Aus dem Institut für Biometrie und Klinische Epidemiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Evaluation of the performance criteria of optimal futility stopping boundaries in flexible designs

Evaluation der Leistungskriterien der optimalen Grenzen für das Stoppen aufgrund der Aussichtslosigkeit bei flexiblen Designs

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Xieran Li
aus Peking

Datum der Promotion: 25.06.2023

Table of content

1 List of figures and tables

List of Figures

List of Tables

2 Abstract

Group sequential design and adaptive design are flexible designs that are frequently applied in clinical trials. Unlike fixed designs, flexible designs allow for statistical inferences on trial endpoints prior to complete data collection. Such early inferences on a trial may lead to different decisions regarding trial continuation after the interim analyses. If the treatment effect can already be shown, the trial may be stopped early for efficacy. On the contrary, if the interim inference indicates a small treatment effect, the trial may be stopped early for futility. Various options for efficacy and futility stopping boundaries have been proposed in the statistical literature. However, futility boundaries are often chosen without the thorough planning of operational characteristics and evaluation of design performance. In this research work, performance criteria in flexible designs with early futility stops are evaluated. Moreover, previous work from Schüler [28] is further developed to select the so-called ”optimal futility boundaries” [17]. The optimization approach is developed for trials with continuous or binary endpoints. Application examples of real clinical trials demonstrate the advantages of the new optimal approach and have evaluated the performance criteria in various flexible designs. The results indicate that the optimal futility stopping boundaries simultaneously minimize the probability of wrongly stopping for futility and power loss. Additionally, boundaries from the optimal approach improve the probability of correctly stopping for futility early. In conclusion, it is recommended to investigate and optimize futility boundaries thoroughly at the planning stage of a clinical trial to achieve greater design efficiency.

3 Zusammenfassung

Gruppensequenzielles Design und adaptives Design sind flexible Designs, die häufig in klinischen Studien angewendet werden. Anders als bei festen Designs, ermöglichen flexible Designs vor der vollständigen Datenerfassung statistische Inferenzen auf Studienendpunkte zu ziehen. Eine solche frühe Inferenz kann zum Zeitpunkt der Zwischenanalysen zu unterschiedlichen Entscheidungen über die Fortsetzung der Studie führen. Bei validiertem Behandlungseffekt kann die Studie wegen Wirksamkeit vorzeitig abgeschlossen werden. Im Gegenteil kann die Studie vorzeitig wegen Aussichtslosigkeit abgebrochen werden, wenn die vorläufige Schlussfolgerung auf einen geringen Behandlungseffekt hinweist. In der statistischen Literatur finden sich bereits diverse Optionen für das Stoppen aufgrund der Wirksamkeit und Aussichtslosigkeit. Die Wahl der Grenzen für das Stoppen aufgrund der Aussichtslosigkeit erfolgt allerdings oft ohne gründliche Planung der operativen Eigenschaften und Evaluation der Güte von Designs. In dieser Forschungsarbeit werden Leistungskriterien in flexiblen Designs mit frühem Stoppen aufgrund der Aussichtslosigkeit evaluiert und frühere Arbeiten von Schüler [28] weiterentwickelt, um sogenannte ”optimale Grenzen für das Stoppen aufgrund der Aussichtslosigkeit” [17] auszuwählen. Der Optimierungsansatz wurde für Studien mit kontinuierlichen oder binären Endpunkten entwickelt. Echte klinische Studien werden als Anwendungsbeispiele verwendet, um die Vorteile des neuen optimalen Ansatzes zu demonstrieren und die Leistungskriterien in verschiedenen flexiblen Designs zu bewerten. Die Ergebnisse zeigen, dass die optimalen Grenzen für das Stoppen aus Aussichtslosigkeit sowohl die Wahrscheinlichkeit eines falschen Stoppens aus Aussichtslosigkeit als auch den Verlust der Trennschärfe gleichzeitig minimieren. Zusätzlich verbessert der optimale Ansatz die Wahrscheinlichkeit, frühzeitig korrekt wegen Aussichtslosigkeit aufzuhören. Schließlich wird empfohlen, die Grenzen in der Planungsphase einer klinischen Studie gründlich zu untersuchen und zu optimieren, um eine höhere Designeffizienz zu erreichen.

4 Synopsis

4.1 Introduction

Designing a clinical trial that balances resources and scientific impact can be challenging. There are constant changes due to new research from scientific communities and new guidelines from regulatory agencies. Critical attention should be given to the protection of the rights, safety, and well-being of trial participants. When attempting to reduce the time and financial resources required for research, clinical trials based on traditional fixed designs are less favorable. For example, major adjustments to an ongoing trial require a trial protocol amendment. On the contrary, trials with flexible designs allow prospectively planned modifications based on accumulated data without changing the protocol.

Although contributing to only 2.6% of the trials on PubMed, phase II and III trials planned with flexible designs are more likely to be completed than those with fixed designs [32]. Such flexible designs allow for trial modification, including early trial termination for efficacy and futility or the adaptation of sample size. The application of early stopping can reduce the cost and patient risk if a trial may already achieve the research objectives at the interim analysis or terminate early for futility.

A multiple testing problem is caused by such prospectively planned interim analyses and trial modifications. Initial methods for controlling type-I error via alpha-spending methods were developed exclusively for group sequential designs by Pocock [23] and O'Brien and Fleming [20]. The key ideas of the alpha-spending method were later extended to diverse flexible applications in terms of the timing and actual information acquired up to the interim analysis [4, 13, 33]. For adaptive designs with sample size re-estimation, the Fisher's method [1] and inverse normal method [15] are commonly used to control overall type-I error by combining the data from each stage.

Although there is an increasing interest in flexible designs, in practice, common applications often only contain the option of early stopping for efficacy. Notably, the method of an early stop for futility is less commonly researched. Between the two purposes of early trial termination at an interim stage, efficacy is considered a positive event because it leads to the successful conclusion of treatment benefits. On the contrary, terminating a trial due to futility is a difficult decision to make. The trial sponsors and investigators must accept such an early stop as a sunk cost of their spent resources. Moreover, a futility stop decreases the possibility of positive secondary findings from a negatively terminated trial [27]. Despite the negative consequences, early stopping for futility is an important trial design feature used to safeguard resources and ethics. Additionally, incorrectly stopping for futility is also a waste of resources and puts patients at risk without scientific impact. A proper boundary for futility stopping should increase the efficiency of a trial design. Therefore, the current research on futility stopping boundaries can be observed in both large phase III trials and small non-controlled phase II trials [30], while the futility boundaries are often different from the efficacy boundaries by being non-binding. For binding futility boundaries, once the result from the interim analysis crosses the boundary, stopping the trial for futility becomes mandatory. When combined with an early stopping for efficacy, the binding futility boundaries may also contribute to the choice of efficacy boundaries if desired, so that the efficacy boundaries can fully exhaust the global significance level [2]. While the non-binding futility boundaries do not have the same features as the binding boundaries, they offer more flexibility. The non-binding boundaries are treated as an optional recommendation, while the decision to stop for futility is weighted alongside other factors (e.g., the secondary analysis and external information). Typically, a data monitoring committee is established to independently evaluate the safety data. This type of committee may also independently evaluate the interim analysis of efficacy data and check the crossing of a futility stopping boundary. Since a data monitoring committee only makes recommendations based on the boundaries, with the trial sponsor takes the final decision, non-binding boundaries are better suited for this process. Therefore, the non-binding type plays a more important role in clinical research practices and deserves greater attention for methodology development.

Methodologies for futility stopping rules were proposed in the literature decades ago. The first group of methods is beta-spending functions, which is classified as a frequentist approach. Beta-spending functions are similar to the alpha-spending functions initially designed for early stopping for efficacy in group sequential designs, except they control for type-II error at β and the stopping boundaries are expressed in probabilities [4, 3, 21, 25]. The second type of approaches is called stochastic curtailment [14, 12]. In these approaches, the boundaries are prospectively chosen based on conditional power (CP), which is defined as the probability of rejecting the null hypothesis at the final analysis based on the observed interim data and certain treatment effect assumptions. Apart from frequentist approaches, there is also a similar Bayesian approach based on predictive power [26, 16, 9].

The performance and operational characteristics of trial designs are further evaluated after the inclusion of futility stopping boundaries at the interim analysis. Some examples of operational characteristics can include the maximum and expected sample size due to time and financial limitation [8, 22]. The boundaries are optimized based on a combination of operational characteristics of investigators' preferences. For example, an investigator might be interested in the actual benefit of stopping a trial early if there is no treatment effect. Alternatively, the overall probability of success for a trial or the expected sample size could be the main focus. However, it remains unclear how these different criteria can be weighted against each other. Liu et al. [19] proposed a performance score of trial designs based on the combination of final sample size and power for adaptive designs. However, the criterion of sample size may not be applicable to all designs and is not specific to early stopping for futility. Different from a single performance score, several performance criteria, including the probability of wrongly stopping, power loss, and probability of correctly stopping, are jointly considered in my project. These criteria are chosen to cover performance at both interim stages locally (namely the probabilities of wrongly and correctly stopping) and the power loss due to futility at global level. Notably, these criteria can easily be understood and communicated between statisticians and physicians. Schüler [28] proposed futility stopping boundaries optimized by such characteristics in a special two-stage group sequential design where the type of primary endpoints is time-to-event and the boundaries are restricted in the scale of probabilities. The approach from

Schüler is extended in this project to continuous endpoints [17] and binary endpoints. Additionally, this project investigates the performance based on various scales of futility boundaries and for several group sequential and adaptive designs.

My research aims to quantify the performance criteria in group sequential and adaptive designs with futility stopping boundaries. Individual criteria can be customized to emphasize different aspects of trial performance. Subsequently, another aim of the project is to provide an algorithm for futility stopping boundary optimization for various designs and endpoint types. Under the framework, open and proactive dialogue is encouraged between statisticians and clinicians in the initial trial design phase so that different designs with optimal futility boundaries can be prospectively compared to achieve maximum trial efficiency.

The dissertation first defines the performance criteria and optimization algorithm in the Methods section. The Methods section is further divided into two parts for continuous data and binary data, respectively. For each type of endpoint, several methods are developed to cover different flexible designs, including a design for non-controlled trials with binary data. Next, the algorithm is applied to hypothetical settings and real clinical trials, as presented in the Results section. The benefits of the optimal approach are also demonstrated. Finally, the conclusions of my research are given in the Discussion section.

4.2 Methods

Continuous data

Consider a randomized controlled trial that compares two treatment groups. T denotes the treatment group and C denotes the control group.

For normally distributed continuous data, the observations are denoted as

$$X_i^T \sim \mathcal{N}(\mu^T, \sigma^2), i = 1 \dots n^T \text{ and } X_i^C \sim \mathcal{N}(\mu^C, \sigma^2), i = 1 \dots n^C \quad (1)$$

As is often the case in practice, the allocation of two groups is balanced so that $n^T = n^C = n$. Additionally, a known common standard deviation σ is often assumed in practice. The trial hypothesis can be written as

$$H_0 : \mu^T - \mu^C \leq 0 \text{ versus } H_1 : \mu^T - \mu^C > 0 \quad (2)$$

The direction of the aforementioned hypothesis indicates an effective treatment group compared to the control group if the endpoint of interest has a higher numeric value. The hypothesis is constructed for superiority testing. If the objective of a trial is to establish non-inferiority, 0 should be replaced by a non-inferiority margin.

Assuming a large enough sample size n, a Z-test based on normal distribution can be used for hypothesis testing. The test statistic Z with n and sample means $\overline{X^T}$ and $\overline{X^C}$ can be expressed as

$$Z = \frac{\overline{X^T} - \overline{X^C}}{\sigma\sqrt{\frac{1}{n^T} + \frac{1}{n^C}}} = \frac{\overline{X^T} - \overline{X^C}}{\sigma} * \sqrt{\frac{n}{2}} \quad (3)$$

If σ is unknown, it can be estimated by the pooled sample standard deviation $S_{pooled} = \sqrt{\frac{S^{T^2} + S^{C^2}}{2}}$ based on the observed data. It is shown that the estimation has a minimal impact on the overall α [24].

After trial completion, the hypothesis H_0 shall be rejected and a treatment benefit is demonstrated if $Z \geq z_{1-\alpha}$. The one-sided significance level α and the power Pow can be formulated in terms of probabilities as

$$\alpha = P_{H_0}(Z \geq z_{1-\alpha}) \text{ and } Pow = 1 - \beta = P_{H_1}(Z \geq z_{1-\alpha}) \quad (4)$$

The sample size n is determined for a pre-specified level of α and Pow (e.g., $\alpha = 0.025$ and $Pow = 0.8$) with an assumed standardized treatment effect of $\frac{\theta}{\sigma} > 0$.

$$n = \frac{2(z_{1-\alpha} + z_{1-\beta})^2}{(\frac{\theta}{\sigma})^2} \tag{5}$$

Compared to the standard one-stage fixed design, group sequential designs offer more flexibility by allowing multiple stages with interim analyses before the final analysis. Different from the fixed design, test statistics from a multi-stage design are based on the data from each stage only or all data cumulatively collected until the final analysis. In this work, the number of stages j is set to 2 for illustration. Let $\overline{X_j^T}$ and $\overline{X_j^C}$ denote the sample means observed at the stage $j = 1, 2$ for the treatment and control groups, based on the data exclusively collected during the stage j. Given a balanced design $n_j^T = n_j^C = n_j$, the stage j test statistics from (3) can similarly be expressed as

$$Z_j = \frac{\overline{X_j^T} - \overline{X_j^C}}{\sigma} * \sqrt{\frac{n_j}{2}} \tag{6}$$

For the final stage test statistics, $\overline{X_{1+2}^T}$ and $\overline{X_{1+2}^C}$ denote the sample means based on all data cumulatively collected at the final stage. The test statistic extends the function (3) to

$$\begin{aligned} Z_{1+2} &= \frac{\overline{X_{1+2}^T} - \overline{X_{1+2}^C}}{\sigma} * \sqrt{\frac{n}{2}} \\ &= \sum_{j=1}^{2} \frac{\overline{X_j^T} - \overline{X_j^C}}{\sigma} * \sqrt{\sum_{j=1}^{2} \frac{n_j}{2}} \\ &= \frac{\sum_{j=1}^{2} \sqrt{n_j} Z_j}{\sqrt{\sum_{j=1}^{2} n_j}} \end{aligned} \tag{7}$$

Therefore, Z_{1+2} can be expressed as a combination of the stage-wise test statistics Z_j. Consequently, the covariance between the final stage and the first stage is fully specified by the information acquired in terms of the sample size n_j and n so that

$$Cov(Z_1, Z_{1+2}) = \sqrt{\frac{n_1}{n}} \tag{8}$$

Stopping for futility

Considering a two-stage trial allowing an early stop for futility, a futility stopping boundary α_f is defined as when

$$Z_1 \leq z_{1-\alpha_f} \tag{9}$$

and the trial may be stopped early for futility.

