Adaptive designs for clinical trials in cardiovascular diseases

Dissertation

zur Erlangung des humanwissenschaftlichen Doktorgrades

in der Medizin

der Georg-August-Universität Göttingen

vorgelegt von

Tobias MÜTZE

aus Bietigheim-Bissingen

Göttingen, 2018

Supervisor:	Prof. Dr. Tim Friede Institut für Medizinische Statistik Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Second Committee Member:	Prof. Dr. Heike Bickeböller Institut für Genetische Epidemiologie Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Third Committee Member:	Prof. Dr. Markus Zabel Klinik für Kardiologie und Pneumologie Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Day of Disputation:	13.07.2018

Declaration of Authorship

I, Tobias MÜTZE, declare that this dissertation titled, "Adaptive designs for clinical trials in cardiovascular diseases" and the work presented in it are my own and that it was written independently with no other sources and aids than quoted.

Abstract

Cardiovascular diseases are diseases of the heart and blood vessels constituting a major cause of death and disability worldwide. Cardiovascular drug development aims to deliver efficacious drugs to address the public health burden of cardiovascular diseases. However, the high costs associated with cardiovascular drug development, for example due to long-running clinical trials, sometimes including thousands of patients, place a high burden on the development of new efficacious treatments for cardiovascular diseases. Proposals for improving the efficiency of cardiovascular drug development include better disease characterization, more defined target populations, and the use of adaptive clinical trial designs. This dissertation focuses on adaptive clinical trial designs for cardiovascular research.

Adaptive clinical trial designs, commonly referred to as adaptive designs, are clinical trial designs with a preplanned modification of design aspects, under some constraints such as preserving integrity and validity of the trials, based on interim data of the ongoing trial. Design aspects which are commonly modified include the sample size, number of doses or treatments, or endpoints. Adaptive designs offer flexibility compared to traditional clinical trials with a fixed design to accommodate newly gained information. However, with the flexibility comes an increased statistical complexity, as adaptive designs require an increased effort to control the probability that the clinical trial declares efficacy of an inefficacious treatment, that is the type I error rate, and to plan the number of patients required such that an efficacious treatment is detected with a high statistical power.

The focus of this dissertation is on two types of adaptive designs: group sequential designs and designs with a nuisance parameter based sample size re-estimation. In group sequential designs, the efficacy of a treatment is tested repeatedly during the conduct of the trial and the trial is stopped early if efficacy of the treatment can be shown with statistical significance. Thus, an efficacious treatment can be detected early in clinical trials with a group sequential design. In designs with a nuisance parameter based sample size re-estimation, the final sample size is adjusted using estimates of the potentially several nuisance parameters based on interim data. Nuisance parameters are for example the outcome variance in trials with continuous outcomes and the overall event rate in trials with count outcomes. The nuisance parameter based sample size re-estimation aims to assure that a clinical trial achieves the target power independently of the initially planned sample size.

The first objective of this dissertation is to study group sequential designs with recurrent events, motivated by clinical trials with patients suffering from chronic heart failure. In clinical trials with patients suffering from chronic heart failure, a common clinical relevant recurrent event outcome is the number of heart failure hospitalizations, which can also be part of a composite endpoint in combination

with cardiovascular death. To model heart failure hospitalizations and the respective composite, a negative binomial model and a more robust semiparametric model have been proposed in the literature. However, group sequential designs have not been studied for these models. Therefore, I propose statistical methods for planning and analyzing group sequential designs for negative binomial models and more robust semiparametric models and study their asymptotic properties. Moreover, I show that the proposed planning and analysis methods result in an appropriate power and type I error rate, respectively, for parameter combinations common in clinical trials with patients suffering from chronic heart failure. I put a particular focus on the longitudinal nature of the recurrent events, i.e., a single subject can experience new events throughout the trial, and its consequential on the group sequential designs. The longitudinal natures of the outcomes distinguishes group sequential designs with recurrent events from group sequential designs for other common models, such a continuous, binary, or survival data.

A second objective of this dissertation is to study nuisance parameter based sample size re-estimation in three-arm trials with normal outcomes; an investigation motivated by clinical trials with patients suffering from hypertension. A common endpoint in these trials modeled as normally distributed is the change of blood pressure between the baseline measurement and the end of the trial. I show that the ideas for nuisance parameter based sample size re-estimation in two-arm trials can be adapted to three-arm trials and highlight that the corresponding approaches do not result in the desired target power. Furthermore, I modify one of the sample size re-estimation procedures such that it results in appropriately powered three-arm clinical trials.

The third objective of this dissertation is to study incorporating prior information on the variance into the nuisance parameter based sample size re-estimation in two-arm trials with normal outcomes. This objective, too, is motivated by clinical trials with patients suffering from hypertension. I propose several ad hoc rules for incorporating prior information into the sample size re-estimation and by means of Monte Carlo simulation studies I show that the incorporation of prior information can reduce the variability of the final sample size when no prior-data conflict is present. However, I illustrate that in the presence of a prior-data conflict, the designs with a sample size re-estimation incorporating prior information do not convey the target power. I also highlight that common approaches of robustifying the prior information cannot completely mitigate the negative effects of a prior-data conflict without also nullifying the benefits of incorporating prior information on the nuisance parameter into the sample size re-estimation.

Acknowledgements

During my work on this dissertation I was fortunate to have had the support of various people who warrant a special mention.

First and foremost I would like to express my gratitude to my advisor Professor Tim Friede. Without his encouragement, I might not have pursued my PhD. I am grateful for his input and his interest in my research, for his time to discuss ideas, for sharing his vast experience in and knowledge of clinical trials, and for providing the funding for my work. All this enabled me to focus wholeheartedly on my research and it made pursuing my PhD an enjoyable learning experience.

I would also like to thank the other members of my thesis committee: Prof. Heike Bickeböller and Prof. Markus Zabel, for their input.

My sincere thanks goes to Dr. Ekkehard Glimm and Dr. Heinz Schmidli for being co-authors on my publications and for supervising my internships at Novartis Pharma AG in Basel. Their open-door policy and willingness to discuss every aspect of my research made my time in Basel exceptionally productive.

Besides my advisors, I would like to thank the staff and my fellow PhD students at the Department of Medical Statistics for providing a productive and friendly working environment. In particular, I would like to thank Christian, David, Markus, and Roland for the fruitful after-lunch discussions as well as Burak and Cynthia for creating a productive working environment in our shared offices on the top floor.

For proofreading this thesis and helpful remarks about the linguistics, I thank Clarissa and Christian.

Last but not least, I would like to express my gratitude to my parents, Clarissa, and my friends for their love and support.

Contents

1 Introduction

1.1 Clinical trials in cardiovascular drug development

Cardiovascular disease (CVD) is a collective term for diseases of the heart and blood vessels which includes, among others, coronary artery disease, cerebrovascular disease, congenital heart disease, and rheumatic heart disease [1]. Cardiovascular diseases are the "leading causes of death and disability in the world" [1, p. 3] with an estimated 17.7 million cardiovascular deaths in 2015, which corresponds to 31% of all deaths worldwide [2]. Moreover, the number of cardiovascular deaths is expected to keep rising to about 24 million by the year 2030 [3]. Even though the burden of cardiovascular diseases on the public health is well recognized, there are concerns within the cardiovascular research community about the decreasing chance of successful clinical trials and a slowdown of investment and interest in developing drugs for cardiovascular diseases [4, 5]. As one reason for unsuccessful clinical trials in cardiovascular drug development, Jackson et al. [5] name the unspecific target population used in the clinical trials; unspecific in the sense that the target population suffers from various cardiovascular diseases. The slowdown of investment in cardiovascular drug development is attributed to the high costs associated with cardiovascular drug development [5, 6]. The high costs are due to the logistical and regulatory requirements as well as the large number of patients required in cardiovascular clinical trials. Reasons for the large number of patients required are the generally small probability for the occurrence of informative events, such as stroke and heart failure, in clinical trials and that the clinical trial participants are often already receiving treatment and thus the experimental drug will only lead to incremental improvements [5]. Both Jackson et al. [5] and Fordyce et al. [6] made several proposals to overcome the difficulties of current drug development for cardiovascular diseases, and among the proposed solutions are the application of adaptive clinical trial designs and the use of genetic markers to identify subpopulations. In the following, I provide a general introduction to adaptive clinical trial designs.

1.2 Adaptive clinical trial designs

Important design aspects of a randomized controlled trial include, but are not limited to, the doses of the experimental treatment, the primary efficacy variable, the sample size, the eligibility criteria, and the type of the control (active or placebo) [7, Chapter 3]. Historically, the main design aspects of a clinical trial were not altered during the course of the trial. However, if during the trial it becomes evident that vital design aspects were based on incorrect assumptions, the continued conduct of the trial may be unethical. For example, a clinical trial is conducted unethically

when patients are exposed unnecessarily to evidently ineffective treatments. An incorrectly planned trial can also have limited scientific merit because, for instance, the primary efficacy variable is inappropriate to measure the treatment effect or the sample size is too small to detect an existing clinically relevant treatment effect with sufficient statistical power. The risk of improperly conducting a clinical trial due to incorrect assumptions in the planning phase can be mitigated by modifying the clinical trial design during the course of the trial based on the accruing data. Clinical trial designs preplanned with the option to modify the design aspects based on interim data are called *adaptive clinical trial designs* or *adaptive designs* in short. Examples for adaptive designs include group sequential designs, sample size re-estimation designs, drop-the-loser designs, adaptive dose finding designs, enrichment designs, and adaptive seamless phase II/III trial designs. While each of the mentioned adaptive designs have benefits compared to non-adaptive designs, the individual goals of the designs differ. For instance, a sample size re-estimation design aims to assure that an existing treatment effect is detected with the desired statistical power by adjusting the final sample size based on interim data. A group sequential design allows for the early discontinuation of a clinical trial for reasons of efficacy or futility based on interim data. For a more comprehensive list and a detailed discussion of the mentioned adaptive designs, we refer to Chow and Chang [8] and Kairalla et al. [9].

Adaptive designs are discussed from the regulatory perspective by the Food and Drug Administration (FDA) [10] and by the European Medicines Agency (EMA) [11]. In their draft guidance document, the FDA defines a clinical trial with an adaptive design as a trial "that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from patients in the study [...]" [10, Section III.A.]. Most importantly from a statistical point of view, the FDA guidance document [10] requires that the type I error rate is controlled and that the bias in estimating the treatment effect is minimized. Moreover, in the FDA guidance document, adaptive designs are separated into well understood and less well understood designs. A clinical trial design is characterized as well understood if the drug development community generally agrees that this design improves the efficiency of a trial while maintaining the trials validity with respect to biases and interpretability. Well understood adaptations include but are not limited to blinded adaptations (i.e., adaptations in which the treatment indicators are not revealed), group sequential methods, and futility stopping. Less well understood adaptations include sample size adaptations based on effect estimates, endpoint selection based on treatment effect estimates, and adaptations in non-inferiority trials. In its reflection paper on adaptive clinical trial designs, the EMA defined an adaptive design as a design whose "[...] statistical methodology allows the modification of a design element [...] at an interim analysis with full control of the type I error." Thus, the EMA already includes the requirement for type I error rate control in its definition of adaptive designs. Furthermore, the EMA reflection paper also highlights the importance of an awareness concerning biases in adaptive designs and the need for confidence intervals which achieve the target coverage probability. Additionally, the EMA reflection paper provides a detailed discussion of selected adaptive designs. In conclusion, the

main concern of regulatory agencies about statistical properties of adaptive designs is the type I error rate control, the reduction of the treatment effect estimate bias, and the coverage probability of confidence intervals.

The perspective of statisticians working in the pharmaceutical industry on adaptive clinical trial designs was highlighted in two major publications from two different expert groups. The first publication was based on a workshop sponsored by the *Statisticians in the Pharmaceutical Industry* (PSI) [12]. Phillips and Keene [12] summarize the points made during the workshop for a number of adaptive designs, which included sample size re-estimation designs, designs that drop or add treatment arms, and designs that change one or more of the following: the primary endpoint, the patient population, the objectives, and the statistical methodology. The second major publication is the executive summary of a white paper from an expert group of the *Pharmaceutical Research and Manufacturers of America* (PhRMA) [13]. In the executive summary, Gallo et al. [13] list the statistical, logistical, and procedural issues of adaptive designs and discuss the benefits of adaptive dose finding trials, seamless phase II/III designs, and sample size re-estimation designs. To summarize, these two publications highlight an agreement among statisticians working in the pharmaceutical industry about the general benefits of adaptive clinical trial designs and also emphasize the importance of type I error control, bias reduction, and the desire to maintain blinding in adaptive designs.

The general regulatory acceptance of blinded clinical trial adaptation warrants a more in-depth discussion of such designs. These are adaptations in which the treatment indicator of the patients remains masked. The most common blinded adaptation is a blinded sample size review in which the sample size is re-estimated based on a blind estimate of the nuisance parameter [14]. Blinded sample size re-estimation methods have been studied for various designs and endpoints, including normally distributed outcomes where the outcome variance is the nuisance parameter [15], binary data where the overall response rate is the nuisance parameter [16], and count data where the overall event rate is the nuisance parameter [17]. Further possible blinded adaptations are the adjustment of the study duration in clinical trials with time-to-event data based on the overall number of events and, as mentioned in the ICH guideline E9 [18], the adaptation of data transformations and the change of the parametric or nonparametric analysis method. Aside from its regulatory acceptance and desired statistical properties, such as type I error rate control, blinded adaptations do not necessarily require an independent data monitoring committee (DMC) [19].

