(\left(\hat{\Delta}_+^{II} - \Delta_{+,\tilde{A}}\right)\sqrt{\frac{n_+^{II}}{2}} \geq \frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_+^I - \Delta_{+,\tilde{A}}\sqrt{\frac{n_+^{II}}{2}} \,\middle|\, Z_+^I\right) \\
&= 1 - \Phi\left(\frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_+^I - \Delta_{+,\tilde{A}}\sqrt{\frac{n_+^{II}}{2}}\right),
\end{aligned}$$

where the Bonferroni correction was incorporated to adjust for multiple testing. From this equation it follows the required sample size n_+^{II} for the conditional power CP and the assumed effect size $\Delta_{+,\tilde{A}}$:

$$\begin{aligned}
z_{1-CP} &= \frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_+^I - \Delta_{+,\tilde{A}}\sqrt{\frac{n_+^{II}}{2}} \\
\Rightarrow n_+^{II} &= \frac{2}{\Delta_{+,\tilde{A}}^2}\left(\frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_+^I - z_{1-CP}\right)^2.
\end{aligned}$$

The term in brackets can become negative if $Z_+^I > z_{1-\alpha/4}/\sqrt{t} - z_{1-CP}\sqrt{(1-t)/t}$. Since this corresponds to extremely large test statistics, negative values are set to zero. Furthermore, $\Delta_{+,\tilde{A}}$ can become negative if the observed interim effect is included in $\Delta_{+,\tilde{A}}$. In this case, $\Delta_{+,\tilde{A}}$ is set to 0. Moreover, the recalculated sample size can get impracticably high if small effects are observed in the interim analysis. Therefore, it may be advisable to restrict the sample size in the second stage to n_{max}^{II}. In particular, the sample size is n_{max}^{II} if the assumed effect in the interim analysis is negative. However, this situation does not occur when using the selection rule based on absolute effect estimates. Here, the study is stopped for futility if negative effects are observed, and sample size recalcu-

lation is not conducted at all. In order to apply the inverse normal combination method, where data in stage II is necessary, also a minimal second-stage sample size n_{min}^{II} should be chosen. With the restrictions described above, the sample size formula is modified to

$$n_+^{II} = \max\left\{n_{min}^{II}, \min\left\{n_{max}^{II}, \frac{2}{(\Delta_{+,\tilde{A}})_+^2}\left(\frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_+^I - z_{1-CP}\right)_+^2\right\}\right\} \quad (5.1.2)$$

where $(x)_+ := \max(x, 0)$.

The conditional power in case the total population is selected, can be derived in the same way, and is given by

$$CP = 1 - \Phi\left(\frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_0^I - \Delta_{0,\tilde{A}}\sqrt{\frac{n_0^{II}}{2}}\right).$$

With the same restrictions as described for recalculating the sample size of the subgroup, this leads to the sample size formula for the second stage in case G_0 is selected:

$$n_0^{II} = \max\left\{n_{min}^{II}, \min\left\{n_{max}^{II}, \frac{2}{(\Delta_{0,\tilde{A}})_+^2}\left(\frac{z_{1-\alpha/4}}{\sqrt{1-t}} - \sqrt{\frac{t}{1-t}}Z_0^I - z_{1-CP}\right)_+^2\right\}\right\}. \quad (5.1.3)$$

It should be noted that the limits of sample size n_{min}^{II} and n_{max}^{II} do not necessarily have to agree in the calculation of n_+^{II} and n_0^{II}. However, for practical reasons, limits are assumed to be equal for each interim decision.

In summary, the following approach is considered in this thesis. When using the selection rule based on estimated effect differences, the sample size for the second stage is calculated using formula (5.1.2) if G_+ is selected as target population and formula (5.1.3) if G_0 is selected. When using the selection rule based on absolute effect estimates, the same procedure is applied if only one population is selected. If the co-primary analysis is chosen, which means patients from G_0 are enrolled in the second stage and both hypotheses are tested in the final analysis, n_0^{II} and n_+^{II} is calculated with the described formulas and finally, the maximum sample size of n_0^{II} and n_+^{II}/p is used overall in the second stage to assure the conditional power for both hypotheses. If the observed effects suggest a futility stop, $n^{II} = 0$.

Moreover, it should be noted that Bonferroni adjustment is applied when using the design including sample size recalculation for the sake of simplicity. Sample size calculation when applying Simes' method for testing the intersection hypothesis within the closed testing procedure is more challenging since it is not clear in the interim analysis which hypothesis is tested at which local significance level.

5.2 Simulation Study

In this section, results of simulation studies are presented for an adaptive enrichment design with sample size reassessment in the interim analysis using conditional power arguments as described in the previous section. Here, the impact of the interim analysis timing on overall sample size distribution is investigated.

5.2.1 Simulation Setup

In the conducted simulation studies, sample sizes are calculated using formulas (5.1.2) and (5.1.3) with a minimum second-stage sample size of $n_{min}^{II} = 10$ (if the trial is not stopped for futility) and a maximum sample size in the second stage of $n_{max}^{II} = 2n_{fix} - n^I$. Hence, the total sample size is never higher than twice the sample size calculated in a fixed design without interim analysis. Moreover, two different methods are considered for choosing the assumed effect size in the interim analysis that is used for sample size recalculation. The first one is to adhere to the assumed effect size from the planning phase, i.e. $\Delta_{+,\tilde{A}} = \Delta_{+,A}$ and $\Delta_{0,\tilde{A}} = \Delta_{0,A}$, while the second considered approach is to use the mean of the effect size from the planning phase and the observed effect in the interim analysis, that is $\Delta_{+,\tilde{A}} = (\Delta_{+,A} + \hat{\Delta}_+)/2$ and $\Delta_{0,\tilde{A}} = (\Delta_{0,A} + \hat{\Delta}_0)/2$. Furthermore, in the simulations, the assumed effect size used in the planning phase is equal to the true effect size, i.e. $\Delta_{+,A} = \Delta_+$ and $\Delta_{0,A} = \Delta_0$. The Bonferroni method is used to handle the underlying multiple testing problem.

Characteristics of the sample size distribution are investigated for interim analysis timings of $t = 0.2, 0.35, 0.5, 0.65, 0.8$. In order to make the distributions comparable for different interim analysis timings, the conditional power is adapted in each scenario to reach an overall power of 80%. The corresponding conditional power is determined using simulations.

Furthermore, results for different effect sizes, prevalences and selection rules are presented. The effect in G_+ is equal to $\Delta_+ = 0.5$ in each scenario, the effect in G_- varies with $\Delta_- = 0, 0.25, 0.5$, and different prevalences are investigated using $p = 0.2, 0.7$. As in the previous chapter, different parameters for the selection rules are considered. For the selection rule based on estimated effect differences, results for $c = 0$ and $c = 0.2$ are presented. When using the selection rule based on absolute effect estimates, $c_0 = 0$ in every scenario and $c_+ = 0.1, 0.3$. For each scenario, $1,000,000$ study results are simulated.

In the following, results are presented for the selection rule based on estimated effect differences and the selection rule based on absolute effect estimates. For both selection rules, firstly, results using the effect size from the planning phase for sample size reassessment and secondly, using the mean of the assumed effect from the planning phase and

the observed effect in the interim analysis are shown.

5.2.2 Selection Rule Based on Estimated Effect Differences

In this section, characteristics of the sample size distribution are investigated for the selection rule based on estimated effect differences. Distribution of sample size is depicted by stacked histograms for both interim decisions, where the total sample size is shown in red color if G_+ is selected, and in blue color if G_0 is selected. It should be noted that histograms are used to demonstrate the distribution of the sample size and frequencies cannot be compared directly for different interim analysis timings since the y-axis scaling is not consistent for different timings. Additionally, the distribution of the sample size irrespective of the interim decision is summarized by boxplots showing median, first and third quartile and mean (marked by a dot). Further characteristics are given in the tables to the right of the figures. Here, the average sample size ± standard deviation (ASS ± SD), the probability for selection of G_+, the adjusted conditional power to reach an overall power of 80%, and the probability for $n > n_{fix}$ is given.

5.2.2.1 Using the Effect Size from the Planning Phase for Sample Size Reassessment

In the first part of this subsection, characteristics of the sample size distribution are investigated using the assumed effect size from the planning phase to recalculate the sample size in the interim analysis. Results for $c = 0$ are presented in Figure 5.1 and Figure 5.2 for $p = 0.2$ and $p = 0.7$, respectively.

If the overall treatment effect Δ_0 is small, which is the case in the considered settings if p and Δ_- are small, the originally planned sample size n_{fix} is high (see Figure 5.1a with $p = 0.2$, $\Delta_- = 0$, $\Delta_0 = 0.1$, $n_{fix} = 1902$). In this case, an early interim analysis clearly leads to the smallest average sample size. The reason is that an early interim analysis of $t = 0.2$ results in a relatively high sample size in the first stage ($n^I = 380$) due to the large n_{fix}. The high sample size and the large difference between Δ_0 and Δ_+ lead to a high probability (99.7%) to select G_+. Due to the high effect in the subgroup, only a small sample size is required for the second stage. This is also reflected by the conditional power of 0 for $t \geq 0.35$ in the described scenario. In this case, the sample size in the first stage is sufficient to detect an effect with a probability of at least 80%. Therefore, only 10 patients per group are included in the second stage, and the power can be higher than 80%. In this scenario, the power is 85.8% for $t = 0.35$, 97.7% for $t = 0.5$, 99.8% for $t = 0.65$, and $> 99.9\%$ for $t = 0.8$. It should be noted that a conditional power of 0 for sample size recalculation means that the minimal sample size of n^{II}_{min} is used and

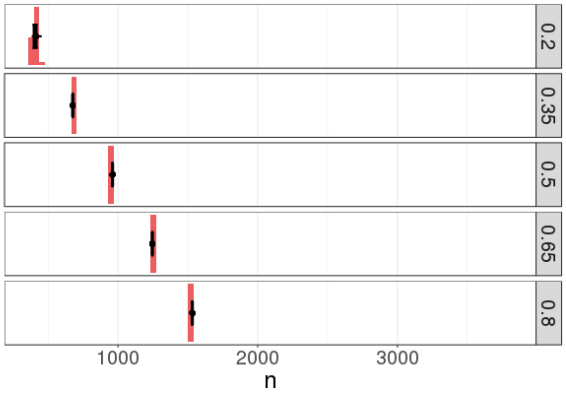
that this sample size is sufficient to reach an overall power of at least 80%. However, the actual conditional power might be larger.

For scenarios with higher Δ_-, median and mean values of the sample size are more similar when comparing different interim analysis timings. For $\Delta_- = 0.25$ (Figure 5.1b), the average sample size is still increasing for increasing t and the standard deviation decreases. However, the probability for $n > n_{fix}$ is smallest for $t = 0.5$. A further advantage of using $t \approx 0.5$ as compared to earlier interim analysis timings is the higher probability to select G_+, which is the correct decision in this scenario. For later interim analysis timings, the sample size in the second stage is the minimum sample size with a probability near 1, which shows that late interim analyses are meaningless.

If the effects in G_- and G_+ are equal (Figure 5.1c), the average sample size is still slightly increasing for increasing t but median sample sizes as well as the probability for $n > n_{fix}$ are very similar for different timings of the interim analysis. In this case, the main difference lies in the distribution of the sample size. While the distribution of the sample size for early interim analysis timings is rather symmetric around the mean, for later interim analysis timings, it becomes more likely that the second stage is conducted with the minimal sample size. This means that if the interim analysis is performed relatively late, it is not possible to get a small overall sample size due to the large first-stage sample size. Furthermore, in this scenario, it is unlikely that a very large sample size is required for the second stage of the trial. When doing the interim analysis early, both options (small and large overall sample size) are possible.

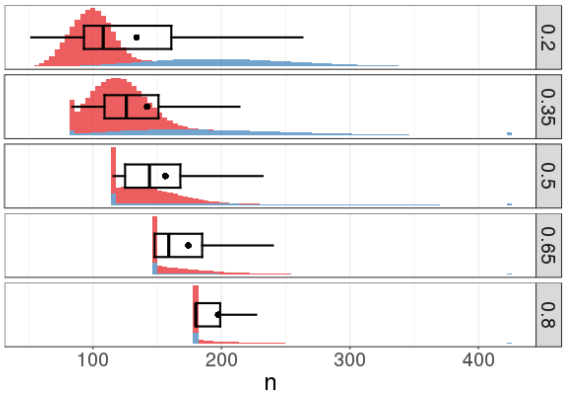
If the prevalence of the subgroup is higher (Figure 5.2), Δ_0 is higher (except for $\Delta_- = 0.5$) and thereby, n_{fix} is smaller. For $\Delta_- = 0$, $p = 0.7$ (Figure 5.2a), the average sample size is more similar across different timings than in the scenario with $p = 0.2$. Although the smallest average sample size occurs for $t = 0.2$, an early interim analysis has some disadvantages: the probability to select G_0, is rather high with 18%, and if G_0 is selected the recalculated sample size is also rather high due to the smaller effect size.

For higher effects, the smallest average sample size occurs for later interim analysis timings compared to the respective scenarios with $p = 0.2$. While for $p = 0.2$, the smallest average sample size is observed for a timing of $t = 0.2$ irrespective of Δ_-, the timing with the smallest average sample size is present for $t = 0.2, 0.35, 0.5$ for $\Delta_- = 0, 0.25, 0.5$ in case $p = 0.7$. This means, if the prevalence is higher, the smallest average sample size is reached for later interim analysis timings. This is reasonable since more data from the first stage can be used if the subgroup is selected in case the prevalence is higher. Thus, a late interim analysis does not lead to a large loss of information as compared to scenarios with a smaller prevalence.



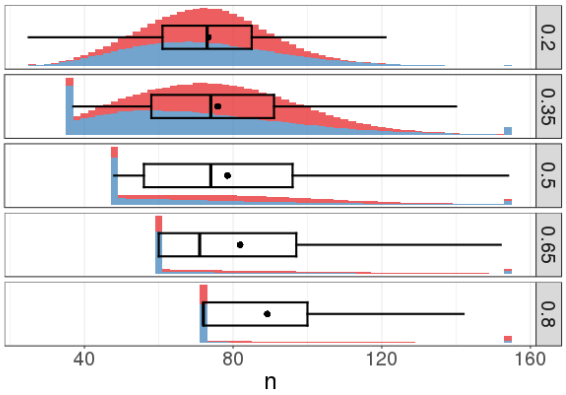
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	412 ± 84	0.997	0.8	0.001
0.35	676 ± 0	> 0.999	0	0
0.5	961 ± 0	> 0.999	0	0
0.65	1246 ± 0	> 0.999	0	0
0.8	1532 ± 0	> 0.999	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c = 0, n_{fix} = 1902$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	134 ± 63	0.678	0.81	0.142
0.35	142 ± 56	0.728	0.8	0.104
0.5	157 ± 49	0.767	0.79	0.087
0.65	174 ± 42	0.797	0.75	0.112
0.8	197 ± 37	0.822	0.65	0.18

(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c = 0, n_{fix} = 212$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	75 ± 19	0.501	0.81	0.43
0.35	76 ± 24	0.499	0.8	0.447
0.5	79 ± 26	0.5	0.79	0.462
0.65	82 ± 26	0.5	0.75	0.429
0.8	89 ± 26	0.501	0.69	0.414

(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0, n_{fix} = 77$

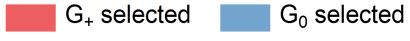
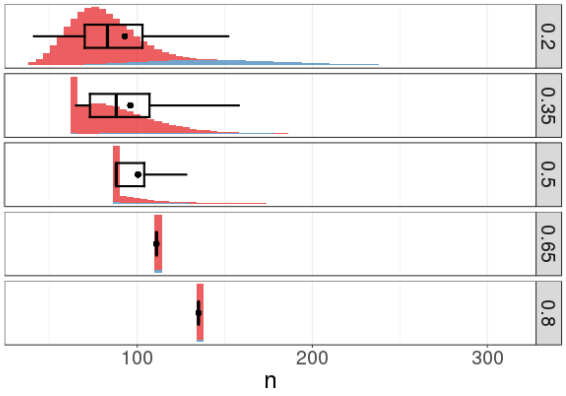
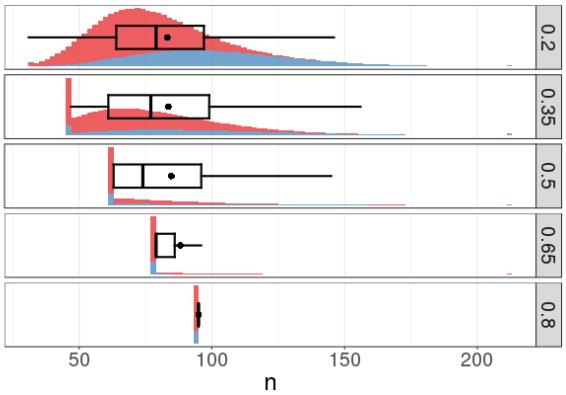


Figure 5.1: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0$; $p = 0.2$. Sample size recalculation is based on the effect size from the planning phase.



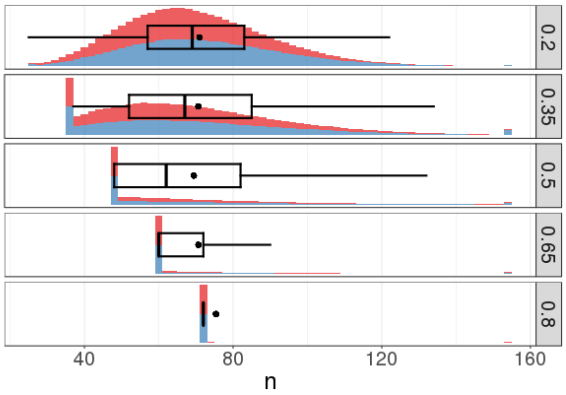
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	95 ± 38	0.817	0.81	0.08
0.35	96 ± 33	0.885	0.8	0.052
0.5	101 ± 24	0.924	0.72	0.033
0.65	111 ± 1	0.948	0.01	0
0.8	135 ± 0	0.965	0	0

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c = 0, n_{fix} = 156$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	85 ± 27	0.645	0.81	0.188
0.35	84 ± 30	0.69	0.8	0.188
0.5	85 ± 28	0.722	0.77	0.174
0.65	88 ± 21	0.749	0.63	0.117
0.8	95 ± 0	0.772	0	0

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c = 0, n_{fix} = 106$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	73 ± 20	0.5	0.81	0.37
0.35	71 ± 25	0.5	0.8	0.34
0.5	69 ± 25	0.501	0.77	0.292
0.65	71 ± 20	0.501	0.67	0.207
0.8	75 ± 13	0.499	0.31	0.101

(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0, n_{fix} = 77$

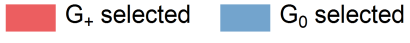
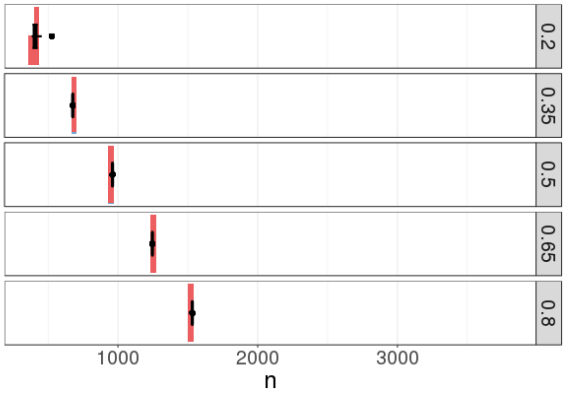


Figure 5.2: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0$; $p = 0.7$. Sample size recalculation is based on the effect size from the planning phase.

For $c = 0.2$ (see Figure 5.3 for $p = 0.2$ and Figure 5.4 for $p = 0.7$), properties of the sample size distribution when comparing different interim analysis timings are similar as compared to the respective scenarios with $c = 0$. However, when using a higher c, the probability to select G_0 in the interim analysis is higher. While the sample size distributions for G_0 and G_+ are very similar when using a higher c compared to scenarios with a smaller c, the sample size distribution for G_0 has a greater contribution to the overall sample size. This behavior leads to an altered overall sample size distribution especially for scenarios where distributions are very different for both interim decisions, namely scenarios with $\Delta_- < 0.5$, i.e. $\Delta_- < \Delta_+$. In these cases, the average sample size is higher for the situation that the total population is selected as compared to the situation that the subgroup is selected due to the smaller effect size Δ_-. For early interim analyses, this leads to a larger overall average sample size in scenarios with $\Delta_- < 0.5$ when using a higher c. In addition, the probability $Pr(n > n_{fix})$ is increased. For late interim analysis timings, the sample size increase for a higher c is not as pronounced since the second-stage sample size is in many cases the minimal sample size of 10. Thus, the smallest average sample size is achieved for later interim analyses when using a higher c. For example, in the scenario $p = 0.7, \Delta_- = 0$ (see Figure 5.4a), the average sample size is 121 for $t = 0.2$ and therefore, considerably larger than 95, which was the value observed in the respective scenario with $c = 0$. Simultaneously, the increase of the average sample size for $t = 0.8$ from 135 to 140 is much smaller.

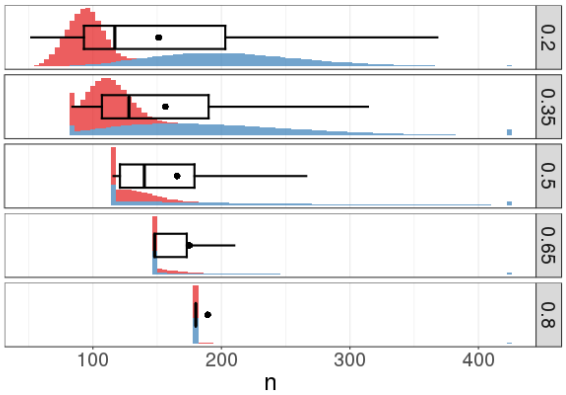
For $\Delta_- = 0.5$ (see Figure 5.3c and Figure 5.4c), sample size distributions for G_0 and G_+ are similar, and a higher value of c, which leads to a larger contribution of the sample size distribution of G_0 to the overall sample size, does not lead to considerable differences between the overall sample size distributions for different c. For example, in the scenario $p = 0.7, \Delta_- = 0.5$ (see Figure 5.4c), the average sample size for $t = 0.2$ is 71 using $c = 0.2$, and 73 for $c = 0$.

To summarize, especially for the case that the effect in the subgroup is larger than in the total population, a larger c tends to shift the minimum average sample size to later timings.



