Group Sequential and Adaptive Designs for Three-Arm 'Gold Standard' Non-Inferiority Trials

Dem Fachbereich 03: Mathematik/Informatik der
Universität Bremen
zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)

eingereichte Dissertation

von
Herrn **Patrick Schlömer, M.Sc.**
geb. am 25.05.1985 in Vechta

1. Gutachter: Prof. Dr. Werner Brannath
2. Gutachter: Prof. Dr. Dr. h.c. Jürgen Timm

Datum der Einreichung: 15.05.2014
Datum des Kolloquiums: 24.06.2014

Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, ohne deren Unterstützung diese Arbeit nicht möglich gewesen wäre.
Zu aller erst möchte ich mich recht herzlich bei Dr. Hermann Kulmann und der gesamten 'Clinical Statistics Europe' Abteilung der Bayer Pharma AG in Berlin bedanken. Durch das mir gewährte Promotionsstipendium bot sich mir die Möglichkeit mich eingehend mit meinem Forschungsprojekt zu befassen und nebenbei in sämtliche Aktivitäten der Abteilung eingebunden zu werden. So bekam ich neben meiner Promotion bereits einen umfangreichen Einblick in die tägliche Arbeit in einem forschenden Pharmaunternehmen. Mein besonderer Dank gilt hierbei Dr. Harald Siedentop und Bodo Kirsch für die inhaltliche Betreuung vor Ort und meiner Zimmerkollegin Dr. Vivian Lanius, die stets ein offenes Ohr für mich hatte und mich des öfteren in fachlich interessante Diskussionen integrierte.
Zu großem Dank verpflichtet bin ich meinem Doktorvater Prof. Dr. Werner Brannath für die umfangreiche Betreuung meines Projektes und die wertvollen Anregungen und Kommentare. Unsere Treffen in Bremen waren während meiner Zeit in Berlin stets eine gelungene Abwechslung, bei denen viele interessante Ideen entstanden sind.
Ein großes Dankeschön gebührt auch meinem Zweitgutachter Prof. Dr. Dr. h.c. Jürgen Timm, der es mir in der letzten Phase meiner Promotion ermöglichte, an der Universität Bremen weitere Erfahrungen in der praktischen Arbeit als Biostatistiker zu erlangen und an der Lehre mitzuwirken. In diesem Zusammenhang möchte ich auch allen Kolleginnen und Kollegen am Kompetenzzentrum für Klinische Studien Bremen für die angenehme Atmosphäre und die interessanten Begegnungen und Gespräche danken. Insbesondere sei hierbei Dr. Sylvia Schmidt gedankt, die mit ihrem akribischen Korrekturlesen zur Verbesserung dieser Dissertation beigetragen hat.
Zu guter Letzt möchte ich noch meiner Familie und meinen Freunden für die gelungene und stets willkommene Ablenkung von der Arbeit und die tollen gemeinsamen Erlebnisse danken. Insbesondere möchte ich meinen Eltern und Tanja Werner danken, ohne deren ständige Unterstützung und Geduld diese Arbeit nicht in dieser Form gelungen wäre.
Der Bayer Pharma AG danke ich recht herzlich für die finanzielle Unterstützung dieses Projektes.

ABSTRACT

This thesis deals with the application of group sequential and adaptive methodology in three-arm non-inferiority trials for the case of normally distributed outcomes. Whenever feasible, use of the three-arm design including a test treatment, an active control and a placebo, is recommended by the health authorities. Nevertheless, especially from an ethical point of view, it is desirable to keep the placebo group size as small as possible.

After giving a short introduction to two-arm non-inferiority trials, we investigate a hierarchical single-stage testing procedure for three-arm trials which starts by assessing the superiority comparison between test and placebo and then proceeds to the test versus control non-inferiority comparison. Based on formulas for the overall power we derive optimal sample size allocations that minimise the overall sample size. Interestingly, the placebo group size turns out to be very low under the optimal allocation. The optimal fixed sample size designs will then serve both as a starting point and a benchmark for the designs determined later.

Subsequently, a general group sequential design for three-arm non-inferiority trials is presented that aims at further minimising the required sample sizes. By choosing different rejection boundaries for the two comparisons we obtain designs with quite different properties. The influence of the boundaries on the operating characteristics such as the expected sample sizes is investigated by means of a comprehensive comparison to the optimal fixed design. Moreover, approximately optimal boundaries are derived for different optimisation criteria such as minimising the placebo group size. It turns out that the implementation of group sequential methodology can further improve the optimal fixed designs, where the potential early termination of the placebo arm is a key advantage that can make the trial more acceptable for patients.

After this, the group sequential testing procedure is extended to adaptive designs that allow data-dependent design changes at the interim analysis. In this context, we discuss optimal mid-trial decision-making based on the observed interim data, with a special focus on sample size re-calculation. In doing so, we will make use of the conditional power and the Bayesian predictive power. Our investigations show the advantages of the proposed adaptive designs over the optimal fixed designs. In particular, the possibility to adapt the sample sizes at interim can help to deal with uncertainties regarding the treatment effects, that often exist in the planning stage of three-arm non-inferiority trials.

We conclude with a discussion of the results and an outlook on possible future work.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations & Symbols

Abbreviations

CHMP	Committee for Medicinal Products for Human Use	6
CPMP	Committee for Proprietary Medicinal Products	10
DSMB	Data Safety Monitoring Board	50
EMA	European Medicines Agency	3
FAS	Full analysis set	13
FDA	Food and Drug Administration	3
FEV$_1$	Forced Expiratory Volume in one second	62
FWER	Familywise error rate	19
ICH	International Conference on Harmonisation	5
IDMC	Independent Data Monitoring Committee	50
ITT	Intention-to-treat	13
l	Litre	62
PP	Per protocol	13
PPS	Per protocol set	13
SPRT	Sequential probability ratio test	38

Symbols

$1-\beta$	Overall statistical power	10
$1-\beta_{TC}$	Statistical power of the test vs. control non-inferiority comparison	29
$1-\beta_{TP}$	Statistical power of the test vs. placebo superiority comparison	29

A_k Event that the group sequential testing procedure rejects $H^{(s)}_{0,TP}$ at stage k ($k = 0$ means no rejection of $H^{(s)}_{0,TP}$ at any stage) and $H^{(n)}_{0,TC}$ is not rejected at stages $k, ..., K$ 53

ASN Overall average sample number at the alternative of the power calculation, i.e. $\theta_{TP} = \delta_{TP}$ and $\theta_{TC} = 0$ 64

$ASN(\theta_{TP}, \theta_{TC})$ Overall average sample number at θ_{TP}, θ_{TC} 57

ASn_C Average sample number of the control group at the alternative of the power calculation, i.e. $\theta_{TP} = \delta_{TP}$ and $\theta_{TC} = 0$ 64

$ASn_C(\theta_{TP}, \theta_{TC})$ Average sample number of the control group at θ_{TP}, θ_{TC} 57

ASn_P Average sample number of the placebo group at the alternative of the power calculation, i.e. $\theta_{TP} = \delta_{TP}$ 64

$ASn_P(\theta_{TP})$ Average sample number of the placebo group at θ_{TP} 57

ASn_T Average sample number of the test group at the alternative of the power calculation, i.e. $\theta_{TP} = \delta_{TP}$ and $\theta_{TC} = 0$ 64

$ASn_T(\theta_{TP}, \theta_{TC})$ Average sample number of the test group at θ_{TP}, θ_{TC} 57

α Significance level 5

$b^{(k)}_{TC}$ Group sequential boundary of the non-inferiority comparison between test and control at stage k 52

$b^{(k)}_{TP}$ Group sequential boundary of the superiority comparison between test and placebo at stage k 52

c_C Quotient of the control divided by the test group sample size 9

$\tilde{c}^{(2)}_C$ Quotient of the control divided by the test group size of the second stage 94

$\tilde{c}^{(2)}_{C,ad}$ Quotient of the control divided by the test group size of the second stage for the adaptive type design 108

c_P Quotient of the placebo divided by the test group sample size 31

$\tilde{c}^{(2)}_P$ Quotient of the placebo divided by the test group size of the second stage 94

$\tilde{c}^{(2)}_{P,ad}$ Quotient of the placebo divided by the test group size of the second stage for the adaptive type design 108

$CP_{TC}(\theta_{TC})$ Conditional power to reject $H^{(s)}_{0,TC}$ at the final analysis when $H^{(s)}_{0,TP}$ has already been rejected at interim and assuming that the true test vs. control treatment difference is θ_{TC} 94

$CP^{fix}_{TP,TC}(\theta_{TP},\theta_{TC})$ Conditional power of the fixed design to reject both $H^{(s)}_{0,TP}$ and $H^{(n)}_{0,TC}$ at the final analysis assuming that the true test vs. placebo and test vs. control treatment differences are θ_{TP} and θ_{TC} 107

$CP_{TP,TC}(\theta_{TP},\theta_{TC})$ Conditional power to reject both $H^{(s)}_{0,TP}$ and $H^{(n)}_{0,TC}$ at the final analysis assuming that the true test vs. placebo and test vs. control treatment differences are θ_{TP} and θ_{TC} 94

$CP_{TP}(\theta_{TP})$ Conditional power to reject $H^{(s)}_{0,TP}$ at the final analysis assuming that the true test vs. placebo treatment difference is θ_{TP} 93

χ^2_ν Chi-squared distribution with ν degrees of freedom 23

Δ Shape parameter of the Wang Tsiatis class of group sequential designs 45

Δ_{TC} Shape parameter of the Wang Tsiatis type boundaries for testing $H^{(n)}_{0,TC}$ 64

Δ_{TP} Shape parameter of the Wang Tsiatis type boundaries for testing $H^{(s)}_{0,TP}$ 64

Δ_{eq} Equivalence margin 4

Δ_{ni} Non-inferiority margin 5

Δ_{sup} Difference representing substantial superiority 7

$\Delta\bar{X}^{(k)}_C$ Sample mean of the control group that is formed from the data of stage k alone 92

$\Delta\bar{X}^{(k)}_P$ Sample mean of the placebo group that is formed from the data of stage k alone 92

$\Delta\bar{X}^{(k)}_T$ Sample mean of the test group that is formed from the data of stage k alone 92

E_k Event that the group sequential testing procedure enters stage k 57

$E_{k_1,k}$ Event that the group sequential testing procedure rejects $H^{(s)}_{0,TP}$ at stage k_1 ($k_1 = 0$ means no rejection of $H^{(s)}_{0,TP}$) and enters stage k 57

$E_\theta(\cdot)$ Conditional expectation given θ 44

f Fraction parameter of the effect preservation hypothesis 26

$f_{HSD}(\cdot)$ Spending function of the error spending family by Hwang, Shih and DeCani 49

$f_{KD}(\cdot)$ Spending function of the Kim DeMets error spending family 49

$f_{OBF}(\cdot)$ Spending function of O'Brien Fleming type error spending designs 49

$f_{POC}(\cdot)$ Spending function of Pocock type error spending designs 49

γ	Spending function parameter of the Hwang Shih DeCani family of error spending designs	49
γ_{TC}	Spending function parameter of the Hwang Shih DeCani family of error spending designs for testing $H_{0,TC}^{(n)}$	67
γ_{TP}	Spending function parameter of the Hwang Shih DeCani family of error spending designs for testing $H_{0,TP}^{(s)}$	67
$H_{0,CP}^{(s)}$	Null hypothesis of the control vs. placebo superiority comparison	28
$H_{0,eq}$	Null hypothesis of an equivalence trial	4
$H_{0,ni}$	Null hypothesis of a non-inferiority trial	5
$H_{0,sup}$	Null hypothesis of a placebo-controlled superiority trial	3
$H_{0,TC}^{(n)}$	Null hypothesis of the test vs. control non-inferiority comparison	28
$H_{0,TC}^{(s)}$	Null hypothesis of the test vs. control superiority comparison	28
$H_{0,TP}^{(s)}$	Null hypothesis of the test vs. placebo superiority comparison	28
$H_{1,eq}$	Alternative hypothesis of an equivalence trial	4
$H_{1,ni}$	Alternative hypothesis of a non-inferiority trial	5
$H_{1,sup}$	Alternative hypothesis of a placebo-controlled superiority trial	3
$H_{1,TC}^{(n)}$	Alternative hypothesis of the test vs. control non-inferiority comparison	28
$H_{1,TP}^{(s)}$	Alternative hypothesis of the test vs. placebo superiority comparison	28
I_0	Set of indices of all true null hypotheses	21
$\mathscr{I}_{max}$	Maximum information level	48
$\mathscr{I}_{TC}^{(k)}$	Information level of the test vs. control non-inferiority comparison at stage k	52
$\mathscr{I}_{TP}^{(k)}$	Information level of the test vs. placebo superiority comparison at stage k	52
K	Number of stages	40
$L(\text{data} \mid \mu)$	Likelihood of some data given a parameter μ	87
μ_C	Treatment effect of the active control treatment	4
$\mu_{C,0}$	Mean of the normal prior distribution of the control treatment effect μ_C	95
μ_C^*	Mean of the normal posterior distribution of the control treatment effect μ_C	95

μ_P	Treatment effect of the placebo treatment	3
$\mu_{P,0}$	Mean of the normal prior distribution of the placebo effect μ_P	95
μ_P^*	Mean of the normal posterior distribution of the placebo effect μ_P	95
μ_T	Treatment effect of the test treatment	3
$\mu_{T,0}$	Mean of the normal prior distribution of the test treatment effect μ_T	95
μ_T^*	Mean of the normal posterior distribution of the test treatment effect μ_T	95
$N(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2	7
$N_k(\boldsymbol{\mu},\boldsymbol{\Sigma})$	k-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$	30
N	Overall sample size	22
N_{ad}	Overall sample size of the adaptive type design	109
N_{bal}	Overall sample size of the balanced fixed sample size design	32
N_{fix}	Overall sample size of the fixed sample size design	46
$N^{(k)}$	Cumulative overall sample size at stage k	59
N_{max}	Maximum overall sample size	46
N_{opt}	Overall sample size of the optimal fixed sample size design	32
n_C	Sample size of the active control group	7
$n_{C,ad}$	Control group size of the adaptive type design	109
$n_{C,bal}$	Control group size of the balanced fixed sample size design	32
$n_{C,fix}$	Control group size of the fixed sample size design	62
$n_C^{(k)}$	Cumulative sample size of the control group at stage k	51
$n_{C,opt}$	Control group size of the optimal fixed sample size design	32
$\tilde{n}_{C,ad}^{(2)}$	Second-stage control group size of the adaptive type design	108
$\tilde{n}_C^{(k)}$	Sample size of the control group in stage k	91
n_P	Sample size of the placebo group	27
$n_{P,ad}$	Placebo group size of the adaptive type design	109
$n_{P,bal}$	Placebo group size of the balanced fixed sample size design	32

$n_{P,fix}$	Placebo group size of the fixed sample size design	62
$n_P^{(k)}$	Cumulative sample size of the placebo group at stage k	51
$n_{P,opt}$	Placebo group size of the optimal fixed sample size design	32
$\tilde{n}_{P,ad}^{(2)}$	Second-stage placebo group size of the adaptive type design	108
$\tilde{n}_P^{(k)}$	Sample size of the placebo group in stage k	91
n_T	Sample size of the test group	7
N_T	Actual sample size of the test group	57
$n_{T,ad}$	Test group size of the adaptive type design	109
$n_{T,bal}$	Test group size of the balanced fixed sample size design	32
$n_{T,fix}$	Test group size of the fixed sample size design	62
$n_T^{(k)}$	Cumulative sample size of the test group at stage k	51
$n_{T,opt}$	Test group size of the optimal fixed sample size design	32
$\tilde{n}_{T,ad}^{(2)}$	Second-stage test group size of the adaptive type design	108
$\tilde{n}_T^{(k)}$	Sample size of the test group in stage k	91
$\mathbb{N}$	Set of natural numbers	49
ν	Number of degrees of freedom	7
P_i	Random p-value	21
p_k	p-value that is formed from the data of stage k alone	112
PP_{TC}	Predictive power to reject $H_{0,TC}^{(n)}$ at the final analysis when $H_{0,TP}^{(s)}$ has already been rejected at interim	96
PP_{TP}	Predictive power to reject $H_{0,TP}^{(s)}$ at the final analysis	95
$PP_{TP,TC}$	Predictive power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ at the final analysis	97
$P_\theta(\cdot)$	Conditional probability given θ	10
$\Phi(\cdot)$	Cumulative distribution function of the standard normal distribution	10
$\phi(\cdot)$	Probability density function of the standard normal distribution	88
$\Phi^{-1}(\cdot)$	Inverse cumulative distribution function of the standard normal distribution	112

$\Phi^{\boldsymbol{\Sigma}}(\cdot)$ Cumulative distribution function of the multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ 30

$\phi^{\boldsymbol{\Sigma}}(\cdot,\cdot)$ Probability density function of the bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ 97

$\Phi^{\boldsymbol{\Sigma}}(\boldsymbol{a};\boldsymbol{b})$ Distribution function of the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at the vector of lower limits $\boldsymbol{a}$ and the vector of upper limits $\boldsymbol{b}$ 56

π_k Probability that a group sequential test falsely rejects the null hypothesis at stage k . 48

$\pi(\mu)$ Prior density of a parameter μ 87

$\pi(\mu \mid \text{data})$ Posterior density of a parameter μ given some data 86

R_{k_1,k_2} Event that the group sequential testing procedure rejects $H_{0,TP}^{(s)}$ at stage k_1 and $H_{0,TC}^{(n)}$ at stage k_2 54

$\mathbb{R}$ Set of real numbers 3

ρ Spending function parameter of the Kim DeMets family of error spending designs .. 49

ρ_{TC} Spending function parameter of the Kim DeMets family of error spending designs for testing $H_{0,TC}^{(n)}$ 67

ρ_{TP} Spending function parameter of the Kim DeMets family of error spending designs for testing $H_{0,TP}^{(s)}$ 67

S_C^2 Unbiased sample variance of the control group 7

S_P^2 Unbiased sample variance of the placebo group 28

S_T^2 Unbiased sample variance of the test group 7

σ^2 Common variance 7

$\hat{\sigma}^2$ Unbiased pooled estimator of σ^2 7

$\sigma_{C,0}^2$ Variance of the normal prior distribution of the control treatment effect μ_C 95

${\sigma_C^*}^2$ Variance of the normal posterior distribution of the control treatment effect μ_C 95

$\sigma_{P,0}^2$ Variance of the normal prior distribution of the placebo effect μ_P 95

${\sigma_P^*}^2$ Variance of the normal posterior distribution of the placebo effect μ_P 95

$\sigma_{T,0}^2$ Variance of the normal prior distribution of the test treatment effect μ_T 95

$\sigma_T^{*\,2}$ Variance of the normal posterior distribution of the test treatment effect μ_T 95

$\mathbf{\Sigma}(i,...,j)$ Matrix that is formed by deleting the rows and columns $i,...,j$ of $\mathbf{\Sigma}$ 44

$t_{\gamma,\nu}$ γ-quantile of the t-distribution with ν degrees of freedom 7

$T_{TC}^{(n)}$ Student's t-test statistic of the non-inferiority comparison between test and control . 28

$T_{TP}^{(s)}$ Student's t-test statistic of the superiority comparison between test and placebo 28

$\mathcal{T}_\nu(\cdot\,|\,\gamma)$ Cumulative distribution function of the noncentral t-distribution with ν degrees of freedom and noncentrality parameter γ 10

$\mathcal{T}_\nu^{\mathbf{\Sigma}}(\cdot\,|\,\boldsymbol{\gamma})$ Cumulative distribution function of the multivariate noncentral t-distribution with ν degrees of freedom, vector of noncentrality parameters $\boldsymbol{\gamma}$ and variance-covariance matrix $\mathbf{\Sigma}$ 30

θ_{CP} Treatment difference between control and placebo group 64

θ_{TC} Treatment difference between test and control group 29

θ_{TP} Treatment difference between test and placebo group 29

$w_{TC}^{(k)}$ Weight assigned to the kth stage of the adaptive test versus control non-inferiority comparison 92

$w_{TP}^{(k)}$ Weight assigned to the kth stage of the adaptive test versus placebo superiority comparison 92

$X_{C,i}$ Observed treatment effect for patient $i \in \{1,2,...,n_C\}$ under control treatment 7

$X_{P,i}$ Observed treatment effect for patient $i \in \{1,2,...,n_P\}$ under placebo treatment 27

$X_{T,i}$ Observed treatment effect for patient $i \in \{1,2,...,n_T\}$ under test treatment 7

$\bar{X}_C$ Sample mean of the observed treatment effects in the control group 7

$\bar{X}_C^{(k)}$ Cumulative sample mean of the observed treatment effects in the control treatment group at stage k 51

$\bar{X}_P$ Sample mean of the observed treatment effects in the placebo group 28

$\bar{X}_P^{(k)}$ Cumulative sample mean of the observed treatment effects in the placebo group at stage k 51

$\bar{X}_T$ Sample mean of the observed treatment effects in the test group 7

$\bar{X}_T^{(k)}$ Cumulative sample mean of the observed treatment effects in the test treatment group at stage k 51

INTRODUCTION

At the end of a drug approval process the drug manufacturer has to provide sufficient evidence regarding the efficacy of the respective therapy. This needs to be accomplished within two adequate clinical studies comparing the experimental treatment with a control treatment. Usually, a placebo control is chosen, so that the studies aim at demonstrating that the treatment effect of the new therapy is greater than the placebo effect, i.e. the test treatment is superior to placebo. However, in situations when there already exists an approved treatment for the respective medical indication, it does not seem to be ethically justifiable to expose patients to placebo. Then, demonstrating the superiority over an approved active comparator would be an obvious approach in order to prove the efficacy of the new treatment. Due to the fact that, nowadays, most treatments on the market are highly efficacious, such a superiority claim is infeasible in most cases as this would require extremely large sample sizes.

This is where the so-called *non-inferiority trials* come into play which serve as an indirect proof of efficacy by demonstrating that the test treatment is not substantially worse than the active control treatment. Certainly the question arises: Is there a need for a new treatment that might be even slightly inferior to those already on the market? In this regard it could be argued that the new therapy might have safety benefits over the standard treatment or it could serve as a second-line treatment when the initial therapy with the standard drug has failed. New routes of application could be another reason for approving a slightly inferior treatment. Besides the evident problem of defining what "substantially worse" means in practical terms, non-inferiority trials are also associated with other methodological problems. One of the major issues is that demonstrating non-inferiority to an active control, in contrast to superiority over placebo, does not necessarily imply a proof of efficacy for the test treatment. Without an existing placebo group it cannot be completely ruled out that the control treatment performs significantly worse than expected, potentially making the non-inferiority claim useless. Moreover, poor study quality could diminish the treatment difference between the two groups, resulting in a bias towards non-inferiority. Consequently, the regulatory authorities state that a "three-armed trial with test, reference, and placebo [...] is therefore the recommended design; it should be used whenever possible." (CHMP, 2005b). Therefore, the three-arm design is also denoted as the "gold standard design" (Koch and Röhmel, 2004).

After giving a short introduction to two-arm non-inferiority trials and the corresponding methodological issues in Chapter 1, the following chapters will deal with such three-arm 'gold standard' non-inferiority trials for the case of normally distributed treatment effects. The main focus will be placed on reducing the overall number of required patients and, in particular, on minimising the number of subjects receiving placebo in order to make better use of the resources and to make the trial more acceptable from an ethical point of view. Therefore, we will derive optimal sample size allocations for three-arm designs with prespecified, fixed sample sizes in Chapter 2. These optimal fixed designs will then serve both as a starting point and a benchmark for the designs presented in the subsequent part of this work.

In most clinical trials we have prespecified, fixed sample sizes and the corresponding statistical analyses are conducted only once, namely after the observations of all patients have been recorded. In contrast, so-called *group sequential designs* allow to repeatedly assess the data at various points in time and to decide on study continuation or termination, either for positive or negative study outcomes, based on the data accrued by then. In Chapter 3 we will derive group sequential designs for three-arm non-inferiority trials together with corresponding power and sample size formulas. Such designs are supposed to improve the optimal fixed designs with respect to overall sample size savings. Moreover, we can make use of the additional possibility to stop allocating patients to placebo once the proof of efficacy has been shown, making the study even more acceptable for patients. After a comprehensive comparison with the optimal fixed designs, approximately optimal group sequential designs will be derived for certain optimality criteria such as minimising the expected placebo group size.

In Chapter 4 we will extend the proposed group sequential designs to so-called *adaptive designs* that allow data-dependent design changes at an interim analysis, such as sample size re-calculations. There are various reasons for such adaptations, although the main one certainly is the ability to make better use of the available resources. In the context of three-arm non-inferiority trials this will also give us the ability to account for uncertainties regarding the treatment effects, which often exist in the planning stage of such trials. Moreover, we can overcome the issue of a possible change in patient population after dropping the placebo group, by reducing the placebo group size to a certain threshold instead of completely closing it. Through this, potential heterogeneities can also be better identified afterwards by comparing the independent results from the different stages. For the proposed adaptive designs we will also discuss optimal mid-trial decision-making based on the observed interim data. In doing so, we will make use of the so-called *conditional power* which is the probability that the null hypothesis will be rejected at the final analysis given the interim observations. The *Bayesian predictive power* will also be considered, which is the Bayesian version of the conditional power.

This thesis ends with a summary and discussion of the results. In particular, we will discuss potential issues associated with the practical application of the proposed procedures. Finally, an outlook will give an insight into further extensions of the presented methods.

CHAPTER 1

Two-Arm Non-Inferiority Trials

This chapter gives an introduction to two-arm non-inferiority trials starting with a short motivation in Section 1.1. Section 1.2 gives an overview on the statistical methodology of a two-arm non-inferiority trial, such as the corresponding statistical tests and sample size calculations. Finally, Section 1.3 deals with the methodological problems that arise during the planning stage and trial conduct, such as the choice of the non-inferiority margin.

The following assumptions are made for this chapter. First of all, confirmatory phase III trials are considered that assess the efficacy with one primary endpoint. Further, it is assumed that higher values represent larger treatment effects. In addition, a zero treatment effect shall represent no effect and the statistical analyses are assumed to be based on the absolute difference between the treatment effects. Thus, a treatment difference of zero means that the two treatments are equally efficacious. With only few modifications the theory described is applicable to trials based e.g. on the relative risk. No treatment difference would then be represented by an estimated relative risk of one.

1.1 Motivation

In general, regulatory authorities such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) require two adequate and well-controlled trials supporting the efficacy of the new drug for approval. Usually, the efficacy of the test treatment is assessed in a placebo-controlled *superiority trial* with the set of hypotheses

$$H_{0,sup}:\ \mu_T - \mu_P = 0 \quad \text{vs.} \quad H_{1,sup}:\ \mu_T - \mu_P \neq 0, \tag{1.1}$$

where $\mu_T,\ \mu_P \in \mathbb{R}$ represent the treatment effect of the test and the placebo treatment, respectively. Superiority of the test treatment over the placebo, also denoted as efficacy of the test treatment, is claimed if $H_{0,sup}$ can be rejected and the point estimate of μ_T is greater than the point estimate of μ_P. Superiority can similarly be shown with the one-sided hypotheses $H_{0,sup} : \mu_T - \mu_P \leq 0$ vs. $H_{1,sup} : \mu_T - \mu_P > 0$ and half of the significance level used for the two-sided hypotheses.

Often, a placebo-controlled trial is not ethically justifiable, e.g. for oncology trials. In addition, there is a large number of proven effective treatments on the market that could also serve as a comparator in a phase III trial. Thus, one possible approach could be to show that the test treatment is superior to an active control treatment that has already been approved and is the standard treatment for the medical indication of interest. However, often the treatments on the market are highly efficacious, in which case it can be very difficult to demonstrate the superiority of the new treatment, even if $\mu_T > \mu_C$, where $\mu_C \in \mathbb{R}$ represents the treatment effect of the control treatment. This is because the treatment difference $\mu_T - \mu_C$ is so small that it would require extremely large sample sizes to obtain a sufficient power for the superiority comparison. In this case one could aim at demonstrating that the treatment effect of the test treatment is comparable to that of the active comparator, which would be an indirect proof of efficacy for the test treatment as the active control has already been shown to be superior to placebo. Certainly, the question arises whether a treatment with a comparable or an only slightly higher treatment effect is urgently needed. However, the new treatment could have safety benefits over the current standard treatment or it could serve as a second-line treatment, e.g. if the initial treatment with the standard drug has failed. A new route of application (e.g. oral) could also be a reason for approval.

A simple ad hoc approach could be to use the same set of hypotheses as for the superiority comparison in (1.1), except that μ_P is replaced by μ_C, and try to *prove* the null hypothesis of no treatment difference (Blackwelder, 1982). That means, if the null hypothesis is not rejected, one should declare that the test treatment effect is equivalent to the treatment effect of the active control. However, this approach is not sensible and probably leads to wrong conclusions, because absence of evidence is not evidence of absence. In particular, it highly depends on the sample size of the study and the true treatment difference $\mu_T - \mu_C$, that means the actual type II error. By means of a small sample size one could thus demonstrate the equivalence of any two treatments with this concept, even if the two treatment effects differ substantially. This is because the type II error increases with decreasing sample size, so that the probability to accept $H_{0,sup}$, although actually $\mu_T \neq \mu_C$ holds, is very high for small sample sizes. Proving the null hypothesis of no treatment difference would thus require an infinitely large clinical trial. However, *equivalence trials* provide an established approach to this problem with the set of hypotheses

$$H_{0,eq} : \left|\mu_T - \mu_C\right| \geq \Delta_{eq} \quad \text{vs.} \quad H_{1,eq} : \left|\mu_T - \mu_C\right| < \Delta_{eq},$$

where $\Delta_{eq} > 0$ represents the so-called *equivalence margin*. The margin is chosen to ensure that there is “no meaningful” difference between the two treatments if the null hypothesis is rejected. A common example is the approval of a generic drug, i.e. a chemically equivalent copy of an already approved drug, where *bioequivalence trials* are carried out in order to show that the generic drug has almost the same pharmacokinetic properties as the original drug.

Another approach to trials with an active control arm are *non-inferiority trials*. First of all, the term “non-inferiority” could be misleading if taken literally, because only a superiority trial can establish real non-inferiority. The aim of a non-inferiority trial is to demonstrate that the test treatment effect is not worse than that of the active control by more than a prespecified amount $\Delta_{ni} > 0$, the so-called *non-inferiority margin*. The corresponding set of hypotheses of the statistical test is given as

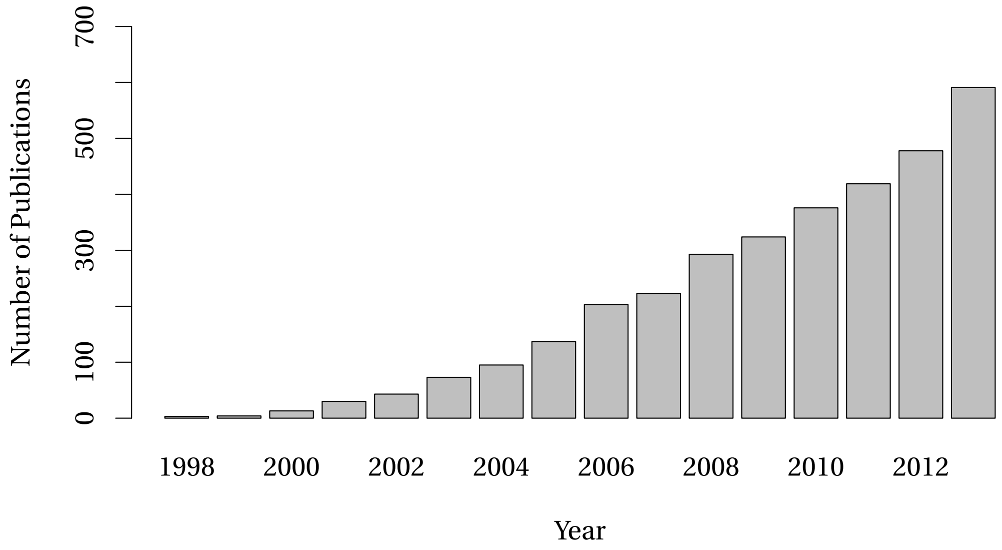
$$H_{0,ni}: \mu_T - \mu_C \leq -\Delta_{ni} \quad \text{vs.} \quad H_{1,ni}: \mu_T - \mu_C > -\Delta_{ni}. \tag{1.2}$$

A claim for non-inferiority should then serve on the one hand as an indirect proof of efficacy, i.e. superiority of the test treatment over putative placebo, and on the other hand as a direct assessment of the similarity to the active comparator. In order to satisfy these two goals, the non-inferiority margin should be determined “based on both statistical reasoning and clinical judgment”, as it is stated in the ICH E10 guideline (International Conference on Harmonisation). The choice of Δ_{ni} is a key element in such trials and it often turns out to be a difficult task. Section 1.3.4 will go into more detail on this problem. Compared with equivalence trials, the null hypothesis $H_{0,ni}$ is also rejected in case that the test treatment effect is substantially larger than that of the active control. Thus, non-inferiority trials are more common in the confirmatory stage (phase III) than equivalence trials, not at least because one can test for superiority without α-adjustment once $H_{0,ni}$ has been rejected. This follows from the *closed testing principle* introduced by Marcus et al. (1976) and the fact that $H_{0,ni}$ is a subset of the corresponding one-sided superiority null hypothesis, i.e. $\{\mu_T - \mu_C \leq -\Delta_{ni}\} \subset \{\mu_T - \mu_C \leq 0\}$. An overview on multiple testing procedures is given in Section 2.1.

Taking a look at the number of publications regarding non-inferiority trials shown in Figure 1.1, their increasing importance becomes apparent. From 1998 to 2013 there have been 3305[1] publications (including methodological work) that are related to non-inferiority trials and the number of publications per year on this topic is sharply increasing. One of the first major contributions on non-inferiority trials was by Röhmel (1998), who especially goes into detail on the methodological problems that are associated with the design and analysis of non-inferiority trials, such as the right choice of the non-inferiority margin Δ_{ni}. Several issues relating to non-inferiority trials are also addressed in the ICH E9 and ICH E10 guidelines, which are two of the most important guidelines for statisticians working in the pharmaceutical industry. Non-inferiority trials have become increasingly important in the recent years for many reasons, especially as they raise challenging statistical problems. Not at least because of numerous

[1] Based on a PubMed search (accessed on May 12, 2014) for the term “non-inferiority OR noninferiority”

Figure 1.1: Number of publications per year regarding non-inferiority trials before 2014. The results are based on a PubMed search (accessed on May 12, 2014) for the term "non-inferiority OR noninferiority".



methodological problems that are related with non-inferiority trials, there are comprehensive regulatory guidelines regarding this topic, e.g. the *Draft Guidance for Industry: Non-Inferiority Clinical Trials* by the FDA (2010b) and the *Guideline on the Choice of the Non-Inferiority Margin* by the Committee for Medicinal Products for Human Use (CHMP, 2005b). Furthermore, active control non-inferiority trials are part of the *FDA's Critical Path Initiative* mentioned by O'Neill (2006), that describes emerging challenges on the critical path of drug development and opportunities for statisticians to make contributions to these problems.

1.2 Statistical Methodology

As we have seen, non-inferiority trials have gained more and more attention in the recent years, especially due to the increasing number of approved effective treatments on the market. Thus, two-arm active control non-inferiority trials should serve as a substitute for placebo-controlled superiority trials in order to obtain drug approval. The following sections will take a closer look at the statistical methodology of two-arm non-inferiority trials with normally distributed outcomes.

1.2.1 Statistical Model and Test Procedure

Let us first assume that all observations of the primary endpoint under the test treatment and the active comparator are mutually independent and normally distributed with common, but unknown variance σ^2, i.e. $X_{T,i} \sim N(\mu_T, \sigma^2)$, $i = 1,2,...,n_T$, and $X_{C,i} \sim N(\mu_C, \sigma^2)$, $i = 1,2,...,n_C$. The corresponding set of hypotheses for a non-inferiority trial is given in (1.2), where the *non-inferiority margin* Δ_{ni} is a small, prespecified amount greater than zero. In other words, it should be demonstrated that the test treatment is not worse than the active comparator by more than Δ_{ni} (with respect to efficacy). The choice of Δ_{ni} is not trivial and requires statistical as well as medical considerations that need to be discussed with health authorities. Section 1.3.4 goes into more detail on this issue.

Often, the additional question arises if the test treatment significantly outperforms the active control. Thus, a subsequent test for superiority, i.e. $\Delta_{ni} = 0$, could be performed once $H_{0,ni}$ has been rejected. For instance, Röhmel (1998) suggests to adopt the following procedure in a non-inferiority or superiority trial with an active control: Starting with non-inferiority and then proceeding to superiority and, once superiority has been demonstrated, to substantial superiority by more than $\Delta_{sup} > 0$ which can be data-driven. According to the closed testing principle (Marcus et al., 1976) it is not necessary to adjust the α-level of the three hypothesis tests, as $\{\mu_T - \mu_C \leq -\Delta_{ni}\} \subset \{\mu_T - \mu_C \leq 0\} \subset \{\mu_T - \mu_C \leq \Delta_{sup}\}$. A short introduction to multiple testing procedures will be given in Section 2.1.

The statistical analysis of $H_{0,ni}$ is usually based on the *Student's t-test* statistic

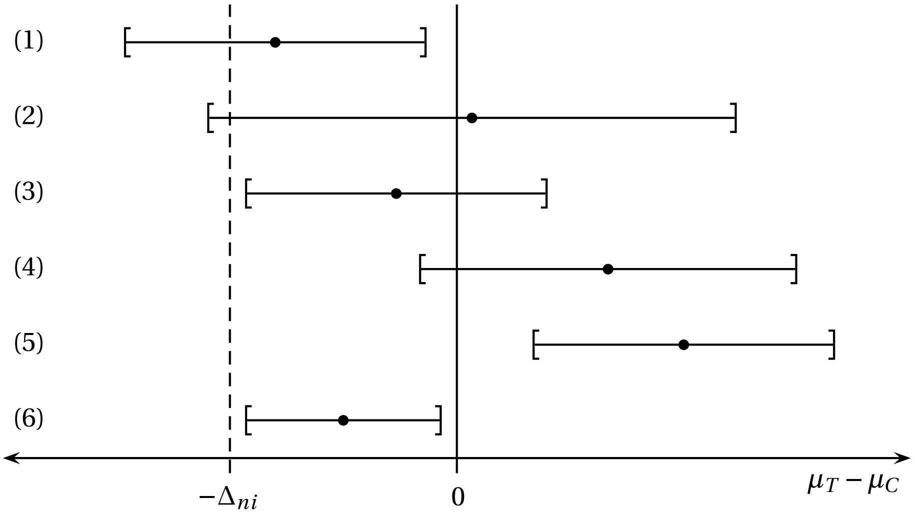
$$T = \frac{\bar{X}_T - \bar{X}_C + \Delta_{ni}}{\hat{\sigma}} \sqrt{\frac{n_T n_C}{n_T + n_C}},$$

where $\bar{X}_T = \frac{1}{n_T} \sum_{i=1}^{n_T} X_{T,i}$ and $\bar{X}_C$ is defined analogously. The common variance σ^2 is estimated by the unbiased pooled estimator $\hat{\sigma}^2 = ((n_T - 1)S_T^2 + (n_C - 1)S_C^2)/(n_T + n_C - 2)$, where $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2, i = T, C,$ denote the unbiased sample variances of the test and the control group, respectively. Given $H_{0,ni}$ is true, it can be shown that T follows or is stochastically smaller than a t-distribution with $\nu = n_T + n_C - 2$ degrees of freedom. Thus, T is compared with $t_{1-\alpha,\nu}$, i.e. the $(1-\alpha)$-quantile of the t-distribution with ν degrees of freedom. In confirmatory phase III trials usually a two-sided significance level of $\alpha = 0.05$ is used. According to that, a significance level of $\alpha = 0.025$ is common for phase III non-inferiority trials due to the one-sided hypothesis testing.

In general, the interpretation of non-inferiority trials is based on a confidence interval for the treatment difference between test and control. This has several benefits over common hypothesis testing and is also recommended by the respective regulatory agencies. In accordance with the significance level of 2.5% for the one-sided hypothesis test in confirmatory phase III trials, one usually calculates a two-sided 95% confidence interval for $\mu_T - \mu_C$ (or equivalently a one-sided 97.5% confidence interval). For normally distributed outcomes with a common but unknown variance σ^2 the corresponding two-sided confidence interval for $\mu_T - \mu_C$ is given as

$$\bar{X}_T - \bar{X}_C \pm t_{1-\alpha, n_T, n_C - 2}\, \hat{\sigma} \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}.$$

Figure 1.2: Six different scenarios for a non-inferiority trial with the corresponding two-sided 95% confidence intervals for the treatment difference $\mu_T - \mu_C$. Note that higher values mean that the test is better than the control treatment.



Note that this is the two-sided $(1-2\alpha)$ confidence interval for $\mu_T - \mu_C$, consistent with the one-sided hypothesis test at level α. Non-inferiority of the test treatment to the active control is stated if the lower bound of the confidence interval exceeds $-\Delta_{ni}$. In addition, the point estimate $\bar{X}_T - \bar{X}_C$ of the treatment difference $\mu_T - \mu_C$ should not be neglected in the interpretation of a non-inferiority trial.

Figure 1.2 shows six different outcomes of a non-inferiority trial represented by the two-sided 95% confidence interval for the treatment difference $\mu_T - \mu_C$. First of all, the scenarios can be divided into two groups: trials, where non-inferiority of the test treatment to the active comparator could (scenarios (3)-(6)) and could not be concluded (scenarios (1) and (2)). However, there are major differences between the scenarios within the two groups, depending on the boundaries of the corresponding confidence interval and the point estimate. For the first two scenarios the lower bound of the confidence interval is less than $-\Delta_{ni}$, indicating that non-inferiority could not be shown. For scenario (1) the test treatment is also significantly inferior to the active comparator as the upper bound of the confidence interval is less than zero. In contrast, the point estimate for scenario (2) suggests that the test treatment might be (slightly) superior to the control. The large confidence interval indicates that the study is underpowered, e.g. due to an underestimation of the standard deviation during the planning stage. In this case a larger sample size might have saved the study resulting in a narrower confidence interval that excludes $-\Delta_{ni}$. In scenarios (3) and (4) the confidence interval lies completely to the right of $-\Delta_{ni}$, therefore non-inferiority can be concluded. Nevertheless, both scenarios differ substantially as the point estimates of these two scenarios are contrary. In scenario (3) the point estimate indicates that the test treatment is slightly inferior to the active control, whereas the point estimate of scenario (4) suggests that the test treatment is even superior to the control. With an only slightly higher sample size even superiority might have been demonstrated, as it is the case for scenario (5) where the confidence interval completely exceeds zero. The last scenario represents an unusual outcome of a non-inferiority study, that is associated with interpretive problems. As we can see, the point estimate favours the active control and even superiority to the test treatment can be demonstrated as the upper bound of the confidence interval is less than zero. Furthermore, the lower bound of the confidence interval is greater than $-\Delta_{ni}$. In other words, the test treatment is at the same time inferior and non-inferior to the active comparator which is a contradiction in terms. Nonetheless, with such an outcome it is valid to state that the test treatment is non-inferior to the active control. If this is not acceptable, one should reconsider the choice of the non-inferiority margin which is probably chosen too large. However, such an outcome is rather rare in practice and would require a very high sample size, as the non-inferiority margin usually is very small. If a far too high standard deviation for the sample size calculation is assumed, such an outcome might occur. Thus, such a scenario should be discussed while planning the trial, especially if there are uncertainties regarding the standard deviation.

1.2.2 Power and Sample Size

Besides the significance level, which is also denoted as the consumer's risk, the statistical power is of main importance for the producer, especially during the planning stage of a clinical trial. The power, which is the probability to correctly reject the null hypothesis, is mainly influenced by the sample size of the study. Thus, a carefully conducted sample size determination will help controlling the producer's risk of falsely accepting the null hypothesis. As soon as initial assumptions on the design parameters are made, e.g. through a literature research, one calculates the required sample size to reject the null hypothesis with a certain amount of power, given a specific alternative.

Let us assume that the sample size of the control group is defined as a fraction of the test group sample size, i.e. $n_C = c_C n_T$ with $c_C > 0$. Usually, $c_C = 1$ is chosen as the balanced design has the largest power of all sample size allocations if the two treatments have equal variances. However, it is sometimes advisable to choose $c_C < 1$, i.e. $n_C < n_T$, in order to collect more safety data on the test treatment or to improve the patients' willingness to participate in the study. The test statistic T of the non-inferiority comparison is noncentral t-distributed with $n_T + n_C - 2$ degrees of freedom and noncentrality parameter $\frac{\mu_T - \mu_C + \Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T + n_C}}$. Thus, the power of the non-inferiority test is given as

$$\begin{aligned}1-\beta &= \mathrm{P}_{\mu_T,\mu_C}\left(T \geq t_{1-\alpha,n_T+n_C-2}\right)\\ &= 1-\mathscr{T}_{n_T+n_C-2}\left(t_{1-\alpha,n_T+n_C-2} \,\middle|\, \frac{\mu_T-\mu_C+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}}\right),\end{aligned}$$

where $\mathscr{T}_\nu(\cdot \mid \gamma)$ denotes the cumulative distribution function of the noncentral t-distribution with ν degrees of freedom and noncentrality parameter γ. With a predefined fraction c_C, the sample sizes n_T and n_C can be found as a solution of this equation. For large sample sizes, which are common for non-inferiority trials, the test statistic is approximately normal distributed with mean $\frac{\mu_T-\mu_C+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}}$ and variance 1. Thus, the power can be approximated by

$$1-\beta \approx \Phi\left(\frac{\mu_T-\mu_C+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}} - z_{1-\alpha}\right),$$

where $\Phi(\cdot)$ and z_γ denote the cumulative distribution function and the γ-quantile of the standard normal distribution, respectively. According to this, the approximate sample sizes of the test and the control group are obtained as

$$n_T = \frac{\left(z_{1-\alpha}+z_{1-\beta}\right)^2 \sigma^2 \left(1+\frac{1}{c_C}\right)}{\left(\mu_T-\mu_C+\Delta_{ni}\right)^2} \quad \text{and} \quad n_C = c_C n_T. \tag{1.3}$$

The sample size calculation of non-inferiority trials is usually carried out under the alternative of no treatment difference, i.e. $\mu_T = \mu_C$. However, if the true test treatment effect is less than that of the active comparator, the sample size calculated for equal treatment effects might be far too small. Furthermore, it becomes obvious that the non-inferiority margin has a strong influence on the sample size. For instance, assuming $\mu_T = \mu_C$ the sample sizes n_T and n_C get quadrupled if Δ_{ni} is cut in half. Equation (1.3) also illustrates why the sample size of a test vs. control non-inferiority comparison usually is much higher than that of a test vs. placebo superiority comparison. Assuming $\mu_T = \mu_C$, the divisor in the sample size formula of the non-inferiority trial is Δ_{ni}^2, whereas for the superiority trial the divisor is $(\mu_T - \mu_P)^2$. As the non-inferiority margin is often chosen as a fraction of the treatment difference $\mu_C - \mu_P$ $(= \mu_T - \mu_P)$ observed in former trials, this results in a much higher sample size in the non-inferiority trial. However, it should be noted that the choice of Δ_{ni} is not trivial and associated with several issues that need to be taken into consideration.

1.3 Methodological Problems

This section addresses methodological problems that can arise in the context of two-arm non-inferiority trials. Although there have been several publications regarding these issues (e.g. Röhmel, 1998; D'Agostino et al., 2003), practitioners still experience difficulties in such trials. For instance, there is often no agreement on the right choice of the non-inferiority margin, which is one of the most crucial points in the planning stage of a non-inferiority trial. Even though the right choice of Δ_{ni} highly depends on certain factors such as the respective medical indication, the regulatory authorities published several guidelines regarding non-inferiority trials (CPMP, 2000; CHMP, 2005b; FDA, 2010b) in order to give general guidance on this topic.

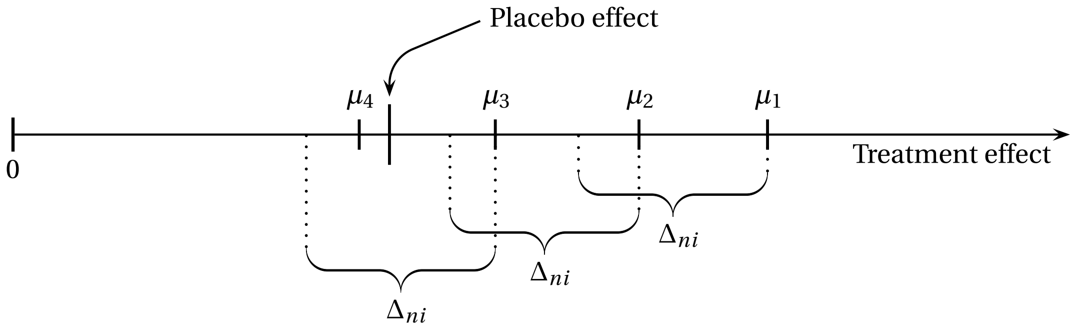
1.3.1 Choice of the Active Control Group

Once non-inferiority has been specified as the trial objective, e.g. for ethical reasons, the first question that arises is which active control group should be chosen. The huge importance of this step is reflected by the fact that the ICH E10 guideline on the *Choice of the Control Group and Related Issues in Clinical Trials* has been developed specifically for this purpose.

Potential control treatments are found through an extensive literature research and need to fulfil certain criteria. First of all, there needs to be sufficient historical evidence for the efficacy of the active control. Ideally, there exist several placebo-controlled superiority trials that consistently support the efficacy of the potential active comparator. If the active control has a volatile treatment effect, i.e. the results of the respective trials are inconsistent with one another, this could create problems in the interpretation of the non-inferiority trial. For instance, the test treatment could be non-inferior to the active comparator, although both treatments do not even outperform the placebo treatment with respect to efficacy. The so-called *constancy assumption* is sometimes questionable, as the placebo and the control treatment effect might have changed over time, e.g. due to improvements in the standard medical care. Thus, the study design and conduct of the non-inferiority trial should be as similar as possible to the historical placebo-controlled superiority trials, e.g. regarding the primary endpoint and the patient population, in order to increase the confidence in the results. However, even if the present and the historical trials are exactly the same with respect to trial conduct and study design, the constancy assumption might be of concern, as stated by Julious and Wang (2008). They found evidence that the placebo effect improves over time, a phenomenon called *placebo creep*, which might lead to a decrease of the drug effect. In addition, relative metrics for the primary endpoint such as the relative risk, or the risk ratio, should be preferred to absolute metrics, as they are more likely to be constant.

Furthermore, care should be taken to ensure that the chosen active comparator is the "best" available treatment on the market with respect to efficacy. Otherwise, a phenomenon called

Figure 1.3: Biocreep illustrated by a hypothetical example where progressively less effective active comparators are used in three consecutive non-inferiority trials.



biocreep might occur over the course of time, which is illustrated in Figure 1.3. Suppose treatment two is shown to be non-inferior to treatment one although it is slightly inferior ($\mu_1 - \Delta_{ni} < \mu_2 < \mu_1$). If treatment two is chosen as the active comparator for a slightly inferior third treatment in another non-inferiority trial and so on, the active controls could become inferior to placebo. Everson-Stewart and Emerson (2010) conducted an extensive simulation study in order to investigate which factors might lead to biocreep. They found that violations of the constancy assumption can lead to high rates of biocreep, although altogether biocreep was fairly rare.

1.3.2 Assay Sensitivity

Another critical point in two-arm non-inferiority trials is the assessment of *assay sensitivity*, which is "a property of a clinical trial defined as the ability to distinguish an effective treatment from a less effective or ineffective treatment" (ICH E10, Section 1.5). In a superiority trial comparing an experimental treatment with an active control, the rejection of the null hypothesis not only implies that there is a treatment difference but also that the study had sensitivity to detect it. In contrast, the rejection of the null hypothesis in a non-inferiority trial could arise from poor study quality as will be seen in the following and it is nearly impossible to assess the degree of assay sensitivity.

The ICH Expert Working Group (2000) mentions two factors that indicate if a two-arm non-inferiority trial has assay sensitivity. Firstly, there should be several similarly designed trials, i.e. same patient population, primary endpoint etc., that were able to demonstrate a treatment difference between two (or more) treatments (*Historical evidence of sensitivity to drug effects*). In addition, the similarity of the trial conduct with that of recent trials should be assessed afterwards to detect potential changes, e.g. of the study population. Note that the constancy assumption also highly depends on this similarity (cf. Section 1.3.1). Secondly, in contrast to superiority trials where poor study quality will bias the trial towards the null hypothesis of no treatment difference, *appropriate trial conduct* is even more important in two-arm non-inferiority trials. For instance, poor compliance could diminish the difference between the two treatments and a substantially inferior treatment might be declared as non-inferior, i.e. effective. Furthermore, there are several other potential reasons for a decrease of assay sensitivity, such as interferences with concomitant medications and a poorly responsive study population. In summary, high study quality and similarity to historical trials that showed sensitivity to drug effects can serve as evidence for assay sensitivity. However, without a present placebo group a lack of assay sensitivity can never be ruled out.

1.3.3 Choice of the Analysis Population

The right choice of the analysis population is another crucial point in two-arm non-inferiority trials that is closely related to the assessment of assay sensitivity. However, this issue has not been discussed in the literature to the same extent as the previously mentioned issues. In general, there exist two types of analysis sets, the full analysis set (FAS) on the one hand and the per protocol set (PPS) on the other.

The FAS is defined according to the intention-to-treat (ITT) principle mentioned in the ICH E9 guideline, where the treatment effect is evaluated on the basis of the intention to treat a patient instead of the actual treatment given. The FAS is as close as possible to this ideal and is generated by the set of all randomised patients with only minimal and justified eliminations defined prior to the trial. For instance, common reasons for an exclusion from the FAS are a failure to take at least one dose of study medication or violations of major inclusion criteria. The use of the FAS preserves the value of randomisation and, moreover, provides results that are more likely to reflect reality.

The PPS is derived as the subset of patients in the FAS who sufficiently complied with the study protocol and is often characterised by the following criteria: 1. Certain minimal exposure to the treatment; 2. Available measurements of the primary variable(s); 3. No major protocol violations such as violations of entry criteria. The main advantage of an analysis based on the PPS is the ability to estimate the drug's efficacy potential under optimal conditions. However, the exclusion of patients who do not adhere to the study protocol breaks the randomisation, so that a per protocol analysis can be biased considerably. Depending on the relationship between adherence to the study protocol and treatment or outcome, this bias can be in both directions. Most often, however, a per protocol (PP) analysis leads to over-optimistic estimates of the treatment effects, as 'problematic' patients tend to be excluded from the PPS.

In superiority trials the FAS is commonly accepted as the primary analysis set of choice, as an ITT analysis provides a conservative analysis approach. Non-compliers will generally diminish the difference between the two treatment groups, resulting in a bias towards the null hypothesis of no treatment difference. Thus, it will usually be more difficult to demonstrate superiority with an ITT analysis than based on the PPS. In contrast, there is still no consensus on the role of the FAS in non-inferiority trials. In the ICH E9 guideline it is stated that "in an equivalence or non-inferiority trial the use of the full analysis set is generally not conservative and its role should be reconsidered very carefully". This has often been mistakenly interpreted to mean that the PPS is a conservative choice and should be the primary analysis set in non-inferiority trials. However, analysing a non-inferiority study based on the PPS is not conservative per se, as e.g. major protocol deviations might be related to the treatment or outcome. For this reason the CPMP (2000) recommends that "in a non-inferiority trial, the full analysis set and the PP analysis set have equal importance and their use should lead to similar conclusions for a robust interpretation". But this strategy also not necessarily guarantees valid conclusions, as it was shown by Sanchez and Chen (2006). Their simulation study revealed that analyses based on the FAS and the PPS in non-inferiority studies can be both conservative and anti-conservative, depending on the types of protocol violations and missingness. They proposed that a so-called *hybrid ITT/PP* analysis, which excludes non-compliant patients as in a PP analysis and properly addresses the missing data as in an ITT analysis, would result in more reliable study results.

Some other interesting issues regarding the choice of the analysis set in non-inferiority trials have been mentioned by Wiens and Zhao (2007). They think that the justifications to use the FAS in superiority trials also carry over to non-inferiority trials. For instance, they argue that an analysis based on the FAS preserves the value of randomisation and estimates the "real-world" effectiveness. Moreover, the use of different analysis sets for superiority and non-inferiority comparisons could lead to inconsistencies. The question arises whether an α-adjustment is necessary for a subsequent superiority test, based on the FAS, after non-inferiority has been demonstrated, based on the PPS. The adequate handling of missing data is also closely related to the right choice of the analysis population, as it was also mentioned by Sanchez and Chen (2006). This topic has become more and more important in the recent years, but publications regarding this matter almost exclusively deal with the superiority objective. Yoo (2010) conducted a comprehensive simulation study to investigate the impact of different types of missingness on six different statistical analyses in a non-inferiority trial. It turned out that none of the six statistical methods uniformly outperformed the others in terms of controlling the type I error rate. Nevertheless, there is a need for further investigations on methods dealing with missing data in non-inferiority trials.

1.3.4 Choice of the Non-Inferiority Margin

Last but not least, the probably most critical step is to determine the non-inferiority margin Δ_{ni}. Due to the major importance of this step, suggestions on the right choice of Δ_{ni} are given in several regulatory guidelines (ICH Expert Working Group, 1998, 2000; CPMP, 2000). The FDA's *Guidance for Industry: Non-Inferiority Clinical Trials* (FDA, 2010b, Draft Version) also contains a whole chapter on choosing the non-inferiority margin and analysing the results of a non-inferiority trial. Moreover, the EMA published the *Guideline on the Choice of the Non-Inferiority Margin* that solely addresses this problem (CHMP, 2005b).

It seems obvious that the non-inferiority margin should be chosen in advance of the study and independently of the significance level. Furthermore, its choice should not depend on the power, because the extent of a clinically acceptable loss of efficacy does not alter with the sample size. There is, however, a general conflict of interest between the pharmaceutical companies and the regulatory authorities. The companies tend to choose a larger margin, whereas the authorities often want smaller margins. In the first instance, it is most important to clarify what exactly is the objective of the trial, because "demonstrating non-inferiority" is no sufficient trial objective. On the one hand, the main focus could be on the *indirect* comparison between the test treatment and placebo, i.e. an ordinary proof of efficacy. On the other hand, the *direct* comparison to the active comparator could be of major interest, e.g. to show that the treatment difference between test and control is negligible from a medical point of view. In practice, usually both the direct and the indirect comparison are of interest and the non-inferiority margin is chosen as a trade-off between the two. According to this, it is stated in the ICH E10 guideline that the choice of the non-inferiority margin should be "based on both statistical reasoning and clinical judgment". Statistical reasoning relates to the indirect comparison, where a margin is chosen based on the historical placebo-controlled trials in order that non-inferiority implies a proof of efficacy, whereas clinical judgement refers to the direct comparison, i.e. the margin is chosen as a clinically acceptable loss of efficacy. This is also in line with the FDA (2010b), who suggest to determine both a conservative estimate of the historical treatment difference between the active control and placebo and the largest clinically acceptable difference between the two treatments, with the non-inferiority margin being the smaller of these two values. In contrast to this, the choice of the equivalence margin in bioequivalence trials is entirely based on medical considerations, as the placebo effect is equal to zero and there is no need for an indirect comparison to placebo.