Extensive literature exists on the statistical considerations for various adaptive clinical trial designs. For instance, Bauer et al. [20] do not only review the methodological work for a number of adaptive designs, but also give a recapitulation of the history of adaptive designs, summarize the industry and regulatory perspectives, discuss several clinical trial applications, and list software available for adaptive designs. An in-depth review of adaptive tests for adaptive designs with a single hypothesis and for designs with multiple hypotheses in the context of planning and analyzing a confirmatory clinical trial was published by Bretz et al. [21]. Bauer and Einfalt [22] and Hatfield et al. [23] quantify the instances of practical applications of adaptive designs in clinical trials.

1.3 Research questions

In the research for this dissertation, I developed adaptive designs and with this I contributed to one of the measures recommended for improving clinical trials in cardiovascular drug development [5, 6]. The focus of my research was on developing two types of adaptive designs, namely group sequential designs and blinded sample size re-estimation designs, since both belong to classes of designs which (a) are accepted by the FDA [10, Chapter V] and (b) are the most commonly applied [5, 23]. In detail, I developed group sequential designs for recurrent events, blinded sample size re-estimation designs for three-arm trials with normal outcomes, and designs in which prior information on a nuisance parameter is incorporated into the nuisance parameter based sample size re-estimation. The three designs are considered separately from each other, as they each deal with different aspects of clinical trials in cardiovascular drug development. In the following, I outline in detail my motivation for researching the aforementioned three adaptive designs.

1.3.1 Group sequential designs for recurrent events

An event is characterized as recurrent when it can be observed repeatedly for a single subject. For example, for patients suffering from chronic heart failure, recurrent heart failure hospitalizations, often as part of a composite endpoint with cardiovascular death, are modeled as recurrent events. In general, recurrent events and their analyses are becoming increasingly important in clinical trials in cardiovascular drug development as the focus shifts away from analyzing only the time to the first event of a subject to analyzing all events of a subject [24, 25]. In group sequential designs the efficacy and the futility of a treatment are assessed through interim analyses while the trial is ongoing. As cardiovascular clinical trials are often conducted over the course of many years [6], group sequential designs in the cardiovascular drug development have the potential to accelerate patient access to treatments and stop trials early if the studied treatment is futile. An example of an ongoing cardiovascular clinical trial with a primary endpoint modeled as a recurrent event and with an interim analysis is the Paragon-HF trial (ClinicalTrials.gov identifier: NCT01920711) [26]. The Paragon-HF trial includes patients suffering from chronic heart failure and the primary endpoint is the composite of cardiovascular death and recurrent heart failure hospitalizations [26]. Rogers et al. [27] compared several models for analyzing recurrent events in clinical trials in chronic heart failure and among the models are the negative binomial model and the Andersen-Gill model with a robust variance estimator (also known as the Lin-Wei-Yang-Ying (LWYY) model [28]). Moreover, in the Paragon-HF trial, the primary analysis is planned to be conducted using the LWYY model [26]. Therefore, I developed group sequential designs for the negative binomial and the LWYY models.

Group sequential designs for recurrent events are distinct from group sequential designs with other outcomes, for instance normal outcomes or time-to-event outcomes, in that one subject can contribute new events to multiple interim analyses. As a consequence, it is not guaranteed that the standard property of group sequential designs holds for group sequential designs with recurrent events. The standard property of group sequential designs is that the joint distribution of the test statistics

from different interim analyses has the canonical form [29]. The canonical form is of interest as a property for the test statistics in group sequential designs because the majority of theoretical considerations for group sequential designs rely on the canonical joint distribution. Aside from that, standard statistical software and standard group sequential software packages generally only provide support for designs with the canonical joint distribution property. Group sequential designs for recurrent events have already been studied for some nonparametric and semiparametric models [30–32] and for these models the joint distribution of the statistics does not have the canonical form. Prior to my research, group sequential designs for the negative binomial model and the LWYY model had not been studied constituting a lack of knowledge about group sequential designs for practically relevant recurrent event models. Thus, my work closed this gap.

The goal of my research was to develop group sequential designs for the negative binomial and the LWYY models, and to assess their performance with respect to type I error rate control and power. In particular, I focused on characterizing the joint distribution of the test statistics from different interim analyses and on studying how this joint distribution relates to the canonical joint distribution, the standard approach in group sequential designs.

1.3.2 Blinded sample size re-estimation in three-arm trials

Hypertension is recognized as a risk factor for cardiovascular diseases [33]. According to the Committee for Medicinal Products for Human Use (CHMP) [34], a recommended endpoint for clinical trials in the field of hypertension is the change in diastolic blood pressure in the weeks after the baseline measurement; an endpoint generally modeled as normally distributed. For this indication, Krum et al. [35] and Elliott et al. [36] published multi-arm trials. The trial published by Krum et al. [35] includes several arms with doses from the experimental treatment as well as an active control and a placebo control. The trial published by Elliott et al. [36] is a three-arm trial comparing two active treatments and a combination of them against each other.

The sample size of a clinical trial is generally planned such that a clinically relevant treatment effect is detected with a prespecified power. The sample size required to achieve a prespecified power depends, among other variables, on the outcome variance. Thus, when planning the sample size of a clinical trial with normal outcomes, assumptions on the outcome variance have to be made. These assumptions are often made based on prior information. For instance, prior information can be available from earlier trials in the same drug development program or from trials with a patient population and an outcome measure similar to the ones of the future trial. However, prior information is not always available or reliable and incorrectly made assumptions on the variance during the sample size planning of a clinical trial can result in an inadequate sample size and therefore in an over- or underpowered clinical trial. Blinded sample size re-estimation adjusts the final sample size mid-trial based on a blinded estimation of the outcome variance and as such aims to ensure that the desired power of the clinical trial is attained irrespectively of the initially assumed variance. While blinded sample size re-estimation has been studied in detail for two-arm trials with normal outcomes [15, 37, 38], blinded sample size

re-estimation for multi-arm trials received very little attention despite its practical relevance. A notable exception is the work by Kieser and Friede [39], who studied an F-test for the equality of means instead of a pairwise comparison scenario, the latter being the more common in clinical trials. Therefore, I developed a blinded sample size re-estimation procedure for a multi-arm trial design that achieves the target power. As the multi-arm trial design, I considered a three-arm trial design with an arm for an experimental treatment, one arm for an active control, and one arm for a placebo control. This design is known as the 'gold standard' design [40].

1.3.3 Incorporating prior information into the sample size re-estimation

As outlined in Chapter 1.3.2, when planning the sample size of a clinical trial, prior information on the outcome variance is regularly available and used to inform the choice of the sample size. Schmidli, Neuenschwander, and Friede [41] formalized incorporating prior information on the variance into the sample size planning of a clinical trial based on meta-analytic-predictive (MAP) priors in the case of normally distributed outcomes. A MAP prior is a prior on a parameter of a future trial which is obtained through a meta-analysis of historical data [42]. When the initial sample size of a clinical trial is planned utilizing prior information, it seems desirable to coherently incorporate prior information also into the sample size re-estimation. However, literature on incorporating prior information on the nuisance parameter into the sample size re-estimation is very sparse. Gould [43] considered incorporating prior information on the overall response rate into the sample size re-estimation in a binomial trial but did not assess the operating characteristics of such a procedure. More recently, Hartley [44] studied incorporating prior information on both the nuisance parameter and the effect into the sample size re-estimation for normal outcomes. To the best of my knowledge, the concept of incorporating prior information on the nuisance parameter into the nuisance parameter based sample size re-estimation is studied in detail for the first time as part of this dissertation. I focused on two-arm clinical trials with a normally distributed endpoint that are planned and analyzed using the frequentist Student's t-test. In these trials, the outcome variance is the nuisance parameter and in I assumed that the prior information on the variance is available through a MAP prior. The goal of my research in this area was to identify different methods of incorporating prior information on the outcome variance into the sample size re-estimation and then to compare the operating characteristics of the resulting re-estimation procedures. The operating characteristics of interest are the power, the final sample size distribution, and the type I error rate. During my research, I put particular emphasis on the robustness of the sample size re-estimation procedures concerning prior-data conflicts, which did not receive any attention in the existing literature [43, 44]. A prior-data conflict is given when the outcome variance observed in the clinical trial does not match the prior information.

A cardiovascular drug development scenario for which incorporating prior information into the sample size planning of a new clinical trial is plausible are clinical trials assessing the efficacy of interventions for blood pressure control in patients with hypertension. For example, Glynn et al. [45] summarized the published trials about interventions – such as self-monitoring, education of patients or health care provider, improvement of delivery of care, etc. – for blood pressure control in

a meta-analysis. Due to the large amount of available information in this setting, clinical trials studying interventions for blood pressure control in patients with hypertension provide a plausible example for the incorporation of prior information into the sample size re-estimation.

1.4 Outline

In this dissertation I addresses the research questions outlined in Chapter 1.3. The results of my research were published or accepted for publication as research articles in peer-reviewed journals [46–49]. In Chapter 2, I present a summary of the published results. Chapter 2 is split into three parts with each part dedicated to my research addressing one of the research questions outlined in Chapter 1.3. In Chapter 3, I critically discuss the proposed statistical models and clinical trial designs, and provide an outlook into future research concerning adaptive designs for clinical trials in the cardiovascular drug development.

2 Proposed adaptive designs for cardiovascular drug development

2.1 Group sequential designs for recurrent events

The results of my research on group sequential designs for recurrent events were published in [48, 49]. The following summary of the main results is split into two parts. In the first part, group sequential designs for recurrent events modeled by a negative binomial distribution are presented and in the second part, group sequential designs for recurrent events modeled by the LWYY model are discussed. Before, I recapitulate the basic statistical concepts for group sequential designs, confer Jennison and Turnbull [29, Chapter 3], and introduce the relevant notation.

The focus is on group sequential designs for two-arm randomized controlled clinical trials assessing the efficacy of an experimental treatment in comparison with a control. Let the parameter θ be the efficacy parameter which is defined such that smaller values correspond to a more efficacious treatment. Then, in clinical trials with a group sequential design, the efficacy of the experimental treatment is assessed while the trial is ongoing during *data looks* by testing of the statistical hypotheses

$$H_0 : \theta \geq 0 \quad \text{versus} \quad H_1 : \theta < 0.$$

In group sequential testing, generally the same test statistic as in the corresponding fixed sample design is applied. However, to ensure that the type I error rate α is controlled even when the null hypothesis is tested repeatedly, the critical values for the test decision at the data looks in a group sequential design are different than in the corresponding fixed sample design. Let $k = 1, \ldots, K$ be the variable indexing the data looks and let T_k be the test statistic for testing the null hypothesis H_0 at data look k defined such that H_0 is rejected if T_k is smaller than or equal to a critical value c_k. Then, to calculate the critical values c_k $(k = 1, \ldots, K)$, the joint distribution of the test statistics $T_1, \ldots, T_K$ must be known. The basic principal of the general statistical methodology for group sequential designs is based on that the test statistics follow the canonical joint distribution under the null hypothesis H_0 . In detail, the sequence $T_1, \ldots, T_K$ of test statistics follows the canonical joint distribution if it is multivariate normally distributed with

$$\begin{aligned} \mathbb{E}[T_k] &= \theta\sqrt{\mathcal{I}_k}, \quad k = 1, \ldots, K, \\ \text{Cov}\,(T_{k_1}, T_{k_2}) &= \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}, \quad 1 \leq k_1 \leq k_2 \leq K. \end{aligned}$$

Here, $\mathcal{I}_k$ $(k = 1, \ldots, K)$ are the information levels. There are various approaches to determine the critical values for the test decision at a data look. I focused on the

error spending approach [29, Chapter 7], which I outline next for the calculation of efficacy boundaries when no mandatory futility boundaries are set. The error spending approach allocates the global type I error rate α to the K data looks by means of an error spending function. An error spending function $f : [0, \infty) \to [0, 1]$ is a non-decreasing function with $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$ and the type I error rate to be spent at data look $k = 1, \ldots, K$ is defined by

$$\begin{aligned} \pi_1 &= f(\nu_1), \\ \pi_k &= f(\nu_k) - f(\nu_{k-1}), \quad k = 2, \ldots, K. \end{aligned} \tag{2.1}$$

Here, ν_k with $\nu_1 < \ldots < \nu_K$ is an increasing measure of the clinical trial's progress which is zero at the beginning of the trial and one at the time of the last planned data look. Common definitions for ν_k are based on the calendar time, the observed information level, or the number of observed events. Then, the critical value c_k for the test decision at data look k is defined through π_k. The critical value c_1 at the first data look is simply the π_1-quantile q_{π_1} of a standard normal distribution. The critical values for the subsequent data looks $k = 2, \ldots, K$ are calculated by solving equation

$$\pi_k = \mathbb{P}_{H_0} \left(T_1 > c_1, \ldots, T_{k-1} > c_{k-1}, T_k \leq c_k \right) \tag{2.2}$$

under the assumption of a canonical joint distribution for the sequence $T_1, \ldots, T_k$. Since the critical values are calculated under the null hypothesis, that is $\theta = 0$, the expected value of the canonical joint distribution is set to zero when solving (2.2). A group sequential testing procedure with the critical values determined as defined above through the error spending approach results in a global type I error rate of α. In other words, the probability to wrongfully reject the null hypothesis at one of the data looks is equal to α.