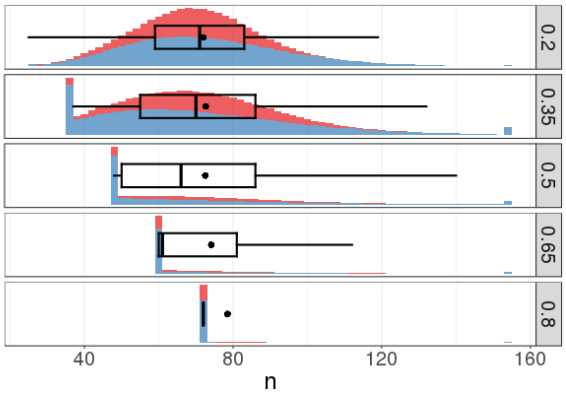
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	527 ± 428	0.916	0.8	0.035
0.35	676 ± 0	0.966	0	0
0.5	961 ± 0	0.985	0	0
0.65	1246 ± 0	0.993	0	0
0.8	1532 ± 0	0.997	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c = 0.2, n_{fix} = 1902$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	151 ± 72	0.5	0.81	0.22
0.35	156 ± 70	0.499	0.8	0.192
0.5	166 ± 66	0.5	0.79	0.173
0.65	175 ± 55	0.5	0.72	0.134
0.8	189 ± 36	0.5	0.4	0.073

(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c = 0.2, n_{fix} = 212$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	72 ± 19	0.39	0.8	0.351
0.35	73 ± 24	0.357	0.8	0.375
0.5	72 ± 25	0.331	0.78	0.349
0.65	74 ± 22	0.308	0.72	0.283
0.8	79 ± 17	0.289	0.52	0.192

(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0.2, n_{fix} = 77$

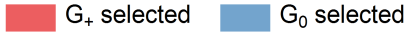
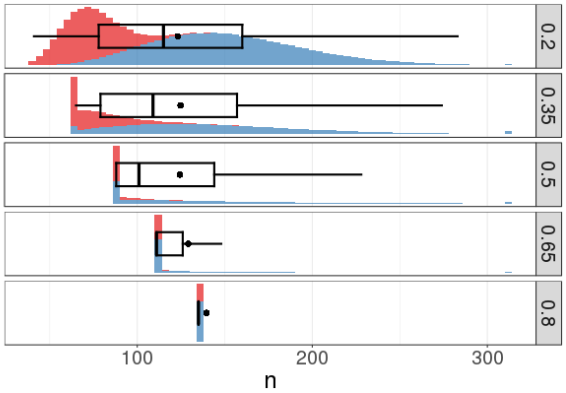
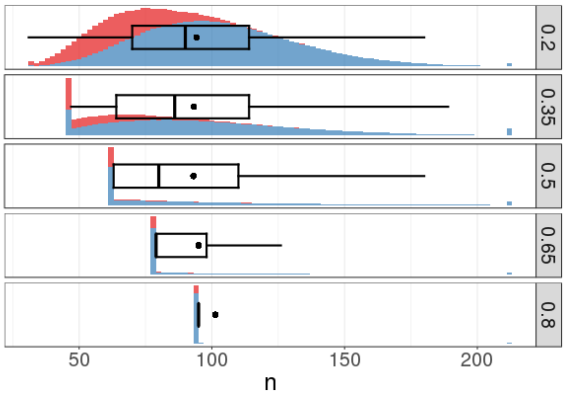


Figure 5.3: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0.2$; $p = 0.2$. Sample size recalculation is based on the effect size from the planning phase.



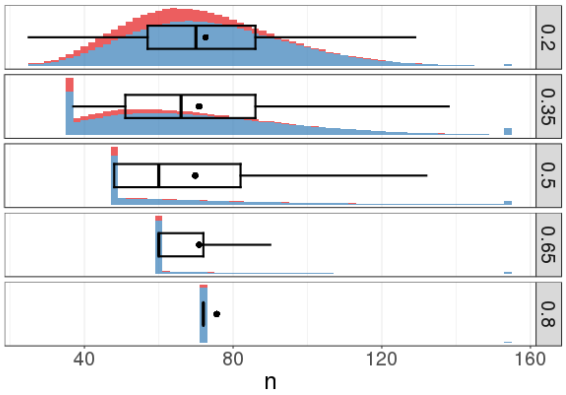
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	121 ± 52	0.382	0.8	0.252
0.35	125 ± 56	0.345	0.8	0.253
0.5	124 ± 51	0.317	0.75	0.203
0.65	129 ± 40	0.293	0.62	0.142
0.8	140 ± 21	0.273	0.17	0.056

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c = 0.2, n_{fix} = 156$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	94 ± 32	0.267	0.81	0.321
0.35	93 ± 37	0.205	0.8	0.307
0.5	93 ± 36	0.163	0.77	0.27
0.65	95 ± 30	0.131	0.67	0.201
0.8	101 ± 20	0.107	0.35	0.112

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c = 0.2, n_{fix} = 106$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	71 ± 21	0.198	0.8	0.346
0.35	71 ± 26	0.131	0.8	0.344
0.5	70 ± 26	0.09	0.77	0.296
0.65	71 ± 21	0.063	0.66	0.207
0.8	76 ± 13	0.045	0.3	0.106

(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0.2, n_{fix} = 77$

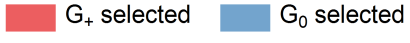


Figure 5.4: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0.2$; $p = 0.7$. Sample size recalculation is based on the effect size from the planning phase.

5.2.2.2 Using the Mean of the Effect Size from the Planning Phase and the Interim Effect Estimate for Sample Size Reassessment

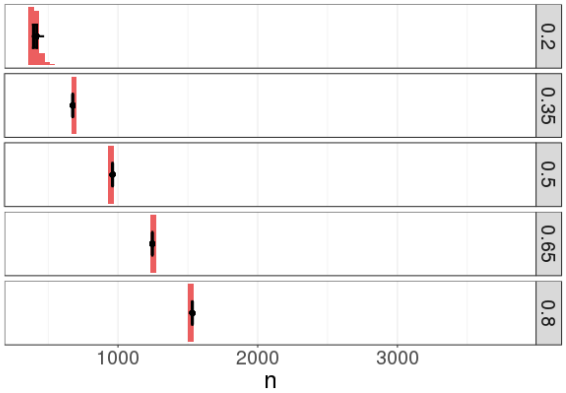
In this section, results are considered when the mean of the assumed effect from the planning phase and the observed effect in the interim analysis is used for sample size recalculation. Results are presented in Figure 5.5 and Figure 5.6 for $c = 0$ and in the Appendix in Figure B.1 and Figure B.2 for $c = 0.2$, for a prevalence of $p = 0.2$ and $p = 0.7$, respectively. Comparing different scenarios regarding the timing with the smallest average sample size show similar behavior as compared to the strategy where the recalculation is only based on the treatment effect assumed in the planning stage. For larger prevalence, larger Δ_-, or larger c, the average sample size occurs for later interim analysis timings. However, in comparison to the strategy where only the assumed effect from the planning stage is used for recalculation, the average sample size and the standard deviation are higher when incorporating the observed effect of the interim analysis to recalculate the sample size. This is the case in particular for early interim analysis timings, where the recalculated sample size is often the maximum possible sample size if G_0 is selected. For example, for $p = 0.2, \Delta_- = 0.25$ (see Figure 5.5b), and a timing of 0.2, the average sample size and standard deviation is 164 ± 117 if the observed effect from the interim analysis is incorporated, and 134 ± 63 if only the assumed effect from the planning phase is used for sample size recalculation. Moreover, in the same scenario, the probability for $n > n_{fix}$ is increased from 0.142 (when using only the effect from the planning phase) to 0.242 (when incorporating the observed effect). For later interim analyses, the difference between sample size distributions of the two recalculation methods become more similar. A reason for the increased standard deviation is that the effect size used for sample size calculation is a random variable and not fixed.

Therefore, early interim analysis timings should especially be avoided when incorporating the observed effect from the interim analysis to recalculate the sample size. Apart from the situation where Δ_0 is very small (see Figure 5.5a) and an early interim analysis is favorable due to the high n_{fix}, the average sample size and the standard deviation are smallest for later interim analysis timings. Furthermore, the probability $Pr(n > n_{fix})$ is also smaller for later interim analysis timings. However, a very late interim analysis is also not advantageous since it is very likely that the minimal sample size is chosen for the second stage due to the small required conditional power.

Results for using a higher c (see Figure B.1 and Figure B.2 for $c = 0.2$ in the Appendix) show similar characteristics as described previously when increasing c from 0 to 0.2: Obviously, the probability to select G_0 increases and moreover, the average sample size

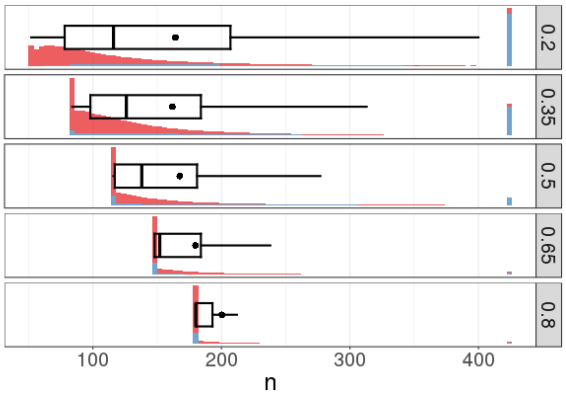
is especially inflated for early timings. This means the smallest average sample size is shifted to later timings.

To summarize, an interim analysis timing around $t = 0.5$ is suitable in most situations, although different prevalences and effect sizes influence the sample size distributions and should be taken into account when determining the timing of the interim analysis.



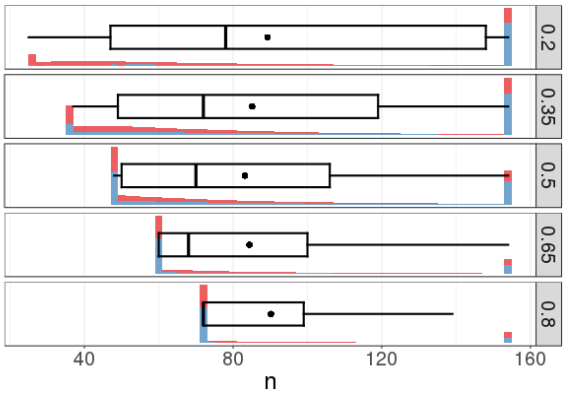
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	417 ± 113	0.997	0.76	0.001
0.35	676 ± 0	> 0.999	0	0
0.5	961 ± 0	> 0.999	0	0
0.65	1246 ± 0	> 0.999	0	0
0.8	1532 ± 0	> 0.999	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c = 0, n_{fix} = 1902$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	164 ± 117	0.677	0.89	0.242
0.35	160 ± 92	0.729	0.83	0.184
0.5	168 ± 75	0.766	0.78	0.166
0.65	180 ± 59	0.797	0.68	0.156
0.8	200 ± 49	0.821	0.52	0.171

(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c = 0, n_{fix} = 212$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	89 ± 47	0.499	0.95	0.504
0.35	85 ± 41	0.5	0.9	0.458
0.5	83 ± 37	0.5	0.85	0.426
0.65	85 ± 32	0.501	0.79	0.396
0.8	90 ± 29	0.501	0.72	0.393

(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0, n_{fix} = 77$

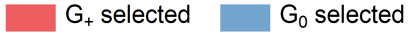
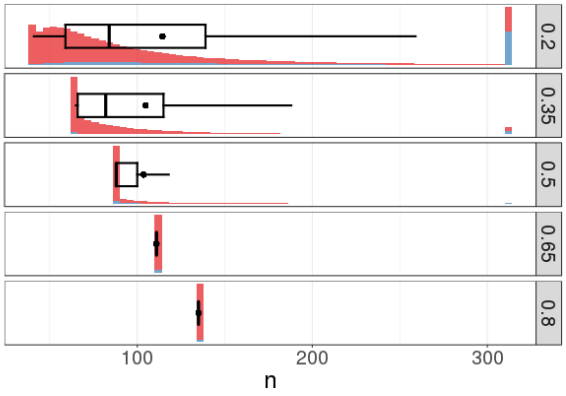
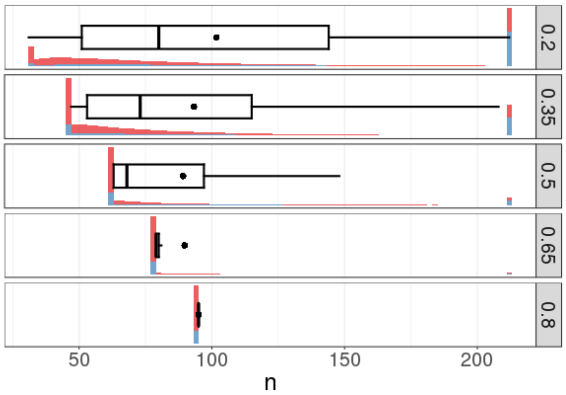


Figure 5.5: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0$; $p = 0.2$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.



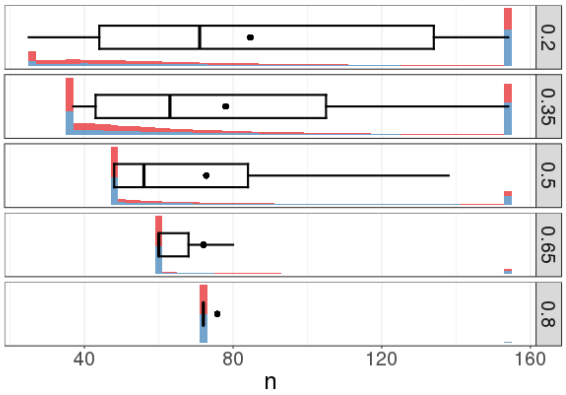
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	115 ± 79	0.817	0.82	0.211
0.35	105 ± 58	0.884	0.75	0.132
0.5	104 ± 38	0.924	0.6	0.07
0.65	111 ± 0	0.949	0	0
0.8	135 ± 0	0.965	0	0

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c = 0, n_{fix} = 156$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	102 ± 62	0.647	0.85	0.36
0.35	93 ± 52	0.689	0.79	0.282
0.5	89 ± 42	0.722	0.7	0.207
0.65	90 ± 28	0.749	0.48	0.118
0.8	95 ± 0	0.772	0	0

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c = 0, n_{fix} = 106$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	85 ± 46	0.5	0.87	0.458
0.35	78 ± 41	0.501	0.81	0.382
0.5	73 ± 35	0.501	0.72	0.293
0.65	72 ± 26	0.499	0.55	0.192
0.8	76 ± 15	0.5	0.15	0.08

(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0, n_{fix} = 77$

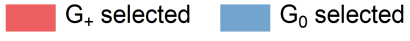


Figure 5.6: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0$; $p = 0.7$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.

5.2.3 Selection Rule Based on Absolute Effect Estimates

In this section, results using the selection rule based on absolute effect estimates are presented. This selection rule additionally includes the option to stop for futility (shown in yellow color) and the option to continue with both populations (shown in green color). The tables to the right of the figures contain the same statistics as in the previous subsection showing results for the selection rule based on estimated effect differences, however, the probability is given for a correct selection instead of the probability to select G_+.

5.2.3.1 Using the Effect Size from the Planning Phase for Sample Size Reassessment

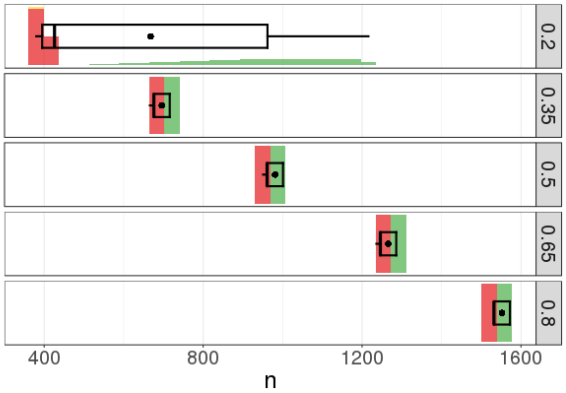
Firstly, results are considered in case the effect size from the planning phase is used for sample size reassessment. Figure 5.7 ($p = 0.2$) and Figure 5.8 ($p = 0.7$) depict results for $c_+ = 0.1$. In the scenario showing distributions for $\Delta_0 = 0.1$ and $p = 0.2$ (Figure 5.7a), the subgroup is selected with a probability $> 99\%$, while with a probability of 50% the total population is selected as well, and the co-primary analysis is conducted. As observed for the selection rule based on estimated effect differences, the smallest average sample size occurs for early interim timings if the difference between Δ_+ and Δ_0 is high. However, the sample size is clearly increased if both populations are selected at $t = 0.2$. For later interim analyses ($t > 0.35$) the sample size in the second stage is always the minimal sample size since the conditional power is 0. Practically, the scenarios with late interim analyses are hardly comparable since the power is larger than 80% ($86.5\%, 97.8\%, 99.8\%, > 99.9\%$ for $t = 0.35, 0.5, 0.65, 0.8$).
For higher Δ_- (Figure 5.7 b and c), interim analysis timings around $t = 0.5$ lead to the smallest average sample size. Moreover, if both populations are selected, the sample size is greatly increased for early interim analysis timings in settings with a small prevalence. The reason is that the maximum of the sample size calculated for the total population and the sample size calculated for the subgroup divided by the prevalence is used, of which the latter can become relatively large for small prevalences. Consequently, also $Pr(n > n_{fix})$ is high, especially if the selection of both populations is likely, as is the case for $\Delta_- = 0.5$. In this situation, the maximum possible sample size is used with a high probability for early timings, and for later timings, the sample size in the second stage is always the smallest possible sample size.
In scenarios with a higher prevalence (Figure 5.8), the sample size's standard deviation is smaller since the sample size distributions are more similar for different decisions. The timing with the smallest average sample size is still around 0.5 for scenarios with $\Delta_- > 0$.

For $\Delta_- = 0$, the smallest average sample size is observed for $t = 0.35$. Hence, for small and high prevalence, the smallest average sample size is achieved by later timings for higher Δ_-.

In each scenario, the correct interim decision is to select both populations, while only in the scenario with $p = 0.2$ and $\Delta_0 = 0.1$, only the subgroup should be selected. As expected, the probability for making a correct decision increases with increasing t. However, in the considered scenarios, a relatively high probability is already achieved for a timing around 0.5 and increases only slightly for higher t.

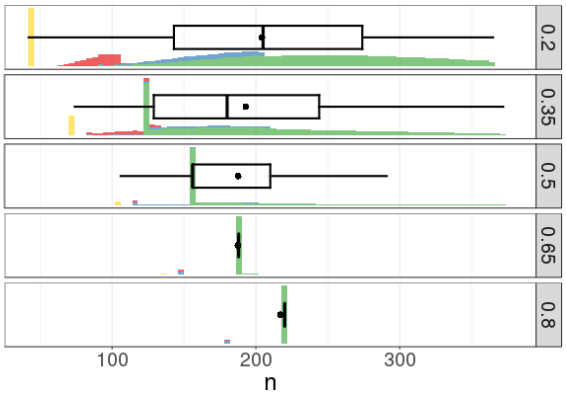
Overall, early interim analyses lead to a large standard deviation of the sample size and the probability to require a sample size that is larger than the sample size in the fixed design is relatively high in most scenarios. Late interim analyses can be deemed meaningless since the sample size of the first stage is already sufficient in many cases and the conditional power required to achieve an overall power of at least 80% is 0. Hence, an interim analysis after approximately half of the sample size that was calculated for a fixed setting seems reasonable in most situations. Only if the effect in the overall population is very small, early interim analyses in relation to the sample size in the fixed design, which is relatively large in this setting, are advantageous.

Results for a higher threshold c_+ are presented in Figure 5.9 for $p = 0.2$ and Figure 5.10 for $p = 0.7$. In this case, the probability to stop for futility and to select only G_0 is higher. Due to the higher probability for a futility stop, the conditional power is generally increased which leads to a higher average sample size compared to scenarios with a smaller c_+. This is especially pronounced for early interim analyses making early timings even less advisable. However, an exception is the scenario shown in Figure 5.9c, where the effects are equal in both populations and the prevalence is small. In this case, the average sample size is sometimes smaller compared to the scenario with $c_+ = 0.1$. In this scenario, both populations are selected with a high probability which often leads to the maximum possible sample size. Nevertheless, comparisons of the characteristics of the sample size distributions for different interim timings are very similar to distributions with a smaller c_+.



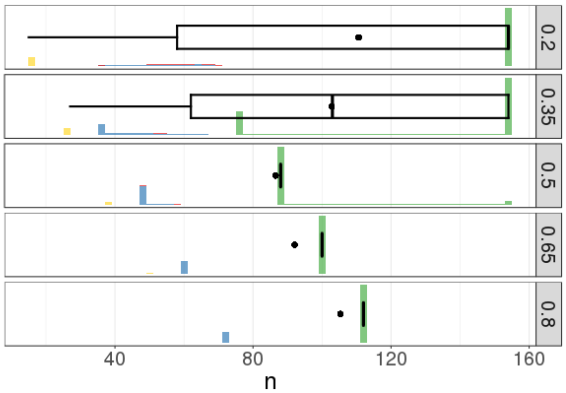
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	668 ± 300	0.494	0.59	0
0.35	696 ± 20	0.5	0	0
0.5	981 ± 20	0.5	0	0
0.65	1266 ± 20	0.5	0	0
0.8	1552 ± 20	0.5	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c_+ = 0.1, n_{fix} = 1902$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	204 ± 88	0.692	0.67	0.472
0.35	193 ± 74	0.791	0.56	0.349
0.5	188 ± 57	0.852	0.4	0.24
0.65	188 ± 19	0.895	0.08	0.058
0.8	217 ± 21	0.926	0	0.926

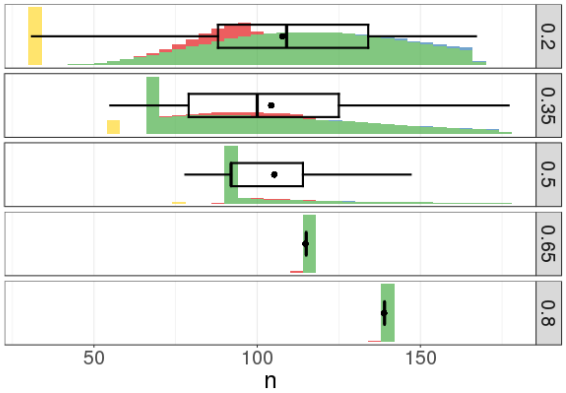
(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c_+ = 0.1, n_{fix} = 212$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	111 ± 52	0.636	0.61	0.629
0.35	102 ± 46	0.715	0.4	0.579
0.5	87 ± 27	0.767	0.12	0.767
0.65	92 ± 16	0.805	0	0.805
0.8	105 ± 15	0.834	0	0.834

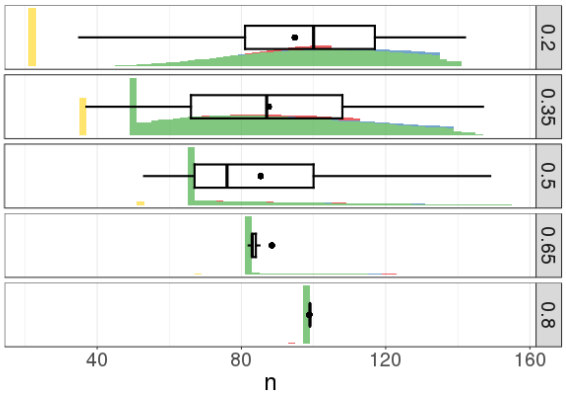
(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.1, n_{fix} = 77$

Figure 5.7: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.1$; $p = 0.2$. Sample size recalculation is based on the effect size from the planning phase.