If the aim of the non-inferiority study is to demonstrate that the test treatment has an effect greater than zero, the margin is chosen on the basis of the historical trials comparing the active comparator with placebo. Ideally, there are more than one of these trials, so that the non-inferiority margin is determined through a meta-analysis. In order to account for uncertainties, the margin is usually chosen as the lower bound of the two-sided 95% confidence interval for the treatment difference between the active comparator and placebo $(\mu_C - \mu_P)$. As it is sometimes desired that the test treatment furthermore retains a specific amount of the control treatment effect, the margin is often chosen as a fraction of this lower bound, e.g. a preservation of at least 50% of the control treatment effect is a common choice in several medical indications. This further decrease of the margin also accounts for potential uncertainties, e.g. regarding the constancy assumption.

The approach described above is denoted as the *fixed margin approach* or the *95%-95% method*, where the first 95% refer to the confidence interval used to determine the margin and the second 95% refer to the confidence interval for the non-inferiority comparison between the test treatment and the active comparator. "Fixed" means that the non-inferiority margin is completely prespecified in advance by means of the historical data. The main advantage of this approach is that it provides a good basis for sample size calculations and the margin is clinically understandable. However, this method is rather conservative and obviously not statistically efficient, so that several other methods have been developed in the recent years to overcome these problems.

In contrast to the fixed margin approach, which is a two-step procedure, the other methods only consist of one step that tests either if the test treatment is superior to placebo or if a certain fraction of the control effect (relative to placebo) is retained by combining the historical and the current data as if they were from one randomised trial. According to this, no fixed non-inferiority margin is specified. As the data from all trials over time are "synthesised", these procedures are also denoted as *synthesis methods*. They might be statistically more efficient than the fixed margin approach, but the constancy assumption is even more crucial for them due to the combination of data from potentially quite different trials. Hung et al. (2009) and the relevant references cited therein give a good overview on the advantages and drawbacks of the two approaches. In general, the use of synthesis methods is limited in practice as medical considerations cannot be incorporated. Thus, the fixed margin approach is the method of choice, which is also reflected by the fact that, to the knowledge of the author, there have been no non-inferiority trials evaluated based on synthesis methods. However, even though the fixed margin approach is rather conservative, an indirect comparison to placebo remains critical due to uncertainties about crucial assumptions.

1.4 Summary

As we have seen, non-inferiority trials have become more and more important in the recent years, not least because of the increasing number of highly efficacious treatments on the market. In many cases it is not ethically justifiable to expose patients to placebo, so that active-controlled non-inferiority trials can serve as a substitute for placebo-controlled superiority trials. Besides the indirect comparison to placebo, i.e. the proof of efficacy, non-inferiority trials furthermore aim at a direct assessment of the similarity to the active comparator. As "demonstrating non-inferiority" is no sufficient trial objective, it is essential in a non-inferiority study to predefine the main objective, i.e. either the indirect proof of efficacy or the assessment of similarity to the control treatment.

The statistical methodology of two-arm non-inferiority trials is straightforward and interpretations are usually based on the corresponding two-sided confidence intervals. Especially switching to a superiority test, once non-inferiority has been demonstrated, is very intuitive using confidence intervals, also because there is no need for an adjustment of the α-level.

However, there are several methodological problems associated with the design and the analysis of two-arm non-inferiority trials. When choosing the active comparator, it is important to assess whether the constancy assumption might be violated. Differing results in the historical studies comparing the control with placebo may indicate that the active comparator has a volatile treatment effect, i.e. the constancy assumption does not hold. Furthermore, the assessment of assay sensitivity remains critical when there is no placebo group included in the trial. Otherwise, it can not be ruled out that the test treatment is demonstrated to be non-inferior to the active comparator, although both treatments are not even superior to placebo.

Thus, the three-arm design, including a test, an active control and a placebo group, is the design of choice for therapeutic indications where these critical assumptions are questionable. Common examples are the treatment of asthma, panic disorder or migraine, where the three-arm design is furthermore advocated in the respective regulatory guidelines of the CPMP (2003) and the CHMP (2005a, 2007a). In addition, the CHMP (2005b) states that a "three-armed trial with test, reference, and placebo allows some within-trial validation of the choice of non-inferiority margin and is therefore the recommended design; it should be used whenever possible.".

Finally, it should be noticed that there are also reasons to include an active comparator into a placebo-controlled superiority trial. For instance, superiority to placebo might be less meaningful if the standard treatment significantly outperforms the experimental treatment as mentioned by Koch and Röhmel (2004). Moreover, the preservation of a specific fraction of the control treatment effect can be adequately addressed in a three-arm trial.

CHAPTER 2

Three-Arm Non-Inferiority Trials

As we have seen in the previous chapter, the three-arm design, including a test treatment, an active control and a placebo, is recommended by the regulatory authorities and “it should be used whenever possible” (CHMP, 2005b). Therefore, it is also denoted as the “gold standard design” (Koch and Röhmel, 2004). In particular, it is the design of choice for medical indications where the constancy assumption and the assessment of assay sensitivity are critical, as e.g. in the treatment of depression.

This chapter gives an overview on the design and analysis of such three-arm non-inferiority trials and related statistical issues. As several hypotheses are of interest in three-arm trials, a short introduction to multiple testing is given in Section 2.1. The subsequent Section 2.2 gives a brief overview on a statistical approach for three-arm non-inferiority trials called the *effect retention approach*. This approach examines whether the test treatment preserves a specific amount of the control treatment effect relative to placebo. However, as the effect retention test has only rarely been used in practice, the main focus of this chapter (and this work) is on procedures with a fixed, prespecified non-inferiority margin Δ_{ni} (cf. Section 1.3.4) which are addressed in Section 2.3. Besides the statistical test procedure as well as power and sample size calculations, optimal sample size allocations are determined for the proposed design. The resulting optimal single-stage design should then serve as a benchmark for the group sequential and adaptive designs derived in the subsequent part of this work.

2.1 Multiple Testing

Obviously, three different comparisons can be of interest in three-arm trials , namely: test vs. placebo, control vs. placebo, test vs. control. Thus, we are confronted with a so-called *multiplicity problem*. Without adequately handling this problem the *familywise error rate* (FWER) might not be controlled, i.e. the probability of committing at least one type I error among the three comparisons mentioned above. In confirmatory clinical trials control of the FWER is of utmost importance and a fundamental prerequisite for study approval.

Multiplicity problems can also occur in several other situations, such as in clinical trials with multiple primary endpoints, interim analyses or subgroup analyses. Switching from non-inferiority to superiority in a two-arm trial as it was described in Section 1.2.1 is another example. In order to cover a large variety of situations the following introduction to multiple testing is formulated in a very general manner.

2.1.1 Motivation

To get an impression on how far the FWER gets inflated for a specific number of statistical tests without adjustment, let us consider the following hypothetical example. Suppose k null hypotheses $H_{0,1},\dots,H_{0,k}$ are tested with *k independent* local level α tests. Assuming the global null hypothesis that all individual null hypotheses $H_{0,1},\dots,H_{0,k}$ are true simultaneously, the probability to reject at least one of the null hypotheses is obtained as

$$\begin{aligned} P\left(\exists\, i \in \{1,\dots,k\} : H_{0,i} \text{ is rejected}\right) &= 1 - P\left(\forall\, i \in \{1,\dots,k\} : H_{0,i} \text{ is not rejected}\right) \\ &= 1 - \bigcap_{i=1}^{k} \underbrace{P\left(H_{0,i} \text{ is not rejected}\right)}_{=1-\alpha} \\ &= 1 - (1-\alpha)^k. \qquad (2.1) \end{aligned}$$

For a differing number of statistical tests at local level $\alpha = 0.05$, the probability of committing at least one type I error is shown in Table 2.1. As we can see, only one additional test at local level $\alpha = 0.05$ results in a doubling of the FWER under the global null hypothesis. Moreover, for $k = 14$ tests at local level $\alpha = 0.05$ the FWER even exceeds 50%, so that committing at least one type I error is more likely than making only correct decisions.

In the example above, the FWER was computed under the global null hypothesis that all individual null hypotheses $H_{0,1},\dots,H_{0,k}$ are true at the same time. Controlling the FWER under the global null hypothesis is also known as *weak control of the FWER*. Assuming that all individual null hypotheses are true simultaneously, however, is not realistic for most clinical trials. For instance, in a dose-finding trial comparing several doses of an experimental treatment with placebo, the treatment effects are very likely to differ across the dose levels. In this case, the FWER should be controlled for any configuration of true or false individual null hypotheses,

Table 2.1: Probability of committing at least one type I error among k independent tests at local level $\alpha = 0.05$, assuming that all corresponding null hypotheses $H_{0,1},\dots,H_{0,k}$ are true.

k	1	2	3	4	5	10	13	**14**	20	50
$1-(1-0.05)^k$	0.05	0.10	0.14	0.19	0.23	0.40	0.49	**0.51**	0.64	0.92

i.e. any possible scenario for the parameters of interest (e.g. the mean treatment differences between each dose and placebo). This is referred to as *strong control of the FWER* and the CPMP (2002) states in their *Points to Consider on Multiplicity Issues in Clinical Trials* that "control of the family-wise type I error in the strong sense [...] is a minimal prerequisite for confirmatory claims".

2.1.2 Šidák and Bonferroni Correction

A very simple approach to control the FWER in the strong sense can be derived by means of (2.1) resulting in local significance levels $1-\sqrt[k]{1-\alpha}$. This adjustment is called the *Šidák correction* (Šidák, 1967) and the proof of strong FWER control is straightforward but assumes *independent* local tests. Suppose, the null hypotheses $H_{0,1},\ldots,H_{0,k}$ are simultaneously tested using the Šidák correction. Let further $P_1,\ldots,P_k$ denote the corresponding *independent* random p-values and let I_0 be the set of indices of all true null hypotheses, so that obviously $|I_0| \leq k$. Then, it follows that

$$\begin{aligned} P\left(\bigcup_{i\in I_0}\left\{P_i \leq 1-\sqrt[k]{1-\alpha}\right\}\right) &= 1-P\left(\bigcap_{i\in I_0}\left\{P_i > 1-\sqrt[k]{1-\alpha}\right\}\right) \\ &= 1-\bigcap_{i\in I_0}\underbrace{P\left(P_i > 1-\sqrt[k]{1-\alpha}\right)}_{=\sqrt[k]{1-\alpha}} = 1-(1-\alpha)^{|I_0|/k} \leq \alpha. \end{aligned}$$

As this holds for any configuration of true and false null hypotheses, the Šidák method controls the FWER in the strong sense (assuming independence). In clinical trials, the assumption of independence often does not hold, e.g. multiple primary endpoints are almost always correlated. However, it can be shown that the Šidák method provides strong FWER control if the corresponding test statistics are positively orthant dependent which holds for many common testing situations (Holland and DiPonzio Copenhaver, 1987).

Without further assumptions on the corresponding test statistics, Dunn (1961) provided a simple but conservative approach for strong FWER control by dividing the local significance levels by the number of statistical tests performed, i.e. local significance levels α/k. The proof of strong FWER control is also straightforward and applies the *Bonferroni inequality* wherefore this approach is generally known as the *Bonferroni correction.* For all possible combinations of true and false null hypotheses the FWER is obtained as

$$P\left(\bigcup_{i\in I_0}\left\{P_i \leq \frac{\alpha}{k}\right\}\right) \leq \sum_{i\in I_0} P\left(P_i \leq \frac{\alpha}{k}\right) = |I_0|\frac{\alpha}{k} \leq k\frac{\alpha}{k} = \alpha. \tag{2.2}$$

Thus the Bonferroni correction provides a simple general approach for strong FWER control, even though it is very conservative in most situations. As $\alpha/k < 1-\sqrt[k]{1-\alpha}$ holds for $k \geq 2$, the Bonferroni method is less powerful than the Šidák method, but the loss in power is negligible.

A generalisation of the Bonferroni method arises when the overall significance level α is not equally distributed among the individual tests, reflecting differing importance of the respective null hypotheses. Suppose that weights $w_i \geq 0, i = 1, ..., k$, are chosen such that $\sum_{i=1}^{k} w_i = 1$ and the local significance levels are defined as $\alpha_i = w_i \alpha$. Then it can be shown analogously to (2.2) that the so-called *weighted Bonferroni method* controls the FWER in the strong sense. This allows to increase the probability to reject more important null hypotheses ($w_i > \alpha/k$), but in turn the rejection probability of less important null hypotheses ($w_i < \alpha/k$) decreases.

2.1.3 Dunnett's Test

In certain situations there is a precise knowledge of the dependencies between the respective individual test statistics. For instance, in *many-to-one* comparisons where several treatments are compared with a control, the joint distribution of the corresponding test statistics is known. Dunnett (1955) exploited this knowledge to derive a test that is uniformly more powerful than the Bonferroni test in case of a many-to-one comparison.

Let us consider Dunnett's approach for simultaneously comparing two treatments (denoted with 1 and 2) with a control (denoted with 0). Let $X_{i,j} \sim N(\mu_i, \sigma^2)$, $j = 1, ..., n_i$, $i = 0, 1, 2$, be the independent observations of the three treatment groups with common but unknown variance σ^2. Let further $N = \sum_{i=0}^{2} n_i$ denote the overall sample size. Without loss of generality larger values of $X_{i,j}$ are assumed to be desirable. The individual one-sided null hypotheses and corresponding tests statistics of the two comparisons are given as

$$H_{0,i}: \theta_i = \mu_i - \mu_0 \leq 0 \quad \text{and} \quad T_i = \frac{\bar{X}_i - \bar{X}_0}{\hat{\sigma}} \sqrt{\frac{n_i n_0}{n_i + n_0}}, \quad i = 1, 2,$$

with $\bar{X}_i = 1/n_i \sum_{j=1}^{n_i} X_{i,j}, i = 0, 1, 2$ and $\hat{\sigma}^2 = \sum_{i=0}^{2} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2 / (N-3)$ denoting the sample means, the unbiased sample variances and the unbiased pooled estimator of σ^2, respectively.

Let us first derive a critical value d_α so that the FWER under the global null hypothesis $H_0 = H_{0,1} \cap H_{0,2}$ equals α, i.e. weak FWER control. d_α can thus be derived as a solution of

$$P_{\boldsymbol{\theta}=\mathbf{0}}(\{T_1 \leq d_\alpha\} \cap \{T_2 \leq d_\alpha\}) = 1 - \alpha, \tag{2.3}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu_1 - \mu_0, \mu_2 - \mu_0)'$ and $\mathbf{0} = (0, 0)'$. Note that all vectors and matrices shall be indicated in bold in the following. The vector of test statistics $(T_1, T_2)'$ can be written as

$$(T_1, T_2)' = \frac{\left(\frac{\bar{X}_1 - \bar{X}_0}{\sigma} \sqrt{\frac{n_1 n_0}{n_1 + n_0}}, \frac{\bar{X}_2 - \bar{X}_0}{\sigma} \sqrt{\frac{n_2 n_0}{n_2 + n_0}}\right)'}{\hat{\sigma}/\sigma} =: \frac{\boldsymbol{Z}}{V}. \tag{2.4}$$

The numerator $\boldsymbol{Z}$ obviously follows a bivariate normal distribution with mean vector $\boldsymbol{\theta}^*$ and variance-covariance matrix $\boldsymbol{\Sigma}^*$ given as

$$\boldsymbol{\theta}^* = \left(\frac{\theta_1}{\sigma} \sqrt{\frac{n_1 n_0}{n_1 + n_0}}, \frac{\theta_2}{\sigma} \sqrt{\frac{n_2 n_0}{n_2 + n_0}}\right)' \quad \text{and}$$

$$\mathbf{\Sigma}^* = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{with} \quad \rho = \sqrt{\frac{1}{\left(1+\frac{n_0}{n_1}\right)\left(1+\frac{n_0}{n_2}\right)}}. \tag{2.5}$$

Interestingly, the correlation ρ between the two test statistics merely depends on the allocation ratios between the three treatment groups, namely n_0/n_1 and n_0/n_2. As further $V^2 \sim \chi^2_{N-3}/(N-3)$ and $\boldsymbol{Z}$ and V are independent it follows from (2.4) that $(T_1, T_2)'$ is bivariate noncentral t-distributed with $N-3$ degrees of freedom, vector of noncentrality parameters $\boldsymbol{\theta}^*$ and variance-covariance matrix $\mathbf{\Sigma}^*$ given in (2.5). By means of the distribution function of the bivariate noncentral t-distribution the left-hand side of Equation (2.3) can be calculated numerically and the critical value d_α is determined as a solution of this equation (depending on α, N and the ratios n_0/n_1 and n_0/n_2).

Under the global null hypothesis, the FWER of the Bonferroni method using the critical value $t_{\alpha/2,N-3}$ is less than (and not equal to) α, whereas the unadjusted test procedure with critical value $t_{\alpha,N-3}$ has a FWER greater than α. Thus, we have the relationship

$$P_{\boldsymbol{\theta}=\mathbf{0}}\left(\bigcap_{i=1}^{2}\{T_i \le t_{\alpha,N-3}\}\right) < \underbrace{P_{\boldsymbol{\theta}=\mathbf{0}}\left(\bigcap_{i=1}^{2}\{T_i \le d_\alpha\}\right)}_{=1-\alpha} < P_{\boldsymbol{\theta}=\mathbf{0}}\left(\bigcap_{i=1}^{2}\{T_i \le t_{\alpha/2,N-3}\}\right). \tag{2.6}$$

As the distribution function of the bivariate t-distribution at (x,x) is strictly increasing in x, it follows from (2.6) that

$$t_{\alpha,N-3} < d_\alpha < t_{\alpha/2,N-3}. \tag{2.7}$$

According to the left inequality in (2.7) we obtain

$$\forall i = 1,2: \; P_{\theta_i=0}(T_i \ge d_\alpha) < P_{\theta_i=0}\left(T_i \ge t_{\alpha,N-3}\right) = \alpha,$$

so that Dunnett's test controls the FWER also in the strong sense. The right inequality in (2.7) shows that Dunnett's test always rejects more null hypotheses than the Bonferroni test, i.e. Dunnett's test is more powerful. A generalisation of Dunnett's test to more than three treatment groups is straightforward (see Dunnett, 1955).

2.1.4 Fixed Sequence Procedure

In many clinical studies there is a natural hierarchy between the corresponding null hypotheses to be tested, e.g. in studies with multiple endpoints some might be clinically more important than others. In such cases the *fixed sequence procedure* (Maurer et al., 1995) provides a simple but powerful approach to deal with the multiplicity issue. It assumes that the null hypotheses $H_{0,1},\ldots,H_{0,k}$ are a-priori ordered in such a way that the lower the index the more important is the corresponding null hypothesis, i.e. if $i < j$ then $H_{0,i}$ is more important than $H_{0,j}$. The procedure starts with testing the first null hypothesis $H_{0,1}$ at level α. If it is rejected, the second null hypothesis $H_{0,2}$ is tested at level α and so on. The procedure stops as soon as an acceptance occurs or all null hypotheses have been rejected.

In order to show that the procedure provides strong FWER control let $k^* = \min_{i \in I_0}(i)$ denote the lowest index of all true null hypotheses, where I_0 is the set of indices of all true null hypotheses. Then obviously

$$P\left(\exists\, i \in I_0 : H_{0,i} \text{ is rejected}\right) \leq P\left(H_{0,k^*} \text{ is rejected}\right) = \alpha,$$

because at least H_{0,k^*} has to be rejected in order to reject any true null hypothesis and hence commit a type I error. In this context $H_{0,i-1}$ is also called a *gatekeeper* for the subsequent null hypothesis $H_{0,i}$, as $H_{0,i-1}$ has to be passed in order to test $H_{0,i}$.

2.1.5 Closed Testing Procedure

The *closed testing procedure* proposed by Marcus et al. (1976) plays a key role in the field of multiple testing procedures and nearly all multiple testing methods applied in clinical trials can be seen as an application of the corresponding *closure principle*.

Suppose k individual null hypotheses $H_{0,1},\ldots,H_{0,k}$ should be simultaneously assessed by means of the closed testing principle. First, all possible $2^k - 1$ intersections of the individual null hypotheses need to be determined with valid local level α tests for each intersection hypothesis. Then an individual null hypothesis $H_{0,i}$ is rejected if and only if all intersection hypotheses containing $H_{0,i}$ are rejected by their local level α tests. Marcus et al. (1976) showed that multiple testing procedures derived by this principle control the FWER in the strong sense at level α. It is easy to see that strong FWER control is provided if we keep in mind that *strong* means *for any configuration of true and false null hypotheses.* This control is ensured by considering all possible combinations of the individual null hypotheses $H_{0,1},\ldots,H_{0,k}$ with local level α tests.

It should be noted that for larger numbers k of individual null hypotheses the closed testing principle can get computationally intensive, as the number of intersection hypotheses $2^k - 1$ becomes very large. Thus so-called *shortcut procedures* have become popular that substantially decrease the number of computational steps (see e.g. Grechanovsky and Hochberg, 1999).

2.1.6 Further Developments

Let us finally give a short overview on further developments in the field of multiple testing methodology. Besides the already mentioned fixed sequence procedure there have been proposed several other so-called *sequentially rejective multiple test procedures*, such as step-down and step-up tests. In such step-wise tests, in contrast to fixed sequence procedures, the individual p-values are ordered *a posteriori* dependent on the observed data. Step-wise versions of the Bonferroni test have been proposed by Holm (1979) and Hochberg (1988), and for Dunnett's test by Naik (1975), Marcus et al. (1976) and Dunnett and Tamhane (1992). These procedures turn out to be uniformly more powerful than the standard versions of Bonferroni and Dunnett's test, respectively.

Often there are logical interrelations between the individual null hypotheses under consideration, e.g. a specific set of null hypotheses can not be true at the same time. Popper Shaffer (1986) described a general procedure to incorporate such logical restrictions. In most applications with restricted combinations the proposed method can lead to a substantial increase in power. The *fallback procedure* proposed by Wiens (2003) is a more flexible approach that combines the weighted Bonferroni method and the fixed sequence procedure and ensures strong control of the FWER. A more general approach for multiple testing problems with logical relations, e.g. a natural hierarchy between the hypotheses, are so-called *gatekeeping procedures*, where the null hypotheses are divided into families of hypotheses. For further information on gatekeeping procedures see Dmitrienko and Tamhane (2007). Recently, Bretz et al. (2009) proposed a general graphical approach to sequentially rejective multiple testing procedures, including gatekeeping/fallback procedures and fixed sequence tests. Their approach is based on directed, weighted graphs and, due to its simplicity, well-suited for communicating complex multiple testing procedures to non-statisticians.

As confidence intervals are of major interest especially in confirmatory clinical trials, there is a need for appropriately adjusted versions. Simultaneous confidence intervals for a variety of multiple testing procedures have been proposed by Strassburger and Bretz (2008) and Guilbaud (2008).

As we have seen, the field of multiple testing methodology is an extensively growing area of research and there is also a variety of books covering this topic (Hochberg and Tamhane, 1987; Westfall and Young, 1993; Hsu, 1996; Dmitrienko et al., 2009). However, up-to-date regulatory guidelines regarding the issues of multiplicity adjustment in clinical trials are urgently needed, as the only guideline by the CPMP (2002) is more than 10 years old. This is reflected by the recently published *Concept paper on the need for a guideline on multiplicity issues in clinical trials* (CHMP, 2012). The FDA is also currently working on an appropriate guideline which is expected to be published soon.

2.2 Effect Retention Approach

The previous section showed that there is a variety of procedures to choose from in order to deal with the multiplicity issue in non-inferiority trials with an additional placebo arm. First of all, the similarities with trials comparing several treatments with a control are conspicuous. In three-arm trials the correlation of the corresponding test statistics is also known, but obviously there is a natural hierarchy between the three individual hypotheses. For instance, non-inferiority of the test treatment to the active control is meaningless without a proof of efficacy for the test treatment, i.e. test is superior to placebo. Therefore, fixed sequence testing is the basis for statistical testing procedures in three-arm non-inferiority trials.

One of the early contributions to the design and analysis of three-arm trials was by Koch and Tangen (1999), who proposed the so-called *effect retention approach* aiming to show that the experimental treatment preserves a certain fraction of the control treatment effect relative to placebo. Therefore, it is sometimes also referred to as the *effect preservation test.* The approach was extensively studied by Pigeot et al. (2003) for the case of normally distributed endpoints with a common but unknown variance. The corresponding hierarchically ordered null hypotheses of the effect retention approach are

$$\begin{aligned} \widetilde{H}_{01} &: \mu_C - \mu_P \leq 0, \\ \widetilde{H}_{02} &: \frac{\mu_T - \mu_P}{\mu_C - \mu_P} \leq f, \end{aligned} \tag{2.8}$$

where f, $0 \leq f \leq 1$, is a prespecified constant and μ_T, μ_C, μ_P denote the mean treatment effect under test, control and placebo, respectively. Obviously, the effect retention hypothesis $\widetilde{H}_{02}$ can be derived from the common non-inferiority hypothesis $H_{0,ni}: \mu_T - \mu_C \leq -\Delta_{ni}$ (cf. Section 1.2.1) by defining $\Delta_{ni} = (1-f)(\mu_C - \mu_P)$, i.e. Δ_{ni} is a fraction of the difference between the control and the placebo effect. Besides the statistical test procedure Pigeot et al. (2003) further derived an optimal sample size allocation and a confidence interval for the parameter f.

The method was further extended to binary outcomes (Tang and Tang, 2004; Kieser and Friede, 2007), heterogeneous variances (Hasler et al., 2008; Dette et al., 2009; Gamalo et al., 2011) and exponentially distributed endpoints (Mielke et al., 2008). In order to deal with weak knowledge of the active control effect, Schwartz and Denne (2006) proposed a two-stage sample size recalculation procedure based on an internal pilot, whereas Li and Gao (2010) suggested a group sequential type design to address uncertainties in the placebo response rate. Besides, Munzel (2009) derived a nonparametric design and Gosh et al. (2011) proposed a Bayesian design to incorporate prior information gained for instance from the historical active control trials.

As already mentioned in Section 1.4 it is also reasonable in many situations to include an active control arm into a placebo-controlled superiority trial. For instance, superiority to placebo might be less meaningful if the standard treatment significantly outperforms the experimental treatment (Koch and Röhmel, 2004). In this context, Hauschke and Pigeot (2005a) stated that the inclusion of a reference arm could serve to investigate the clinical relevance of the test treatment effect, and they adapted the effect retention design for this purpose.

There is a disagreement amongst authors on which hypotheses should be tested in the gold standard design and especially on their hierarchical order (Lewis, 2005; Röhmel, 2005a; Koch, 2005; Mehrotra, 2005; Hung, 2005; Hauschke and Pigeot, 2005b). In general, the direct proof of efficacy for the test treatment ($H_{0,sup}: \mu_T \leq \mu_P$) seems to be the most important step in a three-arm trial. Whether $\widetilde{H}_{01}$ needs to be rejected or not mostly depends on the particular medical setting. However, rejection of $\widetilde{H}_{01}$ is a technical prerequisite for the effect retention test in order to be sure not to divide by zero. An approach to simultaneously deal with the contrasts $\mu_T - \mu_P$

and $\mu_C - \mu_P$ was presented by Röhmel (2005b), who constructed confidence ellipsoids.

As we can see, the design and analysis of the effect preservation approach has been extensively studied in the recent years. Apparently, the approach is very appealing, especially from a theoretical viewpoint. However, to the knowledge of the author, it has not yet been applied for assessing non-inferiority in confirmatory clinical trials. A reason for this might be that the CHMP (2005b) explicitly states that "a noninferiority trial aims to demonstrate that the test product is not worse than the comparator by more than a prespecified, small amount.". The following section goes into more detail on such fixed margin designs that build the basis for the rest of this work.

2.3 Fixed Margin Approach

In contrast to the large number of publications on the effect retention approach there have been comparatively few studies on the design and analysis of three-arm non-inferiority trials with a fixed, prespecified non-inferiority margin Δ_{ni}. The first contribution by Koch and Röhmel (2004) was further modified and extended by Röhmel and Pigeot (2010) resulting in a more powerful procedure. Hida and Tango (2011a) proposed a slightly different approach leading to some critical debate (Röhmel and Pigeot, 2011; Hida and Tango, 2011b). Their approach has been further extended to three-arm trials with multiple new treatments by Kwong et al. (2012). Recently, a general approach to sample size calculations for the 'gold standard' design has been proposed by Stucke and Kieser (2012).

Let us now go into more detail on the procedure proposed by Koch and Röhmel (2004) that forms the basis for the following deliberations presented in this work. After introducing the procedure, the overall power is derived and optimal sample size allocations are calculated according to Stucke and Kieser (2012). This optimal single-stage design should then serve as a benchmark for the designs derived in the subsequent part of this work.

2.3.1 Statistical Model and Test Procedure

Assume that the primary endpoints under the test, control and placebo treatment are mutually independent and normally distributed with common, but unknown variance σ^2, i.e. $X_{T,i} \sim N(\mu_T, \sigma^2)$, $i = 1,2,...,n_T$, $X_{C,i} \sim N(\mu_C, \sigma^2)$, $i = 1,2,...,n_C$, and $X_{P,i} \sim N(\mu_P, \sigma^2)$, $i = 1,2,...,n_P$. As in Chapter 1 it is assumed, without loss of generality, that larger treatment effects are associated with greater benefits and thus are desired. The following two hierarchically ordered sets of hypotheses proposed by Koch and Röhmel (2004) are considered

$$\begin{array}{lrcl} 1. & H_{0,TP}^{(s)}: \mu_T \leq \mu_P & \text{vs.} & H_{1,TP}^{(s)}: \mu_T > \mu_P, \\ 2. & H_{0,TC}^{(n)}: \mu_T \leq \mu_C - \Delta_{ni} & \text{vs.} & H_{1,TC}^{(n)}: \mu_T > \mu_C - \Delta_{ni}. \end{array} \tag{2.9}$$

The superiority comparison between test and placebo acts as a gatekeeper for the subsequent non-inferiority comparison between test and control, hence both comparisons are tested at full level α (cf. Section 2.1.4).

As already mentioned there is a critical debate on the set of hypotheses that should be tested in a three-arm trial. The two hypotheses $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ certainly are of main interest, but in certain situations other hypotheses might also be relevant, e.g. the comparison between control and placebo. When further hypotheses are included in the testing procedure, one can make use of the existing logical interrelations.

For instance, as proposed by Koch and Röhmel (2004), the hypotheses $H_{0,CP}^{(s)} : \mu_C \leq \mu_P$ and $H_{0,TC}^{(s)} : \mu_T \leq \mu_C$ could be added as a third step of the testing procedure, both tested at level α. The reason for this is simple: if $H_{0,TP}^{(s)}$ is false, i.e. $\mu_T > \mu_P$, $H_{0,CP}^{(s)}$ and $H_{0,TC}^{(s)}$ cannot be true at the same time. With the same argument, Röhmel and Pigeot (2011) showed that once $H_{0,TP}^{(s)}$ has been rejected, $H_{0,TC}^{(n)}$ and $H_{0,CP}^{(s)}$ can be assessed simultaneously without adjusting the α-levels. In addition, they showed that further confirmatory testing is possible without α-adjustment leading to sharper test decisions. For instance, if $H_{0,TP}^{(s)}$ has been rejected, one can simultaneously assess whether the control treatment is superior to placebo and whether the test treatment is non-inferior to the control with any non-inferiority margin $\delta \in [0, \Delta_{ni}]$, i.e. even test for superiority of the test treatment over the control.

For specific medical indications, inclusion of $H_{0,CP}^{(s)}$ might be mandatory as mentioned by Hauschke and Pigeot (2005b). For example, in trials regarding the treatment of mild persistent asthma, failure to demonstrate the efficacy of the active control (usually a corticosteroid) will challenge the whole study quality and even the superiority of the test treatment over placebo. Nevertheless, the focus is on $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ as these are the fundamental hypotheses in three-arm non-inferiority trials. $H_{0,CP}^{(s)}$ can be easily included into the procedure according to Röhmel and Pigeot (2010), however, this should be prespecified in the study protocol.

The corresponding test statistics for the null hypotheses $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ are the common Student's t-test statistics

$$\begin{aligned} T_{TP}^{(s)} &= \frac{\bar{X}_T - \bar{X}_P}{\hat{\sigma}} \sqrt{\frac{n_T n_P}{n_T + n_P}}, \\ T_{TC}^{(n)} &= \frac{\bar{X}_T - \bar{X}_C - \Delta_{ni}}{\hat{\sigma}} \sqrt{\frac{n_T n_C}{n_T + n_C}}, \end{aligned} \tag{2.10}$$

where $\bar{X}_T = \frac{1}{n_T} \sum_{i=1}^{n_T} X_{T,i}$ and $\bar{X}_C$ and $\bar{X}_P$ are defined analogously. The common variance σ^2 is estimated by the unbiased pooled estimator $\hat{\sigma}^2 = ((n_T - 1)S_T^2 + (n_C - 1)S_C^2 + (n_P - 1)S_P^2)/(n_T + n_C + n_P - 3)$, where S_T^2, S_C^2 and S_P^2 denote the sample variances of the test, control and placebo group, respectively. Obviously, the test statistics $T_{TP}^{(s)}$ and $T_{TC}^{(n)}$ both follow a noncentral t-distribution with $\nu = N - 3$ degrees of freedom, where $N = n_T + n_C + n_P$ denotes the overall sample size. Thus, the test statistics are compared to $t_{1-\alpha,\nu}$, the $(1-\alpha)$-quantile of the t-distribution with ν degrees of freedom. Altogether, there are three possible outcomes depending on the observed test statistics

$$\begin{aligned} T_{TP}^{(s)} < t_{1-\alpha,\nu} \quad &\Rightarrow \quad \text{Reject neither } H_{0,TP}^{(s)} \text{ nor } H_{0,TC}^{(n)}\,, \\ T_{TP}^{(s)} \geq t_{1-\alpha,\nu} \text{ and } T_{TC}^{(n)} < t_{1-\alpha,\nu} \quad &\Rightarrow \quad \text{Reject } H_{0,TP}^{(s)}\,, \\ T_{TP}^{(s)} \geq t_{1-\alpha,\nu} \text{ and } T_{TC}^{(n)} \geq t_{1-\alpha,\nu} \quad &\Rightarrow \quad \text{Reject } H_{0,TP}^{(s)} \text{ and } H_{0,TC}^{(n)}\,. \end{aligned} \tag{2.11}$$

According to Maurer et al. (1995) this fixed sequence procedure controls the FWER in the strong sense by α (see also Section 2.1.4).

Because of the duality between confidence intervals and statistical tests, one can equivalently calculate the corresponding one-sided $(1-\alpha)$ confidence intervals for the differences $\theta_{TP} = \mu_T - \mu_P$ and $\theta_{TC} = \mu_T - \mu_C$. The hypothesis $H_{0,TP}^{(s)}$ is then rejected if the lower bound of the confidence interval for θ_{TP} lies above or is equal to zero, and $H_{0,TC}^{(n)}$ is rejected if additionally the lower bound of the confidence interval for θ_{TC} is greater than or equal to $-\Delta_{ni}$. However, it should be stated that these are no simultaneous confidence intervals, see e.g. Strassburger and Bretz (2008) and Guilbaud (2008).

2.3.2 Power and Sample Size

Let us now investigate the sample sizes and the overall power $1-\beta$ of the procedure, i.e. the probability to correctly reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$. By means of the Bonferroni inequality, a lower bound for $1-\beta$ can be determined as follows:

$$\begin{aligned} 1-\beta &= P_{\theta_{TP},\theta_{TC}}\left(\left\{T_{TP}^{(s)} \geq t_{1-\alpha,\nu}\right\} \cap \left\{T_{TC}^{(n)} \geq t_{1-\alpha,\nu}\right\}\right) \\ &= 1 - P_{\theta_{TP},\theta_{TC}}\left(\left\{T_{TP}^{(s)} \leq t_{1-\alpha,\nu}\right\} \cup \left\{T_{TC}^{(n)} \leq t_{1-\alpha,\nu}\right\}\right) \\ &\geq 1 - P_{\theta_{TP}}\left(T_{TP}^{(s)} \leq t_{1-\alpha,\nu}\right) - P_{\theta_{TC}}\left(T_{TC}^{(n)} \leq t_{1-\alpha,\nu}\right) = 1 - \beta_{TP} - \beta_{TC}, \end{aligned}$$

where β_{TP} and β_{TC} denote the probabilities of committing a type II error for the superiority comparison between test and placebo and for the non-inferiority comparison of test versus control, respectively.

According to this, a commonly applied *ad hoc* approach to calculate the sample sizes for the proposed procedure takes the following form. First of all, β_{TP} and β_{TC} need to be prespecified with $\beta_{TP} + \beta_{TC} = \beta$ (usually $\beta_{TP} = \beta_{TC}$). Then, the sample sizes of the test and control group (often $n_T = n_C$) are determined to achieve a power of at least $1-\beta_{TC}$ for the non-inferiority comparison, e.g. with the approximate formula in (1.3). Finally, the placebo group size n_P is determined to obtain a power of at least $1-\beta_{TP}$ for the superiority comparison. The overall power of the procedure will then be at least $1-\beta_{TP}-\beta_{TC}$.

This *ad hoc* approach seems very appealing through its simplicity, but it turns out that the overall power $1-\beta$ can also be directly computed. Therefore it might be helpful to notice that the procedure has some similarities with Dunnett's test described in Section 2.1.3, despite the fact that Dunnett's test is based on the union-intersection instead of the intersection-union principle. Analogously, it will now be shown that $(T_{TP}^{(s)}, T_{TC}^{(n)})'$ is bivariate noncentral t-distributed. First of all, the vector of test statistics can be written as

$$\left(T_{TP}^{(s)}, T_{TC}^{(n)}\right)' = \frac{\left(\frac{\bar{X}_T-\bar{X}_P}{\sigma}\sqrt{\frac{n_T n_P}{n_T+n_P}}, \frac{\bar{X}_T-\bar{X}_C+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}}\right)'}{\hat{\sigma}/\sigma} = \frac{\boldsymbol{Z}}{V}.$$

Obviously, the vector $\boldsymbol{Z}$ is bivariate normally distributed, as all linear combinations of the two components also follow a normal distribution (each component itself is a linear combination of normally distributed random variables). As the group means $\bar{X}_T, \bar{X}_C$ and $\bar{X}_P$ are mutually independent, the covariance of the two components of $\boldsymbol{Z}$ is obtained as

$$\begin{aligned} Cov\left(\frac{\bar{X}_T-\bar{X}_P}{\sigma}\sqrt{\frac{n_T n_P}{n_T+n_P}}, \frac{\bar{X}_T-\bar{X}_C+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}}\right) &= \frac{1}{\sigma^2}\sqrt{\frac{n_T n_P}{n_T+n_P}}\sqrt{\frac{n_T n_C}{n_T+n_C}}\underbrace{\mathrm{Var}\left(\bar{X}_T\right)}_{=\frac{\sigma^2}{n_T}} \\ &= \sqrt{\frac{n_C n_P}{(n_T+n_C)(n_T+n_P)}}. \end{aligned}$$

As further $\boldsymbol{Z}$ and $V^2 \sim \chi^2_\nu/\nu$ are independent, it follows that $(T_{TP}^{(s)}, T_{TC}^{(n)})'$ is bivariate noncentral t-distributed with $\nu = n_T + n_C + n_P - 3$ degrees of freedom and vector of noncentrality parameters $\boldsymbol{\theta}^*$ and variance-covariance matrix $\boldsymbol{\Sigma}$ given as

$$\begin{aligned} \boldsymbol{\theta}^* &= \left(\frac{\theta_{TP}}{\sigma}\sqrt{\frac{n_T n_P}{n_T+n_P}}, \frac{\theta_{TC}+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}}\right)', \\ \boldsymbol{\Sigma} &= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ with } \rho = \sqrt{\frac{n_C n_P}{(n_T+n_C)(n_T+n_P)}}. \end{aligned} \tag{2.12}$$

Thus, the overall power is obtained as

$$1-\beta = \mathcal{T}^{\boldsymbol{\Sigma}}_{n_T+n_C+n_P-3}\left(-t_{1-\alpha,n_T+n_C+n_P-3}, -t_{1-\alpha,n_T+n_C+n_P-3} \,\middle|\, -\boldsymbol{\theta}^*\right), \tag{2.13}$$

where $\mathcal{T}^{\boldsymbol{\Sigma}}_\nu(\cdot \mid \boldsymbol{\gamma})$ denotes the cumulative distribution function of the bivariate noncentral t-distribution with ν degrees of freedom, vector of noncentrality parameters $\boldsymbol{\gamma}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

For large sample sizes n_T, n_C and n_P, which are common for non-inferiority trials, the vector of test statistics is asymptotically bivariate normally distributed, i.e. $(T_{TP}^{(s)}, T_{TC}^{(n)})' \sim N_2(\boldsymbol{\theta}^*, \boldsymbol{\Sigma})$ holds approximately. Thus, the overall power can be approximated by

$$1-\beta \approx \Phi^{\boldsymbol{\Sigma}}\left(\frac{\theta_{TP}}{\sigma}\sqrt{\frac{n_T n_P}{n_T+n_P}} - z_{1-\alpha}, \frac{\theta_{TC}+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T n_C}{n_T+n_C}} - z_{1-\alpha}\right), \tag{2.14}$$

where $\Phi^{\boldsymbol{\Sigma}}(\cdot)$ denotes the cumulative distribution function of the bivariate normal distribution with mean vector $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Calculating the right-hand sides of Equations (2.13) or (2.14) turns out to be easy since nowadays many statistical software packages include functions for the cumulative distribution functions of the multivariate normal and t-distribution. For instance, the software environment R (R Core Team, 2013) provides two packages for this purpose, namely `mvtnorm` (Genz et al., 2012) and `mnormt` (Genz and Azzalani, 2012).

By defining the sample sizes of the control and placebo group as fractions of the test group sample size, i.e. $n_C = c_C n_T$ and $n_P = c_P n_T$ with $c_C, c_P > 0$ (cf. Section 1.2.2), the exact and approximate sample sizes can be numerically determined by solving Equation (2.13) and (2.14) for n_T, respectively. Therefore, the non-inferiority margin Δ_{ni}, the (one-sided) significance level α, the overall power $1-\beta$ and the allocation ratios c_C and c_P need to be specified. Furthermore, assumptions on the treatment differences θ_{TP} and θ_{TC} and the standard deviation σ need to be made.

Let us now investigate the validity of the normal approximation in (2.14). For this purpose the standard deviation σ is expressed as a fraction of the treatment difference between control and placebo, that is $\sigma = \epsilon(\mu_C - \mu_P)$, $\epsilon > 0$. It is further assumed that $\mu_T = \mu_C$, which is a common assumption in non-inferiority trials. Table 2.2 gives the required exact and approximate sample sizes per group for the balanced design ($c_C = c_P = 1$) to achieve an overall power of 80% with significance level $\alpha = 0.025$ for different constellations of the parameters ϵ and $\Delta_{ni}/(\mu_C - \mu_P)$. For instance, $\Delta_{ni}/(\mu_C - \mu_P) = 1/2$ means that the non-inferiority margin is half the difference between the control and placebo effect. For some clinical indications this is a common choice for Δ_{ni} and values larger than this are rather scarce in practice.

As we can see, Table 2.2 confirms the validity of the normal approximation, as the required exact and approximate sample sizes per group are almost equal for the considered parameter constellations. Further investigations showed that the normal approximation also stays valid

Table 2.2: Exact (first row) and approximate (second row) sample sizes per group of the balanced design to achieve an overall power of 80% with significance level $\alpha = 0.025$ and assuming $\mu_T = \mu_C$.

$\Delta_{ni}/(\mu_C - \mu_P)$	$\epsilon = \sigma/(\mu_C - \mu_P)$							
	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
1/6	36	142	319	566	884	1273	1732	2262
	36	142	318	566	883	1272	1731	2261
1/5	26	99	222	394	614	884	1203	1571
	25	99	221	393	614	883	1202	1570
1/4	17	64	142	252	394	566	770	1006
	16	63	142	252	393	566	770	1005
1/3	10	36	81	142	222	319	434	566
	9	36	80	142	221	318	433	566
1/2	5	17	36	64	99	142	193	252
	4	16	36	63	99	142	193	252

under the optimal allocation determined in the following section (see Table A.1).

Before proceeding with the determination of an optimal allocation, let us give one comment on the connection between the effect retention and the fixed margin approach. In terms of the effect retention hypothesis in (2.8) the quotient $\Delta_{ni}/(\mu_C - \mu_P)$ is equal to $1 - f$. For example, a choice of $\Delta_{ni}/(\mu_C - \mu_P) = 1/5$ roughly means that the test treatment should preserve at least 80% of the control treatment effect relative to placebo. However, there is a substantial difference between the two approaches, since for the effect retention test the non-inferiority margin is not fixed in advance and depends on the actual treatment effects of control and placebo in the current non-inferiority trial. According to that, one should avoid interpreting a trial based on the fixed margin approach in terms of a retention of the control treatment effect and vice versa.

2.3.3 Optimal Sample Size Allocation

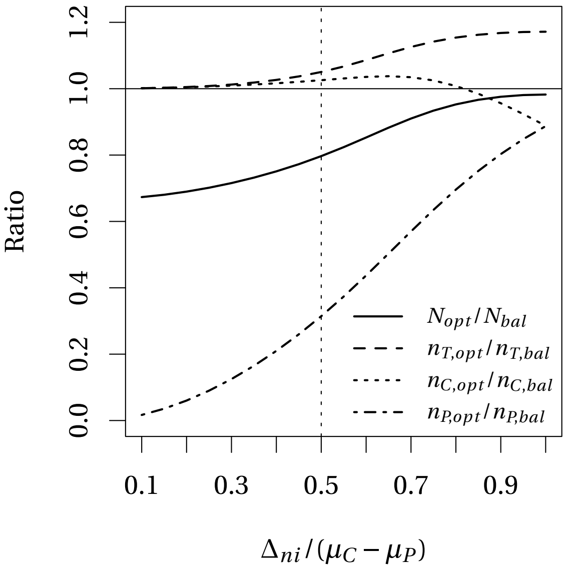
The previous section showed that the normal approximation in Equation (2.14) is valid for a variety of parameter constellations, thus the approximate formula for calculating the overall power will be used from now on.

Taking a look at the sample size allocation between the three groups, it is obvious that a balanced design with equal sample sizes might, for many reasons, not be the best choice in a three-arm non-inferiority trial. For instance, allocating too many patients to placebo might not be ethically justifiable as the control treatment is an already approved effective drug. Furthermore, it might be reasonable to allocate more patients to the test treatment in order to obtain sufficient safety data on the new therapy. As it is also desirable to keep the overall sample size N as small as possible, one can use the fact that the overall power $1 - \beta$ depends, among other variables, on the allocation ratios c_C and c_P.

With a prespecified overall power $1 - \beta$, optimal sample size allocations c_C and c_P that minimise the overall sample size N can be calculated by means of Equation (2.14). This is done by the numerical optimisation method proposed by Byrd et al. (1995) with the constraint c_C, $c_P > 0$. For ease of computation continuous sample sizes n_T, n_C and n_P are assumed.

Figure 2.1 gives an overview on the sample size savings provided by the optimal allocation in comparison to the balanced design with equal sample sizes in the three treatment groups. Therefore, the optimal divided by the balanced sample sizes to achieve an overall power of $1 - \beta = 80\%$ with $\alpha = 2.5\%$ and assuming $\mu_T = \mu_C$ are displayed for differing values of $\Delta_{ni}/(\mu_C - \mu_P)$. For example, a quotient of 0.8 for N means that the optimal design requires 20% less subjects than the balanced design. As we can see, the sample size savings for N and especially n_P become larger with smaller values of $\Delta_{ni}/(\mu_C - \mu_P)$. For scenarios of practical relevance, i.e. $\Delta_{ni}/(\mu_C - \mu_P) \leq 0.5$, the overall sample size can be reduced by more than 20%. Furthermore, the optimal test and control group sizes are almost equal to the balanced sample sizes, whereas the placebo group size is substantially reduced by more than 70%. Further investigations by Stucke and Kieser (2012) also showed that the optimal allocation leads to a substantial sample

Figure 2.1: Quotients of the optimal sample sizes divided by the balanced sample sizes to achieve an overall power of $1-\beta = 80\%$ with $\alpha = 0.025$ and assuming $\mu_T = \mu_C$.

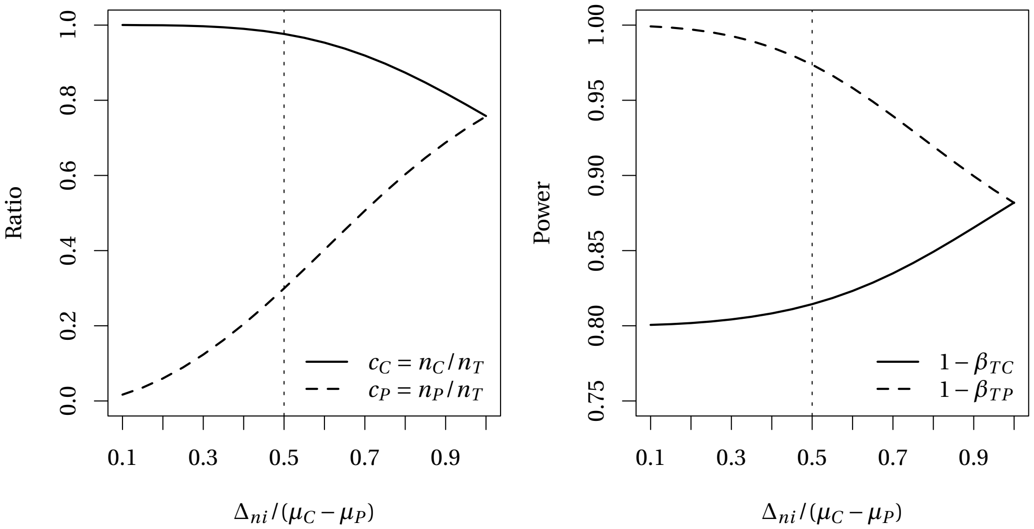


size reduction in comparison to the previously mentioned *ad hoc* method.

The same pattern as in Figure 2.1 can be seen in the left-hand side of Figure 2.2 which shows the allocation ratios c_C and c_P of the optimal design with $1-\beta = 0.8$, $\alpha = 0.025$ and $\mu_T = \mu_C$. For practically relevant values $\Delta_{ni}/(\mu_C - \mu_P) \le 0.5$, the sample size of the control group is almost equal to that of the test group under the optimal allocation, whereas the placebo group size is substantially reduced. For example, assuming $\mu_C - \mu_P = 2\Delta_{ni}$ and an overall power of $1-\beta = 0.80$ the optimal allocation ratio is nearly $n_T : n_C : n_P = 3.3 : 3.3 : 1$, which means that approximately 13% of all patients are assigned to the placebo group. Randomising less patients to placebo than to the active treatment groups is highly desirable from an ethical point of view. In addition, this could reduce patient concerns and as a result enhance the recruitment process of the trial. Moreover, it becomes evident from the right-hand side of Figure 2.2 that the power of the first hypothesis test, i.e. the direct proof of efficacy for the test treatment, is high, while the power of the second hypothesis test is similar to the overall power.

Let us consider a hypothetical example for calculating the sample sizes in a three-arm non-inferiority trial. Suppose that the treatment groups have mean values $\mu_T = \mu_C = 1$ and $\mu_P = 0.6$ with a common standard deviation $\sigma = 0.8$, that means $\epsilon = \sigma/(\mu_C - \mu_P) = 2$. The non-inferiority margin is set to half of the difference between control and placebo effect, thus $\Delta_{ni} = 0.2$. According to Table 2.2 the required sample size per group for the balanced design with $\alpha = 2.5\%$ to achieve $1-\beta = 80\%$ overall power is $n_T = n_C = n_P = 252$, which means that the overall sample size is $N = 756$. The corresponding optimal sample sizes are obtained as $n_T = 264$, $n_C = 258$ and $n_P = 79$ with an overall sample size of $N = 601$. As we can see, the optimal

Figure 2.2: Optimal sample size allocations (left) and power to reject $H_{0,TP}^{(s)}$ or $H_{0,TC}^{(n)}$ under the optimal allocation (right) with $1-\beta = 80\%$, $\alpha = 0.025$ and $\mu_T = \mu_C$.



allocation leads to a substantial reduction of both the overall sample size N and the size of the placebo group n_P compared to the balanced design.

As already mentioned, there are situations where the inclusion of an additional proof of efficacy for the active control is a mandatory requirement, e.g. for trials regarding mild persistent asthma. It can be shown that the vector of test statistics $(T_{TP}^{(s)}, T_{TC}^{(n)}, T_{CP}^{(s)})'$ follows a trivariate noncentral t-distribution (or approximately a trivariate normal distribution), so that all calculations regarding power and sample size can be performed analogously to Section 2.3.2. Our investigations showed, that the optimal allocation ratios are only marginally affected by the inclusion of the null hypothesis $H_{0,CP}^{(s)}: \mu_C \leq \mu_P$ (see Table A.2). Furthermore, the loss of overall power after including $H_{0,CP}^{(s)}$ is negligible under the optimal allocation. For instance, for an optimal design with overall power $1-\beta = 80\%$ to reject both null hypotheses $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ and assuming $\mu_C - \mu_P = 2\Delta_{ni}$, the probability to reject $H_{0,TP}^{(s)}$, $H_{0,TC}^{(n)}$ and $H_{0,CP}^{(s)}$ is 78.39%. In order to achieve 80% power to reject all three null hypotheses the required increase in the overall sample size would be 3.33%. For smaller non-inferiority margins and a higher overall power, e.g. $1-\beta = 90\%$, the loss of power and the necessary increase in the overall sample size are further reduced.

2.4 Summary

In three-arm 'gold standard' non-inferiority trials including an experimental treatment, an active control and a placebo there are three different comparisons of interest. These are the superiority comparison between test and placebo, the non-inferiority comparison between test and control and the superiority comparison between control and placebo, where the former two represent the main study objectives. The application of adequate multiple testing procedures in such trials is therefore fundamental, not least because the regulatory authorities explicitly require strong FWER control for confirmatory claims (CPMP, 2002) and non-inferiority studies usually are of confirmatory nature. It was illustrated that the fixed sequence procedure provides a simple and efficient approach for multiplicity adjustment in three-arm trials, as there is a natural hierarchy between the three hypotheses. For instance, the non-inferiority comparison between test and control is of little sense if the test treatment is not even superior to placebo.

Two different approaches for analysing three-arm non-inferiority trials have been proposed in the literature, on the one hand the effect preservation approach and on the other hand the fixed margin approach. The former one has been extensively studied in the recent years. However, it has not yet been applied in confirmatory clinical trials, as the regulatory guidelines state in particular that a fixed margin Δ_{ni} should be prespecified in advance. Therefore, this chapter concentrated on the fixed margin approach. Although the proof of efficacy for the active control might be mandatory for some clinical indications, the investigations were restricted to the two main objectives in such trials, namely the direct proof of efficacy for the test treatment and assessing the similarity of the test to the control.

For normally distributed endpoints with common but unknown variance it was shown that the exact and approximate overall power of the procedure can be determined by means of the distribution functions of the bivariate t- and normal distribution, respectively. As these functions are implemented in most statistical software packages nowadays, sample size calculations can be easily conducted. Optimal sample size allocations were derived, leading to a substantial reduction of the overall sample size. Furthermore, it turned out that the placebo group size is substantially reduced under the optimal allocation, which is highly desirable from an ethical point of view. For a general approach for sample size calculation in three-arm trials based on maximum likelihood estimators see Stucke and Kieser (2012).

Finally, the question arises whether the optimal design derived in this section can be further improved. Interim analyses with the possibility to stop the trial early either for efficacy or futility might be a useful option for three-arm non-inferiority trials. In particular, the placebo group might be dropped at an interim analysis if there is enough evidence for the efficacy of the test treatment, i.e. test is superior to placebo. This would make the trial even more acceptable for patients and one could also exploit the high power of the superiority comparison between test and placebo under the optimal allocation (see Figure 2.2). The next chapter will deal with such group sequential three-arm non-inferiority designs.

CHAPTER 3

Group Sequential Three-Arm Non-Inferiority Designs

The previous chapter showed that the optimal sample size allocation not only minimises the overall sample size, but also considerably reduces the placebo group size. Besides the ethical advantage of randomising less patients to placebo, this also improves the precision of the active drug comparison, making the trial more acceptable for both patients and investigators (cf. Section 2.1.5.1.1 of ICH E10). Furthermore, it turned out that the power of the superiority comparison between test and placebo is very high under the optimal allocation. The idea is now to utilise this high power by means of implementing a group sequential design. The possibility to terminate the placebo arm when the test treatment is demonstrated to be superior to placebo at an interim analysis, would make the study even more acceptable for patients.

The structure of this chapter is as follows. After an introduction to group sequential designs in Section 3.1, different group sequential designs for three-arm non-inferiority trials are proposed in the subsequent Section 3.2. Formulas for calculation of the overall power and the expected sample sizes are derived and the proposed method is applied to a hypothetical clinical trial. Finally, the proposed designs are compared with the optimal single-stage designs derived in the previous chapter and approximately optimal group sequential rejection boundaries are determined for different optimisation problems, such as minimising the expected overall sample size.

3.1 Group Sequential Designs

In most clinical trials the required sample size is determined in advance and the corresponding statistical analyses are conducted only after the observations of all patients have been recorded. In contrast, sequential designs allow to repeatedly evaluate the data at different points in time while at the same time satisfying the desired error probabilities. At an interim analysis a decision on the termination or continuation of the study is made based on the data accrued by then. Although this approach seems natural, the classical statistical theory is primarily based on prespecified sample sizes. Armitage (1993, p. 392) made an interesting supposition in this regard:

> The classical theory of experimental design deals predominantly with experiments of predetermined size, presumably because the pioneers of the subject, particularly R. A. Fisher, worked in agricultural research, where the outcome of a field trial is not available until long after the experiment has been designed and started. It is interesting to speculate how differently statistical theory might have evolved if Fisher had been employed in medical or industrial research.

The reasons for implementing sequential analyses in clinical studies are numerous. Besides cost and time savings, the application of sequential methodology can result in quicker approval of effective treatments or early detection of harmful therapies. Thus, there are potential benefits for both patients and manufacturers.