There are two types of α_f: binding and non-binding. If a futility boundary is binding, the efficacy boundary α of the final stage in (4) may be adjusted by incorporating the futility boundary α_f from the first stage to improve efficiency if desired. However, if data monitoring committees and trial sponsors do not strictly follow the binding rules, the type-I error is inflated above the predefined α. On the contrary, a non-binding futility stopping boundary only works as a guiding signal so that the decision to stop a trial early due to futility can be made based on the interim result and other information. For example, other secondary endpoints from the trial may suggest a medical and scientific benefit to continuing the trial to the end, even after a non-binding futility boundary is crossed. Even reviewers at the U.S. Food and Drug Administration suggest that trial investigators consider the non-binding type [18]. Therefore, in this work, non-binding futility boundaries are constructed independently, after the sample size being determined. Two-stage designs only allowing early stopping for non-binding futility do not inflate type-I error, but rather reduce it. However, the performance evaluation of a design can only be performed if futility stopping boundaries are considered mandatory. The same critical value $z_{1-\alpha}$ of the fixed design remains valid and the final stage test statistics reject H_0 if $Z_{1+2} \geq z_{1-\alpha}$. The protection of type-I error is feasible because $P_{H_0}(Z_1 > z_{1-\alpha_f} \cap Z_{1+2} \geq z_{1-\alpha}) < \alpha$.

However, the type-II error can be inflated above β due to the additional stop for futility and the power loss Pow_{loss}. Moreover, a futility stop can affect several other trial operational characteristics (e.g., the probability of wrongly stopping for futility π_{wrong}). The trial statistician is responsible for making the clinicians aware of such impacts before calculating the sample size and defining the stopping boundaries in the trial protocol. A futility boundary that is not optimally chosen can lead to undesired trial performance.

In my publication [17], various criteria to optimize operational characteristics were analyzed as an extension of Schüler's work [28]. One of the key criteria focuses on the prevention of mistakenly terminate a trial for futility. Many investigators are particularly concerned about wrongly stopping for futility and thereby leading to an unsuccessful trial. An evaluation of the other secondary endpoints (apart from the primary endpoint) is also affected by premature termination because trials are generally not powered for their secondary endpoints even with complete data collection. However, a small $\alpha_f = 0.10$ makes a correct stop for futility more difficult, especially for a small treatment effect. On the other hand, a generous boundary $\alpha_f = 0.80$ greatly inflates the probability of wrongly stopping for futility and decreases the overall power. Multiple futility boundaries satisfy the conditions for both losses of power and the probability of wrongly stopping for futility. Therefore, other operational characteristics (e.g., the expected sample size) are proposed to derive the optimal futility boundaries among all possible boundaries in the previous research [31]. In the optimal approach, the probability of correctly stopping is chosen as the third performance criteria for optimization. It is motivated by the main objective of futility assessment to correctly save resources when the true treatment effect is not clinically beneficial.

To quantify performance based on the futility stopping boundary, the performance metrics should first be characterized. In an extension of Schüler's work [28], where both early stops for efficacy and futility were allowed at the interim analysis, the first design in my research only considers an early stop for futility. The first two conditions based on the concepts of π_{wrong} and Pow_{loss} are characterized as

Condition 1

$$\pi_{wrong} \geq P_{H_1}(Z_1 \leq z_{1-\alpha_f}) \tag{10}$$

Condition 2

$$Pow_{loss} \geq 1 - \beta - P_{H_1}(Z_1 > z_{1-\alpha_f} \cap Z_{1+2} \geq z_{1-\alpha}) \tag{11}$$

As previously discussed, many α_f fulfill both conditions, while the probability of correctly stopping for futility is included in the next step for optimization. Any smaller treatment effect $\theta_1 \in [0, \theta)$ can be chosen for evaluation. Some investigators may wish to set a minimum level of probability of correctly stopping for futility as a safety net against the continuation of trials involving an ineffective treatment. More generally, the investigators can simply rely on the achievable maximum probability of correctly stopping. Thus, condition 3 for the optimization is characterized as the probability.

Condition 3

$$\pi_{correct,\theta_1} \geq P_{H_{\theta_1}}(Z_1 \leq z_{1-\alpha_f}) \tag{12}$$

Let $A_{\pi_{wrong},Pow_{loss}}$ be the set of all α_f that fulfill conditions 1 and 2. $\alpha_{f,opt}$ denotes the optimal element from the set and constraint of condition 3. The optimal futility boundary $\alpha_{f,opt}$ is found when

$$\alpha_{f,opt} = \max_{\alpha_f \in A_{\pi_{wrong},Pow_{loss}}} \pi_{correct,\theta_1} \tag{13}$$

In my publication [17], the conditions are specified for continuous endpoints to search for the optimal boundary $\alpha_{f,opt}$ based on all three conditions iteratively. In first step, the functions (10) and (11) are transcribed into the standard normal cumulative distribution Φ and the multivariate normal cumulative distribution $MV_{\mu,\Sigma}$ as

Condition 1

$$\pi_{wrong} \geq \Phi(z_{1-\alpha_f} - \frac{\theta}{\sigma}\sqrt{\frac{n_1}{2}}) \tag{14}$$

Condition 2

$$Pow_{loss} \geq MV_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(z_{1-\alpha_f}, z_{1-\alpha}) - \beta \tag{15}$$

The mean $\boldsymbol{\mu}$ and standard deviation Σ matrices of the multivariate normal cumulative distribution are

$$\boldsymbol{\mu} = \quad (\tfrac{\theta}{\sigma}\sqrt{\tfrac{n_1}{2}}, \tfrac{\theta}{\sigma}\sqrt{\tfrac{n}{2}}) \tag{16}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sqrt{\frac{n_1}{n}} & 1 \\ 1 & \sqrt{\frac{n_1}{n}} \end{pmatrix} \tag{17}$$

The iterative search finds the boundary $\alpha_{f,opt}$ by increasing or decreasing α_f so that conditions 1 and 2 are both fulfilled. The search stops at the smallest $\alpha_{f,opt}$ that gives the largest to condition 3 or any pre-specified desired level of $\pi_{correct,\theta_1}$. This concludes the core of the optimal approach in this research, starting with a simple two-stage design with a stop for futility only.

In the trial planning phase, both π_{wrong} and Pow_{loss} should be predefined with a maximum limit deemed acceptable by the investigators. The higher probability of correctly stopping $\pi_{correct,\theta_1}$ corresponds to smaller α_f because it is more difficult to cross a smaller boundary for futility. Therefore, the $\alpha_{f,opt}$ is optimized according to condition 3 in (12) by finding the minimum in the set $A_{Pow_{loss},\pi_{wrong}}$. It is in the interest of the trial investigators to know the probability of correctly stopping during the planning at least. They may also opt to set a minimum acceptable value for $\pi_{correct,\theta_1}$. For example, if conditions 1 and 2 are planned as overly optimistic, the desired condition 3 $\pi_{correct,\theta_1} \geq 0.6$ might yield an $\alpha_{f,opt}$ that is not necessarily the minimum of the set. On the other hand, if $\pi_{correct,\theta_1}$ is too small, the addition of futility stopping to the trial does not provide any benefit and complicates the trial operation.

The optimal futility boundaries $\alpha_{f,opt}$ are probabilities, as presented in the publication [17]. However, there are other popular scales for futility boundaries. In my research, the optimal approach is further extended to show that the method can still be applied if other scales are chosen by the investigators. Since $\alpha_{f,opt}$ is derived based on Z-score, $z_{1-\alpha_{f,opt}}$ is an obvious alternative scale. Another commonly applied boundary is based on CP, which is defined as the probability of rejecting H_0 at the final analysis given the observed interim data [10]. The main advantage of CP over $\alpha_{f,opt}$ or $z_{1-\alpha_{f,opt}}$ scale is the intuitive interpretation of the futility boundary for decisions at the interim stage since $\alpha_{f,opt}$ and $z_{1-\alpha_{f,opt}}$ are more abstract concepts for clinicians. Assuming normality and future data after the interim stage to follow the initial standardized treatment effect $\frac{\theta}{\sigma}$, the CP is derived as

$$CP = \Phi\{\frac{-z_{1-\alpha} + z_{1-\alpha_{f,opt}}\sqrt{\frac{n_1}{n}} + \frac{\theta}{\sigma}\sqrt{\frac{n}{2}}(1-\frac{n_1}{n})}{\sqrt{1-\frac{n_1}{n}}}\} \quad (18)$$

Notably, different treatment effect assumptions could be made for (18). For example, instead of using θ directly, observed data at the interim analysis can be incorporated to estimate θ for future data. Also, σ alone or both θ and σ can be updated given the observed data. The differences in CP caused by the assumptions are not discussed further in this work because my research focuses on the benefits of the optimal approach and shows that the method remains viable regardless of the chosen scale. Since the test statistics Z_1 can be found in the functions of $\alpha_{f,opt}$ and CP scales (9, 18) and directly compared with $z_{1-\alpha_{f,opt}}$, the scales are interchangeable.

Stopping for futility and efficacy

If a two-stage design allows an early stop for efficacy, the rejection of H_0 can occur at either stage 1 if $Z_1 \geq z_{1-\alpha_1}$ or the final stage if $Z_{1+2} \geq z_{1-\alpha_{1+2}}$. Instead of only a single boundary α for efficacy at the final stage, there are two efficacy boundaries α_1 and α_{1+2} with data cumulative collected at stage 1 and the final stage, respectively. The type-I error should be controlled while considering both stages as follows

$$P_{H_0}(Z_1 \geq z_{1-\alpha_1} \cup (Z_1 < z_{1-\alpha_1} \cap Z_{1+2} \geq z_{1-\alpha_{1+2}})) = \alpha \tag{19}$$

To derive α_1 and α_{1+2} from (19), the alpha-spending function method is often applied, with several variations. A function with constant local levels for each interim stage was proposed by Pocock [23]. Another popular function from O'Brien-Fleming [20] tends to spend less local significance at early stages and more at the later stages. Other functions are also available in the literature. For example, Lan and DeMets proposed for more flexibility in the timing of the interim stages [13]. Since methods of efficacy boundaries are not the focus of this research, Pocock's boundaries (where $\alpha_1 = \alpha_{1+2}$) are selected for simplicity.

The method developed in this section allows an early stop for either futility or efficacy. Similar to the design with futility only in the previous section, the type-I error is still controlled after the inclusion of α_f to (19) so that

$$P_{H_0}(Z_1 \geq z_{1-\alpha_1} \cup (z_{1-\alpha_f} < Z_1 < z_{1-\alpha_1} \cap Z_{1+2} \geq z_{1-\alpha_{1+2}})) < \alpha \tag{20}$$

A similar trial design was assumed in the previous work by Schüler for time-to-event endpoints [28] and this is extended to continuous endpoints in the publication presented in this dissertation [17]. In my work, the three conditions for the optimal approach of (10), (11) and (12) are further expanded to incorporate the boundaries α_1 and α_{1+2} instead of α alone. Since condition 1 in (10) and 3 in (11) are only affected by α_f, they remain valid regardless of the addition of an early stop for efficacy. Only condition 2 in (15), as a global criterion, should be adjusted to reflect both stages so that

Condition 2

$$\begin{aligned} 1 - \Phi(z_{1-\alpha_1} - \frac{\theta}{\sigma}\sqrt{\frac{n_1}{2}}) + MV_{\mu,\Sigma}(z_{1-\alpha_1}, z_{1-\alpha_{1+2}}) \\ - MV_{\mu,\Sigma}(z_{1-\alpha_f}, z_{1-\alpha_{1+2}}) \geq 1 - \beta - Pow_{loss} \end{aligned} \tag{21}$$

Stopping for futility and sample size re-estimation

Apart from early termination, other trial features may be altered during the trial under an adaptive design (e.g. sample size re-estimation). The optimal approach from my work not only applies to group sequential designs but also adaptive designs with sample size re-estimation. An adequate sample size is vital to increase the power of the final analysis after the interim analysis, and there are two major categories for methods of sample size adaptation. The first type of method relies on non-comparative results. However, the sample size can also be calculated based on comparative data (e.g., using the observed θ or σ directly or by a certain CP [11]), which is also used by the optimal approach for trials with two groups. The optimal approach further improves the performance of the adaptive design, which is evaluated based on the same three conditions. Since the focus of this work is the futility stopping boundaries, the method focuses on a trial design that combines sample size re-estimation and an early stop for futility, without any early stops for efficacy.

To illustrate the benefit of the optimal approach of futility boundaries in adaptive designs, a sample size of the stage 2 n_2 is recalculated based on the observed interim $\widehat{\theta}$ in this work. In a two-stage design, the recalculated incremental sample size n_2^* after the interim analysis is associated with $2(\frac{\sigma}{\widehat{\theta}})^2(z_{1-\alpha} + z_{1-\beta})^2 - n_1$. Combined with $\alpha_{f,opt}$ from the optimal approach for early futility stopping, the rule for adaptation is

$$n_2^* = \begin{cases} 0 & \text{if } Z_1 \leq z_{1-\alpha_{f,opt}} \\ 2(\frac{\sigma}{\theta})^2(z_{1-\alpha} + z_{1-\beta})^2 - n_1 & \text{if } Z_1 > z_{1-\alpha_{f,opt}} \end{cases} \tag{22}$$

If the optimal boundary $\alpha_{f,opt}$ is crossed, the recruitment may completely stop at the interim stage so that $n_2^* = 0$. Otherwise, it continues with an adapted sample size. Moreover, to avoid recruiting too many patients beyond the capacity of investigators, the pre-specified maximum is set to be $n_2^* \leq 2n_2$.

To derive the $\alpha_{f,opt}$ in an adaptive design, the same three conditions are applied. Conditions 1 and 3 are based solely on the information and assumptions up to the interim analysis and remain unchanged. Due to the change in sample size, the function (7) based on the fixed n is no longer valid for Z_{1+2}. For this purpose, the inverse normal approach [15] is used in this work to combine stage-wise test statistics Z_j. The overall Z_{1+2} and covariance from (7) and (8) as part of the Pow_{loss} can be explicitly expressed with weights w_1 and w_2 as

$$Z_{1+2} = \frac{w_1 Z_1 + w_2 Z_2}{\sqrt{w_1^2 + w_2^2}} \tag{23}$$

$$Cov(Z_1, Z_{1+2}) = \frac{w_1}{\sqrt{w_1^2 + w_2^2}} \tag{24}$$

In an adaptive design, the weights are defined *a priori*. One intuitive choice of w_j is made according to the initially planned sample size at the interim and final stages [11].

Lastly, since the sample size of stage 2 can now vary, based on the result of stage 1, an additional iterative step for each n_2^* over the range of possible Z_1 is implemented as part of the Pow_{loss} from condition 2 to search for the optimal $\alpha_{f,opt}$.

Binary data

Binary response variables can also be the primary endpoint of a trial (e.g., whether a patient is a responder (yes or no) to the treatment within 1 month). The research demonstrated in this section characterizes the optimal approach for the design with two variations. The first part of the methods is dedicated to a typical controlled trial similar to the continuous data section. The second part illustrates the optimal approach for a non-controlled one-group trial, which is often applied in phase II with limited finical resources.