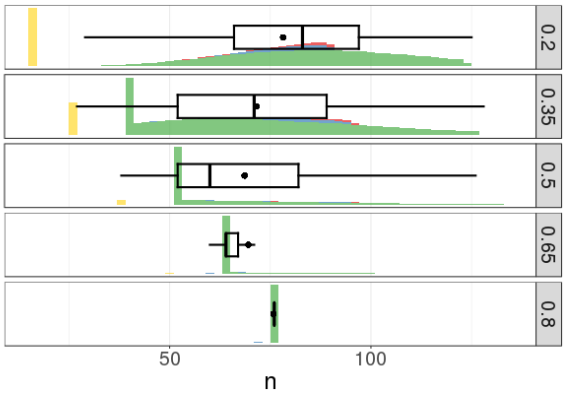
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	108 ± 34	0.818	0.72	0.065
0.35	104 ± 30	0.895	0.63	0.067
0.5	105 ± 21	0.937	0.47	0.04
0.65	115 ± 1	0.961	0	0
0.8	139 ± 1	0.975	0	0

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c_+ = 0.1, n_{fix} = 156$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	93 ± 30	0.808	0.8	0.365
0.35	88 ± 27	0.89	0.72	0.269
0.5	85 ± 23	0.934	0.6	0.197
0.65	88 ± 13	0.96	0.34	0.102
0.8	99 ± 1	0.975	0	0

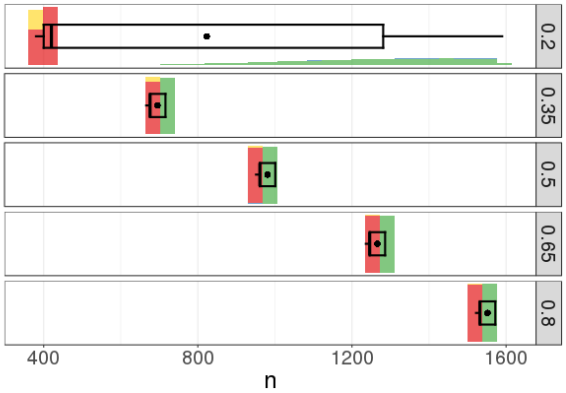
(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c_+ = 0.1, n_{fix} = 106$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	78 ± 30	0.788	0.76	0.591
0.35	72 ± 25	0.87	0.66	0.408
0.5	69 ± 21	0.918	0.53	0.296
0.65	70 ± 13	0.946	0.28	0.163
0.8	76 ± 2	0.964	0	0

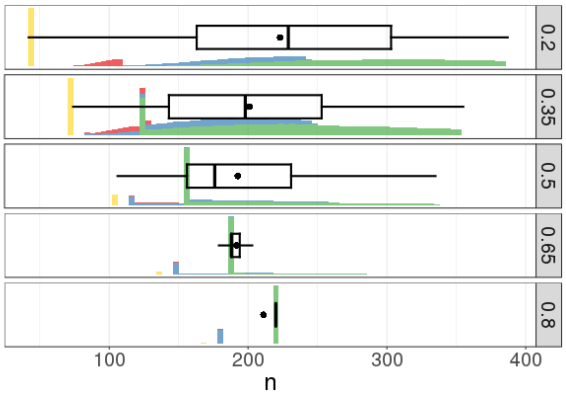
(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.1, n_{fix} = 77$

Figure 5.8: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.1$; $p = 0.7$. Sample size recalculation is based on the effect size from the planning phase.



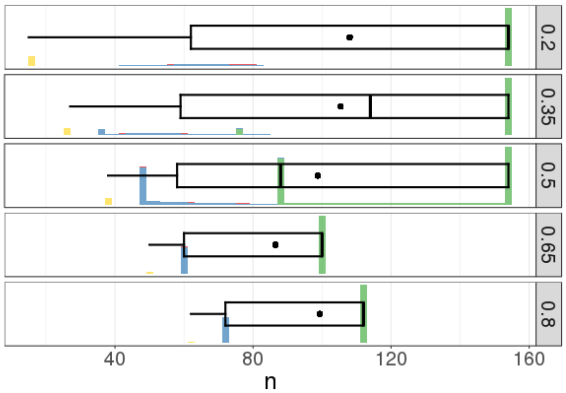
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	823 ± 455	0.413	0.74	0
0.35	695 ± 21	0.457	0	0
0.5	981 ± 20	0.478	0	0
0.65	1266 ± 20	0.489	0	0
0.8	1552 ± 20	0.494	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c_+ = 0.3, n_{fix} = 1902$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	223 ± 98	0.589	0.77	0.2
0.35	198 ± 72	0.663	0.67	0.064
0.5	191 ± 55	0.713	0.57	0.019
0.65	192 ± 31	0.75	0.37	0
0.8	211 ± 17	0.781	0	0

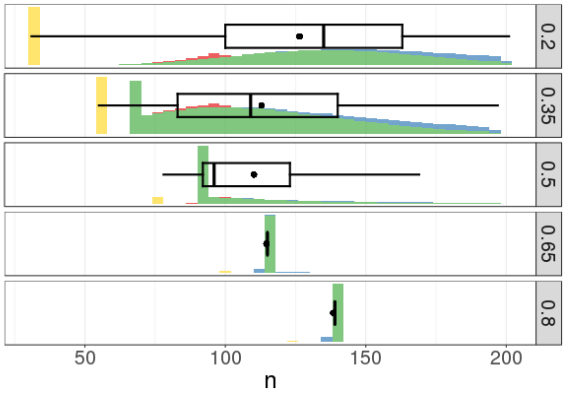
(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c_+ = 0.3, n_{fix} = 212$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	108 ± 52	0.558	0.71	0.6
0.35	105 ± 48	0.609	0.58	0.597
0.5	99 ± 41	0.642	0.41	0.666
0.65	87 ± 19	0.667	0	0.667
0.8	99 ± 19	0.688	0	0.688

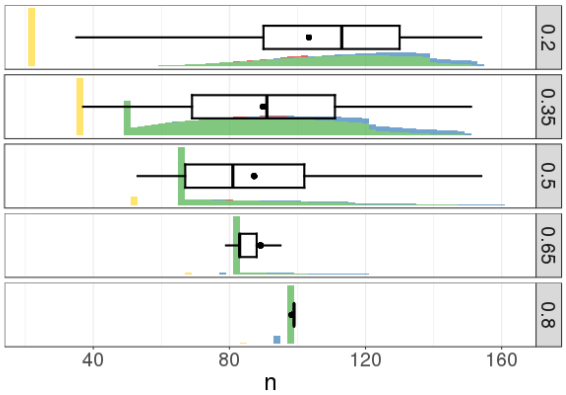
(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.3, n_{fix} = 77$

Figure 5.9: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.3$; $p = 0.2$. Sample size recalculation is based on the effect size from the planning phase.



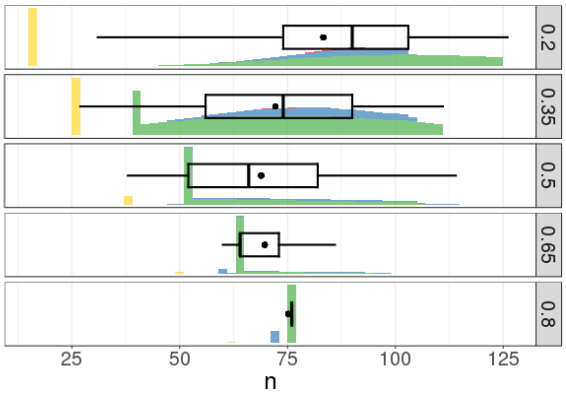
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	126 ± 48	0.717	0.82	0.304
0.35	113 ± 37	0.793	0.71	0.151
0.5	110 ± 27	0.843	0.57	0.092
0.65	115 ± 3	0.877	0.08	0
0.8	138 ± 2	0.903	0	0

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c_+ = 0.3, n_{fix} = 156$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	103 ± 38	0.69	0.87	0.582
0.35	90 ± 28	0.757	0.77	0.306
0.5	87 ± 24	0.802	0.68	0.211
0.65	89 ± 15	0.835	0.47	0.103
0.8	98 ± 2	0.861	0	0

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c_+ = 0.3, n_{fix} = 106$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	83 ± 31	0.668	0.83	0.708
0.35	72 ± 22	0.726	0.73	0.448
0.5	69 ± 18	0.767	0.64	0.328
0.65	70 ± 12	0.799	0.47	0.195
0.8	75 ± 2	0.823	0	0

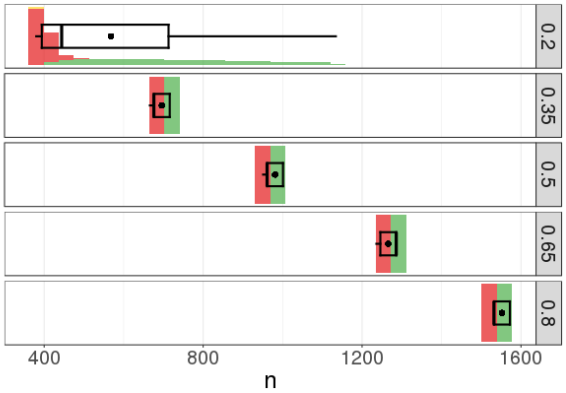
(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.3, n_{fix} = 77$

Figure 5.10: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.3$; $p = 0.7$. Sample size recalculation is based on the effect size from the planning phase.

5.2.3.2 Using the Mean of the Effect Size from the Planning Phase and the Interim Effect Estimate for Sample Size Reassessment

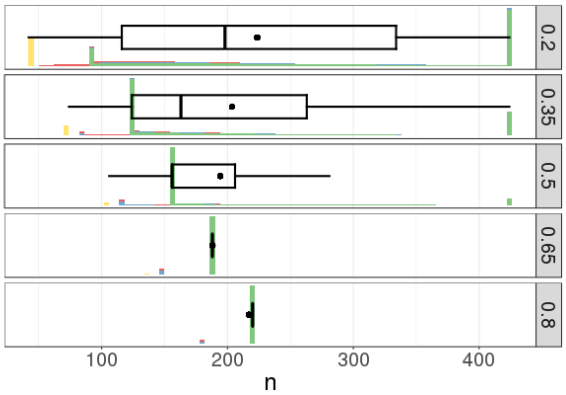
In this section, the mean of the assumed effect size from the planning phase and the interim effect estimate is used for sample size reassessment. Results are presented in Figure 5.11 and Figure 5.12 for $c_+ = 0.1$ and in the Appendix in Figure B.3 and Figure B.4 for $c_+ = 0.3$, for a prevalence of $p = 0.2$ and $p = 0.7$, respectively. As already observed for the other selection rule, when incorporating the observed interim effect for sample size reassessment, the standard deviation of the sample size is increased in many scenarios in comparison to a sample size recalculation which is only based on the effect size assumed in the planning stage. Furthermore, the maximum possible sample size is reached more often. This is especially pronounced for early interim analyses where the variability of the observed effect is higher due to the smaller sample size in the first stage. An exception is the scenario shown in Figure 5.11c. Here, the standard deviation is similar in comparison to the situation where only the effect assumed in the planning phase is used. In this case, the maximum sample size was also reached quite frequently when only using the effect from the planning phase.

The incorporation of the observed interim effect leads to the smallest or largest possible sample size in many cases. Despite the increased variance for early interim analysis timings, the timing with the smallest sample size is still around 0.5 for most of the considered scenarios. Only for the scenario with $p = 0.2$ and $\Delta_- = 0$ (Figure 5.11a), in which the sample size in the fixed design is very high, the sample size of the first stage is already sufficient for early timings in many cases and the reassessed sample size is mostly very small. Hence, the sample size increases with increasing interim analysis timing in this case. When using a higher c_+ (see Figure B.3 and Figure B.4 in the Appendix), characteristics are very similar. However, the probability to select G_+ is smaller and therefore, the probability to stop for futility and the probability to only select G_0 is higher. For early timings, the average sample size is increased when using a larger c_+ in scenarios with $\Delta_+ < 0.5$. Here, the maximum sample size is often reached in case only G_0 is selected. Therefore, when using a higher c_+, the smallest average sample size tends to be present for later interim analysis timings.



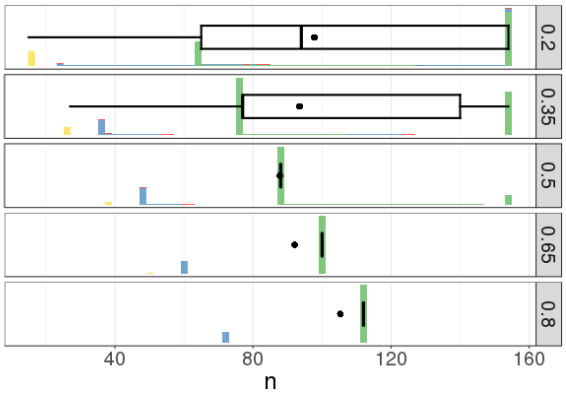
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	568 ± 217	0.494	0.55	0
0.35	696 ± 20	0.499	0	0
0.5	981 ± 20	0.5	0	0
0.65	1266 ± 20	0.499	0	0
0.8	1552 ± 20	0.5	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c_+ = 0.1, n_{fix} = 1902$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	224 ± 127	0.693	0.71	0.462
0.35	204 ± 105	0.789	0.49	0.351
0.5	194 ± 80	0.854	0.28	0.235
0.65	188 ± 24	0.896	0.03	0.05
0.8	217 ± 11	0.925	0	0.925

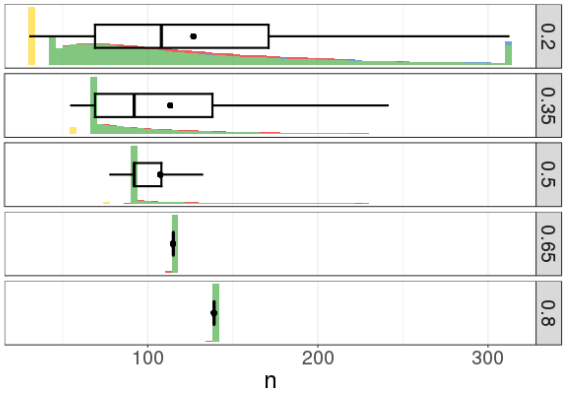
(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c_+ = 0.1, n_{fix} = 212$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	98 ± 49	0.636	0.64	0.576
0.35	94 ± 41	0.715	0.35	0.457
0.5	87 ± 29	0.767	0.07	0.768
0.65	92 ± 16	0.805	0	0.805
0.8	105 ± 15	0.834	0	0.834

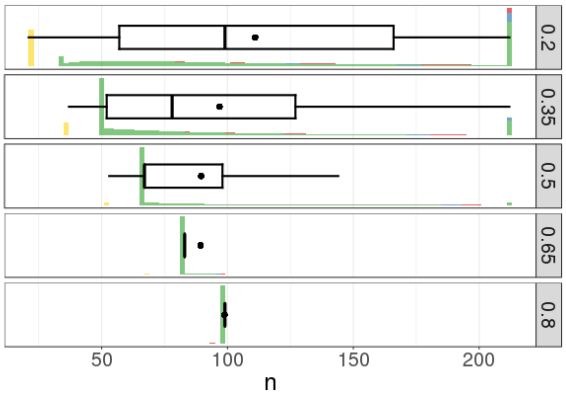
(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.1, n_{fix} = 77$

Figure 5.11: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.1$; $p = 0.2$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.



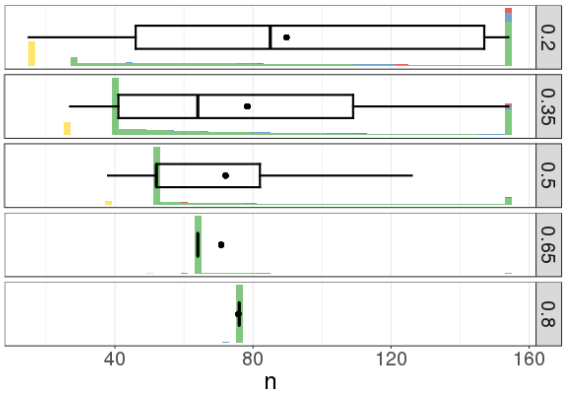
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	127 ± 75	0.818	0.72	0.295
0.35	113 ± 55	0.895	0.55	0.19
0.5	107 ± 31	0.936	0.33	0.089
0.65	115 ± 1	0.961	0	0
0.8	139 ± 1	0.975	0	0

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c_+ = 0.1, n_{fix} = 156$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	111 ± 64	0.808	0.8	0.464
0.35	97 ± 52	0.89	0.63	0.335
0.5	90 ± 38	0.934	0.44	0.216
0.65	90 ± 19	0.96	0.17	0.097
0.8	99 ± 1	0.976	0	0

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c_+ = 0.1, n_{fix} = 106$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	90 ± 49	0.788	0.8	0.54
0.35	78 ± 41	0.87	0.59	0.402
0.5	72 ± 32	0.917	0.39	0.279
0.65	71 ± 19	0.946	0.14	0.143
0.8	76 ± 2	0.964	0	0

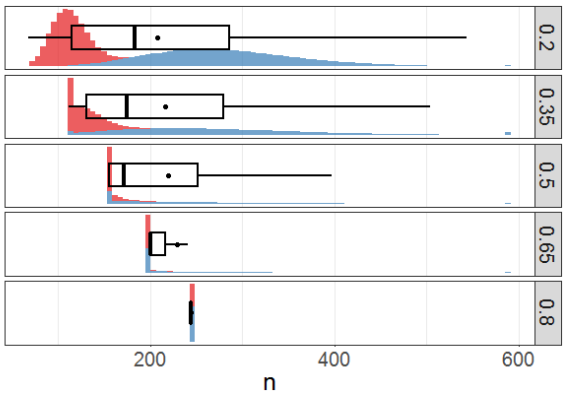
(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.1, n_{fix} = 77$

Figure 5.12: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.1$; $p = 0.7$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.

5.3 Clinical Trial Example

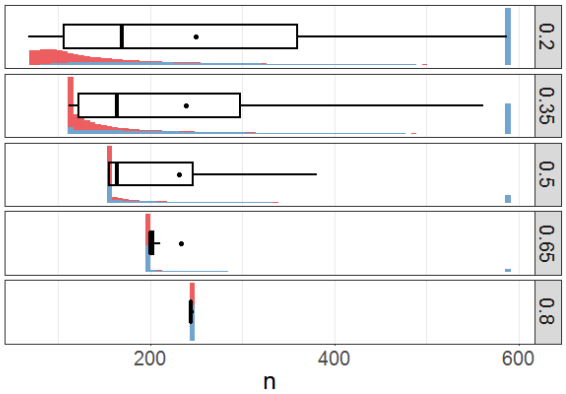
In this section, the choice of the interim analysis timing is investigated for parameters observed in the MILLY trial, which was introduced in Section 4.3. In Figure 5.13 and 5.14, sample size distributions are depicted for the selection rule based on estimated effect differences with $c = 0.2$. In Figure 5.13, the subgroup is defined by high and low periostin levels yielding the effect sizes $\Delta_{per_+} = 0.43$ and $\Delta_{per_-} = 0.08$. With a prevalence of $p = 0.5$, the effect size in the total patient population results in $\Delta_0 = 0.255$. Therefore, using $c = 0.2$, the correct decision is to select the total population as target population. When using the effect size from the planning phase for sample size reassessment (see Figure 5.13a), the average sample size is slightly increasing with increasing interim analysis timing. Regarding the smallest average sample size, an early interim analysis would be sensible. In contrast, the standard deviation of the sample size decreases for later timings. Especially for early timings, the sample size can be very high if G_0 is selected, and the probability $Pr(n > n_{fix})$ is largest.
When recalculating the sample size using the mean of the effect size from the planning phase and the interim effect estimate (see Figure 5.13b), the standard deviation for early interim analyses is clearly increased and also $Pr(n > n_{fix})$ is higher. In this case, the sample size reaches often the maximum possible sample size if G_0 is selected suggesting that a later interim analysis might be more advantageous. In addition, the average sample size is smallest for a timing around $t = 0.5$.
A similar picture is found for the subgroup defined by high and low Th2 level with effect sizes $\Delta_{Th2_+} = 0.34$ and $\Delta_{Th2_-} = 0.25$ (see Figure 5.14). However, the total population is selected with a higher probability, and when using the effect size from the planning phase for sample size reassessment, the average sample size is more similar for different interim analysis timings. When including the interim effect estimate for sample size recalculation, later timings are more advantageous in terms of a smaller average sample size, smaller standard deviation, larger probability to select the correct population G_0 and smaller probability $Pr(n > n_{fix})$.
Sample size distributions for the selection rule based on absolute effect estimates with $c_0 = 0.1$ and $c_+ = 0.3$ are presented in Figure 5.15 and Figure 5.16. For subgroups defined by periostin level ($\Delta_{per_+} = 0.43$, $\Delta_{per_-} = 0.08$), median and mean of the sample size are smallest for a timing around $t = 0.5$ irrespective of whether or not the observed effect from the interim analysis is incorporated in sample size recalculation. Moreover, the standard deviation at $t = 0.5$ is only approximately half the standard deviation for a timing at 0.2, and $Pr(n > n_{fix})$ is considerably reduced, which implies that a timing around 0.5 is appropriate in this situation. If the effects in both subgroups are more simi-

lar, as is the case when subgroups are defined by Th2 level ($\Delta_{Th2_+} = 0.34$, $\Delta_{Th2_-} = 0.25$), an interim analysis conducted at around $t = 0.65$ yields the smallest average sample size.