The development of sequential procedures and their application started in the late 1920s with one of the first contributions by Dodge and Romig (1929), who proposed a two-stage acceptance sampling plan for quality control. The first idea of sequential testing was introduced by Wald (1947) with the sequential probability ratio test (SPRT). In contrast to a fixed sample size design, the SPRT evaluates the data after each new observation, so that the sample size is unknown in advance. Because the procedure is also not bounded, the actual sample size might be relatively large, especially when the true parameter of interest θ lies between the anticipated null and alternative parameters θ_0 and θ_1, respectively. As furthermore a patient-wise evaluation can hardly be implemented in practice, an application of the SPRT in clinical trials is problematic. The term *group sequential designs* was coined by Elfring and Schultz (1973), who suggested a stage-wise procedure for comparing two treatments with binary responses. Several contributions in the following years finally led to the pioneering work of Pocock (1977) and O'Brien and Fleming (1979). The underlying approach can be applied to a large variety of testing situations and provides the basis for all further developments. For a comprehensive summary on group sequential methodology see Jennison and Turnbull (2000) and Proschan et al. (2006). In the following motivation we take a look at the type I error inflation caused by unadjusted interim analyses and a possible solution.

3.1.1 Motivation

Suppose we want to demonstrate the superiority of a new treatment A over another treatment B. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$, $i = 1, ..., n$, and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, ..., n$, denote the mutually independent responses of patients allocated to treatment A and B, respectively, with known common variance σ^2. Assuming that larger responses are associated with greater benefits, the corresponding set of hypotheses is given as

$$H_0 : \mu_A \leq \mu_B \quad \text{vs.} \quad H_1 : \mu_A > \mu_B.$$

Besides the final analysis at one-sided level $\alpha = 0.025$, i.e. critical value $z_{0.975} = 1.96$, an additional analysis is scheduled when 50 percent of the responses, i.e. $n/2$ per treatment group, are observed. The following test procedure shall be adopted:

When $n/2$ responses per group are observed:
 If $Z_{\text{interim}} > 1.96$, reject H_0 and stop the trial.
 If $Z_{\text{interim}} \leq 1.96$, continue recruiting and proceed to the final stage.

When all n responses per group are observed:
 If $Z_{\text{final}} > 1.96$, reject H_0.
 If $Z_{\text{final}} \leq 1.96$, accept H_0.

Here, Z_{interim} and Z_{final} denote the common Z-test statistics at interim and at the end of the study, i.e. $Z_{\text{interim}} = \frac{1}{\sqrt{n\sigma^2}}(\sum_{i=1}^{n/2} X_{A,i} - \sum_{i=1}^{n/2} X_{B,i})$ and $Z_{\text{final}} = \frac{1}{\sqrt{2n\sigma^2}}(\sum_{i=1}^{n} X_{A,i} - \sum_{i=1}^{n} X_{B,i})$. The maximum probability to commit a type I error with this procedure is obtained as

$$\begin{aligned}
&P_{\mu_A=\mu_B}\left(H_0 \text{ is rejected at interim or at the final analysis}\right) \\
&= P_{\mu_A=\mu_B}\left(\{Z_{\text{interim}} > 1.96\} \cup \{Z_{\text{interim}} \leq 1.96, Z_{\text{final}} > 1.96\}\right) \\
&= \underbrace{P_{\mu_A=\mu_B}\left(\{Z_{\text{interim}} > 1.96\}\right)}_{=0.025} + \underbrace{P_{\mu_A=\mu_B}\left(\{Z_{\text{interim}} \leq 1.96, Z_{\text{final}} > 1.96\}\right)}_{>0} \qquad (3.1) \\
&> 0.025.
\end{aligned}$$

Hence, the procedure does not control the overall type I error rate $\alpha = 0.025$. Bearing in mind that repeated significance testing essentially is a multiple testing problem, this is hardly surprising. For instance, the type I error rate could be controlled by using the Bonferroni adjustment mentioned in Section 2.1. However, this would ignore the existing dependencies between the test statistics Z_{interim} and Z_{final}. It can be shown that, given $\mu_A = \mu_B$, the vector of test statistics $(Z_{\text{interim}}, Z_{\text{final}})'$ follows a bivariate normal distribution with mean vector $(0,0)'$ and correlation $1/\sqrt{2}$ (see Sections 3.1.2 and 3.1.3). Consequently, the second term in (3.1) can be directly calculated by means of the bivariate normal distribution function and is given as 0.017, so that the maximum type I error rate of the procedure is 0.042. Table 3.1 gives an overview on the type I error inflation for different numbers of unadjusted and equally spaced interim analyses at one-sided significance level 2.5%. One can see that already two additional unadjusted interim analyses lead to more than a doubling of the type I error rate. Besides calculating the exact type I error rate, the distributional properties of the test statistics can also be used to derive appropriate rejection boundaries for the different stages. In order to control the type I error rate of the above mentioned two-stage procedure by $\alpha = 0.025$, the critical values at interim (c_{interim})

Table 3.1: Maximum probability of committing a type I error for different number of equally spaced interim analyses at unadjusted one-sided significance level $\alpha = 0.025$.

# Interim analyses	0	1	2	3	4	5	6	7	8	9	10
Max. type I error rate (%)	2.5	4.2	5.4	6.3	7.1	7.7	8.3	8.8	9.3	9.7	10.1

and at the final analysis (c_{final}) have to satisfy the equation

$$P_{\mu_A=\mu_B}(\{Z_{\text{interim}} > c_{\text{interim}}\}) + P_{\mu_A=\mu_B}(\{Z_{\text{interim}} \leq c_{\text{interim}}, Z_{\text{final}} > c_{\text{final}}\}) = 0.025.$$

With the constraint of equal stage-wise critical values as proposed by Pocock (1977), this leads to $c_{\text{interim}} = c_{\text{final}} = 2.178$. However, this constant choice is arbitrary, as e.g. the group sequential design with $c_{\text{interim}} = 2.797$ and $c_{\text{final}} = 1.977$ proposed by O'Brien and Fleming (1979) also satisfies the type I error constraint just as multiple other choices.

3.1.2 General Design

Based on the same scenario as in the motivation, i.e. a superiority comparison between two treatments A and B, the general principle of group sequential designs will now be described. Denote by $X_{A,i} \sim N(\mu_A, \sigma^2)$, $i = 1, 2, \ldots$, and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, the mutually independent responses of patients allocated to treatment A and B, respectively, with known common variance σ^2. With larger responses being desirable and defining $\theta = \mu_A - \mu_B$, the corresponding set of hypotheses of the superiority comparison is $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$.

It should be noted that the group sequential procedure described in this subsection can be easily transferred to other testing problems, such as comparison of binary outcomes or survival data. For more information on the general applications of group sequential methodology see Jennison and Turnbull (2000, Chapter 2).

Test Procedure

Suppose that patient entry is divided into K groups and the data are analysed repeatedly after the responses of each new group have been observed. The cumulative sample sizes of the two treatment groups A and B at the different stages shall be denoted as $n_A^{(1)}, \ldots, n_A^{(K)}$ and $n_B^{(1)}, \ldots, n_B^{(K)}$, respectively. Note that there are no restrictions to the sample sizes such as equally sized stage-wise groups or a balanced treatment allocation. The corresponding standardised test statistics after each group of observations are given as

$$Z_k = \frac{\bar{X}_A^{(k)} - \bar{X}_B^{(k)}}{\sigma\sqrt{\frac{1}{n_A^{(k)}} + \frac{1}{n_B^{(k)}}}} = \left(\bar{X}_A^{(k)} - \bar{X}_B^{(k)}\right)\sqrt{\mathscr{I}_k} \quad \text{for } k = 1, \ldots, K, \tag{3.2}$$

where $\bar{X}_A^{(k)}$ and $\bar{X}_B^{(k)}$ denote the stage-wise sample means of treatment group A and B, respectively, i.e. $\bar{X}_D^{(k)} = \frac{1}{n_D^{(k)}} \sum_{i=1}^{n_D^{(k)}} X_{D,k}$ with $D = A, B$ and $k = 1, ..., K$. The variables $\mathscr{I}_k$, $k = 1, ..., K$, are also called *information levels* as they represent the information that is actually available at the respective stage. The group sequential test procedure takes the following form:

$$\begin{aligned}
&\text{At stage } k = 1, ..., K-1 \\
&\qquad \text{if } Z_k \geq b_k, \quad \text{stop and reject } H_0, \\
&\qquad \text{if } Z_k < b_k, \quad \text{continue to stage } k+1. \\
&\text{At stage } K \\
&\qquad \text{if } Z_K \geq b_K, \quad \text{stop and reject } H_0, \\
&\qquad \text{if } Z_K < b_K, \quad \text{stop and accept } H_0.
\end{aligned} \tag{3.3}$$

Type I Error and Rejection Boundaries

The stopping boundaries $b_1, ..., b_K$ are determined in such a way that the overall type I error rate is controlled by α. Therefore, they have to satisfy the equation

$$1 - P_{\theta=0}\left(\bigcap_{k=1}^{K} \{Z_k < b_k\}\right) = \alpha. \tag{3.4}$$

In order to calculate the left-hand side of Equation (3.4) we need to take a look at the distributional properties of the test statistics $Z_1, ..., Z_K$. The means of the test statistics at stage $k = 1, ..., K$ are easily obtained as

$$E(Z_k) = \theta\sqrt{\mathscr{I}_k}.$$

Before assessing the covariance between the test statistics in different stages let us first take a look at the stage-wise sample means. The covariance of the sample means of treatment A at two stages $k_1, k_2 \in \{1, ..., K\}$ is obtained as

$$\begin{aligned}
Cov\left(\bar{X}_A^{(k_1)}, \bar{X}_A^{(k_2)}\right) &= Cov\left(\frac{1}{n_A^{(k_1)}} \sum_{i=1}^{n_A^{(k_1)}} X_{A,i}, \frac{1}{n_A^{(k_2)}} \sum_{i=1}^{n_A^{(k_2)}} X_{A,i}\right) \\
&= \frac{1}{n_A^{(k_1)} n_A^{(k_2)}} \sum_{i=1}^{n_A^{(k_1)}} \sum_{j=1}^{n_A^{(k_2)}} \underbrace{Cov\left(X_{A,i}, X_{A,j}\right)}_{=\begin{cases} \sigma^2 & \text{if } i = j, \\ 0 & \text{else.} \end{cases}} \\
&= \frac{1}{n_A^{(k_1)} n_A^{(k_2)}} n_A^{(\min(k_1,k_2))} \sigma^2 \\
&= \frac{\sigma^2}{n_A^{(\max(k_1,k_2))}}.
\end{aligned} \tag{3.5}$$

Analogously, it follows for treatment B that $Cov(\bar{X}_B^{(k_1)}, \bar{X}_B^{(k_2)}) = \sigma^2 / n_B^{(\max(k_1,k_2))}$, so that the covariance of the test statistics at two stages $k_1, k_2 \in \{1, ..., K\}$ is determined as

$$\begin{aligned}
Cov\left(Z_{k_1}, Z_{k_2}\right) &= \sqrt{\mathscr{I}_{k_1}\mathscr{I}_{k_2}} Cov\left(\bar{X}_A^{(k_1)} - \bar{X}_B^{(k_1)}, \bar{X}_A^{(k_2)} - \bar{X}_B^{(k_2)}\right) \\
&= \sqrt{\mathscr{I}_{k_1}\mathscr{I}_{k_2}} \left[Cov\left(\bar{X}_A^{(k_1)}, \bar{X}_A^{(k_2)}\right) + Cov\left(\bar{X}_B^{(k_1)}, \bar{X}_B^{(k_2)}\right)\right] \\
&= \sqrt{\mathscr{I}_{k_1}\mathscr{I}_{k_2}} \underbrace{\left[\frac{\sigma^2}{n_A^{(\max(k_1,k_2))}} + \frac{\sigma^2}{n_B^{(\max(k_1,k_2))}}\right]}_{=1/\mathscr{I}_{\max(k_1,k_2)}} \\
&= \sqrt{\frac{\mathscr{I}_{\min(k_1,k_2)}}{\mathscr{I}_{\max(k_1,k_2)}}}. \qquad (3.6)
\end{aligned}$$

Obviously, all possible linear combinations of the test statistics Z_k, $k = 1, ..., K$, are normally distributed, because each Z_k itself is a linear combination of normally distributed random variables. Consequently, the vector of test statistics $\boldsymbol{Z} = (Z_1, ..., Z_K)'$ follows a K-variate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ given as

$$\boldsymbol{\mu} = \left(\theta\sqrt{\mathscr{I}_i}\right)'_{1\le i\le K} \quad \text{and} \quad \boldsymbol{\Sigma} = \left(\sqrt{\frac{\mathscr{I}_{\min(i,j)}}{\mathscr{I}_{\max(i,j)}}}\right)_{1\le i,j\le K}. \qquad (3.7)$$

Interestingly, the standard deviation σ can be cancelled out in the items of the covariance matrix, so that $\boldsymbol{\Sigma}$ only depends on the stage-wise sample sizes, or to be exact, on the sample size allocation among the different stages and between the two groups. Consequently, Equation (3.4) can be rewritten as

$$1 - \Phi^{\boldsymbol{\Sigma}}(b_1, ..., b_K) = \alpha, \qquad (3.8)$$

where $\Phi^{\boldsymbol{\Sigma}}(\cdot)$ denotes the cumulative distribution function of the K-variate normal distribution function with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ given in (3.7). As mentioned in Section 2.3.2, $\Phi^{\boldsymbol{\Sigma}}(\cdot)$ is implemented in many statistical software packages nowadays such as R, so that the left-hand side of (3.8) can be easily calculated. Appropriate rejection boundaries $b_1, ..., b_K$ can be determined numerically as a solution of (3.8), so that the overall type I error rate does not exceed α. It should be mentioned that Armitage et al. (1969) presented a recursive version that considerably simplifies determination of the left-hand side of (3.8). However, due to the enormous computing power of today's computers, time gains are only marginal.

Power and Sample Size

After determination of appropriate rejection boundaries $b_1, ..., b_K$ for a specific sample size allocation among the stages and treatment groups, e.g. equal stage and treatment group sizes, the next step is to determine the required sample size in order to have sufficient overall power.

Similar to the type I error rate, the overall power of the group sequential procedure is determined as

$$
\begin{aligned}
1-\beta &= 1 - P_\theta\left(\bigcap_{k=1}^{K}\{Z_k < b_k\}\right)\\
&= 1-\Phi^{\boldsymbol{\Sigma}}\left(b_1-\theta\sqrt{\mathscr{I}_1},\dots,b_K-\theta\sqrt{\mathscr{I}_K}\right). \qquad (3.9)
\end{aligned}
$$

With prespecified sample size allocation among the stages and between the two treatment groups, i.e. $\frac{n_D^{(k)}}{n_D^{(K)}}$, $D = A,B$, $k = 1,\dots,K-1$, and $\frac{n_A^{(K)}}{n_B^{(K)}}$, the right-hand side of (3.9) can be written as a function of $n_A^{(K)}$ (or $n_B^{(K)}$). The required sample sizes are then determined by means of a simple univariate root-finding algorithm such as bisection.

Expected Sample Size

Due to the step-wise nature of a group sequential procedure, the sample sizes actually needed are unknown in the planning stage. Let N_A and N_B denote the actual sample sizes for treatment group A and B, respectively. Similar to the overall power, N_A and N_B are highly dependent on the true treatment difference $\theta = \mu_A - \mu_B$. For treatment differences near zero the group sequential procedure tends to proceed to later stages, resulting in higher actual sample sizes. If, in contrast, the true treatment difference is high, the procedure will more likely stop at an earlier stage with rejection of H_0, so that N_A and N_B are small.

Obviously, the actual sample sizes N_A and N_B are discrete random variables with realisations $n_A^{(1)},\dots,n_A^{(K)}$ and $n_B^{(1)},\dots,n_B^{(K)}$, respectively. Hence, it seems natural to consider the expected values of N_A and N_B. Denote by $\tilde{n}_A^{(k)}$, $k = 1,\dots,K$, the *additional* number of patients that are accrued at stage k, i.e. $\tilde{n}_A^{(1)} = n_A^{(1)}$ and $\tilde{n}_A^{(k)} = n_A^{(k)} - n_A^{(k-1)}$ for $k = 2,\dots,K$. The expected value of N_A for a specific treatment difference θ is then determined as

$$
\begin{aligned}
E_\theta\left(N_A\right) &= \sum_{k=1}^{K} n_A^{(k)} P_\theta\left(N_A = n_A^{(k)}\right)\\
&= \sum_{k=1}^{K}\sum_{i=1}^{k} \tilde{n}_A^{(i)} P_\theta\left(N_A = n_A^{(k)}\right)\\
&= \sum_{k=1}^{K} \tilde{n}_A^{(k)} \underbrace{\sum_{i=k}^{K} P_\theta\left(N_A = n_A^{(i)}\right)}_{=P_\theta\left(N_A \geq n_A^{(k)}\right)}\\
&= \tilde{n}_A^{(1)} \underbrace{P_\theta\left(N_A \geq n_A^{(1)}\right)}_{=1} + \sum_{k=2}^{K} \tilde{n}_A^{(k)} \underbrace{P_\theta\left(N_A \geq n_A^{(k)}\right)}_{P_\theta\left(\bigcap_{i=1}^{k-1}\{Z_i < b_i\}\right)} \qquad (3.10)\\
&= n_A^{(1)} + \sum_{k=2}^{K}\left(n_A^{(k)} - n_A^{(k-1)}\right)\Phi^{\boldsymbol{\Sigma}(k,\dots,K)}\left(b_1-\theta\sqrt{\mathscr{I}_1},\dots,b_{k-1}-\theta\sqrt{\mathscr{I}_{k-1}}\right),
\end{aligned}
$$

where $\boldsymbol{\Sigma}(i,...,j)$ with $i \leq j$ is the matrix that is formed by deleting the rows and columns $i,...,j$ of $\boldsymbol{\Sigma}$. The expectation of N_B is determined analogously and the expected overall sample size is given as the sum of $E_\theta(N_A)$ and $E_\theta(N_B)$. In the literature the expected sample sizes are often also referred to as *average sample numbers*. Thus, the expectations of the actual sample sizes of treatment group A, B and overall for a specific treatment difference θ shall be denoted as $ASn_A(\theta)$, $ASn_B(\theta)$ and $ASN(\theta)$, respectively.

Together with the maximum sample sizes $n_A^{(K)}$ and $n_B^{(K)}$, the average sample numbers are useful performance characteristics of group sequential designs. Since Equation (3.8) obviously has an infinite number of solutions for an appropriate set of rejection boundaries $b_1,...,b_K$, there are infinitely many group sequential designs to choose from. The maximum and expected sample sizes prove to be very useful when it comes to selecting an adequate design for a particular setting. The average sample numbers at $\theta = 0$, i.e. when H_0 is true, or at the treatment difference that was assumed in the sample size calculation are often of particular interest.

3.1.3 Classical Group Sequential Tests

The first contributions regarding group sequential designs, which are often denoted as the *classical* group sequential designs, assumed equal stage-wise sample sizes, i.e. $n_D^{(k)} = \frac{k}{K} n_D^{(K)}$ for $D = A, B$ and $k = 1,...,K-1$. With this assumption the covariance of the test statistics from two stages $k_1, k_2 \in \{1,...,K\}$ simplifies to

$$\begin{aligned} Cov\left(Z_{k_1}, Z_{k_2}\right) &= \sqrt{\frac{\mathscr{I}_{\min(k_1,k_2)}}{\mathscr{I}_{\max(k_1,k_2)}}} \\ &= \frac{\sigma\sqrt{\frac{1}{n_A^{(\max(k_1,k_2))}} + \frac{1}{n_B^{(\max(k_1,k_2))}}}}{\sigma\sqrt{\frac{1}{n_A^{(\min(k_1,k_2))}} + \frac{1}{n_B^{(\min(k_1,k_2))}}}} \\ &= \sqrt{\frac{\frac{K}{\max(k_1,k_2) n_A^{(K)}} + \frac{K}{\max(k_1,k_2) n_B^{(K)}}}{\frac{K}{\min(k_1,k_2) n_A^{(K)}} + \frac{K}{\min(k_1,k_2) n_B^{(K)}}}} \\ &= \sqrt{\frac{\min(k_1,k_2)}{\max(k_1,k_2)}}, \end{aligned} \tag{3.11}$$

so that the variance-covariance matrix of the vector of test statistics $\boldsymbol{Z}$ is given as

$$\tilde{\boldsymbol{\Sigma}} = \left(\sqrt{\frac{\min(i,j)}{\max(i,j)}}\right)_{1 \leq i,j \leq K}.$$

Consequently, each set $(b_1,...,b_K)$ satisfying (3.8) is independent of the maximum sample sizes $n_A^{(K)}$ and $n_B^{(K)}$, so that the rejection boundaries apply for all designs with equal stage sizes. The different proposals for appropriate boundaries made certain assumptions on the boundary value structure that further simplify the determination of $b_1,...,b_K$. For instance, Pocock

(1977) assumed equal boundaries at each stage, whereas O'Brien and Fleming (1979) proposed monotone decreasing critical values, namely

Pocock boundaries: $b_k = b_{POC}$ for $k = 1, \ldots, K$,

O'Brien & Fleming boundaries: $b_k = \sqrt{\frac{K}{k}} b_{OBF}$ for $k = 1, \ldots, K$.

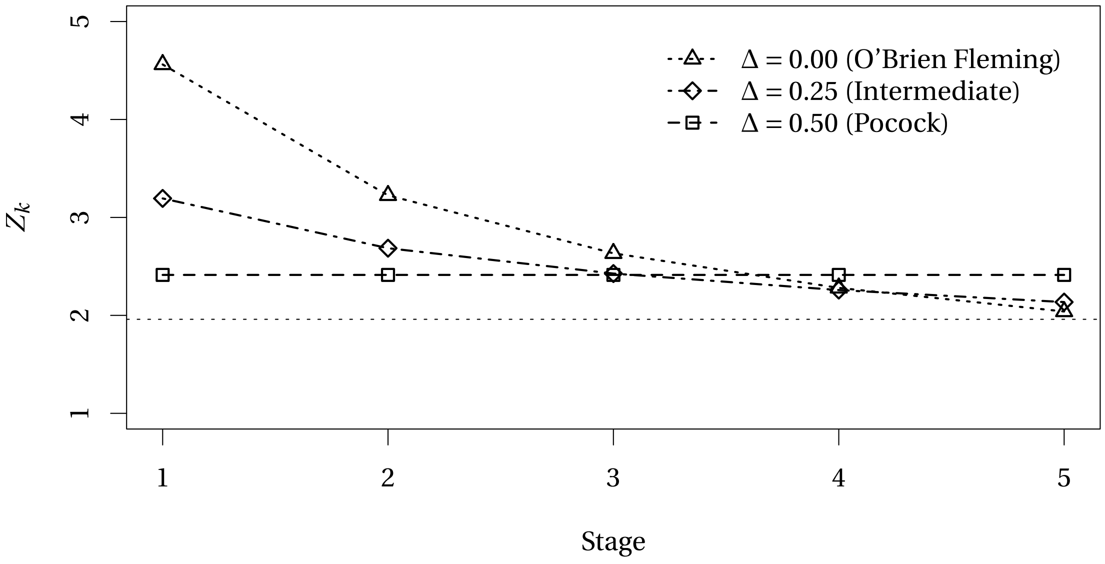
A more general class of group sequential designs was introduced by Wang and Tsiatis (1987) who defined the boundaries in the following way:

Wang & Tsiatis boundaries: $b_k = \left(\frac{k}{K}\right)^{\Delta - \frac{1}{2}} b_{WT}$ for $k = 1, \ldots, K$,

where $\Delta \in \mathbb{R}$ is a prespecified constant usually chosen between 0 and 0.5, which obviously coincide with the designs by O'Brien & Fleming and Pocock, respectively. The boundaries for these designs can be easily determined by solving Equation (3.8) for b_{POC}, b_{OBF} or b_{WT} by means of a simple univariate root-finding method such as regula falsi. For illustrative purposes, the boundary values of the Δ-class according to Wang and Tsiatis for $\Delta = 0$, 0.25, 0.5 with $K = 5$ stages and one-sided significance level of $\alpha = 0.025$ are presented in Figure 3.1.

It becomes apparent that the design by Pocock ($\Delta = 0.5$) has the lowest boundaries at earlier stages, whereas at later stages the boundaries according to O'Brien & Fleming ($\Delta = 0$) are the lowest. Consequently, with Pocock boundaries the group sequential procedure tends to reject H_0 at earlier stages than with O'Brien Fleming boundaries. The O'Brien Fleming design rejects

Figure 3.1: Group sequential rejection boundaries according to the Δ-class by Wang and Tsiatis with $K = 5$ and $\alpha = 0.025$. The thin dotted line indicates the critical value of the common fixed sample size design $z_{0.975} = 1.96$.



H_0 at earlier stages only if there is a substantial treatment difference and its final boundary value is close to the critical value of the common single-stage design. This is why the O'Brien Fleming design is often used in clinical trials, because an early study termination with rejection of H_0 is not always desirable as there might not be enough safety data available at that time point. The intermediate design with $\Delta = 0.25$ is some kind of a trade-off between the designs by Pocock and O'Brien and Fleming.

In order to get an impression on the performance of the three designs represented in Figure 3.1, the corresponding maximum sample sizes (N_{max}) as well as the average sample numbers under H_0 ($ASN(0)$) and the alternative of the sample size calculation ($ASN(\delta)$) were calculated for different amounts of power and number of stages with $\alpha = 0.025$. The sample sizes were not rounded to integers as this has only a small effect on the results. Table 3.2 shows the group sequential sample sizes expressed as percentages of the corresponding single-stage sample size N_{fix}, so that the values apply for all combinations of δ, σ^2 and sample size allocation $\frac{n_A^{(K)}}{n_B^{(K)}}$. Note that the fixed, maximum and expected sample size depend on these parameters in the same way. It should also be considered that the group sequential designs with 80% and 90% power

Table 3.2: Maximum and expected sample sizes of group sequential Δ-class designs according to Wang & Tsiatis represented as percentages of the fixed sample size with $\alpha = 0.025$, power $1-\beta$ at $\theta = \delta$ and K stages.

		$1-\beta = 0.80$			$1-\beta = 0.90$		
	K	$\frac{N_{max}}{N_{fix}}$	$\frac{ASN(0)}{N_{fix}}$	$\frac{ASN(\delta)}{N_{fix}}$	$\frac{N_{max}}{N_{fix}}$	$\frac{ASN(0)}{N_{fix}}$	$\frac{ASN(\delta)}{N_{fix}}$
$\Delta = 0.00$	2	100.8	100.6	90.2	100.7	100.6	85.1
	3	101.7	101.5	85.6	101.6	101.4	79.9
	4	102.4	102.1	83.1	102.2	101.9	76.7
	5	102.8	102.5	81.8	102.6	102.3	75.0
	10	104.0	103.5	79.1	103.7	103.3	71.8
	20	104.8	104.3	78.0	104.5	104.0	70.3
$\Delta = 0.25$	2	103.8	103.4	86.0	103.4	103.0	79.5
	3	105.4	104.9	82.0	105.0	104.4	74.5
	4	106.5	105.8	79.9	105.9	105.3	71.9
	5	107.2	106.5	78.7	106.6	105.9	70.4
	10	108.9	108.1	76.2	108.3	107.5	67.2
	20	110.2	109.3	75.1	109.4	108.5	65.7
$\Delta = 0.50$	2	111.0	110.2	85.3	110.0	109.2	77.6
	3	116.6	115.5	81.9	115.1	113.9	72.1
	4	120.2	118.9	80.5	118.3	117.0	69.7
	5	122.9	121.3	79.9	120.7	119.2	68.5
	10	130.1	128.2	79.5	127.1	125.3	66.6
	20	136.3	134.1	80.6	132.5	130.3	66.5

are compared with different fixed sample size designs.

It becomes apparent that the three designs have quite different operating characteristics. In general, the highest gain in reduction of average sample size under the alternative is observed by adding the first interim analysis ($K = 2$). With each additional interim analysis this gain steadily decreases and the maximum sample size and $ASN(0)$ increase. For the intermediate design there might be some further gains with respect to $ASN(\delta)$ for more than $K = 5$ stages. However, it should also be kept in mind that interim analyses in clinical trials are associated with considerable operational challenges, so that more than five stages are generally not feasible in practice. We will go into a little more detail on this issue at the end of this section. The designs by Pocock ($\Delta = 0.5$) have the highest maximum sample sizes and average sample numbers under H_0, but benefit from very small expected sample sizes at $\theta = \delta$. The average sample numbers under H_1 of the O'Brien Fleming designs are also sufficiently low, while the maximum sample size and $ASN(0)$ are only slightly higher than the fixed sample size. Not surprisingly, the intermediate design with $\Delta = 0.25$ represents a trade-off between the other two designs.

When choosing an appropriate group sequential design from the Δ-class one has to balance low expected against low maximum sample size. Designs with higher values of Δ tend to have higher maximum sample sizes and lower average sample numbers and vice versa for lower Δ values. In general, it can be shown that N_{max} is monotonically increasing with respect to Δ, whereas $ASN(\delta)$ as a function of Δ has a global minimum that can be found by numerical optimisation. Wang and Tsiatis (1987) showed that the optimal designs from the Δ-class have $ASN(\delta)$ values that are almost identical to those of the general optimal designs without any constraints on the boundary value structure that where investigated by Pocock (1982). For $\alpha = 0.025$, $K = 2, ..., 5$ and $1 - \beta = 0.80, 0.90$ these so-called *approximately optimal designs* are obtained for Δ values around 0.4. Another interesting type of designs with low expected and not too high maximum sample size can be determined by minimising the sum $N_{max} + ASN(\delta)$.

The validity of the proposed group sequential tests is based on equal stage sizes, a restriction that is not always sensible. If the trial is planned with unequal stage sizes, appropriate rejection boundaries for the respective sample size allocation among the stages can be calculated by means of Equation (3.4). It also turns out that the usual boundaries based on equal stage sizes are fairly robust with respect to deviations from an equal allocation as it has been shown by Proschan et al. (1992). However, in clinical trials interim analyses are usually conducted at scheduled dates, so that irregular recruitment and drop outs can lead to considerably different stage sizes. Besides the worst case scenario solution that has been proposed by Wassmer (1999), a convenient and flexible solution to deal with unpredictable but data independent stage sizes is the so-called *error spending approach*.

3.1.4 The Error Spending Approach

The key idea of error spending has been first mentioned by Slud and Wei (1982) and derives from the fact that the overall type I error rate can be written as the sum of the K stage-wise rejection probabilities under H_0. Thus, Equation (3.4) is equivalent to

$$\underbrace{P_{\theta=0}(Z_1 \geq b_1)}_{\pi_1} + \underbrace{P_{\theta=0}(Z_1 < b_1, Z_2 \geq b_2)}_{\pi_2} + ... + \underbrace{P_{\theta=0}(Z_1 < b_1, ..., Z_{K-1} < b_{K-1}, Z_K \geq b_K)}_{\pi_K} = \alpha$$

By partitioning the type I error into probabilities prespecified $\pi_1, ..., \pi_K$ which sum up to α, appropriate boundaries $b_1, ..., b_K$ are determined successively at each stage based on the data observed by then. Starting with the first critical value after stage 1, which is easily obtained as

$$b_1 = z_{1-\pi_1}, \tag{3.12}$$

the boundary values at stage $k = 2, ..., K$ can be found as a solution of

$$P_{\theta=0}(Z_1 < b_1, ..., Z_{k-1} < b_{k-1}, Z_k \geq b_k) = \pi_k, \tag{3.13}$$

where the left-hand side can be calculated by means of the distribution function of the k-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix determined as a submatrix of $\mathbf{\Sigma}$ in (3.7).

Although the procedure by Slud and Wei (1982) guarantees type I error control irrespective of the observed information pattern, it has some limitations. On the one hand, the maximum number of analyses K needs to be fixed in advance and the desired power is not always obtained. On the other hand, it seems more desirable to choose the amount of spent error $\pi_1, ..., \pi_K$ based on the actually observed information levels $\mathscr{I}_1, ..., \mathscr{I}_K$. For instance, if at the first interim analysis the information level is low it seems reasonable to spend less type I error, i.e. reduce π_1, in order to save type I error for the subsequent, more informative stages.

The method proposed by Lan and DeMets (1983) is capable of dealing with these issues in a simple but efficient way. First of all, one needs to specify the targeted maximum information level $\mathscr{I}_{max}$, which is the reason why such designs are also called *maximum information designs*. Next, a non-decreasing error spending function $f(t)$ has to be defined, satisfying $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$. The type I error is then partitioned according to this error spending function, in such a way that $f(t)$ specifies the cumulative type I error spent when the information $t \cdot \mathscr{I}_{max}$ is observed. Thus, the type I error is allocated in the following way:

$$\begin{aligned}
&\pi_1 = f(\mathscr{I}_1/\mathscr{I}_{max}), \\
&\pi_k = f(\mathscr{I}_k/\mathscr{I}_{max}) - f(\mathscr{I}_{k-1}/\mathscr{I}_{max}) \text{ for } k = 2, ..., K.
\end{aligned}$$

The corresponding rejection boundaries $b_1, ..., b_K$ are then determined successively as in the method by Slud and Wei (1982) to satisfy Equations (3.12) and (3.13). With these boundary values the testing procedure takes the same form as for the classical group sequential tests.

Interestingly, there is no need to predefine the number of stages K in an error spending design. The final stage simply is the first stage where the observed information level exceeds the maximum information $\mathscr{I}_{max}$, so that $K = \min\{k \mid k \in \mathbb{N}, \mathscr{I}_k/\mathscr{I}_{max} \geq 1\}$. According to the definition of the error spending function with $f(t) = \alpha$ for $t \geq 1$, the type I error is bounded above by α as $\sum_{i=1}^{K} \pi_i = \alpha$.

There have been several proposals on the choice of the error spending function $f(t)$ in the literature. Lan and DeMets (1983) presented the two functions

$$\begin{aligned} f_{POC}(t) &= \min\left\{2 - 2\Phi\left(z_{1-\alpha/2}/\sqrt{t}\right), \alpha\right\} \text{ and} \\ f_{OBF}(t) &= \min\left\{\alpha \log\left(1 + (e-1)t\right), \alpha\right\}, \end{aligned}$$

that result in rejection boundaries similar to those suggested by Pocock and O'Brien & Fleming, respectively. A more general family of error spending functions has been proposed by Kim and DeMets (1987), who defined the ρ-class as

$$f_{KD}(t) = \min\left\{\alpha t^{\rho}, \alpha\right\} \text{ with } \rho > 0.$$

It turns out that the ρ-class represents some kind of an error spending counterpart of the Δ-class by Wang & Tsiatis. For $\rho = 1$ and $\rho = 3$ we obtain critical values that are close to the designs by Pocock and O'Brien & Fleming, respectively, which are special cases of the Δ-class ($\Delta = 0.5$ and $\Delta = 0$). The design with $\rho = 2$ consequently is similar to the intermediate design of the Wang & Tsiatis family with $\Delta = 0.25$. Moreover, the maximum sample size N_{max} is decreasing with increasing ρ and approximately optimal designs with respect to minimising $ASN(\delta)$ can be found via numerical optimisation methods.

Hwang et al. (1990) suggested another family of error spending designs which is in terms of truncated exponential distributions. For a predefined parameter $\gamma \in \mathbb{R}$ they investigated the error spending functions

$$f_{HSD}(t) = \begin{cases} \alpha\left(1 - e^{-\gamma t}\right) / \left(1 - e^{-\gamma}\right) & \text{for } \gamma \neq 0, \\ \alpha t & \text{for } \gamma = 0. \end{cases}$$

As with the ρ-class, it turns out that the designs with $\gamma = 1$ and $\gamma = -4$ have rejection boundaries that are similar to Pocock and O'Brien & Fleming boundaries, respectively. Furthermore, the maximum sample size N_{max} is an increasing function of γ and Hwang et al. (1990) showed that the γ-class also possesses the (approximately) optimal property of minimising the average sample number $ASN(\delta)$ for suitable γ.

3.1.5 Practical Considerations

It should be noted that the group sequential tests presented above rely on the assumption of a known variance σ^2 which is normally not the case in practice. In order to apply the derived group sequential boundaries to data with unknown variance, Pocock (1977) proposed a simple but effective procedure that approximately controls the type I error rate:

1. Estimate the variance in (3.2) with the unbiased pooled variance estimator $\hat{\sigma}^2$.

2. Choose appropriate rejection boundaries $b_1, ..., b_K$ and determine the stage-wise α-levels as $\alpha_k = 1 - \Phi(b_k)$, $k = 1, ..., K$.

3. Define the *adjusted* boundary values by $\tilde{b}_k = t_{1-\alpha_k,\nu}$ with $\nu = n_A^{(k)} + n_B^{(k)} - 2$ degrees of freedom for $k = 1, ..., K$.

By using the adjusted rejection boundaries $\tilde{b}_1, ..., \tilde{b}_K$, we obtain approximate group sequential t-tests with only minor departures from the desired type I error rates. Exact calculations for sequential and group sequential t-tests have been proposed by Jennison and Turnbull (1991), who also dealt with exact sequential χ^2 and F tests. For most applications the tests for known variance are quite robust with regard to type I error control when applied in the unknown variance setting.

The previous subsections dealt with group sequential methodology to reject H_0. In a placebo-controlled superiority trial this corresponds to proving the efficacy of the new drug. It seems reasonable, not only for ethical reasons, to additionally implement so-called *futility boundaries* in such trials, i.e. the possibility of early stopping to accept H_0. The course of a clinical study is usually monitored by a Data Safety Monitoring Board (DSMB) which could also benefit from this statistical monitoring tool. In general, there are two types of futility boundaries, *binding* and *non-binding*. As the names imply, the trial must be stopped with accepting H_0 once a binding futility boundary is crossed, whereas non-binding futility boundaries only serve as a guidance and the study can be continued once the test statistic falls below them. Although this additional flexibility comes along with increased maximum sample sizes, non-binding futility boundaries are more commonly applied in clinical trials, not least because they can be easily determined completely independent of the rejection boundaries.

Last but not least it should not be overlooked that, besides all the benefits such as substantial sample size savings, the realisation of a group sequential design in clinical trials is associated with considerable operational challenges compared with a conventional fixed sample size design. First of all, it is vital to set up an Independent Data Monitoring Committee (IDMC) in order to maintain the validity and integrity of the trial. The information that is revealed by the IDMC after an interim analysis should be furthermore kept to a minimum. The patient population might change due to published information about the treatment efficacy, an issue that should be assessed by comparing the results from the different stages. Another issue in trials with interim analyses is that there often is a delay between start of treatment and observation of the endpoint, so that some patients are still ‘in the pipeline’ at each interim analysis. In accordance with the ITT principle of analysing all randomised patients, these *overrunning* patients should be included in the primary analysis. Differences between an additional analysis excluding these patients should be critically discussed. The error spending approach presented in the previous subsection provides a simple and flexible solution to implement this strategy.

3.2 Group Sequential Designs for Three-Arm Non-Inferiority Trials

The previous section showed that the application of group sequential methods in clinical trials can result in significant sample size and time savings. Particularly because Section 2.3.3 showed that the power of the test vs. placebo superiority comparison is very high under the optimal allocation of the fixed sample size design, the benefits of a group sequential design in a three-arm non-inferiority trial are expected to be substantial. Furthermore, the additional possibility to stop allocating patients to placebo is another key benefit of implementing interim analyses that could help to overcome ethical concerns.

Due to the fact that there is more than one hypothesis under consideration in three-arm non-inferiority trials, the application of group sequential methodology obviously results in two layers of multiplicity, namely multiple hypotheses and repeated significance testing. In order to account for these two multiplicity issues, we will present an appropriate testing procedure that controls the overall probability of committing at least one type I error. After describing the general testing procedure and determining formulas for the overall power and the maximum and expected sample sizes, the subsequent part of this section will provide a detailed comparison of the proposed design with the optimal fixed design derived in Section 2.3.3. A special focus will also be placed on the choice of the rejection boundaries. It should be noted that a part of the results has already been presented in Schlömer and Brannath (2013).

3.2.1 General Design

Denote by $X_{T,i} \sim N(\mu_T, \sigma^2)$, $i = 1, 2, ...$, $X_{C,i} \sim N(\mu_C, \sigma^2)$, $i = 1, 2, ...$, and $X_{P,i} \sim N(\mu_P, \sigma^2)$, $i = 1, 2, ...$, the mutually independent responses of patients allocated to the test, the control and the placebo treatment, respectively, with common known variance σ^2. As in Section 2.3 we want to assess the hierarchically ordered set of hypothesis given in (2.9), i.e. the proof of efficacy for the test treatment ($H_{0,TP}^{(s)} : \mu_T \leq \mu_P$) and the non-inferiority comparison between test and control ($H_{0,TC}^{(n)} : \mu_T \leq \mu_C - \Delta_{ni}$). We now want to apply the group sequential framework in the following way: Start with assessing $H_{0,TP}^{(s)}$ at each interim analysis. If $H_{0,TP}^{(s)}$ is rejected, say at stage k^*, $1 \leq k^* \leq K$, drop the placebo arm and proceed with testing $H_{0,TC}^{(n)}$ at stages $k^*, ..., K$. If furthermore $H_{0,TC}^{(n)}$ is rejected, stop the trial and state that the test treatment is both superior to placebo and non-inferior to the control treatment.

Test Procedure

Let $n_T^{(1)}, ..., n_T^{(K)}$, $n_C^{(1)}, ..., n_C^{(K)}$ and $n_P^{(1)}, ..., n_P^{(K)}$ denote the cumulative sample sizes at stage $1, ..., K$ of the test, control and placebo group, respectively. The cumulative sample means of the three treatment groups shall be denoted as $\bar{X}_T^{(1)}, ..., \bar{X}_T^{(K)}$, $\bar{X}_C^{(1)}, ..., \bar{X}_C^{(K)}$ and $\bar{X}_P^{(1)}, ..., \bar{X}_P^{(K)}$ and defined as $\bar{X}_D^{(k)} = \sum_{i=1}^{n_D^{(k)}} X_{D,i} / n_D^{(k)}$ for $D = T, C, P$ and $k = 1, ..., K$. In accordance with (2.10) and (3.2) the group sequential test statistics $Z_{TP}^{(1)},\dots,Z_{TP}^{(K)}$ for testing the first and $Z_{TC}^{(1)},\dots,Z_{TC}^{(K)}$ for testing the second null hypothesis are calculated as

$$\begin{aligned} Z_{TP}^{(k)} &= \frac{\bar{X}_T^{(k)}-\bar{X}_P^{(k)}}{\sigma\sqrt{\frac{1}{n_T^{(k)}}+\frac{1}{n_P^{(k)}}}} = \left(\bar{X}_T^{(k)}-\bar{X}_P^{(k)}\right)\sqrt{\mathscr{I}_{TP}^{(k)}}, \\ Z_{TC}^{(k)} &= \frac{\bar{X}_T^{(k)}-\bar{X}_C^{(k)}+\Delta_{ni}}{\sigma\sqrt{\frac{1}{n_T^{(k)}}+\frac{1}{n_C^{(k)}}}} = \left(\bar{X}_T^{(k)}-\bar{X}_C^{(k)}+\Delta_{ni}\right)\sqrt{\mathscr{I}_{TC}^{(k)}}, \quad k=1,\dots,K, \end{aligned}$$

where $\mathscr{I}_{TP}^{(k)}$ and $\mathscr{I}_{TC}^{(k)}$ denote the stage-wise information levels of the respective comparison. By combining the fixed sequence procedure of the fixed sample size design in (2.11) and the common group sequential test procedure given in (3.3), the test procedure of the group sequential three-arm non-inferiority design takes the following form:

Step I: At stage $k=1,\dots,K-1$
if $Z_{TP}^{(k)} \geq b_{TP}^{(k)}$, reject $H_{0,TP}^{(s)}$, drop placebo and proceed to step II,
if $Z_{TP}^{(k)} < b_{TP}^{(k)}$, continue to stage $k+1$.
At stage K
if $Z_{TP}^{(K)} \geq b_{TP}^{(K)}$, reject $H_{0,TP}^{(s)}$ and proceed to step II,
if $Z_{TP}^{(K)} < b_{TP}^{(K)}$, accept $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$, stop the trial.
Step II: Suppose $H_{0,TP}^{(s)}$ has been rejected at stage $k^*, 1 \leq k^* \leq K$. (3.14)
At stage $k=k^*,\dots,K-1$
if $Z_{TC}^{(k)} \geq b_{TC}^{(k)}$, reject $H_{0,TC}^{(n)}$, and stop the trial,
if $Z_{TC}^{(k)} < b_{TC}^{(k)}$, continue to stage $k+1$.
At stage K
if $Z_{TC}^{(K)} \geq b_{TC}^{(K)}$, reject $H_{0,TC}^{(n)}$ and stop the trial,
if $Z_{TC}^{(K)} < b_{TC}^{(K)}$, accept $H_{0,TC}^{(n)}$ and stop the trial.

Type I Error and Rejection Boundaries

The group sequential boundaries $b_{TP}^{(1)},\dots,b_{TP}^{(K)}$ and $b_{TC}^{(1)},\dots,b_{TC}^{(K)}$ in (3.14) of the test vs. placebo superiority comparison and the test vs. control non-inferiority test are determined to satisfy the equations

$$1-P_{\theta_{TP}=0}\left(\bigcap_{k=1}^{K}\left\{Z_{TP}^{(k)}<b_{TP}^{(k)}\right\}\right)=\alpha, \tag{3.15}$$

$$1-P_{\theta_{TC}=-\Delta_{ni}}\left(\bigcap_{k=1}^{K}\left\{Z_{TC}^{(k)}<b_{TC}^{(k)}\right\}\right)=\alpha. \tag{3.16}$$

That means, for both null hypotheses $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ we define separate group sequential boundaries each at significance level α. Through this, the family-wise type I error rate is controlled by α in the strong sense.

PROOF: *To show that the testing procedure described in* (3.14) *with rejection boundaries satisfying* (3.15) *and* (3.16) *provides strong family-wise error rate control, we need to consider all possible configurations of true and false null hypotheses $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ or equivalently all constellations of the parameters $\theta_{TP} = \mu_T - \mu_P$ and $\theta_{TC} = \mu_T - \mu_C$.*

$\underline{\theta_{TP} \leq 0:}$ *Irrespective of whether $\theta_{TC} \leq -\Delta_{ni}$ or $\theta_{TC} > -\Delta_{ni}$, the procedure must, due to its hierarchical nature, at least reject $H_{0,TP}^{(s)}$ in order to commit a type I error. Therefore we have*

$$\begin{aligned} P_{\theta_{TP}\leq 0}\left(\textit{Commit a type I error}\right) &\leq P_{\theta_{TP}\leq 0}\left(\bigcup_{k=1}^{K}\left\{H_{0,TP}^{(s)} \textit{ is rejected at stage } k\right\}\right) \\ &= 1 - P_{\theta_{TP}\leq 0}\left(\bigcap_{k=1}^{K}\left\{Z_{TP}^{(k)} < b_{TP}^{(k)}\right\}\right) \overset{(3.15)}{\leq} \alpha. \end{aligned}$$

$\underline{\theta_{TP} > 0, \theta_{TC} \leq -\Delta_{ni}:}$ *According to the hierarchical ordering, both null hypotheses need to be rejected in order to commit a type I error. Denote by A_k, $0 \leq k \leq K$, the event that the procedure rejects $H_{0,TP}^{(s)}$ at stage k ($k = 0$ means no rejection of $H_{0,TP}^{(s)}$ at any stage) and $H_{0,TC}^{(n)}$ is not rejected at stages $k, ..., K$. That means A_k is given as*

$$A_k = \begin{cases} \bigcap\limits_{m=1}^{K}\left\{Z_{TP}^{(m)} < b_{TP}^{(m)}\right\}, & k = 0, \\ \left\{Z_{TP}^{(1)} \geq b_{TP}^{(1)}\right\} \cap \bigcap\limits_{l=1}^{K}\left\{Z_{TC}^{(l)} < b_{TC}^{(l)}\right\}, & k = 1, \\ \bigcap\limits_{m=1}^{k-1}\left\{Z_{TP}^{(m)} < b_{TP}^{(m)}\right\} \cap \left\{Z_{TP}^{(k)} \geq b_{TP}^{(k)}\right\} \cap \bigcap\limits_{l=k}^{K}\left\{Z_{TC}^{(l)} < b_{TC}^{(l)}\right\}, & 2 \leq k \leq K. \end{cases} \tag{3.17}$$

Because $\bigcap\limits_{k=1}^{K}\left\{Z_{TC}^{(k)} < b_{TC}^{(k)}\right\} \subseteq \bigcup\limits_{k=0}^{K} A_k$, the probability of committing a type I error is given as

$$1 - P_{\theta_{TP}>0,\theta_{TC}\leq-\Delta_{ni}}\left(\bigcup_{0\leq k\leq K} A_k\right) \leq 1 - P_{\theta_{TC}\leq-\Delta_{ni}}\left(\bigcap_{k=1}^{K}\left\{Z_{TC}^{(k)} < b_{TC}^{(k)}\right\}\right) \overset{(3.16)}{\leq} \alpha.$$

For $\theta_{TP} > 0$ and $\theta_{TC} > -\Delta_{ni}$ both null hypotheses are false, so that no type I error can be committed. Consequently, the type I error rate is controlled in the strong sense. □

Analogously to Section 3.1.2 it can be easily shown that Equations (3.15) and (3.16) are equivalent to

$$\Phi^{\boldsymbol{\Sigma}_{TP}}\left(b_{TP}^{(1)}, ..., b_{TP}^{(K)}\right) = \Phi^{\boldsymbol{\Sigma}_{TC}}\left(b_{TC}^{(1)}, ..., b_{TC}^{(K)}\right) = 1 - \alpha,$$

where the corresponding variance-covariance matrices $\boldsymbol{\Sigma}_{TP}$ and $\boldsymbol{\Sigma}_{TC}$ can be determined according to (3.6) as

$$\boldsymbol{\Sigma}_{TP} = \left(\sqrt{\frac{\mathscr{I}_{TP}^{(\min(i,j))}}{\mathscr{I}_{TP}^{(\max(i,j))}}} \right)_{1 \le i,j \le K} \quad \text{and} \quad \boldsymbol{\Sigma}_{TC} = \left(\sqrt{\frac{\mathscr{I}_{TC}^{(\min(i,j))}}{\mathscr{I}_{TC}^{(\max(i,j))}}} \right)_{1 \le i,j \le K} . \tag{3.18}$$

As we have seen, the rejection boundaries for both hypotheses can be determined completely independent of each other within the common group sequential framework presented in Section 3.1.2. Most of the established statistical software packages provide tools for designing and analysing group sequential trials nowadays, making it easy to implement the proposed procedure. For instance, SAS/STAT® software (SAS Institute Inc., 2011) contains the procedure `SEQDESIGN` and the package `gsDesign` (Anderson, 2013) is available for the software environment R (R Core Team, 2013).

Power and Sample Sizes

The power for rejecting $H_{0,TP}^{(s)}$, which shall be denoted as $1-\beta_{TP}$, can be easily determined according to Equation (3.9) as

$$\begin{aligned} 1-\beta_{TP} &= 1 - P_{\theta_{TP}}\left(\bigcap_{k=1}^{K} \left\{ Z_{TP}^{(k)} < b_{TP}^{(k)} \right\} \right) \\ &= 1 - \Phi^{\boldsymbol{\Sigma}_{TP}}\left(b_{TP}^{(1)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(1)}}, \ldots, b_{TP}^{(K)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(K)}} \right). \end{aligned} \tag{3.19}$$

Determination of the overall power $1-\beta$ to reject both null hypotheses $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ is more complicated than in the fixed sample size design. With the events A_k, $k = 0, \ldots, K$, defined in (3.17), the overall power of the procedure is given as

$$\begin{aligned} 1-\beta &= 1 - P_{\theta_{TP},\theta_{TC}}\left(\bigcup_{k=0}^{K} A_k \right) \\ &= 1 - P_{\theta_{TP}}(A_0) - \sum_{k=1}^{K} P_{\theta_{TP},\theta_{TC}}(A_k) \\ &= 1 - \beta_{TP} - \sum_{k=1}^{K} P_{\theta_{TP},\theta_{TC}}(A_k), \end{aligned} \tag{3.20}$$

with the power of the test versus placebo superiority comparison $1-\beta_{TP}$ given in Equation (3.19). Because obviously $\bigcup_{k=1}^{K} A_k \not\subseteq \bigcap_{k=1}^{K}\{Z_{TC}^{(k)} < b_{TC}^{(k)}\}$, we can see from Equation (3.20) that the simple intuitive lower bound $1-\beta_{TP}-\beta_{TC}$ for the overall power cannot be easily determined as in the fixed design, when β_{TC} denotes the probability of committing a type II error with a common group sequential design, i.e. $\beta_{TC} = P_{\theta_{TC}}(\bigcap_{k=1}^{K}\{Z_{TC}^{(k)} < b_{TC}^{(k)}\})$. Moreover, the overall power could also be determined by using the events R_{k_1,k_2}, $1 \le k_1 \le k_2 \le K$, that the first hypothesis

$H^{(s)}_{0,TP}$ is rejected at stage k_1 and the second hypothesis $H^{(n)}_{0,TC}$ is rejected at stage k_2, that means

$$R_{k_1,k_2} = \begin{cases} \left\{Z^{(1)}_{TP} \geq b^{(1)}_{TP}\right\} \cap \left\{Z^{(1)}_{TC} \geq b^{(1)}_{TC}\right\}, & k_1 = k_2 = 1, \\ \left\{Z^{(1)}_{TP} \geq b^{(1)}_{TP}\right\} \cap \bigcap_{l=1}^{k_2-1} \left\{Z^{(l)}_{TC} < b^{(l)}_{TC}\right\} \cap \left\{Z^{(k_2)}_{TC} \geq b^{(k_2)}_{TC}\right\}, & 1 = k_1 < k_2 \leq K, \\ \bigcap_{m=1}^{k_1-1} \left\{Z^{(m)}_{TP} < b^{(m)}_{TP}\right\} \cap \left\{Z^{(k_1)}_{TP} \geq b^{(k_1)}_{TP}\right\} \cap \left\{Z^{(k_1)}_{TC} \geq b^{(k_1)}_{TC}\right\}, & 1 < k_1 = k_2 \leq K, \\ \bigcap_{m=1}^{k_1-1} \left\{Z^{(m)}_{TP} < b^{(m)}_{TP}\right\} \cap \left\{Z^{(k_1)}_{TP} \geq b^{(k_1)}_{TP}\right\} \cap \bigcap_{l=k_1}^{k_2-1} \left\{Z^{(l)}_{TC} < b^{(l)}_{TC}\right\} \cap \left\{Z^{(k_2)}_{TC} < b^{(k_2)}_{TC}\right\}, & 1 < k_1 < k_2 \leq K. \end{cases}$$

The overall power would then be given as

$$1 - \beta = P_{\theta_{TP},\theta_{TC}}\left(\bigcup_{1 \leq k_1 \leq k_2 \leq K} R_{k_1,k_2}\right) = \sum_{1 \leq k_1 \leq k_2 \leq K} P_{\theta_{TP},\theta_{TC}}\left(R_{k_1,k_2}\right).$$

This formula seems much more intuitive than Equation (3.20), but it is computationally more intensive as it involves calculating $\frac{K(K+1)}{2}$ instead of $K+1$ multidimensional integrals. Therefore, we will make use of Equation (3.20).

In order to calculate the probabilities $P_{\theta_{TP},\theta_{TC}}(A_k)$, $k = 1, \ldots, K$, in (3.20), we need to take a closer look at the distribution of the vector of test statistics $\boldsymbol{Z} = (Z^{(1)}_{TP}, \ldots, Z^{(K)}_{TP}, Z^{(1)}_{TC}, \ldots, Z^{(K)}_{TC})'$. As each component of $\boldsymbol{Z}$ itself is a linear combination of normally distributed random variables, all linear combinations of the components obviously follow a normal distribution. Consequently, $\boldsymbol{Z}$ follows a $2K$-variate normal distribution, where the mean vector is easily determined as

$$\boldsymbol{\mu} = \left(\mu_i\right)'_{1 \leq i \leq 2K}, \text{ where } \mu_i = \begin{cases} \theta_{TP}\sqrt{\mathscr{I}^{(i)}_{TP}}, & 1 \leq i \leq K, \\ (\theta_{TC} + \Delta_{ni})\sqrt{\mathscr{I}^{(i-K)}_{TC}}, & K+1 \leq i \leq 2K. \end{cases} \tag{3.21}$$

The covariances $Cov(Z^{(k_1)}_{TP}, Z^{(k_2)}_{TP})$ and $Cov(Z^{(k_1)}_{TC}, Z^{(k_2)}_{TC})$ are determined with (3.6) analogously to the common group sequential design. According to (3.5) we have for $k_1, k_2 \in \{1, \ldots, K\}$

$$\begin{aligned} Cov\left(Z^{(k_1)}_{TP}, Z^{(k_2)}_{TC}\right) &= \sqrt{\mathscr{I}^{(k_1)}_{TP}\mathscr{I}^{(k_2)}_{TC}}\,Cov\left(\bar{X}^{(k_1)}_T - \bar{X}^{(k_1)}_P, \bar{X}^{(k_2)}_T - \bar{X}^{(k_2)}_C + \Delta_{ni}\right) \\ &= \sqrt{\mathscr{I}^{(k_1)}_{TP}\mathscr{I}^{(k_2)}_{TC}}\,Cov\left(\bar{X}^{(k_1)}_T, \bar{X}^{(k_2)}_T\right) \\ &= \frac{\sigma^2}{n^{(\max(k_1,k_2))}_T}\sqrt{\mathscr{I}^{(k_1)}_{TP}\mathscr{I}^{(k_2)}_{TC}}, \end{aligned} \tag{3.22}$$

so that the variance-covariance matrix of $\boldsymbol{Z}$ is finally obtained as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{TP} & \boldsymbol{\Sigma}_{TCP} \\ \boldsymbol{\Sigma}'_{TCP} & \boldsymbol{\Sigma}_{TC} \end{pmatrix} \text{ with } \boldsymbol{\Sigma}_{TCP} = \left(\frac{\sigma^2}{n^{(\max(i,j))}_T}\sqrt{\mathscr{I}^{(i)}_{TP}\mathscr{I}^{(j)}_{TC}}\right)_{1 \leq i,j \leq K}, \tag{3.23}$$

where i denotes the index of the rows and j the index of the columns and $\boldsymbol{\Sigma}_{TP}$ and $\boldsymbol{\Sigma}_{TC}$ are given in (3.18).

For a specific k, $1 \leq k \leq K$, calculation of $P_{\theta_{TP},\theta_{TC}}(A_k)$ requires knowledge of the distribution of the vector $(Z_{TP}^{(1)},...,Z_{TP}^{(k)},Z_{TC}^{(k)},...,Z_{TC}^{(K)})'$, which obviously is a subvector of $\boldsymbol{Z}$. Hence, it is $K+1$-variate normally distributed, where the mean vector and covariance matrix are easily determined as a subvector of $\boldsymbol{\mu}$ given in (3.21) and a submatrix of $\boldsymbol{\Sigma}$ in (3.23), respectively. More precisely they are obtained by deleting the entries $k+1,...,K+k-1$ in $\boldsymbol{\mu}$ and the rows and columns $k+1,...,K+k-1$ in $\boldsymbol{\Sigma}$.