Two groups with stopping for futility

If the comparison between the two groups T and C are based on response data, the same notations can be adopted as described for the continuous data from the previous sections. Assuming that the responses follow a Bernoulli distribution, the observed responses are denoted as

$$X_i^T \sim Bernoulli(p^T), i = 1 \ldots n^T \text{ and } X_i^C \sim Bernoulli(p^C), i = 1 \ldots n^C \quad (25)$$

with p^T and p^C representing the proportion of population responses expected in the balanced treatment and control groups, respectively.

The hypothesis in the form of risk difference between the two groups is

$$H_0 : p^T - p^C \leq 0 \text{ versus } H_1 : p^T - p^C > 0 \quad (26)$$

The main difference compared to the continuous variable is that response proportions are found in both the treatment effect in form of risk difference $p^T - p^C$ and the standard deviation σ. In this work, the standard deviation is chosen based on pooled variance $\bar{p}(1 - \bar{p})$ with $\bar{p} = \frac{n^T p^T + n^C p^C}{n^T + n^C} = \frac{p^T + p^C}{2}$.

The test statistic Z for the observed $\widehat{p^T}$ and $\widehat{p^C}$ can be simplified as

$$Z = \frac{\widehat{p^T} - \widehat{p^C}}{\widehat{\sigma}\sqrt{\frac{1}{n^T} + \frac{1}{n^C}}} = \frac{\widehat{p^T} - \widehat{p^C}}{\widehat{\sigma}} * \sqrt{\frac{n}{2}} \text{ , where } \widehat{\sigma} = \sqrt{\widehat{\bar{p}}(1 - \widehat{\bar{p}})} \text{ and } \widehat{\bar{p}} = \frac{\widehat{p^T} + \widehat{p^C}}{2} \quad (27)$$

Trials that intend to compare two groups are more often found in phase III. Phase III trials typically recruit a large number of patients. Therefore, the test statistics can be based on normal approximation. Similar to the continuous data with the normality assumption, H_0 is rejected if $Z \geq z_{1-\alpha}$ for a fixed design, while the probabilities associated with α and Pow can be expressed in the same fashion as in (4). The balanced sample size $n^T = n^C = n$ is derived as

$$n = \frac{2(z_{1-\alpha} + z_{1-\beta})^2 \bar{p}(1-\bar{p})}{(p^T - p^C)^2} \tag{28}$$

For a two-stage design, function (27) should be extended to include stage $j = 1, 2$ and observed proportions $\widehat{p_j^T}$ and $\widehat{p_j^C}$. Test statistics Z_j for each stage and Z_{1+2} for the final stage with all data are expressed as

$$Z_j = \frac{\widehat{p_j^T} - \widehat{p_j^C}}{\widehat{\sigma_j}} * \sqrt{\frac{n_j}{2}} \text{ , where } \widehat{\sigma_j} = \sqrt{\widehat{\bar{p}_j}(1-\widehat{\bar{p}_j})} \text{ and } \widehat{\bar{p}_j} = \frac{\widehat{p_j^T} + \widehat{p_j^C}}{2} \tag{29}$$

$$\begin{aligned} Z_{1+2} = \frac{\widehat{p_{1+2}^T} - \widehat{p_{1+2}^C}}{\widehat{\sigma_{1+2}}} * \sqrt{\frac{n}{2}} \text{ , where } \widehat{\sigma_{1+2}} = \sqrt{\widehat{\bar{p_{1+2}}}(1-\widehat{\bar{p_{1+2}}})} \text{ and} \\ \widehat{\bar{p_{1+2}}} = \frac{\widehat{p_{1+2}^T} + \widehat{p_{1+2}^C}}{2} \end{aligned} \tag{30}$$

It is shown that $Cov(Z_1, Z_{1+2}) = \sqrt{\frac{n_1}{n}}$ of equation (8) holds approximately if $p^T - p^C$ is small [10].

The non-binding futility stopping boundary α_f is defined as per (9). Thus, the trial may stop early for futility if interim analysis $Z_1 \leq z_{1-\alpha_f}$. The performance criteria and optimization process remain the same. However, the three conditions require adaptation for binary variables to accommodate that both the treatment effect and standard deviation contain p^T and p^C. The iterative search for the optimal approach should be characterized with some minor adjustment to functions (14) and (15) as follows

Condition 1

$$\pi_{wrong} \geq \Phi(z_{1-\alpha_f} - \frac{\widehat{p_1^T} - \widehat{p_1^C}}{\sqrt{\widehat{\bar{p}_1}(1-\widehat{\bar{p}_1})}}\sqrt{\frac{n_1}{2}}) \tag{31}$$

Condition 2

$$Pow_{loss} \geq MV_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(z_{1-\alpha_f}, z_{1-\alpha}) - \beta \tag{32}$$

The mean $\boldsymbol{\mu}$ and standard deviation Σ matrices of the multivariate normal cumulative distribution are

$$\boldsymbol{\mu} = \left(\frac{p^T-p^C}{\sqrt{\bar{p}(1-\bar{p})}}\sqrt{\frac{n_1}{2}}, \frac{p^T-p^C}{\sqrt{\bar{p}(1-\bar{p})}}\sqrt{\frac{n}{2}}\right) \quad (33)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sqrt{\frac{n_1}{n}} & 1 \\ 1 & \sqrt{\frac{n_1}{n}} \end{pmatrix} \quad (34)$$

The search finds the boundary $\alpha_{f,opt}$ in the exact same manner as the method for continuous data, given a smaller proportion of responders in the treatment group $p^{T_1} \in [p^C, p^T)$ and $\pi_{correct,p^{T_1}}$ for condition 3.

Regarding other scales of futility boundaries, Z-score is still $z_{1-\alpha_{f,opt}}$ and the CP function (18) requires only replacing θ and σ with the mean and standard deviation based on an approximation [6], as follows

$$CP = \Phi\{\frac{-z_{1-\alpha} + z_{1-\alpha_{f,opt}}\sqrt{\frac{n_1}{n}} + \frac{p^T-p^C}{\sqrt{\bar{p}(1-\bar{p})}}\sqrt{\frac{n}{2}}(1-\frac{n_1}{n})}{\sqrt{1-\frac{n_1}{n}}}\} \quad (35)$$

One group with stopping for futility

Different than the approach for the two-group trials, the method for one-group trials is typically applied in phase II with a limited total number of patients. The normal approximation utilized for sample size and test statistics is no longer appropriate for a small n. The exact method should be considered and the optimal approach in this research is formulated accordingly.

Without the control group, the hypothesis for the response variable X_i^T, formulated with pre-specified null and alternative response proportions p_0 and p_a, is

$$H_0 : p \leq p_0 \text{ versus } H_1 : p \geq p_a \quad (36)$$

where $p = \frac{\sum_{i=1}^{n} X_i^T}{n}$.

Let r denotes the number of responders $r = p * n$. r follows a binomial distribution with probability function $b(r, n, p)$, namely $P_{i=r} = (\frac{n}{i})p^i(1-p)^{n-i}$. Additionally, $B(r, n, p) = \sum_{i=1}^{r}(\frac{n}{i})p^i(1-p)^{n-i}$ denotes the cumulative binomial distribution. Test statistics are also based on the exact binomial test and depend on the exact ratio of r and n, which is evaluated for the decision to reject H_0. Therefore, unlike the methods designed for two-group comparison, the probabilities for testing are directly computed and compared to the decision boundaries α. H_0 shall be rejected whenever $1 - B(\widehat{r}, n, p_0) \leq \alpha$ (i.e., if $\widehat{r}$ or more responses are observed).

Given predefined α and β, the sample size for a fixed design with only one stage is determined by finding the exact n together with r under

$$\alpha \geq 1 - B(r, n, p_0) \text{ and } Pow = 1 - \beta \leq 1 - B(r, n, p_a) \tag{37}$$

For a two-stage design with stage $j = 1, 2$, the hypothesis testing is extended to the stage 1 observed $\widehat{r_1}$ and stage 2 observed $\widehat{r_{1+2}}$ based on all data. If $1 - B(\widehat{r_1}, n_1, p_0) \geq \alpha_f$ then the trial may stop early for futility at the interim analysis. For the final stage, the testing problem relies on the probability conditional on not-terminated stage 1 after at least $r_1 = B^{-1}(1 - \alpha_f, n_1, p_0)$ is observed. The probability of rejecting H_0 is quantified as

$$1 - [B(r_1, n_1, p_0) + \sum_{i=r_1+1}^{min(n_1, \widehat{r})} b(i, n_1, p_0) B(\widehat{r} - i, n_2, p_0)] \leq \alpha \tag{38}$$

In fact, with the parameters p_0, p_a, α, and β (or Pow) as design parameters, the optimal approach is similar to Simon's two-stage designs [30]. Nevertheless, Simon's designs adjust not only r_1 and n_1 but also r and n at the final stage. Moreover, they allow any $n_1 < n$, which often leads to an extreme proportion of the n being distributed to n_1 for stage 1. Therefore, the optimal approach here sets an additional constraint on $n_1 \leq \omega n$. $0 < \omega < 1$ represents the desired stage 1 sample size proportion and improves the balance between the stage 1 and stage 2 sample sizes when compared to Simon's designs. The most important benefit of the optimal approach over Simon's designs is the flexibility offered by non-binding futility stopping boundaries. If the trial is not stopped accordingly under Simon's designs, the type-I error is inflated. Following the general method of the optimal approach, the three conditions should be characterized first in relation to the first stage of early stopping for futility

Condition 1

$$\pi_{wrong} \geq B(r_1, n_1, p_a) \tag{39}$$

Condition 2

$$Pow_{loss} \geq 1 - \beta - [B(r_1, n_1, p_a) + \sum_{i=r_1+1}^{min(n_1,\widehat{r})} b(i, n_1, p_a)B(\widehat{r} - i, n_2, p_a))] \tag{40}$$

Similar to the two-group situations, many α_f with their corresponding r_1 fulfill both conditions. Moreover, Simon showed that there are many possible combinations due to r_1 and n_1 being allowed to vary. The probability of correctly stopping for futility is also a crucial criterion for optimization in the optimal approach and is labeled as ”PET0” in Simon’s designs for performance evaluation under H_0. Thus, $\pi_{correct,p_0}$ from (12) is characterized with the cumulative binomial distribution as

Condition 3

$$\pi_{correct,p_0} \geq B(r_1, n_1, p_0) \tag{41}$$

The optimal $\alpha_{f,opt}$ remains as in (13) by solving the iterative search. Since typical Simon’s designs define the decisions for hypothesis testing based on the corresponding r_1, n_1, r, and n only, the optimal approach for one-group design also provides $\alpha_{f,opt}$ and the set of optimal r_1, n_1, r, and n for the trial investigators. Other scales of futility boundaries are not investigated further since the main benefit of CP (i.e., for easier interpretation than α_f) is already fulfilled by the exact numbers of responses and the sample size, while the Z-score is not applicable.

4.3 Results

In this section, the optimal approach is first demonstrated by various operational characteristic combinations and an evaluation of futility stopping boundaries for both continuous and binary endpoints. Furthermore, combined with either efficacy or sample size re-estimation, applications on real clinical trials are presented to demonstrate the benefit of the optimal approach.

Considering a clinical trial with a continuous endpoint, an interim analysis allowing for early termination due to futility is planned to occur after 50% of the total patients enrolled. Given $\alpha = 0.025$ and $Pow = 1 - \beta = 0.9$, the futility boundaries α_f derived according to the optimal approach are displayed in Table 1.

Table 1: Implementation of a two-stage design with futility stopping only on continuous endpoints, with $n_1 = 0.5n$, $\alpha = 0.025$, and $\beta = 0.1$. The optimal approach is evaluated for the operational characteristics.

Optimal conditions			Optimal boundary	Other scales		Achieved operational characteristics		
π_{wrong}	Pow_{loss}	$\pi_{correct,\theta_1=0.5\theta}$	$\alpha_{f,opt}$	$z_{1-\alpha_{f,opt}}$	CP	global Pow	π_{wrong}	$\pi_{correct,\theta_1=0}$
0.01	0.01	0.12	0.51	-0.03	0.30	0.90	0.01	0.49
0.03	0.01	0.23	0.34	0.41	0.47	0.89	0.03	0.66
0.05	0.01	0.23	0.34	0.41	0.47	0.89	0.03	0.66
0.10	0.01	0.23	0.34	0.41	0.47	0.89	0.03	0.66
0.01	0.03	0.12	0.51	-0.03	0.30	0.90	0.01	0.49
0.03	0.03	0.23	0.34	0.41	0.47	0.89	0.03	0.66
0.05	0.03	0.31	0.26	0.64	0.57	0.88	0.05	0.74
0.10	0.03	0.36	0.22	0.77	0.62	0.87	0.07	0.78
0.01	0.05	0.12	0.51	-0.03	0.30	0.90	0.01	0.49
0.03	0.05	0.23	0.34	0.41	0.47	0.89	0.03	0.66
0.05	0.05	0.31	0.26	0.64	0.57	0.88	0.05	0.74
0.10	0.05	0.44	0.16	0.99	0.69	0.85	0.10	0.84

The first two columns describe the maximum values of conditions 1 and 2, which are allowed and prospectively planned for the trial protocol. The third column displays condition 3, which is searched for the maximum value according to (13). In Table 1, half of the original treatment effect is deemed not clinically beneficial. Thus, the probability of correctly stopping in condition 3 is calculated for such underlying treatment effect $\theta_1 = 0.5\theta$. Although it may seem reasonable to aim for a higher $\pi_{correct} > 0.60$, the combination of the highest $Pow_{loss} = 0.05$ and $\pi_{wrong} = 0.10$ only achieves the correct stopping of 44% of the trials if repeated many times.

The fourth column is the most important one, which gives the optimal futility stopping boundary $\alpha_{f,opt}$ for the stopping decision at the interim analysis. The last three columns show the operational characteristics achieved as a performance evaluation. The actual probability of wrongly stopping under θ can only reach a value up to the maximum allowed π_{wrong} in the first column. The actual global power is reduced due to the inclusion of a futility stop and limited by condition 2, maximum allowed Pow_{loss}. If there is truly no treatment benefit, the probability of early stopping is listed in the last column. Other quantities of performance could be added to the group of achieved operational characteristics. For example, probabilities of correctly stopping assuming another θ_1, not among those listed in the Table 1 (0, θ or 0.5θ), may be further investigated. These are all relevant assessments for the investigators during the planning phase due to the uncertainty associated with the true treatment effect.