It is worthwhile to reflect on the role of the canonical joint distribution in group sequential testing. When a sequence of test statistics in a group sequential design follows the canonical joint distribution, the covariance between the test statistics can be determined easily. Moreover, for the canonical joint distribution, the probability in (2.2) has a computational complexity which is linear in k instead of exponential in k as it is common for k-dimensional integrals, confer Jennison and Turnbull [29, Chapter 19.2]. Group sequential designs can of course also be conducted with test statistics not following the canonical joint distribution. In these cases, the critical values are again determined by solving (2.2), but using the actual distribution of the sequence of test statistics instead of a canonical joint distribution. Since most practically relevant statistical models, such as models with normal outcomes, binary outcomes as well as parametric survival models and the Cox proportional hazards model, fulfill the canonical joint distribution, the critical value calculation for designs without canonical joint distribution is generally not part of standard software packages.

I conclude the introduction into group sequential designs with outlining the determination of the maximum information required in a group sequential design to achieve the desired power. The maximum information $\mathcal{I}_{max}$ is the information level at which the last data look is performed when the trial is not stopped at an earlier data look, i.e., $\mathcal{I}_K = \mathcal{I}_{max}$. As before, the focus is on designs without futility boundaries. For a group sequential design with a maximum of K data looks and a sequence

$c_1, \ldots, c_K$ of critical values, the power for a parameter $\theta^* < 0$ is given by

$$\text{Power} = 1 - \mathbb{P}_{\theta^*}\left(T_1 > c_1, \ldots, T_K > c_K\right). \tag{2.3}$$

For planning purposes, the information fractions $w_k = \mathcal{I}_k/\mathcal{I}_{max}$ with $\mathcal{I}_1 < \mathcal{I}_2 < \ldots < \mathcal{I}_{\max}$, at which the *k-th* data look is performed, are prespecified. With the prespecified information fractions, the expected values and the covariances of the test statistics can be written as

$$\begin{aligned}\mathbb{E}[T_k] &= \theta^* w_k \mathcal{I}_{max}, \quad k = 1, \ldots, K,\\ \text{Cov}\left(T_{k_1}, T_{k_2}\right) &= \sqrt{w_{k_1}/w_{k_2}}\mathcal{I}_{max}, \quad 1 \leq k_1 \leq k_2 \leq K.\end{aligned}$$

Thus, the only remaining unknown variable is the maximum information, which is then determined by solving (2.3) for a given power. Based on the maximum information, other designs parameters such as the sample size, the calendar time of the data looks, or the individual follow-up time in trials with time-to-event and recurrent event endpoints can be determined.

2.1.1 Group sequential designs for negative binomial outcomes

An extended version of the following discourse on group sequential designs for negative binomial outcomes was published by Mütze et al. [48]. Let $j = 1, \ldots, n_i$ index the subjects in treatment group $i = 1, 2$. Moreover, let t_{ijk} be the time since randomization of a subject at data look k and let Y_{ijk} be the number of events a subject experiences between randomization and data look k. The time since randomization is also known as exposure time. To model the recurrent event for a subject, I assume that each subject has a subject specific event rate $\lambda_{ij} > 0$ and that conditional on this event rate, the events of a subject arise from a homogeneous Poisson process. From this it follows that, conditional on the rate λ_{ij}, the number of events Y_{ijk} follows a Poisson distribution with mean $t_{ijk}\lambda_{ij}$, i.e., $Y_{ijk}|\lambda_{ij} \sim \text{Pois}(t_{ijk}\lambda_{ij})$. Then, I modeled the between-patient heterogeneity of the event rates through a Gamma distribution by assuming that the rates λ_{ij} are independent Gamma distributed,

$$\lambda_{ij} \sim \Gamma\left(\frac{1}{\phi}, \frac{1}{\phi\mu_i}\right).$$

Thus, the accumulated number of events Y_{ijk} at each data look follows a negative binomial distribution with mean $t_{ijk}\mu_i$ and shape parameter ϕ, i.e., $Y_{ijk} \sim \text{NB}\left(t_{ijk}\mu_i, \phi\right)$. The negative binomial distribution $\text{NB}\left(\mu, \phi\right)$ is parameterized such that the expected value is equal to μ and the variance is equal to $\mu(1 + \mu\phi)$. The statistical hypothesis testing problem for the efficacy assessment is given by

$$H_0 : \frac{\mu_1}{\mu_2} \geq 1 \quad \text{versus} \quad H_1 : \frac{\mu_1}{\mu_2} < 1.$$

Let $\beta_i = \log(\mu_i)$ denote the log-rate and let $\hat{\beta}_{ik}$ be the maximum likelihood estimator of the log-rate obtained with the data available at data look $k = 1, \ldots, K$. Then, the null hypothesis H_0 can be tested in a group sequential design with negative binomial

outcomes at the data looks $k = 1, \ldots, K$ through the sequence of Wald statistics $T_1, \ldots, T_K$ with

$$T_k = \left(\hat{\beta}_{1k} - \hat{\beta}_{2k}\right) \sqrt{\hat{\mathcal{I}}_k}, \quad k = 1, \ldots, K.$$

Here, $\hat{\mathcal{I}}_k$ is the maximum likelihood estimator for the information level $\mathcal{I}_k$, which is defined through the Fisher information $I_{\beta_i}^{(k)}$ of the log-rates β_i $(i = 1, 2)$ at data look k, i.e.,

$$\begin{aligned} \mathcal{I}_k &= \frac{1}{\frac{1}{I_{\beta_1}^{(k)}} + \frac{1}{I_{\beta_2}^{(k)}}}, \\ I_{\beta_i}^{(k)} &= \sum_{j=1}^{n_i} \frac{t_{ijk} \exp(\beta_i)}{1 + \phi t_{ijk} \exp(\beta_i)}. \end{aligned} \tag{2.4}$$

To define the critical values, the joint distribution of the sequence of test statistics has to be determined. Scharfstein, Tsiatis, and Robins [50] proved that asymptotically the sequence of Wald statistics in a parametric group sequential model converges to the canonical joint distribution. As shown by Mütze et al. [48], the results of Scharfstein, Tsiatis, and Robins [50] can be applied for the negative binomial model even though the negative binomial data in the current model are of longitudinal nature. The negative binomial outcomes are of longitudinal nature in this model because the outcomes of a subject are accumulated over time and as such a subject can experience new events after a data look and, therefore, have a different number of events for different data looks. Since the results of Scharfstein, Tsiatis, and Robins [50] can be applied, the sequence of Wald statistics $T_1, \ldots, T_K$ in the negative binomial model follows asymptotically the canonical joint distribution and the critical values can be calculated by solving (2.2). The information level (2.4) and thus the covariance of the canonical joint distribution for negative binomial outcomes depends on the unknown log-rates and the shape parameter. Therefore, the critical values cannot be calculated prior to the trial, but at each data look, the respective critical value is calculated based on the estimated covariance. Mütze et al. [48] discussed in detail that the resulting group sequential procedure controls the type I error rate asymptotically.

The maximum information for group sequential designs with negative binomial outcomes can be determined as outlined above by solving (2.3). The sample size and the study duration can be planned by equating the maximum information with the information level (2.4) and solving the resulting equation in the sample size and the study duration. To that end, a shape parameter ϕ, log-rates β_1 and β_2, and the accrual process have to be assumed.

The canonical joint distribution only holds asymptotically for the sequence of Wald statistics in group sequential designs with negative binomial outcomes. Therefore, I studied the finite sample size operating characteristics, in particular the type I error rate, of the proposed group sequential procedure by means of Monte Carlo simulation studies. The choice of the parameters for the simulation study of the type I error rate was motivated by the results for the endpoint 'heart failure hospitalizations' in the CHARM-Preserved trial published by Yusuf et al. [51]. The parameters are listed in Table 2.1. It is worth noting that a uniform recruitment within

TABLE 2.1: Parameters considered in the simulation study of the type I error rate in group sequential designs with negative binomial outcomes.

Parameter	**Value**
One-sided significance level α	0.025
Shape parameter ϕ	2, 3, 4, 5
Data looks K	2, 3, 5
Individual follow-up [years]	2.75–4
Recruitment period [years]	1.25
Study duration [years]	4
Maximum sample sizes $n_1 = n_2$	800, 1100, 1400, 1700
Annualized rates $\mu_1 = \mu_2$	0.08, 0.1, 0.12, 0.14

the recruitment period of 1.25 years is assumed for the simulations and that once a subject is randomized, the subject is followed up until the study ends after four years. Thus, the individual follow-up times vary between 2.75 and 4 years. In the simulation study, the *k-th* data look is performed at the calendar time at which the information level $k\mathcal{I}_{\max}/K$ $(k = 1, \ldots, K)$ is attained under the assumed parameter vector $(\mu_1, \mu_2, \phi, n_1, K)$ and the described uniform recruitment. The type I error rate is allocated through the Pocock-type error spending function and the O'Brien-Fleming-type error spending function [52]. In comparison, the Pocock-type error spending function results in larger type I error rate spending during data looks early in the trial, while the O'Brien-Fleming-type error spending function spends more during data looks later in the trial. The simulated type I error rates of the proposed group sequential procedure for negative binomial outcomes are presented in Figure 2.1. Each simulated type I error rate is based on 50 000 Monte Carlo replications. Figure 2.1 shows no practically relevant deviation of the type I error rate from the target level $\alpha = 0.025$. The number of simulated type I error rates outside of the error boundaries, depicted as grey lines, corresponds to what is expected for boundaries defined through two times the simulation error. A simulation study of the power of the proposed group sequential design, which is not reported here, showed that the general method for calculating the maximum information for a group sequential design through solving (2.3) leads to appropriately powered clinical trials with negative binomial outcomes.

In a simulation study of the type I error rate for additional parameter combinations with smaller sample sizes of fewer than 300 subjects per treatment arm, I showed that the proposed group sequential procedure for negative binomial outcomes can have an inflated type I error rate. Therefore, I proposed two modified group sequential procedures for negative binomial outcomes, which have an improved type I error rate control compared to the initial procedure. In the following, I explain the main idea of the two modified procedures. For the first procedure, the Wald statistic for the hypothesis test as well as the distribution used to calculate the critical values when solving (2.2) are modified. The modified Wald statistic uses a variance estimator obtained under the null hypothesis H_0, i.e., the information level estimator $\hat{\mathcal{I}}_k$ for the test statistic is calculated with parameter estimators

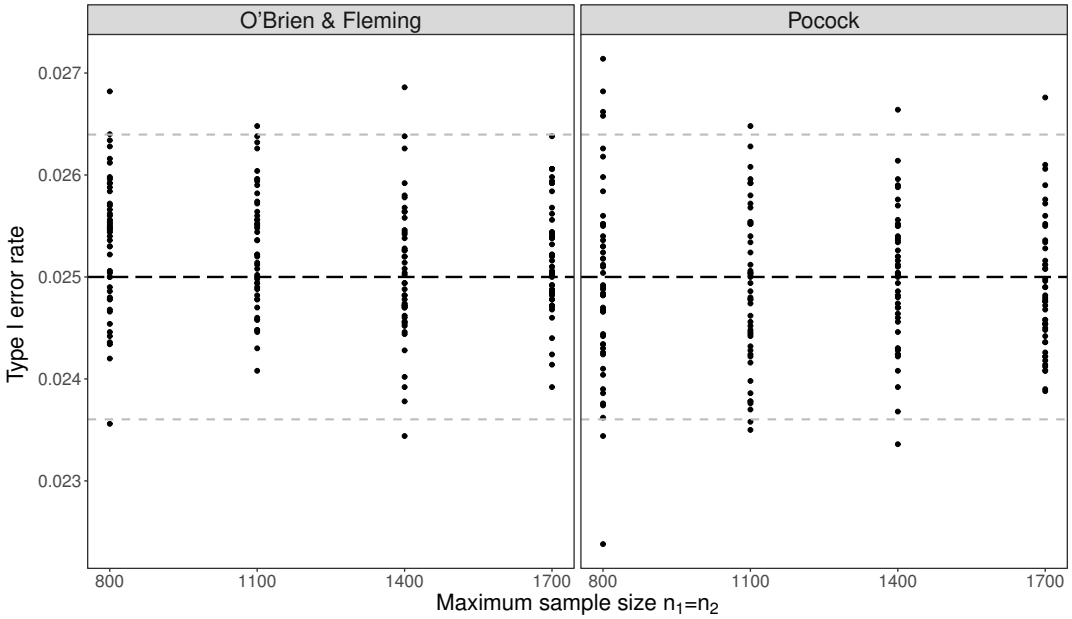


FIGURE 2.1: Simulated type I error rates of the Wald test for negative binomial outcomes in group sequential designs with O'Brien-Fleming-type and Pocock-type error spending functions. The black line depicts the planned one-sided type I error rate of $\alpha = 0.025$. The horizontal grey lines mark $\alpha \pm 2SE$ with SE the simulation error at a simulated type I error rate of 0.025.

obtained from maximizing the likelihood function over the parameter space of the null hypothesis H_0. Moreover, when solving (2.2) to determine the critical values for the data looks, a multivariate t-distribution was considered instead of a multivariate normal distribution. The expected value and the structure of the covariance matrix for the multivariate t-distribution are chosen identical to the canonical joint distribution. The degrees of freedom of the multivariate t-distribution were chosen in a conservative manner as the number of subjects recruited at the first data look. The resulting group sequential procedure with the two modifications clearly improves type I error rate control compared to the initial group sequential procedure for negative binomial outcomes when the sample size is small. The second modified procedure is based on the permutation distribution of the Wald statistics, which itself is not modified compared to the initial group sequential procedure. Let c_k be the critical value at data look k calculated under the assumption of the canonical joint distribution through solving (2.2) and let T_k be the Wald statistic at data look k. Furthermore, let $F(\cdot)$ be the permutation distribution of the Wald statistic at data look k, i.e., the cumulative distribution function obtained when calculating the Wald statistic for every permutation of the data vector at data look k. Then, the initial group sequential procedure is modified by using the transformed critical value $F^{-1}(\Phi(c_k))$ with $\Phi(\cdot)$ the cumulative distribution function of a standard normal distribution. This modified procedure also results in an improved type I error rate control compared to the initial group sequential procedure. However, both modified group sequential procedures still result in some type I error rate inflation when the sample

size is smaller than 100 subjects per treatment arm and, additionally, the overdispersion is large. However, the type I error rate inflation of the modified procedures is considerably smaller than the inflation for the initial group sequential procedure for negative binomial outcomes when the sample sizes are small.