In contrast to previous calculations, the cumulative multivariate normal distribution function cannot be used to calculate $P_{\theta_{TP},\theta_{TC}}(A_k)$, because A_k includes $\{Z_{TP}^{(k)} \geq b_{TP}^{(k)}\}$. Let therefore $\Phi^{\boldsymbol{\Sigma}}(\boldsymbol{a};\boldsymbol{b})$ with $\boldsymbol{a} = (a_i)'_{1\leq i\leq n}, \boldsymbol{b} = (b_i)'_{1\leq i\leq n} \in \mathbb{R}^n$ denote the distribution function of the multivariate normal distribution with mean vector $\boldsymbol{0} = (0)'_{1\leq i\leq n}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ evaluated at lower limits $a_1,...,a_n$ and upper limits $b_1,...,b_n$. That means, we have

$$\Phi^{\boldsymbol{\Sigma}}(\boldsymbol{a};\boldsymbol{b}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^n}} \int_{a_1}^{b_1}\int_{a_2}^{b_2}\cdots\int_{a_n}^{b_n} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}} d\boldsymbol{x},$$

so that the relationship $\Phi^{\boldsymbol{\Sigma}}((-\infty)'_{1\leq i\leq n};\boldsymbol{b}) = \Phi^{\boldsymbol{\Sigma}}(b_1,...,b_n)$ with the cumulative distribution function holds. Through this, we have

$$P_{\theta_{TP},\theta_{TC}}(A_k) = \Phi^{\boldsymbol{\Sigma}(k+1,...,K+k-1)}(\boldsymbol{a}_k;\boldsymbol{b}_k) \quad \text{for } 1 \leq k \leq K, \tag{3.24}$$

where the vectors $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ are defined as

$$\begin{aligned} \boldsymbol{a}_k = \left(a_{k,i}\right)'_{1\leq i\leq K+1} \text{ with } a_{k,i} &= \begin{cases} b_{TP}^{(i)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(i)}}, & i = k, \\ -\infty, & \text{else}, \end{cases} \\ \boldsymbol{b}_k = \left(b_{k,i}\right)'_{1\leq i\leq K+1} \text{ with } b_{k,i} &= \begin{cases} b_{TP}^{(i)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(i)}}, & 1 \leq i \leq k-1, \\ \infty, & i = k, \\ b_{TC}^{(i-1)} - (\theta_{TC} + \Delta_{ni})\sqrt{\mathscr{I}_{TC}^{(i-1)}}, & k+1 \leq i \leq K+1. \end{cases} \end{aligned} \tag{3.25}$$

According to (3.20), the overall power of the procedure can be determined as the difference between the power to reject $H_{0,TP}^{(s)}$ given in (3.19) and the sum of the probabilities $P_{\theta_{TP},\theta_{TC}}(A_k)$, $1 \leq k \leq K$, that are calculated according to (3.24) and (3.25). That means, we have

$$\begin{aligned} 1-\beta = 1 &- \Phi^{\boldsymbol{\Sigma}_{TP}}\left(b_{TP}^{(1)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(1)}},...,b_{TP}^{(K)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(K)}}\right) \\ &- \sum_{k=1}^{K} \Phi^{\boldsymbol{\Sigma}(k+1,...,K+k-1)}(\boldsymbol{a}_k;\boldsymbol{b}_k), \end{aligned} \tag{3.26}$$

where the vectors $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$, $1 \leq k \leq K$, are given in (3.25) and the variance-covariance matrices $\boldsymbol{\Sigma}_{TP}$ and $\boldsymbol{\Sigma}$ are given in (3.18) and (3.23), respectively. With prespecified sample size allocations between the groups and across the stages, the required sample size is obtained as the solution of Equation (3.26).

Expected Sample Sizes

Of particular interest for the proposed design are the expected sample sizes of the placebo group and overall, which shall be denoted as $ASn_P(\theta_{TP})$ and $ASN(\theta_{TP}, \theta_{TC})$, respectively. It can be easily seen that the former only depends on the true treatment difference between test and placebo and can be derived analogously to the common group sequential design. According to (3.10) we have

$$\begin{aligned} ASn_P(\theta_{TP}) &= n_P^{(1)} + \sum_{k=2}^{K} \left(n_P^{(k)} - n_P^{(k-1)}\right) P_{\theta_{TP}} \left(\bigcap_{m=1}^{k-1} \left\{ Z_{TP}^{(m)} < b_{TP}^{(m)} \right\} \right) \qquad (3.27) \\ &= n_P^{(1)} + \sum_{k=2}^{K} \left(n_P^{(k)} - n_P^{(k-1)}\right) \Phi^{\boldsymbol{\Sigma}_{TP}(k,\ldots,K)} \left(b_{TP}^{(1)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(1)}}, \ldots, b_{TP}^{(k-1)} - \theta_{TP}\sqrt{\mathscr{I}_{TP}^{(k-1)}} \right). \end{aligned}$$

In contrast to the average sample number of the placebo group, the expected test and control group sizes, which shall be denoted as $ASn_T(\theta_{TP}, \theta_{TC})$ and $ASn_C(\theta_{TP}, \theta_{TC})$, respectively, depend on the rejection or non-rejection of both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$. Let us therefore denote by E_k, $2 \le k \le K$, the event that the procedure enters stage k. Obviously, there are several constellations of observed test statistics $Z_{TP}^{(1)}, \ldots, Z_{TP}^{(k-1)}$ and $Z_{TC}^{(1)}, \ldots, Z_{TC}^{(k-1)}$ that results in entering stage k. For example, the trial enters stage 2 if $H_{0,TP}^{(s)}$ is not rejected at the first interim analysis. If $H_{0,TP}^{(s)}$ is rejected, but $H_{0,TC}^{(n)}$ is not rejected at the first interim analysis, the trial also enters the second stage. Thus, we have

$$\begin{aligned} E_k &= \bigcup_{k_1=0}^{k-1} E_{k_1,k} \text{ for } 2 \le k \le K, \text{ where} \\ E_{k_1,k} &= \begin{cases} \bigcap_{m=1}^{k-1} \left\{ Z_{TP}^{(m)} < b_{TP}^{(m)} \right\}, & k_1 = 0, \\ \left\{ Z_{TP}^{(1)} \ge b_{TP}^{(1)} \right\} \cap \bigcap_{l=1}^{k-1} \left\{ Z_{TC}^{(l)} < b_{TC}^{(l)} \right\}, & k_1 = 1, \\ \bigcap_{m=1}^{k_1-1} \left\{ Z_{TP}^{(m)} < b_{TP}^{(m)} \right\} \cap \left\{ Z_{TP}^{(k_1)} \ge b_{TP}^{(k_1)} \right\} \cap \bigcap_{l=k_1}^{k-1} \left\{ Z_{TC}^{(l)} < b_{TC}^{(l)} \right\}, & 2 \le k_1 \le k-1. \end{cases} \qquad (3.28) \end{aligned}$$

In other words, $E_{k_1,k}$ corresponds to the event that $H_{0,TP}^{(s)}$ is rejected at stage k_1 ($k_1 = 0$ means no rejection of $H_{0,TP}^{(s)}$), whereas the trial enters stage $k > k_1$, i.e. no rejection of $H_{0,TC}^{(n)}$ at stages k_1 to $k-1$. Note, that the relationship $A_k = E_{k,K+1}$ holds for $1 \le k \le K$ with the events A_k given in (3.17). For given k, $2 \le k \le K$, the events $E_{k_1,k}$ are pairwise disjoint, so that the probability $P_{\theta_{TP},\theta_{TC}}(E_k)$ can be expressed as the sum $\sum_{k_1=0}^{k-1} P_{\theta_{TP},\theta_{TC}}(E_{k_1,k})$. Denote by N_T the actual test group size, then we have according to (3.10)

$$ASn_T(\theta_{TP}, \theta_{TC}) = n_T^{(1)} + \sum_{k=2}^{K} \left(n_T^{(k)} - n_T^{(k-1)}\right) \underbrace{P_{\theta_{TP},\theta_{TC}}\left(N_T \ge n_T^{(k)}\right)}_{P_{\theta_{TP},\theta_{TC}}(E_k)}$$

$$= n_T^{(1)} + \sum_{k=2}^{K} \left(n_T^{(k)} - n_T^{(k-1)} \right) \sum_{k_1=0}^{k-1} P_{\theta_{TP},\theta_{TC}} \left(E_{k_1,k} \right). \tag{3.29}$$

The expected sample size of the control group is determined by replacing $n_T^{(k)}$ with $n_C^{(k)}$, $1 \leq k \leq K$, in (3.29), so that the overall expected sample size is obtained as the sum of $ASn_P(\theta_{TP})$, $ASn_T(\theta_{TP},\theta_{TC})$ and $ASn_C(\theta_{TP},\theta_{TC})$, that means we have

$$\begin{aligned} ASN(\theta_{TP},\theta_{TC}) = {} & ASn_P(\theta_{TP}) + \left(n_T^{(1)} + n_C^{(1)} \right) \\ & + \sum_{k=2}^{K} \left(\left(n_T^{(k)} + n_C^{(k)} \right) - \left(n_T^{(k-1)} + n_C^{(k-1)} \right) \right) \sum_{k_1=0}^{k-1} P_{\theta_{TP},\theta_{TC}} \left(E_{k_1,k} \right), \end{aligned}$$

where $ASn_P(\theta_{TP})$ is calculated through (3.27). Analogous to the overall power of the procedure, the probabilities $P_{\theta_{TP},\theta_{TC}}(E_{k_1,k})$ can be calculated by means of the distribution function of the multivariate normal distribution. According to (3.28), it can be seen that $P_{\theta_{TP},\theta_{TC}}(E_{0,k})$ for $2 \leq k \leq K$ is given as

$$P_{\theta_{TP},\theta_{TC}} \left(E_{0,k} \right) = \Phi^{\boldsymbol{\Sigma}_{TP}(k,\ldots,K)} \left(b_{TP}^{(1)} - \theta_{TP} \sqrt{\mathscr{I}_{TP}^{(1)}}, \ldots, b_{TP}^{(k-1)} - \theta_{TP} \sqrt{\mathscr{I}_{TP}^{(k-1)}} \right)$$

and for $2 \leq k \leq K$ and $1 \leq k_1 \leq k-1$ we have

$$P_{\theta_{TP},\theta_{TC}} \left(E_{k_1,k} \right) = \Phi^{\boldsymbol{\Sigma}_{k_1,k}} \left(\boldsymbol{a}_{k_1,k}; \boldsymbol{b}_{k_1,k} \right),$$

where the variance-covariance matrix $\boldsymbol{\Sigma}_{k_1,k}$ is given as

$$\boldsymbol{\Sigma}_{k_1,k} = \boldsymbol{\Sigma}(k_1+1,\ldots,K+k_1-1,K+k,\ldots,2K) \tag{3.30}$$

and the vectors $\boldsymbol{a}_{k_1,k}$ and $\boldsymbol{b}_{k_1,k}$ are determined according to

$$\begin{aligned} \boldsymbol{a}_{k_1,k} &= \left(a_{k_1,k,i} \right)'_{1 \leq i \leq k} \text{ with } a_{k_1,k,i} = \begin{cases} b_{TP}^{(i)} - \theta_{TP} \sqrt{\mathscr{I}_{TP}^{(i)}}, & i = k_1, \\ -\infty, & \text{else,} \end{cases} \\ \boldsymbol{b}_{k_1,k} &= \left(b_{k_1,k,i} \right)'_{1 \leq i \leq k} \text{ with } b_{k_1,k,i} = \begin{cases} b_{TP}^{(i)} - \theta_{TP} \sqrt{\mathscr{I}_{TP}^{(i)}}, & 1 \leq i \leq k_1 - 1, \\ \infty, & i = k_1, \\ b_{TC}^{(i-1)} - (\theta_{TC} + \Delta_{ni}) \sqrt{\mathscr{I}_{TC}^{(i-1)}}, & k_1 + 1 \leq i \leq k. \end{cases} \end{aligned} \tag{3.31}$$

In conclusion the overall average sample number is given as

$$\begin{aligned} ASN(\theta_{TP},\theta_{TC}) = {} & N^{(1)} + \sum_{k=2}^{K} \left(N^{(k)} - N^{(k-1)} \right) \Phi^{\boldsymbol{\Sigma}_{TP}(k,\ldots,K)} \left(b_{TP}^{(1)} - \theta_{TP} \sqrt{\mathscr{I}_{TP}^{(1)}}, \ldots, b_{TP}^{(k-1)} - \theta_{TP} \sqrt{\mathscr{I}_{TP}^{(k-1)}} \right) \\ & + \sum_{k=2}^{K} \left(\left(n_T^{(k)} + n_C^{(k)} \right) - \left(n_T^{(k-1)} + n_C^{(k-1)} \right) \right) \sum_{k_1=1}^{k-1} \Phi^{\boldsymbol{\Sigma}_{k_1,k}} \left(\boldsymbol{a}_{k_1,k}; \boldsymbol{b}_{k_1,k} \right), \end{aligned} \tag{3.32}$$

where $N^{(k)} = n_T^{(k)} + n_C^{(k)} + n_P^{(k)}$ is the cumulative overall sample size at stage $k = 1, \dots, K$ and the covariance matrix $\boldsymbol{\Sigma}_{k_1,k}$ and the vectors $\boldsymbol{a}_{k_1,k}$ and $\boldsymbol{b}_{k_1,k}$ are given in (3.30) and (3.31), respectively.

3.2.2 Designs with Equal Stage Sizes

Equal stage sizes are a common assumption when applying group sequential methodology in clinical trials. Also in the planning stage of trials implementing error spending designs this assumption is quite common. If the sample sizes of the control and placebo group are furthermore defined as fractions of the test group sample size, it turns out that calculation of the group sequential key characteristics such as maximum and expected sample size can be further simplified. Let us therefore assume that $n_D^{(k)} = \frac{k}{K} n_D^{(K)}$ for $D = T, C, P$ and $n_C^{(K)} = c_C n_T^{(K)}$, $n_P^{(K)} = c_P n_T^{(K)}$, so that we have $n_T^{(k)} = \frac{k}{K} n_T^{(K)}$, $n_C^{(k)} = c_C \frac{k}{K} n_T^{(K)}$ and $n_P^{(k)} = c_P \frac{k}{K} n_T^{(K)}$, $1 \le k \le K$.

Type I Error and Rejection Boundaries

According to (3.11), the covariance matrices $\boldsymbol{\Sigma}_{TP}$ and $\boldsymbol{\Sigma}_{TC}$ simplify to

$$\tilde{\boldsymbol{\Sigma}} = \left(\sqrt{\frac{\min(i,j)}{\max(i,j)}} \right)_{1 \le i,j \le K},$$

so that the rejection boundaries can be determined as a solution of

$$\Phi^{\tilde{\boldsymbol{\Sigma}}}\left(b_{TP}^{(1)}, \dots, b_{TP}^{(K)}\right) = \Phi^{\tilde{\boldsymbol{\Sigma}}}\left(b_{TC}^{(1)}, \dots, b_{TC}^{(K)}\right) = 1 - \alpha. \tag{3.33}$$

As described in Section 3.1.3 on classical group sequential tests, the rejection boundaries can be determined independent of the maximum sample sizes. For instance, the boundaries could be chosen from the Δ-class or also from the ρ- or γ-class of error spending designs, where the boundaries are derived according to (3.12) and (3.13) a priori by assuming equal stage sizes.

Power and Sample Sizes

First of all, it can be easily shown that the information levels $\mathscr{I}_{TP}^{(k)}$ and $\mathscr{I}_{TC}^{(k)}$ simplify to

$$\mathscr{I}_{TP}^{(k)} = \frac{1}{\sigma^2} \frac{c_P}{1 + c_P} \frac{k}{K} n_T^{(K)} \text{ and } \mathscr{I}_{TC}^{(k)} = \frac{1}{\sigma^2} \frac{c_C}{1 + c_C} \frac{k}{K} n_T^{(K)} \text{ for } 1 \le k \le K.$$

Therefore, according to (3.22) we have for $k_1, k_2 \in \{1, \dots, K\}$ that

$$\begin{aligned} Cov\left(Z_{TP}^{(k_1)}, Z_{TC}^{(k_2)}\right) &= \frac{\sigma^2}{n_T^{(\max(k_1,k_2))}} \sqrt{\mathscr{I}_{TP}^{(k_1)} \mathscr{I}_{TC}^{(k_2)}} \\ &= \frac{\sigma^2 K}{\max(k_1,k_2)\, n_T^{(K)}} \sqrt{\frac{1}{\sigma^4} \frac{c_C c_P}{(1+c_C)(1+c_P)} \frac{k_1 k_2}{K^2} n_T^{(K)2}} \end{aligned}$$

$$= \sqrt{\frac{c_C c_P}{(1+c_C)(1+c_P)}} \sqrt{\frac{\min(k_1,k_2)}{\max(k_1,k_2)}}.$$

Consequently, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the vector of test statistics $\boldsymbol{Z} = \left(Z_{TP}^{(1)}, ..., Z_{TP}^{(K)}, Z_{TC}^{(1)}, ..., Z_{TC}^{(K)}\right)'$ are given as

$$\boldsymbol{\mu} = (\mu_i)'_{1\le i\le 2K}, \text{ where } \mu_i = \begin{cases} \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{i}{K}n_T^{(K)}}, & 1 \le i \le K, \\ \frac{\theta_{TC}+\Delta_{ni}}{\sigma}\sqrt{\frac{c_C}{1+c_C}\frac{i-K}{K}n_T^{(K)}}, & K+1 \le i \le 2K, \end{cases}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \tilde{\tilde{\boldsymbol{\Sigma}}} & \rho\tilde{\tilde{\boldsymbol{\Sigma}}} \\ \rho\tilde{\tilde{\boldsymbol{\Sigma}}} & \tilde{\tilde{\boldsymbol{\Sigma}}} \end{pmatrix}, \text{ where } \rho = \sqrt{\frac{c_C c_P}{(1+c_C)(1+c_P)}}.$$

Interestingly, ρ is exactly the correlation of the test statistics $T_{TP}^{(s)}$, $T_{TC}^{(n)}$ of the single-stage design given in (2.12). According to this and (3.19), the power to reject the first null hypothesis $H_{0,TP}^{(s)}$ is given as

$$1-\beta_{TP} = 1-\Phi^{\tilde{\tilde{\boldsymbol{\Sigma}}}}\left(b_{TP}^{(1)} - \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{1}{K}n_T^{(K)}}, ..., b_{TP}^{(K)} - \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{K}{K}n_T^{(K)}}\right).$$

The overall power of the procedure is determined as

$$\begin{aligned} 1-\beta = 1-\Phi^{\tilde{\tilde{\boldsymbol{\Sigma}}}}&\left(b_{TP}^{(1)} - \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{1}{K}n_T^{(K)}}, ..., b_{TP}^{(K)} - \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{K}{K}n_T^{(K)}}\right) \\ &- \sum_{k=1}^{K} \Phi^{\boldsymbol{\Sigma}(k+1,...,K+k-1)}\left(\boldsymbol{a}_k; \boldsymbol{b}_k\right), \end{aligned} \tag{3.34}$$

which follows easily from (3.26), where the vectors $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ are defined as

$$\boldsymbol{a}_k = (a_{k,i})'_{1\le i\le K+1} \text{ with } a_{k,i} = \begin{cases} b_{TP}^{(i)} - \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{i}{K}n_T^{(K)}}, & i = k, \\ -\infty, & \text{else}, \end{cases}$$

$$\boldsymbol{b}_k = (b_{k,i})'_{1\le i\le K+1} \text{ with } b_{k,i} = \begin{cases} b_{TP}^{(i)} - \frac{\theta_{TP}}{\sigma}\sqrt{\frac{c_P}{1+c_P}\frac{i}{K}n_T^{(K)}}, & 1 \le i \le k-1, \\ \infty, & i = k, \\ b_{TC}^{(i-1)} - \frac{\theta_{TC}+\Delta_{ni}}{\sigma}\sqrt{\frac{c_C}{1+c_C}\frac{i-1}{K}n_T^{(K)}}, & k+1 \le i \le K+1. \end{cases}$$

As we can see, the overall power is described as a (increasing) function of the maximum test group size $n_T^{(K)}$, so that sample size calculation could proceed as follows: Based on prior information, e.g. from historical trials, and actual requirements prespecify the variables θ_{TP}, θ_{TC}, Δ_{ni}, σ, c_P, c_C, K and define rejection boundaries $b_{TP}^{(1)}, ..., b_{TP}^{(K)}$, $b_{TC}^{(1)}, ..., b_{TC}^{(K)}$ satisfying (3.33). The required maximum test group size to obtain overall power $1-\beta$ can then be obtained by solving Equation (3.34) for $n_T^{(K)}$ by means of a univariate numerical root finding method. The respective sample sizes at the preceding stages $k = 1, ..., K-1$ are obtained via $n_T^{(k)} = \frac{k}{K}n_T^{(K)}$

and the control and placebo group sample sizes through the relationships $n_C^{(k)} = c_C \frac{k}{K} n_T^{(K)}$ and $n_P^{(k)} = c_P \frac{k}{K} n_T^{(K)}$, $1 \le k \le K$, respectively.

Strictly speaking, as the sample sizes are integers, we need to search for the smallest $n_T^{(K)}$ so that the overall power given in (3.34) is greater than or equal to $1-\beta$ and $n_C^{(1)} = \frac{c_C}{K} n_T^{(K)}$ as well as $n_P^{(1)} = \frac{c_P}{K} n_T^{(K)}$ are also integers, i.e. $n_T^{(K)}$ has to be an integer multiple of $\frac{K}{c_C}$ and $\frac{K}{c_P}$. As this might be difficult to ensure for all situations, it seems appropriate to determine the required sample sizes by rounding the exact solutions of Equation (3.34). This will probably violate the assumption of equal stage sizes, however, it has been mentioned earlier that the classical group sequential tests are fairly robust with respect to these deviations. Moreover, error spending designs could be used just as well, offering the advantage of exact type I error control.

Expected Sample Sizes

For equal stage sizes, the expected placebo group size given in (3.27) can be simplified to

$$\begin{aligned} ASn_P(\theta_{TP}) = \frac{c_P n_T^{(K)}}{K} \Bigg[1 + \sum_{k=2}^{K} \Phi^{\tilde{\boldsymbol{\Sigma}}(k,\dots,K)} \Bigg(& b_{TP}^{(1)} - \frac{\theta_{TP}}{\sigma} \sqrt{\frac{c_P}{1+c_P} \frac{1}{K} n_T^{(K)}}, \\ & \dots, b_{TP}^{(k-1)} - \frac{\theta_{TP}}{\sigma} \sqrt{\frac{c_P}{1+c_P} \frac{k-1}{K} n_T^{(K)}} \Bigg) \Bigg]. \end{aligned} \tag{3.35}$$

Analogously, it can be easily shown with some calculus that the expected overall sample size given in (3.32) simplifies to

$$\begin{aligned} ASN(\theta_{TP}, \theta_{TC}) = \frac{n_T^{(K)}}{K} \Bigg[(1+c_C+c_P) \Bigg(1 + \sum_{k=2}^{K} \Phi^{\tilde{\boldsymbol{\Sigma}}(k,\dots,K)} \Bigg(& b_{TP}^{(1)} - \frac{\theta_{TP}}{\sigma} \sqrt{\frac{c_P}{1+c_P} \frac{1}{K} n_T^{(K)}}, \\ & \dots, b_{TP}^{(k-1)} - \frac{\theta_{TP}}{\sigma} \sqrt{\frac{c_P}{1+c_P} \frac{k-1}{K} n_T^{(K)}} \Bigg) \Bigg) \\ + (1+c_C) \sum_{k=2}^{K} \sum_{k_1=1}^{k-1} \Phi^{\boldsymbol{\Sigma}_{k_1,k}} \left(\boldsymbol{a}_{k_1,k}; \boldsymbol{b}_{k_1,k} \right) \Bigg], \end{aligned} \tag{3.36}$$

where the covariance matrix $\boldsymbol{\Sigma}_{k_1,k}$ and the vectors $\boldsymbol{a}_{k_1,k}$ and $\boldsymbol{b}_{k_1,k}$ for $2 \le k \le K$ and $1 \le k_1 \le k-1$ are given as

$$\begin{aligned} \boldsymbol{\Sigma}_{k_1,k} &= \boldsymbol{\Sigma}(k_1+1,\dots,K+k_1-1,K+k,\dots,2K), \\ \boldsymbol{a}_{k_1,k} &= \left(a_{k_1,k,i}\right)'_{1\le i\le k} \text{ with } a_{k_1,k,i} = \begin{cases} b_{TP}^{(i)} - \frac{\theta_{TP}}{\sigma} \sqrt{\frac{c_P}{1+c_P} \frac{i}{K} n_T^{(K)}}, & i = k_1, \\ -\infty, & \text{else,} \end{cases} \\ \boldsymbol{b}_{k_1,k} &= \left(b_{k_1,k,i}\right)'_{1\le i\le k} \text{ with } b_{k_1,k,i} = \begin{cases} b_{TP}^{(i)} - \frac{\theta_{TP}}{\sigma} \sqrt{\frac{c_P}{1+c_P} \frac{i}{K} n_T^{(K)}}, & 1 \le i \le k_1 - 1, \\ \infty, & i = k_1, \\ b_{TC}^{(i-1)} - \frac{\theta_{TC}+\Delta_{ni}}{\sigma} \sqrt{\frac{c_C}{1+c_C} \frac{i-1}{K} n_T^{(K)}}, & k_1 + 1 \le i \le k. \end{cases} \end{aligned}$$

With the maximum sample size of the test group $n_T^{(K)}$ determined as a solution of Equation (3.34), the expected sample sizes of the placebo group and overall can be calculated by means of Equations (3.35) and (3.36).

3.2.3 Hypothetical Example

In order to illustrate the proposed group sequential procedure, let us consider a hypothetical example of a three-arm non-inferiority trial, where a new therapy for the treatment of bronchial asthma is compared with a control treatment and placebo. Treatment efficacy is assessed by means of changes of the forced expiratory volume in one second (FEV$_1$), which is measured in litre (l). The one-sided significance level is set to $\alpha = 0.025$, which is a regulatory requirement for confirmatory claims. Furthermore, a usual non-inferiority margin for this primary endpoint of $\Delta_{ni} = 0.2l$ is adopted and the overall power to demonstrate non-inferiority of test to control and superiority of test over placebo should be $1-\beta = 0.90$. The assumptions at the planning stage on the treatment effects and the common standard deviation are $\mu_T = \mu_C = 2.4l$, $\mu_P = 2l$ and $\sigma = 1l$, respectively, that means we have $\theta_{TP} = 0.4l$ and $\theta_{TC} = 0l$.

For reference purposes, let us initially consider the optimal single-stage design proposed in Section 2.3.3. Due to practical reasons, we assume a treatment allocation ratio of $n_T : n_C : n_P = 4:4:1$, i.e. $c_C = 1$ and $c_P = 0.25$ which is only slightly different from the optimal allocation determined as $c_C = 0.98$, $c_P = 026$. According to (2.14), the sample sizes of the fixed sample size design are given as $n_{T,fix} = n_{C,fix} = 544$ and $n_{P,fix} = 136$, so that the overall sample size is $N_{fix} = 1224$.

With the same optimal allocation ratio as in the fixed design, we now want to derive a group sequential design with $K = 3$ stages and equal stage sizes, i.e. $n_D^{(k)} = \frac{k}{K} n_D^{(K)}$ for $D = T, C, P$ and $k = 1, \ldots, K$ (cf. Section 3.2.2). The group sequential boundaries are chosen from the Δ-family proposed by Wang and Tsiatis (1987), namely intermediate boundaries ($\Delta = 0.25$) for the proof of efficacy of the test treatment and O'Brien Fleming boundaries ($\Delta = 0$) for the test vs. control non-inferiority comparison, so that we have $(b_{TP}^{(1)}, b_{TP}^{(2)}, b_{TP}^{(3)}) = (2.741, 2.305, 2.083)$ and $(b_{TC}^{(1)}, b_{TC}^{(2)}, b_{TC}^{(3)}) = (3.471, 2.454, 2.004)$. Through this choice, the probability to reject $H_{0,TP}^{(s)}$ at earlier stages is increased, while the second null hypothesis $H_{0,TC}^{(n)}$ will more likely be rejected at later stages. Moreover, the maximum sample size increase compared with the fixed design should be moderate. As it has been mentioned earlier, there are two ways to determine the required sample size to obtain overall power $1-\beta = 0.90$.

Let us first consider the approximative approach by rounding the exact solutions of Equation (3.34). We obtain the maximum test group size $n_T^{(3)} = 555.6$, so that, through the relationships $n_T^{(k)} = n_C^{(k)} = \frac{k}{3} n_T^{(3)}$ and $n_P^{(k)} = 0.25 \frac{k}{3} n_T^{(3)}$, $k = 1, 2, 3$, and by rounding the numbers to integers, we have $(n_T^{(1)}, n_T^{(2)}, n_T^{(3)}) = (n_C^{(1)}, n_C^{(2)}, n_C^{(3)}) = (185, 370, 556)$ and $(n_P^{(1)}, n_P^{(2)}, n_P^{(3)}) = (46, 93, 139)$. With these sample sizes we have an overall power to reject both null hypotheses of $1-\beta = 0.9002$ according to (3.26). Note, that the maximum overall sample size N_{max} of 1251 is only 2.2% higher than the overall sample size of the fixed design. The average sample numbers of the placebo group and overall under the alternative of the power calculation are calculated by means of (3.27) and (3.32) as $ASn_P(0.4) = 77.86$ and $ASN(0.4, 0) = 975.08$, respectively. That means we have an expected reduction of the placebo group size of almost 50% compared with the fixed design, while the overall sample size is reduced (on average) by more than 20%. As it has been mentioned earlier, the classical group sequential tests are quite robust against deviations from the equal stage size assumption, so that this choice would be reasonable with regard to type I error control.

Nevertheless, because exact type I error control is desirable for confirmatory clinical trials, we are interested in determining sample sizes with equal stage sizes. Therefore, we need to search for the smallest $n_T^{(3)}$ which is a multiple of $\frac{K}{c_C} = 3$ and $\frac{K}{c_P} = 12$, so that the overall power given in (3.34) is greater than or equal to 90%. For $n_T^{(3)} = 564$ we have $1 - \beta = 0.9047$, so that the required sample sizes of the test, control and placebo group are given as $(n_T^{(1)}, n_T^{(2)}, n_T^{(3)}) = (n_C^{(1)}, n_C^{(2)}, n_C^{(3)}) = (188, 376, 564)$ and $(n_P^{(1)}, n_P^{(2)}, n_P^{(3)}) = (47, 94, 141)$. That means, the maximum overall sample size is $N_{max} = 1269$, which is a sample size increase of only 3.7% compared with the fixed design. The expected sample sizes are calculated by means of formulas (3.35) and (3.36) as $ASn_P(0.4) = 81.43$ and $ASN(0.4, 0) = 981.58$, which are around 60% of $n_{P,fix}$ and 80% of N_{fix}, respectively.

Via simulation we generated a hypothetical trial assuming the treatment effects and common standard deviation used for sample size determination. Table 3.3 gives an overview on the simulated data and the corresponding test results. At the first interim analysis, the test treatment is shown to be superior to placebo, as the corresponding test statistic $Z_{TP}^{(1)}$ exceeds 2.741. As the test statistic of the non-inferiority comparison between test and control at stage one is less than 3.471, the trial continues with randomising patients to the test and control treatment while the placebo arm is closed. At the second interim analysis, the test statistic $Z_{TC}^{(2)}$ crosses the critical value 2.454, so that the test treatment is demonstrated to be non-inferior to the control and the trial is stopped.

Finally, it should be noted that the test statistics $Z_{TP}^{(k)}$ and $Z_{TC}^{(k)}$ were considered as asymptotically $N(0,1)$-distributed, although the standard deviation σ is estimated by the unbiased pooled variance estimators $\hat{\sigma}^{(k)}$, $k = 1, 2$. The results stay valid if the adjusted boundaries

Table 3.3: Simulated data and test results for a hypothetical trial with $\mu_T = \mu_C = 2.4l$, $\mu_P = 2l$, $\sigma = 1l$ and $\Delta_{ni} = 0.2l$.

Stage	Sample sizes			Simulated data [in l]				Test results			
k	$n_T^{(k)}$	$n_C^{(k)}$	$n_P^{(k)}$	$\bar{X}_T^{(k)}$	$\bar{X}_C^{(k)}$	$\bar{X}_P^{(k)}$	$\hat{\sigma}^{(k)}$	$Z_{TP}^{(k)}$	$b_{TP}^{(k)}$	$Z_{TC}^{(k)}$	$b_{TC}^{(k)}$
1	188	188	47	2.404	2.373	1.968	0.953	2.805	2.741	2.350	3.471
2	376	376	–	2.433	2.417	–	0.990	–	2.305	2.992	2.454
3	–	–	–	–	–	–	–	–	2.083	–	2.004

$(\tilde{b}_{TP}^{(1)}, \tilde{b}_{TP}^{(2)}, \tilde{b}_{TP}^{(3)}) = (2.766, 2.313, 2.087)$ and $(\tilde{b}_{TC}^{(1)}, \tilde{b}_{TC}^{(2)}, \tilde{b}_{TC}^{(3)}) = (3.501, 2.460, 2.006)$ that account for the unknown variance setting would have been used instead.

3.2.4 Design Comparison

The proposed group sequential testing procedure shall now be compared with the optimal single-stage design derived in Section 2.3.3. Therefore we will restrict ourselves to designs with equal stage sizes (see Section 3.2.2) and no stopping for futility will be implemented as the emphasis of the procedure lies on prematurely closing the placebo arm due to proven efficacy of the test treatment. Consequently, the required overall sample size and the expected sample sizes of the placebo group and overall can be determined by means of Equations (3.34), (3.35) and (3.36), respectively. Moreover, we will only assess the expected sample sizes under the alternative of the power calculation, i.e. $\theta_{TP} = \delta_{TP}$ and $\theta_{TC} = 0$ as it is often assumed for non-inferiority comparisons. The expected sample sizes under the null hypothesis are reflected very well by the corresponding maximum sample sizes, as Table 3.2 showed. For better clarity, the respective average sample number of the test, control, placebo group and overall are denoted as ASn_T, ASn_C, ASn_P and ASN, respectively.

Wang Tsiatis Type Designs

Let us start by investigating Wang Tsiatis type designs with shape parameters Δ_{TP} and Δ_{TC} for the test treatments proof of efficacy and the non-inferiority comparison between test and control, respectively. The non-inferiority margin is chosen as half of the difference between control and placebo effect, i.e. $\Delta_{ni} = \frac{\theta_{CP}}{2}$, which is a common choice in clinical practice. The between-group sample size allocation ratios c_C and c_P of the group sequential designs are chosen as the optimal allocation ratios of the fixed sample size design derived in Chapter 2. That means for overall power 80% and 90% we have $c_C = 0.98$, $c_P = 0.30$ and $c_C = 0.98$, $c_P = 0.26$, respectively. Our investigations showed that the optimal group sequential allocations that minimise N_{max} are only slightly different from those of the optimal single-stage design. To be exact, for c_C we have maximum deviations of ± 0.01 and ± 0.04 for the placebo group allocation c_P. Moreover, the additional reductions in maximum overall sample size of only up to 0.56% in absolute terms are negligibly small, so that it seems appropriate, also for simplicity reasons, to adopt the optimal single-stage allocations in the group sequential setting.

The required maximum overall sample sizes and the average sample numbers of the placebo group and of all groups together are given in Table 3.4, represented as percentages of the optimal single-stage sample sizes N_{max} and $n_{P,fix}$ for overall power $1-\beta = 0.80, 0.90$ and $K = 2,3,4,5$ stages, respectively. Thereby, we considered all possible combinations of common choices for the shape parameters, namely $\Delta_{TP}, \Delta_{TC} \in \{0, 0.25, 0.5\}$. It should be noted that the tabulated values were calculated without rounding the sample sizes to the next higher integer values, since this has only a small effect and guarantees a fair comparison of two designs

Table 3.4: Maximum overall sample size (N_{max}) and expected sample sizes of the placebo group (ASn_P) and overall (ASN) for different group sequential designs from the Δ-class represented as percentages of the optimal fixed sample sizes N_{fix} and $n_{P,fix}$ with overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$, K stages and shape parameters Δ_{TP}, Δ_{TC}.

			$1-\beta = 0.80$			$1-\beta = 0.90$		
Δ_{TP}	Δ_{TC}	K	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
0.00	0.00	2	100.9	76.2	91.1	100.8	71.1	86.9
		3	101.9	71.6	85.9	101.7	67.3	80.9
		4	102.5	67.8	83.3	102.3	63.4	77.7
		5	103.0	65.8	81.9	102.7	61.2	76.0
0.25	0.00	2	101.2	69.0	89.2	101.1	64.4	84.7
		3	102.3	63.3	84.2	102.1	58.3	79.0
		4	103.0	60.3	81.8	102.7	55.4	75.9
		5	103.5	58.5	80.4	103.2	53.5	74.3
0.50	0.00	2	102.3	65.0	88.7	102.0	60.9	84.0
		3	104.0	57.3	84.1	103.5	52.2	78.6
		4	105.2	54.1	81.9	104.6	48.7	75.7
		5	106.0	52.3	80.7	105.3	46.7	74.1
0.00	0.25	2	103.8	77.6	90.0	103.4	72.2	84.4
		3	105.5	73.3	85.9	104.9	68.7	79.9
		4	106.5	69.5	83.3	105.9	64.8	76.7
		5	107.2	67.4	81.9	106.5	62.5	75.0
0.25	0.25	2	104.0	70.2	86.9	103.6	65.3	81.1
		3	105.7	64.4	82.9	105.1	59.2	76.4
		4	106.7	61.5	80.7	106.1	56.3	73.7
		5	107.4	59.6	79.3	106.8	54.4	72.0
0.50	0.25	2	104.8	66.0	85.8	104.3	61.8	79.8
		3	107.1	58.1	81.8	106.4	52.9	74.9
		4	108.5	54.8	79.8	107.7	49.2	72.3
		5	109.5	53.0	78.6	108.6	47.2	70.8
0.00	0.50	2	111.2	81.0	92.0	110.0	74.8	84.8
		3	116.8	78.4	91.1	115.1	72.9	83.3
		4	120.4	75.0	89.8	118.2	69.4	81.1
		5	122.9	73.1	89.2	120.5	67.2	79.9
0.25	0.50	2	110.8	72.9	87.8	109.7	67.7	80.6
		3	116.3	67.8	85.8	114.7	62.0	77.3
		4	119.9	65.3	84.9	117.9	59.3	75.6
		5	122.4	63.6	84.4	120.2	57.6	74.7
0.50	0.50	2	111.2	68.6	85.9	110.1	64.1	78.7
		3	116.9	60.7	82.9	115.2	55.0	74.1
		4	120.6	57.2	81.8	118.5	51.1	72.1
		5	123.2	55.3	81.3	120.9	49.0	71.1

with exact overall power $1-\beta$. Furthermore, the values of Table 3.4 apply to all different combinations of true treatment difference θ_{CP} and variance σ^2, because both the sample sizes of the single-stage and the group sequential design, i.e. maximum and expected sample sizes, are proportional to $\frac{\sigma^2}{\theta_{CP}^2}$.

Analogous to the common group sequential design, it can be seen that the maximum sample size increases whereas the average sample sizes ASn_P and ASN decrease with larger number of stages. Moreover, with higher overall power $1-\beta$ the maximum as well as the average sample sizes decrease compared with the required single-stage sample sizes. That means, the higher the targeted overall power the further the advantages of implementing a group sequential design in a three-arm trial are outweighing the disadvantages.

Table 3.4 also shows that the shape parameter of the non-inferiority comparison Δ_{TC} obviously has a much greater effect on the maximum sample size than Δ_{TP}. Due to the larger sample size of the control group compared with the placebo group and $\beta_{TC} \gg \beta_{TP}$ under the optimal allocation this is hardly surprising (cf. Figure 2.2).

Interestingly, the maximum sample size not always increases with higher Δ_{TP}, while for increasing Δ_{TC} also N_{max} increases. For instance, the designs with O'Brien Fleming type boundaries for the first and Pocock type boundaries for the second hypothesis test, i.e. $\Delta_{TP} = 0$ and $\Delta_{TC} = 0.5$, have higher required maximum sample sizes than the designs with $\Delta_{TP} = 0.25$ and $\Delta_{TC} = 0.5$. However, these minor differences might be due to numerical reasons and it can be also seen that anyway, designs with $\Delta_{TP} < \Delta_{TC}$ have less favourable operating characteristics. Especially, the most extreme of these designs with $\Delta_{TP} = 0$ and $\Delta_{TC} = 0.5$ turns out to have by far the highest values for the average sample sizes of the placebo group and overall, while the maximum sample size is also very high. The reason for this can be easily seen and mainly results from the large differences between the separate power of the first and second hypothesis test under the optimal allocation. Let us therefore assume that both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ are false. Choosing O'Brien Fleming boundaries for the proof of efficacy of the test treatment, i.e. $\Delta_{TP} = 0$, will probably result in a rejection of $H_{0,TP}^{(s)}$ at latter analyses, even though $1-\beta_{TP}$ is high under the optimal allocation. At that time, the probability to reject $H_{0,TC}^{(n)}$ will be relatively low because for the Pocock boundaries ($\Delta_{TC} = 0.5$) there will not be much α left to spend.

According to Equation (3.35) the shape parameter Δ_{TC} and the expected sample size ASn_P are only connected through the influence of Δ_{TC} on the maximum sample size $n_T^{(K)}$ to fulfil the overall power requirement, the expected placebo group size increases with increasing Δ_{TC}. Moreover, with increasing Δ_{TP} the average sample number of the placebo group decreases (for the parameter constellation considered), so that the largest average reduction of the placebo group size is obtained for the designs with $\Delta_{TP} = 0.5$ and $\Delta_{TC} = 0$. Namely, for $K = 5$ stages and an overall power of 80% (90%) we have an expected placebo group size of 52.3% (46.7%) of the corresponding optimal single-stage placebo group size. The highest ASN reduction is obtained by using Pocock boundaries for the first and intermediate boundaries for the second hypothesis test, i.e. $\Delta_{TP} = 0.5$ and $\Delta_{TC} = 0.25$. More precisely, with $K = 5$ stages the overall average sample size reductions compared with the optimal fixed designs are 21.4% and 29.2% for an overall power of 80% and 90%, respectively.

In the previous paragraphs we only considered the situation where the non-inferiority margin is chosen as half of the difference between control and placebo effect, i.e. $\Delta_{ni} = \frac{\theta_{CP}}{2}$. This is a common choice for most clinical trials, but sometimes Δ_{ni} is chosen smaller than this. Further investigations showed that for smaller non-inferiority margins the deviations between the optimal single-stage allocation minimising N_{fix} and the optimal group sequential allocation that minimises N_{max} get even smaller. Thus, adopting the optimal single-stage allocation in the group sequential setting is also appropriate for smaller non-inferiority margins.

Furthermore it turns out that the observed pattern in Table 3.4 and the relationships between the shape parameters Δ_{TP} and Δ_{TC} and the maximum and expected sample sizes N_{max}, ASn_P and ASN carry over to group sequential designs with non-inferiority margins smaller than $\frac{\theta_{CP}}{2}$. It can be shown, that the deviations from the numbers given in Table 3.4 get larger with smaller Δ_{ni}, although the maximum deviations for N_{max} and ASN are only around +2% and -3% in absolute terms for $\Delta_{ni} = 0.1 \cdot \theta_{CP}$. The average placebo group size can be reduced to an even greater extent for smaller non-inferiority margins, which is not surprising due to the increasing power $1 - \beta_{TP}$ under the optimal allocation for smaller Δ_{ni} (cf. Table A.2). A reduction of the non-inferiority margin by $0.1 \cdot \theta_{CP}$ results in a decrease of the expected placebo group size by about 2% to 4% absolute (relative to the respective placebo group size of the optimal fixed design). For instance, for $\Delta_{ni} = 0.3 \cdot \theta_{CP}$ the expected placebo group size ASn_P is 4% to 8% lower than the expected placebo group sizes given in Table 3.4.

Nevertheless, it should be kept in mind that the size of the placebo group in the optimal fixed design further decreases with smaller non-inferiority margins, so that a relative average reduction of the placebo group size of 50% might appear to be much greater than it really is in absolute numbers.

Error Spending Designs

Let us now take a look at group sequential designs where the boundaries are determined based on the error spending approach, which provides exact type I error control irrespective of the sample size allocation across the different stages. To enhance comparability, equal stage sizes are assumed and designs from the ρ-class by Kim and DeMets (1987) and the γ-class by Hwang et al. (1990) are considered that are more or less equivalent to Wang Tsiatis type designs with shape parameters $\Delta = 0, 0.25, 0.5$. That means, the spending function parameters of the ρ- and the γ-class for $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ are chosen as $\rho_{TP}, \rho_{TC} \in \{3,2,1\}$ and $\gamma_{TP}, \gamma_{TC} \in \{-4,-2,1\}$ in order to approximate the O'Brien Fleming, the intermediate and the Pocock design, respectively. Furthermore, analogous to the Wang Tsiatis designs we start with investigating the scenario $\Delta_{ni} = \frac{\theta_{CP}}{2}$.

Table 3.5 and Table 3.6 give the corresponding maximum and expected sample sizes represented as percentages of the optimal single-stage sample sizes for overall power $1-\beta = 0.80, 0.90$

Table 3.5: Maximum overall sample size (N_{max}) and expected sample sizes of the placebo group (ASn_P) and overall (ASN) for different group sequential designs from the ρ-class represented as percentages of the optimal fixed sample sizes N_{fix} and $n_{P,fix}$ with overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$, K stages and spending parameters ρ_{TP}, ρ_{TC}.

			$1-\beta=0.80$			$1-\beta=0.90$		
ρ_{TP}	ρ_{TC}	K	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
3	3	2	101.1	75.1	90.4	101.0	70.0	86.0
		3	102.2	68.9	85.1	101.9	64.0	79.7
		4	102.9	65.3	82.5	102.6	60.3	76.5
		5	103.4	63.2	81.0	103.1	58.1	74.7
2	3	2	101.3	70.5	89.1	101.2	65.7	84.4
		3	102.4	64.1	84.0	102.2	59.0	78.4
		4	103.2	60.8	81.5	102.9	55.7	75.3
		5	103.8	58.8	80.0	103.4	53.6	73.5
1	3	2	102.1	66.0	88.2	101.8	61.8	83.3
		3	103.4	58.7	83.3	103.0	53.6	77.4
		4	104.3	55.6	80.9	103.9	50.3	74.4
		5	105.0	53.9	79.5	104.5	48.5	72.7
3	2	2	102.8	75.9	89.5	102.5	70.6	84.2
		3	104.5	69.8	84.8	104.1	64.8	78.7
		4	105.6	66.3	82.4	105.1	61.1	75.7
		5	106.4	64.1	81.0	105.8	58.9	73.9
2	2	2	102.9	71.2	87.7	102.6	66.3	82.2
		3	104.7	64.9	83.2	104.3	59.7	76.9
		4	105.8	61.6	80.9	105.3	56.3	74.0
		5	106.6	59.6	79.5	106.0	54.2	72.3
1	2	2	103.6	66.6	86.4	103.2	62.3	80.7
		3	105.6	59.3	82.0	105.0	54.1	75.4
		4	106.8	56.2	79.9	106.1	50.7	72.7
		5	107.7	54.5	78.6	106.9	48.9	71.1
3	1	2	108.3	78.3	90.3	107.4	72.5	83.6
		3	111.7	72.6	87.0	110.5	67.0	79.5
		4	113.6	69.0	85.1	112.2	63.3	77.0
		5	114.8	66.8	83.9	113.3	61.0	75.4
2	1	2	108.1	73.3	87.8	107.3	68.0	81.1
		3	111.6	67.1	84.6	110.4	61.5	76.8
		4	113.5	63.8	82.8	112.2	58.1	74.5
		5	114.8	61.7	81.7	113.3	55.9	73.0
1	1	2	108.5	68.6	85.8	107.6	64.0	79.0
		3	112.0	61.1	82.4	110.8	55.6	74.3
		4	114.1	57.9	80.8	112.7	52.0	72.1
		5	115.4	56.0	79.8	113.8	50.1	70.8

Table 3.6: Maximum overall sample size (N_{max}) and expected sample sizes of the placebo group (ASn_P) and overall (ASN) for different group sequential designs from the γ-class represented as percentages of the optimal fixed sample sizes N_{fix} and $n_{P,fix}$ with overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$, K stages and spending parameters γ_{TP}, γ_{TC}.

			$1-\beta=0.80$			$1-\beta=0.90$		
γ_{TP}	γ_{TC}	K	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
−4	−4	2	101.1	75.4	90.6	101.0	70.3	86.2
		3	101.8	68.5	85.6	101.6	63.4	80.1
		4	102.3	65.1	83.1	102.1	59.8	77.1
		5	102.7	63.1	81.7	102.5	57.8	75.4
−2	−4	2	101.3	70.0	89.1	101.2	65.2	84.4
		3	102.2	63.1	84.2	101.9	57.9	78.5
		4	102.8	59.9	81.8	102.5	54.5	75.6
		5	103.2	58.0	80.4	102.9	52.6	73.9
1	−4	2	102.6	64.8	88.3	102.3	60.7	83.4
		3	104.1	57.2	83.8	103.6	52.1	77.7
		4	104.9	54.1	81.5	104.4	48.7	74.9
		5	105.5	52.4	80.2	104.9	46.9	73.3
−4	−2	2	103.2	76.3	89.6	102.8	71.0	84.2
		3	104.8	69.7	85.1	104.3	64.2	78.7
		4	105.8	66.3	82.9	105.2	60.8	76.0
		5	106.5	64.3	81.6	105.9	58.7	74.3
−2	−2	2	103.3	70.8	87.4	102.9	65.9	81.8
		3	105.0	64.0	83.1	104.5	58.6	76.5
		4	106.0	60.8	81.0	105.5	55.3	73.9
		5	106.8	58.9	79.8	106.1	53.3	72.4
1	−2	2	104.4	65.5	86.0	103.9	61.4	80.2
		3	106.6	57.9	81.9	105.9	52.7	74.9
		4	107.9	54.7	80.0	107.1	49.2	72.5
		5	108.8	53.1	78.9	107.9	47.4	71.0
−4	1	2	112.4	80.5	91.9	111.1	74.3	84.5
		3	117.2	74.1	89.0	115.4	67.7	80.5
		4	119.7	70.7	87.4	117.7	64.2	78.2
		5	121.3	68.6	86.4	119.1	62.0	76.8
−2	1	2	112.1	74.4	88.6	110.9	68.9	81.2
		3	116.9	67.7	85.8	115.2	61.5	77.1
		4	119.4	64.3	84.4	117.5	58.0	75.0
		5	121.1	62.3	83.5	119.0	56.0	73.8
1	1	2	112.5	68.7	86.0	111.3	64.2	78.7
		3	117.5	60.7	83.0	115.8	55.0	74.1
		4	120.2	57.2	81.7	118.2	51.2	72.1
		5	121.9	55.4	80.9	119.7	49.2	70.9

and $K = 2, 3, 4, 5$ stages of the designs from the ρ- and the γ-class, respectively. Note that the values of ρ in Table 3.5 are listed in descending order to enable an easier comparison with the performance characteristics of the Wang Tsiatis designs listed in Table 3.4, because $\rho = 3$ and $\rho = 1$ correspond to $\Delta = 0$ and $\Delta = 0.5$.

Table 3.6 shows that the relationships between the spending function parameters ρ_{TP} and ρ_{TC} and the performance parameters N_{max}, ASn_P and ASN are similar to the Wang Tsiatis designs, except that the relationships are the other way round. For instance, larger ρ_{TC} result in smaller maximum sample sizes N_{max} and vice versa. In general, the characteristics of the Kim DeMets designs are comparable to those of the Wang Tsiatis type designs. Only for the designs with 'approximative' Pocock boundaries for the non-inferiority test, i.e. $\rho_{TC} = 1$, the former provides better performances than the latter due to smaller maximum sample sizes. However, as it has been mentioned before these designs turn out to be no sensible choice at all, because of relatively large expected sample sizes.

Here, too, the highest expected placebo group size reductions are obtained for (approximative) Pocock spending functions for the first and O'Brien Fleming spending functions for the second hypothesis test, i.e. $\rho_{TP} = 1$ and $\rho_{TC} = 3$. Namely, for $K = 5$ stages and overall power 80% (90%) we have an expected placebo group size of 53.9% (48.5%) of the respective optimal single-stage placebo group size. Compared with the optimal fixed design, the actually required overall sample size can be reduced (on average) by more than 20% for $K = 5$ and overall power 80% ($\rho_{TP} = 1$, $\rho_{TC} = 2$) and by almost 30% if the targeted overall power is 90% ($\rho_{TP} = \rho_{TC} = 1$).

For designs based on the γ-family of error spending design by Hwang et al. (1990) the same pattern can be observed in Table 3.6 as for the corresponding Wang Tsiatis designs in Table 3.4. The relationship between the spending function parameters γ_{TP} and γ_{TC} and the parameters N_{max}, ASn_P and ASN is the same as for the shape parameters Δ_{TP} and Δ_{TC}. With $K = 5$ the placebo group size can be reduced (on average) by around 50% for both overall power 80% and 90%, whereas the expected overall sample size can be reduced to around 80% (70%) of the overall fixed sample size for overall power 80% (90%).

Further investigations showed that the observations for Wang Tsiatis type designs regarding differences between the optimal single-stage and optimal group sequential allocation minimising N_{max} carry over to the presented error spending designs. Consequently, adopting the optimal single-stage allocation is sensible here, too. The same applies for the influence of smaller non-inferiority margins, i.e. $\Delta_{ni} < \frac{\theta_{CP}}{2}$, on the operating characteristics of the group sequential designs. The ratios of maximum/expected overall and the respective fixed overall sample size are only marginally different from the values presented in Tables 3.5 and 3.6. The average reduction of the placebo group size increases with decreasing Δ_{ni}, however, it should be kept in mind, once again, that smaller non-inferiority margins also result in significantly smaller placebo group sizes under the optimal allocation.

The previous sections showed that the application of group sequential methodology in three-arm non-inferiority trials can lead to substantial sample size savings. Especially the expected placebo group size can be considerably reduced with only moderate increase in the maximum overall sample size. As the proposed error spending designs perform very similar to the Wang Tsiatis type designs, the former seem to be a more sensible choice in practice because of their additional flexibility with regard to deviations from the preplanned stage-wise sample size allocation. Moreover, it turned out that the performance of the designs is highly dependent on the choice of the shape or spending function parameters. In particular, it does not seem to be advisable to chose $\Delta_{TP} < \Delta_{TC}$, $\rho_{TP} > \rho_{TC}$ or $\gamma_{TP} < \gamma_{TC}$, as these designs provide by far the worst performance characteristics with high maximum as well as high expected sample sizes.

3.2.5 Design Optimisation

In the following, the proposed designs will be examined more precisely. The emphasis will be on the right choice of the shape parameters with respect to certain optimisation criteria, such as minimal ASN.

Wang Tsiatis Type Designs

As it has been mentioned earlier, designs from the Δ-class proposed by Wang and Tsiatis (1987) possess the (approximately) optimal property of minimising the average sample number. In order to get an impression on how far the average overall sample size can be reduced by the proposed group sequential testing procedure, we will investigate these approximately optimal Δ-class designs for different scenarios. Again, let us start by restricting to non-inferiority margins chosen as half of the difference between the control and placebo effect, i.e. $\Delta_{ni} = \frac{\theta_{CP}}{2}$. We also adopt the optimal single-stage allocations derived in the previous chapter.

By means of the downhill simplex algorithm for non-linear optimisation proposed by Nelder and Mead (1965) we searched for the shape parameters Δ_{TP} and Δ_{TC} that minimise the expected overall sample size ASN. The maximum and expected sample sizes of the respective designs for overall power $1-\beta = 0.80, 0.90$ and $K = 2,3,4,5$ stages are presented in Table 3.7. Again, the group sequential sample sizes are represented as percentages of the corresponding optimal single-stage sample sizes, so that the results apply for all values of θ_{CP} and σ^2.

First of all, it becomes obvious that Table 3.7 confirms the earlier findings of $\Delta_{TP} \geq \Delta_{TC}$ being a sensible restriction leading to “better” designs with reasonable operating characteristics. As it has been already suggested, the highest ASN reduction compared with the optimal single-stage design is obtained roughly by choosing Pocock boundaries for the test treatments proof of efficacy and boundaries between intermediate and Pocock for the non-inferiority comparison, i.e. $\Delta_{TP} \approx 0.5$ and $0.25 \leq \Delta_{TC} \leq 0.5$. Namely, for 80% (90%) overall power we have an average reduction of the overall sample size of around 20% (30%), while the average placebo group size is almost cut in half. For designs with only one or two interim analyses, the optimal shape parameter for the first hypothesis test Δ_{TP} is even above 0.5. Choosing such “aggressive” designs is quite unusual in practice, as they have the undesirable feature of increasing rejection

Table 3.7: Optimal shape parameters of Δ-class designs minimising ASN and corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

	$1-\beta = 0.80$					$1-\beta = 0.90$				
K	Δ_{TP}	Δ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	Δ_{TP}	Δ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
2	0.609	0.392	108.5	66.3	85.2	0.674	0.477	110.2	62.7	78.3
3	0.521	0.332	109.5	58.4	81.6	0.613	0.447	113.5	53.0	73.5
4	0.471	0.297	109.4	55.6	79.6	0.557	0.407	113.6	49.2	71.4
5	0.444	0.280	109.6	54.2	78.4	0.519	0.380	113.2	47.5	70.1

boundaries. Consequently, in practice one usually restricts to designs with $\Delta \leq 0.5$.

With increasing overall power $1-\beta$ we obtain larger optimal shape parameters Δ_{TP}, Δ_{TC} as well as a higher increase in overall sample size relative to the optimal fixed design. The average sample size reductions both in the placebo group and overall obviously increase with higher overall power and additional interim analyses, although the additional reductions for $K > 3$ are relatively small. Increasing the number of stages furthermore results in smaller optimal shape parameters and most interestingly, there is practically no increase in the required overall sample size for more than three stages.

Another very interesting type of designs arises as a result of minimising the sum $ASN + N_{max}$ in order to have a trade-off between low average and low maximum overall sample size (see Table 3.8). Note, that for designs with low maximum overall sample size the expected sample size under the global null hypothesis, i.e. $ASN(0, -\Delta_{ni})$, will also be reasonably low as we have seen in Table 3.2.