The maximum values of conditions 1 and 2 are bounded by small values of $\pi_{wrong} \leq 0.01$ and $Pow_{loss} \leq 0.01$ in the first row. The achieved optimal $\alpha_f = 0.51$ guarantees a global power rounded up to 0.90, with nearly no loss compared to the planned power. The actual π_{wrong} fully exhausts the maximum allowed value. However, it has only a small probability of 0.12, which allows a correct early futility stop for condition 3. Despite $\alpha_f = 0.51$ being optimal and safeguarding the trial, it is questionable whether such interim analysis is necessary when compared to a traditional fixed design. By allowing a higher maximum π_{wrong}, the actual Pow_{loss} quickly reaches the maximum allowed 0.01 set by condition 2 in both rows 2 and 3. The benefit of 11% increase in chance to correctly stop given half of θ is gained, only at a minimal cost of actual $Pow_{loss} = 0.01$ and $\pi_{wrong} = 0.03$. A more extreme case can be observed in the last row. Having an optimal $\alpha_f = 0.16$ fully utilizes the allowed risks, with the actual π_{wrong} and Pow_{loss} being the same as the maximum allowed values for conditions 1 and 2. Condition 3 achieved a high probability of 0.44 to correctly end the trial early because smaller boundaries α_f allow easier crossing over of the boundary if the treatment effect is not large. However, the investigators should consider whether the trade-off between higher π_{wrong} and Pow_{loss} is truly desired. For example, rows 3 and 7 have the same values of condition 1 yet different values of condition 2. Moreover, they give distinct futility boundaries. π_{wrong} plays a more restrictive role in the choice of futility boundary when compared to Pow_{loss} at the same magnitude. Since the actual $\pi_{wrong} = 0.05$ in rows 7 and 11 reach the allowed value of condition 1, the choice of maximum Pow_{loss} does not make any difference on the choice of the optimal boundary. Lastly, the corresponding Z-score and CP for $\alpha_{f,opt} = 0.26$ are 0.64 and 0.57 in the second last row. Some can argue that the Z-score is much greater than 0 and the CP indicates an already good power for success at the final stage. The optimal approach for the performance evaluation indicates that no matter what scales the futility boundaries have, the conditions should be pre-selected and boundaries should not be chosen arbitrarily.

Table 2: Implementation of a two-stage design with futility stopping only on continuous endpoints, with $n_1 = 0.5n$, $\alpha = 0.025$, and $\beta = 0.2$. The optimal approach is evaluated for the operational characteristics.

Optimal conditions			Optimal boundary	Other scales		Achieved operational characteristics		
π_{wrong}	Pow_{loss}	$\pi_{correct,\theta_1=0.5\theta}$	$\alpha_{f,opt}$	$z_{1-\alpha_{f,opt}}$	CP	global Pow	π_{wrong}	$\pi_{correct,\theta_1=0}$
0.01	0.01	0.09	0.63	-0.33	0.13	0.80	0.01	0.37
0.03	0.01	0.19	0.46	0.10	0.25	0.80	0.03	0.54
0.05	0.01	0.25	0.37	0.33	0.32	0.79	0.05	0.63
0.10	0.01	0.25	0.37	0.33	0.32	0.79	0.05	0.63
0.01	0.03	0.09	0.63	-0.33	0.13	0.80	0.01	0.37
0.03	0.03	0.19	0.46	0.10	0.25	0.80	0.03	0.54
0.05	0.03	0.25	0.37	0.33	0.32	0.79	0.05	0.63
0.10	0.03	0.38	0.24	0.71	0.46	0.77	0.10	0.76
0.01	0.05	0.09	0.63	-0.33	0.13	0.80	0.01	0.37
0.03	0.05	0.19	0.46	0.10	0.25	0.80	0.03	0.54
0.05	0.05	0.26	0.37	0.33	0.32	0.79	0.05	0.63
0.10	0.05	0.39	0.24	0.71	0.47	0.77	0.10	0.76

In Table 2, β is set to 0.2. Notably, same trend is observed as in Table 1. Furthermore, the choice of futility boundary is even more constrained by $\pi_{wrong} \leq 0.05$ and less sensitive to the condition 1 Pow_{loss} when compared to a trial with a smaller β.

The timing of a futility stopping boundary could also play a role in the choice of $\alpha_{f,opt}$, as shown in Figure 1. The optimal approach takes the timing of the first stage based on the fraction of patients $\frac{n_1}{n}$ as a pre-specified design parameter because they are often decided for the convenience of trial management and conduct. Nevertheless, the timing would still have an impact on the boundary itself. If more patients with available data can already be included in the first stage, the variability of the interim analysis is decreased. Additionally, it means that fewer patients per n_2 need to be recruited for the second stage. Moreover, the results of the final stage tend to be more consistent with the treatment effect observed in the first stage. With fixed conditions π_{wrong} and Pow_{loss} and achieving the most favorable $\pi_{correct,\theta_1=0}$, the $\alpha_{f,opt}$ becomes stricter with more n_1 acquired.

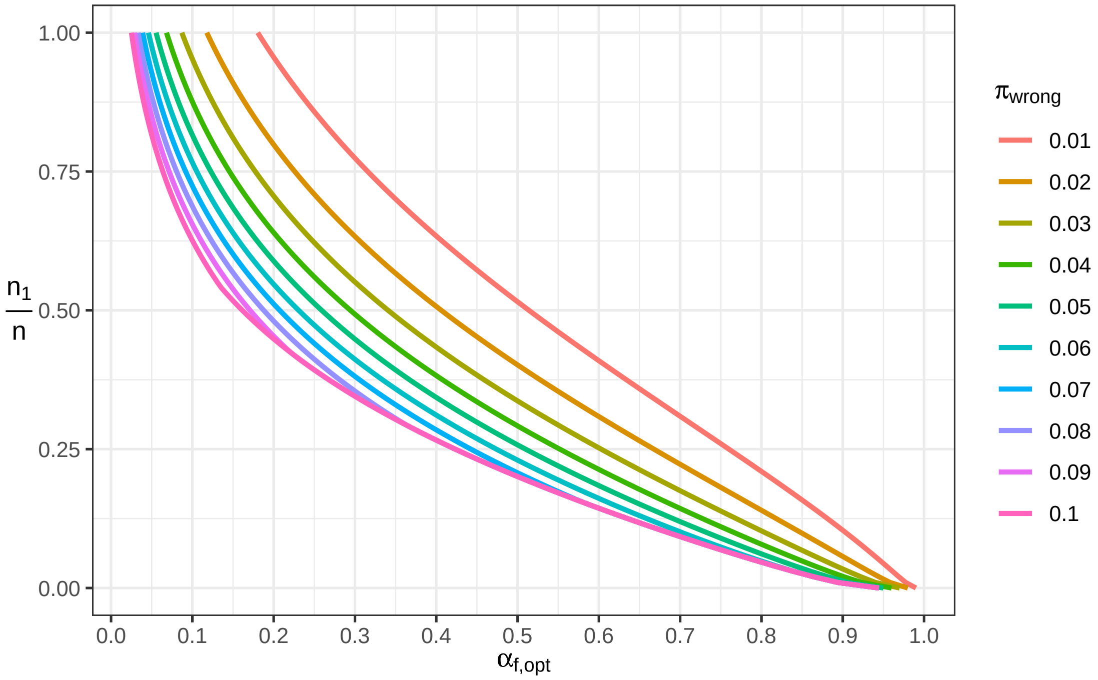


Figure 1: Impact of the timing of the first stage on $\alpha_{f,opt}$, given maximum $Pow_{loss} = 0.05$.

Similar to Table 1 and 2, Table 3 presents the operational characteristics of trials for a binary endpoint with an optimal futility stopping boundary. The trade-off between the three conditions resembles the result from Table 1, which are also

Table 3: Implementation of a two-stage design with futility stopping only on binary endpoints, with $n_1 = 0.5n$, $\alpha = 0.025$, and $\beta = 0.1$. Assuming $p_T = 0.6$ and $p_C = 0.4$.

Optimal conditions			Optimal boundary	Other scales		Achieved operational characteristics		
π_{wrong}	Pow_{loss}	$\pi_{correct,p_{T,1}=0.55}$	$\alpha_{f,opt}$	$z_{1-\alpha_{f,opt}}$	CP	global Pow	π_{wrong}	$\pi_{correct,p_{T,1}=0.4}$
0.01	0.01	0.04	0.51	-0.03	0.30	0.90	0.01	0.49
0.03	0.01	0.09	0.34	0.41	0.47	0.89	0.03	0.66
0.05	0.01	0.09	0.34	0.41	0.47	0.89	0.03	0.66
0.10	0.01	0.09	0.34	0.41	0.47	0.89	0.03	0.66
0.01	0.03	0.04	0.51	-0.03	0.30	0.90	0.01	0.49
0.03	0.03	0.10	0.34	0.41	0.47	0.89	0.03	0.66
0.05	0.03	0.14	0.26	0.64	0.57	0.88	0.05	0.74
0.10	0.03	0.17	0.22	0.77	0.62	0.87	0.07	0.78
0.01	0.05	0.04	0.51	-0.03	0.30	0.90	0.01	0.49
0.03	0.05	0.10	0.34	0.41	0.47	0.89	0.03	0.66
0.05	0.05	0.14	0.26	0.64	0.57	0.88	0.05	0.74
0.10	0.05	0.23	0.16	0.99	0.69	0.85	0.10	0.84

based on a scenario with $\alpha = 0.025$ and $\beta = 0.1$. The minimum probability of correctly stopping is relatively low because it is based on a small 5% decrease in p_T.

Lastly, to demonstrate the benefit of the optimal approach for binary endpoints with only one group, Tables 4 and 5 are created. Simon's designs have two variations: minimax and optimal designs. Both variations do not optimize for π_{wrong} and Pow_{loss} as performed in the optimal approach. Instead, they optimize for the maximum and expected sample size. Additionally, as explained in the Methods section, Simon's designs require binding futility boundaries, which also adjust the total sample size n and number of responses r to fully spend α and β. On the other hand, the optimal approach cannot optimize the choices of n and r and is only able to derive n_1 and r_1 from the boundary $\alpha_{f,opt}$. This difference is shown in Tables 4 and 5, where n and r from Simon's minimax design are smaller than the fixed n and r in the optimal approach designs, ignoring the rounding of values. To highlight the advantage of the optimal approach, non-optimal α_f from Simon's designs are also displayed in Table 4. Furthermore, due to the exact nature of the binomial test, it is not always possible to control type-I and type-II errors exactly at the maximum levels of α and β. Therefore, Pow_{loss} can even become negative. The advantage of the optimal approach is identified for condition 2 with Pow_{loss}, which allows type-II error to be slightly higher than the predefined β instead of an increase of n as compensation to losses of power. Consequently, with minimal loss, some of the unused β could also be spent by setting $Pow_{loss} \leq 0.01$. Regarding condition 1, let $\pi_{wrong} \leq 0.10$ since it is in the range seen in Simon's designs. Condition 3 $\pi_{correct,p_0}$ does not have a pre-specified desired level and is thus used for finding the maximum value. An additional operational characteristic EN_0, the expected sample size assuming p_0, is a basic part of Simon's designs and included in the tables. Depending on p_0 and p_a, the designs using the optimal approach generally do not greatly increase EN_0, but have either better control over π_{wrong} and Pow_{loss} or even higher $\pi_{correct,p_0}$.

Table 4: Comparison of optimal approach designs and Simon's optimal and minimax designs, given $\alpha = 0.1$, $\beta = 0.1$, $\pi_{wrong} = 0.10$ and $Pow_{loss} = 0.01$.

Design	p_0	p_a	r_1	n_1	r	n	α_f	π_{wrong}	Pow_{loss}	$\pi_{correct,p_0}$	α	β	EN_0
Simon's optimal	0.5	0.65	18	35	47	84	0.37	0.0682	-0.0004	0.6321	0.0952	0.0996	53.0
Simon's minimax	0.5	0.65	19	40	41	72	0.56	0.0173	-0.0001	0.4373	0.0956	0.0999	58.0
Optimal approach $\omega = \frac{1}{2}$	0.5	0.65	13	29	41	72	0.64*	0.0206	0.0041	0.3555	0.0944	0.1041	56.7
Optimal approach $\omega = \frac{2}{3}$	0.5	0.65	22	44	41	72	0.44*	0.0289	0.0029	0.5598	0.0942	0.1029	56.3
Simon's optimal	0.7	0.85	14	20	45	59	0.42	0.0673	-0.0010	0.5836	0.0954	0.0990	36.2
Simon's minimax	0.7	0.85	15	22	40	52	0.49	0.0368	-0.0029	0.5058	0.0980	0.0971	36.8
Optimal approach $\omega = \frac{1}{2}$	0.7	0.85	8	13	41	53	0.65*	0.0342	0.0098	0.3457	0.0853	0.1098	39.2
Optimal approach $\omega = \frac{2}{3}$	0.7	0.85	25	34	41	53	0.26*	0.0587	0.0093	0.7323	0.0825	0.1093	39.1

*$\alpha_{f,opt}$

In row 3 of Table 5, where $\omega = \frac{1}{2}$, it is noticeable that no feasible design based on the optimal approach is available. Since the optimal approach has similar r and n to Simon's minimax, it is obvious that in this case, when the proportion $\frac{n_1}{n}$ reaches as high as $\frac{66}{68} = 97\%$, it is difficult to constrain ω below 50%. Even with 67%, the conditions π_{wrong}, Pow_{loss}, and $\pi_{correct,p_0}$ are comparable to Simon's designs, while $EN_0 = 51.6$ is far below Simon's minimax design $EN_0 = 66.1$. Similar settings are displayed in the second block. The optimal approach allows the boundary $\alpha_{f,opt}$ and the corresponding r_1 and n_1 vary, especially with a small $\alpha_{f,opt} = 0.21$ design, while maintaining the EN_0 around 34.4 as per Simon's minimax design.

Table 5: Comparison of optimal approach designs and Simon's optimal and minimax designs, given $\alpha = 0.05$, $\beta = 0.2$, $\pi_{wrong} = 0.10$ and $Pow_{loss} = 0.01$.