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	208 ± 105	0.447	0.80	0.231
0.35	217 ± 107	0.429	0.80	0.222
0.5	220 ± 94	0.415	0.75	0.176
0.65	230 ± 70	0.404	0.59	0.112
0.8	244 ± 0	0.394	0	0

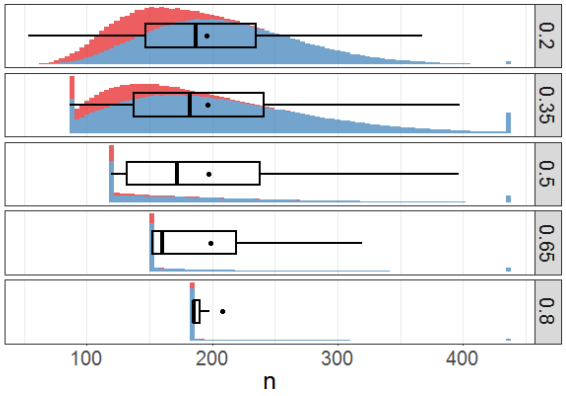
(a) Using the effect size from the planning phase for sample size reassessment



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	250 ± 180	0.447	0.86	0.306
0.35	239 ± 156	0.429	0.8	0.255
0.5	231 ± 127	0.415	0.7	0.193
0.65	234 ± 89	0.404	0.47	0.114
0.8	244 ± 0	0.394	0	0

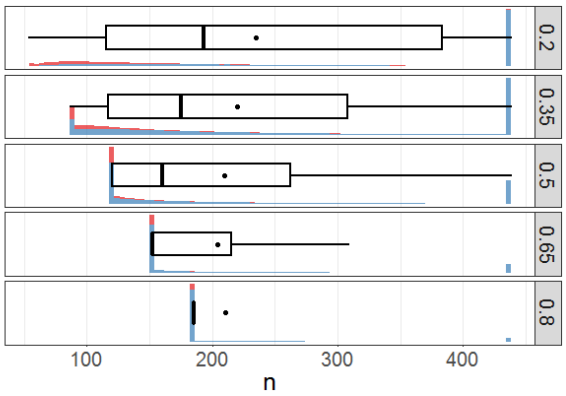
(b) Using the mean of the effect size from the planning phase and the interim effect estimate for sample size reassessment

Figure 5.13: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0.2$. Subgroups defined by periostin level ($\Delta_{per_+} = 0.43$, $\Delta_{per_-} = 0.08$); $n_{fix} = 293$.



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	195 ± 64	0.234	0.8	0.321
0.35	196 ± 77	0.169	0.8	0.326
0.5	197 ± 80	0.126	0.79	0.307
0.65	199 ± 73	0.096	0.73	0.247
0.8	208 ± 57	0.074	0.53	0.168

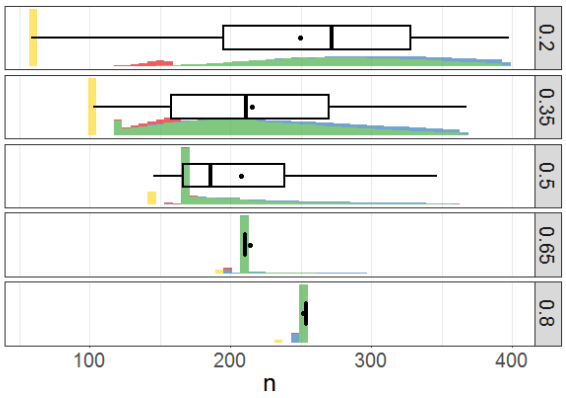
(a) Using the effect size from the planning phase for sample size reassessment



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	235 ± 136	0.234	0.87	0.443
0.35	220 ± 123	0.169	0.81	0.384
0.5	210 ± 109	0.126	0.75	0.322
0.65	204 ± 89	0.096	0.64	0.242
0.8	210 ± 65	0.074	0.39	0.154

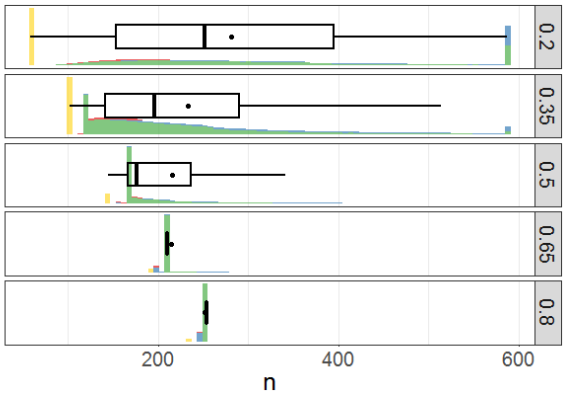
(b) Using the mean of the effect size from the planning phase and the interim effect estimate for sample size reassessment

Figure 5.14: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0.2$. Subgroups defined by Th2 level ($\Delta_{Th2_+} = 0.34$, $\Delta_{Th2_-} = 0.25$); $n_{fix} = 219$.



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	250 ± 101	0.638	0.86	0.406
0.35	216 ± 72	0.709	0.75	0.171
0.5	207 ± 54	0.760	0.63	0.102
0.65	214 ± 18	0.798	0.29	0.007
0.8	252 ± 5	0.828	0	0

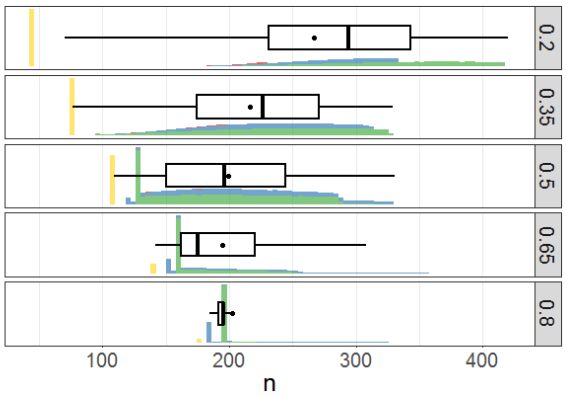
(a) Using the effect size from the planning phase for sample size reassessment



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	282 ± 167	0.638	0.90	0.414
0.35	234 ± 121	0.709	0.73	0.245
0.5	216 ± 82	0.760	0.56	0.151
0.65	215 ± 23	0.798	0.2	0.029
0.8	252 ± 5	0.828	0	0

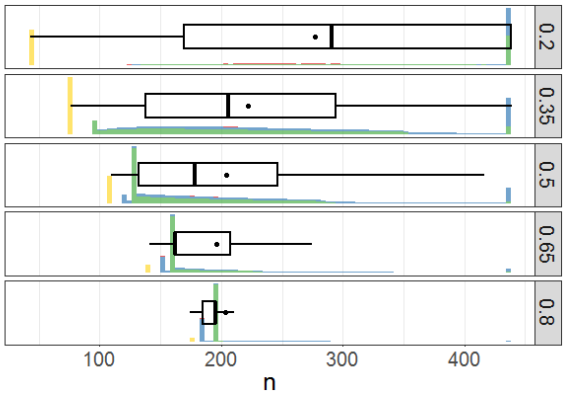
(b) Using the mean of the effect size from the planning phase and the interim effect estimate for sample size reassessment

Figure 5.15: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.3$. Subgroups defined by periostin level ($\Delta_{per_+} = 0.43$, $\Delta_{per_-} = 0.08$); $n_{fix} = 293$.



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	267 ± 111	0.531	0.9	0.781
0.35	216 ± 70	0.560	0.81	0.535
0.5	199 ± 57	0.578	0.75	0.379
0.65	194 ± 46	0.592	0.67	0.250
0.8	202 ± 31	0.603	0.43	0.115

(a) Using the effect size from the planning phase for sample size re- assessment



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	277 ± 141	0.531	0.95	0.651
0.35	222 ± 105	0.560	0.80	0.455
0.5	204 ± 86	0.578	0.70	0.337
0.65	196 ± 67	0.592	0.57	0.203
0.8	204 ± 41	0.603	0.28	0.101

(b) Using the mean of the effect size from the planning phase and the interim effect estimate for sample size reassessment

Figure 5.16: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.3$. Subgroups defined by Th2 level ($\Delta_{Th2_+} = 0.34$, $\Delta_{Th2_-} = 0.25$); $n_{fix} = 219$.

5.4 Chapter Summary

In this chapter, the impact of the interim analysis timing was investigated for an adaptive enrichment design with sample size recalculation based on conditional power arguments. Sample size distributions were simulated for different interim analysis timings and characteristics were compared for different prevalences, effect sizes and selection rules (based on the difference between effect estimates, and based on absolute effect estimates). Furthermore, two different methods to recalculate the sample size were considered (choosing either the assumed effect from the planning phase, or the mean of the effect from the planning phase and the observed interim effect).

In many situations, an interim analysis at $t = 0.5$ shows good properties as the average sample size is small and the standard deviation of the sample size is reduced in comparison to earlier timings. The probability to make a correct interim decision is obviously higher than for earlier timings. Furthermore, for later interim analysis timings, the required sample size in the first stage is very small, since an overall power of 80% is almost or already reached with the observations from the first stage. Only in case of very small Δ_0 (equivalent to small Δ_- and small prevalence), early timings are more advantageous. In this case, n_{fix} is very large and an early interim analysis already includes many observations, which leads to the selection of the subgroup with a high probability. Since the effect in the subgroup is substantially higher than in the total population for which n_{fix} was calculated, a much smaller sample size is needed.

The two different classes of selection rules (selection rule based on estimated effect differences and selection rule based on absolute effect estimates) do not lead to substantial differences as observed for the setting with a fixed sample size investigated in the preceding chapter. However, when using the selection rule based on absolute effect estimates, especially early timings lead to a larger average sample size, where a high sample size is mainly used if the co-primary analysis is selected. Therefore, when using the selection rule based on absolute effect estimates, an interim analysis timing around $t = 0.5$ is sensible in most of the scenarios. When using the selection rule based on estimated effect differences, earlier timings might lead to a smaller average sample size, especially if the effect in G_0 is smaller than in G_+. However, due to an often higher standard deviation and higher probability for $n > n_{fix}$, the benefit of conducting an early interim analysis is questionable.

Moreover, two different methods to specify the assumed effect were considered which also led to differences between sample size distributions. In comparison to the procedure using the effect size from the planning phase, the method that also incorporates the observed effect from the interim analysis tends to lead to a higher average sample size

and standard deviation, especially for early interim analysis timings. This means, if the observed interim effect is used for sample size reassessment, particularly early timings should be avoided due to the higher uncertainty of effect estimation for a smaller sample size.

It should be noted that in the simulation studies, the assumed effect in the planning phase was considered to be equal to the true effect. Results might be different if the assumed effect in the planning phase is not the true effect and a better estimation can be achieved when incorporating the observed interim effect. Furthermore, the sample size of the fixed design depends on the assumed effect in the planning phase and therefore, also the timing of the interim analysis. If a different effect in the planning phase is assumed, the calculated sample size of the fixed design and thus also the timing of the interim analysis would be different. This means that if the assumed effect is too high, n_{fix} is smaller and, thereby, the sample size in the first stage is smaller for the same t. Assuming a too small effect will result in a higher n_{fix} and a larger first-stage sample size for the same t.

Chapter 6

Discussion

Adaptive enrichment designs have become more popular in recent years due to an increased interest in targeted therapies. The methodology offers a useful tool for selecting the patient population with the most promising treatment effect and testing for efficacy in a single trial. However, to profit from the advantages of this design, a careful choice of the interim analysis timing is crucial as this thesis shows.

In this thesis, the impact of the interim analysis timings was investigated for different settings and for two different classes of selection rules using simulation studies. In the first part, a fixed overall sample size is considered where the impact of different timings on the power was investigated. Although the initially proposed adaptive enrichment design assumed a fixed, prespecified sample size of the second stage (see, e.g., Wang et al., 2007; Brannath et al., 2009; Jenkins et al., 2011), adaptive designs in general offer the possibility to reassess the sample size. This might be very useful since, in general, assumptions in the planning phase are uncertain. For example, specification of the treatment effects in different populations might be vague, and in the planning phase it is not known for which population the confirmatory proof of efficacy will be conducted at the end of the trial. Therefore, adaptive enrichment designs including a sample size reassessment for the second stage are considered in the second part of the thesis. In this case, the overall sample size distribution is compared for different interim analysis timings.

When the overall sample size is fixed, results showed that the timing of the interim analysis has indeed an impact on the power of the study. Of course, the degree of influence depends on the specific scenario as different selection rules, the effect sizes, and the prevalences lead to different power characteristics.

In case the power maximum occurs at a medium timing and not at an extreme timing, it is sensible to use the power maximum to define the interim analysis timing. In this case, the sample size in the first stage is not too small and the correct target population can

be selected with a reasonable certainty, and on the other hand, the remaining number of patients for the second stage is large enough to have a decisive influence on the final test statistic and hence, on the power of the study.
When using the selection rule based on estimated effect differences, where either the subgroup or the total population is selected, the power maximum is achieved at extremely early timings if treatment effects are similar for the subgroup and the total population. If one is solely interested in achieving a high power, a very early conduct of the interim analysis would be advantageous. However, in this case, the choice of the target population is more or less random which is not the intention when using an adaptive enrichment design. This characteristic also indicates that the use of the selection rule based on the difference between the effect in the subgroup and the total population may be inappropriate under certain conditions. If the treatment effect in the total population is smaller than the treatment effect in the subgroup by a certain amount but the effect is still relevant, it would be desirable to select both populations.
The selection rule based on absolute effect estimates that was considered includes two further options, namely the possibility to test for efficacy in the subgroup and the total population in the final analysis in case both treatment effects are promising, and to stop for futility in case the interim analysis shows futile effect sizes. The latter option leads to a power loss if the interim analysis is conducted early. This means, early interim analyses cannot be recommended in general using this selection rule. Nevertheless, the futility stop is a useful option to prevent further patients from receiving a potentially inefficient treatment. Whether later timings show a higher or smaller power depends on the scenario. However, in many cases, power is relatively constant after half of the patients are enrolled, which means that a timing of $t = 0.5$ has in general no considerable disadvantages in comparison to later timings.
Overall, for a prespecified, fixed sample size, the selection rule, effect sizes, and also the prevalence have an impact on the power characteristics for different interim analysis timings. However, not only the power should be considered, especially, if the power is maximal for extremely early or late timings. For example, it should be taken into account that the probability for a false decision is increased for an early interim analysis timing. On the other hand, for late timings, it is not possible anymore to influence the conduct of the study to a meaningful extent. Furthermore, if the subgroup is selected, the first stage should not be too large since the number of patients from the complementary group should be as small as possible due to ethical and financial reasons.
If not only the target population is selected in the interim analysis but also the sample size for the second stage is reassessed based on the observed interim results, different selection rules and effect sizes do not lead to substantial differences when comparing the

sample size distribution for different interim analysis timings. In many scenarios, the average sample size is smallest for early to medium timings. However, also the standard deviation of the second-stage sample size is in general high for early timings. For late timings, it was observed that the sample size in the first stage was already sufficient in many cases and hence, the number of patients in the second stage is the smallest possible sample size which indicates that a late interim analysis is not sensible.

Nevertheless, the sample size distribution depends on the scenario, i.e. the effect sizes, prevalence and the applied selection rule, and therefore, the interim analysis timing should be selected individually for each situation.

Furthermore, it makes a difference how the assumed effect for sample size recalculation is chosen. In this thesis, two different approaches were considered: using the assumed effect from the planning phase, and using the mean of the assumed effect from the planning phase and the observed interim effect. The comparison of both approaches has shown, particularly for early interim analyses, that the average sample size and especially the standard deviation of the sample size is increased if the observed effect from the interim analysis is included. The large variability for early interim analyses in case the observed interim effect is incorporated was also shown by Bauer and Koenig (2006) who investigated the impact of the interim analysis timing for common adaptive designs without subgroup selection. Furthermore, the high standard deviation relates in many cases to a high probability that the maximum possible sample size is used, which implies that the upper limit for sample size should be chosen carefully.

The described findings also suggest that it might be advantageous to use the assumed effect from the planning phase for sample size recalculation to prevent extremely large sample sizes. However, it has to be noted that in the conducted simulation studies the used effect from the planning phase was assumed to be the true effect, and results might be different if this assumption is not true, especially if the assumed effect in the planning phase differs considerably from the true effect. If there is high uncertainty about the treatment effect and the observed effect from the planning phase shall be used for sample size recalculation in addition to conditioning on the observed test statistic, early interim analyses should be avoided to obtain more precise treatment effect estimates.

Anyhow, there are some further limitations. At first, simulations were used to investigate characteristics for different interim analysis timings, and no analytical formula could be derived to specify, for example, the power maximum depending on the effect size, the prevalence and the selection rule. Even if the simulated scenarios give a comprehensive insight to the impact of different interim analysis timings, no general rules could be established.

Moreover, only normally distributed endpoints were considered, and, for example, time-

to-event endpoints, which are often considered in oncological trials, were not investigated. However, it is reasonable to assume that results do not considerably differ.
For sake of simplicity, some further assumptions were made. For example, it is assumed that the classification of a patient to the subgroup or the complementary group is correct with a probability of 100%. In practice, this perfection might not be given and incorrect assignment might occur which leads to biased effect estimates for the different populations. In addition, the prevalence of the subgroup was assumed to be fix. Furthermore, known standard deviation was assumed and the z-test was used, which is not practical. However, results should be similar when using, for example, a t-test in case the standard deviation is not known.
To conclude, findings of this thesis show that regardless of whether the overall sample size is fixed in advance, or the sample size of the second stage is recalculated, the interim analysis timing has to be chosen carefully for the specific design features and parameter assumptions at hand since no general rules could be established and no specific timing of the interim analysis can be recommended that uniformly fits to all scenarios. Instead, sensitivity analyses taking the specific design features into account should be conducted in the planning stage of a trial to determine the appropriate timings of an interim analysis.

Chapter 7

Summary

English

This thesis deals with adaptive enrichment designs, which are especially applied in the development of targeted therapies. These designs are devised for the situation in which a higher treatment effect is assumed in a specific subgroup but efficacy cannot be ruled out in the total population. The idea of this two-stage study design is to decide in an interim analysis based on observed treatment effects whether the subgroup or the total population is selected for enrichment in the second stage of the trial, and for which population a test for efficacy is conducted in the final analysis.

The aim of this thesis is to investigate the impact of the interim analysis timing on the power of the study for a normally distributed endpoint. Different effect sizes and prevalences of the subgroup, as well as two different classes of selection rules were considered. The first selection rule is based on the comparison of the estimated effect difference between the subgroup and the total population with a prespecified threshold value, and the subgroup or the total population is selected, respectively. The second selection rule that is considered is based on the absolute effect estimates from the subgroup and the total population, each compared to a prespecified threshold value. Possible options of this selection rule are to select either the subgroup, the total population, both populations, or no population. The latter option leads to termination of the study (stop for futility) without rejecting any hypothesis.

In the first part, the impact of the interim timing on the power of the study is investigated for a fixed overall sample size. Analytical derivation of the power was not possible, and, as a consequence, power was determined using simulation studies. Results showed that the interim timing influences the power of the study to a varying degree for different scenarios. In particular, the chosen selection rule leads to different power characteristics as a function of interim analysis timing. For example, the power is rather small in case

the second selection rule, that is based on absolute effect estimates, is used, which can be explained by the incorporated option to stop for futility. In contrast, for the first selection rule, which is based on the difference between the effect estimates, the smallest power was achieved for early timings in many scenarios. Additionally, the power maximum depends on the effect sizes and the prevalence of the subgroup. This shows that there is no particular interim timing which is optimal with regard to the power in every scenario. When choosing the interim analysis timing, the assumed effect sizes, the prevalence, and the chosen selection rule should be taken into account.

In the second part, an adaptive enrichment design including sample size recalculation is considered. Sample size was recalculated using conditional power arguments, where both the assumed effect from the planning phase and the mean of this effect and the observed effect in the interim analysis was used. The timing of the interim analysis was defined as the ratio of the sample size in the first stage and the sample size that would be required in a corresponding study design for demonstrating efficacy in the total population without interim analysis. In simulation studies, different interim timings were compared based on the distribution of the overall sample size. In particular, the average sample size and its standard deviation as well as the probability to achieve a sample size that is larger than the sample size for the respective design without interim analysis was considered. Results showed that different selection rules, effect sizes and prevalences have a smaller impact, and an interim analysis after half the patients have been enrolled leads to the smallest average sample size in many cases.

For both situations (fixed and adapted sample size), the choice of the interim analysis timing was investigated for a clinical trial example.

In summary, this thesis shows that the choice of the interim analysis timing in adaptive enrichment designs has, in many cases, a substantial effect on the power of the study or the average sample size. However, the most appropriate timing depends on the effect sizes, the prevalence of the subgroup and the chosen selection rule, and should be selected carefully in the planning phase for the specific scenario at hand.