Table 3.8 shows the same relationship between the number of stages K or the overall power $1-\beta$ and the respective optimal shape parameters and operating characteristics. In this context, it should be underlined once again that the required overall sample size increase for ad-

Table 3.8: Optimal shape parameters of Δ-class designs minimising $ASN + N_{max}$ with corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

	$1-\beta = 0.80$					$1-\beta = 0.90$				
K	Δ_{TP}	Δ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	Δ_{TP}	Δ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
2	0.359	0.129	102.6	67.3	87.2	0.419	0.202	103.3	62.4	80.6
3	0.286	0.045	102.8	62.3	83.7	0.358	0.151	103.8	55.5	76.5
4	0.261	0.004	103.1	60.0	81.7	0.323	0.110	104.0	53.2	74.3
5	0.248	−0.016	103.4	58.5	80.5	0.308	0.093	104.3	51.8	72.9

ditional interim analyses is negligibly small. With virtually the same reductions in average placebo group and expected overall sample size these designs have considerably lower maximum sample sizes than the designs minimising ASN. For overall power 80% the optimal designs are roughly equal to intermediate designs for the first and O'Brien Fleming type designs for the second hypothesis test. For 90% overall power the optimal shape parameters are slightly increased, resulting in designs between intermediate and Pocock, and O'Brien Fleming and intermediate for testing $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$, respectively.

Since we adopted the optimal single-stage allocations for the group sequential designs, the question arises if allowing arbitrary allocation ratios c_C and c_P could improve the proposed designs with respect to further reductions of ASN or $ASN+N_{max}$. This leads to a 4-dimensional optimisation problem with the allocation ratios c_C, c_P and the shape parameters Δ_{TP}, Δ_{TC} as function parameters. Table 3.9 shows the optimal allocations and shape parameters together with the operating characteristics of the respective "full" optimal designs minimising either ASN or $ASN + N_{max}$. Analogous to the designs adopting the optimal single-stage allocations, the Nelder Mead algorithm was applied for optimisation. It should be recalled, that for overall power 80% and 90% the optimal single-stage allocation ratios are $c_C = 0.98$, $c_P = 0.30$ and $c_C = 0.98$, $c_P = 0.26$, respectively.

Taking a closer look at the designs minimising the expected overall sample size the following becomes apparent in comparison to the designs given in Table 3.7. First of all, the allocation

Table 3.9: Optimal allocation ratios and shape parameters of Δ-class designs minimising ASN or $ASN + N_{max}$ with corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

	$1-\beta$	K	c_C	c_P	Δ_{TP}	Δ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
min(ASN)	0.80	2	0.97	0.40	0.686	0.383	109.7	79.8	84.0
		3	0.97	0.46	0.648	0.349	113.5	72.8	79.4
		4	0.96	0.49	0.616	0.326	115.4	69.4	77.2
		5	0.96	0.51	0.594	0.314	116.6	67.6	75.8
	0.90	2	0.98	0.34	0.739	0.457	110.7	74.9	77.3
		3	0.98	0.40	0.728	0.448	116.8	67.1	71.4
		4	0.97	0.44	0.699	0.425	119.7	63.1	68.7
		5	0.97	0.47	0.674	0.411	121.6	60.8	67.1
min($ASN + N_{max}$)	0.80	2	0.98	0.34	0.404	0.132	102.9	72.3	86.2
		3	0.97	0.34	0.344	0.057	103.3	65.7	82.5
		4	0.97	0.35	0.320	0.012	103.6	63.0	80.5
		5	0.97	0.35	0.308	−0.008	103.9	61.3	79.3
	0.90	2	0.98	0.30	0.462	0.201	103.5	67.3	79.7
		3	0.98	0.30	0.418	0.159	104.4	58.6	75.2
		4	0.98	0.30	0.384	0.113	104.5	55.5	73.0
		5	0.97	0.30	0.367	0.101	104.9	53.9	71.5

ratio of the control group c_C is almost equal to the optimal single-stage allocation ratio of 0.98. The optimal values of the shape parameter for the non-inferiority comparison Δ_{TC} are also only slightly higher than the respective optimal values given in Table 3.7. In contrast, the allocation ratio of the placebo group c_P and the shape parameter Δ_{TP} are significantly increased. Especially for higher number of stages, the differences to the optimal values of the 2-dimensional optimal designs become more and more apparent. The additional expected reduction in overall sample size by allowing arbitrary allocation ratios is relatively small as it amounts to a maximum of only 3% in absolute terms. At the same time the maximum overall and the placebo group sample size are considerably increased, so that it seems appropriate to opt for the designs using the optimal single-stage allocations, not least because of their simplicity.

The same findings can be observed for the designs minimising the sum $ASN + N_{max}$. To be exact, there virtually are no differences in the control group allocation ratios c_C and the shape parameters Δ_{TC}. Moreover, the differences for c_P and Δ_{TP} are even smaller for these designs. The additional ASN reduction amounts to only up to 1%, while N_{max} and ASn_P are slightly increased. In conclusion, opting for the "simple" optimal designs given in Table 3.8 is reasonable here, too. In general, this confirms the earlier finding that it is sensible to adopt the optimal single-stage allocation also in the group sequential setting.

Another interesting type of group sequential designs arises, if we search for the shape parameters Δ_{TP} and Δ_{TC} that minimise the expected placebo group size ASn_P. As the average placebo group size ASn_P is only connected with Δ_{TC} through the influence of Δ_{TC} on the maximum sample sizes (cf. Equation (3.35)), minimising ASn_P obviously leads to $\Delta_{TC} \to -\infty$. That means the boundary values at analyses 1 to $K-1$ of the second hypothesis test $b_{TC}^{(1)}, \ldots, b_{TC}^{(K-1)}$ converge against ∞ whereas the final boundary $b_{TC}^{(K)}$ becomes $z_{1-\alpha}$, the critical value of the common fixed sample size test. Thus, only the first null hypothesis $H_{0,TP}^{(s)}$ is tested in a group sequential design, while the second null hypothesis $H_{0,TC}^{(n)}$ is tested with a common single-stage test without having the chance to stop the study early with rejection of both null hypotheses. These designs will henceforth be referred to as *partial group sequential designs*, whereas the designs where both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ are tested in a group sequential manner will be denoted as *full group sequential designs*.

Table 3.10 shows the optimal shape parameters Δ_{TP} of the partial group sequential designs minimising either ASn_P or ASN for overall power $1-\beta = 0.80, 0.90$ and $K = 2, 3, 4, 5$ stages. Note that minimising ASN for partial group sequential designs is somewhat similar to minimising $ASN + N_{max}$ in full group sequential designs, as for the partial design we have $ASn_C = n_C^{(K)}$ and $ASn_T = n_T^{(K)}$ so that ASN is a sum of expected and maximum sample sizes. As we can see, the designs that minimise the average placebo group size have rejection boundaries for the first null hypothesis $H_{0,TP}^{(s)}$ that are even more aggressive than a Pocock type design, i.e. we have $\Delta_{TP} > 0.5$. As mentioned previously, this choice is rather unusual and in practice one would choose $\Delta_{TP} = 0.5$. However, the reduction of the placebo group size to more than 50% of the optimal fixed placebo group size for $1-\beta = 0.90$ is remarkable. The corresponding maximum

Table 3.10: Optimal shape parameters of partial group sequential Δ-class designs that minimise ASn_P or ASN with corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

		min(ASn_P)				min(ASN)			
$1-\beta$	K	Δ_{TP}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	Δ_{TP}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
0.80	2	0.762	104.2	63.5	98.9	0.280	100.6	68.1	96.3
	3	0.680	105.8	55.6	99.2	0.282	100.9	61.8	95.8
	4	0.624	106.1	52.8	99.1	0.262	101.0	59.4	95.5
	5	0.586	106.0	51.3	98.8	0.246	101.0	58.0	95.4
0.90	2	0.744	103.5	59.7	98.3	0.268	100.5	63.8	96.2
	3	0.697	105.5	50.5	99.1	0.291	100.8	56.6	95.6
	4	0.651	106.2	47.1	99.3	0.278	100.9	54.0	95.4
	5	0.615	106.3	45.5	99.2	0.261	100.9	52.6	95.3

overall sample size increase is relatively small, whereas the expected overall sample size is almost equal to the overall sample size of the optimal single-stage design.

By minimising the expected overall sample size ASN we obtain more or less intermediate group sequential boundaries for the superiority comparison between test and placebo, i.e. $\Delta_{TP} = 0.25$. The reduction of the placebo group size is still considerably large with the additional benefit that the overall sample size needs to be only marginally increased compared with the optimal single-stage design. However, a reduction of the overall sample size by more than 5% is not possible with the partial group sequential design.

Finally, it should be noticed that all the results presented here were based on the assumption of a non-inferiority margin chosen as half of the difference between control and placebo effect, i.e. $\Delta_{ni} = \frac{\theta_{CP}}{2}$. Further investigations showed that the good performances of the proposed group sequential designs remain unaffected by decreasing the non-inferiority margin. Indeed, the performance of the proposed designs actually gets better for smaller non-inferiority margins. The higher power of the test versus placebo superiority comparison $1-\beta_{TP}$ for smaller Δ_{ni} (cf. Table A.2) results in even smaller expected placebo group sizes. On the other hand, it turns out that the ratios $\frac{N_{max}}{N_{fix}}$ and $\frac{ASN}{N_{fix}}$ are nearly unaffected by decreasing Δ_{ni}. The differences between the "full" optimal and the "simple" optimal design also remain negligible, so that it is sensible to adopt the optimal single-stage allocation in such cases, too.

Error Spending Designs

The same optimisations as for the Wang Tsiatis type designs will now be conducted within the ρ- and γ-family of error spending designs. Thereby, we also start by restricting to the common choice $\Delta_{ni} = \frac{\theta_{CP}}{2}$ for the non-inferiority margin. Furthermore, the optimal single-stage allocation will be adopted here, too.

Table 3.11: Optimal spending parameters of ρ- and γ-family designs minimising ASN and corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

		ρ-family					γ-family				
$1-\beta$	K	ρ_{TP}	ρ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	γ_{TP}	γ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
0.80	2	0.507	1.112	108.5	66.3	85.2	1.728	−0.297	108.5	66.3	85.2
	3	0.669	1.401	109.2	58.4	81.6	1.255	−0.916	109.5	58.3	81.6
	4	0.802	1.627	109.0	55.6	79.7	0.894	−1.304	109.6	55.3	79.8
	5	0.898	1.767	109.0	54.2	78.5	0.670	−1.542	109.6	53.7	78.8
0.90	2	0.383	0.837	110.2	62.7	78.3	2.379	0.480	110.2	62.7	78.3
	3	0.436	0.916	113.1	53.0	73.6	2.347	0.308	113.6	53.0	73.6
	4	0.541	1.087	112.8	49.4	71.5	2.076	0.019	114.3	49.2	71.4
	5	0.629	1.218	112.4	47.7	70.3	1.850	−0.190	114.4	47.3	70.2

Table 3.11 shows the optimal spending function parameters within the ρ- and γ-family of error spending designs that minimise ASN with corresponding operating characteristics. The optimal parameters were found with the same optimisation method that was applied earlier for the Δ-class designs. In turns out, that the optimal error spending designs provide literally the same expected reductions of placebo group size and overall sample size as the Wang Tsiatis type designs. At the same time the respective maximum overall sample sizes are also very similar, except that the ρ-family designs require slightly smaller overall sample sizes than the other two designs for 90% overall power and $K = 4,5$ stages. In practice, however, this difference is definitely negligible. The relationships between the spending function parameters and the operating characteristics can also be carried over from the Wang Tsiatis designs given in Table 3.7. As might be expected, it turns out that the restrictions $\rho_{TP} \leq \rho_{TC}$ and $\gamma_{TP} \geq \gamma_{TC}$ are reasonable resulting in small expected sample sizes.

This finding is confirmed by the respective error spending designs that minimise the sum $ASN + N_{max}$ given in Table 3.12. With overall maximum sample size increases of around 3% compared with the optimal single-stage design, the placebo group size and the overall sample size can be reduced (on average) to less than 60% of $n_{P,fix}$ and 80% of N_{fix}, respectively. Comparisons with the Wang Tsiatis type design that minimise $ASN + N_{max}$ given in Table 3.8 show that for all three families the same observations can be made. The differences in maximum and expected sample sizes are even smaller than for the designs that minimise the average overall sample size ASN. In general, these designs seem to be a more sensible choice than the designs minimising ASN, as they provide almost the same expected reductions while at the same time the required maximum sample size is substantially smaller. This becomes particularly obvious for overall power $1-\beta = 0.90$, where the differences in N_{max} are almost 10% in absolute terms. It should be noted again, that the required sample size increase associated with additional in-

Table 3.12: Optimal spending parameters of ρ- and γ-family designs minimising $ASN + N_{max}$ and corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

		ρ-family					γ-family				
$1-\beta$	K	ρ_{TP}	ρ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	γ_{TP}	γ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
0.80	2	1.226	2.367	102.6	67.3	87.2	−0.583	−2.850	102.6	67.3	87.2
	3	1.591	3.088	102.6	61.9	83.7	−1.328	−3.837	102.5	61.4	83.8
	4	1.796	3.439	102.7	59.6	81.8	−1.730	−4.336	102.6	59.1	82.0
	5	1.934	3.665	102.8	58.2	80.6	−1.987	−4.655	102.6	57.8	80.9
0.90	2	1.014	1.947	103.3	62.4	80.6	−0.039	−2.098	103.3	62.4	80.6
	3	1.269	2.499	103.5	55.3	76.6	−0.671	−3.013	103.3	55.1	76.6
	4	1.457	2.817	103.5	52.8	74.4	−1.057	−3.491	103.3	52.4	74.5
	5	1.568	3.003	103.7	51.4	73.0	−1.287	−3.787	103.4	51.0	73.3

terim analyses within the class of designs minimising $ASN + N_{max}$ is negligibly small. Further investigations showed that allowing arbitrary allocation ratios c_C and c_P instead of adopting the optimal single-stage allocations only leads to small performance gains when minimising ASN or $ASN + N_{max}$ (see Tables A.3 and A.4). Consequently, it is sensible to adopt the optimal single-stage allocation for error spending designs, too.

Optimising the average sample number of the placebo group within the ρ- and γ-family designs leads to the earlier mentioned *partial group sequential designs*, where only the first hypothesis is tested in a group sequential manner while the non-inferiority comparison is assessed with a common fixed sample size test. To be exact, by minimising ASn_P we have $\rho_{TC} \to \infty$ and $\gamma_{TC} \to -\infty$, respectively, which follows easily from Equation (3.35). Tables 3.13

Table 3.13: Optimal spending function parameters of partial group sequential ρ-family designs that minimise ASn_P or ASN with operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

		$\min(ASn_P)$				$\min(ASN)$			
$1-\beta$	K	ρ_{TP}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	ρ_{TP}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
0.80	2	0.253	104.2	63.5	98.9	1.560	100.6	68.1	96.3
	3	0.320	105.2	55.8	98.7	1.705	100.8	61.9	95.7
	4	0.391	105.2	53.1	98.3	1.903	100.9	59.6	95.4
	5	0.450	105.1	51.7	98.0	2.062	100.9	58.3	95.3
0.90	2	0.276	103.5	59.7	98.4	1.616	100.5	63.8	96.2
	3	0.286	104.9	50.6	98.6	1.662	100.7	56.7	95.6
	4	0.334	105.2	47.4	98.4	1.828	100.8	54.2	95.4
	5	0.382	105.1	45.9	98.2	1.967	100.8	52.8	95.2

Table 3.14: Optimal spending function parameters of partial group sequential γ-family designs that minimise ASn_P or ASN with operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

		min(ASn_P)				min(ASN)			
$1-\beta$	K	γ_{TP}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$	γ_{TP}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
0.80	2	3.307	104.2	63.5	98.9	−1.334	100.6	68.1	96.3
	3	3.282	106.2	55.6	99.5	−1.807	100.7	62.1	95.7
	4	3.067	106.8	52.6	99.7	−2.161	100.8	59.7	95.4
	5	2.865	107.0	51.1	99.6	−2.398	100.8	58.4	95.2
0.90	2	3.113	103.5	59.7	98.4	−1.450	100.5	63.8	96.2
	3	3.512	105.9	50.5	99.4	−1.767	100.6	56.9	95.5
	4	3.486	106.9	47.0	99.9	−2.088	100.7	54.3	95.3
	5	3.364	107.3	45.3	100.1	−2.314	100.7	52.9	95.1

and 3.14 show the operating characteristics of the partial group sequential designs that minimise ASn_P or ASN within the ρ- and γ-family of error spending designs, respectively. Again, the differences between the two error spending designs and the respective Wang Tsiatis designs given in Table 3.10 turn out to be negligible. For the designs that minimise ASn_P, the placebo group size can be almost cut in half, while the maximum sample sizes are not too high. With substantially smaller maximum sample sizes, the expected placebo group reductions of the designs minimising ASN are not much smaller. Particularly in situations where the potential risk taken should be as small as possible, these designs are of practical relevance.

The observations of the Wang Tsiatis designs for smaller non-inferiority margins, i.e. $\Delta_{ni} < \frac{\theta_{CP}}{2}$, can be carried over to the proposed error spending designs from the ρ- and γ-family. More precisely that means, the good performances of the error spending designs are almost unaffected by decreasing the non-inferiority margin. Due to the higher power $1-\beta_{TP}$ under the optimal single-stage allocation for smaller Δ_{ni}, the placebo group sizes are reduced to an even greater extent. Moreover, adopting the allocation ratios of the optimal fixed design remains sensible in case that a smaller non-inferiority margin is chosen. In conclusion, as already indicated earlier, the performances of the error spending designs turned out to be comparable to those of the Wang Tsiatis type designs. Due to their ability of adequately handling deviations from the preplanned sampling scheme, choosing the proposed error spending designs seems to be more appropriate in practice.

3.3 Summary

Besides minimising the overall sample size, the optimal allocation for the single-stage design determined at the end of Chapter 2 also leads to considerably low placebo group sizes. Furthermore it turned out that the power of the proof of efficacy for the test treatment is very high under the optimal allocation. This led us to the idea of exploiting the high power by means of implementing a group sequential design, giving us also the possibility to close the placebo arm once the efficacy of the test treatment has been demonstrated.

We proposed a classical group sequential design for normally distributed outcomes where a special emphasis was placed on the right choice of the rejection boundaries. The distributional properties of the group sequential test statistics allowed us exact calculations of the overall power and the expected sample sizes by means of the multivariate normal distribution function. This enabled us to conduct a detailed comparison with the optimal single-stage design and to derive approximately optimal boundaries with respect to certain optimisation criteria such as minimising the expected placebo group size.

First of all, it turned out that it is sensible to adopt the optimal single-stage allocation also in the group sequential setting. Our investigations showed that the performance gain by determining the optimal group sequential allocation is negligible. Furthermore, it became apparent that the operational characteristics of the group sequential designs are highly dependent on the choice of the rejection boundaries. In particular, more aggressive boundaries for the proof of efficacy than for the non-inferiority comparison lead to considerable performance gains. That means, for the respective shape and spending function parameters of the investigated boundary classes the restrictions $\Delta_{TP} \geq \Delta_{TC}$, $\rho_{TP} \leq \rho_{TC}$ and $\gamma_{TP} \geq \gamma_{TC}$ are sensible. The proposed designs can be separated into two approaches, the full and the partial group sequential approach. The former aims at a reduction of both the placebo and the overall sample size, whereas the latter mainly reduces the expected placebo group size while keeping the maximum sample size close to the optimal single-stage sample size. The full group sequential design might be of interest if an early study termination is conceivable and the partial design is a reasonable option when there is a need to collect more safety data on the experimental treatment. In general, the application of group sequential methodology in three-arm non-inferiority trials was demonstrated to have several benefits. Besides reducing the overall sample size and the associated cost and time savings, the potential early termination of the placebo arm is a key advantage that could help to overcome ethical concerns. Furthermore, as mentioned by Li and Gao (2010) the proposed designs could help to deal with uncertainties regarding the placebo effect. For instance, if the placebo effect is overestimated in the planning stage when determining the required sample sizes, the placebo group will most likely be closed early.

It should be noted, however, that the application of a group sequential design also involves operational challenges. In the first instance, it is vital to set up an Independent Data Monitoring Committee (IDMC), so that the validity and integrity of the trial are preserved. In order to maintain the blinding of the study, the information that is revealed by the IDMC should be furthermore kept to a minimum. The information of dropping the placebo arm (or not) should also be limited to a specific group of people. Otherwise the patient population might change if the patients or investigators are aware of the placebo group termination. Thus, it is generally advisable to assess the homogeneity of treatment effects across the different stages in the final analysis.

Another problem observed in group sequential trials is the issue of “overrunning”. Often primary endpoints are not observed immediately, so that some patients might still be under treatment when the interim analysis is conducted. In accordance with the ITT principle of evaluating all randomised patients, a primary analysis should include these patients. In addition, according to the EMA guideline on adaptive designs in confirmatory clinical trials, the “results including and excluding the overrunning patients should be presented and differences between these two analyses should be discussed.” (CHMP, 2007b). The proposed error spending designs from the ρ- and γ-family are a useful option to implement this strategy. As their performances turned out to be comparable to those of the Wang Tsiatis type designs, the error spending designs are a more sensible choice in practice due to their ability to adequately handle deviations from the preplanned sample sizes.

As it has been mentioned above, a major point of criticism on the proposed group sequential testing procedure is that the patient population might change after dropping the placebo group at an interim analysis. This issue might be addressed by keeping the information of terminating the placebo arm under wraps. However, once the recruitment to the placebo group has been stopped, no information on the placebo effect will be collected from then on. Through this, potential differences between the patient population before and after dropping the placebo group might be missed. A possible solution to this is the generalisation to adaptive group sequential designs that provide the possibility of adapting the preplanned sample sizes during trial conduct, e.g. based on the observed treatment effects at an interim analysis. This enables us, after rejection of $H_{0,TP}^{(s)}$, to decrease the placebo group size at the subsequent stages to a certain threshold instead of entirely closing the placebo arm. In addition to that, the proof of efficacy for the control treatment can be assessed at the following stages without type I error inflation. At last, an adaptive design offers the chance to re-calculate the sample sizes and update the sample size allocation so as to deal with uncertainties regarding the treatment effects in the planning stage. In the following chapter the proposed group sequential designs will be extended to adaptive group sequential designs.

CHAPTER 4

Extension to Adaptive Designs

In a classical group sequential design the data observed at an interim analysis, as for example the sample means, cannot be used for adapting the design of the subsequent stages. Otherwise the distributional properties of the group sequential test statistics do not hold any longer, so that using the common group sequential boundaries will typically result in a type I error inflation. For instance, re-calculating the sample sizes of the subsequent stages at an interim analysis based on the observed data can considerably inflate the type I error rate as it has been shown by Proschan and Hunsberger (1995) and Cui et al. (1999). Besides mid-trial sample size re-assessment other design features that might be changed at an interim analysis are the treatment groups, the patient population or the multiple testing strategy, to name but a few. The reasons for such adaptations are manifold, although one of the main reasons certainly is the ability to make better use of the available resources.

In this chapter we will extend the group sequential testing procedure for three-arm non-inferiority trials described in the previous chapter to adaptive designs that allow data-dependent sample size re-calculations at an interim analysis. Through this, commonly present uncertainties regarding the treatment effects in three-arm non-inferiority trials can be addressed by re-calculating the sample sizes based on the observed treatment differences. Because the possibility of data-dependent design changes obviously comes along with increased operational challenges and due to simplicity reasons, we will restrict to designs with only one interim analysis, i.e. $K = 2$ stages.

The chapter is structured as follows. Section 4.1 gives a short introduction to adaptive designs and the related concepts of conditional power and Bayesian predictive power, which are two useful tools for interim decision-making. In the following Section 4.2 the group sequential testing procedure proposed in the previous chapter will be extended to adaptive designs offering the possibility of data-dependent sample size changes at the interim analysis. After deriving the corresponding conditional and predictive power formulas, the proposed procedure will then be applied in a hypothetical example. Finally, we will take a closer look at a special type of adaptive design without early rejection, that uses the observed interim data only to optimise the between-group allocation.

It should be noted that Hartung and Knapp (2009) already proposed adaptive group sequential designs for three-arm non-inferiority trials. However, they did not consider conditional and Bayesian predictive power for assessing the mid-trial data. Moreover, they did not investigate the overall power to reject both null hypotheses as it has been done in the previous chapter. Our proposed adaptive designs are natural extensions of the group sequential designs from Chapter 3, so that all previous findings also apply to the adaptive designs, such as formulas for sample size determination.

4.1 Adaptive Designs

Historically, the idea of adaptively choosing the sample sizes based on mid-trial data goes back to Stein (1945). Although the proposed procedure only uses the sample variance of the first stage to determine the second-stage sample size and does not incorporate test decisions at the interim analysis, it can be seen as the first adaptive procedure. With the introduction of the classical group sequential designs in the late 1970s the idea of data-dependent mid-trial design changes arose again. One of the first comments in this regard was by Fleming et al. (1984), who mentioned that design changes, “where the decision to do so is based upon consideration of interim results, will alter the experimental type 1 error rate”.

The subsequent publications by Bauer (1989), Bauer and Köhne (1994) and Proschan and Hunsberger (1995) can be seen as the basis of procedures nowadays better known as *adaptive* or *flexible designs*, allowing not only sample size re-calculations but also general study design adaptations based on the interim results without inflating the overall type I error rate. The underlying concept of the proposed procedure is surprisingly simple and based on the combination of the p-values that are formed from the data of the respective stage alone, a concept that is also used for combining the results from different studies in meta-analyses. Bauer and Köhne (1994) proposed to use the product of the two stage-wise p-values, leading to Fisher’s combination test. They also mentioned other possible combination methods such as the weighted inverse normal method (Mosteller and Bush, 1954), which was investigated for adaptive designs by Lehmacher and Wassmer (1999). Interestingly, it turns out that use of the weighted inverse normal method leads to adaptive designs that are closely related to group sequential designs. Through this, specific calculations for the group sequential setting such as sample size determination carry over to these type of adaptive designs, which is the reasons why we will use this approach for extending the earlier proposed group sequential testing procedure to adaptive designs.

In the meantime Proschan and Hunsberger (1995) proposed a slightly different approach for adaptive designs by investigating the maximum type I error inflation that can occur when the second-stage sample size is based on the interim data. The critical value of the final stage is then adjusted by means of the so-called *conditional error function*, which gives the probability of committing a type I error at the final stage given the interim data. Vandemeulebroecke (2006) showed that this approach is directly linked to the p-value combination approach and that both concepts can be expressed in terms of the other notation.

Through several publications in the following years adaptive designs became more and more established and finally found their way into clinical trial application. The increasing importance of adaptive designs and the urgent need for specific guidance is also reflected by the recently published reflection paper and draft guideline on this topic from the EMA and the FDA, respectively (CHMP, 2007b; FDA, 2010a). Let us now take a closer look at the weighted inverse normal method for combining the results from two stages as it has been investigated by Lehmacher and Wassmer (1999).

4.1.1 Weighted Inverse Normal Method

Consider the same problem as in Section 3.1.2, i.e. demonstrating the superiority of a treatment A over another treatment B. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$, $i = 1, 2, \ldots$, and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, be the mutually independent responses of patients allocated to treatment A and B, respectively, with known common variance σ^2. Suppose that larger responses are desirable and denote the treatment difference by $\theta = \mu_A - \mu_B$. Then, the corresponding set of hypotheses of the superiority comparison is $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$.

As in the group sequential setting suppose that the data are analysed successively at $K = 2$ different time points, i.e. at an interim and a final analysis. However, instead of using the cumulative test statistics as in group sequential designs, the data from each stage shall be treated separately. Let therefore $\tilde{n}_A^{(1)}, \tilde{n}_A^{(2)}$ and $\tilde{n}_B^{(1)}, \tilde{n}_B^{(2)}$ denote the stage-wise sample sizes of treatment group A and B, respectively. Consequently, we have the relationships $\tilde{n}_D^{(1)} = n_D^{(1)}$ and $\tilde{n}_D^{(2)} = n_D^{(2)} - n_D^{(1)}$ for $D = A, B$, with the cumulative sample sizes defined in Section 3.1.2. Let further $\Delta\bar{X}_A^{(k)}$ and $\Delta\bar{X}_B^{(k)}$ be the sample means formed from the data of stage k alone, that means we have $\Delta\bar{X}_D^{(1)} = \bar{X}_D^{(1)}$ and $\Delta\bar{X}_D^{(2)} = \sum_{i=n_D^{(1)}+1}^{n_D^{(2)}} X_{D,i} / \tilde{n}_D^{(2)}$ for $D = A, B$. The corresponding stage-wise test statistics are given as

$$\tilde{Z}_k = \frac{\Delta\bar{X}_A^{(k)} - \Delta\bar{X}_B^{(k)}}{\sigma} \sqrt{\frac{\tilde{n}_A^{(k)} \tilde{n}_B^{(k)}}{\tilde{n}_A^{(k)} + \tilde{n}_B^{(k)}}} \quad \text{for } k = 1, 2,$$

which obviously coincides with the cumulative test statistic Z_k given in (3.2) for $k = 1$, that means we have $\tilde{Z}_1 = Z_1$. Interestingly, it can be shown that under H_0 the test statistics $\tilde{Z}_1$ and $\tilde{Z}_2$ are independent $N(0, 1)$ variables, irrespective of whether the second-stage sample sizes $\tilde{n}_A^{(2)}$ and $\tilde{n}_B^{(2)}$ depend on the previously observed responses or not. For a formal proof of this property see Brannath et al. (2012). The test statistics based on the *weighted inverse normal* *combination* now take the form

$$Z_k^* = \frac{\sum_{i=1}^{k} w_i \tilde{Z}_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \quad \text{for } k = 1,2, \tag{4.1}$$

where the weights $w_1, w_2 > 0$ need to be prespecified in advance. Note, that definition (4.1) obviously results in $Z_1^* = \tilde{Z}_1 = Z_1$ and $Var(Z_k^*) = 1$ for $k = 1,2$. Furthermore, under H_0 the vector of test statistics $(Z_1^*, Z_2^*)'$ follows a bivariate normal distribution with mean vector $(0,0)'$ and covariance determined as

$$Cov\left(Z_1^*, Z_2^*\right) = \frac{1}{\sqrt{w_1^2 + w_2^2}} Cov\left(Z_1, w_1 Z_1 + w_2 Z_2\right) = \frac{w_1}{\sqrt{w_1^2 + w_2^2}}. \tag{4.2}$$

By applying the group sequential testing procedure given in (3.3) with the test statistics in (4.1), appropriate boundary values b_1 and b_2 in order to control the overall type I error rate can be determined analogous to the group sequential setting. It should be noted, that it is crucial for exact type I error control that the weights w_1 and w_2 are prespecified in advance and are not altered during trial conduct. The sample sizes, however, can be adapted at the interim analysis based on all available information without inflating the type I error (Brannath et al., 2012).

Lehmacher and Wassmer (1999) proposed to choose the weights as $w_1 = w_2 = 1$ so that the two stages are equally weighted. With this choice the covariance in (4.2) becomes $\frac{1}{\sqrt{2}}$ which exactly is the covariance of the group sequential test statistics with equal stage sizes (cf. Section 3.1.3). Consequently, the critical values of the classical group sequential tests such as Pocock or O'Brien Fleming boundaries can be used. Interestingly, by restricting to equal stages sizes, i.e. $\tilde{n}_A^{(1)} = \tilde{n}_A^{(2)}$ and $\tilde{n}_B^{(1)} = \tilde{n}_B^{(2)}$, it can be easily shown that the test statistic of the final stage Z_2^* is equal to the respective cumulative group sequential test statistic Z_2. Therefore, a possible procedure could take the following form. Determine the boundaries b_1, b_2 and the required maximum sample sizes $n_A^{(2)}, n_B^{(2)}$ for a common group sequential design with two equally sized stages. Suppose, that H_0 is not rejected at the interim analysis. Then, if no adaptation occurs, proceed as originally planned with the group sequential design. Otherwise, calculate the final test statistic Z_2^* with the updated sample sizes and compare it with b_2.

Another interesting option is choosing the weights as the square root of the *originally planned* stage-wise information levels, i.e. $w_1 = \sqrt{\mathcal{I}_1}$ and $w_2 = \sqrt{\mathcal{I}_2 - \mathcal{I}_1}$. This seems to be a sensible choice if an unequal sample size distribution across the stages is planned. Clearly, the covariance in (4.2) then becomes $\sqrt{\mathcal{I}_1 / \mathcal{I}_2}$ so that appropriate boundaries can be determined as described in Section 3.1.2. If furthermore a fixed between-treatment allocation ratio at the two stages is specified, i.e. $\frac{\tilde{n}_A^{(1)}}{\tilde{n}_B^{(1)}} = \frac{\tilde{n}_A^{(2)}}{\tilde{n}_B^{(2)}}$, it can be shown that the test statistic Z_2^* coincides with the common group sequential test statistic Z_2, provided that no adaptation occurs.

4.1.2 Conditional Power

The main question that arises now refers to the procedure that specifies how the sample sizes of the second stage are updated at the interim analysis. A simple and very intuitive approach would be to use the observed treatment differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ to determine the sample sizes that would have been required if the study was performed again. However, this would not adequately take into account that the interim observations considerably influence the test results at the final stage. This is where the so-called *conditional power* comes into play, which is the probability that H_0 will be rejected at the final analysis given the observed interim data.

The concept of conditional power has been first mentioned by Lan et al. (1982) who primarily used it for monitoring repeatedly throughout a clinical trial with fixed maximum sample size. Later, there have been several proposals for using conditional power for re-calculating the sample sizes at an interim analysis in the context of adaptive designs (Proschan and Hunsberger, 1995). Besides this, conditional power also turned out to be a useful tool for interim decision-making in a DSMB. For instance, Betensky (1997) proposed to use conditional power for early stopping to accept the null hypothesis, i.e. termination for futility when the conditional power is too low.

In our context, the conditional power at the interim analysis is given as the probability that $Z_2^* \geq b_2$ at the final stage, given the true treatment difference θ and the data accumulated so far, i.e. $X_{A,1},\dots,X_{A,n_A^{(1)}}$ and $X_{B,1},\dots,X_{B,n_B^{(1)}}$. Since at the interim analysis Z_1 is a sufficient test statistic for θ, it can be used instead of the observed data, so that with definition (4.1) the conditional power is determined as

$$
\begin{aligned}
CP(\theta) &= P\left(\frac{w_1 Z_1 + w_2 \tilde{Z}_2}{\sqrt{w_1^2 + w_2^2}} \geq b_2 \,\middle|\, Z_1, \theta \right) \\
&= P\left(\tilde{Z}_2 \geq \frac{\sqrt{w_1^2 + w_2^2}}{w_2} b_2 - \frac{w_1}{w_2} Z_1 \,\middle|\, Z_1, \theta \right) \\
&= \Phi\left(\frac{\theta}{\sigma} \sqrt{\frac{\tilde{n}_A^{(2)} \tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)} + \tilde{n}_B^{(2)}}} + \frac{w_1}{w_2} \frac{\bar{X}_A^{(1)} - \bar{X}_B^{(1)}}{\sigma} \sqrt{\frac{n_A^{(1)} n_B^{(1)}}{n_A^{(1)} + n_B^{(1)}}} - \frac{\sqrt{w_1^2 + w_2^2}}{w_2} b_2 \right).
\end{aligned} \tag{4.3}
$$

As we can see from (4.3) the conditional power highly depends on the true treatment difference θ, which clearly is unknown. Consequently, the question arises how θ should be replaced. An obvious substitute for θ is the estimator $\bar{X}_A^{(1)} - \bar{X}_B^{(1)}$, which does not, however, take into account the information from other comparable studies that were used in the planning stage. Moreover, this choice could potentially lead to a substantial under- or overestimation of the conditional power when the interim analysis is performed at an early point in time where the variance of $\bar{X}_A^{(1)} - \bar{X}_B^{(1)}$ is high. Another sensible substitute for θ in (4.3) is the anticipated treatment difference between treatment group A and B that was used in the planning stage of the trial to determine the required sample sizes. Combinations of these two choices also seem to be reasonable under some circumstances, but in general this issue remains critical as there clearly is no optimal choice.

By defining $\tilde{n}_B^{(2)} = \tilde{c}_B^{(2)}\tilde{n}_A^{(2)}$ with $\tilde{c}_B^{(2)} > 0$, the required sample size of group A to obtain conditional power $CP \in (0,1)$ is easily determined by solving $CP(\theta) = CP$ and is given as

$$\tilde{n}_A^{(2)} = \frac{\sigma^2}{\theta^2}\frac{1+\tilde{c}_B^{(2)}}{\tilde{c}_B^{(2)}}\left(\frac{\sqrt{w_1^2+w_2^2}}{w_2}b_2 + z_{CP} - \frac{w_1}{w_2}\frac{\bar{X}_A^{(1)} - \bar{X}_B^{(1)}}{\sigma}\sqrt{\frac{n_A^{(1)}n_B^{(1)}}{n_A^{(1)}+n_B^{(1)}}}\right)^2 .$$

Sample size re-calculation based on the conditional power seems to be a reasonable option especially when there is a high level of uncertainty regarding the treatment effects and the standard deviation in the planning stage. Moreover, conditional power can be an aid to decision-making, e.g. whether the trial is terminated early for futility or not. However, the crucial point of substituting the true treatment difference θ in (4.3) remains critical. Using the observed treatment difference $\bar{X}_A^{(1)} - \bar{X}_B^{(1)}$ seems to be a natural choice, but might lead to too large second-stage sample sizes when the standard error at interim is high or the observed treatment difference is small. Nevertheless, adopting the anticipated treatment difference from the initial sample size determination might also lead to a substantial under- or overestimation of the true conditional power. In general, it seems advisable to prespecify an upper limit for the second-stage size like e.g. $\max(\tilde{n}_A^{(2)}) = 2n_A^{(1)}$. For a more detailed overview on critical aspects of using conditional power for sample size reassessment at interim analyses see Bauer and König (2006).

4.1.3 Bayesian Predictive Power

In order to avoid the previously discussed critical issue of substituting the true treatment effects, Spiegelhalter et al. (1986) proposed a Bayesian alternative to the conditional power. Let us therefore suppose that the conditional power is written as a function of the two means μ_A and μ_B instead of θ. The simple idea is to average the conditional power over certain reasonable values of μ_A and μ_B by means of a weight function $\pi(\mu_A, \mu_B \mid \text{interim data})$, which is the joint posterior density of μ_A and μ_B. This density is obtained in a Bayesian fashion by updating the joint prior density for μ_A and μ_B, e.g. determined from previous study results, with the information accumulated so far, namely the interim data. Consequently, the predictive power is given as

$$PP = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} CP(\mu_A - \mu_B)\,\pi\left(\mu_A, \mu_B \,\middle|\, X_{A,1},\dots,X_{A,n_A^{(1)}}, X_{B,1},\dots,X_{B,n_B^{(1)}}\right)\mathrm{d}\mu_A\mathrm{d}\mu_B. \tag{4.4}$$

In order to determine the joint posterior density of μ_A and μ_B, let us assume independent prior distributions for μ_A and μ_B, respectively. Under this reasonable assumption the joint prior distribution is given as the product of the two independent prior distributions. Moreover, the joint posterior density of μ_A and μ_B is given as the product of the posterior densities of μ_A and μ_B, that means we have

$$\pi\left(\mu_A,\mu_B \,\middle|\, X_{A,1},\dots,X_{A,n_A^{(1)}},X_{B,1},\dots,X_{B,n_B^{(1)}}\right)=\pi\left(\mu_A \,\middle|\, X_{A,1},\dots,X_{A,n_A^{(1)}}\right)\pi\left(\mu_B \,\middle|\, X_{B,1},\dots,X_{B,n_B^{(1)}}\right).$$

The posterior density of μ_D, $D = A, B$, is then obtained according to Bayes' theorem as

$$\pi\left(\mu_D \,\middle|\, X_{D,1},\dots,X_{D,n_D^{(1)}}\right)=\frac{L\left(X_{D,1},\dots,X_{D,n_D^{(1)}} \,\middle|\, \mu_D\right)\pi\left(\mu_D\right)}{\int\limits_{-\infty}^{\infty} L\left(X_{D,1},\dots,X_{D,n_D^{(1)}} \,\middle|\, \mu_D\right)\pi\left(\mu_D\right)\mathrm{d}\mu_D}, \tag{4.5}$$

where $\pi\left(\mu_D\right)$ denotes the prior density of μ_D and $L(\dots \mid \mu_D)$ is the likelihood of the interim data given μ_D. The term in the denominator is a normalising factor to ensure that the integral of the posterior density over $(-\infty,\infty)$ is equal to one and its value is not of interest. Consequently, Equation (4.5) is often written as

$$\pi\left(\mu_D \,\middle|\, X_{D,1},\dots,X_{D,n_D^{(1)}}\right)\propto L\left(X_{D,1},\dots,X_{D,n_D^{(1)}} \,\middle|\, \mu_D\right)\pi\left(\mu_D\right),$$

which says that the posterior density is proportional to or has the same shape as the product of the likelihood and the prior density. For the likelihood of the interim data it is a well-known fact that

$$L\left(X_{D,1},\dots,X_{D,n_D^{(1)}} \,\middle|\, \mu_D\right)\propto \exp\left(-\frac{1}{2}\frac{\left(\mu_D-\bar{X}_D^{(1)}\right)^2}{\sigma^2/n_D^{(1)}}\right). \tag{4.6}$$

Suppose now that the prior distribution of μ_D is normal with mean $\mu_{D,0}$ and variance $\sigma_{D,0}^2$. Note, that the parameters $\mu_{D,0}$ and $\sigma_{D,0}^2$ are also called *hyperparameters* in order to distinguish them from the parameters we want to make inference about. It can be easily shown that

$$\begin{aligned}\pi\left(\mu_D \,\middle|\, X_{D,1},\dots,X_{D,n_D^{(1)}}\right)&\propto \exp\left(-\frac{1}{2}\frac{\left(\mu_D-\bar{X}_D^{(1)}\right)^2}{\sigma^2/n_D^{(1)}}\right)\exp\left(-\frac{1}{2}\frac{\left(\mu_D-\mu_{D,0}\right)^2}{\sigma_{D,0}^2}\right)\\&\propto \exp\left(-\frac{1}{2}\frac{\left(\mu_D-\mu_D^*\right)^2}{\sigma_D^{*\,2}}\right),\end{aligned}$$

where μ_D^* and $\sigma_D^{*\,2}$ are given as

$$\mu_D^*=\frac{\frac{1}{\sigma_{D,0}^2}\mu_{D,0}+\frac{n_D^{(1)}}{\sigma^2}\bar{X}_D^{(1)}}{\frac{1}{\sigma_{D,0}^2}+\frac{n_D^{(1)}}{\sigma^2}} \quad\text{and}\quad \sigma_D^{*\,2}=\frac{1}{\frac{1}{\sigma_{D,0}^2}+\frac{n_D^{(1)}}{\sigma^2}}. \tag{4.7}$$

That means, the posterior distribution is normal with mean μ_D^* and variance ${\sigma_D^*}^2$. Consequently, the posterior density of μ_D is given as

$$\pi\left(\mu_D \,\middle|\, X_{D,1},\dots,X_{D,n_D^{(1)}}\right) = \frac{1}{\sigma_D^*}\phi\left(\frac{\mu_D-\mu_D^*}{\sigma_D^*}\right), \tag{4.8}$$

where $\phi(\cdot)$ denotes the probability density function of the standard normal distribution. As the prior and the posterior are in the same family of distributions, namely normal with known variance, the prior is also called a *conjugate prior* for the likelihood function. Choosing a conjugate prior has the decisive advantage that the posterior distribution can be easily obtained by simple mathematical operations. Through this, the central practical problem of finding an analytically tractable or numerical solution for the integral in (4.5) is evaded. Moreover, choosing a conjugate prior gives direct insight on how the prior distribution is updated by the observed interim data, thus giving a better understanding about the updating process. In case of a normal prior for normally distributed data, it becomes obvious by (4.7) that the posterior mean is simply the weighted mean of the prior mean and the sample mean of the interim data, where the weights are the inverse of the respective variances $\sigma_{D,0}^2$ and $\frac{\sigma^2}{n_D^{(1)}}$, respectively. That means, if there is little or only uncertain prior information, $\sigma_{D,0}^2$ should be chosen large, whereas a small $\sigma_{D,0}^2$ is adopted if comprehensive prior information is available. Note that the terms *large* and *small* should be understood relative to the sample variance $\frac{\sigma^2}{n_D^{(1)}}$, as $\sigma_{D,0}^2 = \frac{\sigma^2}{n_D^{(1)}}$ obviously means that the prior and interim information are equally weighted.

In case there is no prior information, it seems natural to choose $\sigma_{D,0} \to \infty$, i.e. a so-called *non-informative* prior. Then, according to (4.7) the posterior distribution is completely determined by the interim data, as clearly $\mu_D^* \to \bar{X}_D^{(1)}$ and ${\sigma_D^*}^2 \to \frac{\sigma^2}{n_D^{(1)}}$. Another way to see this, is to adopt the non-informative prior density $\pi(\mu_D) = 1 \ \forall \mu_D \in \mathbb{R}$, which obviously is an improper probability density function as $\int_{-\infty}^{\infty} 1 \mathrm{d}\mu_D = \infty$. However, it follows easily by (4.5) and (4.6) that the posterior density is a proper density function, namely that of the normal distribution with mean and variance equal to $\bar{X}_D^{(1)}$ and $\frac{\sigma^2}{n_D^{(1)}}$, respectively. Note that, for Gaussian distributions with known variance, the improper prior $\pi(\mu_D) = 1 \ \forall \mu_D \in \mathbb{R}$ coincides with *Jeffreys prior*, which is proportional to the square root of the Fisher information. In contrast, if we choose a highly informative prior, i.e. $\sigma_{D,0} \to 0$, then the posterior distribution is completely described by the prior information as we have $\mu_D^* \to \mu_{D,0}$ and ${\sigma_D^*}^2 \to 0$ according to (4.7).

Suppose now that the prior distributions of μ_A and μ_B are normal with means $\mu_{A,0}$ and $\mu_{B,0}$ and variances $\sigma_{A,0}^2$ and $\sigma_{B,0}^2$, respectively. Then, according to (4.4) and (4.8) the predictive power is given as

$$\begin{aligned} PP &= \frac{1}{\sigma_A^*\sigma_B^*}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} CP\left(\mu_A-\mu_B\right)\phi\left(\frac{\mu_A-\mu_A^*}{\sigma_A^*}\right)\phi\left(\frac{\mu_B-\mu_B^*}{\sigma_B^*}\right)\mathrm{d}\mu_A\mathrm{d}\mu_B \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} CP\left(\sigma_A^*x+\mu_A^*-\sigma_B^*y-\mu_B^*\right)\phi(x)\,\phi(y)\,\mathrm{d}x\mathrm{d}y, \end{aligned}$$

where the posterior means μ_A^*, μ_B^* and variances ${\sigma_A^*}^2, {\sigma_B^*}^2$ are defined according to (4.7). By substituting the conditional power given in (4.3) into this formula and carrying out the integration we obtain that

$$PP = \Phi\left(\sqrt{\frac{1}{1+\frac{{\sigma_A^*}^2+{\sigma_B^*}^2}{\sigma^2}\,\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}}}\left[\frac{\mu_A^*-\mu_B^*}{\sigma}\sqrt{\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}}+\frac{w_1}{w_2}\,\frac{\bar{X}_A^{(1)}-\bar{X}_B^{(1)}}{\sigma}\sqrt{\frac{n_A^{(1)}n_B^{(1)}}{n_A^{(1)}+n_B^{(1)}}}-\frac{\sqrt{w_1^2+w_2^2}}{w_2}b_2\right]\right), \tag{4.9}$$

where the relationship $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Phi(a+bx+cy)\phi(x)\phi(y)\mathrm{d}x\mathrm{d}y = \Phi(a/\sqrt{1+b^2+c^2})$ has been used (see Owen, 1980, formula 10,020). Another way to obtain this formula is to realise that the predictive power in (4.4) is just the probability to reject the null hypothesis at the final stage, given the interim data and assuming that the treatment effect μ_D, $D = A, B$, is normally distributed with mean and variance given in (4.7). That means we have

$$\begin{aligned} PP &= P\left(Z_2^* \geq b_2 \,\middle|\, Z_1, \mu_D \sim N\left(\mu_D^*, {\sigma_D^*}^2\right), D = A, B\right) \\ &= P\left(\tilde{Z}_2 \geq \frac{\sqrt{w_1^2+w_2^2}}{w_2}b_2 - \frac{w_1}{w_2}Z_1 \,\middle|\, Z_1, \mu_D \sim N\left(\mu_D^*, {\sigma_D^*}^2\right), D = A, B\right). \end{aligned} \tag{4.10}$$

Let us now derive the so-called *posterior predictive distribution* of the second-stage test statistic $\tilde{Z}_2$, which is the distribution that a new test statistic $\tilde{Z}_2$ would have, given the prior distributions for μ_A and μ_B and the observed interim data. In a frequentist framework we obviously have

$$\tilde{Z}_2 = \frac{\mu_A - \mu_B}{\sigma}\sqrt{\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}} + \epsilon \quad \text{with} \quad \epsilon \sim N(0,1).$$

If we further assume that μ_D is normally distributed with mean μ_D^* and variance ${\sigma_D^*}^2$ given in (4.7), i.e. $\mu_D = \mu_D^* + \epsilon_D$ with $\epsilon_D \sim N(0, {\sigma_D^*}^2)$ and $D = A, B$, this results in

$$\tilde{Z}_2 = \frac{\mu_A^* - \mu_B^*}{\sigma}\sqrt{\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}} + \frac{\epsilon_A - \epsilon_B}{\sigma}\sqrt{\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}} + \epsilon.$$

Consequently, the posterior predictive distribution of $\tilde{Z}_2$ is a normal distribution with mean $\tilde{\mu}^{(2)*}$ and variance $\tilde{\sigma}^{(2)*}$ given as

$$\tilde{\mu}^{(2)*} = \frac{\mu_A^* - \mu_B^*}{\sigma}\sqrt{\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}} \quad \text{and} \quad \tilde{\sigma}^{(2)*} = 1 + \frac{{\sigma_A^*}^2 + {\sigma_B^*}^2}{\sigma^2}\,\frac{\tilde{n}_A^{(2)}\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}.$$

With this knowledge, the predictive power in (4.10) can be directly calculated by means of the cumulative normal distribution function and we easily obtain the formula given in (4.9).

Taking a closer look at the predictive power formula given in (4.9), distinct similarities with the conditional power in (4.3) become apparent. If the true treatment difference $\theta = \mu_A - \mu_B$ in (4.3) is replaced by the posterior mean difference $\mu_A^* - \mu_B^*$, the conditional power is almost equal to the predictive power, despite a certain factor formed by the variances and the second-stage sample sizes. As this factor clearly lies between zero and one, the predictive power can also be seen as a shrinkage of the conditional power towards 0.5. With regard to mid-trial sample size re-calculation this means that the sample sizes determined based on the predictive power will generally be much larger than those calculated based on the conditional power, because the targeted power is usually chosen much larger than 0.5. Intuitively, this is hardly surprising as the predictive power takes into account all uncertainties of the second stage by integrating over a range of potential treatment effects μ_A and μ_B.

As mentioned earlier, by using non-informative priors, i.e. $\sigma_{A,0}, \sigma_{B,0} \to \infty$, we have $\mu_D^* \to \bar{X}_D^{(1)}$ and ${\sigma_D^*}^2 \to \frac{\sigma^2}{n_D^{(1)}}$ for $D = A, B$ and the predictive power becomes

$$PP = \Phi\left(\sqrt{\frac{1}{1 + \frac{n_A^{(1)}+n_B^{(1)}}{n_A^{(1)} n_B^{(1)}} \frac{\tilde{n}_A^{(2)} \tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}}} \left[\frac{\bar{X}_A^{(1)} - \bar{X}_B^{(1)}}{\sigma}\left(\sqrt{\frac{\tilde{n}_A^{(2)} \tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)} + \tilde{n}_B^{(2)}}} + \frac{w_1}{w_2}\sqrt{\frac{n_A^{(1)} n_B^{(1)}}{n_A^{(1)} + n_B^{(1)}}}\right) - \frac{\sqrt{w_1^2 + w_2^2}}{w_2} b_2\right]\right),$$

which is somewhat similar to the conditional power at $\bar{X}_A^{(1)} - \bar{X}_B^{(1)}$, again, except for a factor depending on the first and second-stage sample sizes. If, in contrast, highly informative priors with $\sigma_{A,0}, \sigma_{B,0} \to 0$ are used, we have $\mu_D^* \to \mu_{D,0}$ and ${\sigma_D^*}^2 \to 0$ for $D = A, B$, so that the predictive power becomes the conditional power at the prior mean difference $\mu_{A,0} - \mu_{B,0}$.

If the alternative hypothesis is true, i.e. $\mu_A > \mu_B$, it can be easily seen from (4.3) that the conditional power converges to one for increasing second-stage sample sizes. Unfortunately, this desirable property does not hold for the predictive power. For $\tilde{n}_A^{(2)}, \tilde{n}_B^{(2)} \to \infty$ the first factor on the right-hand side of (4.9) clearly converges to zero. Moreover, we have

$$\sqrt{\frac{\frac{\tilde{n}_A^{(2)} \tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}}{1 + \frac{{\sigma_A^*}^2+{\sigma_B^*}^2}{\sigma^2} \frac{\tilde{n}_A^{(2)} \tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}}} = \sqrt{\frac{1}{\frac{\tilde{n}_A^{(2)}+\tilde{n}_B^{(2)}}{\tilde{n}_A^{(2)} \tilde{n}_B^{(2)}} + \frac{{\sigma_A^*}^2+{\sigma_B^*}^2}{\sigma^2}}} \xrightarrow{\tilde{n}_A^{(2)}, \tilde{n}_B^{(2)} \to \infty} \frac{\sigma}{\sqrt{{\sigma_A^*}^2 + {\sigma_B^*}^2}}.$$

Consequently, it follows for the predictive power that

$$PP \xrightarrow{\tilde{n}_A^{(2)}, \tilde{n}_B^{(2)} \to \infty} \Phi\left(\frac{\mu_A^* - \mu_B^*}{\sqrt{{\sigma_A^*}^2 + {\sigma_B^*}^2}}\right), \tag{4.11}$$

which can be interpreted as one minus the unadjusted p-value at interim that is formed by using the standardised test statistic for the posterior mean difference $\mu_A^* - \mu_B^*$. When a non-informative prior is used, this term clearly becomes $\Phi(Z_1)$, which exactly is one minus the unadjusted interim p-value. That means, some levels of predictive power will never be obtained, however large the second-stage sample sizes are. Besides the considerably larger second-stage sizes when the sample sizes are re-calculated according to the predictive instead of the conditional power, this is another very undesirable property concerning sample size re-assessment. Incorporating prior information will reduce some of these issues since, for example, the term in (4.11) tends towards one for more informative priors, i.e. $\sigma_{A,0}, \sigma_{B,0} \rightarrow 0$. Moreover, the predictive power then converges towards the conditional power calculated at the anticipated prior mean difference $\mu_{A,0} - \mu_{B,0}$. For more information on this topic see Dallow and Fina (2011), who gave a detailed overview on the potential pitfalls when using the predictive power to assess mid-trial data.

It should be kept in mind that, actually, predictive power and classical power calculations are two totally different concepts. For standard sample size determinations in the planning stage the power is calculated at a specific treatment difference and is not averaged over a certain range of possible values as in the predictive power calculation. Not least because of this, the conditional power seems to be more appropriate than the predictive power when it comes to sample size re-calculations at an interim analysis. Nevertheless, if one is aware of the potential perils associated with predictive power, it can be a useful tool for interim decision-making, e.g. with regard to futility stopping. As predictive and conditional power are closely related, they should generally be used jointly when assessing mid-trial data.

As it has been mentioned earlier, when calculating the conditional power in (4.3) it remains critical to choose an sensible substitute for the treatment difference θ. One useful option might also be to use the posterior mean difference $\mu_A^* - \mu_B^*$ that is obtained in a Bayesian fashion as described above by updating prior beliefs from the planning stage with the observed mid-trial data. Note that this is a combination of choosing the preplanned or the observed treatment difference, as the posterior mean difference simply is a weighted average of the two quantities.

4.2 Adaptive Designs for Three-Arm Non-Inferiority Trials

By means of the weighted inverse normal method the group sequential testing procedure proposed in Section 3.2 will now be extended to two-stage adaptive designs that allow data-dependent design changes at the interim analysis. Moreover, corresponding conditional and predictive power formulas will be derived for the proposed adaptive three-arm non-inferiority designs.