Design	p_0	p_a	r_1	n_1	r	n	α_f	π_{wrong}	Pow_{loss}	$\pi_{correct,p_0}$	α	β	EN_0
Simon's optimal	0.5	0.65	15	28	48	83	0.29	0.1428	-0.0015	0.7142	0.0470	0.1985	43.7
Simon's minimax	0.5	0.65	39	66	40	68	0.05	0.1893	-0.0013	0.9456	0.0488	0.1987	66.1
Optimal approach $\omega = \frac{1}{2}$	0.5	0.65	not feasible										
Optimal approach $\omega = \frac{2}{3}$	0.5	0.65	24	45	41	69	0.28*	0.0708	0.0073	0.7243	0.0439	0.2073	51.6
Simon's optimal	0.7	0.85	14	19	46	59	0.28	0.1444	-0.0067	0.7178	0.0494	0.1933	30.3
Simon's minimax	0.7	0.85	16	23	39	49	0.44	0.0463	-0.0008	0.5601	0.0466	0.1992	34.4
Optimal approach $\omega = \frac{1}{2}$	0.7	0.85	17	24	39	49	0.49*	0.0572	0.0020	0.6114	0.0461	0.2020	33.7
Optimal approach $\omega = \frac{2}{3}$	0.7	0.85	24	32	39	49	0.21*	0.0958	0.0065	0.7882	0.0451	0.2065	35.6

$^*\alpha_{f,opt}$

Real trial application 1

The first application of the optimal approach on a real clinical trial allow both the futility and efficacy stopping. The ChroPac Trial [5] was a randomized controlled trial with an interim analysis. The objective was to investigate the efficacy of an intervention surgical procedure compared to a standard surgical procedure treating patients with chronic pancreatitis. The efficacy endpoint, tested for superiority, is the quality of life score from the EORTC QLQ-C30 questionnaire. The total scores of EORTC QLQ-C30 are between 0 and 100. Higher total scores indicate a high level of functioning and quality of life. Although it is a scoring system, the actual calculation is generally performed through a linear transformation. Thus, the data should be treated as a continuous endpoint. During the planning of the ChroPac Trial, the standardized treatment effect $\frac{\theta}{\sigma}$ was assumed to be 0.5 and the hypothesis of treatment efficacy was tested at one-sided with an $\alpha = 0.025$. In the original trial, a fixed design was also planned for $\beta = 0.1$, which set a total sample size of 86 patients for the single final analysis. To illustrate the optimal choice of futility stopping boundaries together with the option to stop early for efficacy, a two-stage design is chosen with 50% of patients enrolled for the interim analysis. Global powers of both 0.9 and 0.8 are commonly applied in clinical research and their corresponding $\beta = 0.1$ and 0.2 are presented in this application. Since the futility stopping boundaries of the optimal approach are non-binding, they are derived after the efficacy stopping boundaries are chosen first. When applying Pocock's alpha-spending function to the overall $\alpha = 0.025$, the local significance levels are $\alpha_1 = \alpha_{1+2} = 0.0147$ at the interim and final analysis for efficacy. Figure 2 shows the set $A_{Pow_{loss},\pi_{wrong}}$ in the two situations. Reasonable conditions of $Pow_{loss} <= 0.05$, $\pi_{wrong} <= 0.05$, and $\pi_{correct,\theta_1=0.5\theta} >= 0.30$ are predefined. The boundaries $\alpha_{f,opt}$ are found at 0.33 and 0.22. Both values of $\alpha_{f,opt}$ show good performance if the true effect size is only half of the initially assumed effect. It is shown that β affects the overall potential choices of α_f, with higher α_f at lower β, thereby making an early futility stop and type-II error less likely.

Additionally, both optimal $\alpha_{f,opt}$ are much lower than an arbitrary futility stopping boundary of 0.50 and, equivalently, a CP of 0.31. Although $\alpha_f = 0.50$ still protects both $Pow_{loss} = 0.0013$ and $\pi_{wrong} = 0.018$, allowing early termination due to futility at the interim stage would be rather unnecessary since the probability of a correct to early stop is only 0.15.

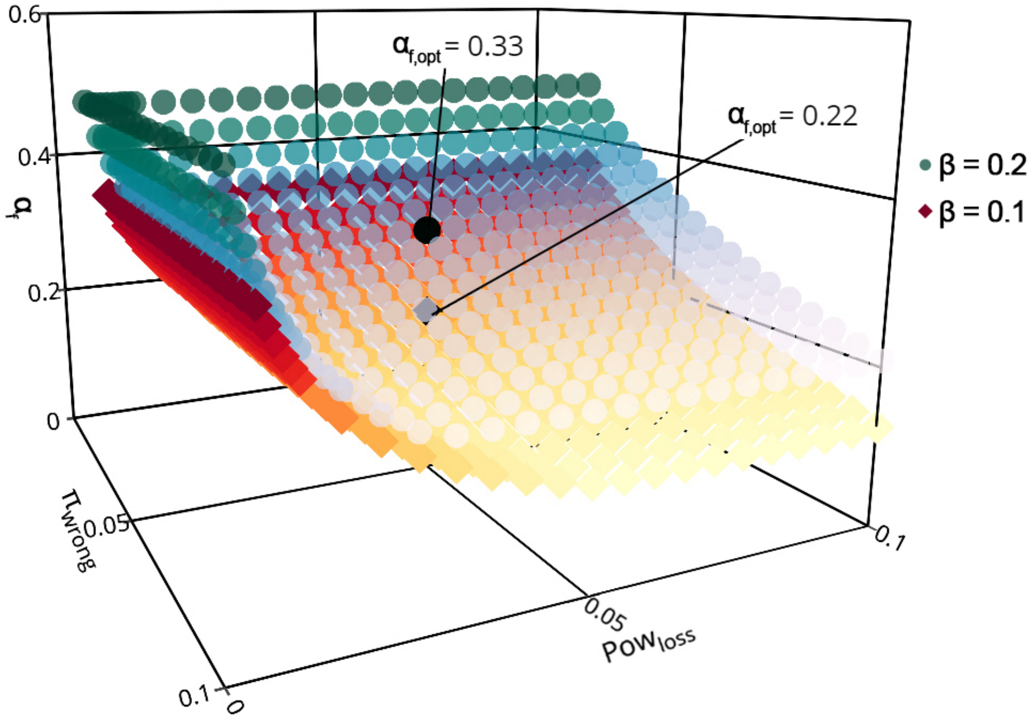


Figure 2: ChroPac Trial application with $\beta = 0.1$ and 0.2. The figure was originally created for the publication [17].

Real trial application 2

In the second trial application, a showcase of the optimal approach is created for an adaptive design with sample size re-estimation at the interim stage. The sample size is recalculated based on the scheme of (22). The PDY6797 trial was a randomized placebo-controlled trial that aimed to test the efficacy of a new treatment on patients with type 2 diabetes [29]. The change in plasma glucose area under the curve from baseline was chosen as the primary efficacy endpoint. This endpoint is often used in linear models without log-transformation due to its approximately normal distribution. Furthermore, values of change from baseline might be negative. The initial sample size per group $n = 11$ is derived based on (22), assuming a $\beta = 0.1$ and one-sided $\alpha = 0.025$ with an expected treatment effect of $\theta = 300$ and standard deviation of $\sigma = 250$. Table 6 shows that adding an option for futility reduces the expected sample size n_{avg} and the smallest n_{avg} is achieved when the optimal approach is applied. Between the two designs with an arbitrarily chosen $\alpha_f = 0.50$ and the optimal $\alpha_{f,opt}$ futility boundaries, the optimal approach controls the three conditions and achieves a desired higher probability of correctly stopping with the pre-specified cost of power loss and probability of wrongly stopping.

Table 6: Comparison of different designs at $\alpha = 0.025$ and $\beta = 0.1$. For the optimal approach, the operational characteristics are predefined as $Pow_{loss} \leq 0.05$, $\pi_{wrong} \leq 0.10$ and $\pi_{correct,\theta_1=0.5\theta} \geq 0.30$.

Design	n_{max}	n_{avg}	global Pow	π_{wrong}	$\pi_{correct,\theta_1=0.5\theta}$	$\pi_{correct,\theta_1=0}$
Adaptive only	33	15.0	0.91	-	-	-
Adaptive + futility with $\alpha_f = 0.50$	33	11.6	0.90	0.02	0.15	0.50
Adaptive + futility with $\alpha_{f,opt} = 0.27$	33	10.0	0.86	0.08	0.35	0.73

4.4 Discussion

The choice of futility boundaries in flexible designs affects various operational characteristics. Conveniently choosing a futility boundary is not equivalent to having a numerically optimal boundary. Even with a relatively small treatment effect that favors the null hypothesis, an early stopping for futility is not guaranteed. Any futility boundary has an impact on the overall power loss and probability of correctly and wrongly stopping under various treatment effect assumptions. Thus, the optimization should not be overlooked.

Previous research has proposed optimization strategies for different designs, including futility stopping boundaries in either group sequential or adaptive designs [8, 7, 22, 30]. A focus on futility stopping optimization alone is carefully investigated in my research. The optimal approach proposed by Schüler on time-to-event endpoints in a two-stage group sequential design [28] is further developed to continuous [17] and binary endpoints in both group sequential and adaptive designs. In the optimal approach, three operational characteristics are jointly evaluated and optimized to derive the appropriate futility boundary. With the optimal approach, trial investigators can fully specify the trade-offs between the desired operational characteristics to increase performance efficiency. For one trial, a 5% probability of wrongly stopping for futility might already be unacceptable due to the importance of the newly developed treatment since the investigators would favor a continuation to the end of the trial whenever possible. Whereas for another trial, a high power loss may not favorable because the trial sponsors have only one chance to run the trial for their innovative treatment and hope for a good chance of success. Also, if the new treatment started with a less promising outlook, a predefined high probability of correctly stopping should be seen as more important than the other characteristics. Thus, the optimal choice of a futility stopping boundary reflects the cautiousness or aggression of the research objectives of trial investigators.

One of the operational characteristics presented in this work is the power loss, which cannot be avoided with non-binding futility boundaries. Some methods of futility stopping design suggest increasing the total sample size to compensate for such losses of power. In comparison, the optimal approach only plans to simply accept a power loss since the power loss comes from the probability of the trial actually being stopped early. An increase in the initial sample size planning automatically treats the futility stopping boundaries as binding, as shown in Simon's designs. This method is quite restrictive if investigators would utilize early futility stopping boundaries merely as a suggestion. Nevertheless, the other performance scores can be evaluated in future research. Real-world situations can be investigated in future work to account for the extra data collected due to overrunning, to apply to other types of adaptation other than the sample size, and to optimize α_f, n_1 and n simultaneously. Additionally, the optimization assumes a balanced allocation between treatment groups, which is not always the case in real trials.

This research aims to create a quantified approach for optimizing futility stopping boundaries based on the evaluation of performance criteria. The criteria are straightforward and can be communicated among trial investigators. With engaging communication and a greater understanding of the operational characteristics, optimal futility stopping boundaries should be applied more often to further increase trial efficiency and drive innovative clinical research.

5 Bibliography

[1] P. Bauer and K. Köhne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4):1029–1041, 1994.

[2] F. Bretz, F. Koenig, W. Brannath, E. Glimm, and M. Posch. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8):1181–1217, 2009.

[3] M. Chang, I. Hwang, and W. Shin. Group sequential designs using both type i and type ii error probability spending functions. *Communications in Statistics - Theory and Methods*, 27(6):1323–1339, 1998.

[4] D. DeMets and J. Ware. Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika*, 67:651–660, 1980.

[5] M. Diener, T. Bruckner, P. Contin, C. Halloran, M. Glanemann, H. Schlitt, J. Mössner, M. Kieser, J. Werner, M. Büchler, and C. Seiler. Chropac-trial: duodenum-preserving pancreatic head resection versus pancreatoduodenectomy for chronic pancreatitis. trial protocol of a randomised controlled multicentre trial. *Trials*, 11(47), 2010.

[6] J. Fleiss, A. Tytun, and H. Ury. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, 36(2):343–346, 1980.

[7] P. Gallo, L. Mao, and V. Shih. Alternative views on setting clinical trial futility criteria. *J Biopharm Stat.*, 24(5):976–993, 2014.

[8] P. He, T. Lai, and O. Liao. Futility stopping in clinical trials. *Statistics and Its Interface*, 5:415–423, 2012.

[9] D. Heitjan. Bayesian interim analysis of phase ii cancer clinical trials. *Statistics in Medicine*, 16(16):1791–1802, 1997.

[10] C. Jennison and B. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton, 2000.

[11] C. Jennison and B. Turnbull. Adaptive sample size modification in clinical trials: start small then ask for more? *Statistics in Medicine*, 34(29):3793–3810, 2015.

[12] J. Lachin. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*, 24(18):2747–2764, 2005.

[13] K. Lan and D. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663, 1983.

[14] K. Lan, R. Simon, and M. Halperin. Stochastically curtailed tests in long–term clinical trials. *Communications in Statistics. Part C: Sequential Analysis*, 1(3): 207–219, 1982.

[15] W. Lehmacher and G. Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290, 1999.

[16] R. Lewis and D. Berry. Group sequential clinical trials: A classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association*, 89(428):1528–1534, 1994.

[17] X. Li, C. Herrmann, and G. Rauch. Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint. *BMC Medical Research Methodology*, 20(274), 2020.

[18] M. Lin, S. Lee, B. Zhen, J. Scott, A. Horne, G. Solomon, and E. Russek-Cohen. Cber's experience with adaptive design clinical trials. *Therapeutic innovation & regulatory science*, 50(2):195–203, 2016.

[19] G. Liu, G. Zhu, and L. Cui. Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Statist Med.*, 27 (4):584–596, 2008.

[20] P. O'Brien and T. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979.

[21] S. Pampallona and A. Tsiatis. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42:19–35, 1994.

[22] M. Pilz, K. Kunzmann, C. Herrmann, G. Rauch, and M. Kieser. A variational approach to optimal two-stage designs. *Stat Med.*, 38(21):4159–4171, 2019.

[23] S. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[24] S. Pocock. Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics*, 38(1):153–162, 1982.

[25] K. Rudser and S. Emerson. Implementing type i & type ii error spending for two-sided group sequential designs. *Contemporary clinical trials*, 29:351–358, 2008.

[26] B. Saville, J. Connor, G. Ayers, and J. Alvarez. The utility of bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, 11(4):485–493, 2014.

[27] D. A. Schoenfeld and M. O. Meade. Pro/con clinical debate: It is acceptable to stop large multicentre randomized controlled trials at interim analysis for futility. *Critical Care*, 9(34), 2004.

[28] S. Schüler, M. Kieser, and G. Rauch. Choice of futility boundaries for group sequential designs with two endpoints. *BMC Medical Research Methodology*, 17(119), 2017.

[29] Y. Seino, A. Takami, G. Boka, E. Niemoeller, and D. Raccah. Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Diabetes, Obesity and Metabolism*, 16:739–747, 2014.

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

[31] A. Tamhane, C. Mehta, and L. Liu. Testing a primary and a secondary endpoint in a group sequential design. *Biometrics*, 66:1174–84, 2010.

[32] The Economist Intelligence Unit. The innovation imperative: The future of drug development part i: Research methods and findings. Technical report, The Economist Group, 2018.