In summary, I outlined the theoretical justification that the canonical joint distribution holds asymptotically for the sequence of Wald statistics in group sequential designs for negative binomial outcomes. Based on this asymptotic property, I proposed a group sequential procedure for negative binomial outcomes which calculates the critical values based on the canonical joint distribution. By means of Monte Carlo simulation studies, I exemplified that the proposed group sequential procedure controls the type I error rate for parameter combinations which are typical for the number of heart failure hospitalizations in clinical trials in chronic heart failure. Moreover, I showed that the proposed group sequential procedure exhibits some type I error rate inflation for small sample sizes and subsequently modified the initial group sequential procedure to achieve a better type I error rate control for small sample sizes. In conclusion, the proposed group sequential procedure for negative binomial outcomes or one of the recommended modifications control the type I error rate sufficiently for being applied in a wide range of practical situations, in particular in clinical trials with patients suffering from chronic heart failure.

For planning group sequential designs with negative binomial outcomes, I implemented the R package *gscounts*, which is available at the Comprehensive R Archive Network (CRAN) [53].

2.1.2 Group sequential designs with the LWYY model

In the following, I summarize the results of my research concerning group sequential designs with the LWYY model which were published by Mütze et al. [49]. Let r_{ij} be the randomization time of subject j in treatment group i and let c_{ij} denote its censoring time. Two time scales are distinguished: the study time s, that is the time since randomization, and the calendar time t, that is time since the start of the trial. Distinguishing the two time scales is important because the treatments are compared on the study time scale, for instance event rates are compared based on the time since randomization. However, when planning the timing of data looks in group sequential designs the calendar time and the closely connected information time are relevant. The subject specific increment $dN_{ij}(s)$ is equal to one if and only if the subject has an event at the study time s. Otherwise, the increment function is equal to zero. For each subject, the indicator function $Y_{ij}(s,t)$ is defined such that it is one if and only if a subject is at risk for experiencing an event at a given study time s and calendar time t, i.e.,

$$Y_{ij}(s,t) = \begin{cases} 1 & \text{if } r_{ij} + s \leq \min(t, c_{ij}) \\ 0 & \text{otherwise} \end{cases}.$$

In other words, the indication function $Y_{ij}(s,t)$ provides a connection between the calendar time and a subjects study time. Let the treatment indicator x_i be zero for group $i = 1$, $x_1 = 0$, and one for group $i = 2$, $x_2 = 1$. Lin et al. [28] proposed a robust semiparametric model for recurrent events which in the case of no covariates other

than the treatment is given by

$$\mathbb{E}\left[dN_{ij}(s)|Y_{ij}(s,t)=1,x_i\right]=\exp(x_i\beta)d\mu_0(s).$$

Here, $\mu_0(s)$ is an unknown nonnegative continuous function. I refer to this model as the LWYY model. With the assumption that smaller mean rates for the recurrent event process correspond to a more efficacious treatment, superiority of treatment $i=2$ over treatment $i=1$ can be formulated as the statistical hypothesis testing problem

$$H_0:\beta\geq 0 \quad \text{versus} \quad H_1:\beta<0.$$

Next, I outline the basics for statistical interference in the LWYY model. For more details, I refer to Lin et al. [28]. The parameter β is estimated based on the partial likelihood score function, which is given by

$$U(t,\beta)=\sum_{i=1}^{2}\sum_{j=1}^{n_i}\int_0^t Y_{ij}(s,t)W_i(s,t,\beta)dN_{ij}(s)$$

with

$$W_i(s,t,\beta)=x_i-\frac{\sum_{j=1}^{n_2}Y_{2j}(s,t)\exp(\beta)}{\sum_{j=1}^{n_1}Y_{1j}(s,t)+\sum_{j=1}^{n_2}Y_{2j}(s,t)\exp(\beta)}.$$

Then, let $\hat{\beta}(t)$ be the parameter estimator for β at a given calendar time t, which is obtained by solving equation $U(t,\beta)=0$ in β. Estimator $\hat{\beta}(t)$ is not to be confused with an estimator of a time varying effect. Here, the parameter β does not depend on the calendar time and by writing the estimator as a function of calendar time t, it is highlighted, that the estimator $\hat{\beta}(t)$ for β is determined at calendar time t. Let β_0 be the true parameter. The parameter estimator $\hat{\beta}(t)$ is asymptotically normally distributed in the sense that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}(t)-\beta_0\right)&\overset{\mathcal{D}}{\sim}\mathcal{N}\left(0,\frac{\mathcal{B}(t,\beta_0)}{\mathcal{A}(t,\beta_0)^2}\right)\qquad\text{with}\\ \mathcal{B}(t,\beta_0)&=n^{-1}\mathbb{E}\left[U(t,\beta_0)^2\right],\\ \mathcal{A}(t,\beta_0)&=n^{-1}\mathbb{E}\left[-\frac{\partial U}{\partial\beta}(t,\beta_0)\right].\end{aligned}$$

Based on the asymptotic variance of the parameter estimator, the information level $\mathcal{I}(t,\beta_0)$ at calendar time t is defined by

$$\mathcal{I}(t,\beta_0)=n\frac{\mathcal{A}(t,\beta_0)^2}{\mathcal{B}(t,\beta_0)}.$$

The terms $\mathcal{A}(t,\beta_0)$ and $\mathcal{B}(t,\beta_0)$ can be estimated consistently and, while I do not provide any details about the estimators in this summary, I denote the consistent estimators by $\hat{A}(t)$ and $\hat{B}(t)$, respectively. Therefore, $\hat{\mathcal{I}}(t)=n\hat{A}(t)^2/\hat{B}(t)$ is a consistent estimator for the information level $\mathcal{I}(t,\beta_0)$ at calendar time t. Based on the

mentioned properties of the parameter estimator and the information level estimator, the Wald statistic

$$T(t) = \hat{\beta}(t)\sqrt{\hat{\mathcal{I}}(t)}$$

is asymptotically standard normally distributed under the null hypothesis, that is for $\beta_0 = 0$, at calendar time t. Thus, an asymptotic level α test for the null hypothesis H_0 can be defined based on the Wald statistic. Therefore, in my research of group sequential designs for the LWYY model, I focused on group sequential testing using the Wald statistic $T(t)$.

Next, I outline the joint distribution of the test statistics from different data looks required to define a group sequential procedure for the LWYY model. Let K be the maximum number of data looks performed at calendar times $t_1 < \ldots < t_K$ through the Wald statistics $T(t_1), \ldots, T(t_K)$. As outlined by Mütze et al. [49] based on results of Lin et al. [28], the joint distribution of the Wald statistics is a multivariate normal distribution with the pairwise covariances

$$\begin{aligned} \text{Cov}\left(T(t_l), T(t_m)\right) &= \frac{\mathcal{B}(t_l, t_m, \beta_0)}{\sqrt{\mathcal{B}(t_l, \beta_0)\mathcal{B}(t_m, \beta_0)}},\ t_l < t_m, \qquad \text{with} \\ \mathcal{B}(t_l, t_m, \beta_0) &= n^{-1}\mathbb{E}\left[U(t_l, \beta_0)U(t_m, \beta_0)\right]. \end{aligned} \tag{2.5}$$

The expected value of the limiting multivariate normal distribution is zero under the null hypothesis $H_0 : \beta = 0$. The covariance structure (2.5) is different from the covariance structure of the canonical joint distribution. In other words, the canonical joint distribution does not hold for group sequential designs with the LWYY model. An asymptotically consistent group sequential procedure for the LWYY model, that is a procedures which maintains a global type I error rate α asymptotically, must calculate the critical values by solving (2.2) under the assumption of a multivariate normal distribution with covariance structure (2.5). Since the covariance matrix is not known, it has to be estimated consistently at every data look to calculate the critical values. For details about estimating the covariances and about calculating the critical values based on the estimated covariance, I refer to Mütze et al. [49].

After I had proposed a consistent group sequential procedure for the LWYY model, two questions arose. The first question was whether the consistent group sequential procedure for the LWYY model controls the type I error rate for finite sample sizes and practically relevant parameter combinations. The second question was how the type I error rate of the group sequential procedure which assumes the canonical joint distribution is affected by the violation of the canonical joint distribution assumption in the LWYY model. This procedure is referred to as *canonical group sequential procedure* and, more precisely, it calculates the critical values based on the canonical joint distribution with the covariance $\text{Cov}(T(t_l), T(t_m))$ estimated through $\sqrt{\hat{\mathcal{I}}(t_l)/\hat{\mathcal{I}}(t_m)}$ for $t_l < t_m$. The type I error rates of the two group sequential procedures were determined by means of Monte Carlo simulations. The setup and results of the Monte Carlo simulation studies are summarized next. In the simulations, the events were generated using a negative binomial process. From a practical perspective, recurrent events generated with a negative binomial process can for

TABLE 2.2: Parameters in the simulation study of the type I error rate for the group sequential procedures with the LWYY model.

Parameter	Value
One-sided significance level α	0.025
Maximum sample sizes $n_1 = n_2$	2300
Shape parameter ϕ	5.2
Maximum number of data looks K	2
Study duration [months]	55
Recruitment period [months]	29
Individual follow-up [months]	26–55
Annualized rate λ_0	0.15
Effect size β under H_0	0

example represent heart failure hospitalizations in clinical trials with subjects suffering from chronic heart failure. The parameter choices for the simulation study were motivated by the settings considered when planning the Paragon-HF trial [26]. The parameters are listed in Table 2.2. In detail, a uniform deterministic recruitment during the recruitment period of 29 months was assumed and once subjects entered the trial, they were followed up until the trial ended after a calendar time of 55 months. Thus, the individual follow-up times varied between 26 and 55 months. As error spending functions, the Pocock-type error spending function and the O'Brien-Fleming-type error spending function were considered. The type I error rate α was spent by means of the calendar time, i.e., α was allocated according to (2.1) with $\nu_k = t_k / t_{\max}$, where $t_{max} = 55$ months the calendar time of the final data look. Here, the focus was on a maximum number of $K = 2$ data looks with the calendar time of the first data look varied, i.e., $t_1 = 6, \ldots, 50$ months, and the second data look performed at $t_2 = t_{max}$. The simulated type I error rates, based on 500 000 Monte Carlo replications, are presented in Figure 2.2. Figure 2.2 shows that the canonical group sequential procedure controls the type I error rate, except for early data looks with an O'Brien-Fleming-type error spending function. Moreover, the consistent group sequential procedure slightly inflates the type I error rate by about 0.0005 for considered practically relevant scenarios. In simulation results not reported here, I showed that the canonical group sequential procedure becomes conservative for large sample sizes, while the type I error rate of the consistent procedure converges to the target type I error rate, as expected. Furthermore, the difference between the two group sequential procedures increases when the number of data looks increases; the consistent procedure becomes more liberal and the canonical procedure becomes slightly conservative. For more detailed results, confer Mütze et al. [49].

The negative binomial process was chosen as a recurrent event generating process to simulate recurrent heart failure hospitalizations in clinical trials in chronic heart failure. Another common endpoint in clinical trials in chronic heart failure is the composite of recurrent heart failure hospitalizations and cardiovascular death. Events for this composite endpoint can be simulated by means of a parametric joint Gamma frailty model with a Poisson process for the hospitalizations, exponentially

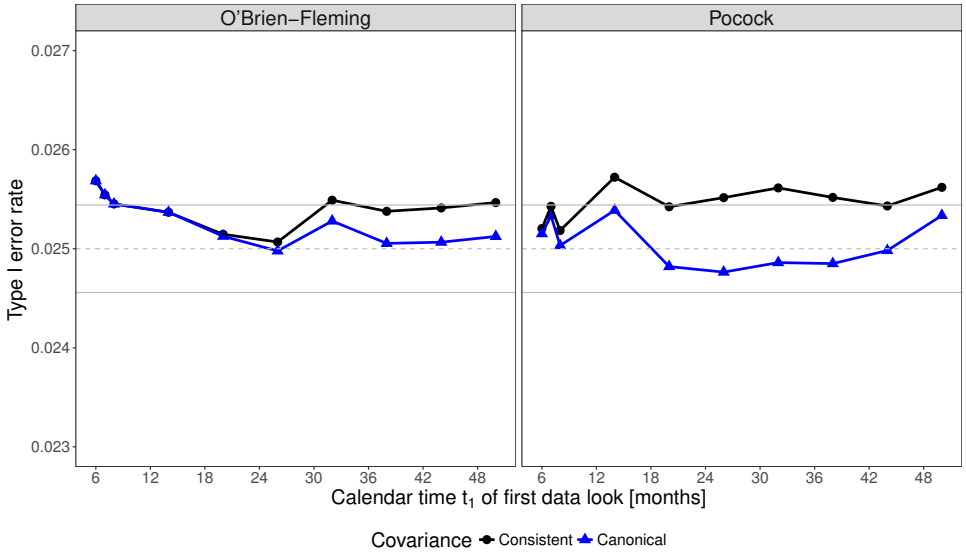


FIGURE 2.2: Simulated type I error rate versus the calendar time t_1 of the first data look for the two group sequential procedures. The maximum sample sizes are $n_1 = n_2 = 2300$. The grey lines mark the area of two times the simulation error around the target type I error rate $\alpha = 0.025$.

distributed death times, and a Gamma distributed frailty term to link the processes. A simulation study with events generated by the joint Gamma frailty model showed that for this event process, both group sequential procedures for the LWYY model control the type I error rate. Additionally, the difference in the type I error rate between the procedures is smaller than in the case of events from a negative binomial process.