Deutsch

Diese Arbeit beschäftigt sich mit adaptiven Enrichment-Designs, die insbesondere in der Entwicklung von zielgerichteten Therapien verwendet werden. Diese sind für die Situation konzipiert, in der ein höherer Therapieeffekt in einer bestimmten Subgruppe vermutet wird, gleichzeitig aber auch eine Wirksamkeit für die Gesamtpopulation nicht ausgeschlossen werden kann. Die Idee dieses zweistufigen Studiendesigns besteht darin, basierend auf den beobachteten Effekten in einer Interimanalyse zu entscheiden, ob die Rekrutierung in der zweiten Stufe aus der Subgruppe oder der Gesamtpopulation erfolgt und für welcher Population in der finalen Auswertung der Wirksamkeitsnachweis durchgeführt werden soll.
In dieser Arbeit wurde der Einfluss des Interimzeitpunktes auf die Power einer Studie für einen normalverteilten Endpunkt untersucht. Dabei wurden verschiedene Effektgrößen und Prävalenzen der Subgruppe sowie zwei verschiedene Entscheidungsregeln betrachtet. Bei der ersten Entscheidungsregel wird die Differenz zwischen den Effektschätzern aus der Subgruppe und der Gesamtpopulation mit einem zuvor spezifizierten Schwellenwert verglichen und entsprechend die Subgruppe oder die gesamte Patientenpopulation ausgewählt. Die zweite Entscheidungsregel, die betrachtet wurde, vergleicht die Effektschätzer aus der Subgruppe und der Gesamtpopulation mit jeweils einem Schwellenwert. Bei dieser Entscheidungsregel können entweder nur die Subgruppe, nur die Gesamtpopulation, beide Populationen, oder auch keine Population ausgewählt werden. Letztere Option führt zu einem frühzeitigen Abbruch der Studie ohne Ablehnung einer Hypothese. Im ersten Teil wurde der Einfluss des Interimzeitpunktes auf die Power der Studie für eine feste Gesamtfallzahl untersucht. Die analytische Berechnung der Power war nicht möglich, sodass die Power mit Hilfe von Simulationsstudien bestimmt wurde. Dabei zeigte sich, dass die Power der Studie je nach Szenario unterschiedlich stark durch den Interimzeitpunkt beeinflusst werden kann. Insbesondere führte die gewählte Entscheidungsregel zu verschiedenen Powercharakteristika in Abhängigkeit von der Zeit. Zum Beispiel, ist die Power bei Verwendung der zweiten Entscheidungsregel, die auf den absoluten Effektschätzern basiert, für frühe Interimanalysen eher gering, was durch die Option des frühzeitigen Abbruchs der Studie erklärt werden kann. Bei der ersten Entscheidungsregel, die auf der Differenz zwischen den Effektschätzern basiert, gab es dagegen einige Szenarien für die eine frühe Zwischenauswertung optimal bezüglich der Power war. Zudem wurde das Powermaximum durch die Effektgrößen und die Prävalenz der Subgruppe beeinflusst. Dies macht deutlich, dass es nicht einen einzigen Interimzeitpunkt gibt, der in jeder Situation optimal ist. Bei der Wahl des Zwischenauswertungszeitpunktes sollten stattdessen die angenommenen Effektgrößen und die Prävalenz sowie die

gewählte Entscheidungsregel berücksichtigt werden.
Im zweiten Teil wurde ein adaptives Enrichment-Design mit Fallzahlrekalkulation betrachtet. Die Fallzahl wurde dabei mit Hilfe der conditional Power rekalkuliert, sowohl basierend auf dem Effekt aus der Planungsphase als auch basierend auf dem Mittelwert aus diesem Wert und dem beobachteten Effekt in der Zwischenauswertung. Der Zeitpunkt der Interimanalyse wurde hier als Verhältnis der Fallzahl in der ersten Stufe und der Fallzahl, die man in einem entsprechenden Studiendesign für den Nachweis eines Effektes in der Gesamtpopulation ohne Interimanalyse benötigen würde, definiert. Verschiedene Interimzeitpunkte wurden anhand der Verteilung der Gesamtfallzahl verglichen, die mit Hilfe von Simulationen bestimmt wurden. Insbesondere wurde die erwartete Fallzahl und deren Standardabweichung als auch die Wahrscheinlichkeit eine Fallzahl zu erhalten, die größer ist als im entsprechenden Design ohne Interimanalyse, untersucht. Hier zeigten verschiedene Entscheidungsregeln, Effektgrößen und Prävalenzen einen weniger großen Einfluss, und eine Interimanalyse nach der Hälfte der Patienten, die man im entsprechenden Design ohne Zwischenauswertung benötigt hätte, weist in vielen Fällen eine geringe erwartete Fallzahl auf.
In beiden Teilen (feste Fallzahl und Fallzahlrekalkulation) wurde die Wahl des Interimzeitpunktes auch für ein konkretes klinisches Beispiel untersucht.
Zusammengefasst zeigt die Arbeit, dass die Wahl des Interimzeitpunktes in adaptiven Enrichment-Designs die Power der Studie oder die erwartete Fallzahl in vielen Fällen wesentlich beeinflusst. Der optimale Zeitpunkt hängt dabei von den Effektgrößen, der Prävelenz der Subgruppe und der gewählten Entscheidungsregel ab, und sollte daher in der Planungsphase in Abhängigkeit der vorliegenden Parameter mit Bedacht gewählt werden.

Bibliography

Baselga, J. (2001), 'Herceptin alone or in combination with chemotherapy in the treatment of her2-positive metastatic breast cancer: Pivotal trials', *Oncology* **61**(suppl 2), 14–21.

Bauer, P. (1989), 'Multistage testing with adaptive designs', *Biometrie und Informatik in Medizin und Biologie* **20**(4), 130–148.

Bauer, P., Bretz, F., Dragalin, V., König, F. and Wassmer, G. (2016), 'Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls', *Statistics in Medicine* **35**, 325–347.

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

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

Benner, L. and Kieser, M. (2018), 'Timing of the interim analysis in adaptive enrichment designs', *Journal of Biopharmaceutical Statistics* **28**, 622–632.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. and Racine-Poon, A. (2009), 'Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology', *Statistics in Medicine* **28**, 1445–1463.

Chen, C., Li, N., Shentu, Y., Pang, L. and Beckman, R. A. (2016), 'Adaptive informational design of confirmatory phase III trials with an uncertain biomarker effect to improve the probability of success', *Statistics in Biopharmaceutical Research* **8**, 237–247.

Corren, J., Lemanske, R. F., Hanania, N. A., Korenblat, P. E., Parsey, M. V., Arron, J. R., Harris, J. M., Scheerens, H., Wu, L. C., Su, Z., Mosesova, S., Eisner, M. D., Bohen, S. P. and Matthews, J. G. (2011), 'Lebrikizumab treatment in adults with asthma', *New England Journal of Medicine* **365**, 1088–98.

European Medicines Agency (2007), 'Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.', *London, UK. http://www.ema.europa.eu.* .

Food and Drug Administration (2018), 'Draft guidance for industry adaptive designs for clinical trials of drug and biologics', *Food and Drug Administration. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER)* .

Friede, T., Parsons, N. and Stallard, N. (2012), 'A conditional error function approach for subgroup selection in adaptive clinical trials', *Statistics in Medicine* **31**, 4309–4320.

Friede, T. and Stallard, N. (2008), 'A comparison of methods for adaptive treatment selection', *Biometrical Journal* **50**(5), 767–781.

Jenkins, M., Stone, A. and Jennison, C. (2011), 'An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints', *Pharmaceutical Statistics* **10**, 347–356.

Kelly, P. J., Stallard, N. and Todd, S. (2005), 'An adaptive group sequential design for phase ii/iii clinical trials that select a single treatment from several', *Journal of Biopharmaceutical Statistics* **15**, 641–658.

Krisam, J. and Kieser, M. (2014), 'Decision rules for subgroup selection based on a predictive biomarker', *Journal of Biopharmaceutical Statistics* **24**(1), 188–202.

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

Marcus, R., Peritz, E. and Gabriel, K. (1976), 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika* **63**(3), 655–60.

Pérez-Herrero, E. and Fernández-Medarde, A. (2015), 'Advanced targeted therapies in cancer: Drug nanocarriers, the future of chemotherapy', *European Journal of Pharmaceutics and Biopharmaceutics* **93**, 52–79.

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

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Renfro, L. A., Coughlin, C. M., Grothey, A. M. and Sargent, D. J. (2014), 'Adaptive randomized phase II design for biomarker threshold selection and independent evaluation', *Chinese Clinical Oncology* **3**(1), pii: 3489.

Sarkar, S. K. and Chang, C.-K. (1997), 'The Simes method for multiple hypothesis testing with positively dependent test statistics', *Statistics in Medicine* **92**, 1601–1608.

Simes, R. J. (1986), 'An improved Bonferroni procedure for multiple tests of significance', *Biometrika* **73**, 751–754.

Wang, S.-J., Hung, H. M. J. and O'Neill, R. T. (2009), 'Adaptive patient enrichment designs in therapeutic trials', *Biometrical Journal* **51**, 358–374.

Wang, S.-J., O'Neill, R. T. and Hung, H. J. (2007), 'Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset', *Pharmaceutical Statistics* **6**, 227–244.

Wassmer, G. and Brannath, W. (2016), *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics. Switzerland: Springer.

Wassmer, G. and Dragalin, V. (2015), 'Designing issues in confirmatory adaptive population enrichment trials', *Journal of Biopharmaceutical Statistics* **25**, 651–669.

Wolfram Reseach, Inc. (2018), Mathematica, version 11.3. Champaign, IL.

Appendix A

Derivation of Power Function

In the following, detailed derivation of the probability shown in formula (4.1.1) is presented referring to the rejection of $H_0^{(0)}$:

$$
\begin{aligned}
&Pr(reject\ H_0^{(0)}\ \cap\ select\ G_0)\\
&= Pr(Z_0 > z_{1-\alpha/4}\ \cap\ \widehat{\Delta}_0 \geq \widehat{\Delta}_+ - c)\\
&= Pr\left(\sqrt{t}\widehat{\Delta}_0\sqrt{\frac{tn}{2}} + \sqrt{1-t}\widehat{\Delta}_0^{II}\sqrt{\frac{(1-t)n}{2}} > z_{1-\alpha/4} \bigcap \widehat{\Delta}_0 \geq \widehat{\Delta}_+ - c\right)\\
&= Pr\left(\widehat{\Delta}_0 > \underbrace{\left(z_{1-\alpha/4} - (1-t)\widehat{\Delta}_0^{II}\sqrt{\frac{n}{2}}\right)\frac{\sqrt{2}}{t\sqrt{n}}}_{=:l(\widehat{\Delta}_0^{II})} \bigcap \widehat{\Delta}_0 \geq \widehat{\Delta}_+ - c\right)\\
&= \int_{\delta_0^{II}=-\infty}^{\infty}\int_{\delta_0=l(\delta_0^{II})}^{\infty}\int_{\delta_+=-\infty}^{\delta_0+c} f_{\widehat{\Delta}_0,\widehat{\Delta}_+}(\delta_0,\delta_+) f_{\widehat{\Delta}_0^{II}}(\delta_0^{II})\ d\delta_+\ d\delta_0\ d\delta_0^{II}\\
&= \int_{\delta_0^{II}=-\infty}^{\infty}\int_{\delta_0=l(\delta_0^{II})}^{\infty}\int_{\delta_+=-\infty}^{\delta_0+c} \frac{tn}{8\pi}\sqrt{\frac{(1-t)np}{(1-p)\pi}}\cdot exp\left\{-\frac{tn}{4(1-p)}\Big((\delta_0-\Delta_0)^2 + p(\delta_+-\Delta_+)^2\right.\\
&\quad \left.-2p(\delta_0-\Delta_0)(\delta_+-\Delta_+) + (\delta_0^{II}-\Delta_0)^2\frac{(1-t)(1-p)}{t}\Big)\right\}\ d\delta_+\ d\delta_0\ d\delta_0^{II}\\
&= \int_{\delta_0^{II}=-\infty}^{\infty}\int_{\delta_0=l(\delta_0^{II})}^{\infty} \Big[-\frac{n}{8\pi}\sqrt{(t-1)t}\cdot exp\Big\{-\frac{n}{4}\left(\Delta_0^2 + \delta_0^{II2} - 2\Delta_0\left(t\left(\delta_0-\delta_0^{II}\right)+\delta_0^{II}\right)\right.\\
&\quad \left.+t\left(\delta_0^2-\delta_0^{II2}\right)\right)\Big\}\cdot Erfi\left\{\frac{\sqrt{ntp}}{2\sqrt{p-1}}\left(-\Delta_0+\Delta_+-\delta_++\delta_0\right)\right\}\Big]_{\delta_+=-\infty}^{\delta_0+c} d\delta_0\ d\delta_0^{II}\\
&= \int_{\delta_0^{II}=-\infty}^{\infty}\int_{\delta_0=l(\delta_0^{II})}^{\infty} -\frac{n}{8\pi}\sqrt{(1-t)t}\cdot exp\Big\{-\frac{n}{4}\left(\Delta_0^2 + \delta_0^{II2} - 2\Delta_0\left(t\left(\delta_0-\delta_0^{II}\right)+\delta_0^{II}\right)\right.
\end{aligned}
$$

$$
\begin{aligned}
&+t(\delta_0^2-\delta_0^{II2})\Big)\Big\}\cdot\left(Erf\left\{\frac{\sqrt{ntp}}{2\sqrt{1-p}}(\Delta_+-\Delta_0-c)\right\}-1\right)\ d\delta_0\ d\delta_0^{II}\\
&=\int_{\delta_0^{II}=-\infty}^{\infty}\left[\frac{\sqrt{n(1-t)}}{8\sqrt{\pi}}\cdot exp\left\{\frac{n}{4}(t-1)\left(\Delta_0-\delta_0^{II}\right)^2\right\}\right.\\
&\quad\cdot\left(1+Erf\left\{\frac{1}{2}\sqrt{\frac{ntp}{1-p}}(c+\Delta_0-\Delta_+)\right\}\right)\\
&\quad\left.\cdot\ Erf\left\{\frac{1}{2}\sqrt{nt}(\delta_0-\Delta_0)\right\}\right]_{\delta_0=l(\delta_0^{II})}^{\infty}\ d\delta_0^{II}\\
&=\int_{\delta_0^{II}=-\infty}^{\infty}\frac{\sqrt{n(1-t)}}{8\sqrt{\pi}}\cdot exp\left\{\frac{n}{4}(t-1)\left(\Delta_0-\delta_0^{II}\right)^2\right\}\\
&\quad\cdot\left(1+Erf\left\{\frac{1}{2}\sqrt{\frac{ntp}{1-p}}(c+\Delta_0-\Delta_+)\right\}\right)\\
&\quad\cdot\left(1-Erf\left\{\frac{1}{2}\sqrt{nt}\left(\left(z_{1-\alpha/4}-(1-t)\delta_0^{II}\sqrt{\frac{n}{2}}\right)\frac{\sqrt{2}}{t\sqrt{n}}-\Delta_0\right)\right\}\right)\ d\delta_0^{II}\\
&=\frac{\sqrt{n(1-t)}}{8\sqrt{\pi}}\left(1+Erf\left\{\frac{1}{2}\sqrt{\frac{ntp}{1-p}}(c+\Delta_0-\Delta_+)\right\}\right)\\
&\quad\int_{\delta_0^{II}=-\infty}^{\infty}exp\left\{\frac{n}{4}(t-1)\left(\Delta_0-\delta_0^{II}\right)^2\right\}\\
&\quad\cdot\left(1-Erf\left\{\frac{1}{2}\sqrt{nt}\left(\left(z_{1-\alpha/4}-(1-t)\delta_0^{II}\sqrt{\frac{n}{2}}\right)\frac{\sqrt{2}}{t\sqrt{n}}-\Delta_0\right)\right\}\right)\ d\delta_0^{II}
\end{aligned}
$$

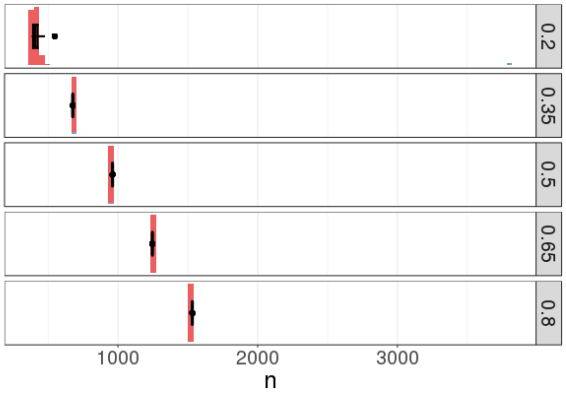
The probability for rejecting the hypothesis $H_0^{(+)}$ shown in formula (4.1.2) can be derived in a similar way:

$$
\begin{aligned}
&Pr(Z_+>z_{1-\alpha/4}\ \cap\ \widehat{\Delta}_0<\widehat{\Delta}_+-c)\\
&=Pr\left(\sqrt{\frac{tp}{tp+1-t}}\widehat{\Delta}_+\sqrt{\frac{tnp}{2}}+\sqrt{\frac{1-t}{tp+1-t}}\widehat{\Delta}_+^{II}\sqrt{\frac{(1-t)n}{2}}>z_{1-\alpha/4}\bigcap\ \widehat{\Delta}_0<\widehat{\Delta}_+-c\right)\\
&=Pr\left(\widehat{\Delta}_+^{II}>\underbrace{\left(z_{1-\alpha/4}\sqrt{\frac{2(tp+1-t)}{n}}-tp\widehat{\Delta}_+\right)\frac{1}{1-t}}_{=:l(\widehat{\Delta}_+)}\bigcap\ \widehat{\Delta}_0<\widehat{\Delta}_+-c\right)\\
&=\int_{\delta_+=-\infty}^{\infty}\int_{\delta_0^{II}=l(\delta_+)}^{\infty}\int_{\delta_0=-\infty}^{\delta_+-c}f_{\widehat{\Delta}_0,\widehat{\Delta}_+}(\delta_0,\delta_+)f_{\widehat{\Delta}_+^{II}}(\delta_+^{II})\ d\delta_0\ d\delta_0^{II}\ d\delta_+
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\delta_+=-\infty}^{\infty} \int_{\delta_+^{II}=l(\delta_+)}^{\infty} \int_{\delta_0=-\infty}^{\delta_+-c} \frac{tn}{8\pi}\sqrt{\frac{(1-t)np}{(1-p)\pi}} \cdot exp\left\{-\frac{tn}{4(1-p)}\Big((\delta_0-\Delta_0)^2 + p(\delta_+-\Delta_+)^2 \right. \\
&\quad \left. -2p(\delta_0-\Delta_0)(\delta_+-\Delta_+) + (\delta_+^{II}-\Delta_+)^2\frac{(1-t)(1-p)}{t}\Big)\right\}\, d\delta_0\, d\delta_+^{II}\, d\delta_+ \\
&= \int_{\delta_+=-\infty}^{\infty} \int_{\delta_+^{II}=l(\delta_+)}^{\infty} \frac{n}{8\pi}\sqrt{(1-t)tp} \\
&\quad \cdot exp\left\{-\frac{n}{4}\left((1+(p-1)t)\,\Delta_+^2 + pt\delta_+^2 + (1-t)\delta_+^2 - 2\Delta_+\left(pt\delta_+ + \delta_+^{II} - t\delta_+^{II}\right)\right)\right\} \\
&\quad \cdot \left(Erf\left\{\frac{\sqrt{nt}}{2\sqrt{1-p}}\left(-\Delta_0 + p(\Delta_+-\delta_+) + \delta_+ - c\right)\right\} + 1\right)\, d\delta_+^{II}\, d\delta_+ \\
&= \int_{\delta_+=-\infty}^{\infty} \frac{1}{8}\sqrt{\frac{npt}{\pi}} \cdot exp\left\{-\frac{npt}{4}(\Delta_+-\delta_+)^2\right\} \\
&\quad \cdot \left(1 + Erf\left\{\frac{\sqrt{nt}}{2\sqrt{1-p}}(-\Delta_0 + p(\Delta_+-\delta_+) + \delta_+ - c)\right\}\right) \\
&\quad \cdot \left(1 + Erf\left\{-\frac{\sqrt{n(1-t)}}{2}\left(-\Delta_+ + \left(z_{1-\alpha/4}\sqrt{\frac{2(tp+1-t)}{n}} - tp\delta_+\right)\frac{1}{1-t}\right)\right\}\right)\, d\delta_+
\end{aligned}
$$

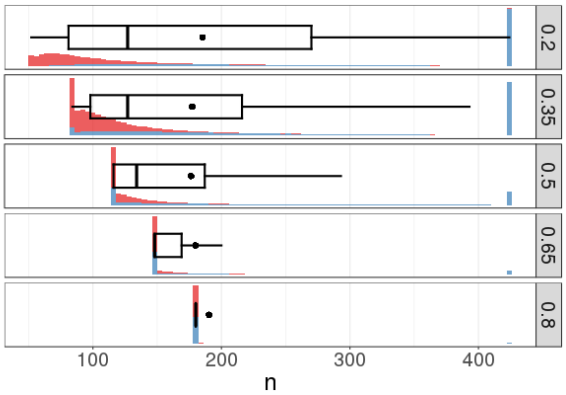
Appendix B

Additional Figures



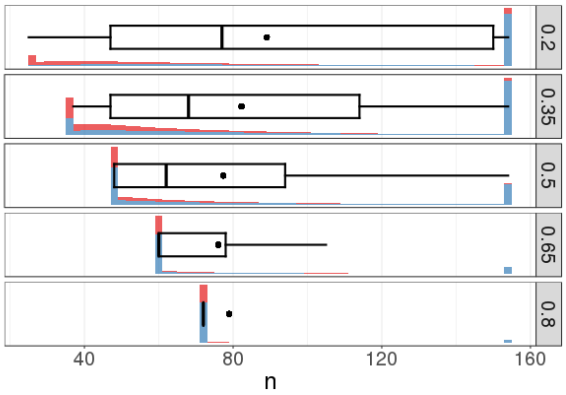
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	548 ± 573	0.916	0.78	0.038
0.35	676 ± 0	0.966	0	0
0.5	961 ± 0	0.986	0	0
0.65	1246 ± 0	0.994	0	0
0.8	1532 ± 0	0.997	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c = 0.2, n_{fix} = 1902$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	182 ± 130	0.5	0.9	0.306
0.35	177 ± 110	0.5	0.85	0.255
0.5	176 ± 91	0.501	0.79	0.205
0.65	180 ± 70	0.501	0.66	0.146
0.8	191 ± 43	0.5	0.27	0.069

(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c = 0.2, n_{fix} = 212$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	89 ± 47	0.391	0.95	0.499
0.35	82 ± 42	0.357	0.89	0.422
0.5	78 ± 36	0.331	0.82	0.347
0.65	76 ± 28	0.309	0.7	0.256
0.8	79 ± 20	0.289	0.42	0.16

(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0.2, n_{fix} = 77$

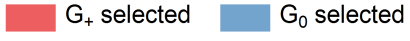
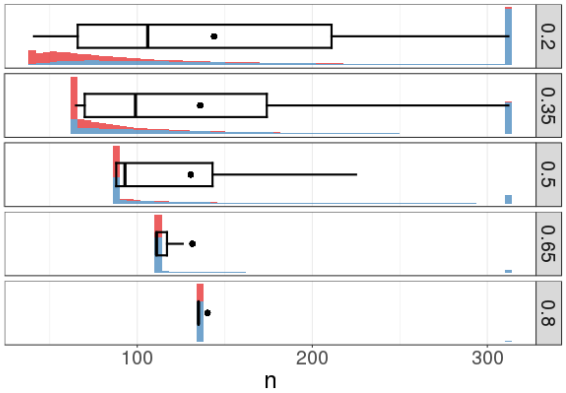
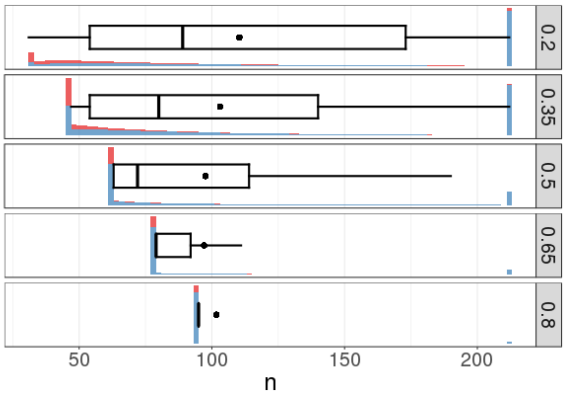


Figure B.1: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0.2$; $p = 0.2$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.