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

6 Statutory declaration

“I, Xieran Li, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic ”Evaluation of the performance criteria of optimal futility stopping boundaries in flexible designs”, ”Evaluation der Leistungskriterien der optimalen Grenzen für das Stoppen aufgrund der Aussichtslosigkeit bei flexiblen Designs”, independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

Date Signature

7 Declaration of own contribution

Xieran Li contributed the following to the below listed publications:

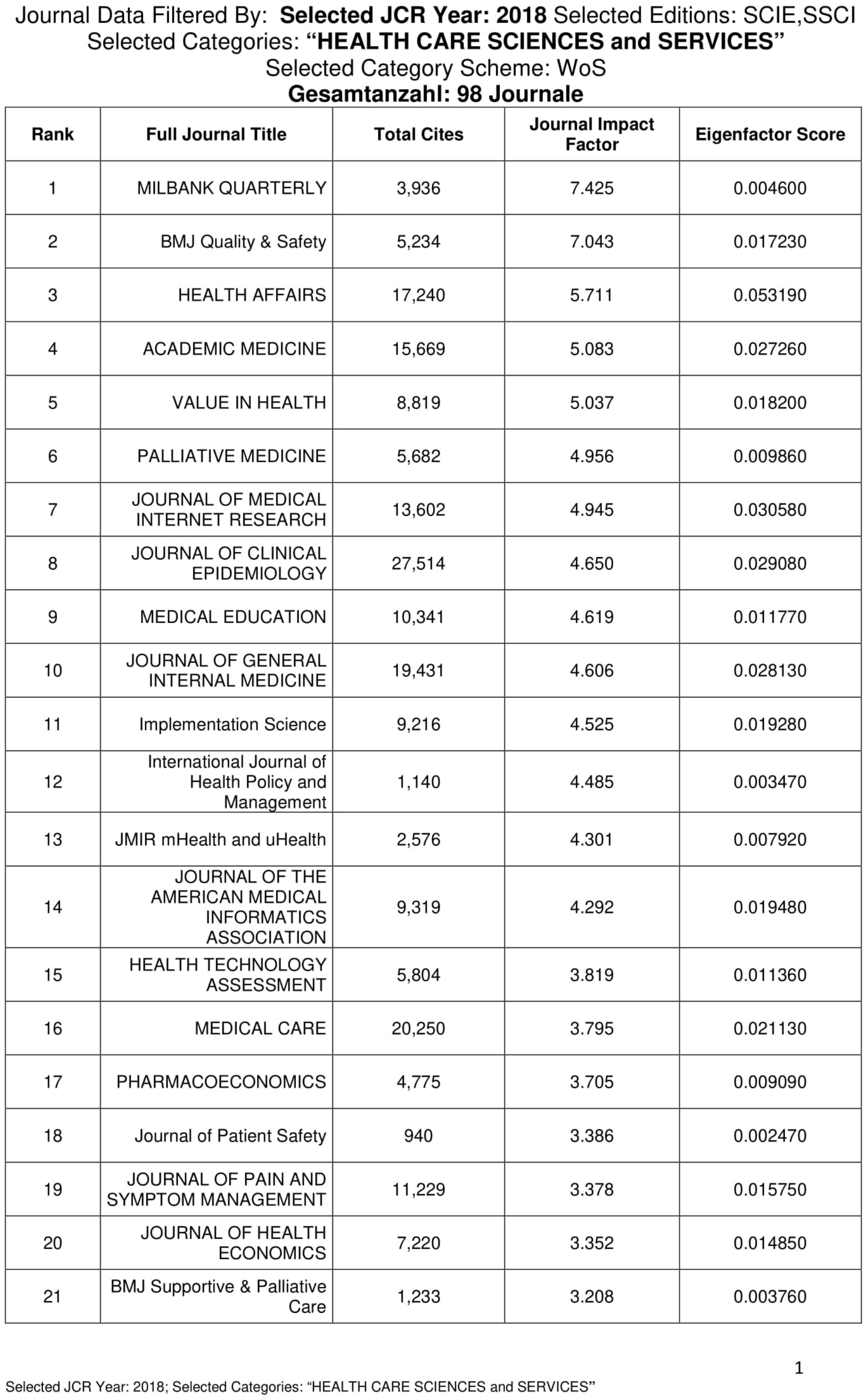
Publication 1: X. Li, C. Hermann, and G. Rauch., Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint, BMC Medical Research Methodology, 2020.

Contribution: I extended the previously published research from Schüler on the topic of optimizing futility stopping boundaries in group sequential designs. I developed the approach further to the continuous endpoint and conceptualized the algorithm theoretically, implemented the program reflecting the algorithm, applied the method and presented and interpreted the results. For the original publication manuscript, I wrote and revised the initial drafts and created all the results including tables and figures, which was then commented and proof-read by the co-authors, and mainly by Prof. Rauch. Prof. Rauch additionally supported the publishing process with highly valuable guidance.

Signature, date and stamp of first supervising university professor / lecturer

Signature of doctoral candidate

8 Extract of journal summary list



Journal Data Filtered By: **Selected JCR Year: 2018** Selected Editions: SCIE,SSCI
Selected Categories: **“HEALTH CARE SCIENCES and SERVICES”**
Selected Category Scheme: WoS
Gesamtanzahl: 98 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	MILBANK QUARTERLY	3,936	7.425	0.004600
2	BMJ Quality & Safety	5,234	7.043	0.017230
3	HEALTH AFFAIRS	17,240	5.711	0.053190
4	ACADEMIC MEDICINE	15,669	5.083	0.027260
5	VALUE IN HEALTH	8,819	5.037	0.018200
6	PALLIATIVE MEDICINE	5,682	4.956	0.009860
7	JOURNAL OF MEDICAL INTERNET RESEARCH	13,602	4.945	0.030580
8	JOURNAL OF CLINICAL EPIDEMIOLOGY	27,514	4.650	0.029080
9	MEDICAL EDUCATION	10,341	4.619	0.011770
10	JOURNAL OF GENERAL INTERNAL MEDICINE	19,431	4.606	0.028130
11	Implementation Science	9,216	4.525	0.019280
12	International Journal of Health Policy and Management	1,140	4.485	0.003470
13	JMIR mHealth and uHealth	2,576	4.301	0.007920
14	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	9,319	4.292	0.019480
15	HEALTH TECHNOLOGY ASSESSMENT	5,804	3.819	0.011360
16	MEDICAL CARE	20,250	3.795	0.021130
17	PHARMACOECONOMICS	4,775	3.705	0.009090
18	Journal of Patient Safety	940	3.386	0.002470
19	JOURNAL OF PAIN AND SYMPTOM MANAGEMENT	11,229	3.378	0.015750
20	JOURNAL OF HEALTH ECONOMICS	7,220	3.352	0.014850
21	BMJ Supportive & Palliative Care	1,233	3.208	0.003760

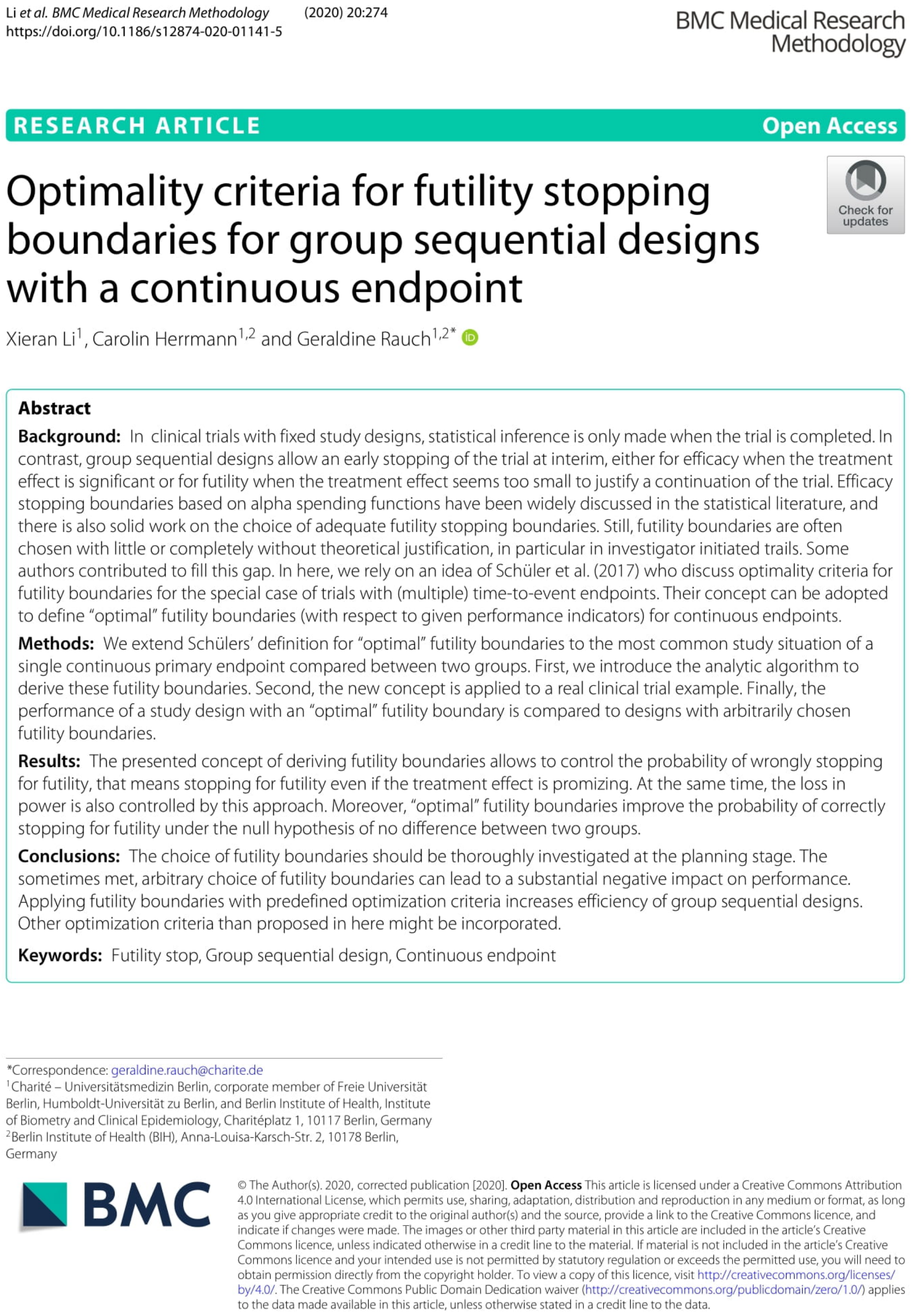
1

Selected JCR Year: 2018; Selected Categories: “HEALTH CARE SCIENCES and SERVICES”

22	Journal of Managed Care & Specialty Pharmacy	1,221	3.024	0.004750
23	BMC Palliative Care	1,522	2.922	0.003880
24	HEALTH EXPECTATIONS	3,199	2.847	0.007740
25	MEDICAL DECISION MAKING	5,281	2.793	0.009000
26	ADVANCES IN HEALTH SCIENCES EDUCATION	2,697	2.761	0.005400
27	SUPPORTIVE CARE IN CANCER	11,975	2.754	0.024130
28	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	4,765	2.731	0.006720
29	HEALTH POLICY AND PLANNING	5,401	2.717	0.010110
30	HEALTH SERVICES RESEARCH	8,061	2.706	0.013670
30	MEDICAL TEACHER	7,977	2.706	0.010530
32	Patient-Patient Centered Outcomes Research	1,008	2.673	0.003090
33	Applied Health Economics and Health Policy	1,126	2.664	0.003350
34	MEDICAL CARE RESEARCH AND REVIEW	2,431	2.577	0.004060
35	BMC Medical Research Methodology	9,832	2.509	0.021050
36	International Journal of Integrated Care	1,137	2.489	0.002010
37	QUALITY OF LIFE RESEARCH	13,192	2.488	0.019050
38	JOURNAL OF PALLIATIVE MEDICINE	5,938	2.477	0.010540
39	JOURNAL OF RURAL HEALTH	1,729	2.471	0.002630
40	EUROPEAN JOURNAL OF CANCER CARE	3,149	2.421	0.005380
41	JOURNAL OF MEDICAL SYSTEMS	4,680	2.415	0.006220
42	STATISTICAL METHODS IN MEDICAL RESEARCH	4,156	2.388	0.012230
43	Health and Quality of Life Outcomes	8,070	2.318	0.012120
44	Health Informatics Journal	691	2.297	0.001450
45	Risk Management and Healthcare Policy	416	2.283	0.001270

Selected JCR Year: 2018; Selected Categories: “HEALTH CARE SCIENCES and SERVICES”

9 Selected publication



Li *et al. BMC Medical Research Methodology* (2020) 20:274
https://doi.org/10.1186/s12874-020-01141-5

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint

Xieran Li[1], Carolin Herrmann[1,2] and Geraldine Rauch[1,2*]

Abstract

Background: In clinical trials with fixed study designs, statistical inference is only made when the trial is completed. In contrast, group sequential designs allow an early stopping of the trial at interim, either for efficacy when the treatment effect is significant or for futility when the treatment effect seems too small to justify a continuation of the trial. Efficacy stopping boundaries based on alpha spending functions have been widely discussed in the statistical literature, and there is also solid work on the choice of adequate futility stopping boundaries. Still, futility boundaries are often chosen with little or completely without theoretical justification, in particular in investigator initiated trails. Some authors contributed to fill this gap. In here, we rely on an idea of Schüler et al. (2017) who discuss optimality criteria for futility boundaries for the special case of trials with (multiple) time-to-event endpoints. Their concept can be adopted to define "optimal" futility boundaries (with respect to given performance indicators) for continuous endpoints.

Methods: We extend Schülers' definition for "optimal" futility boundaries to the most common study situation of a single continuous primary endpoint compared between two groups. First, we introduce the analytic algorithm to derive these futility boundaries. Second, the new concept is applied to a real clinical trial example. Finally, the performance of a study design with an "optimal" futility boundary is compared to designs with arbitrarily chosen futility boundaries.

Results: The presented concept of deriving futility boundaries allows to control the probability of wrongly stopping for futility, that means stopping for futility even if the treatment effect is promizing. At the same time, the loss in power is also controlled by this approach. Moreover, "optimal" futility boundaries improve the probability of correctly stopping for futility under the null hypothesis of no difference between two groups.

Conclusions: The choice of futility boundaries should be thoroughly investigated at the planning stage. The sometimes met, arbitrary choice of futility boundaries can lead to a substantial negative impact on performance. Applying futility boundaries with predefined optimization criteria increases efficiency of group sequential designs. Other optimization criteria than proposed in here might be incorporated.

Keywords: Futility stop, Group sequential design, Continuous endpoint

*Correspondence: geraldine.rauch@charite.de
[1] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany
[2] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany

Background

Conducting clinical trials which fulfil both economical as well as ethical aspects requires extensive efforts in planning. This can be challenging in fixed design clinical trials, as there is no option to react to misspecified planning assumptions during the ongoing trial. Group sequential designs allow for an early stop for either efficacy or futility, thereby, allowing to reduce costs and ethical issues when interim results are either sufficiently convincing or do not justify a further investigation. Group sequential designs are characterized by one or several unblinded interim analyses, thus implying a multiple test problem. Popular methods for alpha adjustment were proposed by Pocock [1] and O'Brien and Fleming [2]. Later, more flexible methods were developed with the idea to define alpha spending functions [3–5]. Following these developments, in the past decades, an increasing number of trials adopted such flexible designs.

Whereas the option for an early efficacy stop is a key feature of group sequential designs, futility stops are not routinely implemented. Stopping a trial early for efficacy implies a successful trial with reduced costs. The probability to stop for efficacy although there is no treatment benefit is naturally controlled by the significance level. In comparison, stopping a trial early for futility means to give up hope for a successful trial based on an interim effect which might have low precision due to small sample sizes at interim. Thereby, the futility stopping boundary is usually defined as a boundary for the interim *p*-value. Valid stopping for futility bounds could reduce costs and avoid involving more patients under unnecessary risks, whereas wrong stopping for futility corresponds to a waste of resources.

Among futility stopping methods of group sequential designs, two main rules are discussed in the literature. Futility stopping rules can either be binding or non-binding, where binding means that stopping is mandatory if the criterion is met and non-binding means that the investigator can freely decide if he or she really wants to stop. Type I error control is guaranteed for both types but there is a decrease in the actual power. In clinical practice, non-binding rules are much more common, as usually it is not only the interim data that affects a decision but also new external data or safety information. When concentrating on binding rules, it is possible to choose larger local significance levels in order to fully exhaust the global significance level [6]. However, this option is usually not applied in practice and more attention should be given to the non-binding option.