Concerning the planning of group sequential designs with the LWYY model, I proposed to calculate the maximum information based on the canonical joint distribution by solving (2.3). Through a Monte Carlo simulation study, I illustrated this approach results in designs achieving the target power for both group sequential procedures for the LWYY model, see Mütze et al. [49] for details.

Summarizing, in my research of group sequential designs for the LWYY model, I outlined that the sequence of Wald statistics in the LWYY model does not follow the canonical joint distribution and I illustrated that a consistent group sequential procedure can be defined based on a consistent estimator for the covariance $\text{Cov}(T(t_l), T(t_m))$ between Wald statistics from different data looks. Through a Monte Carlo simulation study, I highlighted that the proposed consistent group sequential procedure in the LWYY model can result in small type I error rate inflation for scenarios motived by clinical trials in chronic heart failure. Additionally, I demonstrated that the canonical group sequential procedure in the LWYY model is robust against deviations from the canonical joint distribution and that, overall, it results in a slightly better type I error rate control than the consistent group sequential procedure for practically relevant scenarios. Last but not least, I pointed out that the maximum information in group sequential designs with the LWYY model can be planned based on the canonical joint distribution.

In conclusion, my research provides the justification to apply the canonical group sequential procedure, which is implemented in the most common software packages, to the LWYY model when analyzing clinical trials with patients suffering from chronic heart failure and still control the type I error rate even though this model technically does not fulfill the canonical joint distribution. For scenarios in which an asymptotic type I error rate control is of practical importance, I proposed the consistent group sequential procedure.

2.2 Blinded sample size re-estimation in three-arm trials

Next, I summarize my results on blinded sample size re-estimation in three-arm trials with normal outcomes, which address the research question outlined in Chapter 1.3.2. The detailed results were published in [46]. I start with introducing the statistical model. Then, I recapitulate the basic idea for blinded sample size re-estimation based on an estimate of the nuisance parameter and present the proposed blinded sample size re-estimation procedure for three-arm clinical trials with the 'gold standard' design. Following this, I highlight the key performance aspects of the proposed blinded sample size re-estimation procedure.

The 'gold standard' designs for three-arm clinical trials includes one arm for the experimental treatment (E), one arm for the active control, which I refer to as the reference (R), and one arm for the placebo control (P). Here, the outcomes are modeled as normally distributed random variables with identical variance but varying means across the arms, i.e.,

$$X_{ki} \sim \mathcal{N}\left(\mu_k, \sigma^2\right), \quad i = 1, \ldots, n_k, \quad k = E, R, P.$$

Let the total sample size be $n = n_E + n_R + n_P$ and the proportion of the sample size allocated to treatment arm $k = E, R, P$ is denoted by $w_k = n_k/n$. I considered the setting in which a clinical trial is successful if superiority of the experimental treatment and the reference compared to placebo, respectively, can be proven and if non-inferiority of the experimental treatment compared to the reference can be shown. Under the assumption that smaller values of the means μ_k $(k = E, R, P)$ represent a more efficacious treatment and with the non-inferiority margin $\delta_{ER} > 0$, the assessment of non-inferiority of the experimental treatment compared to the reference can be formulated as the statistical hypothesis testing problem

$$H_0^{ER} : \mu_E \geq \mu_R + \delta_{ER} \quad \text{versus} \quad H_1^{ER} : \mu_E < \mu_R + \delta_{ER}.$$

The superiority of the experimental treatment over placebo and the superiority of the reference over placebo can be expressed as the statistical hypothesis testing problems

$$\begin{aligned} H_0^{EP} : \mu_E \geq \mu_P \quad &\text{versus} \quad H_1^{EP} : \mu_E < \mu_P, \\ H_0^{RP} : \mu_R \geq \mu_P \quad &\text{versus} \quad H_1^{RP} : \mu_R < \mu_P. \end{aligned}$$

The three-arm trial is successful if all three hypotheses can be rejected. This results in the intersection-union testing problem

$$H_0 : H_0^{ER} \cup H_0^{EP} \cup H_0^{RP} \quad \text{versus} \quad H_1 : H_1^{ER} \cap H_1^{EP} \cap H_1^{RP}.$$

As proven by Berger [54], testing each of the three local hypotheses H_0^{ER}, H_0^{EP}, and H_0^{RP} with a level α test and rejecting the global hypothesis H_0 when each of the local hypotheses are rejected, results in a test for the global hypothesis H_0, which controls the type I error rate α. This test procedure is in general conservative. In my research, I focused on Student's t-test as a level α test for the local hypotheses. Stucke and Kieser [55] showed that the power of this test procedure for the global hypothesis can be approximated by the cumulative distribution function of a three-dimensional normal distribution. Let $B(n, \sigma^2)$ denote this power approximation as a function of the total sample size n and the variance σ^2. The power also depends on other parameters such as the target type I error rate α, the sample size allocation (w_E, w_R, w_P), and the mean differences in the alternative, i.e., $\delta_{ij}^* = \mu_i - \mu_j$, $(i, j) = (E, R), (R, P), (E, P)$. However, for the sake of readability, not all parameters affecting the power are listed as arguments of the function $B(\cdot, \cdot)$. Thus, for prespecified mean differences in the alternative hypothesis, sample size allocation, and variance σ^2, the sample size for a three-arm trial in the 'gold standard' design required to obtain a power of at least $1 - \beta$ when testing the global hypothesis H_0 is the smallest n solving the inequality $B(n, \sigma^2) \geq 1 - \beta$, i.e.,

$$n = \min \left\{ n \in \mathbb{N} : B(n, \sigma^2) \geq 1 - \beta \right\}. \tag{2.6}$$

The variance assumed in the sample size calculation is generally a guesstimate and a failure to accurately specify the variance during the sample size planning will lead to under- or overpowered clinical trials. Clinical trial designs with a nuisance parameter based sample size re-estimation counteract an inaccurately specified sample size by adjusting the final sample size mid-trial based on an estimate of the variance obtained with data from an internal pilot study. Let the current clinical trial include an internal pilot study of total sample size n_1 and the interim variance estimator $\hat{\sigma}_1^2$ estimates σ^2 with the data from the internal pilot study. Then, the sample size is re-estimated by solving (2.6) with the interim variance estimator $\hat{\sigma}_1^2$ plugged in for the variance parameter σ^2. In mathematical terms, the re-estimated sample size is defined by

$$\hat{n}_{reest} = \min \left\{ n \in \mathbb{N} : B(n, \hat{\sigma}_1^2) \geq 1 - \beta \right\}.$$

If the re-estimated sample size is smaller than the internal pilot study sample size n_1, the internal pilot study sample size n_1 is also the final sample size. Unless stated otherwise, the term *sample size re-estimation* refers to the adjustment of the final sample size based on a variance estimate and *blinded sample size re-estimation* implies that the treatment allocation of subjects from the internal pilot study is unknown when estimating the variance. Next, I present the proposed procedure for blinded sample size re-estimation in three-arm trials with the 'gold standard' design. Let the sample $Y_1, \ldots, Y_{n_1}$ denote the blinded results from an internal pilot study of size n_1. A crucial part of my development of blinded sample size re-estimation procedures for three-arm trials was researching blinded estimators for the variance. When developing said blinded sample size re-estimation procedure, I considered the following blinded variance estimators.

Blinded one-sample variance estimator According to Friede and Kieser [15], the one-sample variance estimator is the recommended estimator for blinded sample size re-estimation in two-arm trials with normal data. The one-sample variance estimator estimates the outcome variance by the sample variance of the blinded data, that is

$$\hat{\sigma}^2_{OS} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_i - \bar{Y})^2 .$$

Blinded adjusted one-sample variance The blinded one-sample variance estimator is generally biased. The blinded adjusted one-sample variance is a version of the one-sample variance estimator, which is adjusted such that it is unbiased under the planning alternative. With $\text{Bias}(\hat{\sigma}^2_{OS}, \sigma^2 | H_1)$ the bias of the one-sample variance estimator under the planning alternative, the blinded adjusted one-sample variance is defined by

$$\hat{\sigma}^2_{OSU} = \hat{\sigma}^2_{OS} - \text{Bias}(\hat{\sigma}^2_{OS}, \sigma^2 | H_1).$$

Blinded variance estimator by Xing and Ganju [56] Xing and Ganju [56] showed that the outcome variance can be estimated unbiased and blinded in a randomized block design. In a design with b_1 balanced blocks of length m from the internal pilot study, an unbiased blind estimator for the nuisance parameter σ^2 is given by the sample variance of the sum T_k of the observations in block k, i.e.,

$$\hat{\sigma}^2_{XG} = \frac{1}{n_1 - m} \sum_{k=1}^{b_1} (T_k - \bar{T})^2 .$$

Here, $\bar{T}$ denotes the mean of the block sums.

With each of the blinded variance estimators, I defined a blinded sample size re-estimation procedure for three-arm trials with the 'gold standard' designs and normal outcomes and then assessed their operating characteristics. Operating characteristics of interest are the power, the sample size distribution, and the type I error rate. The comparison is performed through Monte Carlo simulation studies. In the following, I present the main results concerning the power comparison of the sample size re-estimation procedures. The results presented here are for the parameters listed in Table 2.3. This choice of parameter combination is motivated by the clincal trial in hypertension published by Krum et al. [35]. Mütze and Friede [46] referred to this parameter combination as *Scenario 1*. During the simulation study, the size of the internal pilot study is varied between $n_1 = 30$ and the fixed designs' sample size which is $n_1 = 528$ for the sample size allocation $n_E : n_R : n_P = 1 : 1 : 1$ and $n_1 = 434$ for the sample size allocation $n_E : n_R : n_P = 3 : 3 : 1$. The power of the three blinded sample size re-estimation procedures for the parameters listed in Table 2.3 is presented in Figure 2.3. Figure 2.3 shows that the none of the blinded sample size re-estimation procedures meets the target power for all considered internal pilot study sizes. The sample size re-estimation procedure based on the one-sample variance estimator results in a power larger than the target power $1 - \beta = 0.8$, except

TABLE 2.3: Scenarios for the Monte Carlo simulation study of the sample size re-estimation procedures' operating characteristics.

Parameter	**Value**
One-sided significance level α	0.025
Target power $1 - \beta$	0.8
Non-inferiority margin δ_{ER}	$\delta_{ER} = 0.3$
Means μ_E, μ_R under the alternative H_1	$\mu_E = \mu_R = 0$
Mean μ_P under the alternative H_1	$\mu_P = 0.6$
Standard deviation σ under the alternative H_1	$\sigma = 1$
Sample size allocation $n_E : n_R : n_P$	1:1:1, 3:3:1

for small internal pilot study sizes of $n_1 = 30$ or smaller and an unbalanced sample size. Re-estimating the sample size with the Xing-Ganju variance estimator results in underpowered clinical trials unless the internal pilot study is almost as big as the corresponding clinical trial with a fixed design. Sample size re-estimation based on the adjusted one-sample variance underpowers the clinical trial for internal pilot studies smaller than about $n_1 = 150$ subjects and meets the target power otherwise. However, it is important to note that the adjusted one-sample variance uses information about the alternative, which is unknown in practice, when estimating the variance and as such uses more information than the one-sample variance and the Xing-Ganju variance estimator. The results presented in Figure 2.3 are unexpected in the sense that the blinded sample size re-estimation based on the one-sample variance, which is the recommended method for sample size re-estimation in two-arm trials with normal outcomes and results in the target power in these designs, overpowers three-arm clinical trials with the 'gold standard' design and normal outcomes. Moreover, as I presented in the published manuscript [46], the Xing-Ganju variance estimator results in a higher variability of the final sample size compared to the one-sample variance and its adjusted version. One of the main regulatory requirements about adaptive designs is the control of the type I error rate. All three blinded sample size re-estimation procedures result in a small type I error rate inflation of Student's t-test for the non-inferiority hypothesis H_0^{ER}. The type I error rate of Student's t-test when testing the superiority hypotheses H_0^{EP} and H_0^{RP} is only inflated for the sample size re-estimation based on the Xing-Ganju variance estimator. However, since the global hypothesis H_0 is tested through the intersection-union test procedure, which is conservative, the type I error rate of the test procedure is controlled in the design with blinded sample size re-estimation even though the type I error rates of the local tests might be slightly inflated.