4.2.1 General Design

As in the previous part of this thesis suppose that all observations of the endpoint under the test, control and placebo treatment are mutually independent and normally distributed with common, known variance σ^2, namely $X_{T,i} \sim N(\mu_T, \sigma^2)$, $X_{C,i} \sim N(\mu_C, \sigma^2)$ and $X_{P,i} \sim N(\mu_P, \sigma^2)$ for $i = 1, 2, \ldots$. Let furthermore $\tilde{n}_T^{(1)}, \tilde{n}_T^{(2)}$, $\tilde{n}_C^{(1)}, \tilde{n}_C^{(2)}$ and $\tilde{n}_P^{(1)}, \tilde{n}_P^{(2)}$ denote the stage-wise sample sizes of the test, control and placebo group, respectively. Again, note the difference to the cumulative sample sizes used in the group sequential setting, i.e. we have the relationships $\tilde{n}_D^{(1)} = n_D^{(1)}$ and $\tilde{n}_D^{(2)} = n_D^{(2)} - n_D^{(1)}$ for $D = T, C, P$. The independent stage-wise test statistics of the test versus placebo superiority test and the non-inferiority comparison between the test and control treatment are now given as

$$\tilde{Z}_{TP}^{(k)} = \frac{\Delta\bar{X}_T^{(k)} - \Delta\bar{X}_P^{(k)}}{\sigma}\sqrt{\frac{\tilde{n}_T^{(k)}\tilde{n}_P^{(k)}}{\tilde{n}_T^{(k)} + \tilde{n}_P^{(k)}}} \text{ and } \tilde{Z}_{TC}^{(k)} = \frac{\Delta\bar{X}_T^{(k)} - \Delta\bar{X}_C^{(k)} + \Delta_{ni}}{\sigma}\sqrt{\frac{\tilde{n}_T^{(k)}\tilde{n}_C^{(k)}}{\tilde{n}_T^{(k)} + \tilde{n}_C^{(k)}}} \text{ for } k = 1, 2,$$

where the stage-wise sample means are defined as $\Delta\bar{X}_D^{(1)} = \bar{X}_D^{(1)}$ and $\Delta\bar{X}_D^{(2)} = \sum_{i=n_D^{(1)}+1}^{n_D^{(2)}} X_{D,i}/\tilde{n}_D^{(2)}$ for $D = T, C, P$. According to (4.1) the respective test statistics based on the inverse normal method are given as

$$Z_{TP}^{(k)*} = \frac{\sum_{i=1}^{k} w_{TP}^{(i)}\tilde{Z}_{TP}^{(i)}}{\sqrt{\sum_{i=1}^{k} {w_{TP}^{(i)}}^2}} \text{ and } Z_{TC}^{(k)*} = \frac{\sum_{i=1}^{k} w_{TC}^{(i)}\tilde{Z}_{TC}^{(i)}}{\sqrt{\sum_{i=1}^{k} {w_{TC}^{(i)}}^2}} \text{ for } k = 1, 2, \tag{4.12}$$

with predefined weights $w_{TP}^{(1)}, w_{TP}^{(2)} > 0$ and $w_{TC}^{(1)}, w_{TC}^{(2)} > 0$. By using these test statistics the testing procedure of the adaptive three-arm non-inferiority designs takes the same form as for the group sequential setting given in (3.14). According to (4.2) appropriate rejection boundaries $b_{TP}^{(1)}, b_{TP}^{(2)}$ and $b_{TC}^{(1)}, b_{TC}^{(2)}$ are determined similar to (3.15) and (3.16) as a solution of

$$\Phi^{\mathbf{\Sigma}_{TP}}\left(b_{TP}^{(1)}, b_{TP}^{(2)}\right) = \Phi^{\mathbf{\Sigma}_{TC}}\left(b_{TC}^{(1)}, b_{TC}^{(2)}\right) = 1 - \alpha,$$

where the covariance matrices $\mathbf{\Sigma}_{TP}$ and $\mathbf{\Sigma}_{TC}$ are defined as

$$\mathbf{\Sigma}_{TP} = \begin{pmatrix} 1 & \frac{w_{TP}^{(1)}}{\sqrt{{w_{TP}^{(1)}}^2 + {w_{TP}^{(1)}}^2}} \\ \frac{w_{TP}^{(1)}}{\sqrt{{w_{TP}^{(1)}}^2 + {w_{TP}^{(1)}}^2}} & 1 \end{pmatrix} \text{ and } \mathbf{\Sigma}_{TC} = \begin{pmatrix} 1 & \frac{w_{TC}^{(1)}}{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(1)}}^2}} \\ \frac{w_{TC}^{(1)}}{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(1)}}^2}} & 1 \end{pmatrix}. \tag{4.13}$$

Through this the adaptive testing procedure clearly controls the overall type I error rate in the strong sense by α. The proof can be carried over from the group sequential setting (see page 53). The overall type I error rate is controlled for any data-dependent choice of the second-stage sample sizes, however, it is vital that the weights $w_{TP}^{(1)}, w_{TP}^{(2)}$ and $w_{TC}^{(1)}, w_{TC}^{(2)}$ in (4.12) are prespecified and not altered during trial conduct. Otherwise the type I error rate will be inflated.

As mentioned earlier it seems natural to choose the weights as the square roots of the *originally planned* stage-wise information levels, i.e. $w_{TP}^{(1)} = \sqrt{\mathscr{I}_{TP}^{(1)}}$, $w_{TP}^{(2)} = \sqrt{\mathscr{I}_{TP}^{(2)} - \mathscr{I}_{TP}^{(1)}}$ and $w_{TC}^{(1)} = \sqrt{\mathscr{I}_{TC}^{(1)}}$, $w_{TC}^{(2)} = \sqrt{\mathscr{I}_{TC}^{(2)} - \mathscr{I}_{TC}^{(1)}}$. Then, the covariance matrices in (4.13) become equal to those of the group sequential testing procedure given in (3.18). If we furthermore assume that the between-treatment allocation ratios are equal across the two stages, i.e. $\frac{\tilde{n}_D^{(1)}}{\tilde{n}_T^{(1)}} = \frac{\tilde{n}_D^{(2)}}{\tilde{n}_T^{(2)}}$ for $D = C, P$, the test statistics in (4.12) coincide with those of the group sequential testing procedure suggested in the previous chapter. According to this, we can start by choosing an appropriate group sequential design according to Section 3.2 in the planning stage, including calculation of the respective boundary values and maximum as well as expected sample sizes. If the sample sizes are not altered at the interim analysis, the trial proceeds as in the group sequential setting and uses the cumulative final test statistics $Z_{TP}^{(2)}$ and $Z_{TC}^{(2)}$. Otherwise the weighted inverse normal test statistics $Z_{TP}^{(2)*}$ and $Z_{TC}^{(2)*}$ are used at the final stage.

If a prespecified sample size re-calculation procedure is used at the interim analysis, the actual overall power and expected sample sizes of the proposed adaptive testing procedure are clearly different from those of the group sequential testing procedure. Note that these characteristics are highly dependent on the re-calculation rule used to determine the second-stage sample sizes and can be determined by means of simulation.

4.2.2 Conditional Power

Let us now consider the conditional power of the proposed procedure in order to assess the mid-trial data and potentially re-calculate the sample sizes of the second stage. As there are two null hypotheses being investigated with the adaptive testing procedure, there is also more than one conditional power of interest at the interim analysis. The first one obviously is the conditional power to reject $H_{0,TP}^{(s)}$ at the final analysis, which generally is of interest when $Z_{TP}^{(1)} < b_{TP}^{(1)}$ is observed. According to (4.3) it is obtained as

$$CP_{TP}(\theta_{TP}) = \Phi\left(\frac{\theta_{TP}}{\sigma}\sqrt{\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}} + \frac{w_{TP}^{(1)}}{w_{TP}^{(2)}}\frac{\bar{X}_T^{(1)}-\bar{X}_P^{(1)}}{\sigma}\sqrt{\frac{n_T^{(1)}n_P^{(1)}}{n_T^{(1)}+n_P^{(1)}}} - \frac{\sqrt{{w_{TP}^{(1)}}^2+{w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}} b_{TP}^{(2)} \right). \tag{4.14}$$

Sample size re-calculation based on this conditional power is generally not of interest, as this would completely ignore the sample size of the control group.

The next one is the conditional power to reject $H_{0,TC}^{(n)}$ after the second stage. Due to the hierarchical nature of the testing procedure this probability clearly depends on the interim test statistic of the superiority comparison between test and placebo $Z_{TP}^{(1)}$.

Let us first consider the case, that $H_{0,TP}^{(s)}$ has already been rejected at the interim analysis, i.e. $Z_{TP}^{(1)} \geq b_{TP}^{(1)}$ holds. Then, the conditional power to reject $H_{0,TC}^{(n)}$ can be derived analogously to Section 4.1.2. According to (4.3) we have

$$CP_{TC}(\theta_{TC}) = \Phi\left(\frac{\theta_{TC}+\Delta_{ni}}{\sigma}\sqrt{\frac{\tilde{n}_T^{(2)}\tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}}} + \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}}\frac{\bar{X}_T^{(1)}-\bar{X}_C^{(1)}+\Delta_{ni}}{\sigma}\sqrt{\frac{n_T^{(1)}n_C^{(1)}}{n_T^{(1)}+n_C^{(1)}}} - \frac{\sqrt{{w_{TC}^{(1)}}^2+{w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} \right). \tag{4.15}$$

The required second-stage sample sizes to obtain a specific conditional power can be easily determined by solving $CP_{TC}(\theta_{TC}) = CP_{TC}$, with $CP_{TC} \in (0,1)$ being the targeted conditional power. Let therefore $\tilde{n}_C^{(2)} = \tilde{c}_C^{(2)} \tilde{n}_T^{(2)}$ with $\tilde{c}_C^{(2)} > 0$, then the required test group size is determined as

$$\tilde{n}_T^{(2)} = \frac{\sigma^2}{(\theta_{TC} + \Delta_{ni})^2} \frac{1 + \tilde{c}_C^{(2)}}{\tilde{c}_C^{(2)}} \left(\frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} + z_{CP_{TC}} - \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} \frac{\bar{X}_T^{(1)} - \bar{X}_C^{(1)} + \Delta_{ni}}{\sigma} \sqrt{\frac{n_T^{(1)} n_C^{(1)}}{n_T^{(1)} + n_C^{(1)}}} \right)^2 .$$

For the case that $H_{0,TP}^{(s)}$ has not been rejected at interim, i.e. $Z_{TP}^{(1)} < b_{TP}^{(1)}$, the conditional power to reject $H_{0,TC}^{(n)}$ furthermore depends on the rejection of $H_{0,TP}^{(s)}$ at the final analysis. The respective conditional power to reject both null hypotheses at the final analysis can be determined in a similar fashion as above. Therefore, we need to determine the joint distribution of the second-stage test statistics $\tilde{Z}_{TP}^{(2)}$ and $\tilde{Z}_{TC}^{(2)}$.

It can be easily seen that the vector $(\tilde{Z}_{TP}^{(2)}, \tilde{Z}_{TC}^{(2)})'$ is bivariate normally distributed with mean vector $\tilde{\boldsymbol{\mu}}^{(2)}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}^{(2)}$ determined as

$$\tilde{\boldsymbol{\mu}}^{(2)} = \left(\frac{\theta_{TP}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}}, \frac{\theta_{TC} + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} \right)' \quad \text{and}$$

$$\tilde{\boldsymbol{\Sigma}}^{(2)} = \begin{pmatrix} 1 & \tilde{\rho}^{(2)} \\ \tilde{\rho}^{(2)} & 1 \end{pmatrix} \text{ with } \tilde{\rho}^{(2)} = Cov\left(\tilde{Z}_{TP}^{(2)}, \tilde{Z}_{TC}^{(2)}\right) = \sqrt{\frac{\tilde{n}_C^{(2)} \tilde{n}_P^{(2)}}{\left(\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}\right)\left(\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}\right)}}. \tag{4.16}$$

Consequently, the conditional power to reject both null hypotheses at the final analysis given the data from the first stage is determined as

$$\begin{aligned} CP_{TP,TC}(\theta_{TP}, \theta_{TC}) &= P\left(\left\{Z_{TP}^{(2)*} \geq b_{TP}^{(2)}\right\} \cap \left\{Z_{TC}^{(2)*} \geq b_{TC}^{(2)}\right\} \,\middle|\, Z_{TP}^{(1)}, Z_{TC}^{(1)}, \theta_{TP}, \theta_{TC}\right) \\ &= P\left(\left\{ \tilde{Z}_{TP}^{(2)} \geq \frac{\sqrt{{w_{TP}^{(1)}}^2 + {w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}} b_{TP}^{(2)} - \frac{w_{TP}^{(1)}}{w_{TP}^{(2)}} Z_{TP}^{(1)} \right\}\right. \\ &\qquad \left. \cap \left\{ \tilde{Z}_{TC}^{(2)} \geq \frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} - \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} Z_{TC}^{(1)} \right\} \,\middle|\, Z_{TP}^{(1)}, Z_{TC}^{(1)}, \theta_{TP}, \theta_{TC} \right) \\ &= \Phi^{\tilde{\boldsymbol{\Sigma}}^{(2)}}\left(\frac{\theta_{TP}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}} + \frac{w_{TP}^{(1)}}{w_{TP}^{(2)}} \frac{\bar{X}_T^{(1)} - \bar{X}_P^{(1)}}{\sigma} \sqrt{\frac{n_T^{(1)} n_P^{(1)}}{n_T^{(1)} + n_P^{(1)}}} - \frac{\sqrt{{w_{TP}^{(1)}}^2 + {w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}} b_{TP}^{(2)}, \right. \\ &\qquad \left. \frac{\theta_{TC} + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} + \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} \frac{\bar{X}_T^{(1)} - \bar{X}_C^{(1)} + \Delta_{ni}}{\sigma} \sqrt{\frac{n_T^{(1)} n_C^{(1)}}{n_T^{(1)} + n_C^{(1)}}} - \frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} \right), \end{aligned} \tag{4.17}$$

with the covariance matrix $\tilde{\boldsymbol{\Sigma}}^{(2)}$ given in (4.16). By defining $\tilde{n}_C^{(2)} = \tilde{c}_C^{(2)} \tilde{n}_T^{(2)}$ and $\tilde{n}_P^{(2)} = \tilde{c}_P^{(2)} \tilde{n}_T^{(2)}$ for some $\tilde{c}_C^{(2)}, \tilde{c}_P^{(2)} > 0$, the required second-stage sample size of the test group can be found by numerically solving $CP_{TP,TC}(\theta_{TP}, \theta_{TC}) = CP_{TP,TC}$, with $CP_{TP,TC} \in (0,1)$ being the targeted conditional power. Note, that the covariance of the test statistics in (4.16) then becomes $\tilde{\rho}^{(2)} = \sqrt{\frac{\tilde{c}_C^{(2)}\tilde{c}_P^{(2)}}{(1+\tilde{c}_C^{(2)})(1+\tilde{c}_P^{(2)})}}$.

As it has been mentioned earlier, finding sensible substitutes for the true treatment differences θ_{TP} and θ_{TC} in (4.14), (4.15) and (4.17) remains a critical issue, so that the conditional power must always be assessed with great caution. Useful options are the observed treatment differences at interim, the anticipated treatment differences used for sample size calculation and the posterior mean differences.

4.2.3 Bayesian Predictive Power

As noted in Section 1.3.1, the active comparator used in non-inferiority trials generally is an approved treatment that has already been extensively studied in the past, namely most often in placebo-controlled superiority trials. Therefore, in the planning stage of a three-arm non-inferiority trial, there should be a fairly large amount of information on the control treatment and placebo effect. It is also for this reason that the predictive power seems to be particularly suitable for assessing the interim data in such trials.

Suppose we have mutually independent prior distributions for the treatment effects μ_T, μ_C and μ_P that are normal with means $\mu_{T,0}$, $\mu_{C,0}$ and $\mu_{P,0}$ and variances $\sigma^2_{T,0}$, $\sigma^2_{C,0}$ and $\sigma^2_{P,0}$, respectively. The hyperparameters of the prior distributions could, for example, be determined by means of historical data. Analogous to Section 4.1.3 it can be shown that the respective posterior distribution of μ_D for $D = T, C, P$ at the interim analysis is normal with mean μ_D^* and variance ${\sigma_D^*}^2$ given in (4.7). The respective posterior density of μ_D, $D = T, C, P$ is given in (4.8).

Let us start with the superiority comparison between the test treatment and placebo. According to (4.9) the predictive power to reject $H_{0,TP}^{(s)}$ at the final stage is given as

$$PP_{TP} = \Phi\left(\sqrt{\frac{1}{1+\frac{{\sigma_T^*}^2+{\sigma_P^*}^2}{\sigma^2}\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}}}\left[\frac{\mu_T^*-\mu_P^*}{\sigma}\sqrt{\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}}+\frac{w_{TP}^{(1)}}{w_{TP}^{(2)}}\frac{\bar{X}_T^{(1)}-\bar{X}_P^{(1)}}{\sigma}\sqrt{\frac{n_T^{(1)}n_P^{(1)}}{n_T^{(1)}+n_P^{(1)}}}\right.\right.$$
$$\left.\left.-\frac{\sqrt{{w_{TP}^{(1)}}^2+{w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}}b_{TP}^{(2)}\right]\right). \tag{4.18}$$

By using non-informative priors for the test treatment and placebo effect, i.e. $\sigma_{T,0}, \sigma_{P,0} \to \infty$, the predictive power in (4.18) becomes

$$PP_{TP} = \Phi\left(\sqrt{\frac{1}{1+\frac{n_T^{(1)}+n_P^{(1)}}{n_T^{(1)}n_P^{(1)}}\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}}}\left[\frac{\bar{X}_T^{(1)}-\bar{X}_P^{(1)}}{\sigma}\left(\sqrt{\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}}+\frac{w_{TP}^{(1)}}{w_{TP}^{(2)}}\sqrt{\frac{n_T^{(1)}n_P^{(1)}}{n_T^{(1)}+n_P^{(1)}}}\right)\right.\right.$$
$$\left.\left.-\frac{\sqrt{{w_{TP}^{(1)}}^2+{w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}}b_{TP}^{(2)}\right]\right).$$

If instead highly informative priors are used, it can be easily seen that the predictive power converges to the respective conditional power given in (4.14) calculated at the prior mean difference between test and placebo, i.e. $CP_{TP}(\mu_{T,0} - \mu_{P,0})$. Analogous to Section 4.1.3 it can also be shown that for increasing second-stage sample sizes we have

$$PP_{TP} \xrightarrow{\tilde{n}_T^{(2)}, \tilde{n}_P^{(2)} \to \infty} \Phi\left(\frac{\mu_T^* - \mu_P^*}{\sqrt{{\sigma_T^*}^2 + {\sigma_P^*}^2}} \right).$$

That means the predictive power converges to one minus the unadjusted interim p-value for the test versus placebo supertiority comparison that is formed by using the standardised test statistic for the posterior mean difference $\mu_T^* - \mu_P^*$. When non-informative priors are used this limit clearly becomes $\Phi(Z_{TP}^{(1)})$ which exactly is one minus the unadjusted interim p-value.

As for the conditional power let us now investigate the test versus control non-inferiority comparison for the case that $H_{0,TP}^{(s)}$ has already been rejected at the interim analysis. Then, the predictive power to reject $H_{0,TC}^{(n)}$ can be derived analogously to the predictive power to reject $H_{0,TP}^{(s)}$ and we have

$$\begin{aligned} PP_{TC} = \Phi\Bigg(\sqrt{\frac{1}{1 + \frac{{\sigma_T^*}^2 + {\sigma_C^*}^2}{\sigma^2} \frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}}} \Bigg[& \frac{\mu_T^* - \mu_C^* + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} + \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} \frac{\bar{X}_T^{(1)} - \bar{X}_C^{(1)} + \Delta_{ni}}{\sigma} \sqrt{\frac{n_T^{(1)} n_C^{(1)}}{n_T^{(1)} + n_C^{(1)}}} \\ & - \frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} \Bigg] \Bigg). \end{aligned} \tag{4.19}$$

Furthermore, all considerations on PP_{TP} carry over to PP_{TC}. When non-informative priors are used for the test and control treatment effect, the predictive power in (4.19) becomes

$$\begin{aligned} PP_{TC} = \Phi\Bigg(\sqrt{\frac{1}{1 + \frac{n_T^{(1)} + n_C^{(1)}}{n_T^{(1)} n_C^{(1)}} \frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}}} \Bigg[& \frac{\bar{X}_T^{(1)} - \bar{X}_C^{(1)} + \Delta_{ni}}{\sigma} \left(\sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} + \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} \sqrt{\frac{n_T^{(1)} n_C^{(1)}}{n_T^{(1)} + n_C^{(1)}}} \right) \\ & - \frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} \Bigg] \Bigg). \end{aligned}$$

When highly informative priors are used, i.e. $\sigma_{T,0}, \sigma_{C,0} \to 0$, the predictive power converges to $CP_{TC}(\mu_{T,0} - \mu_{C,0})$. Moreover, for increasing second-stage sample sizes we have

$$PP_{TC} \xrightarrow{\tilde{n}_T^{(2)}, \tilde{n}_C^{(2)} \to \infty} \Phi\left(\frac{\mu_T^* - \mu_C^* + \Delta_{ni}}{\sqrt{{\sigma_T^*}^2 + {\sigma_C^*}^2}} \right),$$

where the limit becomes one minus the unadjusted interim p-value for testing $H_{0,TC}^{(n)}$ when non-informative priors are used.

The predictive power to reject both null hypotheses at the final analysis can be determined by means of averaging the respective conditional power given in (4.17) over the posterior distributions of the treatment effects μ_T, μ_C and μ_P. That means, we have

$$PP_{TP,TC} = \frac{1}{\sigma_T^* \sigma_P^* \sigma_C^*} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} CP_{TP,TC}\left(\mu_T - \mu_P, \mu_T - \mu_C\right)$$
$$\phi\left(\frac{\mu_T - \mu_T^*}{\sigma_T^*}\right) \phi\left(\frac{\mu_P - \mu_P^*}{\sigma_P^*}\right) \phi\left(\frac{\mu_C - \mu_C^*}{\sigma_C^*}\right) \mathrm{d}\mu_T \mathrm{d}\mu_P \mathrm{d}\mu_C.$$

As we will now see the predictive power can also be written as a double integral, which considerably simplifies the computation. Let us therefore consider the posterior distributions of the treatment differences $\theta_{TP} = \mu_T - \mu_P$ and $\theta_{TC} = \mu_T - \mu_C$, which are obtained as

$$\theta_{TP} \Big| X_{T,1}, \ldots, X_{T,n_T^{(1)}}, X_{P,1}, \ldots, X_{P,n_P^{(1)}} \sim N\left(\mu_T^* - \mu_P^*, {\sigma_T^*}^2 + {\sigma_P^*}^2\right) \quad \text{and}$$
$$\theta_{TC} \Big| X_{T,1}, \ldots, X_{T,n_T^{(1)}}, X_{C,1}, \ldots, X_{C,n_C^{(1)}} \sim N\left(\mu_T^* - \mu_C^*, {\sigma_T^*}^2 + {\sigma_C^*}^2\right).$$

It can be easily seen that, conditional on the interim data, the vector $(\theta_{TP}, \theta_{TC})'$ follows a bivariate normal distribution with mean vector $\boldsymbol{\theta}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$ given as

$$\boldsymbol{\theta}^* = \begin{pmatrix} \mu_T^* - \mu_P^* \\ \mu_T^* - \mu_C^* \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} {\sigma_T^*}^2 + {\sigma_P^*}^2 & {\sigma_T^*}^2 \\ {\sigma_T^*}^2 & {\sigma_T^*}^2 + {\sigma_C^*}^2 \end{pmatrix}.$$

Consequently, it follows that the predictive power to reject $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ at the final analysis is given as

$$PP_{TP,TC} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} CP_{TP,TC}\left(\theta_{TP}, \theta_{TC}\right) \phi^{\boldsymbol{\Sigma}^*}\left(\theta_{TP} - \mu_T^* + \mu_P^*, \theta_{TC} - \mu_T^* + \mu_C^*\right) \mathrm{d}\theta_{TP} \mathrm{d}\theta_{TC},$$

where the respective conditional power $CP_{TP,TC}(\cdot,\cdot)$ is given in (4.17) and $\phi^{\boldsymbol{\Sigma}}(\cdot,\cdot)$ denotes the probability density function of the bivariate normal distribution with mean vector $(0,0)'$ and covariance matrix $\boldsymbol{\Sigma}$.

Finally, with the same considerations as in Section 4.1.3, we will now show that the predictive power can be solely expressed by the cumulative bivariate normal distribution function, which even more shortens the computation time. Obviously, the predictive power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ at the final analysis given the interim data is given as

$$\begin{aligned} PP_{TP,TC} &= P\left(\left\{Z_{TP}^{(2)*} \geq b_{TP}^{(2)}\right\} \cap \left\{Z_{TC}^{(2)*} \geq b_{TC}^{(2)}\right\} \Big| \, Z_{TP}^{(1)}, Z_{TC}^{(1)}, \mu_D \sim N\left(\mu_D^*, {\sigma_D^*}^2\right), D = T,C,P\right) \\ &= P\left(\left\{\tilde{Z}_{TP}^{(2)} \geq \frac{\sqrt{{w_{TP}^{(1)}}^2 + {w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}} b_{TP}^{(2)} - \frac{w_{TP}^{(1)}}{w_{TP}^{(2)}} Z_{TP}^{(1)}\right\}\right. \\ &\qquad \left. \cap \left\{\tilde{Z}_{TC}^{(2)} \geq \frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)} - \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} Z_{TC}^{(1)}\right\} \Bigg| \, Z_{TP}^{(1)}, Z_{TC}^{(1)}, \mu_D \sim N\left(\mu_D^*, {\sigma_D^*}^2\right), D = T,C,P\right). \end{aligned} \tag{4.20}$$

In order to calculate the right-hand side of (4.20) we need to determine the joint posterior predictive distribution of the second-stage test statistics $\tilde{Z}_{TP}^{(2)}$ and $\tilde{Z}_{TC}^{(2)}$. Again, analogous to Section 4.1.3, in a frequentist framework the vector of the second-stage test statistics can be written as

$$\begin{pmatrix} \tilde{Z}_{TP}^{(2)} \\ \tilde{Z}_{TC}^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{\mu_T - \mu_P}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}} \\ \frac{\mu_T - \mu_C + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} \end{pmatrix} + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(2)}\right),$$

where $\mathbf{0} = (0,0)'$ and $\tilde{\boldsymbol{\Sigma}}^{(2)}$ is given in (4.16). Let us now assume that the treatment effects μ_T, μ_C and μ_P are normally distributed with means μ_T^*, μ_C^* and μ_P^* and variances ${\sigma_T^*}^2, {\sigma_C^*}^2$ and ${\sigma_P^*}^2$ according to (4.7), i.e. $\mu_D = \mu_D^* + \epsilon_D$ with $\epsilon_D \sim N(0, {\sigma_D^*}^2)$ and $D = T, C, P$. Then we easily obtain

$$\begin{pmatrix} \tilde{Z}_{TP}^{(2)} \\ \tilde{Z}_{TC}^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{\mu_T^* - \mu_P^*}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}} \\ \frac{\mu_T^* - \mu_C^* + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} \end{pmatrix} + \begin{pmatrix} \frac{\epsilon_T - \epsilon_P}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}} \\ \frac{\epsilon_T - \epsilon_C}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} \end{pmatrix} + \boldsymbol{\epsilon},$$

so that the posterior predictive distribution of the vector of second-stage test statistics is also a bivariate normal distribution with mean vector $\tilde{\boldsymbol{\mu}}^{(2)*}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}^{(2)*}$ determined as

$$\tilde{\boldsymbol{\mu}}^{(2)*} = \begin{pmatrix} \frac{\mu_T^* - \mu_P^*}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}} \\ \frac{\mu_T^* - \mu_C^* + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} \end{pmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}^{(2)*} = \begin{pmatrix} \tilde{\sigma}_{TP}^{(2)*} & \tilde{\rho}^{(2)*} \\ \tilde{\rho}^{(2)*} & \tilde{\sigma}_{TC}^{(2)*} \end{pmatrix}, \text{ where}$$

$$\tilde{\sigma}_{TP}^{(2)*} = 1 + \frac{{\sigma_T^*}^2 + {\sigma_P^*}^2}{\sigma^2} \frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}, \quad \tilde{\sigma}_{TC}^{(2)*} = 1 + \frac{{\sigma_T^*}^2 + {\sigma_C^*}^2}{\sigma^2} \frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}} \quad \text{and}$$

$$\tilde{\rho}^{(2)*} = \left(1 + \frac{{\sigma_T^*}^2}{\sigma^2} \tilde{n}_T^{(2)}\right) \sqrt{\frac{\tilde{n}_C^{(2)} \tilde{n}_P^{(2)}}{\left(\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}\right)\left(\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}\right)}}.$$

Consequently, by standardising the random vector $(\tilde{Z}_{TP}^{(2)}, \tilde{Z}_{TC}^{(2)})'$ the predictive power in (4.20) is obtained as

$$PP_{TP,TC} = \Phi^{\boldsymbol{\Sigma}} \left(\frac{\frac{\mu_T^* - \mu_P^*}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}} + \frac{w_{TP}^{(1)}}{w_{TP}^{(2)}} \frac{\bar{X}_T^{(1)} - \bar{X}_P^{(1)}}{\sigma} \sqrt{\frac{n_T^{(1)} n_P^{(1)}}{n_T^{(1)} + n_P^{(1)}}} - \frac{\sqrt{{w_{TP}^{(1)}}^2 + {w_{TP}^{(2)}}^2}}{w_{TP}^{(2)}} b_{TP}^{(2)}}{\sqrt{1 + \frac{{\sigma_T^*}^2 + {\sigma_P^*}^2}{\sigma^2} \frac{\tilde{n}_T^{(2)} \tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_P^{(2)}}}}, \right.$$
$$\left. \frac{\frac{\mu_T^* - \mu_C^* + \Delta_{ni}}{\sigma} \sqrt{\frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}} + \frac{w_{TC}^{(1)}}{w_{TC}^{(2)}} \frac{\bar{X}_T^{(1)} - \bar{X}_C^{(1)} + \Delta_{ni}}{\sigma} \sqrt{\frac{n_T^{(1)} n_C^{(1)}}{n_T^{(1)} + n_C^{(1)}}} - \frac{\sqrt{{w_{TC}^{(1)}}^2 + {w_{TC}^{(2)}}^2}}{w_{TC}^{(2)}} b_{TC}^{(2)}}{\sqrt{1 + \frac{{\sigma_T^*}^2 + {\sigma_C^*}^2}{\sigma^2} \frac{\tilde{n}_T^{(2)} \tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)}}}} \right), \tag{4.21}$$

where $\boldsymbol{\Sigma}$ is a 2×2 matrix with main diagonal elements equal to 1 and off-diagonal elements determined as

$$\rho = \frac{1 + \frac{\sigma_T^{*\,2}}{\sigma^2}\tilde{n}_T^{(2)}}{\sqrt{\left(1 + \frac{\sigma_T^{*\,2}+\sigma_P^{*\,2}}{\sigma^2}\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}\right)\left(1 + \frac{\sigma_T^{*\,2}+\sigma_C^{*\,2}}{\sigma^2}\frac{\tilde{n}_T^{(2)}\tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}}\right)}}\sqrt{\frac{\tilde{n}_C^{(2)}\tilde{n}_P^{(2)}}{\left(\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}\right)\left(\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}\right)}}. \tag{4.22}$$

Taking a closer look at the predictive power formula distinct similarities to the respective conditional power given in (4.17) become apparent. First of all, similar to the differences between $CP_{TC}(\theta_{TC})$ and PP_{TC}, the predictive power in (4.21) particularly differs from the corresponding conditional power by the two quantities in the denominators. Moreover, the correlation in (4.22) differs from that given in (4.16) by a certain factor which is formed by the common variance, the posterior variances and the second-stage sample sizes.

If we use non-informative priors for all treatment effects μ_T, μ_C and μ_P, i.e. $\sigma_{T,0}, \sigma_{C,0}, \sigma_{P,0} \to \infty$, it can be easily shown that the predictive power from (4.21) becomes

$$PP_{TP,TC} = \Phi^{\Sigma}\left(\frac{\frac{\bar{X}_T^{(1)}-\bar{X}_P^{(1)}}{\sigma}\left(\sqrt{\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}}+\frac{w_{TP}^{(1)}}{w_{TP}^{(2)}}\sqrt{\frac{n_T^{(1)}n_P^{(1)}}{n_T^{(1)}+n_P^{(1)}}}\right)-\frac{\sqrt{w_{TP}^{(1)\,2}+w_{TP}^{(2)\,2}}}{w_{TP}^{(2)}}b_{TP}^{(2)}}{\sqrt{1+\frac{n_T^{(1)}+n_P^{(1)}}{n_T^{(1)}n_P^{(1)}}\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}}},\right.$$
$$\left.\frac{\frac{\bar{X}_T^{(1)}-\bar{X}_C^{(1)}+\Delta_{ni}}{\sigma}\left(\sqrt{\frac{\tilde{n}_T^{(2)}\tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}}}+\frac{w_{TC}^{(1)}}{w_{TC}^{(2)}}\sqrt{\frac{n_T^{(1)}n_C^{(1)}}{n_T^{(1)}+n_C^{(1)}}}\right)-\frac{\sqrt{w_{TC}^{(1)\,2}+w_{TC}^{(2)\,2}}}{w_{TC}^{(2)}}b_{TC}^{(2)}}{\sqrt{1+\frac{n_T^{(1)}+n_C^{(1)}}{n_T^{(1)}n_C^{(1)}}\frac{\tilde{n}_T^{(2)}\tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}}}}\right),$$

where the respective correlation from (4.22) is then given as

$$\rho = \frac{1 + \frac{\tilde{n}_T^{(2)}}{n_T^{(1)}}}{\sqrt{\left(1 + \frac{n_T^{(1)}+n_P^{(1)}}{n_T^{(1)}n_P^{(1)}}\frac{\tilde{n}_T^{(2)}\tilde{n}_P^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}}\right)\left(1 + \frac{n_T^{(1)}+n_C^{(1)}}{n_T^{(1)}n_C^{(1)}}\frac{\tilde{n}_T^{(2)}\tilde{n}_C^{(2)}}{\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}}\right)}}\sqrt{\frac{\tilde{n}_C^{(2)}\tilde{n}_P^{(2)}}{\left(\tilde{n}_T^{(2)}+\tilde{n}_C^{(2)}\right)\left(\tilde{n}_T^{(2)}+\tilde{n}_P^{(2)}\right)}}.$$

If instead highly informative priors are used, i.e. $\sigma_{T,0}, \sigma_{C,0}, \sigma_{P,0} \to 0$, it follows from (4.21) and (4.22) that the predictive power converges to $CP_{TP,TC}(\mu_{T,0} - \mu_{P,0}, \mu_{T,0} - \mu_{C,0})$, which is the respective conditional power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ at the final analysis assuming the anticipated prior mean differences.

Finally, we will now investigated how the predictive power in (4.21) behaves for increasing second-stage sample sizes. For $\theta_{TP} > 0$ and $\theta_{TC} > -\Delta_{ni}$ the conditional power in (4.17) obviously approaches one for increasing $\tilde{n}_T^{(2)}, \tilde{n}_C^{(2)}$ and $\tilde{n}_P^{(2)}$, so that any targeted conditional power can be obtained for suitably large second-stage sizes. Unfortunately, as we will now see, this is not the case for the predictive power. Let us therefore assume that the second-stage sample sizes of the control and placebo group are defined as fractions of the second-stage test group size, namely $\tilde{n}_C^{(2)} = \tilde{c}_C^{(2)}\tilde{n}_T^{(2)}$ and $\tilde{n}_P^{(2)} = \tilde{c}_P^{(2)}\tilde{n}_T^{(2)}$ for some $\tilde{c}_C^{(2)}, \tilde{c}_P^{(2)} > 0$. Then the off-diagonal element in (4.22) can be written as

$$\rho = \frac{\frac{1}{\tilde{n}_T^{(2)}} + \frac{\sigma_T^{*\,2}}{\sigma^2}}{\sqrt{\left(\frac{1}{\tilde{n}_T^{(2)}} + \frac{\sigma_T^{*\,2}+\sigma_P^{*\,2}}{\sigma^2}\frac{\tilde{c}_P^{(2)}}{1+\tilde{c}_P^{(2)}}\right)\left(\frac{1}{\tilde{n}_T^{(2)}} + \frac{\sigma_T^{*\,2}+\sigma_C^{*\,2}}{\sigma^2}\frac{\tilde{c}_C^{(2)}}{1+\tilde{c}_C^{(2)}}\right)}}\sqrt{\frac{\tilde{c}_C^{(2)}\tilde{c}_P^{(2)}}{\left(1+\tilde{c}_C^{(2)}\right)\left(1+\tilde{c}_P^{(2)}\right)}}. \tag{4.23}$$

With the same considerations as at the end of Section 4.1.3 it follows for the predictive power in (4.21) that

$$PP_{TP,TC} \xrightarrow{\tilde{n}_T^{(2)},\tilde{n}_C^{(2)},\tilde{n}_P^{(2)}\to\infty} \Phi^{\boldsymbol{\Sigma}}\left(\frac{\mu_T^* - \mu_P^*}{\sqrt{\sigma_T^{*\,2}+\sigma_P^{*\,2}}}, \frac{\mu_T^* - \mu_C^* + \Delta_{ni}}{\sqrt{\sigma_T^{*\,2}+\sigma_C^{*\,2}}}\right),$$

where the off-diagonal elements of the corresponding matrix $\boldsymbol{\Sigma}$ are determined according to (4.23) and we have

$$\rho \xrightarrow{\tilde{n}_T^{(2)},\tilde{n}_C^{(2)},\tilde{n}_P^{(2)}\to\infty} \frac{\sigma_T^{*\,2}}{\sqrt{\left(\sigma_T^{*\,2}+\sigma_P^{*\,2}\right)\left(\sigma_T^{*\,2}+\sigma_C^{*\,2}\right)}}.$$

If non-informative priors are used, it obviously follows that the predictive power converges to $\Phi^{\boldsymbol{\Sigma}}(Z_{TP}^{(1)}, Z_{TC}^{(1)})$ with ρ becoming the correlation of the two first stage test statistics $Z_{TP}^{(1)}$ and $Z_{TC}^{(1)}$, namely $\rho = \sqrt{\frac{n_C^{(1)} n_P^{(1)}}{\left(n_T^{(1)}+n_C^{(1)}\right)\left(n_T^{(1)}+n_P^{(1)}\right)}}$.

With regard to sample size re-calculation this means that both observed interim test statistics need to be sufficiently large in order to have the ability of achieving the targeted predictive power. In other words, if only one of the first-stage test statistics is low, the maximum attainable predictive power will also be low. This is a very undesirable property, so that the predictive power in (4.21) seems to be even less suitable for sample size re-calculation than the predictive power given in (4.19). However, not least because of the large amount of information on the control and placebo treatment that is generally available in the beginning of a three-arm non-inferiority trial, the derived predictive power formulas are useful tools for exploring the interim data in such trials.

4.2.4 Hypothetical Example

Let us now illustrate the proposed adaptive testing procedure by means of reconsidering the hypothetical trial example of the previous chapter. In a three-arm trial a new therapy for the treatment of bronchial asthma should be compared with an already approved active control treatment and placebo. Treatment efficacy will be assessed by means of changes of the forced expiratory volume in one second (FEV$_1$) measured in litre (l). Besides a one-sided significance level of $\alpha = 0.025$, a non-inferiority margin of $\Delta_{ni} = 0.2l$ is adopted.

First of all, let us start by determining a group sequential design with $K = 2$ equally-sized

Table 4.1: Sample sizes, observed data and test results for a hypothetical trial where the proposed adaptive testing procedure is applied.

Stage	Sample sizes			Stage-wise data [in l]			Test results			
k	$\tilde{n}_T^{(k)}$	$\tilde{n}_C^{(k)}$	$\tilde{n}_P^{(k)}$	$\Delta\bar{X}_T^{(k)}$	$\Delta\bar{X}_C^{(k)}$	$\Delta\bar{X}_P^{(k)}$	$Z_{TP}^{(k)*}$	$b_{TP}^{(k)}$	$Z_{TC}^{(k)*}$	$b_{TC}^{(k)}$
1	204	204	68	2.409	2.489	2.021	2.771	2.797	1.212	2.797
2	369	369	38	2.415	2.463	1.913	4.043	1.977	2.317	1.977

stages as described in the previous chapter, were the overall power to demonstrate both superiority over placebo and non-inferiority to the control treatment shall be at least $1-\beta = 0.80$. As rejections at interim should only occur when the level of evidence is high, the rejection boundaries are chosen according to O'Brien Fleming for both comparisons, namely $(b_{TP}^{(1)}, b_{TP}^{(2)}) = (b_{TC}^{(1)}, b_{TC}^{(2)}) = (2.797, 1.977)$.

In previous placebo-controlled superiority trials treatment effects of $2.4l$ and $2l$ were observed for the control treatment and placebo, respectively. The test treatment is assumed to have the same effect as the active control, so that we will assume $\mu_T = \mu_C = 2.4l$ and $\mu_P = 2l$ for the sample size determination, i.e. $\theta_{TP} = 0.4l$ and $\theta_{TC} = 0l$. According to this, the between-group allocation ratio is chosen as $n_T^{(2)} : n_C^{(2)} : n_P^{(2)} = 3 : 3 : 1$, i.e. $c_C = 1$ and $c_P = \frac{1}{3}$, which is, for the assumed treatment effects, close to the respective optimal allocation. For simplicity reasons we further assume that the standard deviation of $\sigma = 1l$ is known. By choosing $(n_T^{(1)}, n_T^{(2)}) = (n_C^{(1)}, n_C^{(2)}) = (204, 408)$ and $(n_P^{(1)}, n_P^{(2)}) = (68, 136)$ the overall power of the group sequential design to show both superiority to placebo and non-inferiority to the active control is 80.2% according to (3.34). In order to account for uncertainties regarding the assumed treatment effects, especially that of the experimental treatment, the study team decides to extend the group sequential to an adaptive design as described in Section 4.2.1, offering the ability of mid-trial sample size re-assessment based on the observed interim data. Therefore, the weights are chosen as $w_{TP}^{(1)} = w_{TP}^{(2)} = w_{TC}^{(1)} = w_{TC}^{(2)} = 1$ so that the two independent stages are equally weighted. Note that for equal stage sizes this is obviously equivalent to choosing the weights as the square roots of the preplanned stage-wise information levels given as $\sqrt{\mathscr{I}_{TP}^{(1)}} = \sqrt{\mathscr{I}_{TP}^{(2)} - \mathscr{I}_{TP}^{(1)}} = 7.141$ and $\sqrt{\mathscr{I}_{TC}^{(1)}} = \sqrt{\mathscr{I}_{TC}^{(2)} - \mathscr{I}_{TC}^{(1)}} = 10.010$, respectively.

Immediately after the primary endpoints of the first 476 patients have been observed the interim analysis is conducted. The corresponding observed data and test results of the two comparisons are given in Table 4.1. Although the test statistic of the test versus placebo superiority comparison $Z_{TP}^{(1)*}$ is relatively large, the corresponding critical boundary $b_{TP}^{(1)}$ is not crossed, so that neither $H_{0,TP}^{(s)}$ nor $H_{0,TC}^{(n)}$ can be rejected at the interim analysis. Moreover, it becomes apparent that the observed means of the test and control group seem to be in line with the assumptions $\mu_T = 2.4l$ and $\mu_P = 2l$, whereas the sample mean of the control group is slightly higher than expected. However, this deviation might also be due to chance.

Let us now take a closer look at the trial data by determining the respective conditional powers when the trial proceeds as planned, i.e. without any design changes such as sample size re-calculation. As the interim test statistic of the test versus placebo superiority comparison is very large, the conditional power to reject $H_{0,TP}^{(s)}$ is expected to be very high. According to (4.14) the conditional power to reject only $H_{0,TP}^{(s)}$ assuming the treatment difference $\theta_{TP} = 0.4l$ from the planning stage is calculated as $CP_{TP}(0.4) = 99.8\%$. The conditional power at the observed treatment difference $\bar{X}_T^{(1)} - \bar{X}_P^{(1)} = 0.388l$ of $CP_{TP}(0.338) = 99.7\%$ is also very high. Thus, the observed data suggest that $H_{0,TP}^{(s)}$ will almost surely be rejected at the final analysis.

The conditional power to reject $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ at the treatment differences assumed in the planning stage is calculated by means of Equation (4.17) as $CP_{TP,TC}(0.4,0) = 66.8\%$. In order to obtain at least 80% conditional power this would require 217 additional patients overall. In contrast, re-calculating the sample sizes based on the conditional power at the observed mean differences, which is only 35.5% for the preplanned design, would result in more than a fourfold increase of the overall second-stage size, namely from 476 to 1911 patients. However, it should be kept in mind that the interim point estimators have a certain degree of uncertainty.

Because there clearly is a lot of prior information from previous trials on the control treatment and placebo effect, it seems natural to implement this prior belief when assessing the interim data. This can be easily done by defining prior distributions for the treatment effects and updating them in a Bayesian fashion with the first-stage observations. For μ_C and μ_P we choose very informative normal priors with means $\mu_{C,0} = 2.4, \mu_{P,0} = 2$ and variances $\sigma_{C,0}^2 = \sigma_{P,0}^2 = \frac{1}{500}$, whereas for the test treatment effect we choose a normal prior distribution which is a little less

Figure 4.1: Marginal prior belief regarding the treatment effects μ_T, μ_C and μ_P (left) and joint prior belief about the treatment differences $\mu_T - \mu_P$ and $\mu_T - \mu_C$ displayed as 1%, 10%, 50%, 90% and 99% of the maximum ordinate (right).

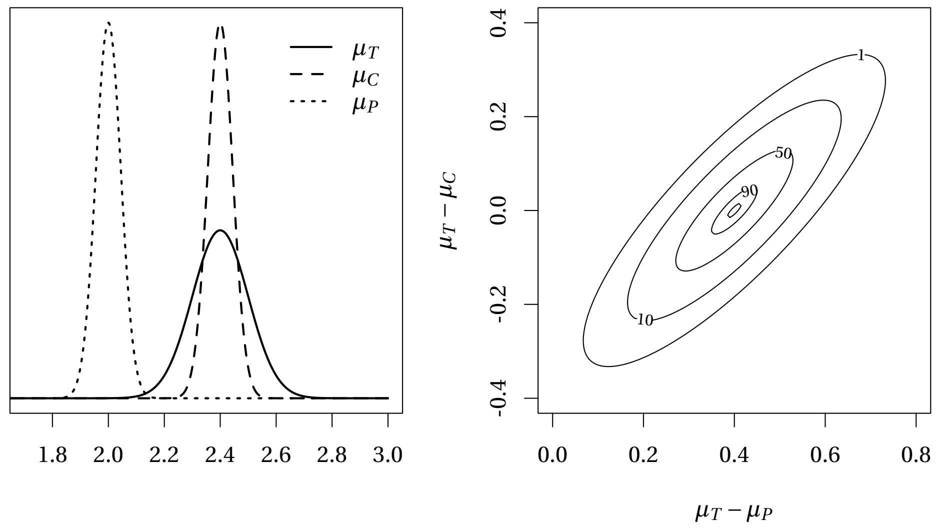


Table 4.2: Summary of the Bayesian updating process including the prior distributions, the observed interim data and the resulting posterior distributions for the treatment effects μ_T, μ_C and μ_P.

Treatment group	Prior hyperparameters		Observed interim data		Posterior hyperparameters	
D	$\mu_{D,0}$	$\sigma^2_{D,0}$	$\bar{X}^{(1)}_D$	$\sigma^2/n^{(1)}_D$	μ^*_D	$\sigma^{*\,2}_D$
T(est)	2.4	1/100	2.409	1/204	2.406	1/304
C(ontrol)	2.4	1/500	2.489	1/204	2.426	1/704
P(lacebo)	2.0	1/500	2.021	1/68	2.003	1/568

informative, namely $\mu_{T,0} = 2.4$ and $\sigma^2_{T,0} = \frac{1}{100}$. Furthermore, we make the simplifying assumption of independent prior distributions, so that the joint prior distribution of the treatment differences $\mu_T - \mu_P$ and $\mu_T - \mu_C$ can be easily determined. Figure 4.1 gives an overview on the prior marginal belief about the treatment effects and the joint prior belief concerning the treatment differences.

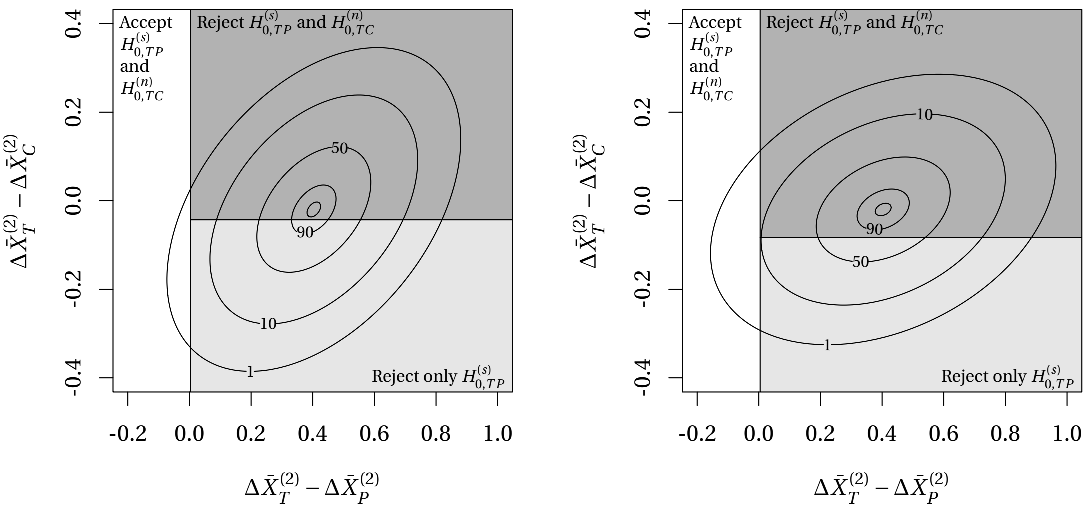
Note that for the control treatment and placebo effect the prior information is given a higher weight than the information gained from the first stage, because we have $\sigma^2_{C,0} < \frac{\sigma^2}{n^{(1)}_C}$ and $\sigma^2_{P,0} < \frac{\sigma^2}{n^{(1)}_P}$. In contrast, as we obviously have $\sigma^2_{T,0} > \frac{\sigma^2}{n^{(1)}_T}$, the first-stage data on the test treatment group has a higher weight than the respective prior information on μ_T.

As it has been mentioned earlier, the respective posterior distributions are also independent normal distributions whose hyperparameters can be calculated by simple mathematical operations according to (4.7). Table 4.2 gives an overview on the respective prior beliefs, the observed first-stage information and the resulting posterior beliefs on the treatment effects μ_T, μ_C and μ_P. As we can see, the posterior means of the test and placebo group are almost the same as those assumed in the planning stage, whereas the posterior mean of the control group is slightly higher than anticipated. By means of the derived posterior hyperparameters from Table 4.2 the predictive power to reject $H^{(s)}_{0,TP}$ and $H^{(n)}_{0,TC}$ with the preplanned second-stage sizes is determined according to (4.21) and (4.22) as $PP_{TP,TC} = 57.6\%$, which is fairly low.

In order to get a better overview on the probabilities of the potential study outcomes based on the currently available information let us now consider the joint posterior predictive distribution of the second-stage mean differences $\Delta\bar{X}^{(2)}_T - \Delta\bar{X}^{(2)}_P$ and $\Delta\bar{X}^{(2)}_T - \Delta\bar{X}^{(2)}_C$. This distribution can be determined analogously to the derivation of the joint posterior predictive distribution of the second-stage test statistics $\tilde{Z}^{(2)}_{TP}$ and $\tilde{Z}^{(2)}_{TC}$ which could also be used instead (cf. Section 4.2.3). However, to the opinion of the author, considering differences is preferable to considering test statistics because they are easier to comprehend, especially by non-statisticians.

The plot on the left-hand side of Figure 4.2, which is in the style of Figure 3 presented in Spiegelhalter et al. (1986), shows the joint posterior predictive distribution of $\Delta\bar{X}^{(2)}_T - \Delta\bar{X}^{(2)}_P$ and

Figure 4.2: Acceptance and rejection regions of the final analysis and corresponding joint posterior predictive distribution of the future mean differences $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ and $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_C^{(2)}$ for the preplanned second-stage sizes $\tilde{n}_T^{(2)} = \tilde{n}_C^{(2)} = 204$ and $\tilde{n}_P^{(2)} = 68$ (left) and the updated second-stage sizes $\tilde{n}_T^{(2)} = \tilde{n}_C^{(2)} = 369$ and $\tilde{n}_P^{(2)} = 38$ (right). Contours shown are 1%, 10%, 50%, 90% and 99% of the maximum ordinate.



$\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_C^{(2)}$ superimposed on the corresponding acceptance and rejection regions at the final analysis when the trial proceeds with the preplanned second-stage sizes. Through this, the figure gives a graphical representation of the predictive probabilities such as the predictive power to reject both null hypotheses, which is obtained by integrating the joint posterior predictive distribution of the second-stage sample mean differences over the dark grey shaded area.

Let us now take a closer look at the left plot in Figure 4.2. With the currently available information it seems very unlikely that none of the null hypotheses is rejected at the final stage because the bulk of plausible future values for $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ lies to the right of 0.004, which is the smallest value of $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ that would result in a rejection of $H^{(s)}_{0,TP}$. Due to the fact that the test statistic $Z^{(1)}_{TP}$ is already very high, this is hardly surprising. The corresponding predictive probability that both null hypotheses will be accepted in the final analysis is as low as 0.5%. Moreover, rejecting both null hypotheses in the end seems to be more plausible than rejecting only $H^{(s)}_{0,TP}$ with corresponding predictive probabilities of 57.6% (= $PP_{TP,TC}$) and 41.8%, respectively.

Based on the observations at the interim analysis the study team decides to continue the trial with an increased second-stage size. In order to account for potentially existing deviations from the treatment effects assumed in the planning stage, the sample sizes of the second stage will be re-calculated based on the conditional power in (4.17) at the posterior mean differences. That means we want to find the smallest sample sizes $\tilde{n}_T^{(2)}, \tilde{n}_C^{(2)}$ and $\tilde{n}_P^{(2)}$ so that $CP_{TP,TC}(\mu_T^* - \mu_P^*, \mu_T^* - \mu_C^*) = CP_{TP,TC}(0.404, -0.02) \geq 0.80$ holds. Note that the predictive power $PP_{TP,TC}$ is not used for re-calculation as this generally results in too large sample sizes (see also Sections 4.1.3 and 4.2.3). Without changing the between-group allocation ratio, i.e. $\tilde{c}_C^{(2)} = 1$ and $\tilde{c}_P^{(2)} = \frac{1}{3}$, the required second-stage sizes to achieve a conditional power of at least 80% to reject both null hypotheses are obtained as $\tilde{n}_T^{(2)} = \tilde{n}_C^{(2)} = 363$ and $\tilde{n}_P^{(2)} = 121$. That means we would have to recruit 371 additional patients which is a second-stage size increase of about 78%. Based on the interim observations it seems natural to put more emphasis on the test versus control non-inferiority comparison as $H_{0,TP}^{(s)}$ will almost surely be rejected. Moreover, the allocation ratio 3:3:1 might no longer be optimal with respect to minimising the overall sample size.

Therefore, by analogy with the optimisations in the previous chapters, among all second-stage sizes $\tilde{n}_T^{(2)}, \tilde{n}_C^{(2)}$ and $\tilde{n}_P^{(2)}$ that result in a conditional power of at least 80%, we search for those with the smallest overall second-stage size $\tilde{n}_T^{(2)} + \tilde{n}_C^{(2)} + \tilde{n}_P^{(2)}$. Practically speaking, this is done as follows: Define $\tilde{n}_C^{(2)} = \tilde{c}_C^{(2)} \tilde{n}_T^{(2)}$ and $\tilde{n}_P^{(2)} = \tilde{c}_P^{(2)} \tilde{n}_T^{(2)}$ for some $\tilde{c}_C^{(2)}, \tilde{c}_P^{(2)} > 0$, so that for each pair $(\tilde{c}_C^{(2)}, \tilde{c}_P^{(2)})$ their is a unique (real number) solution $\tilde{n}_T^{(2)}$ for $CP_{TP,TC}(0.404, -0.02) = 0.80$ with corresponding overall second-stage size $(1 + \tilde{c}_C^{(2)} + \tilde{c}_P^{(2)})\tilde{n}_T^{(2)}$. Then, with an appropriate optimisation routine such as the downhill simplex algorithm by Nelder and Mead (1965) we search for the allocation ratios that minimise the overall second-stage size.

The optimal second-stage sizes are obtained as $\tilde{n}_T^{(2)} = \tilde{n}_C^{(2)} = 363$ and $\tilde{n}_P^{(2)} = 38$ resulting in a conditional power of $CP_{TP,TC}(0.404, -0.02) = 80\%$. Besides saving 71 patients overall compared to re-calculation without changing the between-group allocation ratio, the optimal design furthermore has the desirable feature of considerably reducing the number of patients allocated to the placebo group. Along with an ethical advantage this also reflects the fact that the proof of efficacy for the test treatment could nearly be demonstrated at the interim analysis. It should also be noted that, interestingly, the re-calculated optimal second-stage placebo group size is almost half of the respective preplanned placebo group size.

The influence of the re-calculation on the probability of potential study outcomes is furthermore illustrated in the right plot of Figure 4.2. This plot shows the rejection regions of the final stage together with the joint posterior predictive distribution of $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ and $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_C^{(2)}$ for the re-calculated sample sizes $\tilde{n}_T^{(2)} = \tilde{n}_C^{(2)} = 363$ and $\tilde{n}_P^{(2)} = 38$. First of all, it becomes apparent that the rejection boundary for $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_C^{(2)}$ is slightly decreased compared with the preplanned design, while the respective critical value for $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ is almost unaffected by the re-calculation. The shift from the superiority comparison between test and placebo towards the test versus control non-inferiority comparison is reflected by the fact that the variation of the future sample mean difference $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ is increased, whereas $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_C^{(2)}$ now has a smaller variance than in the preplanned design. Through this, the corresponding predictive power to reject both null hypotheses in the end is considerably increased from 57.5% to 73%. Moreover, we now have a 25.5% predictive probability to reject only $H_{0,TP}^{(s)}$, while the predictive probability that no hypothesis will be rejected is slightly increased to 1.5%, which is, however, still very low.

Shortly after the trial has proceeded with the derived optimal second-stage sizes $\tilde{n}_T^{(2)} = \tilde{n}_C^{(2)} =$

363 and $\tilde{n}_P^{(2)} = 38$, the final analysis is conducted. The second-stage sample means of the three treatment groups and the corresponding test results can be found in Table 4.1. As both test statistics obviously exceed the critical value $b_{TP}^{(2)} = b_{TC}^{(2)} = 1.977$, the trial ends with rejection of both null hypotheses. That means, in addition to its proof of efficacy, the test treatment is furthermore demonstrated to be non-inferior to the active control. Note that the results from the second stage also seem to confirm the impression at the interim analysis of the control treatment effect being slightly higher than assumed. Furthermore it should be noted that, if the same sample means would have been observed with the preplanned second-stage sizes, $H_{0,TC}^{(n)}$ could have not been rejected.

4.2.5 An Optimal Adaptive Type Design without Early Rejection

As we have seen in the hypothetical example above, re-calculation of the between-group sample size allocation based on the observed interim data led to substantial sample size savings, especially in the placebo group. Let us therefore further investigate the option of optimising the between-group allocation at the interim analysis by using the conditional power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$.

In order to reduce the number of influence parameters, such as the group sequential rejection boundaries, we will restrict to adaptive designs without early rejection at the interim analysis. Such designs are easily obtained within the proposed framework by simply setting $b_{TP}^{(1)} = b_{TC}^{(1)} = \infty$ which results in $b_{TP}^{(2)} = b_{TC}^{(2)} = z_{1-\alpha}$, i.e. the critical values of the fixed sample size design from Section 2.3. This choice might be of particular interest when prematurely closing the placebo arm is viewed critically, because there are concerns about a change in patient population after dropping the placebo group. Note that this clearly is one of the major points of criticism regarding the proposed group sequential and adaptive designs.

At the interim analysis, analogous to the optimisation in the hypothetical example, the conditional power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ from (4.17) at the preplanned treatment differences will be used to obtain the optimal second-stage sizes $\tilde{n}_T^{(2)}, \tilde{n}_C^{(2)}$ and $\tilde{n}_P^{(2)}$ that minimise the overall second-stage size. The respective conditional power of the fixed sample size design given the observed sample means $\bar{X}_T^{(1)}, \bar{X}_C^{(1)}$ and $\bar{X}_P^{(1)}$ will be used as the targeted conditional power in the optimisation. Through this, the adaptive type design will obviously have the same overall power as the fixed sample size design, which allows a fair comparison between the two designs.