There exist sound and broad theoretical methodologies on group sequential designs. In particular, theoretically justified choices of futility stopping boundaries were discussed already decades ago [7, 8]. Several authors [9–12] addressed this issue more generally by defining beta-spending functions in analogy to the well-known alpha-spending functions where the latter take account of the multiplicity issue in group sequential designs. The beta-spending function allows to monitor and control the stage-wise and the global power loss induced by the futility stop.

As additional performance measures for futility boundaries, He et al. [13] referred to the conditional and the predictive power. Gallo et al. [14] more generally discussed performance indicators for choosing futility boundaries including the global power loss, the conditional power, the predictive power, and the probability of correctly stopping for futility under the null hypothesis. In another work of Xi et al. [15], an optimal tuple of the futility boundary and the time point for the interim analysis is determined. This tuple is chosen as a solution of an optimization problem given by an objective function with constraints, where a bound for the power loss defines the constraint and the average sample size defines the performance function. Optimization functions with constraints in the context of adaptive designs have also been recently discussed by Pilz et al. [16]. Instead of formulating constraints, Ondra et al. [17] discuss several adaptive designs by means of optimizing prespecified utility functions. Schüler et al. [18] defined "optimal" futility stopping boundaries under predefined optimality criteria, however for the very special case of (multiple) time-to-event endpoints. Thereby, they rely on the performance measures given by power loss, probability of wrongly stopping for futility and probability of correctly stopping for futility. Whereas approaches based on optimization problems with constraints or maximizing utility functions can be seen as more elegant mathematical solutions, the approach by Schüler et al. [18] might have advantages in the communication to clinical researchers as their basic idea for "optimal" futility boundaries is simply understood: For a given sample size and effect under the alternative, the futility bound which preserves a predefined level of a wrong futility classification is determined. This value serves as a starting value for the "optimal" futility boundary. It is enlarged until the power loss is decreased to an acceptable limit. This defines the "optimal" futility boundary.

Despite these important works, the above reported performance indicators are often not investigated when setting the futility boundary in clinical applications. In particular in investigator initiated trials, futility boundaries are often chosen rather arbitrarily. A common choice in a superiority test setting is a futility boundary of 0.5, where the study is stopped whenever the one-sided interim *p*-value lays above this boundary. This corresponds to the situation of the treatment effect pointing in the wrong direction. For example, the software ADDPLAN which implements sample size recalculation for group sequential designs, sets a default value of 0.5 when a futility stop is included [19]. Moreover, within the R-Package `rpact` short examples with this futility boundary of 0.5 are provided for illustration [20]. However, this choice of the futility boundary is usually not justified by design performance characteristics. Note however that other sample size calculation software such as nQuery or Pass implement beta-spending functions as a default, so there is no unique standard [21, 22].

In this work, we aim to adopt the approach by Schüler et al. [18] for the more common case of a controlled trial comparing two groups with a continuous endpoint. Whereas for multiple correlated time-to-event endpoints, the findings of optimal futility boundaries can only be realized by simulations, this more simple case allows a straight forward analytical derivation. Using this common and simple design, we aim to contribute to a more profound discussion on futility boundaries in practice and aim to provide an easy understandable and easily applicable tool to overcome the potential gap between developed theory and clinical practice.

This work is structured as follows: In the Methods Section, we introduce the underlying test problem and the group sequential design. Subsequently, we introduce the definition of "optimal" futility boundaries by Schüler et al. [18] adapted to the situation of a continuous primary endpoint. In the Results Section, we first illustrate the concept of the investigated optimality conditions for futility boundaries for the setting of an exemplary clinical trial. Secondly, we compare the performance characteristics of a study with optimally chosen futility boundaries to those with non-optimal boundaries for various design scenarios, where the expression "optimal" in the following refers to the investigated performance criteria. Finally we discuss our results and provide conclusions and implications for future clinical trials.

Method

Throughout this work, we consider a randomized controlled trial with a continuous primary endpoint which is compared between a new intervention (I) and a control treatment (C)

$$X_i^I \sim \mathcal{N}\left(\mu^I, \sigma^2\right),\; X_i^C \sim \mathcal{N}\left(\mu^C, \sigma^2\right),\; i = 1 \ldots n.$$

For the sake of simplicity, we consider equal standard deviations σ and group sizes n. The test hypotheses are given in terms of a superiority test

$$H_0 : \mu^I - \mu^C \leq 0 \text{ versus } H_1 : \mu^I - \mu^C > 0. \tag{1}$$

Thereby, without loss of generality, a larger value of the endpoint is assumed to be favorable.

Group sequential design

We consider a group sequential design with two sequences, that is with one interim analysis. The total maximal sample size is $N = 2 \cdot n$, the total interim sample size is denoted by $N_1 = 2 \cdot n_1$. The interim test statistic can be formulated as

$$T_1 := \frac{\bar{X}_1^I - \bar{X}_1^C}{S_{pooled,1}} \cdot \sqrt{\frac{n_1}{2}}, \tag{2}$$

with observed interim means $\bar{X}_1^I$, $\bar{X}_1^C$ and a pooled standard deviation at interim $S_{pooled,1}$. This test statistic corresponds to the normal approximation test for continuous endpoints.

The study is stopped for efficacy at the interim stage in case the one-sided interim p-value p_1 is smaller than or equal to the adjusted local one-sided significance level $p_1 \leq \alpha_1$.

The study is stopped for futility if $p_1 > \alpha_0$, where α_0 is the futility boundary.

If the trial is not stopped within the interim analysis, then additional $N_2 = N - N_1$ patients are recruited. The test statistic for the final analysis is then given by

$$T_{1+2} := \frac{w_1 \cdot T_1 + w_2 \cdot T_2}{\sqrt{w_1^2 + w_2^2}}, \tag{3}$$

where T_2 is the independent incremental test statistic including exclusively the data of the second stage and w_1, w_2 are predefined weights which must be fixed at the planning stage. This is also known as the inverse normal combination test [23] as the stage-wise test statistics can be written as the inverse of the normal distribution function applied to the stage-wise p-values $T_i = \Phi^{-}(p_i),\ i = 1, 2$. The combination of p-values provided by the inverse normal method is just one option among others to combine the stage-wise p-values. Another famous approach would be the use of the Fisher combination test [24]. The idea presented in here is also transferable when using another combination function.

A common way to choose the above weights in the inverse normal combination function is to define $w_1 = \sqrt{n_1}$ and $w_2 = \sqrt{n_2}$.

The null hypothesis H_0 is rejected at the final analysis if the corresponding p-value is smaller than or equal to the adjusted local one-sided significance level $p_{1+2} \leq \alpha_{1+2}$. The key idea of the inverse normal approach is that by constructing the final test statistic T_{1+2} from the independent stage-wise test statistics T_1 and T_2, the covariance of the joint distribution of T_1 and T_{1+2} is

$$Cov\left(T_1, T_{1+2}\right) = \sqrt{\frac{n_1}{n}},$$

and thus the joint distribution is fully specified.

The local significance levels for the interim analysis and the final analysis can be specified such that the overall type I error is controlled, that is

$$P_{H_0}\left(p_1 \leq \alpha_1 \cup (\alpha_1 < p_1 \cap p_{1+2} \leq \alpha_{1+2})\right) = \alpha. \qquad (4)$$

If binding futility stopping boundaries are applied, the futility boundary α_0 can be incorporated in the above equation to obtain larger optimized local significance levels α_1 and α_{1+2}. We will not consider this option, as even if a fixed futility stopping rule is incorporated in the trial protocol, there are often external reasons to make exceptions from this binding rule, which is not a problem as long as the local significance levels are chosen according to Eq. (4).

The local significance levels α_1 and α_{1+2} can be derived using various existing methods, such as constant levels as proposed by Pocock [1], increasing levels as given by O'Brien-Flemming [2], or flexible alpha spending functions as e.g. described by Lan and DeMets [4]. In our work, for the sake of simplicity, we rely on Pocock boundaries that is $\alpha_1 = \alpha_{1+2}$. The remaining question is how to choose an adequate value of α_0 already at the planning stage.

Optimality criteria for futility boundaries

The idea of "optimal" futility boundaries proposed by Schüler et al. [18] is to assure a high probability to stop correctly for futility. This means stopping when there is only no or a non-relevant treatment effect, while simultaneously controlling the loss in power and the probability of correctly stopping for futility when in fact, the underlying treatment effect is relevant. In the following, we will use the term "optimal" with respect to these criteria. As discussed in the introduction, there are however various other performance indicators and different approaches to quantify the total performance. Therefore, optimality is not a unique perspective and we do not intend to present the "best" solution. In the following, assume that the trial is powered to detect a standardized effect $\Delta = \frac{\mu^I - \mu^C}{\sigma}$ with power $1 - \beta$ at a global one-sided significance level of α. To introduce the concept of optimal futility boundaries, some additional parameters are required: Let $Pow_{\text{loss}} < 1 - \beta$ denote the admissible overall power loss caused by applying a binding futility boundary. Moreover, the probability to wrongly stop for futility when in fact the underlying standardized treatment effect is given by the relevant effect Δ should be limited by $\pi_{\text{wrong}} \in [0, 1]$. Using these notations, a futility boundary fulfils the so called **admissible conditions** [18] if the following requirements are satisfied:

1. $P_\Delta(p_1 > \alpha_0) \leq \pi_{\text{wrong}}$,
2. $P_\Delta\left(p_1 \leq \alpha_1 \cup (\alpha_1 < p_1 < \alpha_0 \cap p_{1+2} \leq \alpha_{1+2})\right) \geq 1 - \beta - Pow_{\text{loss}}$.

Note that the concept of the optimality parameter Pow_{loss} is similar to the beta-spending approach proposed by several authors [9–12]. The beta-spending approach allows to control the stage-wise power loss induced by futility stopping boundaries. In contrast, we exclusively focus on the global power loss. Note that both approaches guarantee a limited (stage-wise) power loss only for the assumed effect Δ. For smaller effects the power loss can become unacceptably high. Therefore, we strongly recommend to choose Δ as the *minimal* clinically relevant effect and not as the expected effect.

In the following, any futility boundary fulfilling the admissible conditions will be denoted as $\alpha_{0,\text{ad}}$. Note that for a clinical trial with a continuous endpoint and the design specifications given above, the first admissible condition can be translated into

$$\alpha_{0,\text{ad}} \geq 1 - \Phi\left(z_{\pi_{\text{wrong}}} + \Delta \cdot \sqrt{\frac{n_1}{2}}\right),$$

where $\Phi(*)$ denotes the distribution function of the standard normal distribution and $z_{(*)}$ denotes the corresponding quantile of the standard normal distribution. The second admissible condition is equivalent to

$$\begin{aligned} 1 - \Phi&\left(z_{1-\alpha_1} - \Delta \cdot \sqrt{\frac{n_1}{2}}\right) \\ +MV_{\mu,\Sigma}&\left(z_{1-\alpha_1}, z_{1-\alpha_{1+2}}\right) \\ -MV_{\mu,\Sigma}&\left(z_{1-\alpha_{0,\text{ad}}}, z_{1-\alpha_{1+2}}\right) \\ \geq 1 - \beta& - Pow_{\text{loss}}, \end{aligned}$$

where $MV_{\mu,\Sigma}(*)$ is the distribution function of the multivariate normal distribution with expectation

$$\boldsymbol{\mu} = \left(\Delta \cdot \sqrt{\frac{n_1}{2}}; \Delta \cdot \sqrt{\frac{n}{2}}\right)$$

and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sqrt{\frac{n_1}{n}} & 1 \\ 1 & \sqrt{\frac{n_1}{n}} \end{pmatrix}.$$

For predefined parameters Pow_{loss} and π_{wrong}, there exists a whole set of admissible futility boundaries fulfilling the above conditions. Only the probability of correctly stopping for futility is left to further optimize an admissible futility stopping boundary. As the probability to correctly stop for futility increases with decreasing futility boundary, this implies that the optimal futility boundary $\alpha_{0,\text{opt}}$ is the minimum over the set of all admissible futility boundaries $\alpha_{0,\text{ad}}$. With this definition, we can compute the optimal futility boundary at the planning stage of a clinical trial analytically. However, it can happen that the actual achievable probability to correctly stop for futility is still considered as too small. In this case, it might be reasonable to choose slightly larger values of Pow_{loss} and π_{wrong}.

Results

Given predefined design parameters, the optimal futility boundaries can be analytically computed at the planning stage. An R-function which calculates the "optimal" futility boundary for arbitrary design parameters is provided as online supplementary material (see Additional File 1).

A clinical trial example

In the following, we will illustrate the benefit of using an optimal futility boundary compared to an arbitrary choice of a futility boundary by means of a real clinical trial example.

The ChroPac-Trial [25] is a blinded, randomized, controlled clinical trial. The primary endpoint is the quality of life of patients with chronic pancreatitis 24 months after surgical interventions. The intervention group receives a duodenum-preserving pancreatic head resection and is compared to a control group receiving pancreatoduodenectomy. The aim is to show superiority of the intervention. The primary endpoint is measured by the quality of life questionnaire EORTC QLQ-C30, which provides a score for physical functioning. The score ranges from 0 to 100 with a higher score indicating a better quality of life. Although a score is generally seen as an ordinal endpoint, it is a common approach to treat a score with a large range as a continuous endpoint. A score difference of 10 is considered as a clinically relevant treatment difference and 20 is assumed to be the common standard deviation.

The trial was planned to detect a standardized treatment effect $\Delta = \frac{10}{20} = 0.5$ at a one-sided global significance level $\alpha = 0.025$ with power $1 - \beta = 0.90$. This results in a total sample size of 172 patients (86 per group) when the null hypothesis is tested with a standard t-test for independent groups. Note that the original trial was planned with a fixed design. For illustrative purposes, we will now apply a group sequential design to illustrate the new concept.

Applying a two-stage group sequential design with an interim look after 50% of the patients being fully observed and local adjusted significance levels according to Pocock with $\alpha_1 = \alpha_{1+2} = 0.0147$, the above sample size yields a power of 0.88. In order to apply the concept of an optimal futility boundary now, we need specifications of Pow_{loss} and π_{wrong}. A power loss caused by futility stopping of $Pow_{\mathrm{loss}} = 0.05$ is considered reasonable. The probability to wrongly stop for futility should of course be small. Thus, we may choose $\pi_{\mathrm{wrong}} = 0.05$. With these parameter settings, the optimal futility boundary is given by $\alpha_{0,\mathrm{opt}} = 0.22$.

It is also common to anticipate a power of 0.80. Therefore, as a reference design, we will also calculate the optimal futility boundary for the above setting when the global maximal sample size of the group sequential design is only $N = 140$, which results in a power of 0.80 without stopping for futility. In this case, the optimal futility boundary is given by $\alpha_{0,\mathrm{opt}} = 0.33$.

The two admissible parameters, power loss Pow_{loss} and the probability of wrongly stopping for futility π_{wrong}, determine jointly the optimal futility boundary $\alpha_{0,\mathrm{opt}}$. Therefore, $\alpha_{0,\mathrm{opt}}$ can be displayed as a function of these two parameters as illustrated in Fig. 1, which allows to investigate graphically how the optimal futility boundary changes when the admissible parameters are varied.