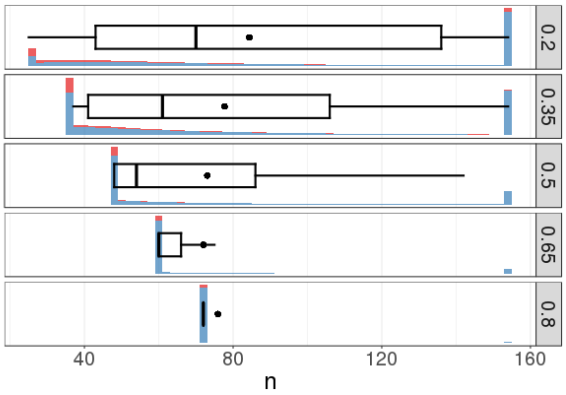
t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	144 ± 95	0.382	0.84	0.342
0.35	136 ± 83	0.345	0.78	0.287
0.5	131 ± 69	0.317	0.69	0.221
0.65	132 ± 50	0.294	0.48	0.138
0.8	140 ± 25	0.273	0.06	0.045

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c = 0.2, n_{fix} = 156$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	111 ± 65	0.268	0.85	0.42
0.35	103 ± 58	0.205	0.79	0.357
0.5	98 ± 49	0.163	0.7	0.277
0.65	97 ± 37	0.132	0.54	0.191
0.8	102 ± 24	0.107	0.19	0.096

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c = 0.2, n_{fix} = 106$



t	ASS± SD	G_+ sel.	CP	$n > n_{fix}$
0.2	84 ± 47	0.198	0.85	0.456
0.35	78 ± 42	0.131	0.78	0.378
0.5	73 ± 35	0.09	0.69	0.294
0.65	72 ± 26	0.063	0.5	0.187
0.8	76 ± 15	0.045	0.13	0.084

(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c = 0.2, n_{fix} = 77$

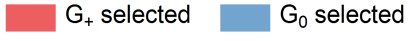
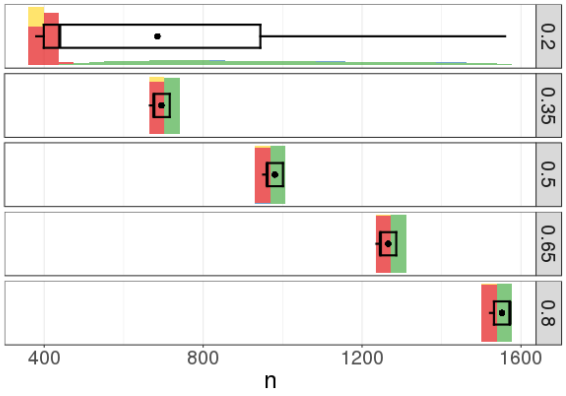
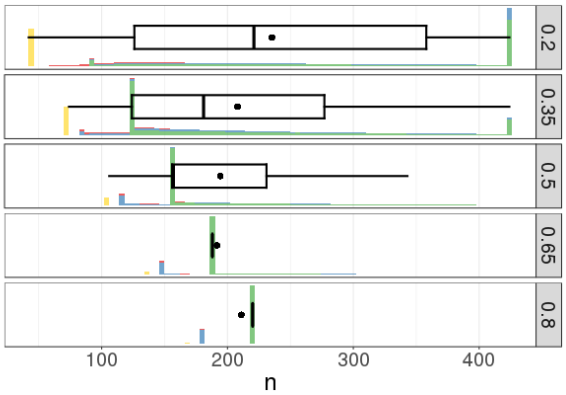


Figure B.2: Distribution of sample size using the selection rule based on estimated effect differences with $c = 0.2$; $p = 0.7$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.



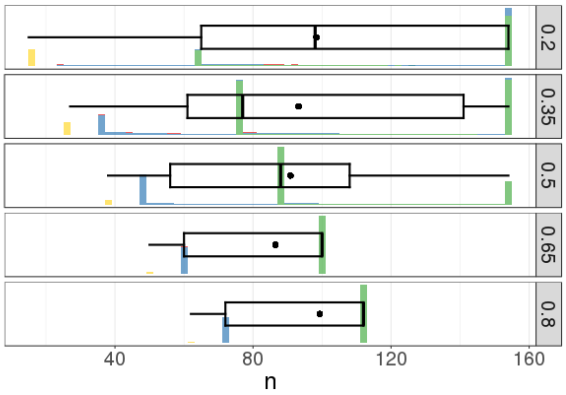
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	686 ± 354	0.412	0.73	0
0.35	695 ± 21	0.457	0	0
0.5	981 ± 20	0.478	0	0
0.65	1266 ± 20	0.488	0	0
0.8	1552 ± 20	0.493	0	0

(a) $p = 0.2, \Delta_- = 0, \Delta_0 = 0.1, c_+ = 0.1, n_{fix} = 1902$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	235 ± 130	0.589	0.83	0.518
0.35	208 ± 101	0.661	0.65	0.403
0.5	195 ± 71	0.713	0.49	0.297
0.65	192 ± 35	0.751	0.26	0.155
0.8	211 ± 17	0.781	0	0.781

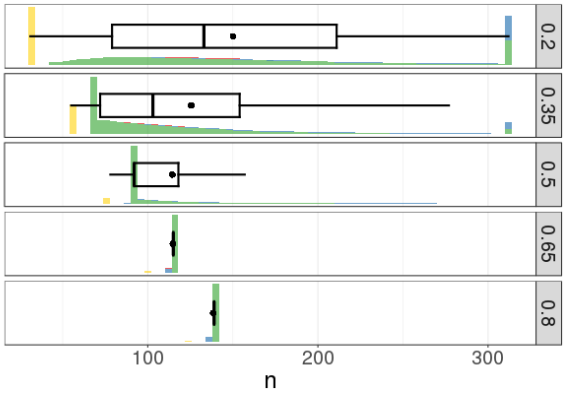
(b) $p = 0.2, \Delta_- = 0.25, \Delta_0 = 0.3, c_+ = 0.1, n_{fix} = 212$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	98 ± 49	0.557	0.75	0.598
0.35	93 ± 43	0.608	0.51	0.495
0.5	91 ± 36	0.643	0.3	0.695
0.65	87 ± 19	0.667	0	0.667
0.8	99 ± 19	0.686	0	0.686

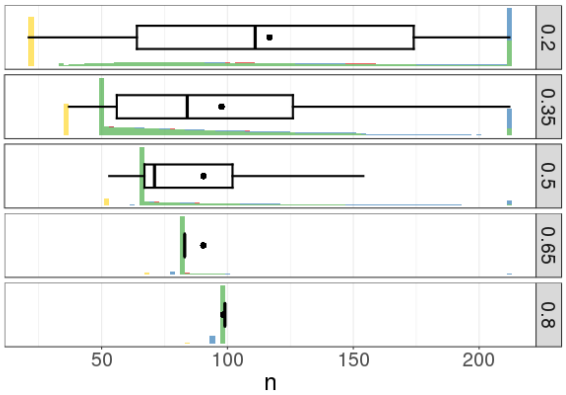
(c) $p = 0.2, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.1, n_{fix} = 77$

Figure B.3: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.3$; $p = 0.2$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.



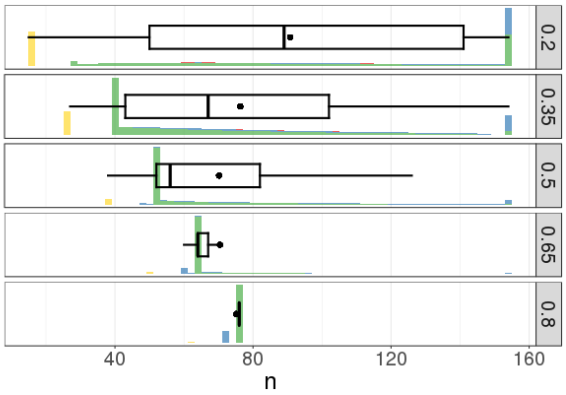
t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	150 ± 90	0.717	0.87	0.404
0.35	126 ± 68	0.793	0.68	0.243
0.5	144 ± 44	0.842	0.46	0.132
0.65	115 ± 4	0.877	0.04	0
0.8	138 ± 2	0.904	0	0

(a) $p = 0.7, \Delta_- = 0, \Delta_0 = 0.35, c_+ = 0.1, n_{fix} = 156$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	119 ± 65	0.69	0.9	0.538
0.35	98 ± 50	0.757	0.72	0.353
0.5	91 ± 38	0.802	0.56	0.225
0.65	90 ± 23	0.836	0.31	0.102
0.8	98 ± 2	0.861	0	0

(b) $p = 0.7, \Delta_- = 0.25, \Delta_0 = 0.425, c_+ = 0.1, n_{fix} = 106$



t	ASS± SD	corr. sel.	CP	$n > n_{fix}$
0.2	91 ± 48	0.667	0.87	0.572
0.35	76 ± 37	0.726	0.68	0.414
0.5	70 ± 27	0.767	0.52	0.29
0.65	70 ± 17	0.798	0.31	0.154
0.8	75 ± 2	0.823	0	0

(c) $p = 0.7, \Delta_- = 0.5, \Delta_0 = 0.5, c_+ = 0.1, n_{fix} = 77$

Figure B.4: Distribution of sample size using the selection rule based on absolute effect estimates with $c_+ = 0.3$; $p = 0.7$. Sample size recalculation is based on the mean of the effect size from the planning phase and the interim effect estimate.