As Figure 2.3 highlights, none of the proposed blinded sample size re-estimation procedures meets the target power across all of the considered internal pilot study sample sizes. Therefore, to obtain a procedure which meets the target power independently of the internal pilot study sample size, I suggested a modification of the sample size re-estimation procedure based on the Xing-Ganju variance estimator. The idea of the modification is to multiply the re-estimated sample size $\hat{n}$ with

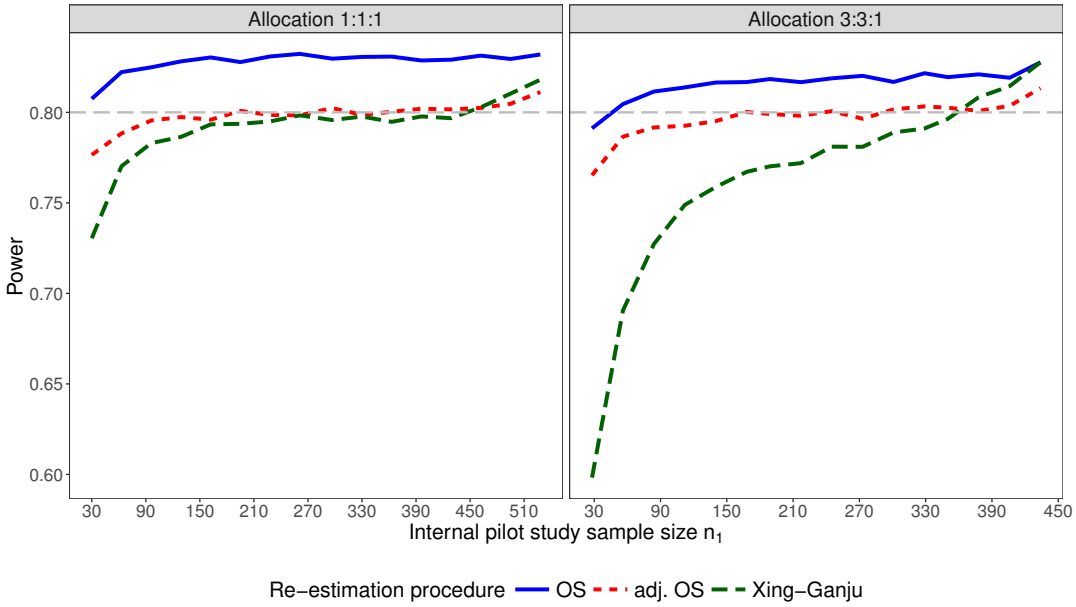


FIGURE 2.3: Simulated power of the 'gold standard' design with blinded sample size re-estimation procedure based on the one-sample variance estimator (OS), the adjusted one-sample variance (adj. OS), and the Xing-Ganju variance estimator (Xing-Ganju). The grey line depicts the target power $1 - \beta$.

a factor ζ, which is determined such that the inflated sample size $\zeta\hat{n}$ results in the target power. In the following, I outline the proposed modification. Let $B(n, \sigma^2)$ be the power approximation and I denote the corresponding sample size (2.6) as a function of the outcome variable σ^2, that is $n(\sigma^2)$. Furthermore, let $f_{\hat{\sigma}^2_{XG}}(\cdot)$ denote to the probability density function of the Xing-Ganju variance estimator. The Xing-Ganju variance estimator based on the results from the internal pilot study follows a stretched χ^2-distribution [46], i.e.,

$$\hat{\sigma}^2_{XG} \sim \frac{m\sigma^2}{n_1 - m}\chi^2_{b_1-1}.$$

Then, the power of the intersection-union test procedure for the hypothesis H_0 in a three-arm trial with a blinded sample size re-estimation can be approximated by

$$\text{Power} \approx \int_0^\infty B\left(\tilde{n}\left(x\right), \sigma^2\right) f_{\hat{\sigma}^2}(x)\mathrm{d}x$$

with the final sample size function $\tilde{n}\left(x\right) = \max\{n(x), n_1\}$. The inflation factor ζ is defined as the positive number which solves

$$\int_0^\infty B\left(\tilde{n}_\zeta\left(x\right), \sigma^2\right) f_{\hat{\sigma}^2_{XG}}(x)\mathrm{d}x = 1 - \beta$$

with $\tilde{n}_\zeta\left(x\right) = \max\{\zeta \cdot n(x), n_1\}$. Since the distribution of the Xing-Ganju variance estimator depends on the variance σ^2, the internal pilot study size n_1, and the length m of the randomized blocks, all three parameters have to be specified when calculating the inflation factor. While the length of the randomized blocks and the internal

pilot study size are generally known, the variance is unknown. However, I showed that the variation of the inflation factor across different values for σ^2 is negligible as long as the variance σ^2 is chosen large enough that the resulting fixed design sample size exceeds the internal pilot study sample size n_1. As a consequence, the choice of σ^2 for calculating the inflation factor does not matter as long as it is sufficiently large. I assessed the operating characteristics of the modified blinded sample size re-estimation based on the Xing-Ganju variance estimator in a simulation study and showed that the modified procedure results in clinical trials which meet the target power for the considered range of internal pilot study sample sizes [46, Section 5]. However, the inflation factor further increases the variability of the final sample size. I also proved that the sample size re-estimation based on the Xing-Ganju variance estimator and its modified version do not result in a biased effect estimate at the end of the trial [46, Section 6]. Last but not least, it is worth highlighting that the presented approach for modifying the sample size re-estimation procedure cannot be easily transferred to the one-sample variance estimator or its adjusted version since the resulting inflation factor is not constant in the variance σ^2.

In conclusion, I showed that blinded sample size re-estimation procedures in three-arm trials can be defined analogously to the sample size re-estimation procedures in two-arm trials. Furthermore, I illustrated that these blinded sample size re-estimation procedures in three-arm trials do not convey the target power for scenarios motivated by a clinical trial in hypertension. Finally, I modified the blinded sample size re-estimation procedure based on the Xing-Ganju variance estimator such that the modified procedure yields the target power.

2.3 Incorporating prior information into the sample size re-estimation

In Chapter 1.3.3, I described the need for studying incorporating prior information on the variance into the nuisance parameter based sample size re-estimation. The detailed results of my research on this topic were published by Mütze, Schmidli, and Friede [47] and in the following, I summarize my main findings. I start by defining the statistical model and then I present an approach for incorporating prior information into the sample size re-estimation and discuss the procedure's operating characteristics.

I focused on two-arm, parallel group superiority trials with normal outcomes. The treatment group is indexed with $i = T$ and the control group is indexed with $i = C$. The groups contain $j = 1, \ldots, n_i$ subjects, the total sample size is $n = n_T + n_C$, and the randomization ratio is denoted by $k = n_C/n_T$. Conditional on the mean μ_i and the variance σ^2, the outcome of the j^{th} patient in group $i = T, C$ is modeled as the normally distributed random variable

$$X_{ij}|\mu_i, \sigma^2 \sim \mathcal{N}\left(\mu_i, \sigma^2\right).$$

The random variables are modeled as independent and larger values of the means μ_i $(i = T, C)$ are better. Then, the statistical hypothesis testing problem to assess the

superiority of the treatment over the control is defined by

$$H_0 : \mu_T \leq \mu_C \qquad \text{versus} \qquad H_1 : \mu_T > \mu_C.$$

The null hypothesis H_0 is tested through a one-sided Student's t-test with a pooled variance estimation and a quantile of a t-distribution with $n-2$ degrees of freedom as a critical value. The statistical power $B(n, \sigma^2, \delta, k)$ of Student's t-test for testing the null hypothesis H_0 against the alternative hypothesis H_1 is a function of the total sample size n, the outcome variance σ^2, the mean difference $\delta = \mu_T - \mu_C$, the randomization ratio k as well as the significance level α. Then, the sample size n required to test the null hypothesis H_0 with a target power of $1-\beta$ for a given mean difference $\delta^* > 0$ is the smallest sample size n which solves the inequality

$$B(n, \sigma^2, \delta^*, k) \geq 1 - \beta \tag{2.7}$$

for a prespecified variance σ^2 and a prespecified allocation ratio k.

In the present design, the nuisance parameter based sample size re-estimation can be facilitated as outlined in Chapter 2.2 by estimating the variance σ^2 based on data from the internal pilot study and then recalculating the final sample size by solving (2.7) with the interim variance estimator $\hat{\sigma}_1^2$ plugged in for the variance parameter. In mathematical terms, the re-estimated sample size is defined by

$$\hat{n}_{reest} = \min\left\{n \in \mathbb{N} : B(n, \hat{\sigma}_1^2, \delta^*, k) \geq 1 - \beta\right\}.$$

Next, I explain the proposed procedure for incorporating prior information into the sample size re-estimation. I assume that prior information on the variance σ^2 is available through an MAP prior, that is a prior distribution for the variance σ^2 obtained through a meta-analysis of sample variances from historical clinical trials. More specifically, prior information about the variance is formulated in terms of an inverse Gamma distribution

$$\sigma^2 \sim \text{InvGamma}(a, b)$$

with shape parameter $a > 0$ and scale parameter $b > 0$. For an inverse Gamma prior distribution, twice the shape parameter, $2a$, can be interpreted as the prior effective sample size (ESS). The effective sample size quantifies the 'amount of information' contained in the prior distribution [57, 58]. For our setting of frequentist planning and analysis of a clinical trial with an MAP prior on the variance, Schmidli, Neuenschwander, and Friede [41] proposed to incorporate prior information into the sample size planning by solving inequality (2.7) with a single value for the variance parameter obtained from its prior distribution. Examples for single values from the prior distribution are the prior mean, the prior median, or a quantile.

I first assumed that the data from the internal pilot study is unblinded because for the unblinded data it is easier to identify the root cause when the re-estimation procedure does not convey the target power. However, it is important to note that blinded sample size re-estimation is recommended from a statistical and a regulatory perspective [10, 11, 15]. To incorporate prior information on the sample variance into the sample size re-estimation, I proposed to update the prior distribution of the

variance using the data from the internal pilot study. Then, the sample size is recalculated by solving inequality (2.7) using a posterior Bayes estimator as the variance in the power formula. Obvious choices for posterior Bayes estimators are the posterior mean and the posterior median. Thus, the crucial step when incorporating prior information into the sample size re-estimation, and therefore explained in more detail in the following, is updating the prior information. Let $p(\mu_T, \mu_C, \sigma^2)$ be the prior density for the parameter vector (μ_T, μ_C, σ^2). In my research, I assumed that the prior distributions of the parameters μ_T, μ_C, and σ^2 are stochastically independent and that the mean' priors are improper and uniform, i.e.,

$$\begin{aligned} &p(\mu_T) \propto 1, \\ &p(\mu_C) \propto 1, \\ &p(\mu_T, \mu_C, \sigma^2) = p(\mu_T)p(\mu_C)p(\sigma^2). \end{aligned}$$

As mentioned above, the variance σ^2 has an inverse Gamma distribution as a prior which is a conjugate prior. Therefore, the posterior distribution of the variance after the internal pilot study also follows an inverse Gamma distribution, i.e.,

$$\sigma^2 \big| \hat{\sigma}^2_{1,pool} \sim \text{InvGamma}\left(a + \frac{n_1 - 2}{2}, b + \frac{n_1 - 2}{2}\hat{\sigma}^2_{1,pool}\right).$$

Here, $\hat{\sigma}^2_{1,pool}$ is the pooled sample variance calculated from the unblinded internal pilot study data and n_1 is the total sample size in the internal pilot study. Then, the sample size is re-estimated by calculating a posterior estimator $\hat{\sigma}^2_{1,Bayes}$ and plugging it into the power formula when calculating the sample size by solving (2.7). Thus, the re-estimated sample size is given by

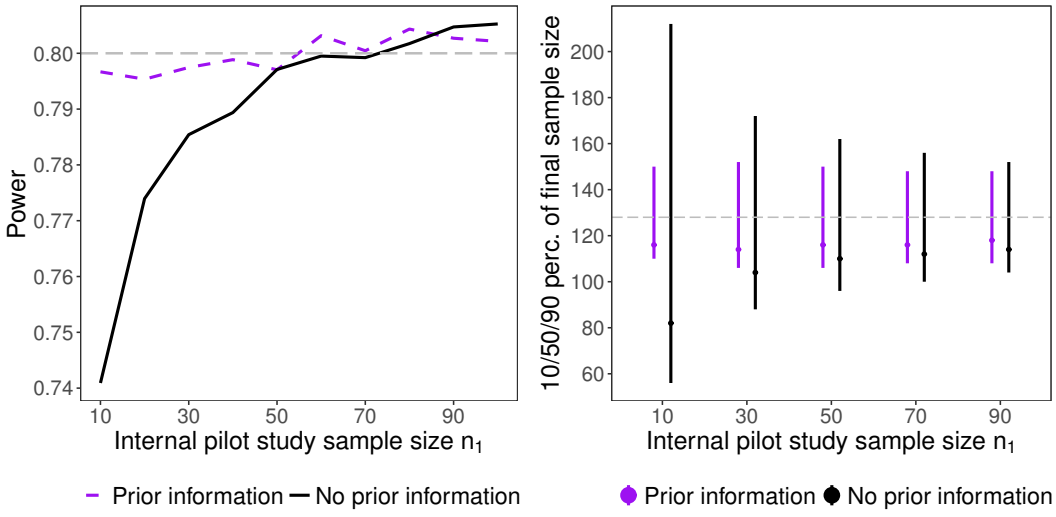
$$\hat{n}_{Bayes} = \min\left\{n \in \mathbb{N} : B(n, \hat{\sigma}^2_{1,Bayes}, \delta^*, k) \geq 1 - \beta\right\}.$$

In the following, I summarize the operating characteristics of the proposed sample size re-estimation procedure incorporating prior information and compare the procedure's performance to the standard unblinded sample size re-estimation procedure based on the pooled sample size, which does not utilize prior information. Here, the focus is on the posterior mean $\hat{\sigma}^2_{1,mean}$ as the posterior estimator for the sample size re-estimation. The considered operating characteristics are the power, the final sample size distribution, and the type I error rate. The operating characteristics are obtained through Monte Carlo simulations. When simulating the power and the final sample size distribution, a one-sided significance level of $\alpha = 0.025$ and a target power of $1 - \beta = 0.8$ are assumed. The data are simulated with the mean difference $\delta^* = 0.5$, which is identical to the mean difference used for calculating the sample size during the sample size re-estimation. The outcome variance is set to $\sigma^2 = 1$ and both study arms are equally sized, that is $k = 1$. In a fixed sample design, this parameter combination requires a total sample size of $n = 128$. In the simulation study, the internal pilot study sample size is varied between $n_1 = 10$ and $n_1 = 100$. The simulation scenarios are listed in Table 2.4. The simulation study is conceptually split into two parts. In the first part, the prior distribution is chosen such that no prior-data conflict exists and in the second part, the prior distribution

TABLE 2.4: Scenarios for the Monte Carlo simulation study of the power and the final sample size distribution.

Parameter	Value
One-sided significance level α	0.025
Target power $1-\beta$	0.8
Margin δ^* under the alternative H_1	0.5
True variance σ^2	1
Internal pilot study size n_1	$10, 20, \ldots, 100$
Sample size ratio k	1

is defined such that a prior-data conflict is present. Here, the absence of a prior-data conflict is interpreted such that the outcome variance $\sigma^2 = 1$ is identical to the expected value of the prior distribution $p_{\sigma^2}(\cdot)$. Thus, in the first part, the parameters of the inverse Gamma prior of the variance are chosen such that the prior's expected value is one and such that the prior has an effective sample size of $ESS = 25$. The results of the simulation study for the case of no prior-data conflict are presented in Figure 2.4. Figure 2.4 highlights that incorporating prior information into the sam-



(A) Power (B) Distribution of final sample size

FIGURE 2.4: Simulated power and final sample size distribution against the internal pilot study sample size for sample size re-estimation procedures with and without prior information when no prior-data conflict is present. The prior effective sample size is equal to $ESS = 25$.

ple size re-estimation is advantageous when no prior-data conflict exists: the sample size re-estimation procedure incorporating prior information has a power closer to the target power than the re-estimation procedure based on the pooled sample variance, which does not incorporate prior information. Moreover, the incorporation of

prior information reduces the variability of the final sample size. Further simulation results reported by Mütze, Schmidli, and Friede [47] show that the benefits of incorporating prior information increase with the prior effective sample size and that the type I error rate of the re-estimation procedure incorporating prior information is inflated but converges to the target level $\alpha = 0.025$ with increasing prior effective sample size.

In the simulation study presented above, I assumed an ideal situation in the sense that the mean of the prior distribution for the variance is identical to the true outcome variance. In practice, discrepancies between the prior mean and the true outcome variance can occur and depending on the magnitude of such discrepancies, a prior-data conflict can be present. What constitutes a prior-data conflict is not uniquely defined. A prior-data conflict can be said to be present if the 95% probability interval of the prior-predictive distribution does not contain the observed variance. Therefore, I also studied the performance of the sample size re-estimation procedures when a prior-data conflict is present. To that end, a prior distribution $p_{\sigma^2}(\cdot)$ for the variance with an expected value of $\sigma^2_{mean} = 0.49$ and a true outcome variance of $\sigma^2 = 1$ are assumed. The other parameters are defined as listed in Table 2.4. In particular, the prior distribution of the variance is again an inverse Gamma distribution with parameters chosen such that the prior effective sample size is equal to $ESS = 25$ and that the expected value is equal to $\sigma^2_{mean} = 0.49$. The probability that this inverse Gamma prior distribution exceeds the value of the true variance $\sigma^2 = 1$ is equal to 0.84%. The corresponding simulation results are presented in Figure 2.5. Figure 2.5 shows that the sample size re-estimation procedure incorpo-

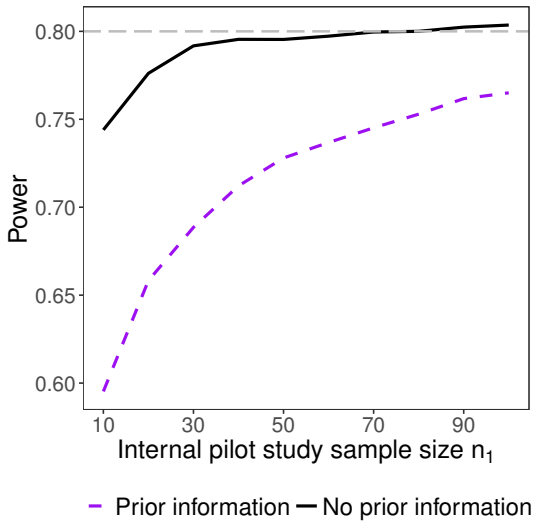


FIGURE 2.5: Simulated power and final sample size distribution against the internal pilot study sample size for sample size re-estimation procedures with and without prior information when a prior-data conflict is present. The prior effective sample size is equal to $ESS = 25$.

rating prior information results in underpowered clinical trials in the presence of a prior-data conflict when the prior mean is smaller than the true variance. The larger

the internal pilot study, the closer is the power of the design with a sample size re-estimation incorporating prior information to the target power. However, even large internal pilot study sizes of close to $n_1 = 100$ subjects cannot completely discount the negative effects of a prior-data conflict on the power. Further simulation results presented in [47] show that the underpowering due to a prior-data conflict increases as the prior effective sample size increases. It is worth emphasizing that the underpowering is due to modeling a prior-data conflict with a prior mean smaller than the outcome variance. If the prior mean is larger than the true outcome variance, the design with a sample size re-estimation incorporating prior information will be overpowered.

Schmidli et al. [42] suggested to make MAP priors more robust against prior-data conflicts by mixing the initial MAP prior $p_{MAP}(\cdot)$ with a vague prior $p_V(\cdot)$. With $w_R \in (0, 1)$ a mixture probability, the robustified MAP prior $p_{rMAP}(\cdot)$ is defined by

$$p_{rMAP}(x) = w_R p_V(x) + (1 - w_R) p_{MAP}(x).$$

I studied whether incorporating a robust MAP prior into the sample size re-estimation results in a re-estimation procedure which maintains the benefits of incorporating prior information while simultaneously being robust against prior-data conflicts. For that, I considered an inverse Gamma distribution with shape parameter $a = 2$ and rate parameter $b = 1$ as the vague prior $p_V(\cdot)$. Results reported in [47] show that to fully mitigate any negative effects of a prior-data conflict on the power when incorporating prior information into the sample size re-estimation using the prior above, the MAP prior must be almost completely discounted, i.e., w_R must be close to one. However, even for smaller weights of $w_R \approx 0.3$, the power-reducing effect of a prior-data conflict is already weakened considerably. Moreover, the larger the weight w_R, the smaller the benefits of incorporating prior information concerning a reduced variability of the re-estimated sample size.

As part of my research, I also studied incorporating prior information into the blinded sample size re-estimation [46, Section 6]. The performance of the proposed blinded procedures are qualitatively the same as the unblinded procedures, i.e., the blinded procedures meet the target power when there is no prior-data conflict but they yield over- or underpowered clinical trials in the presence of a prior-data conflict.

Summarizing, I proposed an ad-hoc approach for incorporating prior information about the outcome variance into the sample size re-estimation. The proposed idea is to update the prior information using data from an internal pilot study and then to re-estimate the sample size by plugging in a Bayes estimator obtained from the posterior into the fixed design sample size calculation. I showed that this ad-hoc approach improves the performance of the sample size re-estimation procedure in comparison to procedures which do not utilize prior information. However, I also showed that the proposed procedure is not robust against prior-data conflicts, but that some robustness can be obtained by robustifying the MAP prior for the sample size re-estimation. In conclusion, incorporating prior information into the sample size re-estimation can be advantageous compared to the traditional sample size re-estimation, but the benefits have to be carefully weighted against the risks on a case-to-case basis.

An implementation of the statistical methodology for incorporating prior information on the variance into the nuisance parameter based sample size re-estimation is provided through the R package *varmap*, which is available on GitHub [59].

3 Discussion

Jackson et al. [5] advised the development of adaptive clinical trial designs to improve the cardiovascular drug development process. In this dissertation, I proposed group sequential designs for clinical trials with subjects suffering from chronic heart failure and designs with nuisance parameter based sample size re-estimation for clinical trials with subjects suffering from hypertension. The main advantages of the studied group sequential designs are that efficacious treatments can be identified earlier than in a fixed sample design, which can shorten the trial duration by up to several years due to the generally long follow-up times in clinical trials in chronic heart failure. A shortened trial proving efficacy accelerates patients' access to the new treatment presuming the treatment's safety. The benefit of the proposed sample size re-estimation for three-arm trials with the 'gold standard' design is that it assures that the trial is appropriately powered even though the nuisance parameter might have been misspecified during the planning phase of the clinical trial. My research on sample size re-estimation incorporating prior information on the variance showed that ad hoc approaches for incorporating the prior information can be beneficial in reducing the variability of the final sample size. However, in the presence of a prior-data conflict, incorporating prior information into the sample size re-estimation results in designs that do not meet the target power.

My research about group sequential designs for recurrent events [48, 49] focused on constructing the efficacy boundaries for the different data looks, planning the maximum information, and assessing the type I error rate and the power of the proposed group sequential procedures. There are practically relevant aspects of the proposed group sequential designs for recurrent events which have not been studied yet, for instance, monitoring group sequential designs with recurrent events. When the calendar time of the data looks is to be set based on the number of observed events or the observed information level, these properties have to be monitored throughout the conduct of the trial. Ideally, monitoring procedures do not require knowledge of the treatment indicator in order that the monitoring can be conducted by the trial statistician, who generally is blinded, i.e., has no knowledge of the treatment indicator. Recently, Friede, Häring, and Schmidli [60] studied blinded continuous information monitoring for two-arm trials with negative binomial outcomes. The proposed blinded continuous information monitoring could also be applied to group sequential designs with negative binomial outcomes. A blinded monitoring of the number of observed events is straight forward in that only the number of observed events has to be counted. No matter which monitoring procedure is used to select the calendar times of the data looks, it is important to study how the monitoring affects the type I error rate and the power of the group sequential design. In addition to monitoring the information level and the number of observed events of a clinical trial, it is also of practical interest to predict the information level and the

number of observed events for a future calendar time, for instance to forecast the timing of the next data looks. A general approach for predicting the information level at a future calendar time using resampling techniques, which can be applied to recurrent events, was established by Scharfstein and Tsiatis [61].

As part of this dissertation, group sequential designs were studied for recurrent events modeled by the LWYY model. This choice of recurrent event model was motivated by the ongoing Paragon-HF trial [26], which includes a planned primary analysis of the composite of heart failure hospitalizations and cardiovascular death with the LWYY model. The LWYY model assumes that the censoring process is independent of the event generating process; a property referred to as *independent censoring*. If the sampling model violates the independent censoring assumption, the rate ratio estimator can be biased and control of the type I error rate is no longer guaranteed from a theoretical perspective. As a death results in censoring, heart failure hospitalizations and cardiovascular death must be assumed to be dependent, statistical models that explicitly account for the presence of a terminal event such as death can also be of practical interest. For instance, Mao and Lin [62] extended the LWYY model to account for the dependence of the censoring process and the event process. Not only are there multiple potential statistical methods, there is also an ongoing discussion on what represents appropriate effect measures and endpoints. The LWYY model and the negative binomial model measure efficacy through the rate ratio. Other possible effect measures can be based on the mean number of cumulative events, which can be analyzed with the nonparametric model proposed by Ghosh and Lin [63]. To improve the interpretability of the composite of heart failure hospitalizations and cardiovascular death, it has been discussed to weight the different parts of the composite endpoint, see Rauch et al. [64]. The extension of the LWYY model by Mao and Lin [62] also allows for different weights of the composite of heart failure hospitalizations and cardiovascular death and as such allows the analysis of the event rate ratio of a weighted composite endpoint. Pocock et al. [65] proposed the win ratio approach for analyzing the composite of heart failure hospitalizations and cardiovascular death. The win ratio approach does not define the treatment effect based on recurrent event rates. Instead, to calculate the win ratio, the subjects' outcomes are compared between the two groups in a pairwise manner concerning a prespecified order of the possible outcomes of the composite endpoint. Each comparison does have a *winner* and a *loser*, or the comparison is *tied*. Then, the win ratio is the number of winners in the treatment group compared to the number of losers in the treatment group. Therefore, depending on which primary analysis method will be considered in future clinical trials in chronic heart failure, further development of adaptive designs are required, since, to the best of my knowledge, adaptive designs have not been studied for any of the mentioned alternative analyses.

The proposed rule for selecting the final sample size when incorporating prior information on a nuisance parameter into the sample size re-estimation is an ad hoc adaptation of the frequentist rule for selecting the final sample size in a the nuisance parameter based sample size re-estimation. From a practical perspective, such a rule makes sense as it is easy to implement and thus would be the first choice when trying to incorporate prior information into the sample size re-estimation. I showed

that this ad hoc rule can result in under- or overpowered clinical trials when a prior-data conflict is present [47]. From a practical perspective, this makes the ad hoc rule undesirable as it increases the risk of conducting an unsuccessful clinical trial or a too large trial thereby defeating its purpose. However, the fact that the ad hoc rule can result in under- or overpowered clinical trials does not exclude the possibility that better rules for selecting the final sample size in a design with sample size re-estimation incorporating prior information exist. For instance, a decision theoretic approach for selecting the final sample size could results in a sample size re-estimation procedure incorporating prior information with more desirable characteristics. The concept of using decision theory to determine the sample size is not new, see Parmigiani and Inoue [66], and it has also been considered in the drug development context by Stallard [67], who presented a decision theoretic approach for calculating the sample size of a phase II clinical trial.

Bibliography

[1] Mendis, S., Puska, P., and Norrving, B. *Global Atlas on Cardiovascular Disease Prevention and Control*. 2011. URL: http://whqlibdoc.who.int/publications/2011/9789241564373_eng.pdf (visited on 12/17/2017).

[2] World Health Organization (WHO). *Cardiovascular diseases – Fact sheet*. 2017. URL: http://www.who.int/mediacentre/factsheets/fs317/en/ (visited on 12/17/2017).

[3] Fuster, V. "Global Burden of Cardiovascular Disease". In: *Journal of the American College of Cardiology* 64.5 (2014), pp. 520–522.

[4] Calvo, G., McMurray, J. J., Granger, C. B., Alonso-Garcia, Á., Armstrong, P., Flather, M., Gómez-Outes, A., Pocock, S., Stockbridge, N., Svensson, A., et al. "Large streamlined trials in cardiovascular disease". In: *European Heart Journal* 35.9 (2014), pp. 544–548.

[5] Jackson, N., Atar, D., Borentain, M., Breithardt, G., Eickels, M. van, Endres, M., Fraass, U., Friede, T., Hannachi, H., Janmohamed, S., et al. "Improving clinical trials for cardiovascular diseases: a position paper from the Cardiovascular Round Table of the European Society of Cardiology". In: *European Heart Journal* 37.9 (2015), pp. 747–754.

[6] Fordyce, C. B., Roe, M. T., Ahmad, T., Libby, P., Borer, J. S., Hiatt, W. R., Bristow, M. R., Packer, M., Wasserman, S. M., Braunstein, N., et al. "Cardiovascular drug development: is it dead or just hibernating?" In: *Journal of the American College of Cardiology* 65.15 (2015), pp. 1567–1582.

[7] Chow, S.-C. and Liu, J.-P. *Design and analysis of clinical trials: concepts and methodologies*. Vol. 2. John Wiley & Sons, 2008.

[8] Chow, S.-C. and Chang, M. "Adaptive design methods in clinical trials – a review". In: *Orphanet Journal of Rare Diseases* 3.1 (2008), pp. 1–13.

[9] Kairalla, J. A., Coffey, C. S., Thomann, M. A., and Muller, K. E. "Adaptive trial designs: a review of barriers and opportunities". In: *Trials* 13.145 (2012).

[10] Food and Drug Administration (FDA). *Guidance for Industry – Adaptive Design Clinical Trials for Drugs and Biologics*. 2010. URL: https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf (visited on 12/17/2017).

[11] European Medicines Agency (EMA). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive designs*. 2017. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf (visited on 12/17/2017).

[12] Phillips, A. J. and Keene, O. N. "Adaptive designs for pivotal trials: discussion points from the PSI adaptive design expert group". In: *Pharmaceutical Statistics* 5.1 (2006), pp. 61–66.

[13] Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. "Adaptive designs in clinical drug development – an executive summary of the PhRMA working group". In: *Journal of Biopharmaceutical Statistics* 16.3 (2006), pp. 275–283.

[14] Friede, T. and Kieser, M. "Sample size recalculation in internal pilot study designs: a review". In: *Biometrical Journal* 48.4 (2006), pp. 537–555.

[15] Friede, T. and Kieser, M. "Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation". In: *Pharmaceutical Statistics* 12.3 (2013), pp. 141–146.

[16] Friede, T., Mitchell, C., and Müller-Velten, G. "Blinded Sample Size Reestimation in Non-Inferiority Trials with Binary Endpoints". In: *Biometrical Journal* 49.6 (2007), pp. 903–916.

[17] Friede, T. and Schmidli, H. "Blinded sample size reestimation with count data: methods and applications in multiple sclerosis". In: *Statistics in Medicine* 29.10 (2010), pp. 1145–1156.

[18] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). *Statistical Principles for Clinical Trials E9*. 1998. URL: `http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf` (visited on 12/17/2017).

[19] Committee for Medicinal Products for Human Use (CHMP). *Guideline on data monitoring committees*. 2005. URL: `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003635.pdf` (visited on 12/17/2017).

[20] Bauer, P., Bretz, F., Dragalin, V., König, F., and Wassmer, G. "Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls". In: *Statistics in Medicine* 35.3 (2016), pp. 325–347.

[21] Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. "Adaptive designs for confirmatory clinical trials". In: *Statistics in Medicine* 28.8 (2009), pp. 1181–1217.

[22] Bauer, P. and Einfalt, J. "Application of adaptive designs – a review". In: *Biometrical Journal* 48.4 (2006), pp. 493–506.

[23] Hatfield, I., Allison, A., Flight, L., Julious, S. A., and Dimairo, M. "Adaptive designs undertaken in clinical research: a review of registered clinical trials". In: *Trials* 17.150 (2016), pp. 1–13.

[24] Anker, S. D. and McMurray, J. J. "Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials?" In: *European Heart Journal* 33.22 (2012), pp. 2764–2765.

[25] Anker, S. D., Schroeder, S., Atar, D., Bax, J. J., Ceconi, C., Cowie, M. R., Crisp, A., Dominjon, F., Ford, I., Ghofrani, H.-A., et al. "Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency". In: *European Journal of Heart Failure* 18.5 (2016), pp. 482–489.

[26] Solomon, S. D., Rizkala, A. R., Gong, J., Wang, W., Anand, I. S., Ge, J., Lam, C. S., Maggioni, A. P., Martinez, F., Packer, M., et al. "Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: rationale and design of the PARAGON-HF trial". In: *JACC: Heart Failure* 5.7 (2017), pp. 471–482.

[27] Rogers, J. K., Pocock, S. J., McMurray, J. J., Granger, C. B., Michelson, E. L., Östergren, J., Pfeffer, M. A., Solomon, S. D., Swedberg, K., and Yusuf, S. "Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved". In: *European Journal of Heart Failure* 16.1 (2014), pp. 33–40.

[28] Lin, D., Wei, L., Yang, I., and Ying, Z. "Semiparametric regression for the mean and rate functions of recurrent events". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4 (2000), pp. 711–730.

[29] Jennison, C. and Turnbull, B. W. *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

[30] Cook, R. J. and Lawless, J. F. "Interim monitoring of longitudinal comparative studies with recurrent event responses". In: *Biometrics* 52.4 (1996), pp. 1311–1323.

[31] Jiang, W. "Group sequential procedures for repeated events data with frailty". In: *Journal of Biopharmaceutical Statistics* 9.3 (1999), pp. 379–399.

[32] Cook, R. J., Grace, Y. Y., and Lee, K.-A. "Sequential Testing with Recurrent Events over Multiple Treatment Periods". In: *Statistics in Biosciences* 2.2 (2010), pp. 137–153.

[33] Poulter, N., Prabhakaran, D., and Caulfield, M. "Hypertension". In: *The Lancet* 386.9995 (2015), pp. 801–812.

[34] Committee for Medicinal Products for Human Use (CHMP). *Guideline on clinical investigation of medicinal products in the treatment of hypertension*. 2016. URL: `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2016/07/WC500209943.pdf` (visited on 12/17/2017).

[35] Krum, H., Viskoper, R. J., Lacourciere, Y., Budde, M., and Charlon, V. "The effect of an endothelin-receptor antagonist, bosentan, on blood pressure in patients with essential hypertension". In: *New England Journal of Medicine* 338.12 (1998), pp. 784–791.

[36] Elliott, W. J., Whitmore, J., Feldstein, J. D., and Bakris, G. L. "Efficacy and safety of perindopril arginine+ amlodipine in hypertension". In: *Journal of the American Society of Hypertension* 9.4 (2015), pp. 266–274.

[37] Kieser, M. and Friede, T. "Simple procedures for blinded sample size adjustment that do not affect the type I error rate". In: *Statistics in Medicine* 22.23 (2003), pp. 3571–3581.

[38] Friede, T. and Kieser, M. "Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis". In: *Pharmaceutical Statistics* 10.1 (2011), pp. 8–13.

[39] Kieser, M. and Friede, T. "Blinded sample size reestimation in multiarmed clinical trials". In: *Drug information journal* 34.2 (2000), pp. 455–460.

[40] Koch, A. and Röhmel, J. "Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference". In: *Journal of Biopharmaceutical Statistics* 14.2 (2004), pp. 315–325.

[41] Schmidli, H., Neuenschwander, B., and Friede, T. "Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials". In: *Computational Statistics & Data Analysis* 113 (2017), pp. 100–110.

[42] Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. "Robust meta-analytic-predictive priors in clinical trials with historical control information". In: *Biometrics* 70.4 (2014), pp. 1023–1032.

[43] Gould, A. L. "Interim analyses for monitoring clinical trials that do not materially affect the type I error rate". In: *Statistics in Medicine* 11.1 (1992), pp. 55–66.

[44] Hartley, A. M. "Adaptive blinded sample size adjustment for comparing two normal means—a mostly Bayesian approach". In: *Pharmaceutical Statistics* 11.3 (2012), pp. 230–240.

[45] Glynn, L. G., Murphy, A. W., Smith, S. M., Schroeder, K., and Fahey, T. "Interventions used to improve control of blood pressure in patients with hypertension". In: *The Cochrane Library* 3.CD005182 (2010). URL: `http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD005182.pub4/abstract`.

[46] Mütze, T. and Friede, T. "Blinded sample size re-estimation in three-arm trials with 'gold standard' design". In: *Statistics in Medicine* 36.23 (2017), pp. 3636–3653.

[47] Mütze, T., Schmidli, H., and Friede, T. "Sample size re-estimation incorporating prior information on a nuisance parameter". In: *Pharmaceutical Statistics* 17.2 (2018), pp. 126–143.

[48] Mütze, T., Glimm, E., Schmidli, H., and Friede, T. "Group sequential designs for negative binomial outcomes". In: *Statistical Methods in Medical Research* (accepted) (2018). URL: `https://doi.org/10.1177/0962280218773115`.

[49] Mütze, T., Glimm, E., Schmidli, H., and Friede, T. "Group sequential designs with robust semiparametric recurrent event models". In: *Statistical Methods in Medical Research* (accepted) (2018). URL: `https://doi.org/10.1177/0962280218780538`.

[50] Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. "Semiparametric efficiency and its implication on the design and analysis of group-sequential studies". In: *Journal of the American Statistical Association* 92.440 (1997), pp. 1342–1350.

[51] Yusuf, S. et al. "Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial". In: *The Lancet* 362.9386 (2003), pp. 777–781.

[52] Lan, G. and DeMets, D. "Discrete sequential boundaries for clinical trials". In: *Biometrika* 70.3 (1983), pp. 659–663.

[53] Mütze, T. *gscounts: Group Sequential Designs with Negative Binomial Outcomes*. R package version 0.1-1. 2018. URL: `https://CRAN.R-project.org/package=gscounts`.

[54] Berger, R. L. "Multiparameter hypothesis testing and acceptance sampling". In: *Technometrics* 24.4 (1982), pp. 295–300.

[55] Stucke, K. and Kieser, M. "A general approach for sample size calculation for the three-arm 'gold standard' non-inferiority design". In: *Statistics in Medicine* 31.28 (2012), pp. 3579–3596.

[56] Xing, B. and Ganju, J. "A method to estimate the variance of an endpoint from an on-going blinded trial". In: *Statistics in Medicine* 24.12 (2005), pp. 1807–1814.

[57] Morita, S., Thall, P. F., and Müller, P. "Determining the effective sample size of a parametric prior". In: *Biometrics* 64.2 (2008), pp. 595–602.

[58] Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. "Summarizing historical information on controls in clinical trials". In: *Clinical Trials* 7.1 (2010), pp. 5–18.

[59] Mütze, T. *varmap: Tools Related to MAP Priors for Variances*. R package version 0.0-1. 2017. URL: `https://github.com/tobiasmuetze/varmap`.

[60] Friede, T., Häring, D., and Schmidli, H. "Blinded continuous monitoring in clinical trials with recurrent event endpoints". In: *Submitted* (2018).

[61] Scharfstein, D. O. and Tsiatis, A. A. "The use of simulation and bootstrap in information-based group sequential studies". In: *Statistics in Medicine* 17.1 (1998), pp. 75–87.

[62] Mao, L. and Lin, D. "Semiparametric regression for the weighted composite endpoint of recurrent and terminal events". In: *Biostatistics* 17.2 (2016), pp. 390–403.

[63] Ghosh, D. and Lin, D. Y. "Nonparametric analysis of recurrent events and death". In: *Biometrics* 56.2 (2000), pp. 554–562.

[64] Rauch, G., Rauch, B., Schüler, S., and Kieser, M. "Opportunities and challenges of clinical trials in cardiology using composite primary endpoints". In: *World Journal of Cardiology* 7.1 (2015), p. 1.

[65] Pocock, S. J., Ariti, C. A., Collier, T. J., and Wang, D. "The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities". In: *European Heart Journal* 33.2 (2011), pp. 176–182.

[66] Parmigiani, G. and Inoue, L. *Decision theory: principles and approaches*. Vol. 812. John Wiley & Sons, 2009.

[67] Stallard, N. "Sample size determination for phase II clinical trials based on Bayesian decision theory". In: *Biometrics* (1998), pp. 279–294.