From Fig. 1 it can nicely be seen that the optimal futility boundary also depends on the sample size, where a larger sample size results in a smaller futility boundary. It can be seen that for $N = 140$, the optimal futility boundary is mainly determined by the parameter π_{wrong}, whereas for $N = 188$ the influence of Pow_{loss} grows. In order to quantitatively assess the impact of variations of the admissible parameter settings, Table 1 shows the resulting optimal futility boundaries for selected parameter values of Pow_{loss} and π_{wrong} for both sample size settings $N = 188$ and 140.

Column 1 displays the underlying sample size. Columns 2 and 3 show the specification of the admissible condition parameters Pow_{loss} and π_{wrong}. The resulting optimal futility boundary is displayed in Column 4. Columns 5 to 8 show the performance of the design by various performance measures such as the actually achieved power including stopping for futility (Column 5), the probability of wrongly stopping for futility under the anticipated relevant effect Δ (Column 6), as well as the probability of correctly stopping for futility under a small non-relevant effect, which is half the size of the anticipated effect and under the null hypothesis with no effect (Columns 7 and 8).

It can be seen from Table 1 that a very low value of $\pi_{\mathrm{wrong}} = 0.01$ results in high optimal futility boundaries, which may be close to the often arbitrarily chosen value of $\alpha_0 = 0.5$ (Row 1) or even larger (Row 8). However, the probability of correctly stopping for futility is relatively low in these scenarios.

Looking at the parameter settings where the resulting probability of correctly stopping for futility under either the null hypothesis or half of the relevant treatment effect is at least above 20%, it can be deduced that a slightly larger value of, e.g. $\pi_{\mathrm{wrong}} = 0.05$, is a better choice.

The admissible power loss Pow_{loss} is generally often not exhausted, especially for smaller values of π_{wrong}. For example, a change in the parameter Pow_{loss} does not have an impact on the optimal futility boundary when π_{wrong} is fixed to either 0.01 or 0.05 for $N = 140$.

Note that a conventional choice of the futility boundary is $\alpha_0 = 0.5$. Looking at Table 1 it can be seen that for the favorable settings, where the probability of correctly stopping for futility is not too small and lays above 20%, the optimal futility boundaries are considerably smaller than the conventional choice of $\alpha_0 = 0.5$. For $N = 188$ the optimal futility boundaries range between $\alpha_{0,\text{opt}} = 0.13$ and $\alpha_{0,\text{opt}} = 0.46$, for $N = 140$ between $\alpha_{0,\text{opt}} = 0.21$ and $\alpha_{0,\text{opt}} = 0.59$.

Discussion

Although efficacy boundaries in group sequential designs are widely discussed in the literature, the choice of futility boundaries gains much less attention in clinical applications. A naive choice choice of a futility boundary of $\alpha_0 = 0.5$, where an interim p-value of $p_1 > 0.5$ suggests an early stopping for futility, means that at interim, as soon as the treatment effect points into the adverse direction, the trial is stopped. Although this is intuitive, the implications of this futility boundary choice on the design performance are not always investigated. However, the choice of the futility boundary naturally influences the power of the study design. Moreover, a large futility boundary implies that the probability to stop the study, when indeed there is no or only a non-relevant effect (correct stopping for futility), can be small. In contrast, a too low futility boundary can imply that the probability of wrongly stopping for futility, when there is a relevant treatment effect, is considered as too large.

Some authors have proposed adaptive design strategies to optimize design parameters, like the value of the futility boundary and the number and timing of interim looks [11, 13, 14, 16, 17]. Thereby, different concept were proposed, e.g. optimization problems with constraints [14, 16] or maximization of utility functions [17]. A comparison between these different approaches is still lacking. The optimality criteria initially proposed by Schüler et al. [18] allow to define a relatively simple concept of "optimal" futility boundaries, which was originally proposed in the context of (composite) time-to-event endpoints, by balancing the performance characteristics of global power loss and the probability of correctly and wrongly stopping for futility. The task of this work was to adapt this concept to the more general case of a group sequential design with a continuous endpoint. We showed that with the concept of optimal futility boundaries, it is possible to quantify the performance characteristics and the implications of a futility boundary already at the planning stage. By a clinical trial example, we demonstrated that arbitrarily choosing $\alpha_0 = 0.5$ can lead to very unfavorable performance characteristics in some situations. However, there are also trial settings, where the choice of $\alpha_0 = 0.5$ is close to or even smaller than the optimal one. This highlights the necessity to investigate the implications of different futility stopping boundaries already at the planning stage.

The concept of optimal futility boundaries fits the regulatory guidance documents provided by the U.S. Food and Drug Administration [26] and European Medicines Agency [27] for confirmatory trials. If a trial sponsor aims at applying our method in a confirmatory trial, the power loss and probability of wrongly and correctly stopping for futility can be predefined as two additional design parameters in the clinical trial protocol. Simulations are not required as the operating characteristics

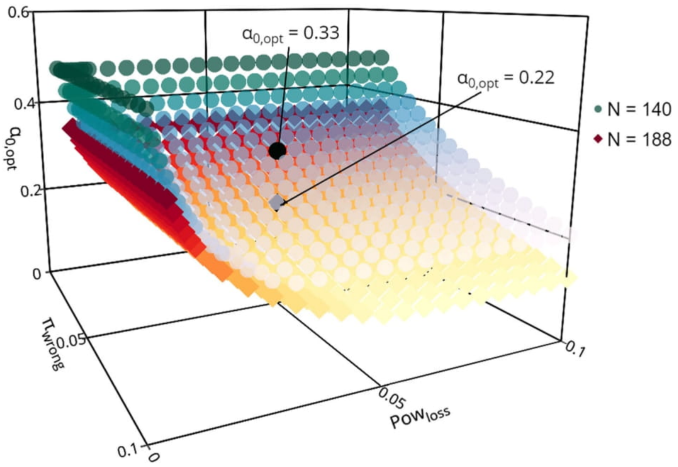


Fig. 1 The "optimal" futility boundary $\alpha_{0,\text{opt}}$ as a function of the admissible parameters Pow_{loss} and π_{wrong} for $N = 140$ (blue dots) and $n = 188$ (red squares). The black symbols highlight the "optimal" futility boundaries for $Pow_{\text{loss}} = 0.05$ and $\pi_{\text{wrong}} = 0.05$

Table 1 Performance characteristics for the group sequential design with "optimal" futility boundaries based on different admissible condition parameters for $N = 182$ and $N = 140$. The last lines in the two sample size settings show the performance characteristics for the arbitrary choice of $\alpha_0 = 0.5$

Sample	Admissible condition		"Optimal" futility	Actual	Probability of wrongly	Probability of correctly	
size	parameters		boundary	power	stopping for futility	stopping for futility	
n	Pow_{loss}	π_{wrong}	$\alpha_{0,\text{opt}}$	under $\Delta = 0.5$	$P_{\Delta_{\text{true}}=0.5}\ (p_1 > \alpha_0)$	$P_{\Delta_{\text{true}}=0.25}\ (p_1 > \alpha_0)$	$P_{\Delta_{\text{true}}=0.0}\ (p_1 > \alpha_0)$
188	0.01	0.01	0.46	0.90	0.01	0.13	0.54
	0.05	0.01	0.46	0.90	0.01	0.13	0.54
	0.01	0.05	0.29	0.89	0.03	0.26	0.71
	0.05	0.05	0.22	0.89	0.05	0.33	0.78
	0.01	0.10	0.29	0.89	0.03	0.26	0.71
	0.05	0.10	0.13	0.85	0.10	0.47	0.87
	0.0013	0.008	0.50	0.90	0.01	0.11	0.50
140	0.01	0.01	0.59	0.80	0.01	0.10	0.41
	0.05	0.01	0.59	0.80	0.01	0.10	0.41
	0.01	0.05	0.33	0.79	0.05	0.27	0.67
	0.05	0.05	0.33	0.79	0.05	0.27	0.67
	0.01	0.10	0.32	0.79	0.05	0.28	0.68
	0.05	0.10	0.21	0.77	0.10	0.41	0.79
	0.0013	0.018	0.50	0.80	0.02	0.15	0.50

can be derived analytically for continuous endpoints. An R-code providing the analytical solution is provided as online supplementary material (see Additional File 1). Thus, the design modifications are easily calculated and communicated which is a requirement of the FDA guidance [26].

A possible limitation of the presented futility concept is that the choice of the admissible condition parameters, which limits the power loss and controls the probability of wrongly stopping for efficacy, is to a certain extend arbitrary. We therefore recommend to calculate the optimal futility boundaries for a range of plausible admissible condition parameters and to investigate the performance characteristics. In particular, the probability of correctly stopping for futility should be reasonably high (above 20% as a rule of thumb). This approach can lead to a reasonable choice of the futility boundary that provides a fair balance between the different performance characteristics.

In this work, we concentrated on a two-stage group sequential design with a continuous endpoint with local significance levels adjusted according to Pocock [1]. The corresponding R-source code (see Additional File 1) can be easily adapted to use other alpha spending functions and other *p*-value combination tests. Moreover, the concept can equivalently be adopted to binary endpoints, which will be the task of future work.

An attractive argument for the presented approach lays in the simplicity of the key idea. In particular within investigator initiated trials, there often exist not theoretically founded recommendations for choosing futility bounds. One main aim of this article is thus to encourage the theoretical justification of futility boundaries in practical applications. There are different ways to do so of which our approach is only one option.

Conclusions

While other trial design parameters and operational characteristics are routinely investigated in the planning stage of group sequential designs, futility boundaries should not be neglected. The concept of an "optimal" futility boundary method as introduced in here allows to control the power loss and the probability of wrongly stopping for futility, while maximizing the probability of correctly stopping for futility. We recommend to investigate futility boundaries following our approach over a range of parameter settings and to carefully compare the resulting futility boundaries to the arbitrary choice of $\alpha_0 = 0.5$ when planning a trial with a group sequential design.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-020-01141-5.

Additional file 1: The file contains the R-source code for the calculation of the "optimal" futility boundary for a predefined design setting and admissible parameters.

Acknowledgements
Not applicable.

Authors' contributions

XL implemented the R-programs, produced the results and wrote the first draft of the manuscript. GR contributed to all parts of the manuscript. CH checked the analytical derivations and revised the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. Geraldine Rauch recieved funding from the German Research Foundation (**RA 234714-1**).

Availability of data and materials
The R-source code for the calculation of the optimal futility boundary is available (see Additional File 1).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 17 May 2020 Accepted: 7 October 2020
Published online: 05 November 2020

References

1. Pocock S. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191–9.
2. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. Biometrics. 1979;35(3):549–56.
3. DeMets D, Ware J. Group sequential methods for clinical trials with a one-sided hypothesis. Biometrika. 1980;67:651–60.
4. Lan K, DeMets D. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659–63.
5. Wang S, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics. 1987;43:193–9.
6. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Stat Med. 2009;28(8):1181–217.
7. DeMets D, Ware J. Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika. 1982;69(3):661–3.
8. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. Biometrics. 1983;39:227–36.
9. Chang M, Hwang I, Shin W. Group sequential designs using both type I and type II error probability spending functions. Commun Stat Theory Methods. 1998;27(6):1323–39.
10. Pampallona S, Tsiatis A. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. J Stat Plan Infer. 1994;42:19–35.
11. Pampallona S, Tsiatis AA, Kim K. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. Drug Inf J. 2001;35(4):1113–21.
12. Rudser K, Emerson S. Implementing type I & type II error spending for two-sided group sequential designs. Contemp Clin Trials. 2008;29:351–8.
13. He P, Lai T, Liao O. Futility stopping in clinical trials. Stat Interface. 2012;5: 415–23.
14. Gallo P, Mao L, Shih VH. Alternative views on setting clinical trial futility criteria. J Biopharm Stat. 2014;24(5):976–93.
15. Xi D, Gallo P, Ohlssen D. On the optimal timing of futility interim analyses. Stat Biopharm Res. 2977;9:293–301.
16. Pilz M, Kunzmann K, Herrmann C, Rauch G, Kieser M. A variational approach to optimal two-stage designs. Stat Med. 2019;38(21):4159–71.
17. Ondra T, Jobjörnsson S, Beckman RA, Burman C-F, König F, Stallard N, Posch M. Optimized adaptive enrichment designs. Stat Methods Med Res. 2019;28(7):2096–111.
18. Schüler S, Kieser M, Rauch G. Choice of futility boundaries for group sequential designs with two endpoints. BMC Med Res Methodol. 2017;17(119):.
19. ICON. ADDPLAN: Adaptive Designs - Plans and Analyses. 2014. https://www.iconplc.com/innovation/addplan/.
20. Wassmer G, Pahlke F. Rpact: Confirmatory Adaptive Clinical Trial Design and Analysis. 2018. https://rdrr.io/cran/rpact/man/rpact.html.
21. Statsols: Statistical Solutions Ltd. nQuery: Sample Size and Power Calculation. 2017. https://www.statsols.com/nquery.
22. NCSS. PASS: Power Analysis and Sample Size Software. 2019. https://www.ncss.com/software/pass/.
23. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. 1029–41. 1994.
24. Fisher RA. "Statistical methods for research workers". Statistical methods for research workers. 1934;5th Ed:.
25. Diener M, Bruckner T, Contin P, Halloran C, Glanemann M, Schlitt H, Mössner J, Kieser M, Werner J, Büchler M, Seiler C. ChroPac-trial: duodenum-preserving pancreatic head resection versus pancreatoduodenectomy for chronic pancreatitis. Trial protocol of a randomised controlled multicentre trial. Trials. 2010;11:47.
26. Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics - guidance for industry. Food and Drug Administration. 2019. https://www.fda.gov/media/78495/download.
27. European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency. 2007. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

10 Curriculum vitae

My curriculum vitae does not appear in the electronic version of my paper for reasons of data protection.

11 List of publications

X. Li, C. Hermann, and G. Rauch. Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint. *BMC Medical Research Methodology*, 20(274), 2020.

12 Acknowledgements

First of all, I would like to express my deepest thanks to my first supervisor Professor Geraldine Rauch. I am very grateful for her guidance and granting me with the great opportunity to work at the institute and complete my studies. Her passion and vision for clinical research and statistics motivated me to conduct the better research and her support and encouragement helped me to truly become a scientist. Also, I want to thank Dr. Rene Schmidt for his support as my co-supervisor.

I want to express my sincere gratitude to Carolin Hermann and Michaela Maria Freitag for providing valuable support and the fruitful discussions. There are also many great colleagues at the institute who supported me throughout the journey. I want to thank them for their motivating words throughout the doctorate program. It was a great pleasure to have colleagues who listened to my problems and cheered me up. I am also very grateful to my dear friends, who kept my spirits up to overcome various challenges during my studies.

Lastly, a very special and deep thank to my amazing family for their unconditional moral support. My parents always listened to me and believed in me. Their love guided me through whenever it was tough. I am also very blessed to have a loving partner full of patience, care and wisdom. Thank you for being there when I needed most.