Appendix C

Selected R Program Code

C.1 Code for Design with Fixed Sample Size

```
library(MASS)
library(xlsx)
library(matrixcalc)
library(RColorBrewer)
library(ggplot2)
library(tidyr)

#####################################################################
# Function to calculate selection and rejection probabilities
# for selection rule based on estimated effect differences
#####################################################################

#####################################################################
# Parameters to specify:
# p = prevalence of subgroup
# n = overall sample size per treatment group
# delta1 = standardized effect in subgroup
# delta2 = standardized effect in complement
# tv = vector of interim analysis timings
# m = number of simulated studies
# c = threshold of selection rule
# adj = method to control for multiplicity, either "ClosedTesting" or
#       "Bonferroni"
# alpha = significance level
# seed = seed value
#####################################################################

#####################################################################
# Returned values (vectors for different t)
# prSub    = probability to select subgroup
# powerSub = probability to reject null hypothesis of subgroup
# powerTot = probability to reject null hypothesis of total population
# powerOverall = overall power
#####################################################################

Power_Sel1 <- function(p, n, delta1, delta2, tv, m, c, adj, alpha, seed){
  set.seed(seed)
```

```
delta0 <- p*delta1 + (1-p)*delta2

prSub <- as.numeric()
powerSub <- as.numeric()
powerTot <- as.numeric()
powerOverall <- as.numeric()
i <- 0

for(t in tv){

  i <- i+1

  n1 <- t*n    # sample size stage I
  n2 <- n - n1 # sample size stage II

  if(is.positive.definite(matrix(c(2/n1, 2/n1, 2/n1, 2/(n1*p)), 2, 2),
     tol = 1e-8)){

    # ------------------------
    # Stage I
    # ------------------------

    Delta_1 <- mvrnorm(n=m, mu=c(delta0, delta1), Sigma = matrix(c(2/n1,
        2/n1, 2/n1, 2/(n1*p)), 2, 2))

    # subgroup
    Z1_1 <- Delta_1[,2] * sqrt(n1*p/2)
    p1_1 <- 1 - pnorm(Z1_1)

    # total population
    Z0_1 <- Delta_1[,1] * sqrt(n1/2)
    p0_1 <- 1 - pnorm(Z0_1)

    # combination of p-values (Closed Testing Procedure)
    p01_1 <- apply(rbind(2 * apply(rbind(p1_1, p0_1), 2, min),
                         apply(rbind(p1_1, p0_1), 2, max)), 2, min)

    selSub <- ifelse(Delta_1[,1] + c < Delta_1[,2], 1, 0)


    # ------------------------
    # Stage II
    # ------------------------

    # if subgroup is selected
    delta1_2 <- rnorm(n = m, mean = delta1, sd = sqrt(2/n2))

    Z1_2_sub <- delta1_2 * sqrt(n2/2)
    U1_sub <- sqrt(p*n1/(p*n1+n2)) * Z1_1 + sqrt(n2/(p*n1+n2)) * Z1_2_
       sub
    p1_sub <- 1 - pnorm(U1_sub)

    # combination of p-values (Closed Testing Procedure)
    U01_sub <- sqrt(n1/n) * qnorm(1-p01_1) + sqrt(n2/n) * Z1_2_sub
    p01_sub <- 1 - pnorm(U01_sub)


    # if total population is selected
```

```
      if(is.positive.definite(matrix(c(2/n2, 2/n2, 2/n2, 2/(n2*p)),2,2),
         tol=1e-8)){

        Delta_2 <- mvrnorm(n = m, mu = c(delta0, delta1), Sigma = matrix(c
           (2/n2, 2/n2, 2/n2, 2/(n2*p)), 2, 2))

        Z0_2_tot <- Delta_2[,1] * sqrt(n2/2)
        U0_tot <- sqrt(n1/n) * Z0_1 + sqrt(n2/n) * Z0_2_tot
        p0_tot <- 1 - pnorm(U0_tot)

        # combination of p-values (Closed Testing Procedure)
        U01_tot <- sqrt(n1/n) * qnorm(1-p01_1) + sqrt(n2/n) * Z0_2_tot
        p01_tot <- 1 - pnorm(U01_tot)

      }
    }

    p_single <- ifelse(selSub == 1, p1_sub, p0_tot)
    p_closed <- ifelse(selSub == 1, p01_sub, p01_tot)

    prSub[i] <- sum(selSub)

    if(adj == "Bonferroni"){
      powerSub[i] <- sum(p_single[selSub == 1] < alpha/4)/m
      powerTot[i] <- sum(p_single[selSub == 0] < alpha/4)/m
    }

    if(adj == "ClosedTesting"){
      powerSub[i] <- sum(p_closed[selSub == 1] < alpha/2 & p_single[selSub
          == 1] < alpha/2)/m
      powerTot[i] <- sum(p_closed[selSub == 0] < alpha/2 & p_single[selSub
          == 0] < alpha/2)/m
    }
  }
  powerOverall <- powerSub + powerTot

  return(list(prSub = prSub, powerSub = powerSub, powerTot = powerTot,
     powerOverall = powerOverall, n = n))
}



#########################################################################
# Function to calculate sample size to assure a power of 80%
# at a specific timing t
#########################################################################

#########################################################################
# Parameters to specify:
# p = prevalence of subgroup
# delta1 = standardized effect in subgroup
# delta2 = standardized effect in complement
# m = number of simulated studies
# c = threshold for selection rule
# t = interim analysis timing
# adj = method to control for multiplicity, either "ClosedTesting" or
#       "Bonferroni"
# alpha = significance level
```

```
# seed = seed value
#########################################################################

#########################################################################
# Returned value:
# n = sample size per group
#########################################################################

calculate_n_Sel1 <- function(p, delta1, delta2, m, c, t, adj, alpha, seed)
   {

  n_start <- 5
  powerOverall <- 0

  while(powerOverall < 0.8){
    n_start <- n_start + 50
    res <- Power_Sel1(p, n_start, delta1, delta2, t, m, c, adj, alpha,
        seed)
    powerOverall <- res$powerOverall
  }

  n_start2 <- n_start - 50
  powerOverall <- 0

  while(powerOverall < 0.8){
    n_start2 <- n_start2 + 1
    res <- Power_Sel1(p, n_start2, delta1, delta2, t, m, c, adj, alpha,
        seed)
    powerOverall <- res$powerOverall
  }
  n <- n_start2
  return(n)
}



#########################################################################
# Function to calculate selection and rejection probabilities
# for selection rule based on absolute effect estimates
#########################################################################

#########################################################################
# Parameters to specify:
# p = prevalence of subgroup
# n = overall sample size per treatment group
# delta1 = standardized effect in subgroup
# delta2 = standardized effect in complement
# tv = vector of interim analysis timings
# m = number of simulated studies
# c1, c0 = threshold values for selection rule
# adj = method to control for multiplicity, either "ClosedTesting" or
#       "Bonferroni"
# alpha = significance level
# seed = seed value
#########################################################################

#########################################################################
# Returned values (vectors for different t):
```

```
# prSub, prTot, prCop, prFut = probability to select subgroup,
#       total population, co-primary analysis, futility stop
# powerSub = probability to reject null hypothesis of subgroup
# powerTot = probability to reject null hypothesis of total population
# powerOverall = overall power
# powerBoth_cp = probability to reject both hypotheses
# powerSub_cp = probability to reject the null hypothesis of subgroup and
#               co-primary analysis is selected
# powerTot_cp = probability to reject the null hypothesis of total pop.
#               and co-primary analysis is selected
# powerOverall_cp = probability to reject at least one hypothesis and
#                   co-primary analysis is selected
# powerSub_Sub = probability to reject the null hypothesis of subgroup
#                and subgroup only is selected
# powerTot_Tot = probability to reject the null hypothesis of total pop.
#                and total pop. is selected
# n = sample size
#########################################################################

Power_Sel2 <- function(p, n, delta1, delta2, tv, m, c1, c0, adj, alpha,
    seed){

  set.seed(seed)
  z_alpha <- qnorm(1-alpha/2)
  delta0 <- p*delta1 + (1-p)*delta2

  prSub <- as.numeric()
  prTot <- as.numeric()
  prCop <- as.numeric()
  prFut <- as.numeric()
  powerSub <- as.numeric()
  powerTot <- as.numeric()
  powerOverall <- as.numeric()
  powerSub_cp <- as.numeric()
  powerTot_cp <- as.numeric()
  powerBoth_cp <- as.numeric()
  powerOverall_cp <- as.numeric()
  powerSub_Sub <- as.numeric()
  powerTot_Tot <- as.numeric()

  i <- 0

  for(t in tv){
    i <- i + 1

    n1 <- t * n  # sample size per group in stage I
    n2 <- n - n1 # sample size per group in stage II

    # ----------------------
    # Stage I
    # ----------------------

    Delta_1 = mvrnorm(n=m, mu=c(delta0, delta1), Sigma = matrix(c(2/n1, 2/
        n1, 2/n1, 2/(n1*p)), 2, 2))

    # subgroup
    Z1_1 = Delta_1[,2] * sqrt(n1*p/2)
    p1_1 = 1 - pnorm(Z1_1)
```

```
# total population
Z0_1 = Delta_1[,1] * sqrt(n1/2)
p0_1 = 1 - pnorm(Z0_1)

#  combination of p-values (Closed Testing Procedure)
p01_1 = apply(cbind(2 * apply(cbind(p0_1, p1_1), 1, min), apply(cbind(
    p0_1, p1_1), 1, max)), 1, min)


# ----------------------
# Stage II
# ----------------------

# interim decision:
selSub <- ifelse(Delta_1[,2] > c1 & Delta_1[,1] <= c0, 1, 0)
selTot <- ifelse(Delta_1[,2] <= c1 & Delta_1[,1] > c0, 1, 0)
selCop <- ifelse(Delta_1[,2] > c1 & Delta_1[,1] > c0, 1, 0)
selFut <- ifelse(Delta_1[,2] <= c1 & Delta_1[,1] <= c0, 1, 0)


# if subgroup is selected
delta1_2 <- rnorm(n = m, mean = delta1, sd = sqrt(2/n2))

Z1_2 <- delta1_2 * sqrt(n2/2)
p1_2 <- 1 - pnorm(Z1_2)

U1 <- sqrt(p*n1/(p*n1+n2))*qnorm(1-p1_1) + sqrt(n2/(n1*p+n2))*qnorm(1-
    p1_2)
U01 <- sqrt(n1/n) * qnorm(1-p01_1) + sqrt(n2/n) * qnorm(1-p1_2) #
    closed testing

if(adj == "ClosedTesting"){
  sigSub <- U01 > z_alpha & U1 > z_alpha
}

if(adj == "Bonferroni"){
  sigSub <- 1 - pnorm(U1) < alpha/4
}


# if total population is selected
Delta_2 <- mvrnorm(n = m, mu = c(delta0, delta1), Sigma = matrix(c(2/
    n2, 2/n2, 2/n2, 2/(n2*p)), 2, 2))

Z0_2 <- Delta_2[,1] * sqrt(n2/2)
p0_2 <- 1 - pnorm(Z0_2)

U0 <- sqrt(n1/n) * Z0_1 + sqrt(n2/n) * Z0_2
U01 <- sqrt(n1/n) * qnorm(1-p01_1) + sqrt(n2/n) * Z0_2

if(adj == "ClosedTesting"){
  sigTot <- U01 > z_alpha & U0 > z_alpha
}

if(adj == "Bonferroni"){
  sigTot <- 1 - pnorm(U0) < alpha/4
}
```

```
  # if co-primary analysis is selected
  # subgroup
  Z1_2 <- Delta_2[,2] * sqrt(n2*p/2)
  p1_2 <- 1 - pnorm(Z1_2)

  # combination of p-values (Closed Testing Procedure)
  p01_2 <- apply(cbind(2 * apply(cbind(p0_2, p1_2), 1, min), apply(cbind
      (p0_2, p1_2), 1, max)), 1, min)

  U0 <- sqrt(n1/n) * Z0_1 + sqrt(n2/n) * Z0_2
  U1 <- sqrt(n1/n) * Z1_1 + sqrt(n2/n) * Z1_2
  U01 <- sqrt(n1/n) * qnorm(1-p01_1) + sqrt(n2/n) * qnorm(1-p01_2)

  if(adj == "ClosedTesting"){
    sigTot_cp <- U01 > z_alpha & U0 > z_alpha
    sigSub_cp <- U01 > z_alpha & U1 > z_alpha
  }

  if(adj == "Bonferroni"){
    sigTot_cp <- 1 - pnorm(U0) < alpha/4
    sigSub_cp <- 1 - pnorm(U1) < alpha/4
  }


  # Selection Probabilities
  prSub[i] <- sum(selSub, na.rm = TRUE) / m
  prTot[i] <- sum(selTot, na.rm = TRUE) / m
  prCop[i] <- sum(selCop, na.rm = TRUE) / m
  prFut[i] <- sum(selFut, na.rm =TRUE) / m

  # Power
  # unconditionally
  powerSub[i] <- sum(sigSub * selSub | sigSub_cp * selCop, na.rm = TRUE)
      / m
  powerTot[i] <- sum(sigTot * selTot | sigTot_cp * selCop, na.rm=TRUE) /
      m
  powerOverall[i] <- sum(sigTot * selTot | sigSub * selSub | sigSub_cp *
      selCop |
                           sigTot_cp * selCop, na.rm = TRUE) / m

  # if co-primary analysis is selected
  powerBoth_cp[i] = sum(selCop * sigSub_cp * sigTot_cp, na.rm = TRUE) /
      m
  powerSub_cp[i] = sum(selCop * sigSub_cp, na.rm = TRUE) / m
  powerTot_cp[i] = sum(selCop * sigTot_cp, na.rm = TRUE) / m
  powerOverall_cp[i] = sum(selCop * (sigTot_cp | sigSub_cp), na.rm =
      TRUE) / m

  # if subgroup only is selected
  powerSub_Sub[i] = sum(sigSub * selSub, na.rm = TRUE) / m

  # if total population is selected
  powerTot_Tot[i] = sum(sigTot * selTot, na.rm=TRUE) / m

}
```

```
  return(list(prSub = prSub, prTot = prTot, prCop = prCop, prFut = prFut,
              powerSub = powerSub, powerTot = powerTot, powerOverall =
                 powerOverall,
              powerBoth_cp = powerBoth_cp, powerSub_cp = powerSub_cp,
                 powerTot_cp = powerTot_cp, powerOverall_cp = powerOverall
                 _cp,
              powerSub_Sub = powerSub_Sub, powerTot_Tot = powerTot_Tot, n
                 = n))
}



#########################################################################
# Function to calculate sample size to assure a specific power
# at a specific timing t
#########################################################################

#########################################################################
# Parameters to specify:
# p = prevalence of subgroup
# delta1 = standardized effect in subgroup
# delta2 = standardized effect in complement
# m = number of simulated studies
# c1, c0 = threshold values for selection rule
# t = interim analysis timing
# adj = method to control for multiplicity, either "ClosedTesting" or
#       "Bonferroni"
# alpha = significance level
# seed = seed value
#########################################################################

#########################################################################
# Returned value:
# n = sample size per group
#########################################################################

calculate_n_Sel2 <- function(p, delta1, delta2, m, c1, c0, t, adj, alpha,
   seed){
  set.seed(seed)

  n_start <- 0
  powerOverall <- 0

  while(powerOverall < 0.8){
    n_start <- n_start + 50
    res <- Power_Sel2(p, n_start, delta1, delta2, t, m, c1, c0, adj, alpha
       , seed)
    powerOverall <- res$powerOverall
  }

  n_start2 <- n_start - 50
  powerOverall <- 0

  while(powerOverall < 0.8){
    n_start2 <- n_start2 + 1
    res <- Power_Sel2(p, n_start2, delta1, delta2, t, m, c1, c0, adj,
       alpha, seed)
    powerOverall <- res$powerOverall
```

```
  }
  n <- n_start2
  return(n)
}



#####################################################################
# Function to plot graphic showing power for different interim analysis
# timings and effect sizes in the complementary group
#####################################################################

#####################################################################
# Parameters to specify:
# p = prevalence of subgroup
# delta1 = standardized effect in subgroup
# delta2v = vector of standardized effects in complement
# tv = vector of interim analysis timings
# m = number of simulated studies
# c = threshold of selection rule (vector (c1, c0) if sel_rule = 2)
# adj = method to control for multiplicity, either "ClosedTesting" or
#       "Bonferroni"
# alpha = significance level
# seed = seed value
# sel_rule = 1 for selection rule based on estimated effect differences,
#          = 2 for selection rule based on absolute effect estimates
#####################################################################

Plot_Power <- function(p, delta1, delta2v, tv, m, c, adj, alpha, seed, sel
    _rule){

  power_m <- matrix(rep(NA, length(tv) * length(delta2v)), length(delta2v)
      , length(tv))
  diff_03_07 <- rep(NA, length(delta2v))
  ma_vec <- rep(NA, length(delta2v))
  mi_vec <- rep(NA, length(delta2v))
  s <- 0

  for(delta2 in delta2v){

    s <- s+1
    delta0 <- p*delta1 + (1-p)*delta2

    if(sel_rule == 1){
      n <- calculate_n_Sel1(p, delta1, delta2, m, c, t, adj, alpha, seed)
      res <- Power_Sel1(p, n, delta1, delta2, tv, m, c, adj, alpha, seed)
    }

    if(sel_rule == 2){
      n <- calculate_n_Sel2(p, delta1, delta2, m, c[1], c[2], t, adj,
          alpha, seed)
      res <- Power_Sel2(p, n, delta1, delta2, tv, m, c[1], c[2], adj,
          alpha, seed)
    }

    i30 <- which(tv == 0.3)
    i70 <- which(tv == 0.7)
    t30_70 <- tv[i30:i70]
```

```
  mi <- min(res$powerOverall[i30:i70])
  ma <- max(res$powerOverall[i30:i70])
  mi_vec[s] <- min(res$powerOverall[i30:i70])-0.8
  ma_vec[s] <- max(res$powerOverall[i30:i70])-0.8
  t_mi <- t30_70[which(res$powerOverall[i30:i70] == min(res$powerOverall
      [i30:i70]))]
  t_ma <- t30_70[which(res$powerOverall[i30:i70] == max(res$powerOverall
      [i30:i70]))]
  diff_03_07[s] <- round(ma-mi,4)
  power_m[s,] <- res$powerOverall

}

col <- c("red4", "red3", "red", rgb(1,0.4,0,1), rgb(1,0.6, 0,1), rgb
    (1,0.8, 0,1),
          rgb(1, 0.9, 0,1), rgb(1,1,0,0.4), rgb(1,1,0,0.1))

power_col_1 <- ifelse(power_m < 0.73, col[9], 0)
power_col_2 <- ifelse(power_m < 0.75 & power_m >= 0.73, col[8], power_
    col_1)
power_col_3 <- ifelse(power_m < 0.77 & power_m >= 0.75, col[7], power_
    col_2)
power_col_4 <- ifelse(power_m < 0.79 & power_m >= 0.77, col[6], power_
    col_3)
power_col_5 <- ifelse(power_m < 0.81 & power_m >= 0.79, col[5], power_
    col_4)
power_col_6 <- ifelse(power_m < 0.83 & power_m >= 0.81, col[4], power_
    col_5)
power_col_7 <- ifelse(power_m < 0.85 & power_m >= 0.83, col[3], power_
    col_6)
power_col_8 <- ifelse(power_m < 0.87 & power_m >= 0.85, col[2], power_
    col_7)
power_col_9 <- ifelse(power_m >= 0.87, col[1], power_col_8)

op <- par(mar = c(5,6,0,2) + 0.1)
plot(0.5, 1, xlim = c(0,1.3), ylim = c(0.3,length(delta2v)+1.6),
     xlab = "t␣␣␣␣␣␣␣␣␣␣␣␣", ylab = "", yaxt = "n", xaxt = "n", las =
         1,
     cex.lab = 2, cex.main = 2)
title(ylab = bquote(Delta["-"]), cex.lab = 2, line = 4.5)
axis(2, at = 1:length(delta2v), labels = delta2v, las = 1, cex.axis =
    1.5)
axis(1, at = c(0, 0.2, 0.4, 0.6, 0.8, 1), labels = c("0.0", "0.2", "0.4"
    , "0.6", "0.8", "1.0"), las = 1, cex.axis = 1.5)
for(j in 1:length(delta2v)){
  for(i in 1:length(tv)){
    lines(c(tv[i] - 0.0125, tv[i] + 0.0125), c(j,j), col = power_col_9[j
        ,i],
          lwd = 20, lend = "butt")
  }
}
text(rep(1.2,length(delta2v)), 1:length(delta2v), paste(format(round(
    diff_03_07*100,1), nsmall = 1), "%"),pos = 2, cex = 1.5)
text(0.95, length(delta2v) + 1.6, "power␣range", pos = 4, cex = 1.4)
text(0.95, length(delta2v) + 0.8, expression(paste("for␣", t %in% "[0.3,
    ␣0.7]")), pos = 4, cex = 1.4)
```

```
}



#######################################################################
# Function to plot graphic showing selection and rejection probabilities
# for different interim analysis timings
#######################################################################

#######################################################################
# Parameters to specify:
# p = prevalence of subgroup
# delta1 = standardized effect in subgroup
# delta2v = vector of standardized effects in complement
# tv = vector of interim analysis timings
# m = number of simulated studies
# c = threshold of selection rule (vector (c1, c0) if sel_rule = 2)
# adj = method to control for multiplicity, either "ClosedTesting" or
#       "Bonferroni"
# alpha = significance level
# seed = seed value
# sel_rule = 1 for selection rule based on estimated effect differences,
#          = 2 for selection rule based on absolute effect estimates
#######################################################################

Plot_SelProb <- function(pv, delta1, delta2v, tv, m, c, adj, alpha, seed,
    sel_rule){

  p_long <- as.numeric()
  delta2_long <- as.numeric()
  tv_long <- as.numeric()

  prSub <- as.numeric()
  prTot <- as.numeric()
  prCop <- as.numeric()
  prFut <- as.numeric()
  powerSub <- as.numeric()
  powerTot <- as.numeric()
  powerOverall <- as.numeric()
  powerSub_Sub <- as.numeric()
  powerTot_Tot <- as.numeric()
  powerOverall <- as.numeric()
  powerOverall_cp <- as.numeric()

  if(sel_rule == 1){

    for(p in pv){
      for(delta2 in delta2v){

        p_long <- c(p_long, rep(p, length(tv)))
        delta2_long <- c(delta2_long, rep(delta2, length(tv)))
        tv_long <- c(tv_long, tv)

        n <- calculate_n_Sel1(p, delta1, delta2, m, c, 0.5, adj, alpha,
            seed)
        res <- Power_Sel1(p, n, delta1, delta2, tv, m, c, adj, alpha, seed
            )
```

```
      prSub <- c(prSub, res$prSub)
      powerSub <- c(powerSub, res$powerSub)
      powerTot <- c(powerTot, res$powerTot)
      powerOverall <- c(powerOverall, res$powerOverall)

    }
  }

  res_tab <- data.frame(prSub = prSub/m, prTot = 1 - prSub/m, powerSub =
      powerSub, powerTot = powerTot,
                        powerOverall = powerOverall, p_long = p_long,
                           delta2_long = delta2_long, tv_long = tv_long)
  res_long <- gather(res_tab, var, prob, prSub:powerOverall, factor_key
     = TRUE)

  coltyp <- brewer.pal(5, "Set1")
  coltyp <- coltyp[c(1, 2, 1, 2, 3)]
  ltyp <- c(2, 2, 1, 1, 1)
}

if(sel_rule == 2){

  for(p in pv){
    for(delta2 in delta2v){

      p_long <- c(p_long, rep(p, length(tv)))
      delta2_long <- c(delta2_long, rep(delta2, length(tv)))
      tv_long <- c(tv_long, tv)

      n <- calculate_n_Sel2(p, delta1, delta2, m, c[1], c[2], 0.5, adj,
         alpha, seed)
      res <- Power_Sel2(p, n, delta1, delta2, tv, m, c[1], c[2], adj,
         alpha, seed)

      prSub <- c(prSub, res$prSub)
      prTot <- c(prTot, res$prTot)
      prCop <- c(prCop, res$prCop)
      prFut <- c(prFut, res$prFut)
      powerSub_Sub <- c(powerSub_Sub, res$powerSub_Sub)
      powerTot_Tot <- c(powerTot_Tot, res$powerTot_Tot)
      powerOverall_cp <- c(powerOverall_cp, res$powerOverall_cp)
      powerOverall <- c(powerOverall, res$powerOverall)

    }
  }

  res_tab <- data.frame(prSub = prSub, prTot = prTot, prCop = prCop,
     prFut = prFut,
      powerSub_Sub = powerSub_Sub, powerTot_Tot = powerTot_Tot,
         powerOverall_cp = powerOverall_cp, powerOverall = powerOverall,
      p_long = p_long, delta2_long = delta2_long, tv_long = tv_long)
  res_long <- gather(res_tab, var, prob, prSub:powerOverall, factor_key
     = TRUE)

  coltyp <- brewer.pal(5, "Set1")
  coltyp <- coltyp[c(1, 2, 4, 5, 1, 2, 4, 3)]
  ltyp <- c(2, 2, 2, 2, 1, 1, 1, 1)
}
```

```
  p.labs <- paste("p = ", pv, sep = "")
  names(p.labs) <- pv
  delta2.labs <- as.character()
  for(i in 1:length(delta2v)){
    delta2.labs <- c(delta2.labs, bquote(Delta["-"]*plain(" = ")*.(delta2v
        [i])))
  }
  names(delta2.labs) <- delta2v

  res_long %>%
    ggplot(aes(x = tv_long, y = prob, colour = var, linetype = var, group
        = var)) +
    geom_line(size=0.6) +
    facet_grid(p_long ~ delta2_long, labeller = label_bquote(cols = Delta[
        "-"] ~ "=" ~ .(delta2_long), rows = p~"="~ .(p_long))) +
    scale_linetype_manual(values = ltyp) +
    scale_colour_manual(values = coltyp) +
    xlab("t") +
    ylab("Probability") +
    ylim(c(0, 1)) +
    theme_bw() +
    theme(legend.position = "none", axis.text = element_text(size = 11),
        axis.title = element_text(size = 13),
           strip.text.x = element_text(size = 11), strip.text.y = element_
               text(size = 11))

}
```