For simplicity reasons, let us assume that the sample size re-calculation will be conducted when the primary endpoint has been observed for half of the patients, so that we have $n_D^{(1)} = \tilde{n}_D^{(2)}$ for $D = T, C, P$. By setting $w_{TP}^{(1)} = w_{TP}^{(2)} = w_{TC}^{(1)} = w_{TC}^{(2)} = 1$ it can be easily shown that the inverse normal test statistics of the final analysis coincide with the respective cumulative test statistics of the group sequential design, i.e. we have $Z_{TP}^{(2)*} = Z_{TP}^{(2)}$ and $Z_{TC}^{(2)*} = Z_{TC}^{(2)}$. Consequently, the derived conditional and predictive power formulas can also be used for group sequential and even fixed sample size designs. Moreover, one can use the required sample sizes of the fixed design as a starting point for the adaptive type design.

Suppose we have a fixed design with corresponding sample sizes $n_{T,fix}, n_{C,fix}$ and $n_{P,fix}$. Then, after half of the patients have been observed with sample means $\bar{X}_T^{(1)}, \bar{X}_C^{(1)}$ and $\bar{X}_P^{(1)}$, according to (4.17) the corresponding conditional power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ in the end is obtained as

$$\begin{aligned} CP_{TP,TC}^{fix}(\theta_{TP}, \theta_{TC}) = \Phi^{\boldsymbol{\Sigma}_{fix}}\Bigg(&\frac{\theta_{TP} + \bar{X}_T^{(1)} - \bar{X}_P^{(1)}}{\sqrt{2}\sigma}\sqrt{\frac{n_{T,fix} n_{P,fix}}{n_{T,fix} + n_{P,fix}}} - \sqrt{2} z_{1-\alpha}, \\ &\frac{\theta_{TC} + \bar{X}_T^{(1)} - \bar{X}_C^{(1)} + 2\Delta_{ni}}{\sqrt{2}\sigma}\sqrt{\frac{n_{T,fix} n_{C,fix}}{n_{T,fix} + n_{C,fix}}} - \sqrt{2} z_{1-\alpha}\Bigg), \end{aligned} \tag{4.24}$$

where the corresponding covariance matrix is easily determined according to (4.16) as

$$\boldsymbol{\Sigma}_{fix} = \begin{pmatrix} 1 & \rho_{fix} \\ \rho_{fix} & 1 \end{pmatrix} \text{ with } \rho_{fix} = \sqrt{\frac{n_{C,fix} n_{P,fix}}{\left(n_{T,fix} + n_{C,fix}\right)\left(n_{T,fix} + n_{P,fix}\right)}}. \tag{4.25}$$

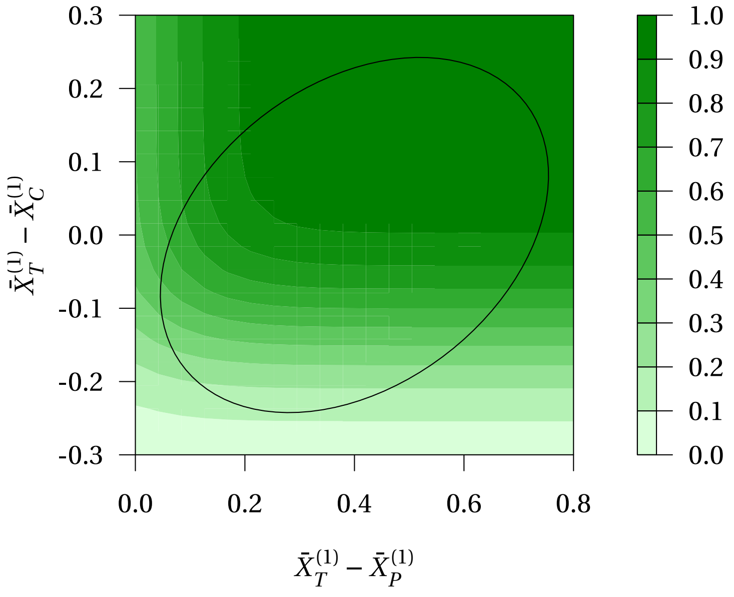
As mentioned earlier, this conditional power will be used as the targeted conditional power for calculating the optimal second-stage sizes. By means of applying the adaptive type design to the previously considered hypothetical trial example we will now investigate its properties in comparison to the optimal fixed design.

Just as a quick reminder, the assumed treatment differences from the planning stage and the common known standard deviation were $\theta_{TP} = 0.4, \theta_{TC} = 0$ and $\sigma = 1$, respectively. The non-inferiority margin was chosen as $\Delta_{ni} = 0.2$, the one-sided significance level was $\alpha = 2.5\%$ and the targeted overall power was set to $1 - \beta = 80\%$. First of all, let us derive the respective optimal fixed design that should serve both as a starting point and a reference for the adaptive type design. The corresponding optimal sample sizes are determined as $n_{T,fix} = 414, n_{C,fix} = 402$ and $n_{P,fix} = 124$ resulting in an overall power of 80.1% to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ assuming the treatment differences $\theta_{TP} = 0.4$ and $\theta_{TC} = 0$.

Before examining the adaptive type design let us first take a closer look at the respective conditional power $CP_{TP,TC}^{fix}(0.4, 0)$ calculated by means of (4.24) and (4.25), that is used as the targeted conditional power when re-calculating the sample sizes at interim. Figure 4.3 illustrates the dependency of $CP_{TP,TC}^{fix}(0.4, 0)$ on the mid-trial sample mean differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$. Moreover, under the assumption that the true treatment differences are $\theta_{TP} = 0.4$ and $\theta_{TC} = 0$, it gives the respective 95% prediction ellipsoid for $(\bar{X}_T^{(1)} - \bar{X}_P^{(1)}, \bar{X}_T^{(1)} - \bar{X}_C^{(1)})'$, which is the two-dimensional extension of the prediction interval.

It can be seen that the conditional power is already very high when the interim observations are as anticipated, namely for $\bar{X}_T^{(1)} - \bar{X}_P^{(1)} = 0.4$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)} = 0$ the conditional power is almost 90%. Furthermore, it becomes apparent that, for plausible sample mean differences, the conditional power $CP_{TP,TC}^{fix}(0.4, 0)$ is mainly influenced by the treatment difference between the test and control group $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$. This is primarily caused by the fact that the power for

Figure 4.3: Conditional power of the optimal fixed design to reject $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ at $\theta_{TP} = 0.4$ and $\theta_{TC} = 0$ depending on the observed treatment differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ of the first half of patients. The ellipsoid represents the respective 95% prediction area for the vector of observed mean differences $(\bar{X}_T^{(1)} - \bar{X}_P^{(1)}, \bar{X}_T^{(1)} - \bar{X}_C^{(1)})'$.

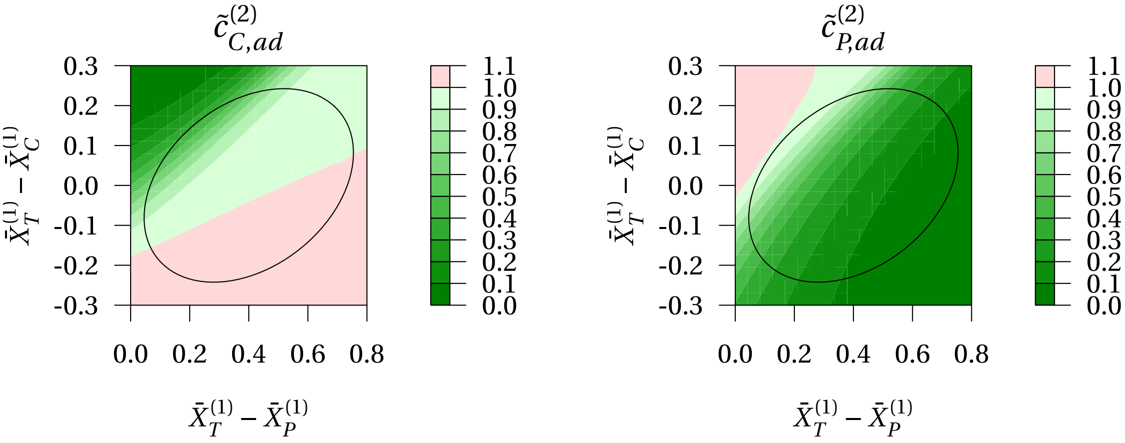


rejecting $H_{0,TP}^{(s)}$ is already very high under the optimal allocation. As a result there is almost no difference in conditional power whether the observed difference $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ is 0.2 or larger. Moreover, if $\bar{X}_T^{(1)} - \bar{X}_C^{(1)} \geq 0$ is observed, the conditional power will most surely be higher than 50% as this only requires a slightly positive sample mean difference between test and placebo, which obviously is more than likely under the assumption $\theta_{TP} = 0.4$. If we have $\bar{X}_T^{(1)} - \bar{X}_C^{(1)} \geq -0.1$, the conditional power will be at least 50% for any observed difference $\bar{X}_T^{(1)} - \bar{X}_P^{(1)} \geq 0.1$.

Let us now take a closer look at the behaviour of the optimal adaptive type design by examining how the optimal second-stage allocation depends on the observed sample mean differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$. Let therefore $\tilde{n}_{T,ad}^{(2)}, \tilde{n}_{C,ad}^{(2)}$ and $\tilde{n}_{P,ad}^{(2)}$ denote the re-calculated second-stage sample sizes of the adaptive type design for the test, control and placebo group, respectively. Furthermore, assume that the second-stage sizes of the control and placebo group are defined as fractions of the respective test group size, i.e. $\tilde{n}_{C,ad}^{(2)} = \tilde{c}_{C,ad}^{(2)} \tilde{n}_{T,ad}^{(2)}$ and $\tilde{n}_{P,ad}^{(2)} = \tilde{c}_{P,ad}^{(2)} \tilde{n}_{T,ad}^{(2)}$ for some parameters $\tilde{c}_{C,ad}^{(2)}, \tilde{c}_{P,ad}^{(2)} > 0$. Figure 4.4 shows the second-stage allocation ratios $\tilde{c}_{C,ad}^{(2)}$ and $\tilde{c}_{P,ad}^{(2)}$ of the optimal adaptive type design for different combinations of observed treatment differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$. Again, the 95% prediction ellipsoid for the vector of mean differences is superimposed in order to get an insight on plausible interim values when the true treatment differences are indeed $\theta_{TP} = 0.4$ and $\theta_{TC} = 0$.

It becomes apparent that the optimal second-stage allocations are mainly influenced by the

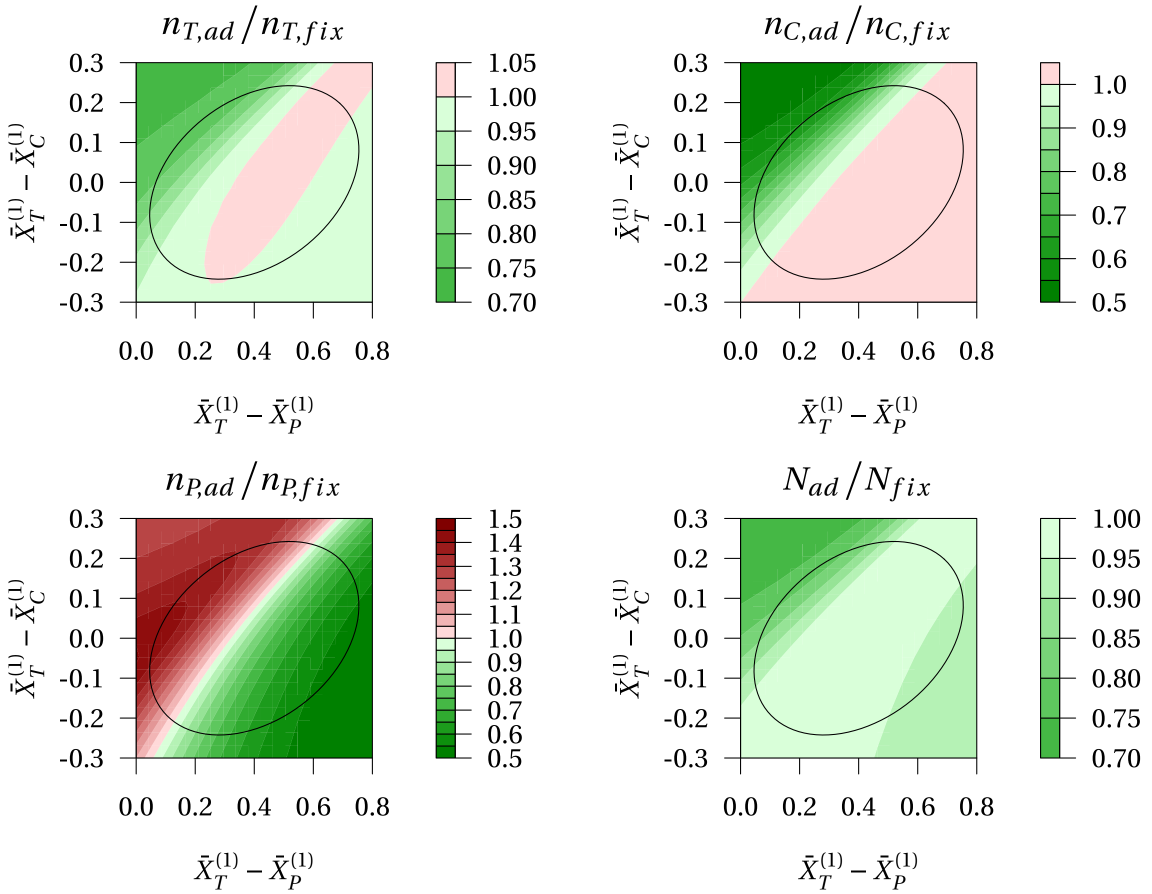
Figure 4.4: Second-stage allocation ratios of the adaptive type design depending on the observed treatment differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ of the first half of patients. The ellipsoid represents the respective 95% prediction area for the vector of observed mean differences $(\bar{X}_T^{(1)} - \bar{X}_P^{(1)}, \bar{X}_T^{(1)} - \bar{X}_C^{(1)})'$.



observed sample means of the control and placebo group $\bar{X}_C^{(1)}$ and $\bar{X}_P^{(1)}$, respectively, whereas $\bar{X}_T^{(1)}$ alone has only a small effect on $\tilde{c}^{(2)}_{C,ad}$ and $\tilde{c}^{(2)}_{P,ad}$. In particular, the lines of the right contour plot in Figure 4.4 are almost parallel to the line through the origin with slope one, illustrating the negligible influence of $\bar{X}_T^{(1)}$ alone on $\tilde{c}^{(2)}_{P,ad}$. For increasing interim placebo effect $\bar{X}_P^{(1)}$ the allocation ratio of the placebo group $\tilde{c}^{(2)}_{P,ad}$ also increases, while at the same time $\tilde{c}^{(2)}_{C,ad}$ decreases. If instead the observed sample mean of the placebo group is small, the second-stage size of the placebo group will also be small and the respective control group size will be large, each relative to the second-stage test group size. For the interim estimator of the control treatment effect $\bar{X}_C^{(1)}$ the dependencies are simply the other way around. That means for increasing $\bar{X}_C^{(1)}$ we have smaller $\tilde{c}^{(2)}_{P,ad}$ and larger $\tilde{c}^{(2)}_{C,ad}$, whereas we have larger $\tilde{c}^{(2)}_{P,ad}$ and smaller $\tilde{c}^{(2)}_{C,ad}$ when $\bar{X}_C^{(1)}$ is decreasing. Consequently, the smallest $\tilde{c}^{(2)}_{P,ad}$ are observed for large $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and small $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$, while, in contrast, for small $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and large $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ we will have a small control group. Practically speaking this means that the optimal adaptive type design decreases the number of patients allocated to the control (placebo) when the performance of the control (placebo) at the interim analysis compared with the other two treatments is poorer than expected. It should be noted that such a behaviour is also desirable from an ethical viewpoint.

Let us now investigate the performance of the adaptive type design in comparison to the optimal single-stage design by considering the quotients of the adaptive divided by the corresponding optimal single-stage sample sizes. Figure 4.5 shows these quotients for different combinations of interim sample mean differences together with the 95% prediction ellipsoid for the vector of observed mean differences. First of all, a quite similar pattern as in Figure 4.4 can be observed, i.e. the sample sizes of the adaptive type design mainly depend on $\bar{X}_C^{(1)}$ and

Figure 4.5: Sample sizes of the adaptive type design divided by those of the optimal single-stage design depending on the observed treatment differences $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ of the first half of patients. The ellipsoid represents the respective 95% prediction area for the vector of observed mean differences $(\bar{X}_T^{(1)} - \bar{X}_P^{(1)}, \bar{X}_T^{(1)} - \bar{X}_C^{(1)})'$.



$\bar{X}_P^{(1)}$, while $\bar{X}_T^{(1)}$ alone has only a minor influence. Moreover, for plausible values of $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ the overall sample size reductions compared with the optimal fixed design are only moderate. Only when $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ is small and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$ is large, an overall sample size reduction of up to 30% is possible. In these situations also the test and the control group size can be reduced to up to 70% and 50% of the respective optimal single-stage sample size, respectively. Among the three treatment groups the sample size of the placebo group reacts most sensitive to different values of $\bar{X}_T^{(1)} - \bar{X}_P^{(1)}$ and $\bar{X}_T^{(1)} - \bar{X}_C^{(1)}$. For plausible interim sample mean differences the placebo group size is reduced or increased by up to 50%, which means that the second-stage placebo group size is either completely reduced to zero or doubled. Increasing the sample size of the placebo group might not be desirable in many cases, however, in these situations we benefit from greater reductions of the overall sample size.

Last but not least we are interested in the expected performance of the proposed adaptive type design, or in other words: What is the expected additional benefit of the second optimisation of the between-group sample size allocation? Let us therefore assume that the derived single-stage design is indeed the true optimal design, i.e. the assumed treatment differences $\theta_{TP} = 0.4$ and $\theta_{TC} = 0$ are the true treatment differences. By means of simulation with 1 million iterations we determined the expected sample sizes of the adaptive type design for the test, control and placebo group as 406.1, 394.2 and 111.1, so that the expected overall sample size is given as 911.4. Compared with the sample sizes of optimal fixed design this means we have an average reduction of the test, control and placebo group of around 2%, 2% and 10%, while the overall sample size is reduced on average by about 3%. In absolute terms this is an overall expected reduction of about 29 patients, with an average decrease of almost 8, 8 and 13 patients in the test, control and placebo group, respectively.

4.3 Summary

The group sequential testing procedure proposed in Chapter 3 turned out to improve the optimal single-stage design from Chapter 2 by means of minimising the required overall sample size and, in particular, the number of patients allocated to placebo. However, one major point of criticism of the group sequential design is the potential change in patient population once the placebo arm is dropped. Moreover, often uncertainty exists in the planning stage of three-arm trials regarding the true treatment effects, so that the determined sample sizes might be either too small or too large. As a natural extension of the group sequential testing procedure, the adaptive design proposed in this chapter provides a simple way of dealing with the two mentioned issues. Through the ability of data-dependent sample size changes at the interim analysis, the second-stage placebo group size could be reduced to a certain threshold once $H_{0,TP}^{(s)}$ has been rejected, instead of completely closing the placebo arm. In doing so, potential heterogeneities across the stages due to a change in patient population might be decreased and can be assessed at the final analysis by comparing the independent results from the two stages. Re-calculating the sample sizes of the second stage based on the observed interim data could also help to overcome uncertainties concerning the treatment effects.

Due to its similarity to the proposed group sequential testing procedure, all findings from the previous chapter can be carried over to the adaptive testing procedure, such as, for example, formulas for the overall power. Thus, an obvious approach is to start with determining an appropriate group sequential design, which serves as a basis for the adaptive design. The derived formulas for the conditional power can then be used at the interim analysis to decide on whether to stop the trial for futility or to continue with preplanned or re-calculated sample sizes. As there often is a large amount of information regarding the control and placebo treatment in three-arm non-inferiority trials, we also determined formulas for the corresponding Bayesian predictive power. In particular, it turned out that an illustration such as Figure 4.2

with the rejection regions of the final stage and the respective joint posterior predictive distribution of the second-stage differences provides a comprehensive overview on the interim knowledge about the further course of the study. Together with the corresponding predictive probabilities such a figure can aid the decision-making process during trial conduct.

Note that the above-mentioned figure and the derived formulas for the conditional and the predictive power can also be used for the group sequential testing procedure derived in the previous chapter. Therefore, simply replace the weights $w_{TP}^{(1)}, w_{TP}^{(2)}$ and $w_{TC}^{(1)}, w_{TC}^{(2)}$ in formulas (4.15), (4.17), (4.19) and (4.20) by the square roots of the corresponding stage-wise information levels $\sqrt{\mathscr{I}_{TP}^{(1)}}, \sqrt{\mathscr{I}_{TP}^{(2)} - \mathscr{I}_{TP}^{(1)}}$ and $\sqrt{\mathscr{I}_{TC}^{(1)}}, \sqrt{\mathscr{I}_{TC}^{(2)} - \mathscr{I}_{TC}^{(1)}}$, respectively. In the same way these calculations can also be used to monitor three-arm non-inferiority trials with fixed sample sizes.

At the end of this chapter we furthermore investigated a special type of adaptive design without early rejection, where the interim analysis is only used to optimise the sample size allocation between the three treatment groups in order to further minimise the overall sample size. The proposed adaptive type design has the same overall power as the optimal single-stage design and at the same time always results in a smaller overall sample size. The average sample size reductions compared with the fixed design turned out to be relatively low, although in certain situations the overall sample size can be reduced by up to 30%. Moreover, when the observed interim sample mean difference between test and placebo is larger than expected, the placebo group size is considerably reduced to almost 50% of the optimal single-stage placebo group size. Interestingly, the resulting allocation rule is also reasonable from an ethical viewpoint, because it reduces the sample size of the control or placebo group when it apparently performs poorer than expected.

Finally, let us give a short comment on the assumption of a known common variance σ^2. Usually the sample sizes in three-arm non-inferiority trials are relatively large, so that the proposed adaptive design as well as the group sequential testing procedure will also be (at least approximately) valid for the unknown variance setting. However, exact type I error control is often required in confirmatory clinical trials. An exact solution for the unknown variance setting that can be easily applied to the proposed adaptive testing procedure was proposed by Lehmacher and Wassmer (1999), who considered the final test statistic

$$Z_2^* = \frac{w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)}{\sqrt{w_1^2 + w_2^2}},$$

where p_1 and p_2 are the independent p-values of the first and the second-stage, respectively, and $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function of the $N(0,1)$-distribution. Note that this also provides another solution for group sequential testing with an unknown variance. When assessing the conditional power of such a design at the interim analysis, a sensible substitute for σ^2 must be found, introducing further uncertainty. With respect to Bayesian calculations it can be shown that the respective conjugate prior for a normal distribution with unknown variance is the normal inverse-gamma distribution.

Discussion and Outlook

In this thesis we dealt with the design and the analysis of three-arm 'gold standard' non-inferiority trials for the case of a normally distributed endpoint. Although the superiority comparison between the active comparator and placebo might be of interest in some situations, we restricted to the two main objectives in such trials, namely the direct proof of efficacy for the experimental treatment ($H_{0,TP}^{(s)}$) and the non-inferiority comparison between the test and the control treatment ($H_{0,TC}^{(n)}$).

We started by investigating the hierarchical fixed sample size design proposed by Koch and Röhmel (2004), beginning with the test versus placebo superiority comparison and proceeding to the non-inferiority test once $H_{0,TP}^{(s)}$ has been rejected. In addition to exact and approximate formulas for the overall power to reject both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$, we determined optimal sample size allocations that minimise the overall sample size. Interestingly, it turned out that the placebo group size is also considerably reduced under the optimal allocation, which clearly is desirable from an ethical point of view as this makes the trial more acceptable for patients. The fact that the power of the superiority comparison between the experimental treatment and placebo is very high under the optimal allocation, brought us to the idea of utilising this high power by means of implementing a group sequential design.

We then proposed a general group sequential three-arm non-inferiority design that allows dropping the placebo arm once the efficacy of the test treatment has been demonstrated. We derived corresponding formulas for the overall power that can be used for sample size planning. Moreover, formulas for the expected placebo group size and the expected overall sample size were determined. It turned out that the implementation of group sequential methodology in three-arm non-inferiority trials can lead to considerable sample size savings, especially in the placebo group. This makes the trial even more acceptable from an ethical point of view. Besides the full group sequential designs where both hypotheses are tested group sequentially, we also investigated partial group sequential designs that test only $H_{0,TP}^{(s)}$ in a group sequential manner. Such designs lead to the largest reductions in the expected placebo group size and might be of interest in situations when an early study termination with rejection of both $H_{0,TP}^{(s)}$ and $H_{0,TC}^{(n)}$ is not desirable, e.g. in some situations it might be required to collect more safety data on the experimental treatment. Further investigations showed that it is reasonable to adapt the optimal single-stage allocations also in the group sequential setting. Moreover, the derivation of approximately optimal group sequential boundaries showed that it seems advisable to choose more aggressive rejection boundaries for the proof of efficacy than for the non-inferiority comparison as this leads to designs with favourable properties. With respect to practical application the discussed error spending designs are preferable to the classical group sequential tests because they turned out to have comparable performances and provide the additional flexibility to adequately handle deviations from the preplanned stage-wise sample size allocation.

As it has been mentioned earlier, when applying the proposed group sequential testing procedure it is vital to restrict the knowledge about dropping the placebo arm. Otherwise the patient population might change and as a result the whole study might be called into question. The proposed adaptive testing procedure, which can be seen as a natural extension of the group sequential designs, offers another possibility of dealing with this issue. Instead of completely closing the placebo arm, it could be reduced to a certain threshold, giving us also a better chance to detect potential heterogeneities across the different stages. By means of sample size re-calculations at an interim analysis, the adaptive designs also enable us to account for uncertainties concerning the treatment effects, which often exist in three-arm non-inferiority trials. Besides sample size re-assessment the derived conditional power formulas are also suitable for decision-making regarding stopping for futility.

The determined Bayesian predictive power formulas represent another monitoring tool that is useful especially in three-arm non-inferiority trials, because in the planning stage of such trials we already have a large amount of information about the control treatment and the placebo effect. However, similar to Dallow and Fina (2011) our investigations showed that using the predictive power for sample size re-calculation is not advisable as this generally leads to too large sample sizes. Nevertheless, a certain graphical representation turned out to provide a comprehensive overview on the interim knowledge about the further course of the study, namely the contour plot of the joint posterior predictive distribution of the second-stage sample mean differences $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_P^{(2)}$ and $\Delta\bar{X}_T^{(2)} - \Delta\bar{X}_C^{(2)}$ superimposed on the respective rejection regions of the final analysis. Note that the proposed conditional and predictive power formulas as well as the above-mentioned figure can also be used for monitoring trials with a group sequential or even a fixed sample size design.

Finally, we proposed an optimal adaptive design without early rejection where the interim analysis is used solely for updating the between-group allocation ratio in a way that the overall sample size is minimised. With the same overall power as the optimal fixed design, the proposed adaptive design always results in smaller overall sample sizes. Although the expected sample size savings are only marginal, in certain situations the sample sizes, especially that of the placebo group, can be reduced considerably. Moreover, interestingly, the optimal adaptive design behaves in an ethical manner, as it reduces the sample size of the control or the placebo group when the interim observations suggest that the respective treatment performs poorer than expected.

As we have seen, there are a number of benefits associated with the application of group sequential and adaptive methodology in three-arm non-inferiority trials. In particular, they can help to minimise the number of patients treated with placebo and to better exploit the resources. Besides an ethical advantage, this could also reduce patient concerns to participate in the trial and as a result enhance the recruitment process of the study. Moreover, the determined conditional and predictive power formulas are useful monitoring tools that can aid the decision-making process during trial conduct.

However, one should always bear in mind that the implementation of interim analyses is associated with additional operational efforts. In order to preserve the integrity and validity of the trial, it is vital to set up an Independent Data Monitoring Committee (IDMC). Furthermore, the amount of information revealed by the IDMC during trial conduct should be reduced to a minimum, so that the blinding is guaranteed. Allowing sample size changes at the interim analysis will introduce additional efforts, especially when the allocation between the treatment groups is altered. A centralised online randomisation could be a simple and effective way of dealing with this issue. Irrespective of whether the placebo arm was dropped or the placebo group size was reduced at the interim analysis, it is generally advisable to assess potential differences between the independent stages (CHMP, 2007b). As it has been mentioned earlier, the error spending designs can be used for dealing with the problem of overrunning patients. Here, it is important that the "results including and excluding the overrunning patients should be presented and differences between these two analyses should be discussed." (CHMP, 2007b).

Let us conclude with an outlook on potential subjects of future research based on the methods presented in this thesis. First of all, it seems natural to implement the control versus placebo superiority comparison ($H_{0,CP}^{(s)}$) in the proposed group sequential and adaptive testing procedure, because the proof of efficacy for the control treatment might be mandatory under some circumstances. In doing so, one could make use of the fact that $H_{0,CP}^{(s)}$ and $H_{0,TC}^{(n)}$ cannot be true at the same time if $H_{0,TP}^{(s)}$ is false (Koch and Röhmel, 2004). Dropping the placebo arm then seems reasonable only if both the test and the control treatment are demonstrated to be superior to placebo. Thus, the presented formulas for the overall power and the expected sample sizes, as well as those for the conditional and the predictive power have to be adjusted.

Implementing futility boundaries might also be of interest and would require only minor changes of the presented formulas. In this regard, it would be worth investigating the loss in overall power introduced by applying futility boundaries.

As we only dealt with normally distributed data, transferring the proposed methods to other distributions could also be the subject of future research. Especially the case of Poisson distributed outcomes would be of interest as e.g. in trials for the treatment of migraine the primary endpoint often is the number of attacks, and in this medical indication the three-arm design is furthermore recommended by the health authorities (CHMP, 2007a). In a similar fashion as Stucke and Kieser (2012) a general approach based on maximum likelihood estimators could be derived, covering a wide variety of distributions.

APPENDIX A

TABLES

Table A.1: Exact (first row) and approximate (second row) overall sample sizes of the optimal design to achieve an overall power of 80% with significance level $\alpha = 0.025$ and assuming $\mu_T = \mu_C$.

$\Delta_{ni}/(\mu_C - \mu_P)$	$\epsilon = \sigma/(\mu_C - \mu_P)$							
	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
1/6	75	292	654	1161	1812	2609	3550	4636
	73	290	652	1159	1810	2607	3548	4634
1/5	53	206	459	815	1271	1830	2490	3251
	51	204	457	813	1269	1828	2488	3249
1/4	36	135	300	531	828	1192	1621	2117
	34	133	298	529	826	1190	1619	2115
1/3	22	80	176	310	483	695	945	1234
	20	77	174	308	481	693	943	1232
1/2	12	40	87	153	237	340	462	603
	10	38	85	151	235	338	460	601

Table A.2: Comparison of the optimal allocations for achieving 80% overall power to reject $H^{(s)}_{0,TP} \cup H^{(n)}_{0,TC}$ and $H^{(s)}_{0,TP} \cup H^{(n)}_{0,TC} \cup H^{(s)}_{0,CP}$ with $\alpha = 0.025$ and $\mu_T = \mu_C$.

	$H^{(s)}_{0,TP} \cup H^{(n)}_{0,TC}$				$H^{(s)}_{0,TP} \cup H^{(n)}_{0,TC} \cup H^{(s)}_{0,CP}$				
$\Delta_{ni}/(\mu_C - \mu_P)$	c_C	c_P	$1-\beta_{TP}$	$1-\beta_{TC}$	c_C	c_P	$1-\beta_{TP}$	$1-\beta_{TC}$	$1-\beta_{CP}$
0.1	1.00	0.02	99.9%	80.1%	1.00	0.02	99.9%	80.1%	99.9%
0.2	1.00	0.06	99.7%	80.2%	1.00	0.06	99.8%	80.2%	99.8%
0.3	1.00	0.12	99.3%	80.4%	1.01	0.14	99.6%	80.5%	99.6%
0.4	0.99	0.20	98.5%	80.8%	1.02	0.23	99.1%	81.0%	99.1%
0.5	0.98	0.30	97.4%	81.4%	1.04	0.35	98.4%	81.8%	98.4%

Table A.3: Optimal allocation ratios and spending parameters of ρ-family designs minimising ASN or $ASN + N_{max}$ with corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

	$1-\beta$	K	c_C	c_P	ρ_{TP}	ρ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
min(ASN)	0.80	2	0.97	0.40	0.363	1.141	109.7	79.8	84.0
		3	0.97	0.46	0.365	1.327	113.2	72.9	79.4
		4	0.96	0.49	0.392	1.481	114.8	69.7	77.2
		5	0.96	0.51	0.423	1.584	115.7	67.9	76.0
	0.90	2	0.98	0.34	0.283	0.885	110.8	75.0	77.3
		3	0.98	0.40	0.235	0.908	116.5	67.2	71.4
		4	0.97	0.44	0.245	0.993	119.2	63.5	68.8
		5	0.97	0.47	0.264	1.067	120.7	61.5	67.3
min($ASN + N_{max}$)	0.80	2	0.98	0.34	1.065	2.355	102.9	72.3	86.2
		3	0.98	0.35	1.301	3.032	103.1	65.5	82.5
		4	0.97	0.35	1.458	3.378	103.3	62.7	80.5
		5	0.97	0.35	1.565	3.600	103.4	61.1	79.3
	0.90	2	0.98	0.30	0.874	1.956	103.5	67.3	79.7
		3	0.98	0.30	1.008	2.456	104.0	58.5	75.2
		4	0.98	0.31	1.137	2.770	104.1	55.3	73.0
		5	0.98	0.31	1.224	2.953	104.3	53.6	71.6

Table A.4: Optimal allocation ratios and spending parameters of γ-family designs minimising ASN or $ASN + N_{max}$ with corresponding operating characteristics for overall power $1-\beta$, $\alpha = 0.025$, $\theta_{TC} = 0$, $\Delta_{ni} = \frac{\theta_{CP}}{2}$ and K stages.

	$1-\beta$	K	c_C	c_P	γ_{TP}	γ_{TC}	$\frac{N_{max}}{N_{fix}}$	$\frac{ASn_P}{n_{P,fix}}$	$\frac{ASN}{N_{fix}}$
min(ASN)	0.80	2	0.97	0.40	2.503	−0.373	109.7	79.8	84.0
		3	0.97	0.46	2.844	−0.802	113.4	72.8	79.4
		4	0.96	0.50	3.012	−1.053	115.5	69.6	77.3
		5	0.96	0.52	3.010	−1.228	116.6	67.6	76.0
	0.90	2	0.98	0.34	3.063	0.332	110.8	75.0	77.3
		3	0.98	0.40	3.968	0.279	116.7	67.1	71.4
		4	0.97	0.44	4.401	0.150	120.0	63.2	68.7
		5	0.97	0.47	4.643	0.064	122.2	60.9	67.2
min($ASN + N_{max}$)	0.80	2	0.98	0.34	−0.176	−2.830	102.9	72.3	86.2
		3	0.98	0.35	−0.708	−3.755	103.1	65.3	82.5
		4	0.98	0.35	−1.050	−4.258	103.1	62.5	80.6
		5	0.98	0.35	−1.269	−4.553	103.2	60.9	79.5
	0.90	2	0.98	0.30	0.365	−2.115	103.5	67.3	79.7
		3	0.98	0.31	0.013	−2.962	103.9	58.6	75.2
		4	0.98	0.31	−0.287	−3.423	103.9	55.2	73.0
		5	0.98	0.31	−0.485	−3.712	104.0	53.4	71.8

APPENDIX B

R FUNCTIONS

B.1 Overview

Chapter 2: Three-Arm Non-Inferiority Trials

- Exact and approximate overall power of the fixed sample size design
→ `ThreeArmSingleStagePower()`

- Required sample sizes with predefined c_C, c_P to achieve a certain overall power
→ `ThreeArmSingleStageDesign()`

- Optimal sample sizes to achieve a certain overall power
→ `ThreeArmSingleStageOptDesign()`

Chapter 3: Group Sequential Three-Arm Non-Inferiority Designs

- Boundaries of Wang Tsiatis type designs with equal stage-sizes
→ `WangTsiatis()`

- Boundaries of Kim DeMets and Hwang Shih DeCani error spending designs with equal stage-sizes
→ `KD()`, `HSD()`, `ErrorSpending()`

- Overall power of the group sequential testing procedure
→ `ThreeArmGroupSeqPower()`

- Required sample sizes with predefined c_C, c_P to achieve a certain overall power
→ `ThreeArmGroupSeqDesign()`

- Expected sample size of the placebo group and overall
→ `ThreeArmGroupSeqASN()`

Chapter 4: Extension to Adaptive Designs

- Conditional power to reject either $H_{0,TC}^{(n)}$ alone or both null hypotheses at the final stage given the interim data
  → `ThreeArmAdaptiveCP()`
- Required second-stage sample sizes with predefined $\tilde{c}_C^{(2)}, \tilde{c}_P^{(2)}$ to achieve a certain conditional power
  → `ThreeArmAdaptiveReCalcCP()`
- Optimal second-stage sample sizes to achieve a certain conditional power
  → `ThreeArmAdaptiveOptReCalcCP()`
- Predictive power to reject either $H_{0,TC}^{(n)}$ alone or both null hypotheses at the final stage given the interim data
  → `ThreeArmAdaptivePP()`

B.2 Source Code

```
##################################################################################
##############                                                      ##############
############## Calculation of the overall power in a single stage   ##############
############## three-arm non-inferiority trial                      ##############
##############                                                      ##############
##################################################################################
#                                                                                #
# Author:       Patrick Schlömer                                                 #
# Last update:  13/May/2014                                                      #
#                                                                                #
##################################################################################
#                                                                                #
# REQUIRED PACKAGES:   - mnormt                                                  #
# -----------------     - mvtnorm                                                #
#                                                                                #
# REQUIRED FUNCTIONS:  NONE                                                      #
# ------------------                                                             #
#                                                                                #
##################################################################################
#                                                                                #
# THIS FUNCTION:                                                                 #
# --------------                                                                 #
# Calculates the power to reject H_0,TP^(s), the separate powers to reject       #
# H_0,TC^(n) and H_0,CP^(s) and the overall power to reject all null            #
# hypotheses.                                                                    #
#                                                                                #
##################################################################################

ThreeArmSingleStagePower <- function(nT,nC,nP,thetaTP,thetaTC,sigma,DeltaNI,
                                     alpha,method="approx",H0CP=FALSE) {

##################################################################################
#                                                                                #
# INPUT-PARAMETERS:                                                              #
# -----------------                                                              #
#________________________________________________________________________________#
#             |            |           |                                         #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                             #
#_____________|____________|___________|_________________________________________#
#             |            |           |                                         #
# nT          | float      | >0        | Sample size of the test group           #
#             |            |           |                                         #
# nC          | float      | >0        | Sample size of the control group        #
#             |            |           |                                         #
# nP          | float      | >0        | Sample size of the placebo group        #
#             |            |           |                                         #
# thetaTP     | float      |           | True treatment difference between       #
#             |            |           | test and placebo group                  #
#             |            |           |                                         #
# thetaTC     | float      |           | True treatment difference between       #
#             |            |           | test and control group                  #
#             |            |           |                                         #
# sigma       | float      | >0        | Common standard deviation               #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Separate significance level           #
#             |            |           |                                       #
# method      | string     | "approx", | Defines the method to calculate the   #
#             |            | "exact"   | overall power: either approximative   #
#             |            |           | by means of the normal distribution   #
#             |            |           | (DEFAULT) exact with the t-distrib.   #
#             |            |           |                                       #
# HOCP        | logical    | TRUE,     | Defines if H_0,CP^(s) also has to be  #
#             |            | FALSE     | rejected (TRUE) or not (FALSE), which #
#             |            |           | is the default value. Including       #
#             |            |           | H_0,CP^(s) is only allowed for        #
#             |            |           | method="approx".                      #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# PowerTP     | float      | >0 & <1   | Separate power to reject H_0,TP^(s)   #
#             |            |           |                                       #
# PowerTC     | float      | >0 & <1   | Separate power to reject H_0,TC^(n)   #
#             |            |           |                                       #
# PowerCP     | float      | >0 & <1   | Separate power to reject H_0,CP^(s)   #
#             |            |           |                                       #
# Power       | float      | >0 & <1   | Overall power of the procedure        #
#             |            |           |                                       #
################################################################################

  if (method=="exact") {
    # exact power using t-distribution

    # critical value
    crit <- qt(p=1-alpha,df=max(round(nT+nC+nP-3),1))

    # separate power to reject H_0,TP^(s)
    PowerTP <- 1-pt(q=crit,df=max(round(nT+nC+nP-3),1),
                    ncp=thetaTP/sigma*sqrt(nT*nP/(nT+nP)))

    # separate power to reject H_0,TC^(n)
    PowerTC <- 1-pt(q=crit,df=max(round(nT+nC+nP-3),1),
                    ncp=(thetaTC+DeltaNI)/sigma*sqrt(nT*nC/(nT+nC)))

    # separate power to reject H_0,CP^(s)
    PowerCP <- 1-pt(q=crit,df=max(round(nT+nC+nP-3),1),
                    ncp=(thetaTP-thetaTC)/sigma*sqrt(nC*nP/(nC+nP)))

    # mean vector of test statistics
    theta <- c(thetaTP/sigma*sqrt(nT*nP/(nT+nP)),
               (thetaTC+DeltaNI)/sigma*sqrt(nT*nC/(nT+nC)))
    # covariance matrix of test statistics
    rho <- sqrt(nC*nP/((nT+nC)*(nT+nP)))
    cov <- matrix(data=c(1,rho,rho,1),nrow=2,ncol=2)

    # overall power to reject both null hypotheses
    Power <- sadmvt(lower=c(-Inf,-Inf),upper=c(-crit,-crit),mean=-theta,S=cov,
                    df=max(round(nT+nC+nP-3),1))[1]

  } else if (method=="approx") {
    # approximative power using normal distribution

    # critical value
    crit <- qnorm(p=1-alpha)

    # separate power to reject H_0,TP^(s)
    PowerTP <- pnorm(q=thetaTP/sigma*sqrt(nT*nP/(nT+nP))-crit)

    # separate power to reject H_0,TC^(s)
    PowerTC <- pnorm(q=(thetaTC+DeltaNI)/sigma*sqrt(nT*nC/(nT+nC))-crit)

    # separate power to reject H_0,CP^(s)
    PowerCP <- pnorm(q=(thetaTP-thetaTC)/sigma*sqrt(nC*nP/(nC+nP))-crit)

    if (H0CP==FALSE) {
      # without H_0,CP^(s)

      # mean vector of test statistics
      theta <- c(thetaTP/sigma*sqrt(nT*nP/(nT+nP)),
                 (thetaTC+DeltaNI)/sigma*sqrt(nT*nC/(nT+nC)))

      # covariance matrix of test statistics
      rho <- sqrt(nC*nP/((nT+nC)*(nT+nP)))
      cov <- matrix(data=c(1,rho,rho,1),nrow=2,ncol=2)

      # overall power to reject both nullhypotheses
      Power <- sadmvn(lower=rep(-Inf,2),upper=theta-rep(crit,2),mean=rep(0,2),
                      varcov=cov)[1]

    } else if (H0CP==TRUE) {
      # with H_0,CP^(s)

      # mean vector of test statistics
      theta <- c(thetaTP/sigma*sqrt(nT*nP/(nT+nP)),
                 (thetaTC+DeltaNI)/sigma*sqrt(nT*nC/(nT+nC)),
                 (thetaTP-thetaTC)/sigma*sqrt(nC*nP/(nC+nP)))

      # covariance matrix of test statistics
      rhoTPTC <- sqrt(nC*nP/((nT+nC)*(nT+nP)))
      rhoTPCP <- sqrt(nT*nC/((nT+nP)*(nC+nP)))
      rhoTCCP <- -sqrt(nT*nP/((nT+nC)*(nC+nP)))
      cov <- matrix(c(1,rhoTPTC,rhoTPCP,rhoTPTC,1,rhoTCCP,rhoTPCP,rhoTCCP,1),3,3)

      # overall power to reject both nullhypotheses
      Power <- pmvnorm(lower=rep(-Inf,3),upper=theta-rep(crit,3),mean=rep(0,3),
                       sigma=cov,algorithm=TVPACK(abseps=1e-6))[1]
    }

  }

  return(list(PowerTP=PowerTP,PowerTC=PowerTC,PowerCP=PowerCP,Power=Power))

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmSingleStagePower(nT=544,nC=544,nP=136,thetaTP=0.4,thetaTC=0,           #
                         sigma=1,DeltaNI=0.2,alpha=0.025)                      #
# $PowerTP                                                                     #
# [1] 0.9865279                                                                #
#                                                                              #
# $PowerTC                                                                     #
# [1] 0.9096366                                                                #
#                                                                              #
# $PowerCP                                                                     #
# [1] 0.9865279                                                                #
#                                                                              #
# $Power                                                                       #
# [1] 0.9000693                                                                #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the required sample sizes for a      #############
############## single stage three-arm non-inferiority trial        #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------    - mvtnorm                                               #
#                                                                              #
# REQUIRED FUNCTIONS:  - ThreeArmSingleStagePower()                            #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# -------------                                                                #
# Calculates the required sample sizes to obtain a specific overall power      #
# 1-beta with prespecified allocation ratios cC=nC/nP and cP=nP/nT.            #
#                                                                              #
################################################################################

ThreeArmSingleStageDesign <- function(thetaTP,thetaTC,sigma,DeltaNI,alpha,beta,
                                      cC,cP,method="approx",H0CP=FALSE) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# thetaTP     | float      |           | True treatment difference between     #
#             |            |           | test and placebo group                #
#             |            |           |                                       #
# thetaTC     | float      |           | True treatment difference between     #
#             |            |           | test and control group                #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Separate significance level           #
#             |            |           |                                       #
# beta        | float      | >0 & <1   | Targeted type II error rate           #
#             |            |           |                                       #
# cC          | float      | >0        | Relative size of the placebo group    #
#             |            |           | (cC=nC/nT)                            #
#             |            |           |                                        #
# cP          | float      | >0        | Relative size of the control group     #
#             |            |           | (cP=nP/nT)                             #
#             |            |           |                                        #
# method      | string     | "approx", | Defines the method to calculate the    #
#             |            | "exact"   | overall power: either approximative    #
#             |            |           | by means of the normal distribution    #
#             |            |           | (DEFAULT) exact with the t-distrib.    #
#             |            |           |                                        #
# HOCP        | logical    | TRUE,     | Defines if H_0,CP^(s) also has be      #
#             |            | FALSE     | rejected (TRUE) or not (FALSE), which  #
#             |            |           | is the default value. Including        #
#             |            |           | H_0,CP^(s) is only allowed for         #
#             |            |           | method="approx".                       #
#             |            |           |                                        #
#------------------------------------------------------------------------------ #
#                                                                               #
# OUTPUT-PARAMETERS:                                                            #
# ------------------                                                            #
#_______________________________________________________________________________#
#             |            |           |                                        #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                            #
#_____________|____________|___________|________________________________________#
#             |            |           |                                        #
# nT          | float      | >0        | Sample size of the test group          #
#             |            |           |                                        #
# nC          | float      | >0        | Sample size of the control group       #
#             |            |           |                                        #
# nP          | float      | >0        | Sample size of the placebo group       #
#             |            |           |                                        #
# N           | float      | >0        | Overall sample size                    #
#             |            |           |                                        #
#             |            |           |                                        #
# PowerTP     | float      | >0 & <1   | Separate power to reject H_0,TP^(s)    #
#             |            |           |                                        #
# PowerTC     | float      | >0 & <1   | Separate power to reject H_0,TC^(n)    #
#             |            |           |                                        #
# PowerCP     | float      | >0 & <1   | Separate power to reject H_0,CP^(s)    #
#             |            |           |                                        #
# Power       | float      | >0 & <1   | Overall power of the procedure         #
#             |            |           |                                        #
#################################################################################

  # new environment
  func.env <- new.env()

  # root finding function
  solvenT <- function(nT) {
    assign("nT",nT,envir=func.env)
    assign("nC",cC*nT,envir=func.env)
    assign("nP",cP*nT,envir=func.env)
    assign("Powers",ThreeArmSingleStagePower(nT=nT,nC=get("nC",envir=func.env),
                                             nP=get("nP",envir=func.env),
                                             thetaTP=thetaTP, thetaTC=thetaTC,
                                             sigma=sigma,DeltaNI=DeltaNI,
                                             alpha=alpha,
                                             method=method,
                                             H0CP=H0CP),envir=func.env)
    return(get("Powers",envir=func.env)$Power-(1-beta))
  }

  # determine required sample sizes & power
  uniroot(solvenT,lower=1,upper=1e6)

  nT <- get("nT",envir=func.env)
  nC <- get("nC",envir=func.env)
  nP <- get("nP",envir=func.env)
  Powers <- get("Powers",envir=func.env)

  return(list(nT=nT,nC=nC,nP=nP,N=nT+nC+nP,PowerTP=Powers$PowerTP,
              PowerTC=Powers$PowerTC,PowerCP=Powers$PowerCP,Power=Powers$Power))

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmSingleStageDesign(thetaTP=0.4,thetaTC=0,sigma=1,DeltaNI=0.2,           #
                          alpha=0.025,beta=0.1,cC=1,cP=0.25)                   #
# $nT                                                                          #
# [1] 543.8802                                                                 #
#                                                                              #
# $nC                                                                          #
# [1] 543.8802                                                                 #
#                                                                              #
# $nP                                                                          #
# [1] 135.97                                                                   #
#                                                                              #
# $N                                                                           #
# [1] 1223.73                                                                  #
#                                                                              #
# $PowerTP                                                                     #
# [1] 0.986512                                                                 #
#                                                                              #
# $PowerTC                                                                     #
# [1] 0.9095774                                                                #
#                                                                              #
# $PowerCP                                                                     #
# [1] 0.986512                                                                 #
#                                                                              #
# $Power                                                                       #
# [1] 0.9                                                                      #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the optimal sample sizes for a       #############
############## single stage three-arm non-inferiority trial        #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------    - mvtnorm                                               #
#                                                                              #
# REQUIRED FUNCTIONS:  - ThreeArmSingleStagePower()                            #
# ------------------   - ThreeArmSingleStageDesign()                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the optimal sample sizes to obtain a specific overall power       #
# 1-beta that minimise the overall sample size.                                #
#                                                                              #
################################################################################

ThreeArmSingleStageOptDesign <- function(thetaTP,thetaTC,sigma,DeltaNI,alpha,
                                         beta,H0CP=FALSE) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# thetaTP     | float      |           | True treatment difference between     #
#             |            |           | test and placebo group                #
#             |            |           |                                       #
# thetaTC     | float      |           | True treatment difference between     #
#             |            |           | test and control group                #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Separate significance level           #
#             |            |           |                                       #
# beta        | float      | >0 & <1   | Targeted type II error rate           #
#             |            |           |                                       #
# H0CP        | logical    | TRUE,     | Defines if H_0,CP^(s) also has be     #
#             |            | FALSE     | rejected (TRUE) or not (FALSE), which #
#             |            |           | is the default value.                 #
#             |            |           |                                       #
#---------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT          | float      | >0        | Sample size of the test group         #
#             |            |           |                                       #
# nC          | float      | >0        | Sample size of the control group      #
#             |            |           |                                       #
# nP          | float      | >0        | Sample size of the placebo group      #
#             |            |           |                                       #
# N           | float      | >0        | Overall sample size                   #
#             |            |           |                                       #
# PowerTP     | float      | >0 & <1   | Separate power to reject H_0,TP^(s)   #
#             |            |           |                                       #
# PowerTC     | float      | >0 & <1   | Separate power to reject H_0,TC^(n)   #
#             |            |           |                                       #
# PowerCP     | float      | >0 & <1   | Separate power to reject H_0,CP^(s)   #
#             |            |           |                                       #
# Power       | float      | >0 & <1   | Overall power of the procedure        #
#             |            |           |                                       #
################################################################################

  # optimisation function
  optAlloc <- function(c) {
    cC <- c[1]
    cP <- c[2]
    optDesign <<- ThreeArmSingleStageDesign(thetaTP=thetaTP,thetaTC=thetaTC,
                                            sigma=sigma,DeltaNI=DeltaNI,
                                            alpha=alpha,beta=beta,cC=cC,cP=cP,
                                            H0CP=H0CP)
    return(optDesign$N)
  }

  # find optimal design
  optim(c(0.5,0.5),optAlloc,lower=rep(0.01,2),upper=c(3,3),method="L-BFGS-B")

  return(optDesign)

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE) {                                                                   #
ThreeArmSingleStageOptDesign(thetaTP=0.4,thetaTC=0,sigma=1,DeltaNI=0.2,        #
                             alpha=0.025,beta=0.1)                             #
# $nT                                                                          #
# [1] 546.2521                                                                 #
#                                                                              #
# $nC                                                                           #
# [1] 533.5604                                                                  #
#                                                                               #
# $nP                                                                           #
# [1] 142.6462                                                                  #
#                                                                               #
# $N                                                                            #
# [1] 1222.459                                                                  #
#                                                                               #
# $PowerTP                                                                      #
# [1] 0.989109                                                                  #
#                                                                               #
# $PowerTC                                                                      #
# [1] 0.9075568                                                                 #
#                                                                               #
# $PowerCP                                                                      #
# [1] 0.9888057                                                                 #
#                                                                               #
# $Power                                                                        #
# [1] 0.9                                                                       #
}                                                                               #
#################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the group sequential boundaries      #############
############## according to Wang & Tsiatis (1987)                  #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the group sequential rejection boundaries for the Delta-class     #
# proposed by Wang & Tsiatis (1987). Equal stage sizes are assumed.           #
#                                                                              #
################################################################################

WangTsiatis <- function(K,alpha,Delta){

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# K           | float      | >1        | Number of stages                      #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Significance level                    #
#             |            |           |                                       #
# Delta       | float      |           | Parameter defining the shape of the   #
#             |            |           | rejection boundaries. Delta=-Inf      #
#             |            |           | returns the boundary values of the    #
#             |            |           | common single stage design.           #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# bounds      | Kx1 vector |           | Stage-wise rejection boundaries       #
#             | (floats)   |           |                                       #
#             |            |           |                                       #
################################################################################

  # new environment
  func.env <- new.env()

  if (Delta==-Inf) {

    # Delta==-Inf returns the boundaries of the common single stage design
    assign("bounds",c(rep(Inf,K-1),qnorm(1-alpha)),envir=func.env)

  } else {

    # covariance matrix of the test statistics (equal stage sizes!)
    cov <- sapply(1:K,function(j)
                      sapply(1:K,function(i,j) sqrt(min(i,j)/max(i,j)), j=j))

    # funtion to solve for bWT
    solvebWT <- function(bWT){
      assign("bounds",(1:K/K)^(Delta-1/2)*bWT,envir=func.env)
      typeIerror <- 1-sadmvn(lower=rep(-Inf,K),upper=get("bounds",envir=func.env),
                             mean=rep(0,K),varcov=cov)[1]
      return(typeIerror-alpha)
    }

    # determine boundaries
    uniroot(solvebWT,lower=1e-10,upper=1e10)

  }

  return(get("bounds",envir=func.env))

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
WangTsiatis(K=3,alpha=0.025,Delta=0)                                           #
# [1] 3.471086 2.454429 2.004033                                               #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the error spending function proposed #############
############## by Kim & DeMets (1987)                              #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the error spending function proposed by Kim & DeMets (1987) with  #
# shape parameter rho.                                                         #
#                                                                              #
################################################################################

KD <- function(t,spendpar,alpha){

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# t           | float      |           | Time parameter                        #
#             |            |           |                                       #
# spendpar    | float      | >0        | Parameter defining the shape of the   #
#             |            |           | spending function                     #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Significance level                    #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# f           | float      | >=0       | Cumulative type I error spent at      #
#             |            | <=alpha   | time t                                #
#             |            |           |                                       #
################################################################################

  if (t<=0) {

    f <- 0

  } else if (t>0 && t<1) {

    f <- alpha*t^spendpar

  } else {

    f <- alpha

  }

  return(f)

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
KD(t=0.5,spendpar=1,alpha=0.025)                                               #
# [1] 0.0125                                                                   #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the error spending function proposed #############
############## by Hwang, Shih & DeCani (1990)                      #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the error spending function proposed by Hwang, Shih & DeCani      #
# (1990) with shape parameter gamma.                                           #
#                                                                              #
################################################################################

HSD <- function(t,spendpar,alpha){

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# t           | float      |           | Time parameter                        #
#             |            |           |                                       #
# spendpar    | float      | >0        | Parameter defining the shape of the   #
#             |            |           | spending function                     #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Significance level                    #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# f           | float      | >=0       | Cumulative type I error spent at      #
#             |            | <=alpha   | time t                                #
#             |            |           |                                       #
################################################################################

  if (t<=0) {

    f <- 0

  } else if (t>0 && t<1) {

    if (spendpar==0){

      f <- alpha*t

    } else {

      f <- alpha*(1-exp(-spendpar*t))/(1-exp(-spendpar))

    }

  } else {

    f <- alpha

  }

  return(f)

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
HSD(t=0.5,spendpar=1,alpha=0.025)                                              #
# [1] 0.01556148                                                               #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the group sequential boundaries for  #############
############## error spending designs                              #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  - KD()                                                  #
# ------------------   - HSD()                                                 #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the group sequential rejection boundaries for the rho- and gamma- #
# class of error spending designs proposed by Kim & DeMets (1987) and Hwang,   #
# Shih & DeCani (1990), respectively. Equal stage sizes are assumed.           #
#                                                                              #
################################################################################

ErrorSpending <- function(K,alpha,spendfunc,spendpar){

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#------------------------------------------------------------------------------#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#-------------|------------|-----------|---------------------------------------#
#             |            |           |                                       #
# K           | float      | >1        | Number of stages                      #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Significance level                    #
#             |            |           |                                       #
# spendfunc   | function   | KD, HSD   | Defines the family of error spending  #
#             |            |           | functions (KD = Kim & DeMets, HSD =   #
#             |            |           | Hwang, Shih & DeCani)                 #
#             |            |           |                                       #
# spendpar    | float      | KD: >0    | Parameter defining the shape of the   #
#             |            |           | spending function. spendpar=Inf for   #
#             |            |           | KD designs and spendpar=-Inf for HSD  #
#             |            |           | designs return the boundary values    #
#             |            |           | of the common single stage design,    #
#             |            |           | i.e. Inf at stages 1,...,K-1 and      #
#             |            |           | qnorm(1-alpha) at stage K.            #
#             |            |           |                                       #
```

```
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# bounds      | Kx1 vector |           | Stage-wise rejection boundaries       #
#             | (floats)   |           |                                       #
#             |            |           |                                       #
################################################################################

  # new environment
  func.env <- new.env()

  if (as.character(substitute(spendfunc))=="KD" & spendpar==Inf) {

    # for KD designs with spendpar==Inf, return single stage design
    assign("bounds",c(rep(Inf,K-1),qnorm(1-alpha)),envir=func.env)

  } else if (as.character(substitute(spendfunc))=="HSD" & spendpar==-Inf) {

    # for HSD designs with spendpar==-Inf, return single stage design
    assign("bounds",c(rep(Inf,K-1),qnorm(1-alpha)),envir=func.env)

  } else {

    # covariance matrix of the test statistics (equal stage sizes!)
    cov <- sapply(1:K,function(j)
                      sapply(1:K,function(i,j)
                                 sqrt(min(i,j)/max(i,j)),j=j))

    # calculate boundaries
    calcbounds <- function(k){
      if (k==1){
        pk <- spendfunc(1/K,spendpar=spendpar,alpha=alpha)
        assign("bounds",qnorm(1-pk),envir=func.env)
      } else {
        solvebk <- function(bk){
          k <- length(get("bounds",envir=func.env))+1
          pk <- spendfunc(k/K,spendpar=spendpar,alpha=alpha)-
                spendfunc((k-1)/K,spendpar=spendpar,alpha=alpha)
          prob <- sadmvn(lower=c(rep(-Inf,k-1),bk),
                         upper=c(get("bounds",envir=func.env),Inf),
                         mean=rep(0,k),varcov=cov[1:k,1:k])[1]
          return(prob-pk)
        }
        assign("bounds",
               c(get("bounds",envir=func.env),
                 uniroot(solvebk,lower=1e-10,upper=1e10)$root),envir=func.env)
      }
    }

    sapply(1:K,calcbounds)
  }

  return(get("bounds",envir=func.env))

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ErrorSpending(K=3,alpha=0.025,spendfunc=KD,spendpar=1)                         #
# [1] 2.393980 2.293769 2.199902                                               #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the overall power of a group         #############
############## sequential three-arm non-inferiority design         #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the power to reject H_0,TP^(s) and the overall power to reject    #
# both null hypotheses.                                                        #
#                                                                              #
################################################################################

ThreeArmGroupSeqPower <- function(K,nT,nC,nP,thetaTP,thetaTC,sigma,DeltaNI,
                                  bTP,bTC) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# K           | integer    | >1        | Number of stages                      #
#             |            |           |                                       #
# nT          | Kx1 vector | >0        | Cumulative sample sizes of the test   #
#             | (floats)   |           | group                                 #
#             |            |           |                                       #
# nC          | Kx1 vector | >0        | Cumulative sample sizes of the        #
#             | (floats)   |           | control group                         #
#             |            |           |                                       #
# nP          | Kx1 vector | >0        | Cumulative sample sizes of the        #
#             | (floats)   |           | placebo group                         #
#             |            |           |                                       #
# thetaTP     | float      |           | True treatment difference between     #
#             |            |           | test and placebo group                #
#             |            |           |                                       #
# thetaTC     | float      |           | True treatment difference between     #
#             |            |           | test and control group                #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# bTP         | Kx1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | Kx1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# PowerTP     | float      | >0 & <1   | Power to reject H_0,TP^(s)            #
#             |            |           |                                       #
# Power       | float      | >0 & <1   | Overall power of the procedure        #
#             |            |           |                                       #
################################################################################

  # fisher informations of the test statistics:
  I_TP <- (sigma^2*(1/nT+1/nP))^-1
  I_TC <- (sigma^2*(1/nT+1/nC))^-1

  # covariance matrix of the vector (Z_TP^(1),..,Z_TP^(K),Z_TC^(1),..,Z_TC^(K))'
  covTP <- sapply(1:K,function(j)
                    sapply(1:K,function(i,j)
                               sqrt(I_TP[min(i,j)]/I_TP[max(i,j)]),j=j))
  covTC <- sapply(1:K,function(j)
                    sapply(1:K,function(i,j)
                               sqrt(I_TC[min(i,j)]/I_TC[max(i,j)]),j=j))
  covTCP <- sapply(1:K,function(j)
                     sapply(1:K,function(i,j)
                                sigma^2/nT[max(i,j)]*sqrt(I_TP[i]*I_TC[j]),j=j))
  cov <- rbind(cbind(covTP,covTCP),cbind(t(covTCP),covTC))

  # power to reject H_0,TP^(s)
  PowerTP <- 1-sadmvn(lower=rep(-Inf,K),upper=bTP-thetaTP*sqrt(I_TP),
                    mean=rep(0,K),varcov=covTP)[1]

  # probabilities P(A_k), 1<=k<=K (A_k = H_0,TP^(s) rejected at stage k and
  # H_0,TC^(n) is not rejected at stages k,...,K)
  CalcProbAk <- function(k){
    if (k==1){
      # k=1
      ProbAk <- sadmvn(lower=c(bTP[1]-thetaTP*sqrt(I_TP[1]),rep(-Inf,K)),
                       upper=c(Inf,bTC[1:K]-(thetaTC+DeltaNI)*sqrt(I_TC[1:K])),
                       mean=rep(0,K+1),
                       varcov=cov[-(2:K),-(2:K)])[1]
    } else {
      # 2<=k<=K
      lower <- rep(-Inf,K+1)
      lower[k] <- bTP[k]-thetaTP*sqrt(I_TP[k])
      upper <- c(bTP[1:(k-1)]-thetaTP*sqrt(I_TP[1:(k-1)]),Inf,
                 bTC[k:K]-(thetaTC+DeltaNI)*sqrt(I_TC[k:K]))
      varcov <- cov[-((k+1):(K+k-1)),-((k+1):(K+k-1))]
      ProbAk <- sadmvn(lower=lower,upper=upper,mean=rep(0,K+1),varcov=varcov)[1]
    }
    return(ProbAk)
  }

  # overall power
  Power <- PowerTP - sum(sapply(1:K,CalcProbAk))

  return(list(PowerTP=PowerTP,Power=Power))

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmGroupSeqPower(K=3,nT=cumsum(rep(188,3)),nC=cumsum(rep(188,3)),         #
                      nP=cumsum(rep(47,3)),thetaTP=0.4,thetaTC=0,sigma=1,      #
                      DeltaNI=0.2,bTP=c(2.741,2.305,2.083),                    #
                      bTC=c(3.471,2.454,2.004))                                #
# $PowerTP                                                                     #
# [1] 0.9861338                                                                #
#                                                                              #
# $Power                                                                       #
# [1] 0.9047309                                                                #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the required sample size for a group #############
############## sequential three-arm non-inferiority design         #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  - WangTsiatis()                                         #
# ------------------   - KD()                                                  #
#                      - HSD()                                                 #
#                      - ErrorSpending()                                       #
#                      - ThreeArmGroupSeqPower()                               #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the required sample sizes to obtain a specific overall power      #
# 1-beta with prespecified allocation ratios cC=nC/nP and cP=nP/nT. Equal      #
# stage sizes are assumed, i.e. nD^(k)=k/K*nD^(K) for D=T,C,P.                 #
#                                                                              #
################################################################################

ThreeArmGroupSeqDesign <- function(K,thetaTP,thetaTC,sigma,DeltaNI,alpha,beta,
                                   cC,cP,type,parTP,parTC) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# K           | integer    | >1        | Number of stages                      #
#             |            |           |                                       #
# thetaTP     | float      |           | True treatment difference between     #
#             |            |           | test and placebo group                #
#             |            |           |                                       #
# thetaTC     | float      |           | True treatment difference between     #
#             |            |           | test and control group                #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# alpha       | float      | >0 & <1   | Separate significance level           #
#             |            |           |                                       #
# beta        | float      | >0 & <1   | Targeted type II error rate           #
#             |            |           |                                       #
# cC          | float      | >0        | Relative size of the placebo group    #
#             |            |           | (cC=nC/nT)                            #
#             |            |           |                                       #
# cP          | float      | >0        | Relative size of the control group    #
#             |            |           | (cP=nP/nT)                            #
#             |            |           |                                       #
# type        | string     | "WT","KD",| Defines the type of the rejection     #
#             |            | "HSD"     | boundaries that are calulated ("WT"=  #
#             |            |           | Wang Tsiatis, "KD"= Kim & DeMets      #
#             |            |           | error spending, "HSD"= Hwang, Shih &  #
#             |            |           | DeCani error spending)                #
#             |            |           |                                       #
# parTP       | float      | type="KD":| Parameter that defines the rejection  #
#             |            | >0        | boundaries for H_0,TP^(s)             #
#             |            |           |                                       #
# parTC       | float      | type="KD":| Parameter that defines the rejection  #
#             |            | >0        | boundaries for H_0,TC^(n)             #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# bTP         | Kx1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | Kx1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
# nT          | Kx1 vector | >0        | Cumulative sample sizes of the test   #
#             | (floats)   |           | group                                 #
#             |            |           |                                       #
# nC          | Kx1 vector | >0        | Cumulative sample sizes of the        #
#             | (floats)   |           | control group                         #
#             |            |           |                                       #
# nP          | Kx1 vector | >0        | Cumulative sample sizes of the        #
#             | (floats)   |           | placebo group                         #
#             |            |           |                                       #
# N           | Kx1 vector | >0        | Cumulative overall sample sizes       #
#             | (floats)   |           |                                       #
#             |            |           |                                       #
# PowerTP     | float      | >0 & <1   | Power to reject H_0,TP^(s)            #
#             |            |           |                                       #
# Power       | float      | >0 & <1   | Overall power of the procedure        #
#             |            |           |                                       #
################################################################################

  # new environment
  func.env <- new.env()

  # rejection boundaries
  if (type=="WT") {
    bTP <- WangTsiatis(K=K,alpha=alpha,Delta=parTP)
    bTC <- WangTsiatis(K=K,alpha=alpha,Delta=parTC)
  } else if (type=="KD") {
    bTP <- ErrorSpending(K=K,alpha=alpha,spendfunc=KD,spendpar=parTP)
    bTC <- ErrorSpending(K=K,alpha=alpha,spendfunc=KD,spendpar=parTC)
  } else if (type=="HSD") {
    bTP <- ErrorSpending(K=K,alpha=alpha,spendfunc=HSD,spendpar=parTP)
    bTC <- ErrorSpending(K=K,alpha=alpha,spendfunc=HSD,spendpar=parTC)
  }

  # root finding function
  solvenTK <- function(nTK) {
    assign("nT",1:K/K*nTK,envir=func.env)
    assign("nC",cC*1:K/K*nTK,envir=func.env)
    assign("nP",cP*1:K/K*nTK,envir=func.env)
    assign("Powers",ThreeArmGroupSeqPower(K=K,nT=get("nT",envir=func.env),
                                          nC=get("nC",envir=func.env),
                                          nP=get("nP",envir=func.env),
                                          thetaTP=thetaTP,thetaTC=thetaTC,
                                          sigma=sigma,DeltaNI=DeltaNI,
                                          bTP=bTP,bTC=bTC),envir=func.env)
    return(get("Powers",envir=func.env)$Power-(1-beta))
  }

  # determine required sample sizes & power
  uniroot(solvenTK,lower=1,upper=1e6)

  nT <- get("nT",envir=func.env)
  nC <- get("nC",envir=func.env)
  nP <- get("nP",envir=func.env)
  Powers <- get("Powers",envir=func.env)

  return(list(bTP=bTP,bTC=bTC,nT=nT,nC=nC,nP=nP,N=nT+nC+nP,
              PowerTP=Powers$PowerTP,Power=Powers$Power))

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmGroupSeqDesign(K=3,thetaTP=0.4,thetaTC=0,sigma=1,DeltaNI=0.2,          #
                       alpha=0.025,beta=0.1,cC=1,cP=0.25,type="WT",parTP=0.25, #
                       parTC=0)                                                #
# $bTP                                                                         #
# [1] 2.741137 2.305013 2.082814                                               #
#                                                                              #
# $bTC                                                                         #
# [1] 3.471086 2.454429 2.004033                                               #
#                                                                              #
# $nT                                                                          #
# [1] 185.2007 370.4013 555.6020                                               #
#                                                                              #
# $nC                                                                          #
# [1] 185.2007 370.4013 555.6020                                               #
#                                                                              #
# $nP                                                                          #
# [1] 46.30016  92.60033 138.90049                                             #
#                                                                              #
# $N                                                                           #
# [1] 416.7015  833.4029 1250.1044                                             #
#                                                                              #
# $PowerTP                                                                     #
# [1] 0.9849846                                                                #
#                                                                              #
# $Power                                                                       #
# [1] 0.9                                                                      #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the expected sample size for a group #############
############## sequential three-arm non-inferiority design         #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the expected sample sizes of a group sequential three-arm        #
# non-inferiority design at a specific alternative thetaTP and thetaTC.       #
#                                                                              #
################################################################################

ThreeArmGroupSeqASN <- function(K,nT,nC,nP,thetaTP,thetaTC,sigma,DeltaNI,
                                bTP,bTC) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# K           | integer    | >1        | Number of stages                      #
#             |            |           |                                       #
# nT          | Kx1 vector | >0        | Cumulative sample sizes of the test   #
#             | (floats)   |           | group                                 #
#             |            |           |                                       #
# nC          | Kx1 vector | >0        | Cumulative sample sizes of the        #
#             | (floats)   |           | control group                         #
#             |            |           |                                       #
# nP          | Kx1 vector | >0        | Cumulative sample sizes of the        #
#             | (floats)   |           | placebo group                         #
#             |            |           |                                       #
# thetaTP     | float      |           | Treatment difference between test and #
#             |            |           | placebo for which ASN is calculated   #
#             |            |           |                                       #
# thetaTC     | float      |           | Treatment difference between test and #
#             |            |           | control for which ASN is calculated   #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# bTP         | Kx1 vector |           | Stage-wise rejection boundaries for  #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | Kx1 vector |           | Stage-wise rejection boundaries for  #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# ASnP        | float      | >0        | Expected placebo group size           #
#             |            |           |                                       #
# ASN         | float      | >0        | Expected overall sample size          #
#             |            |           |                                       #
################################################################################

  # fisher informations of the test statistics:
  I_TP <- (sigma^2*(1/nT+1/nP))^-1
  I_TC <- (sigma^2*(1/nT+1/nC))^-1

  # covariance matrix of the vector (Z_TP^(1),..,Z_TP^(K),Z_TC^(1),..,Z_TC^(K))'
  covTP <- sapply(1:K,function(j)
                    sapply(1:K,function(i,j)
                               sqrt(I_TP[min(i,j)]/I_TP[max(i,j)]),j=j))
  covTC <- sapply(1:K,function(j)
                    sapply(1:K,function(i,j)
                               sqrt(I_TC[min(i,j)]/I_TC[max(i,j)]),j=j))
  covTCP <- sapply(1:K,function(j)
                     sapply(1:K,function(i,j)
                                sigma^2/nT[max(i,j)]*sqrt(I_TP[i]*I_TC[j]),j=j))
  cov <- rbind(cbind(covTP,covTCP),cbind(t(covTCP),covTC))

  # probabilities P(E_k1,k), 2<=k<=K, 0<=k1<=k-1
  CalcProbEk1k <- function(k1,k){
    if (k1==0){
      # no rejection of H_0,TP^(s)
      ProbEk1k <- sadmvn(lower=rep(-Inf,k-1),
                         upper=bTP[1:(k-1)]-thetaTP*sqrt(I_TP[1:(k-1)]),
                         mean=rep(0,k-1),varcov=covTP[-(k:K),-(k:K)])[1]
    } else if (k1==1) {
      # rejection of H_0,TP^(s) at the first stage
      ProbEk1k <- sadmvn(lower=c(bTP[1]-thetaTP*sqrt(I_TP[1]),rep(-Inf,k-1)),
                         upper=c(Inf,bTC[1:(k-1)]-
                                     (thetaTC+DeltaNI)*sqrt(I_TC[1:(k-1)])),
                         mean=rep(0,k),
                         varcov=cov[-c(2:K,(K+k):(2*K)),-c(2:K,(K+k):(2*K))])[1]
    } else {
      # rejection of H_0,TP^(s) at stage k1, 2<=k1<=k-1
      lower <- rep(-Inf,k)
      lower[k1] <- bTP[k1]-thetaTP*sqrt(I_TP[k1])
      upper <- c(bTP[1:(k1-1)]-thetaTP*sqrt(I_TP[1:(k1-1)]),Inf,
                 bTC[k1:(k-1)]-(thetaTC+DeltaNI)*sqrt(I_TC[k1:(k-1)]))
      varcov <- cov[-c((k1+1):(K+k1-1),(K+k):(2*K)),
                    -c((k1+1):(K+k1-1),(K+k):(2*K))]
      ProbEk1k <- sadmvn(lower=lower,upper=upper,mean=rep(0,k),varcov=varcov)[1]
    }
    return(ProbEk1k)
  }

  # probabilities P(E_k), 2<=k<=K
  CalcProbEk <- function(k){
    sum(sapply(0:(k-1),CalcProbEk1k,k=k))
  }

  # average sample number of the test and control group
  ASnT <- nT[1] + sum((nT[2:K]-nT[1:(K-1)])*sapply(2:K,CalcProbEk))
  ASnC <- nC[1] + sum((nC[2:K]-nC[1:(K-1)])*sapply(2:K,CalcProbEk))


  # average sample number of placebo group
  ASnP <- nP[1] + sum((nP[2:K]-nP[1:(K-1)])*sapply(2:K,CalcProbEk1k,k1=0))

  # overall average sample number
  ASN <- ASnT + ASnC + ASnP

  return(list(ASnT=ASnT,ASnC=ASnC,ASnP=ASnP,ASN=ASN))

}

#################################################################################
# EXAMPLE:                                                                      #
# --------                                                                      #
if (FALSE){                                                                     #
ThreeArmGroupSeqASN(K=3,nT=cumsum(rep(188,3)),nC=cumsum(rep(188,3)),            #
                    nP=cumsum(rep(47,3)),thetaTP=0.4,thetaTC=0,sigma=1,         #
                    DeltaNI=0.2,bTP=c(2.741,2.305,2.083),                       #
                    bTC=c(3.471,2.454,2.004))                                   #
# $ASnT                                                                         #
# [1] 450.0797                                                                  #
#                                                                               #
# $ASnC                                                                         #
# [1] 450.0797                                                                  #
#                                                                               #
# $ASnP                                                                         #
# [1] 81.42504                                                                  #
#                                                                               #
# $ASN                                                                          #
# [1] 981.5844                                                                  #
}                                                                               #
#################################################################################
```

```
################################################################################
##############                                                    ##############
############## Calculation of the conditional power for a         ##############
############## two-stage adaptive group sequential three-arm      ##############
############## non-inferiority design                             ##############
##############                                                    ##############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the conditional power given the interim data to reject either     #
# both null hypotheses or only H_0,TC^(n) at the final analysis. The latter    #
# might e.g. be of interest, if H_0,TP^(s) has already been rejected at the    #
# interim analysis.                                                            #
#                                                                              #
################################################################################

ThreeArmAdaptiveCP <- function(nT,nC,nP,thetaTP,thetaTC,sigma,DeltaNI,diffTP,
                               diffTC,bTP,bTC,wTP,wTC,H0TP) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT          | 2x1 vector | >0        | Stage-wise sample sizes of the test   #
#             | (floats)   |           | group                                 #
#             |            |           |                                       #
# nC          | 2x1 vector | >0        | Stage-wise sample sizes of the        #
#             | (floats)   |           | control group                         #
#             |            |           |                                       #
# nP          | 2x1 vector | >0        | Stage-wise sample sizes of the        #
#             | (floats)   |           | placebo group                         #
#             |            |           |                                       #
# thetaTP     | float      |           | Treatment difference between test and #
#             |            |           | placebo for which CP is calculated    #
#             |            |           |                                       #
# thetaTC     | float      |           | Treatment difference between test and #
#             |            |           | control for which CP is calculated    #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# diffTP      | float      |           | Observed treatment difference between #
#             |            |           | test and placebo at interim           #
#             |            |           |                                       #
# diffTC      | float      |           | Observed treatment difference between #
#             |            |           | test and control at interim           #
#             |            |           |                                       #
# bTP         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
# wTP         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TP^(s)             #
#             |            |           |                                       #
# wTC         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TC^(n)             #
#             |            |           |                                       #
# H0TP        | boolean    | 0 or 1    | Defines if the conditional power to   #
#             |            |           | reject both null hypotheses (H0TP=1)  #
#             |            |           | or only H_0,TC^(n) (H0TP=0) is        #
#             |            |           | calculated.                           #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# CP          | float      | >0 & <1   | Conditional power                     #
#             |            |           |                                       #
################################################################################

  if (H0TP==0) {
    # conditional power to reject H_0,TC^(n)

    # test statistic of the first stage
    ZTC1 <- ((diffTC+DeltaNI)/sigma)*sqrt((nT[1]*nC[1])/(nT[1]+nC[1]))

    CP <- pnorm((thetaTC+DeltaNI)/sigma*sqrt((nT[2]*nC[2])/(nT[2]+nC[2]))
                +wTC[1]/wTC[2]*ZTC1-sqrt(wTC[1]^2+wTC[2]^2)/wTC[2]*bTC[2])

  } else if (H0TP==1) {
    # conditional power to reject both H_0,TP^(s) and H_0,TC^(n)

    # test statistics of the first stage
    ZTP1 <- (diffTP/sigma)*sqrt((nT[1]*nP[1])/(nT[1]+nP[1]))
    ZTC1 <- ((diffTC+DeltaNI)/sigma)*sqrt((nT[1]*nC[1])/(nT[1]+nC[1]))

    # variance-covariance matrix for test statistics of second stage
    rho <- sqrt((nC[2]*nP[2])/((nT[2]+nC[2])*(nT[2]+nP[2])))
    varcov <- matrix(c(1,rho,rho,1),ncol=2)

    # upper limits for integration
    upper <- c(thetaTP/sigma*sqrt((nT[2]*nP[2])/(nT[2]+nP[2]))
               +wTP[1]/wTP[2]*ZTP1-sqrt(wTP[1]^2+wTP[2]^2)/wTP[2]*bTP[2],
               (thetaTC+DeltaNI)/sigma*sqrt((nT[2]*nC[2])/(nT[2]+nC[2]))
               +wTC[1]/wTC[2]*ZTC1-sqrt(wTC[1]^2+wTC[2]^2)/wTC[2]*bTC[2])

    # conditional power
    CP <- sadmvn(lower=rep(-Inf,2),upper=upper,mean=rep(0,2),varcov=varcov)[1]

  }

  return(CP)

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmAdaptiveCP(nT=rep(275,2),nC=rep(275,2),nP=rep(69,2),thetaTP=0.3,       #
                   thetaTC=-0.02,sigma=1,DeltaNI=0.2,diffTP=0.3,diffTC=-0.02,  #
                   bTP=c(2.423862,2.038216),bTC=c(2.796511,1.977432),          #
                   wTP=c(1,1),wTC=c(1,1),H0TP=1)                               #
# [1] 0.8769092                                                                #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Re-calculation of the second stage sample sizes for #############
############## a two-stage adaptive group sequential three-arm     #############
############## non-inferiority design based on conditional power   #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  - ThreeArmAdaptiveCP()                                  #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the second stage sample sizes required to obtain a specific       #
# conditional power based on the interim data. If the targeted conditional     #
# power cannot be obtained with the defined maximum overall sample size of the #
# second stage (maxN2), the obtained conditional power for the respective      #
# maximum sample sizes is calculated.                                          #
#                                                                              #
################################################################################

ThreeArmAdaptiveReCalcCP <- function(nT1,nC1,nP1,cC2,cP2,Power,maxN2,thetaTP,
                                     thetaTC,sigma,DeltaNI,diffTP,diffTC,bTP,
                                     bTC,wTP,wTC,HOTP) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT1         | float      | >0        | First stage sample size of the test   #
#             |            |           | group                                 #
#             |            |           |                                       #
# nC1         | float      | >0        | First stage sample size of the        #
#             |            |           | control group                         #
#             |            |           |                                       #
# nP1         | float      | >0        | First stage sample size of the        #
#             |            |           | placebo group                         #
#             |            |           |                                       #
# cC2         | float      | >0        | Relative control group size of the    #
#             |            |           | second stage (cC2=nC2/nT2)            #
#             |            |           |                                       #
# cP2         | float      | >0        | Relative placebo group size of the    #
#             |            |           | second stage (cP2=nP2/nT2)            #
#             |            |           |                                       #
# Power       | float      | >0 & <1   | Targeted conditional power            #
#             |            |           |                                       #
# maxN2       | float      | >0        | Maximum overall sample size of the    #
#             |            |           | second stage                          #
#             |            |           |                                       #
# thetaTP     | float      |           | Treatment difference between test and #
#             |            |           | placebo for which CP is calculated    #
#             |            |           |                                       #
# thetaTC     | float      |           | Treatment difference between test and #
#             |            |           | control for which CP is calculated    #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# diffTP      | float      |           | Observed treatment difference between #
#             |            |           | test and placebo at interim           #
#             |            |           |                                       #
# diffTC      | float      |           | Observed treatment difference between #
#             |            |           | test and control at interim           #
#             |            |           |                                       #
# bTP         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
# wTP         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TP^(s)             #
#             |            |           |                                       #
# wTC         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TC^(n)             #
#             |            |           |                                       #
# H0TP        | boolean    | 0 or 1    | Defines if the conditional power to   #
#             |            |           | reject both null hypotheses (H0TP=1)  #
#             |            |           | or only H_0,TC^(n) (H0TP=0) is        #
#             |            |           | calculated.                           #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT2         | float      | >0        | Second stage sample size of the test  #
#             |            |           | group to achieve conditional power CP #
#             |            |           |                                       #
# nC2         | float      | >0        | Second stage sample size of the       #
#             |            |           | control group to achieve conditional  #
#             |            |           | power CP                              #
#             |            |           |                                       #
# nP2         | float      | >0        | Second stage sample size of the       #
#             |            |           | placebo group to achieve conditional  #
#             |            |           | power CP (only given for H0TP=1!)     #
#             |            |           |                                       #
# CondPower   | float      | >0        | Actual conditional power              #
#             |            |           |                                       #
################################################################################

  if (H0TP==0) {
    # sample size recalculation based on conditional power to reject only
    # H_0,TC^(n) => placebo group size is irrelevant
    cP2 <- 0

    # test statistic of the first stage
    ZTC1 <- ((diffTC+DeltaNI)/sigma)*sqrt((nT1*nC1)/(nT1+nC1))

    # analytical solution for required second stage sample sizes is available
    nT2 <- min(maxN2/(1+cC2+cP2),
               sigma^2/(thetaTC+DeltaNI)^2*(1+cC2)/cC2*
               (sqrt(wTC[1]^2+wTC[2]^2)/wTC[2]*bTC[2]+qnorm(Power)
                -wTC[1]/wTC[2]*ZTC1)^2)
    nC2 <- cC2*nT2
    CondPower <- ThreeArmAdaptiveCP(nT=c(nT1,nT2),nC=c(nC1,nC2),nP=c(nP1,nP2),
                                    thetaTP=thetaTP,thetaTC=thetaTC,
                                    sigma=sigma,DeltaNI=DeltaNI,diffTP=diffTP,
                                    diffTC=diffTC,bTP=bTP,bTC=bTC,wTP=wTP,
                                    wTC=wTC,H0TP=H0TP)

    return(list(nT2=nT2,nC2=nC2,CondPower=CondPower))

  } else if (H0TP==1) {
    # sample size recalculation based on conditional power to reject both
    # null hypotheses

    # test statistics of the first stage
    ZTP1 <- (diffTP/sigma)*sqrt((nT1*nP1)/(nT1+nP1))
    ZTC1 <- ((diffTC+DeltaNI)/sigma)*sqrt((nT1*nC1)/(nT1+nC1))

    # new environment
    func.env <- new.env()

    # root finding function
    solvenT2 <- function(nT2) {
      assign("nC2",cC2*nT2,envir=func.env)
      assign("nP2",cP2*nT2,envir=func.env)
      assign("CondPower",ThreeArmAdaptiveCP(nT=c(nT1,nT2),
                                            nC=c(nC1,get("nC2",envir=func.env)),
                                            nP=c(nP1,get("nP2",envir=func.env)),
                                            thetaTP=thetaTP,thetaTC=thetaTC,
                                            sigma=sigma,DeltaNI=DeltaNI,
                                            diffTP=diffTP,diffTC=diffTC,
                                            bTP=bTP,bTC=bTC,wTP=wTP,
                                            wTC=wTC,H0TP=H0TP),envir=func.env)
      return(get("CondPower",envir=func.env)-Power)
    }

    # maximum overall sample size maxN2 provides conditional power less than
    # the targeted conditional power
    if (solvenT2(maxN2/(1+cC2+cP2)) < 0) {

      nT2 <- maxN2/(1+cC2+cP2)
      nC2 <- cC2*nT2
      nP2 <- cP2*nT2
      CondPower <- ThreeArmAdaptiveCP(nT=c(nT1,nT2),nC=c(nC1,nC2),nP=c(nP1,nP2),
                                      thetaTP=thetaTP,thetaTC=thetaTC,
                                      sigma=sigma,DeltaNI=DeltaNI,diffTP=diffTP,
                                      diffTC=diffTC,bTP=bTP,bTC=bTC,wTP=wTP,
                                      wTC=wTC,H0TP=H0TP)

    } else {

      # determine required sample sizes & conditional power
      nT2 <- uniroot(solvenT2,lower=1e-10,upper=maxN2/(1+cC2+cP2))$root

      nC2 <- get("nC2",envir=func.env)
      nP2 <- get("nP2",envir=func.env)
      CondPower <- get("CondPower",envir=func.env)

    }

    return(list(nT2=nT2,nC2=nC2,nP2=nP2,CondPower=CondPower))

  }

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmAdaptiveReCalcCP(nT1=275,nC1=275,nP1=69,cC2=1,cP2=0.5,Power=0.9,       #
                         maxN2=500,thetaTP=0.4,thetaTC=0,sigma=1,DeltaNI=0.2,  #
                         diffTP=0.2,diffTC=0,bTP=c(2.423862,2.038216),         #
                         bTC=c(2.796511,1.977432),wTP=c(1,1),wTC=c(1,1),H0TP=1)#
# $nT2                                                                         #
# [1] 185.4249                                                                 #
#                                                                              #
# $nC2                                                                         #
# [1] 185.4249                                                                 #
#                                                                              #
# $nP2                                                                         #
# [1] 92.71244                                                                 #
#                                                                              #
# $CondPower                                                                   #
# [1] 0.9                                                                      #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Re-calculation of the optimal second stage sample   #############
############## size for a two-stage adaptive group sequential      #############
############## three-arm non-inferiority design based on           #############
############## conditional power                                   #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  - ThreeArmAdaptiveCP()                                  #
# ------------------   - ThreeArmAdaptiveReCalcCP()                            #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# --------------                                                               #
# Calculates the optimal second stage sample sizes required to obtain a        #
# specific conditional power based on the interim data. Here, 'optimal' means  #
# minimizing the the overall second stage sample size. If the targeted         #
# conditional power cannot be obtained with the defined maximum overall sample #
# size of the second stage ('maxN2'), the function determines the second stage #
# sample sizes that sum up to 'maxN2' and give the highest conditional power.  #
#                                                                              #
################################################################################

ThreeArmAdaptiveOptReCalcCP <- function(nT1,nC1,nP1,Power,maxN2,thetaTP,
                                        thetaTC,sigma,DeltaNI,diffTP,diffTC,bTP,
                                        bTC,wTP,wTC,H0TP) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT1         | float      | >0        | First stage sample size of the test   #
#             |            |           | group                                 #
#             |            |           |                                       #
# nC1         | float      | >0        | First stage sample size of the        #
#             |            |           | control group                         #
#             |            |           |                                       #
# nP1         | float      | >0        | First stage sample size of the        #
#             |            |           | placebo group                         #
#             |            |           |                                       #
# Power       | float      | >0 & <1   | Targeted conditional power            #
#             |            |           |                                       #
# maxN2       | float      | >0        | Maximum overall sample size of the    #
#             |            |           | second stage                          #
#             |            |           |                                       #
# thetaTP     | float      |           | Treatment difference between test and #
#             |            |           | placebo for which CP is calculated    #
#             |            |           |                                       #
# thetaTC     | float      |           | Treatment difference between test and #
#             |            |           | control for which CP is calculated    #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# diffTP      | float      |           | Observed treatment difference between #
#             |            |           | test and placebo at interim           #
#             |            |           |                                       #
# diffTC      | float      |           | Observed treatment difference between #
#             |            |           | test and control at interim           #
#             |            |           |                                       #
# bTP         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
# wTP         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TP^(s)             #
#             |            |           |                                       #
# wTC         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TC^(n)             #
#             |            |           |                                       #
# H0TP        | boolean    | 0 or 1    | Defines if the conditional power to   #
#             |            |           | reject both null hypotheses (H0TP=1)  #
#             |            |           | or only H_0,TC^(n) (H0TP=0) is        #
#             |            |           | calculated.                           #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT2         | float      | >0        | Second stage sample size of the test  #
#             |            |           | group to achieve conditional power CP #
#             |            |           |                                       #
# nC2         | float      | >0        | Second stage sample size of the       #
#             |            |           | control group to achieve conditional  #
#             |            |           | power CP                              #
#             |            |           |                                       #
# nP2         | float      | >0        | Second stage sample size of the       #
#             |            |           | placebo group to achieve conditional  #
#             |            |           | power CP (only given for H0TP=1!)     #
#             |            |           |                                       #
# CondPower   | float      | >0        | Actual conditional power              #
#             |            |           |                                       #
################################################################################

  if (H0TP==0){
    # sample size recalculation based on conditional power to reject only
    # H_0,TC^(n) => placebo group size is irrelevant and the optimal design is
    # given as the balanced design (nT2=nC2)

    optdesign <- ThreeArmAdaptiveReCalcCP(nT1=nT1,nC1=nC1,nP1=nP1,cC2=1,cP2=1,
                                          Power=Power,maxN2=maxN2,
                                          thetaTP=thetaTP,thetaTC=thetaTC,
                                          sigma=sigma,DeltaNI=DeltaNI,
                                          diffTP=diffTP,diffTC=diffTC,bTP=bTP,
                                          bTC=bTC,wTP=wTP,wTC=wTC,H0TP=H0TP)
    return(optdesign)

  } else if (H0TP==1) {
    # sample size recalculation based on conditional power to reject both
    # null hypotheses

    # function that returns minus the conditional power for a specific
    # allocation and the maximum overall second stage sample size maxN2
    optmaxcC2cP2 <- function(alloc){
      -ThreeArmAdaptiveCP(nT=c(nT1,maxN2/(1+alloc[1]+alloc[2])),
                          nC=c(nC1,alloc[1]*maxN2/(1+alloc[1]+alloc[2])),
                          nP=c(nP1,alloc[2]*maxN2/(1+alloc[1]+alloc[2])),
                          thetaTP=thetaTP,thetaTC=thetaTC,
                          sigma=sigma,DeltaNI=DeltaNI,diffTP=diffTP,
                          diffTC=diffTC,bTP=bTP,bTC=bTC,wTP=wTP,
                          wTC=wTC,H0TP=H0TP)
    }
    # seach for the allocation that maximizes conditional power
    optmaxAlloc <- optim(c(0.5,0.5),optmaxcC2cP2,lower=rep(0.01,2),upper=c(3,3),
                         method="L-BFGS-B")

    if (-optmaxAlloc$value < Power) {
      # with the defined maximum overall second stage sample size maxN2 the
      # targeted conditional power cannot be obtained

      nT2 <- maxN2/(1+optmaxAlloc$par[1]+optmaxAlloc$par[2])
      nC2 <- optmaxAlloc$par[1]*nT2
      nP2 <- optmaxAlloc$par[2]*nT2
      CondPower <- -optmaxAlloc$value
      return(list(nT2=nT2,nC2=nC2,nP2=nP2,CondPower=CondPower))

    } else {
      # the targeted conditional power can be obtained. thus, search for the
      # optimal second stage sample sizes that minimize the overall second
      # stage sample size

      # new environment
      func.env <- new.env()
      # function that returns the overall second stage sample size to obtain
      # the targeted conditional power for a specific allocation
      optcC2cP2 <- function(alloc){
        assign("optDesign",
               ThreeArmAdaptiveReCalcCP(nT1=nT1,nC1=nC1,nP1=nP1,cC2=alloc[1],
                                        cP2=alloc[2],Power=Power,maxN2=1e10,
                                        thetaTP=thetaTP,thetaTC=thetaTC,
                                        sigma=sigma,DeltaNI=DeltaNI,
                                        diffTP=diffTP,diffTC=diffTC,bTP=bTP,
                                        bTC=bTC,wTP=wTP,wTC=wTC,H0TP=H0TP),
               envir=func.env)
        return(get("optDesign",envir=func.env)$nT2+
               get("optDesign",envir=func.env)$nC2+
               get("optDesign",envir=func.env)$nP2)
      }
      # search for the optimal allocation that minimizes the overall second
      # stage sample size
      optim(optmaxAlloc$par,optcC2cP2,lower=rep(0.01,2),upper=c(3,3),
            method="L-BFGS-B")

      return(get("optDesign",envir=func.env))

    }

  }

}


################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmAdaptiveOptReCalcCP(nT1=275,nC1=275,nP1=69,Power=0.9,maxN2=500,        #
                            thetaTP=0.4,thetaTC=0,sigma=1,DeltaNI=0.2,         #
                            diffTP=0.2,diffTC=0,bTP=c(2.423862,2.038216),      #
                            bTC=c(2.796511,1.977432),wTP=c(1,1),wTC=c(1,1),    #
                            H0TP=1)                                            #
# $nT2                                                                         #
# [1] 183.6715                                                                 #
#                                                                              #
# $nC2                                                                         #
# [1] 158.5484                                                                 #
#                                                                              #
# $nP2                                                                         #
# [1] 110.8326                                                                 #
#                                                                              #
# $CondPower                                                                   #
# [1] 0.9                                                                      #
}                                                                              #
################################################################################
```

```
################################################################################
##############                                                     #############
############## Calculation of the predictive power for a two-stage #############
############## adaptive group sequential three-arm non-inferiority #############
############## design                                              #############
##############                                                     #############
################################################################################
#                                                                              #
# Author:       Patrick Schlömer                                               #
# Last update:  13/May/2014                                                    #
#                                                                              #
################################################################################
#                                                                              #
# REQUIRED PACKAGES:   - mnormt                                                #
# -----------------                                                            #
#                                                                              #
# REQUIRED FUNCTIONS:  NONE                                                    #
# ------------------                                                           #
#                                                                              #
################################################################################
#                                                                              #
# THIS FUNCTION:                                                               #
# -------------                                                                #
# Calculates the predictive power given the interim data to reject either      #
# both null hypotheses or only H_0,TC^(n) at the final analysis. The prior     #
# distributions for the treatment effects of the test, control and placebo     #
# group are assumed to be mutually independent and normal with known variance. #
#                                                                              #
################################################################################

ThreeArmAdaptivePP <- function(nT,nC,nP,sigma,DeltaNI,muT0,muC0,muP0,sigmaT0,
                               sigmaC0,sigmaP0,meanT,meanC,meanP,bTP,bTC,wTP,
                               wTC,H0TP) {

################################################################################
#                                                                              #
# INPUT-PARAMETERS:                                                            #
# -----------------                                                            #
#______________________________________________________________________________#
#             |            |           |                                       #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                           #
#_____________|____________|___________|_______________________________________#
#             |            |           |                                       #
# nT          | 2x1 vector | >0        | Stage-wise sample sizes of the test   #
#             | (floats)   |           | group                                 #
#             |            |           |                                       #
# nC          | 2x1 vector | >0        | Stage-wise sample sizes of the        #
#             | (floats)   |           | control group                         #
#             |            |           |                                       #
# nP          | 2x1 vector | >0        | Stage-wise sample sizes of the        #
#             | (floats)   |           | placebo group                         #
#             |            |           |                                       #
# sigma       | float      | >0        | Common standard deviation             #
#             |            |           |                                       #
# DeltaNI     | float      | >0        | Non-inferiority margin                #
#             |            |           |                                       #
# muT0        | float      |           | Mean of the prior distribution for    #
#             |            |           | the test treatment effect             #
#             |            |           |                                       #
# muC0        | float      |           | Mean of the prior distribution for    #
#             |            |           | the control treatment effect          #
#             |            |           |                                       #
# muP0        | float      |           | Mean of the prior distribution for    #
#             |            |           | the placebo effect                    #
#             |            |           |                                       #
# sigmaT0     | float      | >0        | Standard deviation of the prior       #
#             |            |           | distribution for the test treatment   #
#             |            |           | effect (sigmaT0=Inf for a             #
#             |            |           | non-informative prior)                #
#             |            |           |                                       #
# sigmaC0     | float      | >0        | Standard deviation of the prior       #
#             |            |           | distribution for the control          #
#             |            |           | treatment effect (sigmaC0=Inf for a   #
#             |            |           | non-informative prior)                #
#             |            |           |                                       #
# sigmaP0     | float      | >0        | Standard deviation of the prior       #
#             |            |           | distribution for the placebo effect   #
#             |            |           | (sigmaP0=Inf for a non-informative    #
#             |            |           | prior)                                #
#             |            |           |                                       #
# meanT       | float      |           | Observed treatment effect of the test #
#             |            |           | group at interim                      #
#             |            |           |                                       #
# meanC       | float      |           | Observed treatment effect of the      #
#             |            |           | control group at interim              #
#             |            |           |                                       #
# meanP       | float      |           | Observed treatment effect of the      #
#             |            |           | placebo group at interim              #
#             |            |           |                                       #
# bTP         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TP^(s)                            #
#             |            |           |                                       #
# bTC         | 2x1 vector |           | Stage-wise rejection boundaries for   #
#             | (floats)   |           | H_0,TC^(n)                            #
#             |            |           |                                       #
# wTP         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TP^(s)             #
#             |            |           |                                       #
# wTC         | 2x1 vector | >0        | Weights used for combining the test   #
#             | (floats)   |           | statistics for H_0,TC^(n)             #
#             |            |           |                                       #
# H0TP        | boolean    | 0 or 1    | Defines if the conditional power to   #
#             |            |           | reject both null hypotheses (H0TP=1)  #
#             |            |           | or only H_0,TC^(n) (H0TP=0) is        #
#             |            |           | calculated.                           #
#             |            |           |                                       #
#----------------------------------------------------------------------------- #
#                                                                              #
# OUTPUT-PARAMETERS:                                                           #
# ------------------                                                           #
#______________________________________________________________________________#
```

```
#             |            |           |                                      #
# VARIABLE    | FORMAT     | RANGE     | DESCRIPTION                          #
#_____________|____________|___________|______________________________________#
#             |            |           |                                      #
# PP          | float      | >0 & <1   | Predictive power                     #
#             |            |           |                                      #
###############################################################################

  if (H0TP==0) {
    # predictive power to reject H_0,TC^(n)

    # test statistic of the first stage
    ZTC1 <- ((meanT-meanC+DeltaNI)/sigma)*sqrt((nT[1]*nC[1])/(nT[1]+nC[1]))

    # posterior means and standard deviations for the test and control
    # treatment effect
    muTpost <- (1/sigmaT0^2*muT0+nT[1]/sigma^2*meanT)/
               (1/sigmaT0^2+nT[1]/sigma^2)
    muCpost <- (1/sigmaC0^2*muC0+nC[1]/sigma^2*meanC)/
               (1/sigmaC0^2+nC[1]/sigma^2)
    sigmaTpost <- sqrt(1/(1/sigmaT0^2+nT[1]/sigma^2))
    sigmaCpost <- sqrt(1/(1/sigmaC0^2+nC[1]/sigma^2))

    # 'posterior' test statistic of the second stage
    ZTC2post <- (muTpost-muCpost+DeltaNI)/sigma*
                sqrt((nT[2]*nC[2])/(nT[2]+nC[2]))

    PP <- pnorm(sqrt(sigma^2/(sigmaTpost^2+sigmaCpost^2)/
                     (sigma^2/(sigmaTpost^2+sigmaCpost^2)+
                      (nT[2]*nC[2])/(nT[2]+nC[2])))*
                (ZTC2post+wTC[1]/wTC[2]*ZTC1-
                 sqrt(wTC[1]^2+wTC[2]^2)/wTC[2]*bTC[2]))

  } else if (H0TP==1) {
    # predictive power to reject both H_0,TP^(s) and H_0,TC^(n)

    # posterior means and standard deviations for the test, control and placebo
    # treatment effect
    muTpost <- (1/sigmaT0^2*muT0+nT[1]/sigma^2*meanT)/
               (1/sigmaT0^2+nT[1]/sigma^2)
    muCpost <- (1/sigmaC0^2*muC0+nC[1]/sigma^2*meanC)/
               (1/sigmaC0^2+nC[1]/sigma^2)
    muPpost <- (1/sigmaP0^2*muP0+nP[1]/sigma^2*meanP)/
               (1/sigmaP0^2+nP[1]/sigma^2)
    sigmaTpost <- sqrt(1/(1/sigmaT0^2+nT[1]/sigma^2))
    sigmaCpost <- sqrt(1/(1/sigmaC0^2+nC[1]/sigma^2))
    sigmaPpost <- sqrt(1/(1/sigmaP0^2+nP[1]/sigma^2))

    # covariance matrix of the joint posterior predictive distribution of the
    # standardised second stage test statsitics
    rho <- (1+sigmaTpost^2/sigma^2*nT[2])/
           sqrt((1+(sigmaTpost^2+sigmaPpost^2)/sigma^2*
                 (nT[2]*nP[2])/(nT[2]+nP[2]))*
                (1+(sigmaTpost^2+sigmaCpost^2)/sigma^2*
                 (nT[2]*nC[2])/(nT[2]+nC[2])))*
           sqrt((nC[2]*nP[2])/((nT[2]+nC[2])*(nT[2]+nP[2])))
    varcov <- matrix(c(1,rho,rho,1),ncol=2)

    # upper limits for integration
    upper <- c(((muTpost-muPpost)/sigma*sqrt((nT[2]*nP[2])/(nT[2]+nP[2]))+
                wTP[1]/wTP[2]*(meanT-meanP)/sigma*
                sqrt((nT[1]*nP[1])/(nT[1]+nP[1]))-
                sqrt(wTP[1]^2+wTP[2]^2)/wTP[2]*bTP[2])/
               sqrt(1+(sigmaTpost^2+sigmaPpost^2)/sigma^2*
                      (nT[2]*nP[2])/(nT[2]+nP[2])),
               ((muTpost-muCpost+DeltaNI)/sigma*
                sqrt((nT[2]*nC[2])/(nT[2]+nC[2]))+
                wTC[1]/wTC[2]*(meanT-meanC+DeltaNI)/sigma*
                sqrt((nT[1]*nC[1])/(nT[1]+nC[1]))-
                sqrt(wTC[1]^2+wTC[2]^2)/wTC[2]*bTC[2])/
               sqrt(1+(sigmaTpost^2+sigmaCpost^2)/sigma^2*
                      (nT[2]*nC[2])/(nT[2]+nC[2]))
    )

    PP <- sadmvn(lower=rep(-Inf,2),upper=upper,mean=rep(0,2),varcov=varcov)[1]

  }

  return(PP)

}

################################################################################
# EXAMPLE:                                                                     #
# --------                                                                     #
if (FALSE){                                                                    #
ThreeArmAdaptivePP(nT=rep(275,2),nC=rep(275,2),nP=rep(69,2),sigma=1,           #
                   DeltaNI=0.2,muT0=0,muC0=0,muP0=0,sigmaT0=Inf,               #
                   sigmaC0=Inf,sigmaP0=Inf,meanT=2.4,meanC=2.42,               #
                   meanP=2.1,bTP=c(2.423862,2.038216),bTC=c(2.796511,1.977432),#
                   wTP=c(1,1),wTC=c(1,1),H0TP=1)                               #
# [1] 0.7504354                                                                #
}                                                                              #
################################################################################
```

BIBLIOGRAPHY

Anderson, K. (2013). *gsDesign: Group Sequential Design.* R package version 2.7-06.

Armitage, P. (1993). Interim analyses in clinical trials. In Hoppe, F. M., editor, *Multiple Comparisons, Selection, and Applications in Biometry*, pages 391–402. Marcel Dekker, New York.

Armitage, P., McPherson, C., and Rowe, B. C. (1969). Repeated significance tests on accumultating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244.

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

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4):1029–1041.

Bauer, P. and König, F. (2006). The reassessment of trial prespectives from interim data – a critical view. *Statistics in Medicine*, 25(1):23–36.

Betensky, R. A. (1997). Early stopping to accept H_0 based on conditional power: Approximations and comparisons. *Biometrics*, 53(3):794–806.

Blackwelder, W. C. (1982). “Proving the null hypothesis” in clinical trials. *Controlled Clinical Trials*, 3:345–353.

Brannath, W., Gutjahr, G., and Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association*, 107(498):824–832.

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(6):1190–1208.

CHMP (2005a). *Guideline on clinical investigation of medicinal products indicated for the treatment of panic disorder.* London.

CHMP (2005b). *Guideline on the choice of the non-inferiority margin.* London.

CHMP (2007a). *Guideline on clinical investigation of medicinal products for the treatment of migraine.* London.

CHMP (2007b). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.* London.

CHMP (2012). *Concept paper on the need for a guideline on multiplicity issues in clinical trials.* London.

CPMP (2000). *Points to consider on switching between superiority and non-inferiority.* London.

CPMP (2002). *Points to consider on multiplicity issues in clinical trials.* London.

CPMP (2003). *Note for guidance on the clinical investigation of medicinal products indicated in the treatment of asthma.* London.

Cui, L., Hung, H. M. J., and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 55(3):853–857.

D'Agostino, R. B., Massaro, J. M., and Sullivan, L. M. (2003). Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine*, 22(2):169–186.

Dallow, N. and Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, 10(4):311–317.

Dette, H., Trampisch, M., and Hothorn, L. A. (2009). Robust designs in noninferiority three-arm clinical trials with presence of heteroscedasticity. *Statistics in Biopharmaceutical Research*, 1(3):268–278.

Dmitrienko, A. and Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*, 6(3):171–180.

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2009). *Multiple Testing Problems in Pharmaceutical Statistics.* Chapman & Hall/CRC, New York.

Dodge, H. F. and Romig, H. G. (1929). A method of sampling inspection. *Bell System Technical Journal*, 8(4):613–631.

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

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):482–491.

Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, 87(417):162–170.

Elfring, G. L. and Schultz, J. R. (1973). Group sequential designs for clinical trials. *Biometrics*, 29(3):471–477.

Everson-Stewart, S. and Emerson, S. S. (2010). Bio-creep in non-inferiority clinical trials. *Statistics in Medicine*, 29(27):2769–2780.

FDA (2010a). *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics – Draft Guidance.*

FDA (2010b). *Guidance for Industry: Non-Inferiority Clinical Trials – Draft Guidance.*

Fleming, T. R., Harrington, D. P., and O'Brien, P. C. (1984). Designs for group sequential tests. *Controlled Clinical Trials*, 5(4):348–361.

Gamalo, M. A., Muthukumarana, S., Ghosh, P., and Tiwari, R. C. (2011). A generalized p-value approach for assessing noninferiority in a three-arm trial. *Statistical Methods in Medical Research.* [Epub ahead of print].

Genz, A. and Azzalani, A. (2012). *mnormt: The multivariate normal and t distribution.* R package version 1.4-5.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2012). *mvtnorm: Multivariate Normal and t Distributions.* R package version 0.9-9994.

Gosh, P., Nathoo, F. S., Gönen, M., and Tiwari, R. C. (2011). Assessing noninferiority in a three-arm trial using the Bayesian approach. *Statistics in Medicine*, 30(15):1795–1808.

Grechanovsky, E. and Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference*, 76(1–2):79–91.

Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. *Biometrical Journal*, 50(5):678–692.

Hartung, J. and Knapp, G. (2009). Adaptive controlled noninferiority group sequential trials. *TU Dortmund - Eldorado.*

Hasler, M., Vonk, R., and Hothorn, L. A. (2008). Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity. *Statistics in Medicine*, 27(4):490–503.

Hauschke, D. and Pigeot, I. (2005a). Establishing efficacy of a new experimental treatment in the 'gold standard' design. *Biometrical Journal*, 47(6):782–786.

Hauschke, D. and Pigeot, I. (2005b). Rejoinder to "Establishing efficacy of a new experimental treatment in the 'gold standard' design"'. *Biometrical Journal*, 47(6):797–798.

Hida, E. and Tango, T. (2011a). On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Statistics in Medicine*, 30(3):224–231.

Hida, E. and Tango, T. (2011b). Response to Joachim Röhmel and Iris Pigeot. *Statistics in Medicine*, 30(26):3165–3165.

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

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

Holland, B. S. and DiPonzio Copenhaver, M. (1987). An improved sequentially rejective bonferroni test procedure. *Biometrics*, 43(2):417–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, London.

Hung, H. M. J. (2005). Discussion on “Establishing efficacy of a new experimental treatment in the ’gold standard’ design”. *Biometrical Journal*, 47(6):795–796.

Hung, H. M. J., Wang, S.-J., and O’Neill, R. (2009). Challenges and regulatory experiences with non-inferiority trial design without placebo arm. *Biometrical Journal*, 51(2):324–334.

Hwang, I. K., Shih, W. J., and De Cani, J. S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9(12):1439–1445.

ICH Expert Working Group (1998). *Statistical Prinicples for Clinical Trials E9*.

ICH Expert Working Group (2000). *Choice of Control Group and Related Issues in Clinical Trials E10*.

Jennison, C. and Turnbull, B. W. (1991). Exact calculations for sequential t, χ^2 and F tests. *Biometrika*, 78(1):133–141.

Jennison, C. and Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC, New York.

Julious, S. A. and Wang, S.-J. (2008). How biased are indirect comparisons, particularly when comparisons are made over time in controlled trials? *Drug Information Journal*, 42:625–633.

Kieser, M. and Friede, T. (2007). Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine*, 26(2):253–273.

Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, 74(1):149–154.

Koch, A. (2005). Discussion on “Establishing efficacy of a new experimental treatment in the ’gold standard’ design”. *Biometrical Journal*, 47(6):792–793.

Koch, A. and Röhmel, J. (2004). Hypothesis testing in the ’gold standard’ design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of Biopharmaceutical Statistics*, 14(2):315–325.

Koch, G. G. and Tangen, C. M. (1999). Nonparametric analysis of covariance and its role in noninferiority clinical trials. *Drug Information Journal*, 33(4):1145–1159.

Kwong, K. S., Cheung, S. H., Hayter, A. J., and Wen, M.-J. (2012). Extension of three-arm noninferiority studies to trials with multiple new treatments. *Statistics in Medicine*, 31(24):2833–2843.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trial. *Communications in Statistics. Part C: Sequential Analysis*, 1(3):207–219.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Lewis, J. A. (2005). Discussion on “Establishing efficacy of a new experimental treatment in the ’gold standard’ design”. *Biometrical Journal*, 47(6):787–789.

Li, G. and Gao, S. (2010). A group sequential type design for three-arm non-inferiority trials with binary endpoints. *Biometrical Journal*, 52(4):504–518.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660.

Maurer, W., Hothorn, L. A., and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. *Vollmar J (Ed): Biometrie in der chemisch-pharmazeutischen Industrie.*

Mehrotra, D. V. (2005). Discussion on “Establishing efficacy of a new experimental treatment in the ’gold standard’ design”. *Biometrical Journal*, 47(6):794–794.

Mielke, M., Munk, A., and Schacht, A. (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Statistics in Medicine*, 27(25):5093–5110.

Mosteller, F. and Bush, R. (1954). Selected quantitative techniques. In Lindzey, G. and Aronson, E., editors, *Handbook of Social Psychology*, volume 1, pages 289–334. Addison-Wesley, Camebridge.

Munzel, U. (2009). Nonparametric non-inferiority analyses in the three-arm design with active control and placebo. *Statistics in Medicine*, 28(29):3643–3656.

Naik, U. (1975). Some selection rules for comparing p processes with a standard. *Communication in Statistics*, Series A(6):519–535.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, 7:308–313.

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

O'Neill, R. T. (2006). FDA's critical path initiative: A perspective on contributions of biostatistics. *Biometrical Journal*, 48(4):559–564.

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419.

Pigeot, I., Schäfer, J., Röhmel, J., and Hauschke, D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, 22(6):883–899.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics*, 38(1):153–162.

Popper Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831.

Proschan, M. A., Follmann, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48(4):1131–1143.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51(4):1315–1324.

Proschan, M. A., Lan, K. K. G., and Turk Wittes, J. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Statistics for Biology and Health. Springer, New York.

PubMed (accessed on May 12, 2014). *Bethesda (MD): National Library of Medicine (US)*. http://www.ncbi.nlm.nih.gov/pubmed/.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Röhmel, J. (1998). Therapeutic equivalence investigations: Statistical considerations. *Statistics in Medicine*, 17(15-16):1703–1714.

Röhmel, J. (2005a). Discussion on "Establishing efficacy of a new experimental treatment in the 'gold standard' design". *Biometrical Journal*, 47(6):790–791.

Röhmel, J. (2005b). On confidence bounds for the ratio of net differences in the "gold standard" design with reference, experimental, and placebo treatment. *Biometrical Journal*, 47(6):799–806.

Röhmel, J. and Pigeot, I. (2010). A comparison of multiple testing procedures for the gold standard non-inferiority trial. *Journal of Biopharmaceutical Statistics*, 20(5):911–926.

Röhmel, J. and Pigeot, I. (2011). Statistical strategies for the analysis of clinical trials with an experimental treatment, an active control and placebo, and a prespecified fixed non-inferiority margin for the difference in means. *Statistics in Medicine*, 30(26):3162–3164.

Sanchez, M. M. and Chen, X. (2006). Choosing the analysis population in non-inferiority studies: Per protocol or intent-to-treat. *Statistics in Medicine*, 25(7):1169–1181.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC.

Schlömer, P. and Brannath, W. (2013). Group sequential designs for three-arm 'gold standard' non-inferiority trials with fixed margin. *Statistics in Medicine*, 32(28):4875–4889.

Schwartz, T. A. and Denne, J. S. (2006). A two-stage sample size recalculation procedure for placebo- and active-controlled non-inferiority trials. *Statistics in Medicine*, 25(19):3396–3406.

Slud, E. and Wei, J. T. (1982). Two-sample repeated significance tests based on the modified wilcoxon statistic. *Journal of the American Statistical Association*, 77(380):862–868.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7(1):8–17.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3):243–258.

Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine*, 27(24):4914–4927.

Stucke, K. and Kieser, M. (2012). A general approach for sample size calculation for the three-arm 'gold standard' non-inferiority design. *Statistics in Medicine*, 31(28):3579–3596.

Tang, M.-L. and Tang, N.-S. (2004). Tests of noninferiority via rate difference for three-arm clinical trials with placebo. *Journal of Biopharmaceutical Statistics*, 14(2):337–347.

Vandemeulebroecke, M. (2006). An investigation of two-stage tests. *Statistica Sinica*, 16(3):933–951.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–200.

Wassmer, G. (1999). Group sequential monitoring with arbitrary inspection times. *Biometrical Journal*, 41(2):197–216.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons, Inc., New York.

Wiens, B. L. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2(3):211–215.

Wiens, B. L. and Zhao, W. (2007). The role of intention to treat in analysis of noninferiority studies. *Clinical Trials*, 4:286–291.

Yoo, B. (2010). Impact of missing data on type 1 error rates in non-inferiority trials. *Pharmaceutical Statistics*, 9:87–99.