C.2 Code for Design with Sample Size Reassessment

```
library(MASS)
library(ggplot2)
library(RColorBrewer)

#######################################################################
# Function to calculate sample size distribution
# for selection rule based on estimated effect differences
#######################################################################

#######################################################################
# Parameters to specify:
# p = prevalence of subgroup
# t = interim analysis timing
# delta1 = standardized effect in subgroup
# delta2 = standardized effect in complement
# m = number of simulated studies
# c = threshold for selection rule
# beta = type II error rate
# alpha = significance level
# CP = conditional power for effect size from planning phase (="pp")
#      or mean between effect size from planning phase and observed effect
#      from interim analysis (="ppia")
# adj = method to control for multiplicity, either "ClosedTesting"
```

```
#       or "Bonferroni"
# seed = seed value
# H0 = TRUE if effects in both populations are 0 (delta1 and delta2 have
#      to be specified then as the assumed effects in the planning phase
#      to calculate n_fix
#####################################################################

#####################################################################
# Returned values:
# 1. data frame: for each simulated study:
## selSub = 1 if subgroup is selected, 0 otherwise
## sig_all = 1 if significant result, 0 otherwise
## n2 = sample size in second stage
## n = overall sample size
# 2. data frame: input parameters plus
# n_fix = sample size in fixed design
# conP = adjusted conditional power to reach an overall power of 80%
# sel_rule = 1 (selection rule)
#####################################################################

Samplesize_Sel1 <- function(p, t, delta1, delta2, m, c, beta, alpha, CP,
   adj, seed, H0 = FALSE){

  set.seed(seed)

  delta0 <- p*delta1 + (1-p)*delta2
  n_fix <- ceiling(2 * (qnorm(1-alpha/4) + qnorm(1-beta))^2 / delta0^2) #
     sample size per group in a fixed design

  if(H0 == TRUE) {
    delta1 <- 0
    delta2 <- 0
    delta0 <- 0
  }

  n1 <- t*n_fix  # sample size stage I

  # ----------------------
  # Stage I
  # ----------------------

  Delta_1 <- mvrnorm(n = m, mu = c(delta0, delta1), Sigma = matrix(c(2/n1,
      2/n1, 2/n1, 2/(n1*p)),2,2))

  # subgroup
  Z1_1 <- Delta_1[,2] * sqrt(n1*p/2)
  p1_1 <- 1 - pnorm(Z1_1)

  # total population
  Z0_1 <- Delta_1[,1] * sqrt(n1/2)
  p0_1 <- 1 - pnorm(Z0_1)

  # combination of p-values (Closed Testing Procedure)
  p01_1 <- apply(rbind(2 * apply(rbind(p1_1, p0_1), 2, min),
                       apply(rbind(p1_1, p0_1), 2, max)), 2, min)

  selSub <- ifelse(Delta_1[,1] + c < Delta_1[,2], 1, 0)
```

```
# ------------------------
# Stage II
# ------------------------

if(CP == "pp"){  # conditional power using the effect size from the
    planning stage
  mu0 <- rep(delta0, m)
  mu1 <- rep(delta1, m)
}

if(CP == "ppia"){  # conditional power using the mean of the effect size
    from the planning stage and the observed interim effect
  mu0 <- (rep(delta0, m) + Delta_1[,1]) / 2
  mu1 <- (rep(delta1, m) + Delta_1[,2]) / 2
}


poweri <- 0
beta_CP_v <- (100:1)/100

if(H0 == TRUE) {
  beta_CP_v <- beta
}

for(beta_CP in beta_CP_v){

  # if subgroup is selected
  # sample size for stage II
  br <- (qnorm(1-alpha/4) /sqrt(1-t) - qnorm(beta_CP) - sqrt(t/(1-t)) *
      Z1_1)
  br <- ifelse(br < 0, 0, br)
  fz <- cbind(br^2 * 2 / mu1^2, 2*n_fix - n1)
  n2_sub <- apply(fz, 1, min)
  n2_sub <- ifelse(n2_sub < 10, 10, n2_sub)

  # tests in stage II
  delta1_2 <- rnorm(n = m, mean = delta1, sd = sqrt(2/n2_sub))
  Z1_2_sub <- delta1_2 * sqrt(n2_sub/2)
  U1_sub <- sqrt(t)*Z1_1 + sqrt(1-t)*Z1_2_sub

  # combination of p-values (Closed Testing Procedure)
  U01_sub <- sqrt(t) * qnorm(1-p01_1) + sqrt(1-t)*Z1_2_sub
  p01_sub <- 1 - pnorm(U01_sub)

  if(adj == "ClosedTesting"){
    sig_sub <- ifelse(p01_sub < alpha/2 & 1 - pnorm(U1_sub) < alpha/2,
        1, 0)
  }
  if(adj == "Bonferroni"){
    sig_sub <- ifelse(1 - pnorm(U1_sub) < alpha/4, 1, 0)
  }


  # if total population is selected

  # sample size for stage II
  br <- (qnorm(1-alpha/4) /sqrt(1-t) - qnorm(beta_CP) - sqrt(t/(1-t)) *
      Z0_1)
```

```
    br <- ifelse(br < 0, 0, br)
    fz <- cbind(br^2 * 2 / mu0^2, 2*n_fix - n1)
    n2_tot <- apply(fz, 1, min)
    n2_tot <- ifelse(n2_tot < 10, 10, n2_tot)

    # tests in stage II
    delta0_2 <- rnorm(n = m, mean = delta0, sd = sqrt(2/n2_tot))
    Z0_2_tot <- delta0_2 * sqrt(n2_tot/2)
    U0_tot <- sqrt(t)*Z0_1 + sqrt(1-t)*Z0_2_tot

    # combination of p-values (Closed Testing Procedure)
    U01_tot <- sqrt(t)*qnorm(1-p01_1) + sqrt(1-t)*Z0_2_tot
    p01_tot <- 1 - pnorm(U01_tot)

    if(adj == "ClosedTesting"){
      sig_tot = ifelse(p01_tot < alpha/2 & 1 - pnorm(U0_tot) < alpha/2, 1,
          0)
    }
    if(adj == "Bonferroni"){
      sig_tot = ifelse(1 - pnorm(U0_tot) < alpha/4, 1, 0)
    }

    # sample size
    n2_v <- round(ifelse(selSub == 1, n2_sub, n2_tot))
    n_v <- round(n1 + n2_v)

    # power
    sig_all <- ifelse(selSub == 1, sig_sub, sig_tot)
    poweri <- mean(sig_all)
    if (poweri >= 0.8) break
  }

  return(list(data.frame(selSub = selSub, sig_all = sig_all, n2 = n2_v, n
      = n_v),
    data.frame(p = p, t = t, delta1 = delta1, delta2 = delta2, m = m, c =
        c,
    beta = beta, alpha = alpha, CP = CP, n_fix = n_fix, conP = 1-beta_CP,
        sel_rule = 1)))

}



######################################################################
# Function to calculate sample size distribution
# for selection rule based on absolute effect estimates
######################################################################

######################################################################
# Parameters to specify:
# p = prevalence of subgroup
# t = interim analysis timing
# delta1 = standardized effect in subgroup
# delta2 = standardized effect in complement
# m = number of simulated studies
# c = threshold for selection rule
# beta = type II error to calculate sample size of fixed design
# alpha = significance level
```

```
# CP = conditional power for effect size from planning phase (="pp")
#      or mean between effect size from planning phase and observed
#      effect from interim analysis (="ppia")
# adj = method to control for multiplicity, either "ClosedTesting"
#       or "Bonferroni"
# seed = seed value
# H0 = TRUE if effects in both populations are 0 (delta1 and delta2 have
#      to be specified then as the assumed effects in the planning phase
#      to calculate n_fix
#####################################################################

#####################################################################
# Returned values:
# 1. data frame: for each simulated study:
## selSub = 1 if subgroup is selected, 0 otherwise
## sig_all = 1 if significant result, 0 otherwise
## n2 = sample size in second stage
## n = overall sample size
# 2. data frame: input parameters plus
# n_fix = sample size in fixed design
# conP = adjusted conditional power to reach an overall power of 80%
# sel_rule = 2 (selection rule)
#####################################################################

Samplesize_Sel2 <- function(p, t, delta1, delta2, m, c0, c1, beta, alpha,
   CP, adj, seed, H0 = FALSE){

  set.seed(seed)

  delta0 <- p*delta1 + (1-p)*delta2
  n_fix <- ceiling(2*(qnorm(1-alpha/4) + qnorm(1-beta))^2 / delta0^2)

  if(H0 == TRUE) {
    delta1 <- 0
    delta2 <- 0
    delta0 <- 0
  }

  n1 <- t * n_fix

  # ----------------------
  # Stage I
  # ----------------------

  Delta_1 <- mvrnorm(n = m, mu = c(delta0, delta1), Sigma = matrix(c(2/n1,
      2/n1, 2/n1, 2/(n1*p)), 2, 2))

  # subgroup
  Z1_1 <- Delta_1[,2] * sqrt(n1*p/2)
  p1_1 <- 1 - pnorm(Z1_1)

  # total population
  Z0_1 <- Delta_1[,1] * sqrt(n1/2)
  p0_1 <- 1 - pnorm(Z0_1)

  # combination of p-values (Closed Testing Procedure)
  p01_1 <- apply(rbind(2*apply(rbind(p1_1, p0_1), 2, min),
                       apply(rbind(p1_1, p0_1), 2, max)), 2, min)
```

```
# ----------------------
# Stage II
# ----------------------

# interim decision:
selSub <- ifelse(Delta_1[,2] >  c1 & Delta_1[,1] <= c0, 1, 0)
selTot <- ifelse(Delta_1[,2] <= c1 & Delta_1[,1] >  c0, 1, 0)
selCop <- ifelse(Delta_1[,2] >  c1 & Delta_1[,1] >  c0, 1, 0)
selFut <- ifelse(Delta_1[,2] <= c1 & Delta_1[,1] <= c0, 1, 0)

if(CP == "pp"){
  mu0 <- rep(delta0, m)
  mu1 <- rep(delta1, m)
}

if(CP == "ppia"){
  mu0 <- (rep(delta0, m) + Delta_1[,1])/2
  mu1 <- (rep(delta1, m) + Delta_1[,2])/2
}

poweri = 0
beta_CP_v = (100:1)/100

if(H0 == TRUE) {
  beta_CP_v <- beta
}

for(beta_CP in beta_CP_v){

  # if subgroup is selected
  # sample size for stage II
  br <- (qnorm(1-alpha/4) /sqrt(1-t) - qnorm(beta_CP) - sqrt(t/(1-t)) *
      Z1_1)
  br <- ifelse(br < 0, 0, br)
  fz <- cbind(br^2 * 2/mu1^2, 2*n_fix - n1)
  n2_sub <- apply(fz, 1, min)
  n2_sub <- ifelse(n2_sub < 10, 10, n2_sub)

  # tests in stage II
  delta1_2 <- rnorm(n = m, mean = delta1, sd = sqrt(2/n2_sub))
  Z1_2_sub <- delta1_2 * sqrt(n2_sub/2)
  U1_sub <- sqrt(t)*Z1_1 + sqrt(1-t)*Z1_2_sub

  # combination of p-values (Closed Testing Procedure)
  U01_sub <- sqrt(t)*qnorm(1-p01_1) + sqrt(1-t)*Z1_2_sub
  p01_sub <- 1 - pnorm(U01_sub)

  if(adj == "ClosedTesting"){
    sig_sub <- ifelse(p01_sub < alpha/2 & 1 - pnorm(U1_sub) < alpha/2,
        1, 0)
  }

  if(adj == "Bonferroni"){
    sig_sub <- ifelse(1 - pnorm(U1_sub) < alpha/4, 1, 0)
  }
```

```
# if total population is selected

# sample size in stage II
br <- (qnorm(1 - alpha/4) / sqrt(1-t) - qnorm(beta_CP) - sqrt(t/(1-t))
    *  Z0_1)
br <- ifelse(br < 0, 0, br)
fz_tot <- cbind(br^2 * 2/mu0^2, 2*n_fix - n1)
n2_tot <- apply(fz_tot, 1, min)
n2_tot <- ifelse(n2_tot < 10, 10, n2_tot)

# tests in stage II
delta0_2 <- rnorm(n = m, mean = delta0, sd = sqrt(2/n2_tot))
Z0_2_tot <- delta0_2 * sqrt(n2_tot/2)
U0_tot <- sqrt(t)*Z0_1 + sqrt(1-t)*Z0_2_tot

# combination of p-values (Closed Testing Procedure)
U01_tot <- sqrt(t)*qnorm(1-p01_1) + sqrt(1-t)*Z0_2_tot
p01_tot <- 1 - pnorm(U01_tot)

if(adj == "ClosedTesting"){
  sig_tot <- ifelse(p01_tot < alpha/2 & 1 - pnorm(U0_tot) < alpha/2,
     1, 0)
}

if(adj == "Bonferroni"){
  sig_tot <- ifelse(1 - pnorm(U0_tot) < alpha/4, 1, 0)
}


# if co-primary analysis is selected

# sample size in stage II
n2_cop <- cbind(ifelse(n2_tot*p > n2_sub, n2_tot, n2_sub/p), 2*n_fix -
    n1)
n2_cop <- apply(n2_cop, 1, min)

# tests in stage II
Delta_2 <- matrix(rep(NA, m*2), m, 2)
for(j in 1:m){
  Delta_2[j,] <- mvrnorm(n = 1, mu = c(delta0, delta1), Sigma= matrix(
     c(2/n2_cop[j], 2/n2_cop[j], 2/n2_cop[j], 2/(n2_cop[j]*p)), 2, 2))
}

# subgroup
Z1_2 <- Delta_2[,2] * sqrt(n2_cop*p/2)
p1_2 <- 1 - pnorm(Z1_2)
U1 <- sqrt(t)*Z1_1 + sqrt(1-t)*Z1_2
p1 <- 1 - pnorm(U1)

# total population
Z0_2 <- Delta_2[,1] * sqrt(n2_cop/2)
p0_2 <- 1 - pnorm(Z0_2)
U0 <- sqrt(t)*Z0_1 + sqrt(1-t)*Z0_2
p0 <- 1 - pnorm(U0)

# combination of p-values (Closed Testing Procedure)
p01_2 = apply(cbind(2*apply(cbind(p0_2, p1_2), 1, min), apply(cbind(p0
```

```
      _2, p1_2), 1, max)), 1, min)
    U01 = sqrt(t)*qnorm(1-p01_1) + sqrt(1-t)*qnorm(1-p01_2)

    if(adj == "ClosedTesting"){
      sig_cop = ifelse((U01 > qnorm(1-alpha/2) & U0 > qnorm(1-alpha/2)) |
         (U01 > qnorm(1-alpha/2) & U1 > qnorm(1-alpha/2)), 1, 0)
    }
    if(adj == "Bonferroni"){
      sig_cop = ifelse(p0 < alpha/4 | p1 < alpha/4, 1, 0)
    }

    # sample size
    n2_v <- ifelse(selSub == 1, n2_sub, n2_tot)
    n2_v <- ifelse(selCop == 1, n2_cop, n2_v)
    n2_v <- round(ifelse(selFut == 1, 0, n2_v))
    n_v <- round(n1 + n2_v)


    # Signifikanz
    sig_all <- ifelse(selSub == 1, sig_sub, sig_tot)
    sig_all <- ifelse(selCop == 1, sig_cop, sig_all)
    sig_all <- ifelse(selFut == 1, 0, sig_all)
    poweri <- mean(sig_all)
    if (poweri >= 0.8) break
  }

  return(list(data.frame(selSub = selSub, selTot = selTot, selCop = selCop
     , selFut = selFut,
                         sig_all = sig_all, n2 = n2_v, n = n_v),
              data.frame(p = p, t = t, delta1 = delta1, delta2 = delta2, m
                  = m, c0 = c0, c1 = c1,
                         beta = beta, alpha = alpha, CP = CP, n_fix = n_
                            fix, conP = 1 - beta_CP, sel_rule = 2)))

}



#########################################################################
# Function that saves results of Samplesize_Sel1 / Samplesize_Sel2
# for different t in a matrix
#########################################################################

#########################################################################
# Parameters to specify:
# same as for amplesize_Sel1 / Samplesize_Sel2, but:
# t = vector of interim analysis timing
#########################################################################


Samplesize_Sel1_t <- function(p, t, delta1, delta2, m, c, beta, alpha, CP,
    adj, seed){
  sim <- as.numeric()

  for(ti in t) {
    sim <- rbind(sim, Samplesize_Sel1(p, ti, delta1, delta2, m, c, beta,
       alpha, CP, adj, seed))
  }
```

```
  sim
}


Samplesize_Sel2_t  <- function(p, t, delta1, delta2, m, c0, c1, beta,
   alpha, CP, adj, seed){
  sim <- as.numeric()
  for(ti in t) {
    sim <- rbind(sim, Samplesize_Sel2(p, ti, delta1, delta2, m, c0, c1,
        beta, alpha, CP, adj, seed))
  }
  sim
}



#########################################################################
# Function to plot sample size distribution (histogram plus boxplot)
# for both selection rules
#########################################################################

#########################################################################
# Parameters to specify:
# sim = simulation results received from function Samplesize_Sel1_t or
#       Samplesize_Sel2_t
#########################################################################

Plot_Samplesize  <- function(sim){

  simm <- as.numeric()
  for(i in 1:dim(sim)[1]){
    simm <- rbind(simm, cbind(sim[[i,1]], t = sim[[i,2]]$t))
  }

  sel_rule <- sim[[1,2]]$sel_rule

  if(sel_rule == 1){
    selection <- as.factor(simm$selSub)
    selection <- relevel(selection, "1")
    col <- brewer.pal(5,"Set1")[1:2]
  }

  if(sel_rule == 2){
    selection <- rep(0, dim(simm)[1])
    selection <- ifelse(simm$selSub == 1, 1, selection)
    selection <- ifelse(simm$selTot == 1, 2, selection)
    selection <- ifelse(simm$selCop == 1, 3, selection)
    selection <- as.factor(selection)
    col <- brewer.pal(5,"Set1")
    col2 <- brewer.pal(6,"Set2")
    col <- c(col2[6], col[1:3])
  }

  n_fix <- as.numeric(sim[[1,2]]["n_fix"])
  m <- sim[[1,2]]$m

  P <- ggplot(simm, aes(x = simm$n, fill = selection)) +
    geom_histogram(position = "stack", binwidth = round(n_fix/100)*2,
```

```
      alpha = 0.7) +
  facet_grid(simm$t ~ ., scales = "free") +
  xlab("n") +
  ylab("") +
  theme_bw() +
  scale_y_continuous(breaks = NULL) +
  theme(legend.position = "none", text = element_text(size = 25)) +
  scale_fill_manual(values = col)

# draw boxplot
means <- as.numeric()
medians <- as.numeric()
b <- boxplot(simm$n ~ simm$t, plot = FALSE)
box_stat <- b$stats
mh <- as.numeric()
wu <- as.numeric()
wo <- as.numeric()
bu <- as.numeric()
bo <- as.numeric()
bb <- as.numeric()
m_middle <- as.numeric()
m_up <- as.numeric()
m_low <- as.numeric()
r <- 0.3 # width of the boxplot in relation to plot hight

panel <- 0
for (k in levels(as.factor(simm$t))){
  panel <- panel + 1

  # mean
  me <- rep(mean(simm$n[which(simm$t == k)]), m)
  means <- c(means,me)

  # median
  md <- rep(median(simm$n[which(simm$t == k)]), m)
  medians <- c(medians,md)

  # box width
  maxh <- max(ggplot_build(P)$data[[1]]$ymax[ggplot_build(P)$data[[1]]$
      PANEL == panel])
  mh_ <- rep(maxh*r, m)
  mh_[1] <- maxh*(1-r)
  mh <- c(mh,mh_)

  # middle of box
  m_middle_ <- rep(maxh * 0.5, m)
  m_middle <- c(m_middle, m_middle_)

  # upper and lower end of box
  m_up_ <- rep(maxh * (1-r), m)
  m_up <- c(m_up, m_up_)
  m_low_ <- rep(maxh * r, m)
  m_low <- c(m_low, m_low_)

  # whisker
  wu_ <- box_stat[1, panel]
  wu_ <- rep(wu_, m)
  wu_[1] <- box_stat[2, panel]
```

```
    wu <- c(wu,wu_)

    wo_ <- box_stat[4, panel]
    wo_ <- rep(wo_, m)
    wo_[1] <- box_stat[5, panel]
    wo <- c(wo,wo_)

    # box height
    bb_ <- box_stat[2, panel]
    bb_ <- rep(bb_, m)
    bb_[1] <- box_stat[4, panel]
    bb <- c(bb,bb_)

    # box (first and third quartile)
    bu_ <- rep(box_stat[2, panel], m)
    bu <- c(bu, bu_)
    bo_ <- rep(box_stat[4, panel], m)
    bo <- c(bo, bo_)

  }

  P <- P + geom_point(aes(x=means, y = m_middle), size = 2, shape = 19) +
    geom_line(aes(x = medians, y = mh), size = 1.5) +
    geom_line(aes(x = wu, y = m_middle), size = 1, lineend = "square") +
    geom_line(aes(x = wo, y = m_middle), size = 1, lineend = "square") +
    geom_line(aes(x = bb, y = m_up), size = 1, lineend = "square") +
    geom_line(aes(x = bb, y = m_low), size = 1, lineend = "square") +
    geom_line(aes(x = bu, y = mh), size = 1, lineend = "square") +
    geom_line(aes(x = bo, y = mh), size = 1, lineend = "square")

  plot(P)

}
```

Own Publications

Partial Results of this thesis were published in the following publication:

Benner, L. and Kieser, M. (2018), 'Timing of the interim analysis in adaptive enrichment designs', *Journal of Biopharmaceutical Statistics* **28**(4), 622-632.

The main results described in Section 4.2 originate from this publication. Consequently, also the described design and testing procedures described in Chapter 3 as well as some content of the introduction and discussion was described in this publication. My contribution to this publication was programming the simulation study, creating graphics, the description and interpretation of results, and preparation of the manuscript draft.

Further Own Publications:

Alt, C. D., Benner, L., Mokry, T., Lenz, F., Hallscheidt, P., Sohn, C., Kauczor, H. U. and Brocker, K. A. (2018), 'Five-year outcome after pelvic floor reconstructive surgery: evaluation using dynamic magnetic resonance imaging compared to clinical examination and quality-of-life questionnaire', *Acta Radiologica* **59**(10), 1264-1273.

Arians, N., Kieser, M., Benner, L., Rochet, N., Katayama, S., Sterzing, F., Herfarth, K., Schubert, K., Schröder, L., Leitzen, C., Schneeweiss, A., Sohn, C., Debus, J. and Lindel, K. (2017), 'Adjuvant intensity modulated whole-abdominal radiation therapy for high-risk patients with ovarian cancer (International Federation of Gynecology and Obstetrics Stage III): First results of a prospective phase 2 study', *International Journal of Radiation Oncology Biology Physics* **99**(4), 912-920.

Bahrmann, A., Benner, L., Christ, M., Bertsch, T., Sieber, C. C., Katus, H. and Bahrmann P. (2019), 'The Charlson Comorbidity and Barthel Index predict length of hospital stay, mortality, cardiovascular mortality and rehospitalization in unselected

older patients admitted to the emergency department', *Aging Clinical and Experimental Research* (epublised ahead of print, DOI 10.1007/s40520-018-1067-x).

Benner, L., Kirchner, M., Krisam, J., Kunzmann, K. and Sander, A. (2019), *Auswertung klinischer Studien mit SPSS*, Springer Verlag, Wiesbaden.

De La Garza, J. R., Schmidt, M. W., Kowalewski, K. F., Benner, L., Müller, P. C., Kenngott, H. G., Fischer, L., Müller-Stich, B. P. and Nickel, F. (2019), 'Does rating with a checklist improve the effect of E-learning for cognitive and practical skills in bariatric surgery? A rater-blinded, randomized-controlled trial', *Surgical Endoscopy* **33**(5), 1532-1543.

Kowalewski, K. F., Garrow, C. R., Schmidt, M. W., Benner, L., Müller-Stich, B. P. and Nickel F. (2019), 'Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying', *Surgical Endoscopy* (epublised ahead of print, DOI 10.1007/s00464-019-06667-4).

Kowalewski, K. F., Minassian, A., Hendrie, J. D., Benner, L., Preukschas, A. A., Kenngott, H. G., Fischer, L., Müller-Stich, B. P. and Nickel, F. (2019), 'One or two trainees per workplace for laparoscopic surgery training courses: results from a randomized controlled trial', *Surgical Endoscopy* **33**(5), 1523-1531.

Kulu, Y., Müller-Stich, B. P., Ghamarnejad, O., Khajeh, E., Polychronidis, G., Golriz, M., Nickel, F., Benner, L., Knebel, P., Diener, M., Morath, C., Zeier, M., Büchler, M. W. and Mehrabi, A. (2018), 'Hand-assisted laparoscopic donor nephrectomy periumbilical versus Pfannenstiel incision and return to normal physical activity (HAPERPACT): study protocol for a randomized controlled trial', *Trials* **19**(377).

Kunzmann, K., Benner, L. and Kieser, M. (2017), 'Point estimation in adaptive enrichment designs', *Statistics in Medicine* **36**(25), 3935-3947.

Müller, P. C., Dube, A., Steinemann, D. C., Senft, J. D., Gehrig, T., Benner, L., Nickel, F., Müller-Stich, B. P. and Linke, G. R. (2018), 'Contamination after disinfectant rectal washout in left colectomy as a model for transrectal NOTES: a randomized controlled trial', *Journal of Surgical Research* **232**, 635-642.

Nickel, F., Tapking, C., Benner, L., Schüler, S., Ottawa, G. B., Krug, K., Müller-Stich,

B. P. and Fischer, L. (2019), 'Video teaching leads to improved attitudes towards obesity - a randomized study with 949 participants', *Obesity Surgery* **29**(7), 2078-2086.

Nickel, F., Tapking, C., Benner, L., Sollors, J., Billeter, A. T., Kenngott, H. G., Bokhary, L., Schmid, M., von Frankenberg, M., Fischer, L., Müller, S. and Müller-Stich, B. P. (2018), 'Bariatric surgery as an efficient treatment for non-alcoholic fatty liver disease in a prospective study with 1-year follow-up: BariScan study', *Obesity Surgery* **28**, 1342-1350.

Roch, P. J., Friedrich, M., Kowalewski, K. F., Schmidt, M. W., De la Garza Herrera, J., Müller, P. C., Benner, L., Romero, P., Müller-Stich, B. P. and Nickel, F. (2017), 'Neue Wege zum chirurgischen Nachwuchs - Studierendenforum für Minimal Invasive Chirurgie', *Zentralblatt für Chirurgie* **142**, 560-565.

Roch, P. J., Rangnick, H. M., Brzoska, J. A., Benner, L., Kowalewski, K. F., Müller, P. C., Kenngott, H. G., Müller-Stich, B. P. and Nickel, F. (2018), 'Impact of visual-spatial ability on laparoscopic camera navigation training', *Surgical Endoscopy* **32**(3), 1174-1183.

Swartman, B., Benner, L., Grechenig, S., Franke, J., Grützner, P. A. and Schnetzke M. (2019), 'Normal values of distal radioulnar translation assessed by three-dimensional C-arm scans: a cadaveric study', *Journal of Hand Surgery (European Volume)* **44**(5), 503-509.

Schäfer, F., Benner, L., Borzych-Duzalka, D., Zaritsky, J., Xu, H., Rees, L., Antonio, Z. L., Serdaroglu, E., Hooman, N., Patel, H., Sever, L., Vondrak, K., Flynn, J., Rébori, A., Wong, W., Hölttä, T., Yildirim, Z. Y., Ranchin, B., Grenda, R., Testa, S., Drozdz, D., Szabo, A. J., Eid, L., Basu, B., Vitkevic, R., Wong, C., Pottoore, S. J., Müller, D., Dusunsel, R., Gonzalez Celedon, C., Fila, M., Sartz, L., Sander, A., Warady, B. A. and International Pediatric Peritoneal Dialysis Network (IPPN) Registry (2019), 'Global variation of nutritional status in children undergoing chronic peritoneal dialysis: a longitudinal study of the International Pediatric Peritoneal Dialysis Network', *Scientific Reports* **9**, 4886.

Schmidt, M. W., Kowalewski, K. F., Schmidt, M. L., Wennberg, E., Garrow, C. R., Paik, S., Benner, L., Schijven, M. P., Müller-Stich, B. P. and Nickel F. (2019), 'The Heidelberg VR Score: development and validation of a composite score for laparoscopic

virtual reality training', *Surgical Endoscopy* (e-published online, DOI 10.1007/s00464-018-6480-x).

Schmidt, M. W., Kowalewski, K. F., Trent, S. M., Benner, L., Müller-Stich, B. P. and Nickel F. (2019), 'Self-directed training with e-learning using the first-person perspective for laparoscopic suturing and knot tying: a randomised controlled trial: Learning from the surgeon's real perspective', *Surgical Endoscopy* (epublised ahead of print, DOI 10.1007/s00464-019-06842-7).

Curriculum Vitae

Personal information:

Name:	Laura Benner (née Kohlhas)
Date of birth:	08.04.1990
Place of birth:	Hachenburg

Education:

Since 01/2015	Doctoral student at the University of Heidelberg
09/2012 - 11/2014	Studies of Applied Mathematics (M.Sc.) at the University of Applied Sciences, RheinAhrCampus Remagen
09/2009 - 09/2012	Studies of Biomathematics (B.Sc.) at the University of Applied Sciences, RheinAhrCampus Remagen
08/2006 - 03/2009	Privates Gymnasium Marienstatt (Abitur)
08/2000 - 06/2006	Graf-Heinrich Realschule, Hachenburg (qualifizierter Sekundarabschluss I)
09/1996 - 06/2000	Grundschule am Schloss, Hachenburg

Professional experience:

Since 01/2015	Research fellow at the Institute of Medical Biometry and Informatics, University of Heidelberg
03/2012 - 03/2014	Student assistant at the Institute of Medical Statistics and Bioinformatics, University of Cologne

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Meinhard Kieser for offering me the possibility to write this thesis at the Institute of Medical Biometry and Informatics, and for proposing the subject of the thesis. I am very grateful for his excellent support and advice, and for his valuable and constructive suggestions that contributed greatly to the development of this thesis.

I also gratefully acknowledge the support for this work by the program 'Mathematics for innovations in industry and services' of the German Federal Ministry of Education and Research (BMBF) under grant 05M13VHC.

Furthermore, I would like to thank my colleagues at the Institute of Medical Biometry and Informatics for their helpfulness, and for creating a very pleasant working atmosphere.

Lastly, I wish to thank my family for their constant encouragement, for always believing in me, and for supporting me in all my pursuits.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema 'Choice of the Interim Analysis Timing in Adaptive Enrichment Designs' handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift