Confidence Intervals for the Treatment Effects in Adaptive Enrichment Designs

Jinyu Zhu, BBA, MSc

Mathematics and Statistics Department

Lancaster University

A thesis submitted for the degree of

Doctor of Philosophy

August, 2023

Abstract

While safety is the primary objective in Phase I designs of pharmacological or novel treatment development clinical trials, the focus shifts to detecting the effectiveness of the experimental treatment in confirmatory seamless Phase II/III designs. The presence of patient heterogeneity in modern medication development is widely acknowledged. The response of patients to the same treatment might vary depending on factors such as their gender, age, lifestyle, or genetic diversity. Therefore, it is necessary to determine which group of the population is more likely to benefit from the experimental treatment in the Phase II/III designs. In order to save time and cost, the adaptive enrichment design was proposed. The adaptive enrichment design concentrates resources on promising subgroups by allowing modifications based on the interim analysis results. However, the adaptive nature of the procedure complicates the estimation of the treatment effects and makes the quantification of uncertainty in treatment effects challenging. In particular, confidence intervals based on the naive maximum likelihood estimate and corresponding Fisher information will tend to have incorrect coverage. Focusing on a two-stage design with two disjoint subgroups, we develop a general method based on devising an appropriate p-value function. We derive the conditional confidence intervals for selected subgroups by inverting their corresponding conditional p-value functions, which are obtained using stage-wise, score, and MLE sample space orderings methods. Comparing the confidence intervals produced from the aforementioned space ordering methods reveals that score ordering treats each stage more evenly. Additionally, we construct the unconditional p-value function for each subgroup and utilize the classic Bonferroni, Bonferroni-Holm, and parameter-dependent weighted Bonferroni multiple testing procedures to create simultaneous confidence intervals at the end of the trial. We demonstrate that Bonferroni-Holm is most effective at detecting actual treatment effects, but its confidence interval for rejected hypotheses is uninformative when not all hypotheses are rejected. In contrast, the traditional Bonferroni simultaneous confidence intervals provide information regarding the magnitude of

the real treatment effect but are less effective at rejecting false null hypotheses. The weighted parameter-dependent Bonferroni method compromises between informativeness and power. The confidence interval construction approach is illustrated through the application of two adaptive enrichment designs. Simulation studies show that our approach constructs confidence intervals with exact asymptotic coverage probabilities. Our method may be extended to k-stage m-subgroup adaptive enrichment design with $k \geq 3$ and $m \geq 3$; although, the computation cost will also increase.

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography. A rough estimate of the word count is 25805.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Dr Fang Wan and Professor Andrew Titman. Their guidance, wisdom, and expertise have been instrumental in shaping the trajectory of this research. They have always been there when I needed their valuable advice or direction. Their mentorship has enriched my understanding and honed my skills, and I am genuinely grateful for their patience and dedication. I extend my sincere appreciation to the members of my thesis committee, Dr. Juhyun Park and Dr. Gareth Ridall, for their invaluable insights, constructive feedback, and scholarly guidance that have immensely contributed to the refinement of this work. I would also like to express my gratitude to my external examiner, Dr. Peter Kimani, who provided valuable comments on my thesis and contributed to shaping its final version.

I am thankful to the faculty and staff in Mathematics and Statistics department at Lancaster University for providing a stimulating academic environment, abundant resources, and opportunities for intellectual growth. A special note of appreciation goes to Mrs. Sharon Bryan and the IT help desk, who have consistently demonstrated patience whenever I faced challenges and provided selfless assistance when I needed it the most.

I wish to acknowledge the Mathematics and Statistics department, Faculty of Science and Technology for their financial support, which has enabled me to wholeheartedly engage in this research and achieve my dream.

I am also deeply grateful for the support, understanding, and encouragement from my parents and grandparents. Their presence has been pivotal in enabling me to overcome challenges and persist in my journey. Their unwavering support has been my cornerstone, and I will forever be grateful for their belief in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the widespread adoption of human genome sequencing techniques, the necessity to identify patient heterogeneity in medical practice has become apparent (Hodson, 2016). As a result, precision medicine has become an appealing concept in clinical treatment development. The increasing recognition of heterogeneity in clinical trials has led to the realization that the traditional one-size-fits-all approach to treatment is insufficient (Knottnerus and Tugwell, 2013). Patients may respond differently to the same treatment based on factors such as their gender, age, genes, lifestyle, or environment. Precision medicine involves a series of decision-making actions, including treatment selection and dose-finding, aimed at identifying the best treatment for specific patients (Kosorok and Laber, 2019; FDA, 2018). For instance, breast tumors can be classified into at least three subtypes: luminal, human epidermal growth factor receptor 2+ (HER2+), and basal-like. Luminal tumors respond well to hormonal interventions, whereas HER2+ tumors should be treated with anti-HER2+ therapies. As for basal-like tumors, still, no targeted therapy is currently available (Polyak et al., 2011). Another example is the medication to treat depression. Fournier et al. (2010) found that the antidepressant is more effective for patients with HDRS scores above 25. For those patients with HDRS scores below 25, the medication is not superior compared to the placebo. Therefore, it is essential to identify the most appropriate patient population group before introducing a new treatment to the market. In order to screen out the promising population of an experimental medication, the adaptive enrichment design was introduced in Phase II/III clinical trials. The enrichment design allows for various modifications based on the interim analysis, such as sample size re-estimation and subgroup selection. However, those adaptive modifications inevitably introduce bias and difficulties in parameter

inference. This thesis concentrates on the adaptive enrichment design with two stages and two subgroups, which includes one experimental arm and one control arm. Specifically, the design proposed by Magnusson and Turnbull (2013) is given particular attention.

Because only promising subgroups are enriched and no data are collected for dropped subgroups in subsequent stages, unbiased point estimates for all subgroups have not yet been established. Point estimates do not account for uncertainty and inference validity. Therefore the FDA (2007) recommends that "appropriate measures include estimates of sensitivity and specificity pairs, likelihood ratio of positive and negative result pairs, and ROC (Receiver Operating Characteristic) analysis along with confidence intervals" should be reported. Additionally, the Consolidated Standard of Reporting Trials (CONSORT, 2010) states "for each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)" should be reported. Although there is a significant amount of research on constructing bias-reduced point estimates for the adaptive enrichment design, few papers focus on constructing confidence intervals with coverage probability close to a nominal level. Therefore, we propose a relatively straightforward approach that constructs confidence intervals by inverting p-values that are a function of the treatment effect parameter.

In Chapter 3, we derive p-value functions for specific subgroups using the enrichment design proposed by Magnusson and Turnbull (2013), based on the interim analysis decision results and three sample space ordering methods. By inverting these functions, we obtain conditional confidence intervals for the chosen subpopulation. In Chapter 4, we initially construct unconditional p-value functions for individual subgroups regardless of selection results. To construct simultaneous confidence intervals for all subgroups, we utilize multiple testing procedures, such as Bonferroni, Holm, and parameter-dependent weighted Bonferroni. Our methodology is further expanded in Chapter 5, where we apply it to the sample size re-estimation design proposed by Lin et al. (2021). Finally, Chapter 6 summarizes our research and highlights possible areas for future work.

Chapter 2

Literature review

2.1 Adaptive enrichment design

Traditional single-stage fixed sample size designs conduct analysis at the termination of the trial only. Such designs are straightforward but inflexible since they do not allow desirable or necessary changes to occur during the trial (Pallmann et al., 2018). In addition to that, the traditional fixed sample design usually requires larger patient recruitment (at least on average) than the group sequential design to achieve the desired power (Kjaersgaard et al., 1994). Subgroup analysis, which aims to determine whether the treatment effect differs across one or more patient subgroups, is a common secondary analysis within clinical trials. However, traditional subgroup analysis is prone to spurious false positive or negative results and lacks statistic power (Wagner et al., 2009). Adaptive enrichment design (AED) is a novel approach in clinical trials that aims to improve trial efficiency and accelerate drug development for situations where it is unclear whether the whole patient population will benefit from the new treatment. The design allows for modification of the study protocol during the trial based on accumulating data and aims to enrich the study population with patients who are most likely to benefit from the intervention. In this thesis, we focus on inferring the treatment effect of one targeted experimental treatment. The clinical trial recruits patients from the whole population, however, patients might respond differently to the novel treatment. Patients in certain subgroups might be more likely to have positive response outcomes while others probably have no response at all. Therefore, the aim of this type of enrichment design is to detect the subgroups that respond better to the targeted experimental treatment via interim analyses. In the interim analyses, if the subgroup meets the

eligible criteria, the subgroup will be retained. Otherwise, the design drops the subgroup and there is no further data available on that subgroup. For instance, the diagram in Figure 2.1 illustrates the procedure of the two-stage two-arm (experimental and control) enrichment design with two disjoint subgroups. Recent studies have shown that adaptive enrichment design can increase the efficiency of clinical trials by reducing the number of patients required for a successful trial (Burnett and Jennison, 2021; Lin et al., 2021; Pallmann et al., 2018). Moreover, adaptive enrichment designs can allow for more efficient testing of interventions by incorporating a smaller number of patients in the initial stages of the trial and gradually enrolling additional patients as the study progresses. There is already a large body of research on enrichment designs, such as the approach proposed by Wang et al. (2007, 2009) which considers adaption in sample size and futility stopping in the first interim analysis. Wang et al. (2009) discussed three enrichment design scenarios in their paper: total sample size fixed with futility stopping, total sample sized adapted with futility stopping and total sample size adapted without futility stopping. Their conclusion is that all of the enrichment design scenarios outperform the traditional one-size-fits-all fixed design in terms of the probability of seeking out the responsive subsets. Magnusson and Turnbull (2013) extended this approach to a k-stage design with multiple subpopulations. Rather than allowing only one subgroup to be selected in the first interim analysis, Magnusson and Turnbull's design considers cases in which more than one subgroup treatment effect exceeds the futility threshold and proceeds to subsequent stages. Magnusson and Turnbull (2013) assume that the sampling rule following selection is fixed. In other words, for every possible selection result, the sample size in subsequent stages should be prespecified. Based on Magnusson and Turnbull's approach, Lin et al. (2021) proposed a design involving sample size re-estimation for stage 2 that depends on the observed statistic values in stage 1 to ensure the conditional power is maintained at a desired level. Simon and Simon (2013) proposed an enrichment design that does not pre-specify subgroups but rather updates the eligibility criteria by computing the threshold that maximizes the log-likelihood of the data for the binary outcome.

Using decision-theoretic approaches to find the decision boundaries, sample sizes or both in enrichment designs is on an upward trend recently. Ondra et al. (2019) and Burnett and Jennison (2021) proposed Bayesian optimal rules for subgroup selection that maximize or improve expected utilities at the interim analysis. Rosenblum et al. (2020) use sparse linear programming to optimize the decision rule for subgroup selection and multiple testing procedures. The sample space is divided into a large number of grid rectangles, and then a sparse linear programming problem is solved to find the action associated with each rectangle.

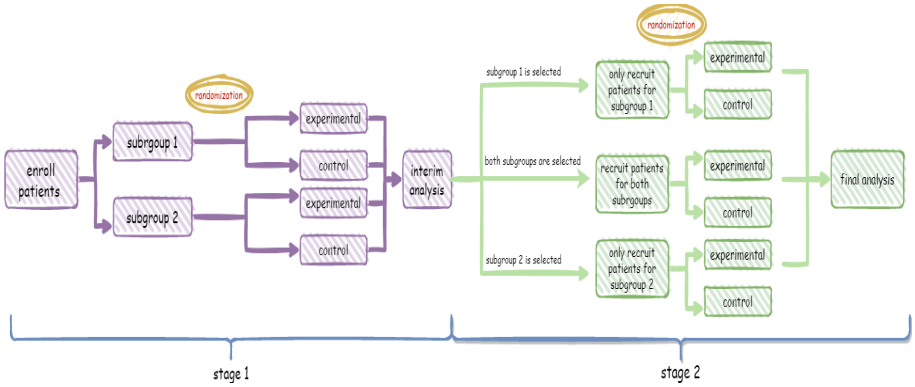


Figure 2.1: The general design procedure for the two-stage two-subgroup adaptive enrichment design.

2.1.1 Group sequential adaptive enrichment design

Heterogeneity frequently presents in patient populations recruited for clinical trials. A specific group of a population might respond positively to the experimental treatment whereas some other group may not. To accommodate this fact, Magnusson and Turnbull (2013) put forward an adaptive design that supports subgroups that respond in the first stage. This design focuses the limited sample size on the subpopulation that has a large possibility to respond and enhances the possibility of acknowledging the efficacy of the target treatment based on score statistics. Magnusson and Turnbull (2013) developed the design in which statistics exit at the previous stage should be more extreme than those terminates at succeeding stages. Additionally, spending error functions are adopted to determine the acceptance and rejection boundaries. The sampling rule of Magnusson and Turnbull (2013) design is prespecified which depends on the selection results.

Basic setups

This section introduces some basic setups of the design. We broadly follow the notation in Magnusson and Turnbull (2013)'s paper but focus on the specific case of a two-stage, two-group trial with normally distributed responses. Assume that the whole patient population Ω_0 can be separated into several disjoint subpopulations $\Omega_1, \Omega_2,$, which means $\Omega_0 = \bigcup_{j=1,2} \Omega_j$. Meanwhile, we denote $\mathcal{P}$ as the index set of the subgroups under consideration, $\mathcal{P} = \{0, 1, 2\}$. Let $\mathcal{S}^*$ be the random variable of the index of the subset that is selected in the interim analysis and $\mathcal{S}$ be the index of the subset that is observed being chosen, $\mathcal{S} \subseteq \mathcal{P}$. Let ρ_j represent the prevalence of each individual subgroup j, and it is assumed that patients are recruited into the first stage of the trial

with probabilities that correspond to these prevalences. For example, suppose that we recruit N_1 patients in total at the first stage of the two-subgroup trial, then the expected number of patients allocated to subgroup 1 and subgroup 2 will be $N_{1,1} = \rho_1 N_1$ and $N_{1,2} = \rho_2 N_1$ respectively. As for N_2, in the Magnusson and Turnbull (2013) design, the total sample size in the second stage is fixed and equals to N_1 but the sample size for each individual subgroup depends on the random specific selection results in the interim analysis. In addition to that, we also consider cases where the sample size in the second stage is different from the sample size in the first stage in our simulation studies. $M_{k,j}$ and $M_{k,\mathcal{S}^*}$ are the accumulated number of patients recruited for individual subgroup j and retained subpopulation $\mathcal{S}$ until stage k, where $k \in \{1, 2, ..., K\}$. Hence $M_{k,j} = \sum_{i=1}^{k} N_{i,j}$ and $M_{k,\mathcal{S}^*} = \sum_{i=1}^{k} N_{i,\mathcal{S}^*}$. We need to point out that the termination stage index K is a random variable in our settings. For instance, if the trial consists of two stages with two subgroups, K could either equal 1 or 2, which implies that the trial terminates at stage 1 or stage 2. Furthermore, suppose that we have two arms (experimental and control) in both subgroups and patients are randomly assigned to each arm with equal possibilities, the expected sample size increment in the experimental arm and control arm in subgroup j at stage k will be $N^E_{k,j} = N^C_{k,j} = N_{k,j}/2$.

Let $Z^E_{k,j,1}, Z^E_{k,j,2}, ..., Z^E_{k,j,N^E_{k,j}}$ be the patients' outcomes for subgroup j at stage k in the experimental arm and let $Z^C_{k,j,1}, Z^C_{k,j,2}, ..., Z^C_{k,j,N^C_{k,j}}$ be the patients' outcomes for subgroup j at stage k in the control arm. Assuming that $Z^E_{k,j,i} \sim N(\mu^E_j, \sigma^2_0)$ and $Z^C_{k,j,i} \sim N(\mu^C_j, \sigma^2_0)$, we define the treatment effect for subgroup j as $\theta_j = \mu^E_j - \mu^C_j$. For the selected subpopulation, we denote the treatment effect as $\theta_{\mathcal{S}^*} = \sum_{j \in \mathcal{S}^*} \rho_j \theta_j$. We construct the accumulated score statistics:

$$Y_{k,j} = (\hat{\mu}^E_{k,j} - \hat{\mu}^C_{k,j}) \times (M_{k,j}/4\sigma^2_0)$$

where $\hat{\mu}^E_{k,j}$ and $\hat{\mu}^C_{k,j}$ are sample means of final outcomes for patients in subgroup j at stage k in the experimental and control arm respectively. σ^2_0 represents the common and known true variance of both $Z^E_{k,j}$ and $Z^C_{k,j}$. Let the $\mathcal{I}_{k,j} = M_{k,j}/4\sigma^2_0$ be the Fisher information, therefore,

$$Y_{k,j} \sim N(\theta_j \mathcal{I}_{k,j}, \mathcal{I}_{k,j})$$

In terms of the above distribution, the combined statistic of the selected subpopulation $\mathcal{S}^*$ is defined as $Y_{k,\mathcal{S}^*} = \sum_{j \in \mathcal{S}^*} Y_{k,j}$. Obviously,

$$Y_{k,\mathcal{S}^*} \sim N(\theta_{\mathcal{S}^*} \mathcal{I}_{k,\mathcal{S}^*}, \mathcal{I}_{k,\mathcal{S}^*})$$

where $\mathcal{I}_{k,\mathcal{S}^*} = \sum_{j \in \mathcal{S}^*} \mathcal{I}_{k,j}$. As statistics accumulate throughout stages in the adaptive enrichment design, we denote $X_{k,j}$ and $X_{k,\mathcal{S}^*}$ as the statistic increment for the individual subgroup and the

chosen subpopulation $\mathcal{S}^*$ at stage k, $k \in \{1, 2, ..., K\}$. Therefore,

$$X_{k,j} = Y_{k,j} - Y_{k-1,j}, \text{ for } k = \{2, ..., K\};$$
$$X_{k,\mathcal{S}^*} = Y_{k,\mathcal{S}^*} - Y_{k-1,\mathcal{S}^*}, \text{ for } k = \{2, ..., K\}.$$

Also, we define $Y_{1,j} = X_{1,j}$ and $Y_{1,\mathcal{S}^*} = X_{1,\mathcal{S}^*}$. Moreover, we denote $\boldsymbol{Y}_j = (Y_{1,j}, ..., Y_{K,j})$ and $\boldsymbol{Y}_{\mathcal{S}^*} = (Y_{1,\mathcal{S}^*}, ..., Y_{K,\mathcal{S}^*})$, where K is the termination stage. Let $\delta_{k,j}$ and $\delta_{k,\mathcal{S}^*}$ be the Fisher information increment. Then $\delta_{k,j}$ and $\delta_{k,\mathcal{S}^*}$ are defined as

$$\delta_{k,j} = \mathcal{I}_{k,j} - \mathcal{I}_{k-1,j} = N_{k,j}/4\sigma_0^2, \text{ for } k = \{2, ..., K\};$$
$$\delta_{k,\mathcal{S}^*} = \mathcal{I}_{k,\mathcal{S}^*} - \mathcal{I}_{k-1,\mathcal{S}^*} = N_{k,\mathcal{S}^*}/4\sigma_0^2, \text{ for } k = \{2, ..., K\}.$$

Similarly, $\delta_{1,j} = \mathcal{I}_{1,j}$ and $\delta_{1,\mathcal{S}^*} = \mathcal{I}_{1,\mathcal{S}^*}$.

If subpopulation $\mathcal{S}^*$ is chosen in the first interim analysis, we denote $H_{0,\mathcal{S}^*} : \theta_{\mathcal{S}^*} = 0$ as the null hypothesis. Meanwhile, we denote $H_{a,\mathcal{S}^*} : \theta_{\mathcal{S}^*} > 0$ as its corresponding one-sided alternative hypothesis for subset $\mathcal{S}^*$. Analogously, the null and alternative hypotheses for an individual subgroup are defined as $H_{0,j} : \theta_j = 0$ and $H_{a,j} : \theta_j > 0$ respectively.

At each stage, k, upper and lower boundaries, u_k and l_k are defined to establish stopping rules for efficacy and futility, respectively. The boundaries are chosen in order to control the family-wise error rate (FWER) specified beforehand. Here we use the same error spending approach proposed by Magnusson and Turnbull (2013) to determine standardized upper and lower boundaries l_k and u_k. In order to ensure that the trial will terminate at the final stage K, we make $l_K = u_K$. We illustrate more details of boundary calculation in the following sections.

Magnusson and Turnbull (2013) also proposed two decision rules to screen out responsive subpopulations in the first interim analysis. We use the first decision rule which assumes no prior ordering in treatment effects. In other words, we assume a prior that all subgroups have the same response to the target experimental treatment. Let $\mathcal{S}^*$ be the index of selected subset, then $\mathcal{S}^* = \{j \in \mathcal{P} : Y_{1,j} > l_1\sqrt{\mathcal{I}_{1,j}}\}$. We drop subgroups with $Y_{1,j} \leq l_1\sqrt{\mathcal{I}_{1,j}}$ and retain the remainder. In this case, the sample size and Fisher information increment for the retained subgroups are defined as $N_{k,\mathcal{S}^*} = \sum_{j\in\mathcal{S}^*} N_{k,j}$ and $\delta_{k,\mathcal{S}^*} = \sum_{j\in\mathcal{S}^*} \delta_{k,j}$. The first stage adjusted upper boundary for the retaining subgroups is defined as

$$\tilde{u}_{1,\mathcal{S}^*} = u_1\sqrt{\mathcal{I}_{1,\mathcal{S}^*}}.$$

Only subgroups in the remaining population will be sampled in succeeding stages. Therefore the

scaled boundaries at stage $k \geq 2$ will be

$$\tilde{l}_{k,\mathcal{S}^*} = l_k\sqrt{\mathcal{I}_{k,\mathcal{S}^*}} \text{ and } \tilde{u}_{k,\mathcal{S}^*} = u_k\sqrt{\mathcal{I}_{k,\mathcal{S}^*}}.$$

After subset $\mathcal{S}^*$ is chosen at the interim analysis, if we observe $Y_{1,\mathcal{S}^*} > \tilde{u}_{1,\mathcal{S}^*}$, we will terminate the trial and declare that the treatment is effective in subpopulation $\mathcal{S}^*$. Otherwise, we proceed the subpopulation $\mathcal{S}^*$ to the next stage. For example, if we focus on the two-stage enrichment design here, when the accumulated statistic $Y_{2,\mathcal{S}^*}$ is greater than $\tilde{u}_{2,\mathcal{S}^*}$, we declare treatment efficacy in subset $\mathcal{S}^*$. If the accumulated statistic $Y_{2,\mathcal{S}^*}$ is smaller than $\tilde{u}_{2,\mathcal{S}^*}$, we declare treatment futility in subset $\mathcal{S}^*$.

Figure 2.2 presents how the design works regarding to the sample space. As shown in Figure 2.2, the entire sample space in the Magnusson and Turnbull enrichment design could be partitioned into seven disjoint subspaces. If the observed paired statistics locate in region 1, the trial terminates at the first stage and we declare futility in both subgroups. When the observed paired statistics fall into region 2 or region 3, one subgroup is enriched in the second stage while the other one stops for futility in the first stage. If the observed statistics belong to region 5 or region 6. one subgroup stops for efficacy while the other one stops for futility at stage 1. Region 4 represents that both subgroups are enriched in the second stage and region 7 represents that both subgroups stop for efficacy at stage 1.

Let $\mathcal{P}^{NR} = \{i \in \mathcal{P} : \theta_i = 0\}$ be the index set of non-responsive population, Magnusson and Turnbull (2013) define the strongly restricted FWER as

$$\begin{aligned} FWER &= \sup_{\boldsymbol{\theta}\in\Theta} \Pr[\text{Reject at least one } H_{\mathcal{S}}, \mathcal{S} \subseteq \mathcal{P}^{NR}] \\ &= \sup_{\boldsymbol{\theta}\in\Theta} \sum_{\mathcal{S}\subseteq\mathcal{P}^{NR}} P_{\boldsymbol{\theta}}(\mathcal{S}^* = \mathcal{S} \text{ and reject } H_{\mathcal{S}} \text{ in subsequently stages}). \end{aligned}$$

We employ the same definition of FWER as proposed by Magnusson and Turnbull (2013), define the probability that the trial ends at stage k, and demonstrate that the treatment is effective for participants in $\Omega_{\mathcal{S}}$ as follows:

$$\begin{aligned} &\psi_{k,\mathcal{S}}(l_1, u_1, ..., l_k, u_k; \boldsymbol{\theta}) \\ =&\Pr[\text{Select } \mathcal{S} \text{ and the trial terminates for effectiveness at stage } k] \\ =&P_{\boldsymbol{\theta}}[\mathcal{S}^* = \mathcal{S}, Y_{1,\mathcal{S}} < \tilde{u}_{1,\mathcal{S}}, Y_{2,\mathcal{S}} \in (\tilde{l}_{2,s}, \tilde{u}_{2,\mathcal{S}}), ..., Y_{k,\mathcal{S}} \geq \tilde{u}_{k,\mathcal{S}}]. \end{aligned} \tag{2.1}$$

Similarly, the probability that the trial terminates at stage k and shows the treatment is ineffec-

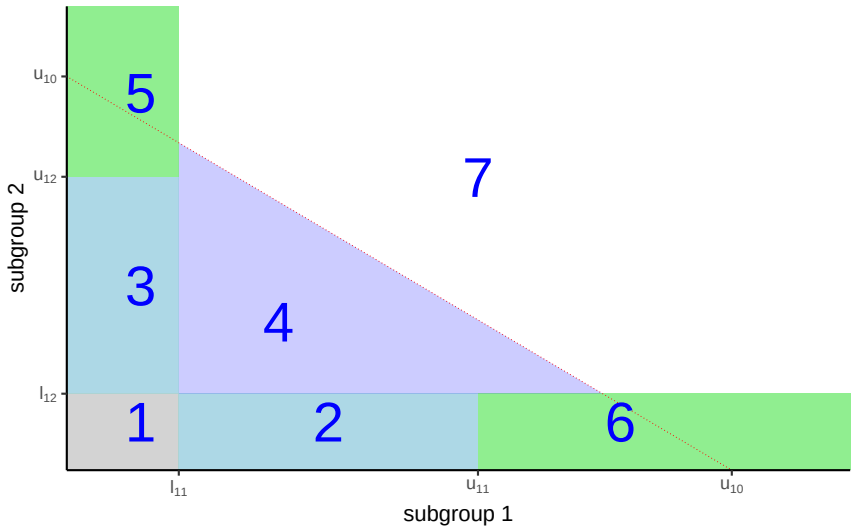


Figure 2.2: Sample space partition of conditional confidence interval construction in the Magnusson and Turnbull design.

tive for subjects in $\Omega_{\mathcal{S}}$ could be written as:

$$\begin{aligned}
&\xi_{k,\mathcal{S}}(l_1, u_1, ..., l_k, u_k; \boldsymbol{\theta}) \\
=&\Pr[\text{Select } \mathcal{S} \text{ and the trial terminates for futility at stage } k] \\
=&P_{\boldsymbol{\theta}}[\mathcal{S}^* = \mathcal{S}, Y_{1,\mathcal{S}} < \tilde{u}_{1,\mathcal{S}}, Y_{2,\mathcal{S}} \in (\tilde{l}_{2,s}, \tilde{u}_{2,\mathcal{S}}), ..., Y_{k,\mathcal{S}} < \tilde{l}_{k,\mathcal{S}}].
\end{aligned} \tag{2.2}$$

As we only retain subpopulations whose statistics exceed their corresponding lower boundary $\tilde{l}_{1,j}$, the probability of selecting $\mathcal{S}$ is

$$\begin{aligned}
\Pr[\mathcal{S}^* = \mathcal{S}] = &\prod_{j \in \mathcal{S}} \left[1 - \Phi \left(\frac{\tilde{l}_{1,j} - \theta_j \times \delta_{1,j}}{\sqrt{\delta_{1,j}}} \right) \right] \\
&\times \prod_{j \notin \mathcal{S}} \Phi \left(\frac{\tilde{l}_{1,j} - \theta_j \times \delta_{1,j}}{\sqrt{\delta_{1,j}}} \right)
\end{aligned} \tag{2.3}$$

where $\Phi(\cdot)$ is the cumulative function for the standard normal distribution. Given Equation (2.1), (2.2) and (2.3), the marginal probability of stopping the trial at stage K due to efficacy and futility can be expressed as

$$\psi_K(l_1, u_1, ..., l_K, u_K; \boldsymbol{\theta}) = \sum_{\mathcal{S} \subseteq \mathcal{P}} \psi_{K,\mathcal{S}}(l_1, u_1, ..., l_K, u_K; \boldsymbol{\theta}) \times \Pr[\mathcal{S}^* = \mathcal{S}] \text{ and}$$

$$\xi_K(l_1, u_1, ..., l_K, u_K; \boldsymbol{\theta}) = \sum_{\mathcal{S} \subseteq \mathcal{P}} \xi_{K,\mathcal{S}}(l_1, u_1, ..., l_K, u_K; \boldsymbol{\theta}) \times \Pr[\mathcal{S}^* = \mathcal{S}].$$

As for termination thresholds, Magnusson and Turnbull (2013) proposed an approach that uses non-decreasing spend error functions specified beforehand to determine upper and lower boundaries at each stage. When $FWER = \alpha$, they define $\gamma_U : [0,1] \to [0,\alpha]$ and $\gamma_L : [0,1] \to [0, 1-\alpha]$ as the upper and lower spending error functions with $\gamma_U[0] = \gamma_L[0] = 0$, $\gamma_U[1] = \alpha$ and $\gamma_L[1] = 1 - \alpha$. Let t_k be the analysis time points that range from 0 to 1, we can search for u_k and l_k that satisfy

$$\psi_k(l_1, u_1, ..., l_k, u_k; \mathbf{0}) = \gamma_U[t_k] - \gamma_U[t_{k-1}] \text{ and}$$
$$\xi_k(l_1, u_1, ..., l_k, u_k; \mathbf{0}) = \gamma_L[t_k] - \gamma_L[t_{k-1}].$$

by plugging in Equation (2.1) and Equation (2.2) under the null hypothesis. In the two-stage trial, we assume that the "error" is equally spent between analysis points. In other words, $\gamma_U[t_2] - \gamma_U[t_1] = \gamma_U[t_1] = \alpha/2$ and $\gamma_L[t_2] - \gamma_L[t_1] = \gamma_L[t_1] = (1-\alpha)/2$.

Next, we establish equations to find the required maximum sample size (i.e. sample size required if the trial proceeds to the second stage) which achieves the power at the desired level. Let θ^* be the clinically effective treatment effect and $1-\beta$ be the power level we desire, $\beta \in (0,1)$. Magnusson and Turnbull (2013) has proposed various power criteria such as:

1. Assuming that the treatment effect is homogeneous across all subgroups, find N_{max} to ensure

   $$\Pr[\text{reject } H_{0,\mathcal{S}}, \mathcal{S} \in \mathcal{P} | \theta_j = \theta^*, \forall j \in \mathcal{P}] = 1 - \beta.$$

   In other words,

   $$\sum_{\mathcal{S} \subseteq \mathcal{P}} \sum_{k=1}^{K} \psi_{k,\mathcal{S}}(l_1, u_1, l_2, u_2; \boldsymbol{\theta}^*) = 1 - \beta.$$

   The above equation means that we could deem at least some subpopulation $\mathcal{S}$ effective with $1 - \beta$ probability.

2. Suppose that $\theta_j = \theta^*$ for $j \in \mathcal{P}^*$ and $\mathcal{P}^* \subseteq \mathcal{S}$. For $j \notin \mathcal{P}^*$, $\theta_j = 0$. Then we define the power as:

   $$\Pr[\text{reject } H_{0,\mathcal{S}} \text{ for } \mathcal{P}^* \subseteq \mathcal{S} \subseteq \mathcal{P} | \theta_j = \theta^*, j \in \mathcal{P}^* \text{ and } \theta_j = 0 \text{ else}]$$

   that is

   $$\sum_{\mathcal{P}^* \subseteq \mathcal{S}^{\subseteq} \mathcal{P}} \sum_{k=1}^{2} \psi_{k,\mathcal{S}}(l_1, u_1, l_2, u_2; \boldsymbol{\theta}^*) = 1 - \beta.$$

This criteria means that given the experimental treatment is only responsive in subpopulation $\mathcal{P}^*$ and $\mathcal{P}^* \subseteq \mathcal{S} \subseteq \mathcal{P}$, the probability with which we declare efficacy in subpopulation $\mathcal{S}$ is guaranteed with $1-\beta$.

Other criteria might be applicable as well. But in our work, we primarily use the first criteria to determine the maximum sample sizes in our simulation studies.

Given $X_{1,j} \sim N(\theta_j \delta_{1,j}, \delta_{1,j})$ and $\tilde{l}_{1,j} = l_1\sqrt{\delta_{1,j}}$ where $\delta_{1,j} = N_{1,j}/4\sigma_0^2$, we denote

$$f_{1,j}(x_{1,j}|\theta_j, \mathcal{S}^* = \{j\}) = \frac{1}{\sqrt{\delta_{1,j}}}\psi(\frac{y_{1,j} - \theta_j\delta_{1,j}}{\sqrt{\delta_{1,j}}}) \left[1 - \Phi\left(\frac{\tilde{l}_{1,j} - \theta_j\delta_{1,j}}{\sqrt{\delta_{1,j}}}\right)\right]^{-1} \times I(x_{1,j} > \tilde{l}_{1,j})$$

the conditional density function of $X_{1,j}|x_{1,j} > \tilde{l}_{1,j}$. $\psi(\cdot)$ is the density function of the standard normal distribution. Moreover,

$$I(x_{1,j} > \tilde{l}_{1,j}) = \begin{cases} 1 & \text{if } x_{1,j} > \tilde{l}_{1,j} \\ 0 & \text{if } x_{1,j} \leq \tilde{l}_{1,j} \end{cases}.$$

In the two-stage two-subgroup design, if both subgroups are chosen at the first interim analysis, it indicates that $y_{1,1} > \tilde{l}_{1,1}$ and $y_{1,2} > \tilde{l}_{1,2}$. Since $y_{1,0} = y_{1,1} + y_{1,2}$, we define the density function for $y_{1,0}$ at the first stage given the overall group is chosen as

$$f_{1,0}(y_{1,0}|\theta_{\mathcal{S}^*}, \mathcal{S}^* = \{0\}) = \int_{\tilde{l}_{1,1}}^{y_{1,0}-\tilde{l}_{1,2}} f_{1,1}(y_{1,1}|\theta_1) f_{1,2}(y_{1,0} - y_{1,1}|\theta_2) \mathrm{d}y_{1,1} \tag{2.4}$$

As for the conditional density function in the second stage, though statistics accumulated when subgroups proceed to the second stage, the density of $y_{2,\mathcal{S}}$ only depends on $y_{1,\mathcal{S}}$ given $\mathcal{S}^* = \mathcal{S}$. Therefore $X_{2,\mathcal{S}}|x_{1,\mathcal{S}} \sim N(\theta_{\mathcal{S}}\delta_{2,\mathcal{S}}, \delta_{2,\mathcal{S}})$ and $X_{2,\mathcal{S}} = Y_{2,\mathcal{S}} - Y_{1,\mathcal{S}}$. Since $\delta_{2,\mathcal{S}} = N_{2,\mathcal{S}}/4\sigma_0^2$, we denote the conditional density function of $Y_{2,\mathcal{S}}|y_{1,\mathcal{S}}$ as

$$f_{2,\mathcal{S}}(y_{2,\mathcal{S}}|y_{1,\mathcal{S}}, \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) = \frac{1}{\sqrt{\delta_{2,\mathcal{S}}}}\psi\left(\frac{(y_{2,\mathcal{S}} - y_{1,\mathcal{S}}) - \theta_{\mathcal{S}}\delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}}\right). \tag{2.5}$$

We notice that in Equation (2.4), the density function for the overall group relies on the individual treatment effect of subgroups (i.e. θ_1 and θ_2); but in Equation (2.5), it depends on the combined treatment effect (i.e. $\theta_{\mathcal{S}}$). The difference of inference subject in density functions might cause inaccuracy of confidence interval construction when the true treatment effect is heterogeneous among subgroups. Recall that we assume that the sampling of subgroups is proportional to the population prevalence. However, if the prevalence of the sampling is different from the real prevalence of the population (though it rarely happens), the inference of the treatment effect would be more inaccurate when $\theta_1 \neq \theta_2$.

2.1.2 Sample size re-estimation adaptive enrichment design

Lin et al. (2021) proposed an enrichment design for two-arm trials that re-estimates the sample size in the second stage to compare a targeted treatment with a standard control treatment. In the Magnusson and Turnbull design, the sample size assigned to each subgroup is determined by the power requirement and the selection results of the interim analysis. We calculate the maximum total sample size by using the criteria we mentioned in Section 2.1.1 and assign an equal number of patients to each stage. The number of patients assigned to each subgroup depends on which subpopulation is enriched in the subsequent stages. In contrast, Lin et al. (2021) proposed that the sample size in the second stage should be determined by the observed statistic values in the first stage rather than merely the screening results in the interim analysis. In their paper, they mainly focus on the clinical trial including two stages and two subgroups, which is the usual case in practice.

Again, let $Z^E_{k,j,i}$ and $Z^C_{k,j,i}$ be the observed outcome of patient i in subgroup j at stage k in the experimental arm and control arm correspondingly. Assume that $Z^E_{k,j,i} \sim N(\mu^E_j, \sigma^2_0)$ and $Z^C_{k,j,i} \sim N(\mu^C_j, \sigma^2_0)$, the treatment effect for subgroup j is defined as $\theta_j = \mu^E_j - \mu^C_j$. Suppose that N_k is the total number of patients recruited at stage k, then $N_{k,j} = \rho_j N_k$ patients will be allocated to subgroup j. Each patient will be assigned to the experimental or control arm with equivalent probability. Based on these setups, they construct the Wald statistics for subgroup 1, subgroup 2, and the overall group in the first stage:

$$
\begin{aligned}
X_{1,1} &= \frac{\hat{\mu}^E_1 - \hat{\mu}^C_1}{\sqrt{4\sigma^2_0/N_{1,1}}}, \\
X_{1,2} &= \frac{\hat{\mu}^E_2 - \hat{\mu}^C_2}{\sqrt{4\sigma^2_0/N_{1,2}}}, \text{ and} \\
X_{1,0} &= \frac{\hat{\mu}^E_0 - \hat{\mu}^C_0}{\sqrt{4\sigma^2_0/N_1}}.
\end{aligned}
$$

where $\hat{\mu}^E_j$ and $\hat{\mu}^C_j$ are sample means of the treatment effects in the experimental arm and the control arm. Obviously, $X_{1,j} \sim N(\theta_j/\sqrt{4\sigma^2/N_{1,j}}, 1)$ for $j = \{1, 2\}$ and $X_{1,0} = \sqrt{\rho_1}X_{1,1} + \sqrt{\rho_2}X_{1,2}$. For the entire population, they denote the treatment effect as $\theta_0 = \rho_1\theta_1 + \rho_2\theta_2$ where ρ_j is the prevalence for subgroup j. In practice and simulation studies, the pooled sample variance σ^2 is used in place of σ^2_0. When the trial proceeds to the second stage, the test statistic for the chosen subpopulation is defined as:

$$
X_{2,\mathcal{S}} = \frac{\hat{\mu}^E_{\mathcal{S}} - \hat{\mu}^C_{\mathcal{S}}}{\sqrt{4\sigma^2/N_2}}
$$

where $\mathcal{S} \in \{0, 1, 2\}$. $\hat{\mu}_{\mathcal{S}}^{E}$ and $\hat{\mu}_{\mathcal{S}}^{C}$ is the sample mean for the experimental treatment arm and control arm within chosen group $\mathcal{S}$. The accumulated statistic from stage 1 and stage 2 is given by

$$y_{2,\mathcal{S}} = \sqrt{\omega} Y_{1,\mathcal{S}} + \sqrt{1-\omega} X_{2,\mathcal{S}}$$

where $\omega = N_{1,\mathcal{S}}/(N_{1,\mathcal{S}} + N_2)$ is the information fraction and $Y_{1,\mathcal{S}} = X_{1,\mathcal{S}}$. Under $\theta_{\mathcal{S}}$, the true treatment effect for the target treatment for selected group $\mathcal{S}$, the stage 2 statistic increment is normally distributed:

$$X_{2,\mathcal{S}} | \theta_{\mathcal{S}} \sim N(\frac{\theta_{\mathcal{S}}}{\sqrt{4\sigma^2/N_{2,\mathcal{S}}}}, 1).$$

Lin et al. (2021) specify a conditional error function $E(\cdot)$ to control FWER, which satisfy

$$\int_{-\infty}^{\infty} E(x_{1,\mathcal{S}}) f(x_{1,\mathcal{S}}) \mathrm{d}x_{1,\mathcal{S}} = \alpha. \tag{2.6}$$

By plugging the circular conditional error function

$$E(x_{1,\mathcal{S}}) = \begin{cases} 0 & \text{if } x_{1,\mathcal{S}} \leq l_1, \\ 1 - \Phi(\sqrt{u_1^2 - x_{1,\mathcal{S}}^2}) & \text{if } l_1 < x_{1,\mathcal{S}} < u_1, \\ 1 & \text{if } x_{1,\mathcal{S}} \geq u_1 \end{cases}$$

proposed by Proschan and Hunsberger (1995) into Equation (2.6), we are able to determine the stopping boundaries l_k and u_k for $k = 1, 2$. As for the sample size in the second stage, it is decided by the conditional power function proposed by Lin et al. (2021):

$$\Pr(Y_{2,\mathcal{S}} > u | y_{1,\mathcal{S}}, \theta_{\mathcal{S}}) = 1 - \Phi(\frac{u - \sqrt{\omega} y_{1,\mathcal{S}}}{\sqrt{1-\omega}} - \frac{\theta_{\mathcal{S}}}{\sqrt{4\sigma^2/N_{2,\mathcal{S}}}}).$$

We can determine the sample size in the second stage that maintains the power at a level of $1-\beta$ by setting $\Pr(Y_{2,\mathcal{S}} > u | y_{1,\mathcal{S}}, \theta_{\mathcal{S}}) = 1 - \beta$.

Let $g_{\max} = \mathrm{argmax}_j(y_{1,j}; j = 1, 2)$ and $g_{\min} = \mathrm{argmin}_j(y_{1,j}; j = 1, 2)$. According to Lin et al. design, the trial will be carried out as below:

1. If $y_{1,g_{\min}} \geq u_1$, stop the trial and declare treatment efficacy in both subgroups.

2. If $y_{1,g_{\max}} \geq u_1$ and $y_{1,g_{\min}} \leq l_1$, stop the trial and declare treatment efficacy in subgroup $g_{\max}$ and treatment futility in subgroup $g_{\min}$.

3. if $y_{1,g_{\max}} \geq u_1$ and $l_1 < y_{1,g_{\min}} < u_1$, stop the trial and declare treatment efficacy in subgroup $g_{\max}$ and inconclusive treatment effect in subgroup $g_{\min}$.

4. if $l_1 < y_{1,g_{\max}} < u_1$ and $y_{1,g_{\min}} \leq l_1$, the trial proceeds to the second stage and we only enrich subgroup $g_{\max}$.

5. if $l_1 < y_{1,g_{\min}} \leq y_{1,g_{\max}} < u_1$, the trial proceeds to the second stage and we enrich both subgroups.

6. if $y_{1,g_{\max}} \leq l_1$, stop the trial and declare treatment futility in both subgroups.

In Figure 2.3, if $(y_{1,1}, y_{1,2})$ falls into region 1, we conclude that both subgroups stop for futility in the first stage. If $(y_{1,1}, y_{1,2})$ locates in regions 2, 3, or 4, it represents that at least one subgroup will be enriched in the second stage. Region 5 and 9 show cases where one subgroup terminates for efficacy while the other one exits for futility. If $(y_{1,1}, y_{1,2})$ belongs to region 7, then we can declare that the experimental treatment is effective for the entire population. Note that in regions 2, 3, and 4, when statistics are close to the lower boundary at stage 1, the sample size required in the second stage is large. By contrast, when statistics approach the upper boundary at stage 1, we are supposed to recruit fewer patients at stage 2 as we believe that the experimental treatment has a promising effect based on the statistics we observed in the first stage.

2.2 Point estimation of treatment effects

Developing an unbiased or consistent point estimator of the treatment effect remains a significant research area because of the impact of treatment or subgroup selection characteristics in adaptive enrichment. As the naive maximum likelihood estimate fails to account for the selection bias in the initial stage, it often yields an overestimation of the actual treatment effect. Consequently, multiple researchers have proposed different unbiased or bias-reduced point estimators to address this issue. For instance, Magnusson and Turnbull (2013) evaluated the conditional and unconditional bias of the naive maximum likelihood estimate of the treatment effect. However, they also note the absence of a perfectly unbiased estimator and suggest utilizing the bootstrap method to reduce bias.

Kimani et al. (2013) proposed two estimators for a two-stage multi-arm enrichment design, where the most effective treatment in the first stage proceeds to the second stage, and any ineffective treatments are dropped at the first stage for futility. One of the estimators is an extension of the uniformly minimum variance conditionally unbiased estimator (UMVCUE) proposed by Cohen and Sackrowitz (1989). However, Cohen and Sackrowitz (1989) assumed that the design would always continue to the second stage, whereas Kimani et al. (2013)'s approach allows for

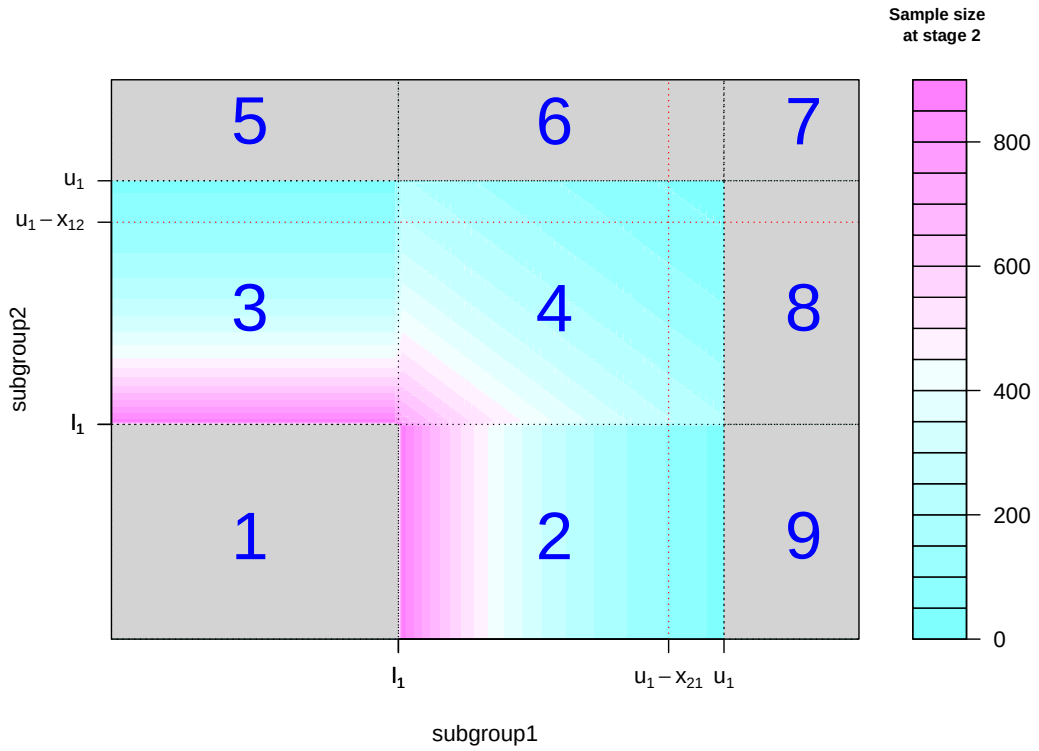


Figure 2.3: Lin et al. design's sample space partition manner. l_1 and u_1 are standardized boundaries. The red dotted lines are the adjusted boundaries for the individual subgroups. In the Lin et al. design, region 1 in Figure 2.3 corresponds to Ω_1^c, region 2 corresponds to Ω_2^c, region 3 corresponds to Ω_3^c, region 4 corresponds to Ω_4^c, region 5 and region 6 corresponds to Ω_6^c, region 8 and region 9 correspond to Ω_5^c, and region 7 corresponds to Ω_7^c.

an early stop in the first stage. The other estimator proposed by Kimani et al. (2013) is the bias-adjusted estimator, which extends the estimator proposed by Stallard and Todd (2005). Kimani et al. (2013) defined the selection time, denoted as t, as the ratio of the sample size in the first stage to the sum of sample sizes in both stages ($n_1/(n_1 + n_2)$, where n_1 and n_2 are the sample sizes at stage 1 and stage 2, respectively). According to their simulation study, when $t > 0.6$, the unbiased estimator and the bias-adjusted estimator performed similarly in terms of mean squared error. However, they noticed that the bias-adjusted estimator was negatively biased, while the unbiased UMVCUE estimator was practically unbiased.

However, the UMVCUE estimator is possibly unavailable for the dropped subgroups as no

unbiased second-stage data is available. Kunzmann et al. (2017) propose hybrid estimators for the two-subgroup two-stage design. In their design settings, they only compare one target experimental treatment with the control treatment. However, the entire population is partitioned into two disjoint subgroups. The decision rule is that either the most promising subgroup or the entire population will proceed to the subsequent stage. In other words, the less responsive subgroup never proceeds alone. They also assume that the patients in each subgroup are equally likely to be assigned to the experimental arm and the control arm, hence each arm has an equal sample size. For the selected subgroup $\mathcal{S}$, the UMVCUE for the true treatment effect $\theta_{\mathcal{S}}$ can be seen as the expectation of the second stage statistic given its statistic value observed in both stages, the interim analysis result, and the statistic value observed for the complementary subgroup. As the UMVCUE is unavailable for the unselected subgroup $\mathcal{S}^{'}$ when only $\mathcal{S}$ is selected, due to the lack of data in the second stage, Kunzmann et al. (2017) propose to use the conditional moment estimator (CME) in this situation. The main idea of the CME is that the conditional expectation of the statistic of the target subgroup $\mathcal{S}$ given interim analysis result and the observed statistic of the complimentary subgroup is a function of the true treatment effect $\theta_{\mathcal{S}}$ and does not depend on $\theta_{\mathcal{S}'}$. Let D be the random selection options and d be the observed selection result. Assuming that $X_{k,\mathcal{S}}$ and $X_{k,\mathcal{S}'}$ to be the target subgroup statistic and its complementary subgroup statistic at stage k, and they define $X_{\mathcal{S}} = \sum_{k\in\{1,2\}} \omega_k X_{k,\mathcal{S}}$ with information fraction ω_k. The conditional moment estimator can be solved from the following equation:

$$\boldsymbol{E}_{\theta_{\mathcal{S}}}[X_{\mathcal{S}}|D = d, X_{1,\mathcal{S}'} = x_{1,\mathcal{S}'}] = x_{\mathcal{S}}$$

where $x_{\mathcal{S}}$ and $x_{k,\mathcal{S}'}$ is the observed value for $X_{\mathcal{S}}$ and $X_{k,\mathcal{S}'}$ respectively. Kunzmann et al. (2017) also compared the naive maximum likelihood estimator, empirical Bayesian estimator, and parametric bootstrap estimator with the hybrid estimator by running simulation studies. Simulation results show that both the empirical Bayesian estimator and parametric bootstrap estimator underestimate the treatment effect for the responsive subgroup and overestimate the treatment effect for the non-responsive subgroup when the true treatment effect of the responsive subgroup is positive. The CME and the hybrid estimator reduce the bias significantly, however, variance and root mean squared error are increased. Additionally, only the CME and the hybrid estimator take the decision rule into account.

2.3 Confidence interval construction

An adaptive enrichment design may involve terminating a trial prematurely if a particular subgroup shows a high likelihood of treatment success. However, this determination is based on interim data, which carries a risk that the estimate of the treatment effect may be unreliable or inaccurate. To better comprehend the uncertainty associated with the estimate of the treatment effect and to make more informed decisions regarding the trial design and whether to proceed with the study, confidence intervals can be established. These intervals help estimate the probable range in which the true treatment effect lies, with a certain level of confidence. The use of point estimates alone neglects the uncertainty and validity of parameter inference, which is why many regulations mandate reporting confidence intervals for all treatment effects in clinical trials. Furthermore, the ICH E9 guideline (1998) requires that "Estimates of treatment effects should be accompanied by confidence intervals, whenever possible, and the way in which these will be calculated should be identified". To address this, numerous studies have focused on developing confidence interval construction for various different types of adaptive designs. One such method is the confidence region approach proposed by Posch et al. (2005) for the flexible group sequential design, which utilizes the close testing procedure to adjust p-values at each stage and combines them using various combination functions. Stallard and Todd (2005) adopt the straightforward p-value inversion approach to construct confidence intervals, however, their design only allows the most effective treatment to be chosen at the interim analysis.

Specifically, in the case of their adaptive enrichment design, Magnusson and Turnbull (2013) suggested using a double bootstrap technique for constructing confidence intervals. This approach commences with the basic maximum likelihood estimators and generates the initial set of bootstrap samples by simulating new datasets assuming the MLE values are correct. Next, they compute the mean of the bootstrap maximum likelihood estimators for each subset. The second set of bootstrap samples is then produced using the bias-corrected simulated bootstrap estimate obtained from the first sample run. Once again, the final estimate is determined by correcting the bias based on the first round of simulated bootstrap estimates. Finally, the $1-\alpha$ confidence interval is formed by using the α quantile of the last simulated bootstrap estimates. However, a simulation study showed that the coverage probability of this method is poor.

2.3.1 Space orderings for p-value function construction

It is complicated to construct confidence intervals via inverting p-value functions in group sequential adaptive designs as the distribution of the test statistic relies on the null and alternative hypotheses of interest (Emerson and Fleming, 1990). Therefore, it is necessary to define the sample space ordering before deriving the p-value functions. In this thesis, we mainly focus on three sample space ordering methods proposed by Armitage (1957), Rosner and Tsiatis (1988), and Emerson and Fleming (1990).

Armitage's method for ordering a sample space involves a step-by-step process where priority decreases as the stages progress. This means that the statistic observed in each subsequent stage should be less extreme than the one observed in the previous stage. To determine which statistic is more extreme, we use the following criteria: if both X_1 and X_2 are observed in the same stage, and X_1 is greater than X_2, then X_1 is more extreme than X_2; if X_1 stops at an earlier stage than X_2 and X_1 crosses the upper boundary in its stage, then X_1 is more extreme; and if X_1 stops at a later stage than X_2 and X_2 crosses the lower boundary in its stage, then X_1 is more extreme. It is evident that this method favours earlier stages over later ones. In the rest of this thesis, we call this method "stage-wise ordering".

In contrast, the ordering method presented by Rosner and Tsiatis (1988) gives equal weight to all stages. To begin with, if we observed standardized statistics, all statistics are transformed into score statistics by multiplying the standardized statistic with the square root of its corresponding Fisher information. The score statistic with the higher value is then deemed the most extreme. Hence, this method is named "score ordering".

Emerson and Fleming (1990) proposed an alternative approach for ordering which relies on the maximum likelihood estimate (MLE) of statistics. This method entails regarding the statistic with a higher MLE value as more extreme, similar to the score ordering method. This method is referred to as MLE ordering or sample mean ordering (Hsu, 1996).

2.3.2 P-value inversion approach

In the p-value inversion method, the goal is to estimate the parameter values of a model that best explains the observed p-value. This is done by using the inverse cumulative distribution function (ICDF) of the test statistic to determine the treatment effect that corresponds to the observed p-value. This critical value is then used to estimate the true treatment effect. Whitehead (1997) has described an approach to construct confidence intervals based on the relationship between

hypothesis testing and confidence intervals. Assuming the parameter to be estimated is denoted by θ, Kimani et al. (2014) summarized the general p-value function based on such relationship as

$$p(\theta, \boldsymbol{x}) = \Pr(\boldsymbol{X} \geq \boldsymbol{x}; \theta) \tag{2.7}$$

where $\boldsymbol{X}$ is the possible data set and $\boldsymbol{x}$ is the observed data set. Whitehead (1997) also described that if the value of $p(\theta, \boldsymbol{x})$ is monotonically increasing on θ and $\theta_\alpha(\boldsymbol{x})$ is defined by $p(\theta_\alpha(\boldsymbol{x}), \boldsymbol{x}) = \alpha$, then

$$\Pr(\theta \leq \theta_\alpha(\boldsymbol{X})) = \alpha \tag{2.8}$$

Equation (2.8) provides a method for obtaining a distinct value of θ for a given data set $\boldsymbol{x}$ with a minimum coverage probability of $1 - \alpha$.

This p-value inversion approach is adopted by Stallard and Todd (2005) in their paper to construct the confidence region for the multiple-arm enrichment design that selects the most effective treatment arm in the first interim analysis. Their p-value function is based on the ordering method proposed by Armitage (1957) and Fairbanks and Madsen (1982) which prioritizes subgroups that stop at the earlier stage for efficacy over those that stop at the later stages. When two subgroups stop at the same stage, the statistic with a greater value is considered more extreme than the one with a smaller value. Stallard and Todd (2005) define the overall p-value function for k treatments as:

$$p(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta}, \boldsymbol{x}) I(\mathcal{S} = j). \tag{2.9}$$

The function $I(\cdot)$ is an indicator function that evaluates to 1 when the jth treatment is chosen and proceeds to the following stages, and evaluates to 0 otherwise. The value of the p-value function follows a standard uniform distribution $U[0, 1]$ for the true value of θ. Therefore, the confidence region for $\boldsymbol{\theta}$ with one-sided coverage probability α is given by $\{\boldsymbol{\theta} : p(\boldsymbol{\theta}, \boldsymbol{x}) \in (-\infty, \alpha)\}$. However, one limitation of this approach is that when we reduce the confidence region for all treatment arms to the confidence interval for the selected subgroup j, we are supposed to assume that the treatment effect for subgroup $j \neq \mathcal{S}$ equals 0 or their maximum likelihood estimate. This leads to inaccurate coverage probabilities for the confidence intervals which are caused by the bias of the estimators.

Stallard and Todd (2005)'s approach focuses on designs in which only the most effective experimental treatment arm is selected. For those designs that allow flexible selection of treatment arms, Magirr et al. (2013) proposed an approach that utilizes the closed testing principle

and p-value combination functions to construct a confidence region for all experimental treatment arms which strongly controls the family-wise error rate at the desired level. Let θ_j be the treatment difference between the experimental treatment j and the control treatment and Q be the combination function that combines p-values at stage 1 and stage 2. The confidence region obtained for all experimental treatments can be reduced to simultaneous confidence intervals. Their algorithm is described below:

1. Perform the closed test procedure.

2. If not all null hypotheses $H_{0,j} : \theta_j \leq 0$ of the experimental treatment arms which proceed to the second stage are rejected at the second stage, then the lower bound for the rejected experimental treatment equals 0 while the lower bound for the accepted experiment treatment should be below 0.

3. Else if all null hypotheses of the selected experimental arms are rejected at the second stage, we compute $p_{\max} = \max_{i \subset (T_1/T_2)} p_i^{(1)}$ where T_1 is the index set of all experimental treatment arms, T_2 is the index set of experimental arms that proceeds to stage 2. When $i = \varnothing$, $p_{\max} = 0$. Then the lower bound for treatment j is defined as

$$l_j = \max\left(0, \sup\{\theta : Q[\max\{p_{\max}, p_j^{(1)}(\theta)\}, p_j^{(2)}(\theta)]\}\right)$$

   where $p_j^{(1)}$ and $p_j^{(2)}$ are adjusted p-values at stage 1 and stage 2.

From the algorithm above, we notice that the lower bounds will only be informative when the selected treatment arms are rejected in the second stage. Also, this design takes no account of the early termination of selected treatment arms in the first stage.

Kimani et al. (2020) developed a method for constructing two-sided confidence intervals for time-to-event data using the confidence region construction method proposed by Magirr et al. (2013). Another feature of the design proposed in their paper is that the subgroup partition is not prespecified but depends on the observed outcomes of patients. The approach involves constructing score statistics and using the closed testing procedure to adjust p-values in each stage. A combination function is then used to merge the adjusted p-values across all stages. Nevertheless, similar to Magirr et al.'s confidence intervals, Kimani et al.'s confidence intervals do not offer information for rejected hypotheses when just a subset of hypotheses are rejected, which potentially contributes to the conservativeness of the confidence region.

In addition to proposing an adaptive design that permits early stopping for efficacy or futility, as well as the addition of a new treatment in the first stage, Posch et al. (2005) also suggest a method for constructing confidence intervals for this flexible selection design. They create a rectangular confidence region that encompasses the simultaneous confidence intervals for two experimental treatments. Assume that k experimental treatments are compared with the control treatment, the hypothesis for the jth treatment is

$$H_{j,0} : \theta_j \leq \mu_j \text{ against } H_{j,a} > \mu_j$$

where θ_j is the true treatment effect for experiment treatment j. Assuming p_j^{adj} and q_j^{adj} are adjusted p-values obtained from the Bonferroni, Simes, or Sidak test for the first and second stage, let a and b be the decision thresholds for the first stage. If p_j^{adj} is less than a, treatment j is stopped due to efficacy, while if p_j^{adj} is greater than b, treatment j is stopped due to futility. For cases where p_j^{adj} is within the range (a, b), treatment j proceeds to the second stage. Likewise, if q_j^{adj} is less than c, treatment j is considered effective, while if q_j^{adj} is greater than or equal to c, treatment j is considered ineffective. Then Posch et al. (2005) define that the one-sided confidence interval for treatment j is

$$I_j = \{\mu_j | \psi_C(p_j^{adj}(\mu_j), q_j^{adj}(\mu_j)) = 0\}$$

where

$$\psi_C(p_j^{adj}, q_j^{adj}) = \begin{cases} 0 & \text{if } p_j^{adj} \leq a \text{ or both } p_j^{adj} \leq b \text{ and } C(p_j^{adj}, q_j^{adj}) \leq c \\ 1 & \text{otherwise} \end{cases}$$

$C(\cdot)$ is the p-value combination function (i.e. $C(p, q) = pq$). From the simulation study, they conclude that the conservativeness of the confidence interval is affected by the p-value adjustment approach used. A limitation of their method is that the simultaneous confidence intervals they construct may not necessarily be compatible with the closed testing procedure. Additionally, these intervals may be inconsistent with the test decision.

2.4 Multiple testing procedures

The control of Type I errors in multiple testing is a topic that is presently of great interest. Various techniques have been proposed by researchers, such as those introduced by Bauer and Kohne (1994), Bauer and Kieser (1999), Posch et al. (2005), and Rosenblum et al. (2016). These methods include the closed testing procedure (Magirr et al., 2013), the p-value combination

method, conditional power (Lin et al., 2021), and combinations of these methods, among others (Burnett and Jennison, 2021; Ondra et al., 2019; Rosenblum et al., 2016, 2020). Magnusson and Turnbull (2013) have developed an approach that is similar to sequential designs, with boundaries in the current stage being dependent on boundaries in all previous stages, but they use the property of the joint distribution of statistics throughout all stages to control the familywise error rate (FWER).

Dynamic programming is utilized by both Ondra et al. (2019) and Rosenblum et al. (2020) to identify the optimal adaption rule under certain constraints. Specifically, in the approach presented by Rosenblum et al. (2020), the goal is to minimize a pre-defined loss function based on observed data, while also ensuring that the FWER is preserved and the Bayesian risk is limited. To achieve this objective, the problem is transformed into a sparse linear programming problem.

The Bonferroni method is regarded as the most straightforward multiple comparison procedure, enabling the simultaneous testing of many hypothesis statements while ensuring that the overall type I error rate remains below a certain level. In earlier research, this method was primarily employed in ANOVA scenarios, where specific sets of pairwise comparisons were pre-selected. According to Hsu (1996), the Bonferroni method is applicable for both equal and unequal sample sizes, and it is founded on Boole's inequality, which is expressed as follows:

$$\Pr(\bigcup_{i=1}^{m} E_i) \leq \sum_{i=1}^{m} \Pr(E_i) \tag{2.10}$$

To elaborate, suppose E_i denotes the event that the i-th confidence level constructed does not encompass the true value. Then, the probability of at least one interval missing its true value is represented by the left-hand side of the inequality. Similarly, the right-hand side of the inequality is the sum of probabilities of each interval failing to capture its true value. Therefore, to restrict the family-wise error rate of multiple interval estimates to α, it is necessary to restrict the type I error rate of each interval to α/m, where m is the number of comparisons or statements. However, the Bonferroni procedure tends to be a bit conservative (Hsu, 1996). To demonstrate this, we can calculate the probability of observing at least one significant result. Let $\alpha = 0.05$. Assuming that there are 15 hypotheses to be tested, and the individual p-values are independent of each other, the probability of observing at least one significant result can be calculated as

follows:

$$\begin{aligned}\Pr(\text{at least one significant result}) &= 1 - \Pr(\text{no significant result}) \\ &= 1 - (1 - \frac{0.05}{15})^{15} \\ &= 0.0485\end{aligned}$$

The probability is less than 0.05, which highlights the conservative nature of this method. To obtain the lower bound of the one-sided Bonferroni confidence interval for all subgroups, we can set the individual p-value function equal to α/m. Although all the resulting Bonferroni confidence intervals are informative, they are conservative as we stated above.

The fundamental concept of the step-down method is to determine whether there is sufficient evidence to suggest that the target treatment, which has the smallest p-value, or, in other words, the treatment that appears to be the most significant, is indeed superior (Holm, 1979). The power of the step-down method is greater than that of the Bonferroni method due to the fact that, given the same assumptions and observation values, the critical value of the Bonferroni method is always greater than or equal to the critical value of the step-down method. As a result, the step-down method tends to reject more hypotheses than the Bonferroni method. However, it is worth noting that the lower bounds of the Bonferroni method are more informative than those of the step-down method, since many of the bounds in the stepwise procedure are zero, as noted by Strassburger and Bretz (2008).

Although rejections using single-step simultaneous confidence intervals such as the Bonferroni method or single-step Dunnet method are typically informative, the simultaneous confidence intervals used in stepwise multiple procedures are often non-informative. Specifically, these confidence intervals only offer information on the parameters of rejected hypotheses in situations where all null hypotheses are rejected. As a consequence, Brannath and Schmidt (2014) have introduced a novel approach known as the weighted Bonferroni procedure, wherein the weight assigned to each hypothesis is contingent on its respective parameter values. By incorporating the penalizing function, this method successfully trades-offs between power and informative rejections. Compared to the Bonferroni procedure, it rejects more null hypotheses, while compared to the Holm, the confidence interval shrinkage implies that Brannath and Schmidt's procedure provides more information in a partial rejection case. Nevertheless, even though this method's algorithm is relatively simple to conduct, it still has its drawback: the choice of the weight function is difficult in practice.

2.5 Conclusion

Having already addressed the shortcomings of point estimates in adaptive enrichment designs, the remainder of this thesis employs the p-value inversion approach to establish both conditional and simultaneous confidence intervals for group sequential designs featuring subgroup selection during interim analysis. Initially, we establish conditional confidence intervals for chosen subgroups in the Magnusson and Turnbull (2013) design, utilizing three sample space ordering methods discussed earlier in this chapter. Subsequently, we generate unconditional intervals for all subgroups by using a similar idea to the CME discussed in Section 2.2 and by implementing multiple testing procedures such as Bonferroni, Holm, and Brannath and Schmidt (2014). Lastly, we extend the construction method to general two-group two-stage enrichment designs and provide an illustration of the generalized approach utilizing the Lin et al. (2021) design.

Chapter 3

Conditional confidence interval for selected subgroups

As we mentioned in the literature review chapter, a large body of literature has investigated the topic of how to construct point estimators in the enrichment design. However, none of them is unbiased. In order to address the uncertainty and bias issue of the parameter inference, many regulation institutions(i.e. FDA and CONSORT) require to include the confidence interval in clinical trial reports. Magnusson and Turnbull (2013) presented a double bootstrap method but the coverage is poor. In this chapter, we suggest a p-value inversion method to establish and verify the coverage probability of the nominal level. We concentrate on the construction of confidence intervals conditional on certain subpopulations being selected at the initial interim analysis. We also concentrate on the case of a two-stage design with two possible subgroups. Nevertheless, as in Section 8, we show our approach can be extended to the three-stage design in generality.

Based on the aforementioned effectiveness exit probability (Equation (2.1)), we can define the probability of observing a statistic greater than u at stage k as:

$$\begin{aligned} &\psi^*_{k,\mathcal{S}}(l_1, u_1, ..., l_k, u_k; \boldsymbol{\theta}, u) \\ =&\Pr[\text{Select } \mathcal{S} \text{ and observing a statistic greater than } u \text{ at stage } k] \\ =&P_{\boldsymbol{\theta}}[\mathcal{S}^* = \mathcal{S}, Y_{1,\mathcal{S}} < \tilde{u}_{1,\mathcal{S}}, Y_{2,\mathcal{S}} \in (\tilde{l}_{2,s}, \tilde{u}_{2,\mathcal{S}}), ..., Y_{k,\mathcal{S}} \geq u]. \end{aligned} \tag{3.1}$$

In this case, if $\mathcal{S}$ is not an empty set, the p-value function which conditions on the event that

$\mathcal{S}^* = \mathcal{S}$ can be written as:

$$\begin{aligned}
&\psi^{\dagger}_{k,\mathcal{S}}(l_1, u_1, ..., l_k, u_k; \boldsymbol{\theta}, u) \\
=&\Pr[\text{observing a statistic greater than } u \text{ at stage } k \text{ conditional on selecting } \mathcal{S}] \\
=&P_{\boldsymbol{\theta}}[Y_{1,\mathcal{S}} < \tilde{u}_{1,\mathcal{S}}, Y_{2,\mathcal{S}} \in (\tilde{l}_{2,s}, \tilde{u}_{2,\mathcal{S}}), ..., Y_{k,\mathcal{S}} \geq u | \mathcal{S}^* = \mathcal{S}].
\end{aligned} \tag{3.2}$$

By incorporating the above basic p-value functions with various space ordering methods, we are able to construct the conditional p-value functions for the corresponding conditional confidence intervals. Throughout the rest of this thesis, the term "conditional" refers to the creation of confidence intervals specifically for the subpopulation that is retained following the interim analysis, based on its selection via the screening rule.

3.1 P-value functions based on specific space ordering methods

As we mentioned in Chapter 2, the p-value function construction is dependent on the space ordering approach (Emerson and Fleming, 1990). Here we first use stage-wise ordering to construct p-value functions which are defined as the p-value of statistics that is observed in the previous stage should be more extreme than whose statistics observed in the afterward stage. In addition to stage-wise space ordering methods, we also introduce p-value functions constructed under the score and MLE ordering methods in this section.

Let I be the stage reached, Y be the statistic, and (i, y) be the pair of observed values of (I, Y) when the trial terminates. Then, we define the probability of observing values as extreme or more extreme than (i, y) under the null hypothesis $H_{0,\mathcal{S}}$ given $\mathcal{S}^* = \mathcal{S}$ as

$$\begin{aligned}
p_{H_{0,\mathcal{S}}} &= \Pr[\text{Observe pair of values as extreme} \\
&\quad \text{or more extreme than } (i, y) | \mathcal{S}^* = \mathcal{S}, \boldsymbol{\theta}] \\
&= \Pr((I, Y) \gg (i, y) | \mathcal{S}^* = \mathcal{S}, \boldsymbol{\theta})
\end{aligned}$$

Note that the definition of "extreme" is different under different space ordering methods. If (i_1, y_1) is ranked higher than (i_2, y_2) given certain space ordering method, we denote $(i_1, y_1) \gg (i_2, y_2)$.

3.1.1 Stage-wise ordering

When we apply the stage-wise space ordering method to the p-value function construction, the stage at which the trial terminates takes precedence. If any of the following three conditions holds:

- $i_2 = i_1$ and $y_2 \geq y_1$
- $i_2 < i_1$ and $y_2 \geq u_{i_2}$
- $i_2 > i_1$ and $y_1 \leq l_{i_1}$

we draw the conclusion that $(i_2, y_2) \gg (i_1, y_1)$. u_{i_2} is the upper bound at stage i_2; l_{i_1} is the lower bound at stage i_1. For instance, consider the test of $H_0 : \theta = 0$ against alternative hypothesis $H_a : \theta > 0$ with 4 interim analyses. We know that $u_1 = 4.6, u_2 = 3.2, u_3 = 2.9$ and $u_4 = 2.1$. If the trial stops at interim 4 with $y_4 = 4$, given stage-wise ordering, then the one-sided p-value is

$$\begin{aligned} &\Pr(\text{observing statistics greater than } y_4) \\ =&\Pr\{Y_1 \geq 4.6 \text{ or } Y_2 \geq 3.2 \text{ or } Y_3 \geq 2.9 \text{ or } Y_4 \geq 4\}. \end{aligned}$$

According to the above definition, we define the p-value function for the two-stage enrichment design that conditions on subgroup $\mathcal{S}$ is chosen as:

$$\begin{aligned} &\Pr(Y_{\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\ =&\Pr(Y_{1,\mathcal{S}} > \max(u, \tilde{u}_{1,\mathcal{S}})|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \times I(\text{trial terminates at stage 1}) \\ +&[\Pr(Y_{1,\mathcal{S}} > \tilde{u}_{1,\mathcal{S}}|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) + \Pr(Y_{2,\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})] \\ &\times I(\text{trial terminates at stage 2}). \end{aligned} \tag{3.3}$$

Recall that $X_{k,\mathcal{S}}$ follows a normal distribution with mean $\theta_{\mathcal{S}}$ and variance $\delta_{k,\mathcal{S}}$, we explicate Equation (3.3) as:

$$\begin{aligned} &\Pr(Y_{\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\ =&\int_{\max(u,\tilde{u}_{1,\mathcal{S}})}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}})\mathrm{d}y_{1,\mathcal{S}} \times I(\text{trial terminates at stage 1}) \\ +&\Big\{\int_{\tilde{u}_{1,\mathcal{S}}}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}})\mathrm{d}y_{1,\mathcal{S}} \\ &+\int_{l_{1,\mathcal{S}}}^{u_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \left[1 - \Phi\left(\frac{(u - y_{1,\mathcal{S}}) - \theta_{\mathcal{S}}\delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}}\right)\right] \mathrm{d}y_{1,\mathcal{S}}\Big\} \\ &\times I(\text{trial terminates at stage 2}). \end{aligned} \tag{3.4}$$

If $\delta_{k,\mathcal{S}} = N_{k,\mathcal{S}}/4\sigma^2$, then

$$f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) = \begin{cases} \frac{\frac{1}{\sqrt{\delta_{1,\mathcal{S}}}}\psi(\frac{y_{1,\mathcal{S}}-\theta_{\mathcal{S}}\delta_{1,\mathcal{S}}}{\sqrt{\delta_{1,\mathcal{S}}}})}{1-\Phi\left(\frac{\tilde{l}_{1,\mathcal{S}}-\theta_{\mathcal{S}}\delta_{1,\mathcal{S}}}{\sqrt{\delta_{1,\mathcal{S}}}}\right)} \times I(x_{1,\mathcal{S}} > \tilde{l}_{1,\mathcal{S}}) & \mathcal{S} \in \{1,2\} \\ \int_{\tilde{l}_{1,1}}^{y_{1,\mathcal{S}}-\tilde{l}_{1,2}} f_{1,1}(y_{1,1}) f_{1,2}(y_{1,\mathcal{S}}|y_{1,1}) \mathrm{d}y_{1,1} & \mathcal{S} = \{0\} \end{cases} . \tag{3.5}$$

$\psi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution functions of the standard normal distribution.

3.1.2 Score ordering

Let y_i be the standardized statistic. According to the score ordering method, if (i_2, y_2) and (i_1, y_1) satisfy

$$y_2\sqrt{\mathcal{I}_2} > y_1\sqrt{\mathcal{I}_1},$$

we declare $(i_2, y_2) \gg (i_1, y_1)$ (Rosner and Tsiatis, 1988). Here, $\mathcal{I}_1$ and $\mathcal{I}_2$ are the cumulative Fisher information corresponding to i_1 and i_2. Under the score ordering method, we only focus on the observed score statistic value, regardless of the stage at which the trial terminates. Because we solely rely on statistical data within our p-value function, it is essential to account for every potential scenario in which the trial concludes. To clarify, the p-value computed based on score ordering encompasses the probability of observing a statistic that exceeds the threshold (denoted as u) throughout all stages of the trial.

We use a similar example as in the stage-wise ordering method where: $u_1\sqrt{\mathcal{I}_1} = 4.6$, $u_2\sqrt{\mathcal{I}_2} = 3.2$, $u_3\sqrt{\mathcal{I}_3} = 2.9$, $u_4\sqrt{\mathcal{I}_4} = 2.1$ and $y_4\sqrt{\mathcal{I}_4} = 4$. Since $y_4\sqrt{\mathcal{I}_4}$ is smaller than $u_1\sqrt{\mathcal{I}_1}$ but greater than $u_2\sqrt{\mathcal{I}_2}$, $u_3\sqrt{\mathcal{I}_3}$ and $u_4\sqrt{\mathcal{I}_4}$, the p-value function could be defined as below based on score ordering method:

$$\begin{aligned} &\Pr(\text{observing statistics greater than } y_4\sqrt{\mathcal{I}_4}) \\ =&\Pr\{Y_1\sqrt{\mathcal{I}_1} \geq 4.6 \text{ or } Y_2\sqrt{\mathcal{I}_2} \geq 4 \text{ or } Y_3\sqrt{\mathcal{I}_3} \geq 4 \text{ or } Y_4\sqrt{\mathcal{I}_4} \geq 4\}. \end{aligned}$$

In our two-stage adaptive enrichment design, we denote the probability of observing an outcome greater than u conditional on selecting $\mathcal{S}$ at the first interim analysis under the score ordering method as:

$$\begin{aligned} &\Pr(Y_{\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\ =&\Pr(Y_{1,\mathcal{S}} > \max(u, \tilde{u}_{1,\mathcal{S}})|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\ +&\Pr(Y_{2,\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \end{aligned} \tag{3.6}$$

By plugging in density functions in Equation (3.6), we explicate the conditional p-value function as:

$$\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\int_{\max(u, \tilde{u}_{1,\mathcal{S}})}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S}} \\
+&\int_{l_{1,\mathcal{S}}}^{u_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \times \left[1 - \Phi\left(\frac{(u - y_{1,\mathcal{S}}) - \theta_{\mathcal{S}}\delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}}\right)\right] \mathrm{d}y_{1,\mathcal{S}}
\end{aligned}$$

where $f_{1|\mathcal{S}}(\cdot)$ is the first stage density function given subpopulation $\mathcal{S}$ is chosen.

3.1.3 MLE ordering

Under the MLE ordering method, outcomes are ordered according to the value of the MLE when the trial terminates. We denote $\hat{\theta}_{MLE} = Y_i/\mathcal{I}_i$, where Y_i is the score scaled statistic and $\mathcal{I}_i$ is the cumulative Fisher information at stage i. If (i_2, y_2) satisfy

$$y_2/\mathcal{I}_2 > y_1/\mathcal{I}_1,$$

then we can conclude that $(i_2, y_2) \gg (i_1, y_1)$ (Emerson and Fleming, 1990).

Note that the definition of MLE ordering involves a transformation of the score statistic. To simplify the calculation, we convert the MLEs back into score statistics after we found their corresponding MLE-scaled statistics. However, for the second converting process, we need to take the possible termination stage into account. If subset $\mathcal{S}$ exits at stage 1, then the converted statistic will be $Y_{1,\mathcal{S},mle} = Y_{\mathcal{S}}/\mathcal{I}_{1,\mathcal{S}} \times \mathcal{I}_{1,\mathcal{S}}$ and $Y_{2,\mathcal{S},mle} = Y_{\mathcal{S}}/\mathcal{I}_{1,\mathcal{S}} \times \mathcal{I}_{2,\mathcal{S}}$ corresponding to stage 1 and stage 2 respectively. If the original statistic $Y_{\mathcal{S}}$ proceeds to stage 2, the converted statistics are $Y_{1,\mathcal{S},mle} = Y_{\mathcal{S}}/\mathcal{I}_{2,\mathcal{S}} \times \mathcal{I}_{1,\mathcal{S}}$ and $Y_{2,\mathcal{S},mle} = Y_{\mathcal{S}}/\mathcal{I}_{2,\mathcal{S}} \times \mathcal{I}_{2,\mathcal{S}}$ corresponding to stage 1 and stage 2 respectively. Analogously, let $c_{1,mle}$ and $c_{2,mle}$ be the corresponding converted observed values at stage 1 and stage 2. Next, we define the conditional p-value function as

$$\begin{aligned}
&\Pr(Y_{mle,\mathcal{S}} > c_{mle} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\Pr(Y_{1,\mathcal{S}} > \max(c_{1,mle}, \tilde{u}_{1,\mathcal{S}}) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
+&\Pr(Y_{2,mle,\mathcal{S}} > c_{2,mle} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})
\end{aligned} \tag{3.7}$$

By plugging in density functions of $Y_{mle,\mathcal{S}}$, Equation (3.7) is equivalent to

$$\begin{aligned}
&\Pr(Y_{mle,\mathcal{S}} > c_{mle} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\int_{\max(c_{1,mle}, \tilde{u}_{1,\mathcal{S}})}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S},mle}|\theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S},mle} \\
+&\int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S},mle}|\theta_{\mathcal{S}}) \left[1 - \Phi\left(\frac{c_{2,mle} - y_{1,\mathcal{S},mle} - \theta_{\mathcal{S}}\delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}}\right)\right] \mathrm{d}y_{1,\mathcal{S},mle}
\end{aligned} \tag{3.8}$$

One property of the converted statistics is that they are more likely to be greater than the rejection threshold in the first stage, therefore the MLE ordering prioritizes the first stage in the same way as the stage-wise ordering approach.

3.1.4 Conditional confidence intervals

Recall that we are testing the one-sided hypothesis $H_{0,\mathcal{S}} : \theta_{\mathcal{S}} = 0$ vs $H_{a,\mathcal{S}} : \theta_{\mathcal{S}} > 0$. All p-value functions we derived above are nondecreasing on θ, thereby we can find certain $\hat{\theta}_{\mathcal{S}}$ that satisfy $\Pr(Y_{\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \hat{\theta}_{\mathcal{S}}) = \alpha$ based on certain ordering manner by using *uniroot* function which implements a line search to find the root assuming there is one within a specified interval. The *uniroot* function adopts the algorithm proposed by Brent (2013), which guarantees a superlinear convergence. Then, the conditional confidence interval for subgroup $\mathcal{S}$ is $(\hat{\theta}_{\mathcal{S}}, \infty)$.

For the two-sided confidence interval, assuming that each side is tested at $\alpha/2$ significance level, we can find a unique solution corresponding to the equation $\Pr(Y_{\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \hat{\theta}_{\alpha/2,\mathcal{S}}) = \alpha/2$ and $\Pr(Y_{\mathcal{S}} > u|\mathcal{S}^* = \mathcal{S}, \hat{\theta}_{1-\alpha/2,\mathcal{S}}) = 1 - \alpha/2$. Therefore we denote the two-sided confidence interval conditional subpopulation $\mathcal{S}$ selected as $(\hat{\theta}_{\alpha/2,\mathcal{S}}, \hat{\theta}_{1-\alpha/2,\mathcal{S}})$.

3.2 Simulation study

In this section, we assess the coverage properties of the proposed conditional confidence intervals by simulating datasets from adaptive enrichment trials following the Magnusson and Turnbull (2013) design. We test the one-sided hypothesis $H_{0,\mathcal{S}} : \theta_{\mathcal{S}} = 0$ at 0.025 significance level. The maximum total sample size, determined to be 1250 based on the first power criteria in Section 2.1.1, allows for maintaining the power at 90%. This is based on the assumption that the clinically effective treatment effect for subgroups 1 and 2 is 0.2 and 0.2, respectively. Suppose that patients are equally allocated in each stage, hence $N_1 = N_2 = 625$. Given the prevalence of subgroup 1 ρ_1 equals 0.6, we randomly generate the sample size of each subgroup by drawing from a binomial distribution. For each space ordering method, we also compare the proportion of trials for which the result is inconsistent between the confidence interval and the design decision. By using the spending error functions in Section 2.1.1, the standardized boundaries are determined as follows:

$$(l_1, u_1) = (0.5192, 2.5529); (l_2, u_2) = (2.4072, 2.4072).$$

We consider two possible situations here. The first one is that only subgroup 1 is chosen in the first interim analysis. In other words, in the simulations, we only retain trials that $Y_{1,1}$ exceeds

its lower adjusted boundary $\tilde{l}_{1,1}$. When the trial proceeds to the second stage, we only recruit patients for subgroup 1. At the end of the second stage, the statistic for subgroup 1 accumulates in the pattern of $X_{2,1} = Y_{2,1} - Y_{1,1}$. The second case is that the entire population is selected, which means both $Y_{1,1}$ and $Y_{1,2}$ are greater than their corresponding lower boundaries. Then, when we make the decision of whether to reject the null hypothesis for the overall group at the first stage, we combine the two statistics as $Y_{1,0} = Y_{1,1}+Y_{1,2}$ and compare it with $\tilde{u}_{1,0} = u_1\sqrt{\mathcal{I}_{1,0}}$. If the trial proceeds to the second stage, we use the same prevalence as stage 1 to assign patients between subgroups. Then the accumulated statistic of the overall group is $X_{2,0} = Y_{2,0} - Y_{1,0}$ where $X_{2,0} \sim N(\theta_0\delta_{2,0}, \delta_{2,0})$. We illustrate a specific trial in Table 3.1. From the definition in Section 2.1.1, we know $\delta_{2,0} = N_{2,0}/4\sigma_0^2$. Throughout our simulation studies in this section and subsequent sections, we assume a common variance σ_0^2 as 1, yet we estimate it using the pooled variance σ^2. Let n_{1j} and n_1 be the observed number of patients recruited for subgroup j and the total number of patients recruited in the first stage respectively. s_{1j}^E and s_{1j}^C are the sample variance of the experimental arm and control arm of subgroup j. The trial's pooled sample variance is calculated from the formula below

$$\sigma^2 = \frac{(n_{11}/2-1)s_{11}^E + (n_{11}/2-1)s_{11}^C + (n_{12}/2-1)s_{12}^E + (n_{12}/2-1)s_{12}^C}{n_1 - 4}$$

which equals 1.0166 in this simulation study. Since $X_{1,1}$ is greater than $\tilde{l}_{1,1}$ and $X_{1,2}$ is smaller than $\tilde{l}_{1,2}$, we only retain subgroup 1 after the first interim analysis. At the second stage, the final statistic equals 59.1081 which is greater than the adjusted second stage upper boundary of 37.1767, hence we reject the null hypothesis and deem that there is a treatment difference between the experimental treatment and the control treatment in subgroup 1.

3.2.1 One-sided conditional confidence interval with equal sample sizes assigned to two stages

Next, we simulate 10,000 runs of trials under three different scenarios and show their coverage probabilities of conditional confidence intervals in Table 3.2. Suppose that $\sigma_0^2 = 1$ in the patient's outcome distribution. Again, in the score statistic distribution $X_{k,j} \sim N(\theta_j\delta_{k,j}, \delta_{k,j})$ where $\delta_{k,j} = N_{k,j}/4\sigma_0^2$, σ_0^2 is estimated by the pooled sample variance σ^2. The three scenarios represent three possible situations. Under the null scenario, the fact is that the target treatment cause no difference from the placebo treatment in the entire population i.e. $\boldsymbol{\theta} = (0, 0)$. In the second scenario, $\boldsymbol{\theta} = (0.2, 0)$ which means that subgroup 1 is more promising compared to subgroup 2. The last scenario $(0.2, 0.2)$ represents that the target treatment is effective for the entire

Table 3.1: Result of one simulated trial described in Section 3.2.

stage 1	$X_{1,j}$	$N_{1,j}$	$\delta_{1,j}$	$\tilde{l}_{1,j}$	$\tilde{u}_{1,j}$
Ω_1	11.5565	370	90.9896	4.9532	24.3521
Ω_2	-7.6191	255	62.7090	4.1120	20.2165
stage 2	$X_{2,j}$	$N_{2,j}$	$\delta_{2,j}$	$\tilde{l}_{2,j}$	$\tilde{u}_{2,j}$
Ω_1	47.5516	625	153.6986	37.6563	37.6563
Ω_2	\	\	\	\	\
Total	$Y_{2,\mathcal{S}}$	$N_{\mathcal{S}}$	$\mathcal{I}_{2,j}$	$\tilde{u}_{1,\mathcal{S}}$	$\tilde{u}_{2,\mathcal{S}}$
Ω_1	59.1081	995	244.6882	24.3521	37.6563

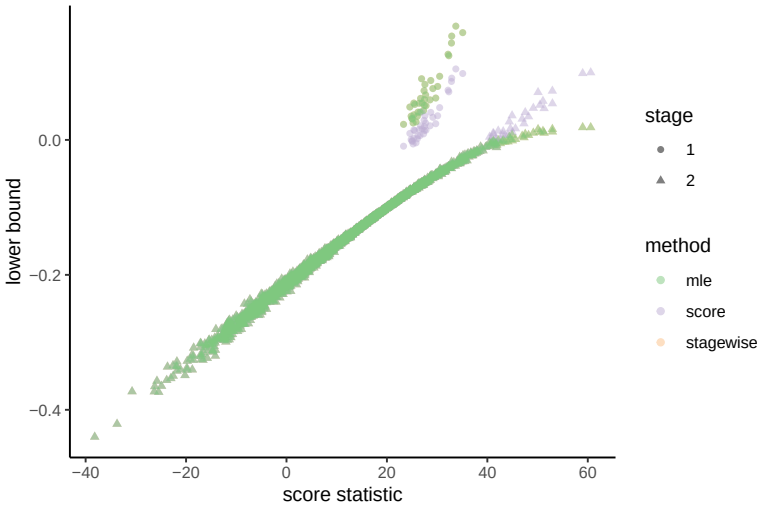
population and treatment effect is homogeneous among them.

Table 3.2: The empirical coverage probability and power of nominal 97.5% one-sided confidence intervals in Magnusson and Turnbull design conditioning on subgroup 1 is selected.

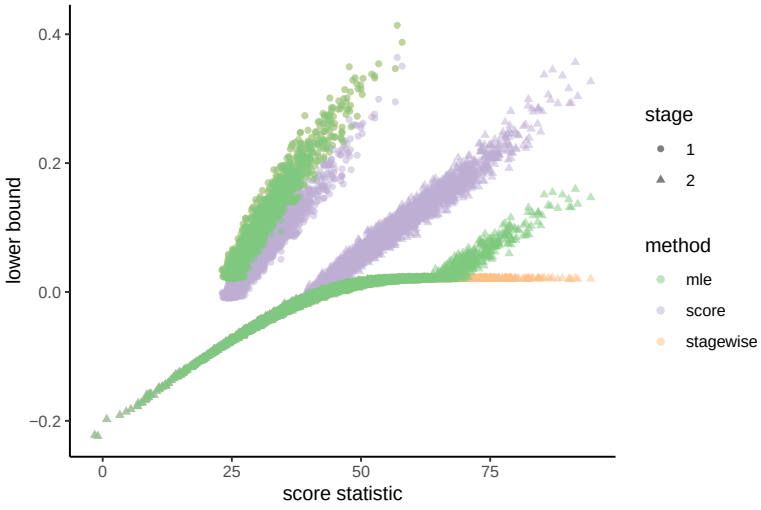
scenario	number of cases where the overall group is selected	coverage probability				power			
		naive	stage-wise	score	MLE	naive	stage-wise	score	MLE
$\theta = (0, 0)$	2074	0.9161	0.9745	0.9711	0.9745	0.0839	0.0255	0.0289	0.0255
$\theta = (0.2, 0)$	6369	0.9677	0.9731	0.9742	0.9731	0.9146	0.7233	0.7365	0.7233
$\theta = (0.2, 0.2)$	1270	0.9583	0.9646	0.9717	0.9646	0.9110	0.7205	0.7276	0.7205

As shown in Table 3.2, confidence intervals constructed under all three space ordering methods have coverage probabilities close to the nominal level for all scenarios. One interesting point is that the stage-wise confidence intervals are close to MLE confidence intervals as both of them prioritize stage 1 over stage 2. When we observe an extreme value, stage-wise ordering prefers to believe that it occurs in the first stage while score ordering does not have such preference. For the MLE method, since the converted statistic is more likely to be extreme at the first stage, it also favors stage 1 over stage 2. Figure 3.1 presents the scatter plots of confidence intervals given subgroup 1 is selected in terms of three different scenarios. Under all scenarios, when the trial terminates at stage 2, the confidence intervals derived from score ordering shrink upward compared to the stage-wise and MLE ordering approach.

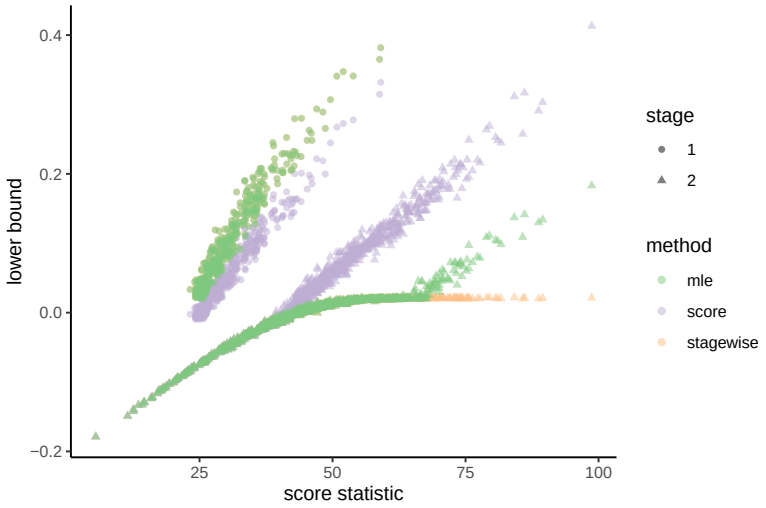
Histograms of lower bounds under different scenarios are shown in Figure 3.2. Each row



(a) $\boldsymbol{\theta} = (0, 0)$



(b) $\boldsymbol{\theta} = (0.2, 0)$



(c) $\boldsymbol{\theta} = (0.2, 0.2)$

Figure 3.1: Lower bounds of one-sided confidence intervals conditioning on subgroup 1 is chosen in the two-stage two-subgroup Magnusson and Turnbull design. The circle and triangle dots indicate that the trials end at stages 1 and 2, respectively. Lower bounds obtained from the MLE, score, and stage-wise ordering approaches are represented by the dots filled in light green, light purple, and light orange, respectively.

displays lower bounds of confidence intervals obtained under scenario $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ respectively given subgroup 1 is chosen. Each column is based on stage-wise, score, and MLE ordering approaches respectively. The red vertical line in every single histogram is the 97.5% quantile. Obviously, they are located around the true treatment effect of subgroup 1. In other words, approximately 2.5% of lower bounds are above the true treatment effect which also indicates that the coverage probability of those confidence intervals nearly reaches the nominal level.

However, one limitation of our p-value inversion approach is that the conclusion obtained from the conditional confidence interval probably does not necessarily agree with the design procedure proposed by Magnusson and Turnbull (2013). Table 3.3 shows the proportion of trials that has different results given that only subgroup 1 is chosen in the interim analysis under three scenarios. We notice that the design procedure rejects more null hypotheses than the confidence interval approach in general and the minimum inconsistency appears in the null scenario. Additionally, the stage-wise ordering method usually draws the same conclusions as the

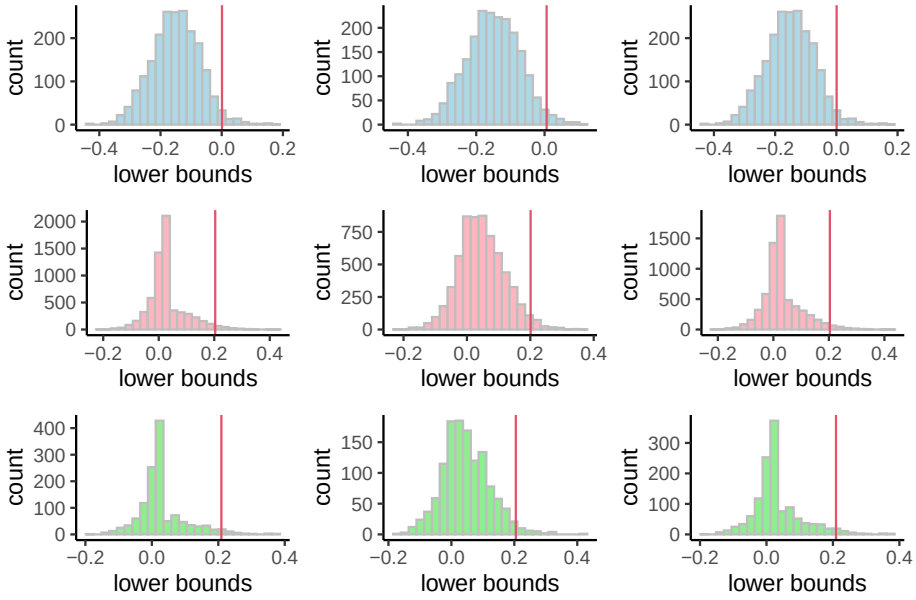


Figure 3.2: Histogram for the lower bounds of nominal one-sided 97.5% confidence intervals given subgroup 1 is selected in the two-stage design.

MLE ordering method. The confidence interval is uniformly more conservative than Magnusson and Turnbull (2013) test. There are several possible explanations for this result. One of those reasons is that these confidence intervals are controlling Type I error conditional on the selection which is stronger than the Magnusson and Turnbull (2013) test which controls the unconditional Type I error. Furthermore, none of the proposed ordering methods directly aligns with the testing boundaries. Therefore, we propose a new ordering approach in Section 3.3 which ensures the alignment between the two-stage upper boundaries by finding the appropriate power of the Fisher information.

There are 932 and 7965 out of 10,000 trials that both subgroups are retained after the interim analysis under scenario $\boldsymbol{\theta} = (0, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ respectively. When both subgroups are selected at the first interim, Table 3.4 shows that under the same treatment effect scenario, the coverage probability is again close to nominal. However, when the treatment effect is heterogeneous across subgroups, the situation becomes complicated. One possible solution is to find the smallest rectangle which contains the contour of the possible pairs of treatment effects. However, obviously, this approach is more conservative compared with the straightforward p-value inversion method. Moreover, in this case, the p-value is based on the joint distribution of the

Table 3.3: The proportion of simulated trials that have different conclusions regarding the design procedure and conditional confidence intervals (CI) under three scenarios.

$\boldsymbol{\theta} = (0, 0)$	proportion of trials rejected by design but accepted by CI
stage-wise	0.0135
score	0.0101
MLE	0.0135
$\boldsymbol{\theta} = (0.2, 0)$	proportion of trials rejected by design but accepted by CI
stage-wise	0.1061
score	0.0930
MLE	0.1061
$\boldsymbol{\theta} = (0.2, 0.2)$	proportion of trials rejected by design but accepted by CI
stage-wise	0.1079
score	0.1007
MLE	0.1078

statistics from all subgroups, which probably is not uniformly distributed as in the univariate case.

Table 3.4: The empirical coverage probability and power of nominal 97.5% one-sided confidence intervals in the Magnusson and Turnbull design conditioning on all subgroups are selected.

scenario	number of cases where the overall group is selected	coverage probability				power			
		naive	stage-wise	score	MLE	naive	stage-wise	score	MLE
$\theta = (0, 0)$	816	0.8155	0.9742	0.9764	0.9763	0.1845	0.0258	0.0236	0.0258
$\theta = (0.2, 0)$	236	0.9079	0.9520	0.9520	0.9520	0.8208	0.2670	0.3599	0.2777
$\theta = (0.2, 0.2)$	3371	0.9702	0.9778	0.9775	0.9778	0.9789	0.4645	0.6645	0.5231

Additionally, we formulated the simple one-sided confidence interval upon concluding the trial, which can be mathematically represented as:

$$\theta_{\mathcal{S}} > (y_{K,\mathcal{S}} - 1.96\sqrt{\mathcal{I}_{K,\mathcal{S}}})/\mathcal{I}_{K,\mathcal{S}}.$$

Upon analyzing the data presented in Table 3.2 and Table 3.4, we can infer that the naive confidence intervals exhibit subpar performance in terms of coverage probability. This deficiency stems from the fact that these intervals fail to account for the intricacies of selection behavior during their construction.

3.2.2 Two-sided conditional confidence interval with equal sample sizes assigned to two stages

In this section, we conduct simulation studies to assess the coverage probability and power of the dual-sided confidence intervals constructed under the score ordering approach. We assume a significance level that is constrained to be at or below 0.025. The outcomes for the three scenarios are detailed in Table 3.5. Coverage probabilities exhibit favorable performance across all scenarios, particularly when retaining only subgroup 1 after the interim analysis. The power of the two-sided confidence intervals for subgroup 1 is slightly lower than that of the one-sided confidence intervals, yet it remains above 55% for scenarios $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$. For the two-sided confidence intervals that account for both selected subgroups, coverage probabilities remain consistently near 97.5% across all scenarios. However, the power is merely 25.63% for the scenario involving a true difference in treatment effects. Nevertheless, this outcome aligns with the findings from the one-sided conditional confidence intervals that are conditioned solely on

the selection of subgroup 1. The construction of the naive two-sided confidence intervals takes the form:

$$(y_{K,\mathcal{S}} - 2.24\sqrt{\mathcal{I}_{K,\mathcal{S}}})/\mathcal{I}_{K,\mathcal{S}} < \theta_{\mathcal{S}} < (y_{K,\mathcal{S}} + 2.24\sqrt{\mathcal{I}_{K,\mathcal{S}}})/\mathcal{I}_{K,\mathcal{S}}$$

where K is the ultimate stage of trial termination. The naive confidence interval disregards any selection process that takes place during the interim analysis. The simulation outcomes for the naive two-sided confidence intervals are also detailed in Table 3.5. We note that their coverage probabilities deviate more significantly from the intended level, yet their statistical power exceeds that of the conditional confidence intervals constructed through the score sample space ordering.

Table 3.5: The coverage probability and power of the conditional two-sided confidence intervals under the score ordering method when the sample sizes assigned to stage 1 and stage 2 are equal. They are compared with the naive two-sided confidence intervals.

scenarios	$S^* = \{1\}$				$S^* = \{1,2\}$			
	coverage probability		power		coverage probability		power	
	score	naive	score	naive	score	naive	score	naive
$\boldsymbol{\theta} = (0,0)$	0.9754	0.9484	0.0246	0.0516	0.9753	0.8777	0.0247	0.1223
$\boldsymbol{\theta} = (0.2,0)$	0.9727	0.9750	0.5966	0.8684	0.9742	0.9459	0.2563	0.7358
$\boldsymbol{\theta} = (0.2,0.2)$	0.9717	0.9701	0.5858	0.8646	0.9749	0.9833	0.5406	0.9584

3.2.3 One-sided conditional confidence interval with unequal sample sizes assigned to two stages

Simulation studies are conducted to evaluate the performance of the conditional confidence interval when varying the number of patients assigned to stage 1 and stage 2. Table 3.6 presents a summary of the coverage probability and power for different scenarios, with one-third and two-thirds of the total patients allocated to stage 1. With one-third of the total sample size allocated to the first stage, the boundaries are set as follows:

$$(l_1, u_1) = (0.1766, 2.6585), (l_2, u_2) = (2.2917, 2.2917).$$

To ensure a 90% power level, a maximum total sample size of 1203 is necessary to detect the minimum significant treatment effect of $\boldsymbol{\theta} = (0.2, 0.2)$. When allocating two-thirds of the total

sample size to the first stage, the boundaries are defined as follows:

$$(l_1, u_1) = (0.8641, 2.4688), (l_2, u_2) = (2.5104, 2.5104).$$

Likewise, in order to achieve a 90% power level for detecting the minimum significant treatment effect of $\boldsymbol{\theta} = (0.2, 0.2)$, a maximum total sample size of 1290 is required. The results indicate that coverage probabilities are consistent across different sample size allocations. Nonetheless, in scenario $\boldsymbol{\theta} = (0.2, 0)$, the power is relatively reduced for the combined overall group. Again, this discrepancy can be attributed to the underlying assumption of a consistent treatment effect in the p-value function.

Table 3.6: The coverage probability and power of the conditional confidence interval when the sample sizes assigned to stage 1 and stage 2 are unequal.

scenario	$S^* = \{1\}$		$S^* = \{1, 2\}$	
$N_1 = 401, N_2 = 802$	coverage probability	power	coverage probability	power
$\boldsymbol{\theta} = (0, 0)$	0.9783	0.0217	0.9761	0.0239
$\boldsymbol{\theta} = (0.2, 0)$	0.9751	0.8359	0.9779	0.4027
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9793	0.8289	0.9746	0.7932
$N_1 = 833, N_2 = 417$	coverage probability	power	coverage probability	power
$\boldsymbol{\theta} = (0, 0)$	0.9785	0.0215	0.9764	0.0236
$\boldsymbol{\theta} = (0.2, 0)$	0.9763	0.7059	0.9572	0.4198
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9719	0.7014	0.9758	0.7000

3.3 Generalized space ordering method

As we mentioned in Section 3.2, we notice an inconsistency between the design procedure decision and our confidence interval conclusion. This is partly due to none of the ordering methods ensuring that a statistic at u_1 in stage 1 is treated as equivalent to a statistic at u_2 in stage 2. Hence, we propose a generalized space ordering method that can ensure there is an alignment between the upper rejection boundaries at stage 1 and stage 2 which is achieved by considering different ways to scale the statistic by a power of the Fisher information.

Let $Z_{k,\mathcal{S}}$ be the standardized statistic for the selected subpopulation $\mathcal{S}$ at stage k and we

denote the adjusted statistic as

$$X^*_{k,\mathcal{S}} = Z_{k,\mathcal{S}} I_k^{\lambda_\mathcal{S}}.$$

where $\lambda_\mathcal{S}$ is a control parameter that can be chosen. Note that I_k will increase between stages 1 and 2 meaning the relative weighting given to stage 1 versus stage 2 is controlled by the value of $\lambda_\mathcal{S}$. When $\lambda_\mathcal{S} = 0$, $X^*_{k,\mathcal{S}}$ is the standardized score, while choosing $\lambda_\mathcal{S} = -0.5$, $X^*_{k,\mathcal{S}}$ is the MLE of θ. Suppose we choose $\lambda_\mathcal{S}$ so that

$$u_1 I_1^{\lambda_\mathcal{S}} = u_2 I_2^{\lambda_\mathcal{S}} \tag{3.9}$$

where u_1 and u_2 are upper boundaries at stage 1 and stage 2 in the Magnusson and Turnbull (2013) design. Equation (3.9) implies that

$$\lambda_\mathcal{S} = \frac{\log u_1 - \log u_2}{\log I_2 - \log I_1}. \tag{3.10}$$

Using this value of $\lambda_\mathcal{S}$ then ensures that the value of $X^*_{k,\mathcal{S}}$ is the same at the stage 1 and stage 2 boundaries. However, as seen in Section 3.2, the conditional confidence intervals had uniformly lower power than the unconditional Magnusson-Turnbull test, and using the proposed ordering does not remedy this issue. Instead, it is possible, in terms of Equation (3.9) and Equation (3.10), and under the null hypothesis, to define the adjusted significance level $\alpha_\mathcal{S}$ as

$$\Pr[X^*_\mathcal{S} > u_1 I^{\lambda_\mathcal{S}} | \theta_\mathcal{S} = 0] = p_\mathcal{S}(u_1; \theta_\mathcal{S} = 0, I_1) = p_\mathcal{S}(u_2; \theta_\mathcal{S} = 0, I_2) = \alpha_\mathcal{S}.$$

Hence, we can say that confidence intervals based on this ordering have the property that a $100(1-\alpha_\mathcal{S})\%$ one-sided confidence interval for $\theta_\mathcal{S}$ conditional on $\mathcal{S}^* = \mathcal{S}$ will be consistent with the one-sided α Magnusson-Turnbull test that defined the boundaries u_1 and u_2. Note, however, that the value of $\lambda_\mathcal{S}$ and also $\alpha_\mathcal{S}$ are themselves dependent on the selected subgroup $\mathcal{S}$ and so it is only possible to get complete consistency between the test and confidence intervals when the α-level for the conditional confidence intervals itself depends on $\mathcal{S}$.

For instance, if $u_1 = 2.5530$, $u_2 = 2.4072$, and the Fisher information for the individual group in the first stage is $I_{11} = 6$ and $I_{12} = 4$ respectively. Therefore, for $\mathcal{S} = \{1\}$, $I_1 = 6$, $I_2 = 16$, and

$$\lambda_{\{1\}} = \frac{\log(2.5530) - \log(2.4072)}{\log(16) - \log(6)} = 0.0600$$

Hence, in this case, the scaling means stage 2 is favoured over stage 1 to a greater degree than either the stage-wise or MLE orderings.

Similarly, first of all, we need to convert the raw statistics at stage 1 and stage 2 into standardized statistics $Z_\mathcal{S}$. If the trial stops at the first stage, the adjusted statistic will be $Z^*_{1,\mathcal{S}} = Z_\mathcal{S} I_1^{\lambda_\mathcal{S}}$;

if the trial terminates at stage 2, the adjusted statistic will be $Z^*_{1,\mathcal{S}} = Z_{\mathcal{S}} I_2^{\lambda_{\mathcal{S}}}$. Then, we need to convert them back to standardized statistics. However, each adjusted score statistic corresponds to more than one converted score statistic due to the different cumulative Fisher information in each stage. Let the converted score statistic be $\tilde{Z}_{1,\mathcal{S}}$ and $\tilde{Z}_{2,\mathcal{S}}$ for stage 1 and stage 2 respectively. Therefore, if the subpopulation $\mathcal{S}$ exits at stage 1,

$$\begin{aligned}\tilde{Z}_{1,\mathcal{S}} &= Z_{\mathcal{S}} \times I_1^{\lambda_{\mathcal{S}}} = Z^*_{1,\mathcal{S}} \\ \tilde{Z}_{2,\mathcal{S}} &= Z_{\mathcal{S}} \times I_2^{\lambda_{\mathcal{S}}} = Z^*_{1,\mathcal{S}} / I_1^{\lambda_{\mathcal{S}}} \times I_2^{\lambda_{\mathcal{S}}};\end{aligned}$$

while if the score statistics proceed to stage 2,

$$\begin{aligned}\tilde{Z}_{1,\mathcal{S}} &= Z_{\mathcal{S}} \times I_1^{\lambda_{\mathcal{S}}} = Z^*_{1,\mathcal{S}} / I_2^{\lambda_{\mathcal{S}}} \times I_1^{\lambda_{\mathcal{S}}} \\ \tilde{Z}_{2,\mathcal{S}} &= Z_{\mathcal{S}} \times I_2^{\lambda_{\mathcal{S}}} = Z^*_{2,\mathcal{S}}.\end{aligned}$$

Let $\zeta_{1,\mathcal{S}}$ and $\zeta_{2,\mathcal{S}}$ be the observed value after converting, then we define the p-value function conditioning on subgroup $\mathcal{S}$ selected as

$$\begin{aligned}\Pr(\tilde{Z}_{\mathcal{S}} > \zeta_{\mathcal{S}} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) = &\int_{\max(\zeta_{1,\mathcal{S}}, \tilde{u}_{1,\mathcal{S}})}^{\infty} f_{1|\mathcal{S}}(\tilde{z}_{1,\mathcal{S}} | \theta_{\mathcal{S}}) \mathrm{d}\tilde{z}_{1,\mathcal{S}} \\ &+ \int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(\tilde{z}_{1,\mathcal{S}} | \theta_{\mathcal{S}}) \left[1 - \Phi \left(\frac{\zeta_{2,\mathcal{S}} - \tilde{z}_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}\tilde{z}_{1,\mathcal{S}}\end{aligned}$$

By plugging in corresponding density functions in the above equation and inverting it based on the adjusted significance level, we obtain conditional confidence intervals that agree with the trial decision. However, it would be difficult to extend this idea to other adaptive enrichment designs.

3.3.1 Numerical study

Here we use similar assumptions and setups as in Section 3.2. Firstly, we calculate the adjusted significance level and Fisher information power corresponding to different selection results given $N_{1,1} = 375$, $N_{1,2} = 250$, and $\sigma^2 = 1$ (shown in Table 3.7). Obviously, the adjusted significance level is greater than the normal significance level ($\alpha = 0.025$). The inflation is due to its dependence on the selection results.

Table 3.8 shows the proportion of inconsistent cases conditioning on the event that only subgroup 1 is selected. The inconsistent rate in the first row is obtained by using the adjusted significance level while the inconsistent rate in the second row is got from the unadjusted significance level (0.025). The inconsistency is completely eliminated when the significant levels

depend on the corresponding selection results. Table 3.9 indicates the same conclusion for the overall group selected cases.

Table 3.7: Adjusted significance level and Fisher information power

Selected population	$\alpha_{\mathcal{S}}$	$\lambda_{\mathcal{S}}$
$\mathcal{S}$={1}	0.0385	0.0600
$\mathcal{S}$={2}	0.0500	0.0391
$\mathcal{S}$={1,2}	0.0974	0.0847

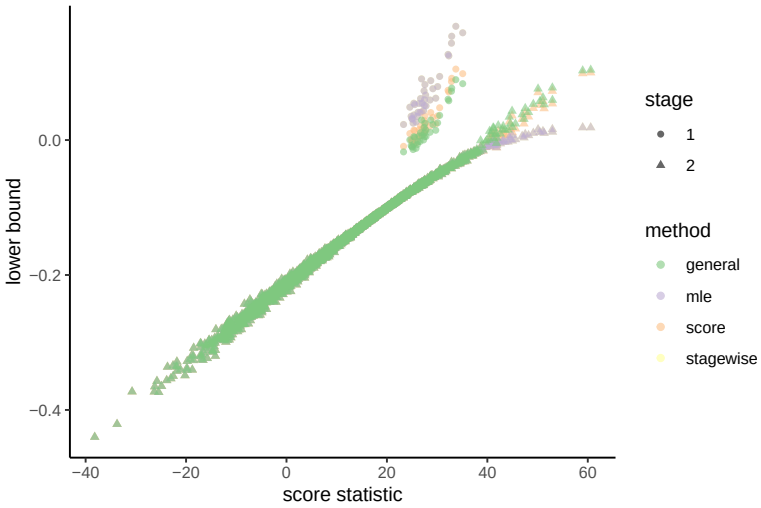
Table 3.8: The proportion of simulated trials that have different conclusions between the design procedure and conditional confidence intervals (only subgroup 1 is selected) using the new ordering method when the true treatment effect is $\boldsymbol{\theta} = (0.2, 0)$.

Scenario: $\theta = (0.2, 0)$	the proportion of trials rejected by the design but accepted by the CI	the proportion of trials rejected by the CI but accepted by the design
new method: $\alpha_{\mathcal{S}} = 0.0385$	0	0
new method: $\alpha = 0.025$	0.0947	0
stage-wise	0.1061	0
score	0.0930	0
MLE	0.1061	0

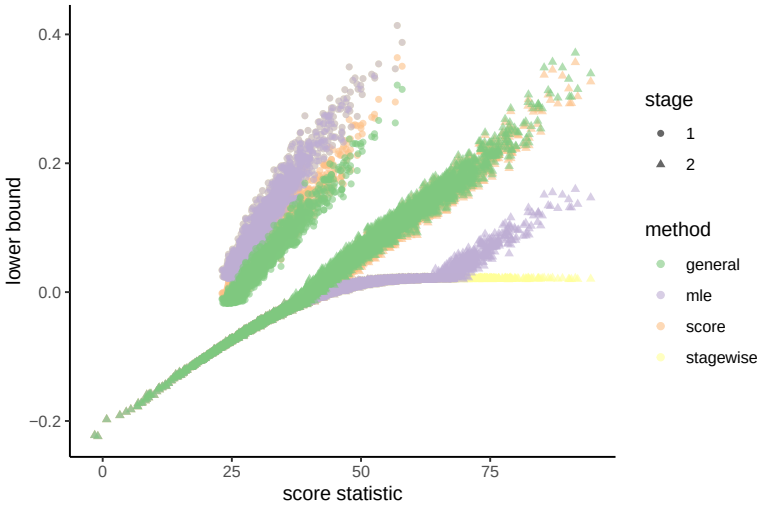
The histogram in Figure 3.4 displays the lower bounds from conditional confidence intervals using the new ordering approach. The histogram filled in light green is the distribution of the lower bounds we obtained when the significance level equals 0.025 while the histogram filled in light yellow is the lower bounds we obtained when the significance level equals $\alpha_{\mathcal{S}}$. Histograms in the upper row present cases that subgroup 1 is selected at the first interim analysis whilst histograms in the lower row present cases that both subgroups are selected at the first interim analysis. The vertical red line is the $1 - \alpha$ quantile. In general, we may claim that the coverage probabilities of both significant level unadjusted and adjusted confidence intervals are close to the nominal level since all of the $1 - \alpha$ quantiles are located around the true treatment effect of 0.2.

Figure 3.3 reveals that despite the positive relationship between the lower bounds of the confidence intervals and the observed final statistic value, the generalized ordering lower bounds obtained from stage 1 and stage 2 are closer to each other in all scenarios compared to stage-wise

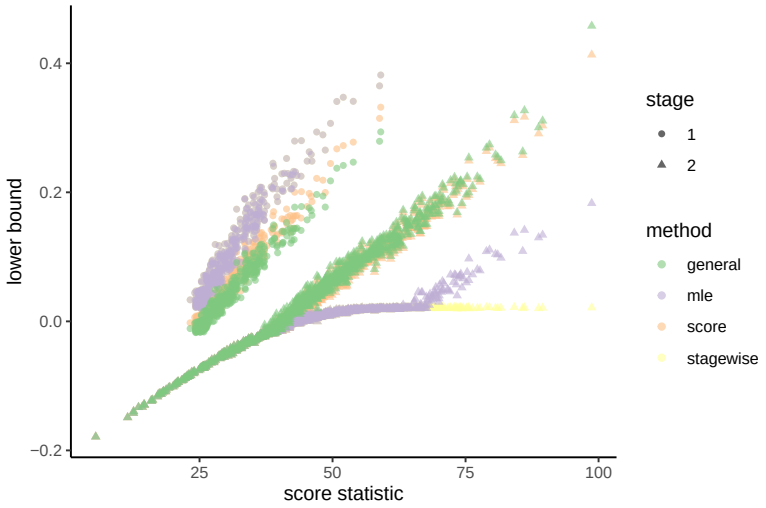
and MLE ordering methods, indicating that the generalized space ordering method favors stage 2 over stage 1 to a greater degree. In other words, the generalized ordering approach treats each stage more evenly.



(a) $\boldsymbol{\theta} = (0, 0)$



(b) $\boldsymbol{\theta} = (0.2, 0)$



(c) $\boldsymbol{\theta} = (0.2, 0.2)$

Figure 3.3: Lower bounds of one-sided confidence intervals conditioning on subgroup 1 is chosen in the two-stage two-subgroup Magnusson and Turnbull design. The circle and triangle dots indicate that the trials end at stages 1 and 2, respectively. Lower bounds obtained from the general, MLE, score, and stage-wise ordering approaches are represented by the dots filled in light green, light purple, light orange, and light yellow respectively.

Table 3.9: The proportion of simulated trials that have different conclusions between the design procedure and conditional confidence intervals (both subgroups are selected) using the new ordering method when the true treatment effect is $\boldsymbol{\theta} = (0.2, 0.2)$.

Scenario: $\theta = (0.2, 0.2)$	the proportion of trials rejected by the design but accepted by the CI	the proportion of trials rejected by the CI but accepted by the design
new method: $\alpha_{\mathcal{S}} = 0.0974$	0	0
new method: $\alpha = 0.025$	0.2841	0
stage-wise	0.4765	0
score	0.2765	0
MLE	0.4178	0

3.4 Conditional confidence intervals in the three-stage Magnusson and Turnbull design

Up to now, we have focused on the two-stage Magnusson and Turnbull (2013) design with two disjoints subgroups. However, the p-value function inversion approach can be extended to a three-

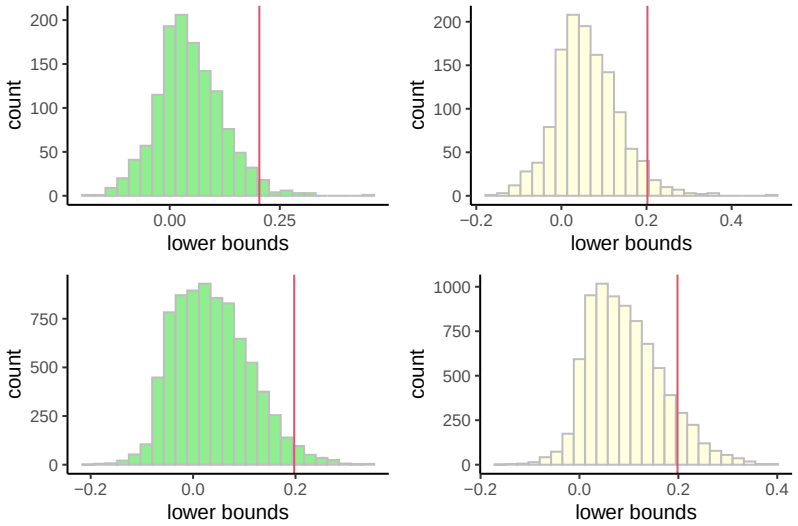


Figure 3.4: Distribution of lower bounds under the scenario $\boldsymbol{\theta} = (0.2, 0.2)$ based on generalized ordering approach. The upper row shows lower bounds for subgroup 1 given only subgroup 1 is chosen. The bottom row displays lower bounds for the overall group given the entire population is selected. The red vertical lines are $1-\alpha$ and $1-\alpha_{\mathcal{S}}$ quantiles.

stage adaptive enrichment design relatively easily. Suppose that the entire population consists of two disjoint subgroups as before, we construct one-sided conditional confidence intervals given that only subgroup 1 is selected and the entire population is chosen. We derive conditional p-value functions based on three different ordering methods as previously. This section solely illustrates the establishment of conditional confidence intervals using the score ordering approach. Appendix A contains the stage-wise and MLE ordering p-value derivation.

According to the definition of the score ordering approach, the conditional p-value function could be defined as

$$\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\Pr(Y_{1,\mathcal{S}} > \max(u, \tilde{u}_{1,\mathcal{S}})) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
+&\Pr(Y_{2,\mathcal{S}} > \max(u, \tilde{u}_{2,\mathcal{S}}) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})] \\
+&\Pr(\min(u, \tilde{l}_{2,\mathcal{S}}) < Y_{2,\mathcal{S}} < \tilde{l}_{2,\mathcal{S}} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})] \\
+&\Pr(Y_{3,\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})].
\end{aligned} \tag{3.11}$$

Hence, the probability of observing a value greater than u can be written as

$$
\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
&= \int_{\max(u, \tilde{u}_{1,\mathcal{S}})}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S}} \\
&+ \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \left[1 - \Phi \left(\frac{\max(u, \tilde{u}_{2,\mathcal{S}}) - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}y_{1,\mathcal{S}} \\
&+ \Big\{ \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \left[1 - \Phi \left(\frac{\min(u, \tilde{l}_{2,\mathcal{S}}) - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}y_{1,\mathcal{S}} \\
&- \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \left[1 - \Phi \left(\frac{\tilde{l}_{2,\mathcal{S}} - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}y_{1,\mathcal{S}} \Big\} \\
&+ \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{2,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) f_{2|\mathcal{S}}(y_{2,\mathcal{S}}|y_{1,\mathcal{S}}, \theta_{\mathcal{S}}) \\
&\times \left[1 - \Phi \left(\frac{u - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{3,\mathcal{S}}}{\sqrt{\delta_{3,\mathcal{S}}}} \right) \right] \mathrm{d}y_{2,\mathcal{S}} \mathrm{d}y_{1,\mathcal{S}}.
\end{aligned}
$$

where $f_{1|\mathcal{S}}(\cdot)$ and $f_{2|\mathcal{S}}(\cdot)$ is same to Equation (3.5) and (2.5) respectively.

By searching for possible $\hat{\theta}_{\mathcal{S}}$ that satisfy $\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \hat{\theta}_{\mathcal{S}}) = \alpha$, we construct the one-sided confidence interval that conditions on subpopulation $\mathcal{S}$ being enriched in succeeding stages as $(\hat{\theta}_{\mathcal{S}}, \infty)$. In principle, the p-value inversion approach could be extended to a multi-stage design with an arbitrary number of stages ($k \geq 3$), but the computation cost will increase, primarily due to the need to numerically evaluate integrals of dimension $k - 1$.

3.4.1 Numerical study

Here we again use the similar settings as we mentioned in Section 3.2. We simulate 10,000 runs of trials for the three-stage Magnusson and Turnbull (2013) design with dividing the entire population into two disjoint subpopulations. The prevalence of subgroup 1 and subgroup 2 are 0.6 and 0.4 respectively. We totally assign 450 patients to each stage and the power is maintained at 90% given $\boldsymbol{\theta} = (0.2, 0.2)$.

Suppose that the significance level is 0.025, we calculate the boundaries in each stage using the spending error functions we mentioned in Section 2.1.1 but extend them to three-stage formulas. The spending error function is defined as:

$$\gamma_U[t_1] = 0.025/3, \gamma_U[t_2] = 2 \times 0.025/3, \gamma_U[t_3] = 0.025;$$

and

$$\gamma_L[t_1] = 0.975/3, \gamma_L[t_2] = 2 \times 0.975/3, \gamma_L[t_3] = 0.975.$$

By recursively solving the above equations, we get the boundaries:

$$(l_1, u_1) = (0.1766, 2.6585), (l_2, u_2) = (0.5867, 2.5635), (l_3, u_3) = (2.4265, 2.4265).$$

Table 3.10 displays the coverage probabilities and powers of three space ordering methods conditioning on two possible selection results. Only subgroup 1 is retained after the first interim analysis in 2467, 5201, and 1141 of 10,000 simulated trials, respectively, under scenarios $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$, and $\boldsymbol{\theta} = (0.2, 0.2)$. The full population is retained after the first interim analysis in 1875, 4130, and 8725 out of 10,000 simulated trials, respectively, under scenarios $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$. Apparently, coverage probabilities of scenario $\boldsymbol{\theta} = (0.0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ are close to the nominal level, which agrees with the conclusion we drew from the two-stage design. However, again, the coverage probability and power of scenario $\boldsymbol{\theta} = (0.2, 0)$ are less satisfying. Each row of Figure 3.5 shows the distribution of lower bounds of conditional confidence intervals obtained under scenarios $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ given subgroup 1 is chosen in the three-stage Magnusson and Turnbull (2013) design, respectively. Histograms in each column present the distribution of lower bounds derived from stage-wise, score, and MLE ordering approaches, respectively. The red vertical line is the 97.5% quantile. Figure 3.5 also indicates that coverage probabilities are guaranteed as around 2.5% of lower bounds in each histogram are greater than the treatment effect. Additionally, we notice that the lower bounds of the stage-wise and MLE ordering approaches are generally above the score ordering lower bounds in the first stage, but the ordering is reversed at stage 2 and stage 3 (see Figure 3.6). It implies that the score ordering treats each stage more evenly, while the other two ordering methods clearly favor stage 1 over succeeding stages. This may also explain why the score ordering method has greater power than the stage-wise and MLE ordering methods across all scenarios in Table 3.10.

3.5 Conclusion

In this chapter, we concentrated on deriving p-value functions conditional on the first interim analysis decision. Taking cases where a single subgroup is selected into consideration or the complete population is enrolled, by inverting its corresponding p-value functions, we showed

Table 3.10: The coverage probability and power of conditional confidence intervals for three-stage Magnusson and Turnbull design.

selection result: only subgroup 1 is chosen						
scenario	coverage probability			power		
	stage-wise	score	MLE	stage-wise	score	MLE
$\boldsymbol{\theta} = (0, 0)$	0.9732	0.9728	0.9732	0.0268	0.0272	0.0268
$\boldsymbol{\theta} = (0.2, 0)$	0.9742	0.9706	0.9746	0.8381	0.8391	0.8379
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9754	0.9702	0.9754	0.8755	0.8773	0.8755
selection result: subgroup 1 and subgroup 2 are chosen						
scenario	coverage probability			power		
	stage-wise	score	MLE	stage-wise	score	MLE
$\boldsymbol{\theta} = (0, 0)$	0.9743	0.9754	0.9743	0.0257	0.0246	0.0257
$\boldsymbol{\theta} = (0.2, 0)$	0.9683	0.9678	0.9683	0.3019	0.4058	0.3092
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9753	0.9719	0.9753	0.6286	0.7554	0.6579

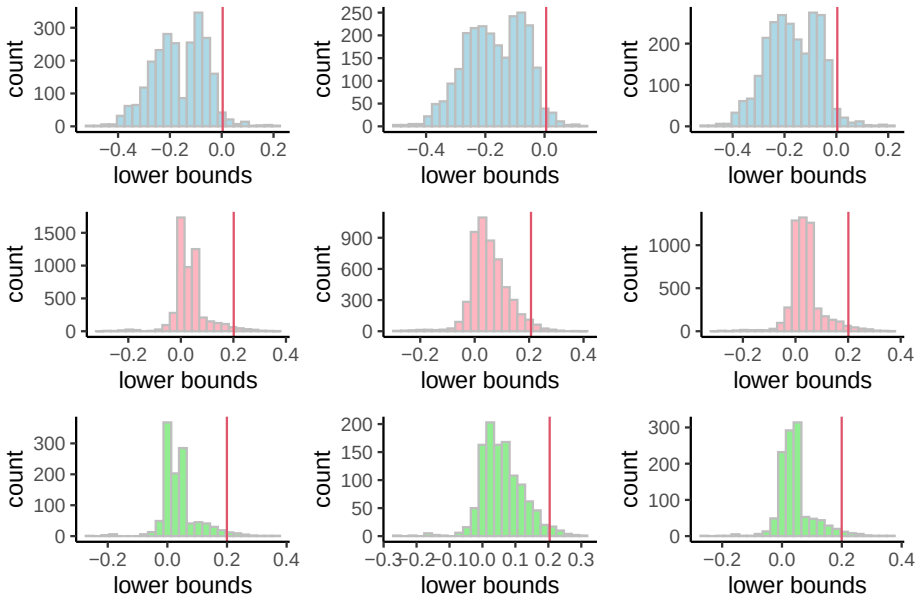
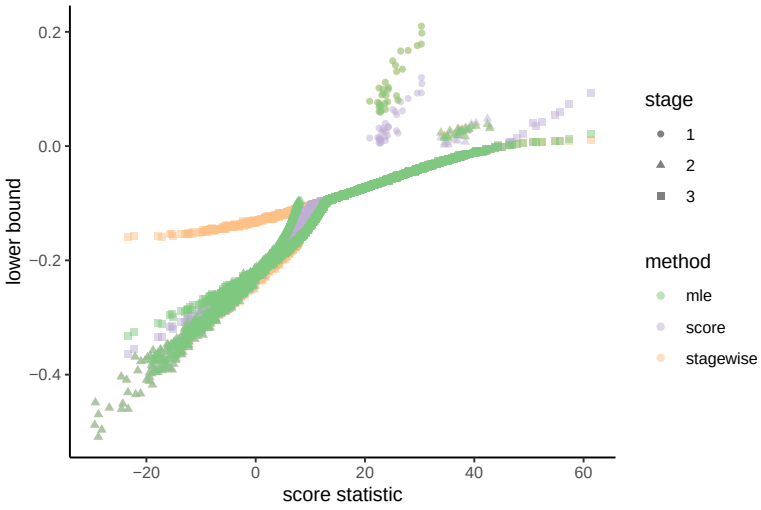
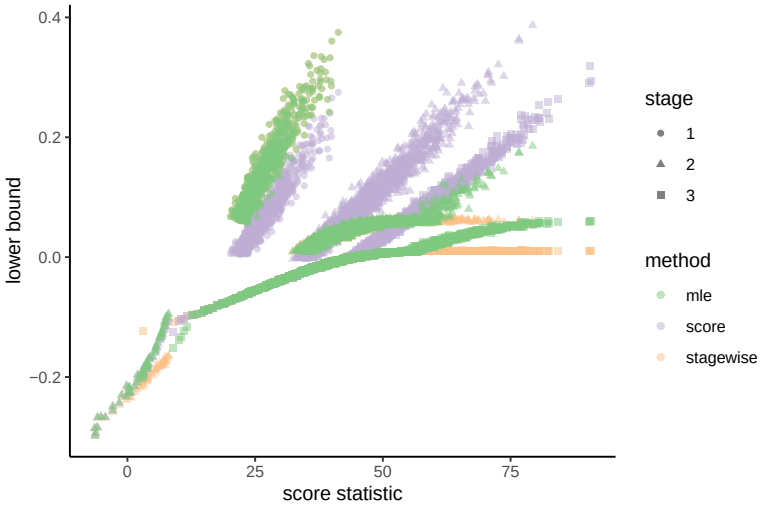


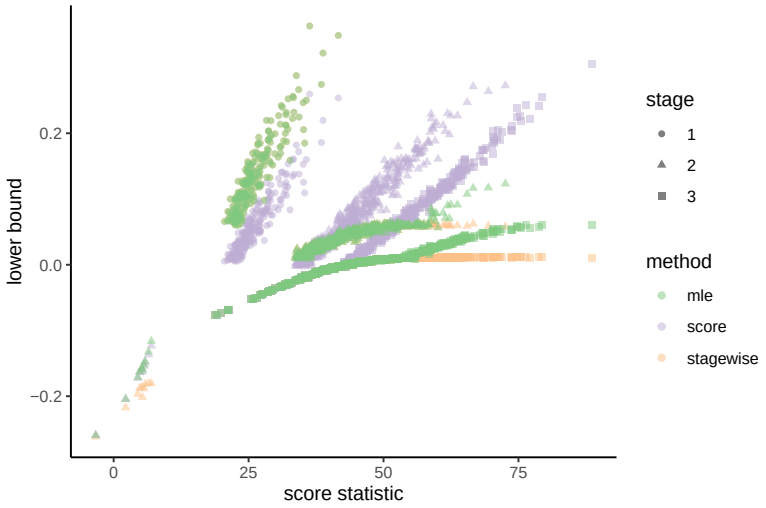
Figure 3.5: Distribution of the confidence interval lower bounds given subgroup 1 is chosen in the three-stage Magnusson and Turnbull design. The red vertical line is the 97.5% quantile.



(a) $\boldsymbol{\theta} = (0, 0)$



(b) $\boldsymbol{\theta} = (0.2, 0)$



(c) $\boldsymbol{\theta} = (0.2, 0.2)$

Figure 3.6: The lower bounds of one-sided confidence intervals conditional on subgroup 1 are chosen in the three-stage two-subgroup Magnusson and Turnbull design. The circle, triangle, and square dots indicate that the trials end at stages 1, 2, and 3 respectively. Lower bounds obtained from the MLE, score, and stage-wise ordering approaches are represented by the dots filled in light green, light purple, and light orange, respectively.

that conditional confidence intervals can be obtained that have coverage probabilities close to the nominal level. As the construction of the p-value function relies on the space ordering criterion, three commonly utilized ordering approaches were compared: stage-wise, score, and MLE. Despite the fact that each of them provides intervals with good coverage probabilities, the score ordering dominates the other two methods. This is because the stagewise and MLE ordering usually prefers situations where the trial terminates at an earlier stage as compared to succeeding stages, unlike the score ordering which treats every stage approximately equally.

One limitation common to all three ordering methods is the inconsistency between the confidence interval conclusion (i.e. based on whether the interval contains 0) and the decision implied by the testing procedure proposed in the design. Hence, we proposed a generalized ordering approach to reduce the disagreement. While it is not possible to give a conditional 97.5% confidence interval that agrees with the 2.5% Magnusson and Turnbull (2013) test result, it is possible to construct a (1-$\alpha_{\mathcal{S}}$)100% confidence interval that will agree with the test result. Finally, we showed that our p-value inversion method that can be extended to a multiple-stage

design ($k \geq 3$). Simultaneously, the computation cost would increase as well.

When treatment effects are diverse across subgroups, the inference subject differs between the initial and subsequent stages. At the initial stage, we construct density functions based on the individual treatment effect; however, in the following stages, we combine the treatment effect for the selected subgroups. The true treatment effect would then be a weighted average of θ_1 and θ_2. One potential solution is to construct a joint p-value function for θ_1, θ_2 and find the smallest rectangle that encloses the non-linear interval given by the following equation:

$$\begin{aligned} &\{(\theta_1, \theta_2) : \Pr[Y_1 > y_1, Y_2 > y_2 | \theta_1, \theta_2] \geq \alpha\} \\ =&\{(\theta_1, \theta_2) : \bigcup_{\Omega_i^{\subseteq} \Omega_0} \iint_{\Omega_i \cap\ \Omega_{obs}} f(y_{1,1}, y_{1,2}) \mathrm{d}y_{2,1} \mathrm{d}y_{1,1} \geq \alpha\} \end{aligned} \tag{3.12}$$

where Ω_{obs}, Ω_j, and Ω_0 is the sample space for the observed value, the statistic of individual subgroup, and the entire population respectively. $f(\cdot)$ is the joint p-value density function of $y_{1,1}$ and $y_{1,2}$. However, this method is excessively conservative because the joint p-value is not generally uniformly distributed given the true treatment effect. An alternative way to address this issue is to find the critical value $c_\alpha(\theta_1, \theta_2)$ that satisfies $\Pr(p_{\boldsymbol{\theta}}(Y_1 > y_1, Y_2 > y_2 | \theta_1, \theta_2) > c_\alpha(\theta_1, \theta_2)) = \alpha$. The confidence region will encompass every potential pair of (θ_1, θ_2) such that $p_{\boldsymbol{\theta}}(Y_1 > y_1, Y_2 > y_2 | \theta_1, \theta_2) > c_\alpha(\theta_1, \theta_2)$. This approach is, however, computationally demanding.

Chapter 4

Unconditional confidence interval for an individual subgroup

4.1 Introduction

So far this thesis has focused on constructing confidence intervals given a certain subpopulation has been chosen in the first interim analysis. Due to the multiple subgroups involved in the design, we need to test more than one null hypothesis (i.e. $H_{0,j} : \theta_j = 0, j = \{1, 2, .., m\}$) at the same time and construct simultaneous confidence intervals for all subgroups. We will again use the straightforward p-value inversion approach here, hence, first of all, we derive the unconditional individual p-value function for a single subgroup. The term "unconditional" here is used to refer to the confidence intervals for the individual subgroup irrespective of the selection result in the first interim analysis. However, in general, the distribution of the score statistic for a given subgroup will depend on the whole vector of treatment effects. To circumvent this issue, we will therefore work with the conditional distribution of the score statistic for a given subgroup given the stage 1 score statistics for the other groups. The confidence interval construction approach we propose in the following sections is similar to the main idea of the conditional moment estimator we mentioned in Chapter 2, which is the expectation of $X_{\mathcal{S}}$ conditioning on subgroup $\mathcal{S}$ chosen at the interim analysis and the observed statistic of the dropped subgroup $X_{\mathcal{S}'}$ is a function of the treatment effect $\theta_{\mathcal{S}}$ and does not depend on $\theta_{S'}$ (Kunzmann et al., 2017). Here, our p-value for the individual subgroup j will be only conditional on observed statistics of other subgroups at stage 1 instead of θ_j. Therefore, we can find the bounds of the confidence intervals for the

true treatment by inverting the corresponding p-value function. Assuming that the data from every disjoint subgroup is independent of each other (usually, it is also the case in practice), we adopt the classic Bonferroni, Bonferroni-Holm and Brannath and Schmidt procedure to construct simultaneous confidence intervals for the two-stage Magnusson and Turnbull (2013) enrichment design with two disjoint subgroups and illustrate their simulation study results. In the interest of comparing these three simultaneous confidence interval construction procedures, we extend the simulation study to the three-subgroup case in Section 4.2.3.

4.2 Simultaneous confidence intervals in the two-stage Magnusson and Turnbull design

In Chapter 3, we constructed confidence intervals given certain subpopulation is chosen regarding the corresponding conditional p-value function. Similarly, before constructing an unconditional confidence interval for an individual group, it is necessary to derive its corresponding unconditional p-value function in the first place. Based on the individual p-value function, in Section 4.2.2, we construct simultaneous confidence intervals based on the classic Bonferroni procedure, Bonferroni-Holm stepdown procedure, and a new weighted Bonferroni approach proposed by Brannath and Schmidt (2014).

4.2.1 P-value function and the worked example of Magnusson and Turnbull design

In this section, we first introduce the p-value function for an individual subgroup which only conditions on the observed statistics of remaining subgroups in the selected subpopulation. Then, we provide a worked example for the Magnusson and Turnbull (2013) enrichment design, with individual rejection bounds calculated in the meanwhile.

p-value function

We again focus on the two-arm two-stage Magnusson and Turnbull (2013) design with two disjoint subgroups and use similar setups as in conditional confidence interval construction in Chapter 3. Recall that in subgroup j, the outcome of the ith patient in both the experimental and control arm follows the normal distribution with mean μ^E and μ^C respectively, and a common variance σ^2. Hence the treatment effect of subgroup j is defined as $\theta_j = \mu_j^E - \mu_j^C$ where μ_j^E and μ_j^C

are the treatment effects in the experimental arm and control arm. Given that we recruit the same number of patients for each stage (i.e. $N_1 = N_2 =, ..., = N_K, k = \{1, 2, ..., K\}$) and the prevalence of subgroup j is ρ_j, the number of patients allocated to subgroup j at stage k would be $N_{k,j} = \rho_j N_k$. We assign an equal number of patients to the experimental and the control arm. Therefore, the score statistic of subgroup j at stage k is distributed as follows:

$$X_{k,j} \sim N(\theta_j \delta_{k,j}, \delta_{k,j}).$$

where $\delta_{k,j} = N_{k,j}/4\sigma_0^2$. We define that $Y_{1,j} = X_{1,j}$ and the score statistic is accumulated in the second stage, the statistic for individual subgroup j at stage 2 can be written as $Y_{2,j} = Y_{1,j} + X_{2,j}$. According to the above definition, we can write the first stage density function for subgroup j as

$$f_{1,j}(y_{1,j}|\theta_j) = \frac{1}{\sqrt{\delta_{1,j}}} \psi \left(\frac{y_{1,j} - \theta_j \delta_{1,j}}{\sqrt{\delta_{1,j}}} \right)$$

where $\psi(\cdot)$ is the density function for the standardized normal distribution. Since the second stage accumulated statistic $y_{2,j}$ only depends on $y_{1,j}$, the density function at stage 2 is explicated as

$$f_{2,j}(y_{2,j}|y_{1,j}, \theta_j) = \frac{1}{\sqrt{\delta_{2,j}}} \psi \left(\frac{y_{2,j} - y_{1,j} - \theta_j \delta_{2,j}}{\sqrt{\delta_{2,j}}} \right).$$

We notice that the density function for the conditional and unconditional p-value functions are different in the first stage but are the same in the second stage.

Despite the first stage density function, another notable difference is the upper boundary in the first stage. As we are deriving the p-value function for an individual subgroup, we are supposed to find the individual boundaries for the target subgroup based on different selection results. Here we use the approach similar to the conditional moment estimator proposed by Kunzmann et al. (2017). The fundamental idea of Kunzmann et al. (2017)'s point estimator is that the conditional expectation of Y_j given the interim analysis result and observed statistics of other subgroups is a function in the true treatment effect. In our case, rather than seeking the expectation of statistics, our focus shifts towards investigating the probability of encountering a statistic exceeding a predetermined threshold. Drawing inspiration from the notion of the conditional moment estimator, instead of conditioning on the true treatment effects, we derive the individual p-value function for subgroup j based on the observed outcomes of the rest chosen subgroups. As the adjusted upper boundaries combined all the information from all the chosen subgroups by adding them up in the Magnusson and Turnbull (2013) design, we need to segregate information for the target subgroup. This is accomplished by subtracting the observed outcomes

of the remaining selected subgroups from the boundaries scaled based on selection results. For instance, if the overall group is chosen, the upper boundary for subgroup 1 would be $\tilde{u}_{1,0} - y_{1,2}$ where $\tilde{u}_{1,0}$ is the adjusted boundary for the entire population.

In Chapter 3, we introduced three ordering methods to construct confidence intervals given a specific subpopulation is selected, and the simulation results showed that the score ordering is superior to other ordering methods. Hence we choose to use the score ordering to construct the p-value function for the individual subgroup here. The probability of observing a statistic greater than u could be described as

$$\begin{aligned}\Pr(Y_j \geq u|y_{1,\ell}, \theta_j) = &\int_{\max(\max(u,\tilde{l}_{1,j}),\tilde{u}_{1,\mathcal{S}}-y_{1,\ell})}^{\infty} \frac{1}{\sqrt{\delta_{1,j}}}\psi\left(\frac{y_{1,j}-\theta_j\delta_{1,j}}{\sqrt{\delta_{1,j}}}\right)\mathrm{d}y_{1,j} \\ &+ \int_{\tilde{l}_{1,j}}^{\tilde{u}_{1,\mathcal{S}}-y_{1,\ell}} \frac{1}{\sqrt{\delta_{1,j}}}\psi\left(\frac{y_{1,j}-\theta_j\delta_{1,j}}{\sqrt{\delta_{1,j}}}\right) \\ &\times\left[1-\Phi\left(\frac{u-y_{1,j}-\theta_j\delta_{2,j}}{\sqrt{\delta_{2,j}}}\right)\right]\mathrm{d}y_{1,j} \times I(u \geq \tilde{l}_{1,j})\end{aligned} \tag{4.1}$$

where $\{\ell : \forall j \notin \mathcal{S}\}$ and $I(\cdot)$ is the indicator function. Assume that α is the significance level, we search for possible $\hat{\theta}_j$ that satisfy $\Pr(Y_j \geq u|y_{1,\ell}, \hat{\theta}_j) = \alpha$. Let $\Pr(Y_j \geq u|y_{1,\ell}, \theta_j) = p_j(\theta_j)$, we assume that $p_j(\theta_j)$ satisfied the following property: $p_j(\theta_j)$ is a non-decreasing function on $\theta_j \in \mathbb{R}$, and $p_j(\theta_j) \in (0, 1)$. Hence $CI = (\hat{\theta}_j, \infty)$ is the unconditional one-sided $1-\alpha$ confidence interval for subgroup j. If we are supposed to test a two-sided hypothesis, let $\Pr(Y_j \geq u|y_{1,\ell}, \hat{\theta}_{\alpha/2,j}) = \alpha/2$ and $\Pr(Y_j \geq u|y_{1,\ell}, \hat{\theta}_{1-\alpha/2,j}) = 1-\alpha/2$, the unconditional confidence interval for subgroup j will be $(\hat{\theta}_{1-\alpha/2,j}, \hat{\theta}_{\alpha/2,j})$.

Worked example

In this section, we still focus on a two-stage Magnusson and Turnbull (2013) design with two disjoint subgroups. Assuming that the family-wise error rate is required to be controlled under 0.025, we simulate 10,000 runs of trials and construct simultaneous confidence intervals for each individual subgroup in every single trial. As we mentioned in Section 4.2.1, suppose that each patient's outcome follows a normal distribution, we construct the score statistic for the treatment effect which distributes with mean $\theta_j\mathcal{I}_{k,j}$ and variance $\mathcal{I}_{k,j}$ where $\mathcal{I}_{k,j} = M_{k,j}/4\sigma_0^2$. Note that in all numerical studies, we report in this chapter, we estimate the variance σ_0^2 by the pooled sample variance. To maintain a power of 0.9, 625 patients are recruited in each stage, assuming the clinically effective treatment effect is $\boldsymbol{\theta} = (0.2, 0.2)$. 60% of them are assigned to subgroup 1 and 40% of them are assigned to subgroup 2 in the first stage. However, the sample size for

each subgroup in every trial is not fixed, we use the binomial distribution to randomize them according to the prespecified proportion. If we apply the first decision rule at the first interim analysis, then when only subgroup j is selected, we enrich it by recruiting 625 more patients. If both subgroups exceed their corresponding lower bounds in the first stage, then we use the same prevalences observed in the first stage to determine the sample size assignment between the two subgroups in the second stage. Block randomization is adopted here, which means that each patient is equally likely to be assigned to the experimental arm and control arm. By solving the spending error equation we mentioned in Chapter 3 Section 2.1.1, the boundaries are determined to be

$$(l_1, u_1) = (0.5192, 2.5529); (l_2, u_2) = (2.4072, 2.4072).$$

A worked example of the two-stage two-subgroup Magnusson and Turnbull (2013) design in which the overall group is selected at the first interim analysis is presented in Table 4.1. The pooled sample variance σ^2 in this simulated trial is 0.94823. The Fisher information scaled upper boundary in the first stage equals 33.6557, thereby the adjusted boundary for subgroup 1 and subgroup 2 are 28.7783 and 9.9998 respectively. Note that while each of the individual p-values will have a marginal $U[0, 1]$ distribution for the true value of θ_j, the p-values will not necessarily be independent and have a complicated distribution. We summarize the correlation between subgroups under different scenarios in Table 4.2. In situations when the trial is more likely to terminate at the second stage and more than one subgroup is selected, the dependence between subgroups is greater. One possible explanation is that the Magnusson and Turnbull design defines that when more than one subgroup proceeds to stage 2, their statistics are combined and then accumulated hence the dependence is plausibly increased. However, we notice the correlations are not substantial as all their magnitudes are less than 0.05.

In their study, Magnusson and Turnbull (2013) proposed a double bootstrap method for creating simultaneous confidence intervals for all subgroups in the design. The first round of bootstrapping employs the maximum likelihood estimates from the trial as the true treatment effect and calculates bias-corrected estimates using the sampled data. In the second round, the bias-corrected estimates are used as the treatment effect parameter inputs, and higher-order bias-corrected estimates are computed again. This process is repeated for multiple trials, and the lower bound of the simultaneous confidence interval is determined by taking the α quantile of the bias-corrected estimates. Furthermore, we present the simulation outcomes for the naive simultaneous confidence intervals in Table 4.3. To construct these naive simultaneous confidence intervals, we employ a simple Bonferroni correction, thereby subjecting each subgroup to a

Table 4.1: Result of one simulated trial under scenario $\boldsymbol{\theta} = (0.2, 0.2)$.

stage 1	$X_{1,j}$	$N_{1,j}$	$\delta_{1,j}$
Ω_1	23.6559	367	96.7531
Ω_2	4.8774	258	71.7434
stage 2	$X_{2,j}$	$N_{2,j}$	$\delta_{2,j}$
Ω_1	22.3110	367	96.7531
Ω_2	28.4639	258	71.7435
	Ω_1	Ω_2	
adjusted u_1 for individual subgroup j	28.7783	9.9998	

Table 4.2: Correlation coefficients between subgroup 1 and subgroup 2 using Kendall tau test.

scenario	correlation	proportion of trials terminate at stage 2	proportion of trials select both subgroups
$\boldsymbol{\theta} = (0, 0)$	0.0014	0.4997	0.0932
$\boldsymbol{\theta} = (0.2, 0)$	0.0023	0.6489	0.2790
$\boldsymbol{\theta} = (0.2, 0.2)$	0.0364	0.4655	0.7965
$\boldsymbol{\theta} = (2, 2)$	-0.0107	0	1

Table 4.3: Coverage probabilities and powers for double bootstrap sampling confidence intervals and naive simultaneous confidence intervals.

scenario	double bootstrap		naive	
	overall coverage probability	power	overall coverage probability	power
$\theta = (0, 0)$	0.9874	0.0126	0.9328	0.0672
$\theta = (0.2, 0)$	0.9123	0.4526	0.8318	0.7861
$\theta = (0.2, 0.2)$	0.6559	0.6213	0.8173	0.8813

significance level of 0.0125. Consequently, the formulation of the naive one-sided simultaneous confidence intervals is as follows:

$$\theta_1 > (y_{K,1} - 2.24\sqrt{\mathcal{I}_{K,1}})/\mathcal{I}_{K,1}$$
$$\theta_2 > (y_{K,2} - 2.24\sqrt{\mathcal{I}_{K,2}})/\mathcal{I}_{K,2}.$$

Unfortunately, the coverage probabilities and powers perform poorly under all scenarios as shown in Table 4.3, indicating the need to investigate alternative methods for constructing simultaneous confidence intervals in subsequent sections.

4.2.2 Simultaneous confidence intervals

At the termination of the trial, we need to test null hypotheses for all individual subgroups simultaneously, therefore it is necessary to apply multiple comparison procedures, such as the classic Bonferroni correction (Hsu, 1996), Bonferroni-Holm (Strassburger and Bretz, 2008), and Brannath and Schmidt (2014) method to construct the simultaneous confidence intervals. However, the classic Bonferroni procedure is less powerful in rejecting hypotheses while the lower bounds of the Holm method are uninformative where only part of the hypotheses are rejected (Strassburger and Bretz, 2008). In this context, the term "uninformative" characterizes situations where the confidence interval offers no insights into the magnitude of the actual treatment effect, whereas "informative" denotes cases where the confidence interval imparts some level of understanding about the true treatment effect. To illustrate, consider the Holm procedure, which yields identical lower bounds for rejected hypotheses when not all hypotheses are rejected, hence referred to as uninformative lower bounds. Conversely, the conventional Bonferroni procedure yields lower bounds that differ based on the actual treatment effect for rejected hypotheses,

making them informative. Therefore, Brannath and Schmidt (2014) proposed a new weighted Bonferroni approach in which the weight depends on parameters. In Section 4.2.3, we compare all three methods and show that the new weighted approach has uniform improvement in the number of rejections and informativeness.

Classic Bonferroni procedure

The Bonferroni correction is one of the most commonly used methods for controlling the family-wise error rate in multiple comparisons. It allows many comparison hypothesis statements to be made simultaneously while ensuring the overall type I error is strongly restricted under a certain level. The Bonferroni method is valid for both equal and unequal sample sizes and it is based on Boole's inequality (Hsu, 1996) which is presented in Equation (2.10) where E_i is the event that the ith constructed confidence level does not contain the true value. In this case, the left-hand side of the inequality is the probability that at least one interval does not cover its true value while the right-hand side of the inequality is the sum of the probabilities of each of the intervals missing their true values. Hence, if the family-wise error rate of the multiple interval estimates is desired to be restricted under α, then we can restrict the type I error rate of each interval by α/m where m is the number of statements or comparisons. For instance, we focus on the two-stage Magnusson and Turnbull design with two disjoint subgroups here thereby $m = 2$ and we are going to test the null hypothesis $H_{0,j} : \theta_j = 0$ at $\alpha/2$ level for $j = \{1, 2\}$.

A natural property of Equation (4.1) is that the p-value functions are continuously increasing from 0 to 1 in θ_j hence we can find the unique solution of the equation $\Pr(Y_j \geq u|y_{1,\ell}, \theta_j) = p_j(\theta_j) = \alpha/m$. Therefore $p_j^{-1}(\alpha/m)$ is the lower bound of the one-sided simultaneous confidence interval for subgroup j, which can be explicated as

$$\hat{\theta}_j^{BF} = p_j^{-1}(\alpha/m). \tag{4.2}$$

We denote $CI_j^{BF} = (\hat{\theta}_j^{BF}, \infty)$ as the one-sided simultaneous confidence interval for subgroup j. If $0 \notin CI_j^{BF}$, we reject $H_j : \theta_j = 0$ and declare treatment effective in subgroup j; otherwise, we accept H_j and declare futility in subgroup j.

The classic Bonferroni procedure makes no assumption about the dependence among hypotheses. However, the main limitations of the Bonferroni correction method are the inflation of the Type II error rate. If there are $m \geq 2$ comparisons, the coverage probability of the Bonferroni is defined as $(1 - \alpha/m)^m$ which is obviously greater than $1 - \alpha$. When the number of comparisons increases, the corresponding confidence interval becomes more conservative. In

addition, compared to the Bonferroni-Holm method proposed by Strassburger and Bretz (2008), the standard Bonferroni method rejects fewer hypotheses under the same scenario.

Numerical study Here we utilize the same setups as we proposed in the worked example. However, like the simulation study we presented in Chapter 3, we ran simulations under three scenarios: $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$. According to the classic Bonferroni approach, the significant level assigned to each subgroup is $\alpha/2$. Table 4.4 compares the coverage probability, overall power, and average number of rejections in each trial of three scenarios. We notice that all of those coverage probabilities are close to the nominal level we desired but not all of them are greater than 97.5%. Theoretically, by adopting the classic Bonferroni correction, the FWER should be more conservative and equal $(1 - (1 - \alpha/2)^2) = 0.0248$. However, under the null scenario, the FWER is less conservative than expected although it is not far away from the nominal level. The most likely causes of the lack of conservation are the limited number of comparisons (i.e.we only have two subgroups here), the use of pooled sample variance, and not accounting for random in the simulation procedure (i.e. the prevalence in the simulation procedure is inconsistent with the true prevalence).

Table 4.4: The coverage probability, overall power, and average number of rejections by applying the classic Bonferroni procedure under three scenarios.

scenarios	coverage probability	overall power	average number of rejections in one trial
$\boldsymbol{\theta} = (0, 0)$	0.9737	0.0263	0.0263
$\boldsymbol{\theta} = (0.2, 0)$	0.9752	0.7236	0.7290
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9758	0.7740	0.9245

Histograms for the distribution of the simultaneous confidence interval lower bounds are presented in Figure 4.1. Each row displays lower bounds of confidence intervals obtained under scenario $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ respectively. The left column lists all lower bounds from subgroup 1 simultaneous confidence intervals and the right column lists those from subgroup 2. What can be clearly seen in Figure 4.1 is that the 98.75% quantiles (vertical red line) are approximately located around the true treatment effect for every case which also implies that our individual p-value functions ensure the individual confidence intervals have coverage probabilities close to the nominal level.

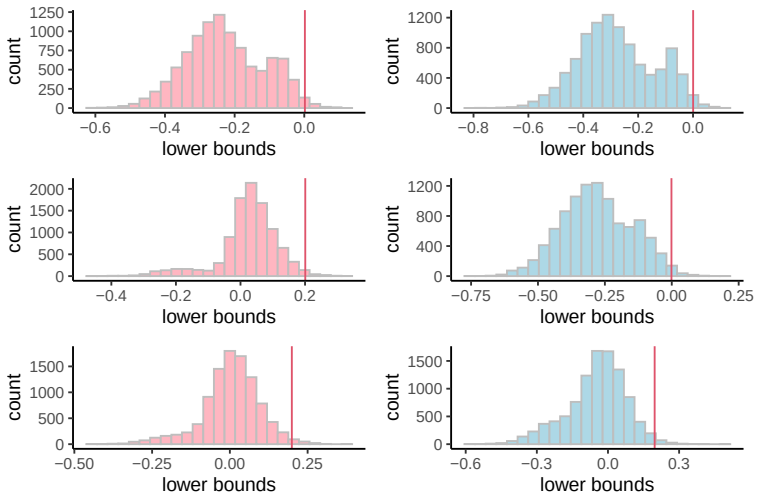


Figure 4.1: Distribution of the classic Bonferroni simultaneous confidence interval lower bounds with FWER constrained at or below 0.025. The vertical red lines are the 98.75% quantiles.

Bonferroni-Holm procedure

In order to increase the power of the multiple tests, Holm (1979) proposed a step-down procedure that rejects more null hypotheses while still preserving the coverage probability. The reason for naming this procedure "step-down" is that the first step is to order the individual p-value of each test. We denote $p_{(1)} \leq p_{(2)} \leq, ..., \leq p_{(m)}$ as the ordered p-values and $H_{0,(1)}, H_{0,(2)}, ..., H_{0,(m)}$ as the corresponding individual null hypotheses. According to the classic version of the Holm step-down procedure, we compare each p-value with $\alpha/(m-r+1)$ where m is the number of subgroups and r is the rank of the corresponding individual p-value ordering. The scheme can be described as follows:

1. Check whether $p_{(1)} < \alpha/m$. If so, we reject $H_{0,(1)}$, and proceed to step 2; otherwise, accept all hypotheses and stop.

2. Check whether $p_{(2)} < \alpha/(m-1)$. If so, we reject $H_{0,(1)}$ and $H_{0,(2)}$, then proceed to step 3; otherwise, we reject $H_{0,(1)}$, accept $H_{0,(2)}, ..., H_{0,(m)}$ and stop.

   $\vdots$

m. Check whether $p_{(m)} < \alpha$. If so, we reject all hypotheses and stop; otherwise, we reject $H_{0,(1)}, ..., H_{0,(m-1)}$, accept $H_{0,(m)}$ and stop.

The above scheme suggests that the rejection threshold increases as the procedure continues, therefore the Holm test is more likely to reject null hypotheses compared to the classic Bonferroni method. However, if we regard the coefficients: $1/m, 1/(m-1), ..., 1$ as the weight of FWER, we can say that the classic Bonferroni test assigns equal weight $(1/m)$ to every test while the Holm procedure assigns increasing weights to tests. Therefore, we also call the classic Holm procedure the "Bonferroni-Holm" procedure.

According to the definition of the Bonferroni-Holm testing procedure, Strassburger and Bretz (2008) computed the lower bounds of the one-sided simultaneous confidence intervals for subgroup j:

$$\hat{\theta}_j^{BH} = \begin{cases} 0 & \text{if } H_{0,j} \text{ is rejected but } H_{0,i\neq j} \text{ is accepted} \\ p_j^{-1}(\frac{\alpha}{m-r}) & \text{if } H_{0,j} \text{ is accepted} \\ \max(0, p_j^{-1}(\frac{\alpha}{m})) & \text{if all hypotheses are rejected} \end{cases} \tag{4.3}$$

where r is the number of rejected hypotheses and α is the nominal FWER. Although the classic Bonferroni method is conservative and lacks power compared to the Bonferroni-Holm method, its corresponding confidence intervals contain more informative rejections. Equation (4.3) indicates that the lower bound of the Bonferroni-Holm method is positive only when all hypotheses are rejected. If only part of the hypotheses is rejected, the lower bound of rejected hypotheses is 0 by definition whereas the lower bound of the accepted hypotheses is negative in terms of the Holm testing procedure (Strassburger and Bretz, 2008). $\hat{\theta}_j^{BH} > 0$ always implies $\hat{\theta}_j^{BF} > 0$, however, when $\hat{\theta}_j^{BH}$ equals zero, $\hat{\theta}_j^{BF}$ might still probably greater than zero. The zero lower bound shows no additional information about the true treatment effect (i.e. how far the treatment effect is from zero), thereby we say that the Bonferroni-Holm method is less informative.

Numerical study In order to show that the Bonferroni-Holm approach gives an improvement in the power of rejection, we simulate trials based on the same setups we proposed in Section 4.2.2. In total, 1,250 patients are recruited and half of them are assigned to the first stage. We randomly allocate 60% of them to subgroup 1 and 40% of them to subgroup 2. Again, suppose that we use block randomization in simulating trials, thereby each patient is equally likely to be assigned to the experimental or control arm. However, the Bonferroni-Holm test differs from the classic Bonferroni test in that it requires ordering p-values in advance. Therefore, after the simulation process, we calculate individual p-values for each subgroup under the null scenario and order them from small to large in every trial.

After running 10,000 trials, we compute and present the coverage probability, the proportion

of rejecting either hypothesis (over power), and the average number of rejected hypotheses in one trial in Table 4.5. What stands out in the table is the mean of the number of rejections, which is apparently greater than the classic Bonferroni method (shown in Table 4.4), although the overall powers (in terms of rejecting any null hypothesis) of these two methods are very close. Perhaps counter-intuitively, the coverage probability is very conservative when the true treatment effects are not both zero. A possible explanation is that we order the individual p-value according to the null hypothesis because they are what we intend to test; however, the lower bounds are computed based on the statistics simulated under true treatment effects. Another limitation of the step-down method is also clearly shown in Figure 4.2. Compared to the classic correction way, we barely see positive informative lower bounds under all scenarios when adopting the Bonferroni-Holm test to construct simultaneous confidence intervals. This implies the difficulty to have an idea about the scale of the true treatment effect in the Bonferroni-Holm testing procedure.

Table 4.5: The coverage probability, overall power, and average number of rejections by applying the Bonferroni-Holm procedure under three scenarios.

scenarios	coverage probability	overall power	average number of rejections in one trial
$\boldsymbol{\theta} = (0, 0)$	0.9737	0.0263	0.0265
$\boldsymbol{\theta} = (0.2, 0)$	0.9817	0.7236	0.7368
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9874	0.7742	1.0267

Weighted Bonferroni approach

From the above sections, we can say that the Bonferroni method rejects fewer hypotheses while the simultaneous confidence intervals for rejected hypotheses are informative. On the other hand, the Bonferroni-Holm method rejects more hypotheses while the lower bounds for rejected hypotheses are informative unless all hypotheses are rejected (i.e. the simultaneous confidence interval for the rejected hypothesis is $(0, \infty)$ when not all hypotheses are rejected). Therefore, Brannath and Schmidt (2014) proposed a new weighted Bonferroni approach that trades-off informativeness against the proportion of hypotheses rejected in one trial.

Brannath and Schmidt (2014) fix a penalizing function $\lambda_j : \mathbb{R} \to [0, \infty)$ to construct a weight function that depends on the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)$. λ_j can be any continuously non-decreasing functions. For instance, we specify it as $\lambda_j(x) = \exp(\max(0, ax))$ in the two-

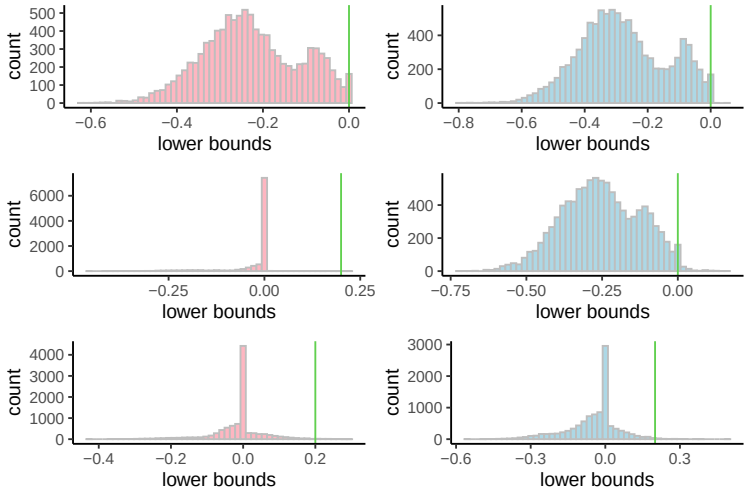


Figure 4.2: Distribution of the Bonferroni-Holm simultaneous confidence interval lower bounds with FWER constrained at or below 0.025. The vertical green lines are the true treatment effects.

stage two-subgroup simulation studies. Then, the weight function depending on the parameter $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_m\}$ for subgroup j will be:

$$\omega_j(\boldsymbol{\theta}) = \frac{\lambda_j(\theta_j)^{-1}}{\sum_{i=1}^{m} \lambda_i(\theta_i)^{-1}}. \tag{4.4}$$

Note that in Equation (4.4) the value of $\lambda_j(\theta_j)$ increases, the weight assigned to $H_{0,j}$ in the overall p-value $p(\boldsymbol{\theta})$ decreases hence why λ_j is referred to as the penalizing function. In Brannath and Schmidt's procedure, hypothesis $H_{0,j}$ is tested at $\omega_j(\boldsymbol{\theta})\alpha$ level, therefore we let

$$p_j(\theta_j) = \omega_j(\boldsymbol{\theta})\alpha \tag{4.5}$$

Equation (4.5) indicates that

$$p_j(\theta_j)\lambda_j(\theta_j) = \left(\sum_{i=1}^{m} \lambda_i(\theta_i)^{-1}\right)^{-1} \alpha, \text{ for all } j = 1, ..., m, \tag{4.6}$$

where the right hand of the equation is a constant. Therefore, we have $p_1(\theta_1)\lambda_1(\theta_1) =, ..., = p_m(\theta_m)\lambda_m(\theta_m)$. Based on this property, Brannath and Schmidt (2014) define the overall p-valaue for each intersection hypothesis $H_{\boldsymbol{\theta}} = H_{0,1} \cap ... \cap H_{0,m}$ as:

$$p(\boldsymbol{\theta}) = \min\left(\left(\min_{j=1,..m}\{p_j(\theta_j)\lambda_j(\theta_j)\}\sum_{i=1}^{m}\lambda_i(\theta_i)^{-1}\right), 1\right) \tag{4.7}$$

Let $\Theta = \{\theta \in \mathbb{R} : p_1(\theta_1)\lambda_1(\theta_1) =, .., = p_m(\theta_m)\lambda_m(\theta_m)\}$, Brannath and Schmidt (2014) proved that $p(\boldsymbol{\theta})$ is increasing on Θ, thereby we can find a unique solution $\hat{\boldsymbol{\theta}}$ to $p(\boldsymbol{\theta}) = \alpha$. Note that

if we fix one element in Θ, the rest elements could be recursively solved according to Equation (4.6).

We use a straightforward algorithm that is different from Brannth and Schmidt's bisection searching algorithm. Our algorithm relies on *uniroot* zero root searching function in R, which we mentioned in the previous chapter. *uniroot* function is constructed based on the root searching algorithm proposed by Brent (2013) which is "never much slower than bisection, but which has the advantage of superlinear convergence to a simple zero of a continuously differentiable function". By utilizing *uniroot* function multiple times, our algorithm can be described below:

1. Let $\boldsymbol{\vartheta} = (\vartheta_1, ..., \vartheta_m)$ be the m-dimensional set of treatment effect that satisfies $p(\boldsymbol{\vartheta}) = \alpha$, then $p_1(\vartheta_1)\lambda_1(\vartheta_1) =, ..., = p_m(\vartheta_m)\lambda_m(\vartheta_m)$.

2. Then, if we choose ϑ_j, every ϑ_i for $i \neq j$ can be expressed as a function of ϑ_j by applying *uniroot* function in R. Hence the dimension of the root search is reduced from m to 1.

3. Finally, we search for possible value of ϑ_j in (a, b) that satisfies $p(\vartheta_1, ..., \vartheta_j, ..., \vartheta_m) = \alpha$.

The above algorithm gives us a set of unique roots $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ of equation $p(\boldsymbol{\theta}) = \alpha$ on $\hat{\theta}_j \in (a, b)$.

Note that when $\lambda_j(\theta_j) = 1$ for all $j = 1, ..., m$, we get the Bonferroni bound $CI_j^{BF} = p_j^{-1}(\alpha/m)$. Also, Bonferroni-Holm bounds can be obtained by letting $\lambda_i(\theta_i) = 1$ for accepted hypotheses and letting $\lambda_i(\theta_i) \to \infty$ for rejected hypotheses. Since there are no clear definitions of weights when all hypotheses are rejected, we define $w_i(\theta_i) = \frac{1}{m}$ in this case. If only part of the hypotheses is rejected, we define $\lambda_i^{-1}(\theta_i) = 0$ and allocate zero weight to those hypotheses. This means we obtain no power from individual p-value functions $p_i(\delta_i)$ in this situation.

Brannath and Schmidt (2014) proved that the proposed weighted Bonferroni method provides a uniform improvement of the classic Bonferroni confidence interval. Consider the penalizing function $\lambda_j(x) = \exp(\max(0, ax))$, a continuous nondecreasing function that has the range $\lambda_i : \mathbb{R} \to [1, \infty)$. Obviously,

$$\omega_j(\boldsymbol{\theta}) = \frac{\lambda_j(\theta_j)^{-1}}{\sum_{i=1}^{m} \lambda_i(\theta_i)^{-1}} \geq \frac{1}{m},$$

which means the Brannath and Schmidt (2014) approach always tests the hypothesis $H_{0,j}$ at a level greater than the Bonferroni method does in terms of Equation (4.5). This implies all hypotheses rejected by the Bonferroni test will be rejected by the new weighted Bonferroni test. Namely, it provides more informative rejections than Bonferroni bounds $\hat{\theta}_i^{BF} = p_i^{-1}(\frac{\alpha}{m})$. When

$\hat{\theta}_i^{BF} > 0$, then $\hat{\theta}_i^{BS} > 0$ for all $i = 1, .., m$ at every point. Note that the above conclusion only implies that the $\hat{\theta}_i^{BS}$ are greater than zero whenever the $\hat{\theta}_i^{BF}$ is but does not necessarily imply that $\hat{\theta}_i^{BS} > \hat{\theta}_i^{BF}$. In fact, compared to the Bonferroni lower bound, the Brannath and Schmidt lower bound shrink close to each other. For instance, we consider two subgroups design and let the penalty function be $\lambda_j(x) = \exp(ax)$. Then the weight for each subgroup will be

$$\omega_1(\boldsymbol{\theta}) = \frac{1}{1 + \exp(a(\theta_1 - \theta_2))} \text{ and } \omega_2(\boldsymbol{\theta}) = \frac{1}{1 + \exp(a(\theta_2 - \theta_1))}$$

respectively. Apparently, if $\theta_1 > \theta_2$, then $\omega_1 < \alpha/2$, and vice versa. When $\hat{\theta}_1^{BF} > \hat{\theta}_2^{BF}$, we have $p_1(\hat{\theta}_1^{BS}) < \alpha/2 = p_1(\hat{\theta}_1^{BF})$. Since $p_j(\theta_j)$ is an increasing function on θ_j, we conclude that $\hat{\theta}_1^{BS} < \hat{\theta}_1^{BF}$. Analogously, as $p_2(\hat{\theta}_2^{BS}) > \alpha/2 = p_2(\hat{\theta}_2^{BF})$, we say that $\hat{\theta}_2^{BS} > \hat{\theta}_2^{BF}$. Therefore $\hat{\boldsymbol{\theta}}^{BS}$ are closer to each other than $\hat{\boldsymbol{\theta}}^{BF}$. This shrinkage property helps to explain the improvement of Brannath and Schmidt (2014) approach compared to the classic Bonferroni method. When at least one lower bound among subgroups are positive, the parameter-dependent weighted approach decreases the bounds (though still positive) to detect more rejections.

A main limitation of the Bonferroni-Holm confidence intervals is that when only part of the null hypotheses is rejected, the lower bounds of the one-sided simultaneous confidence intervals are 0 for those rejected subgroups. Brannath and Schmidt improve this by producing simultaneous confidence intervals that have positive lower bounds on the above occasion. However, when $\hat{\theta}_j^{BH} = 0$, the Brannath and Schmidt approach allows its simultaneous confidence intervals $\hat{\theta}_j^{BS}$ below 0 which means the new procedure rejects less null hypotheses the Holm procedure.

Until now, the new parameter-dependent weighted Bonferroni approach showed improvement regarding the power of rejection and informativeness theoretically. In Section 4.2.3, we will compare the three methods based on the simulation study that involves three subgroups in the Magnusson and Turnbull (2013) design.

Numerical study In this section, we adopt similar setups as in the above numerical studies. We still recruit 625 patients for each stage. The prevalence of subgroup 1 and subgroup 2 is 0.6 and 0.4 respectively. What distinguishes the Brannath and Schmidt approach from the classic Bonferroni confidence interval is the introduction of the parameter-dependent weight function. Here, we let the penalizing function $\lambda_j(x)$ be $\exp(\max(0, ax))$ and take $a = 1$. In the same way, as the classic Bonferroni and Bonferroni-Holm numeric studies did, we simulate 10,000 trials based on three scenarios (i.e.$\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$) with coverage probability constrained equivalent or above 0.975. Recall that we are testing one-sided hypotheses for two

disjoint subgroups, the unconditional confidence interval for subgroup j will be $(\hat{\theta}_j, \infty)$.

From Table 4.4 and Table 4.6, we notice that the Brannath and Schmidt (2014) approach has a similar overall power compared with the classic Bonferroni method. As for the coverage probability, still, no significant improvement is detected. Under the null scenario, we find no difference in the power of rejection between the classic method and the new weighted approach, however, in the cases where the treatment is effective, the average number of hypothesis rejections is slightly greater than the classic Bonferroni method.

Table 4.6: The coverage probability, overall power, and average number of rejections by applying the Brannath and Schmidt procedure under three scenarios.

scenarios	coverage probability	overall power	average number of rejections in one trial
$\boldsymbol{\theta} = (0, 0)$	0.9737	0.0263	0.0263
$\boldsymbol{\theta} = (0.2, 0)$	0.9759	0.7236	0.7293
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9778	0.7742	0.9291

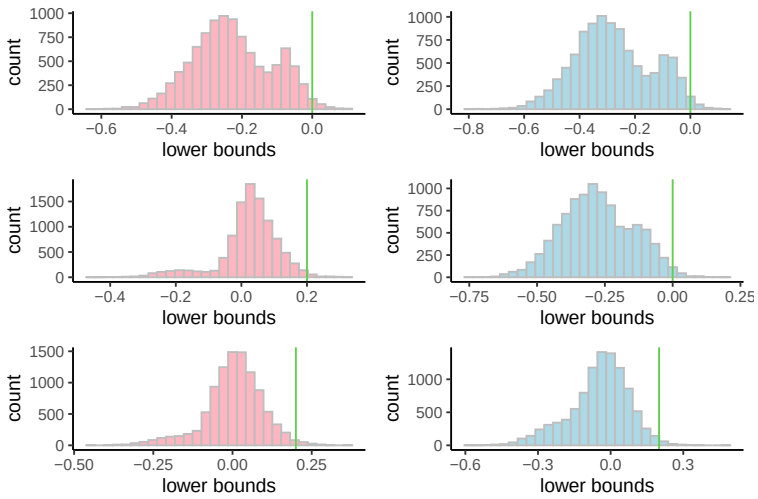


Figure 4.3: Distribution of the Brannath and Schmidt simultaneous confidence interval lower bounds for two subgroups with FWER constrained at or below 0.025. The green vertical lines are the true treatment effects.

Distributions of the simultaneous confidence interval lower bounds are presented in Figure 4.3. The light pink histograms illustrate the distribution of lower bounds for subgroup 1 and

the light blue histograms show the distribution of lower bounds for subgroup 2. Compared with histograms we got from the classic Bonferroni correction (Figure 4.1), the parameter-dependent weighted Bonferroni approach has more non-negative lower bounds when the true treatment effect is greater than zero. This implies that the new weighted Bonferroni approach rejects more null hypotheses. In contrast, the Brannath and Schmidt procedure provides a greater number of positive bounds than the Bonferroni-Holm procedure, suggesting that its confidence intervals are more informative. Nevertheless, the Brannath and Schmidt method generates fewer non-negative bounds than the Bonferroni-Holm method, which may indicate lower efficacy in rejecting hypotheses.

4.2.3 Comparison of the three simultaneous confidence interval construction approaches in the two-stage three-subgroup Magnusson and Turnbull design

From Sections 4.2.2, we know that the new weighted Bonferroni approach proposed by Brannath and Schmidt has a uniform improvement in hypotheses rejection compared with the classic Bonferroni method. On the other hand, the new approach provides more informative rejections than the Bonferroni-Holm method. However, theoretically, the improvement is more substantial when the number of comparisons increases. Until now, we have only focused on a two-stage two-subgroup design. In order to enhance the improvement of the Brannath and Schmidt approach, we consider the two-stage Magnusson and Turnbull (2013) design with three subgroups.

We start from the simple design which assumes that the statistics of subgroup j follows the normal distribution $N(\theta_j, 1)$ with $j = \{1, 2, 3\}$. We are interested in the case where only subgroup 1 and subgroup 2 are promising, thereby we fix the individual p-value for subgroup 3 as 0.5. Suppose that the FWER needs to be restricted at or below 0.05, Figure 4.4 displays the rejection regions for all three multiple test procedures. The pink rectangle is the region where both $H_{0,1}$ and $H_{0,2}$ are rejected by the classic Bonferroni test. The green region indicates cases, where $H_{0,1}$ and $H_{0,2}$, are rejected by the Brannath and Schmidt test and the Bonferroni-Holm test but at least one of the null hypotheses is accepted by the classic Bonferroni test. The blue region includes occasions on which both of the null hypotheses are only rejected by the Bonferroni-Holm test procedure. It is apparent from the figure that the rank of the power of the three tests to detect the true treatment effects is: $Power_{Holm} > Power_{Brannath} > Power_{Bonferroni}$. Furthermore, when the value of a in the penalizing function $\lambda(x) = \exp(\max(0, ax))$ increases, we notice

that the rejection boundary of the Brannath and Schmidt procedure approaches the Bonferroni-Holm's boundary. This agrees with the conclusion that when a approaches infinity, the power of the Brannatha and Schmidt procedure increases and becomes closer to the Bonferroni-Holm procedure. However, the cure of the boundary of the Brannath and Schmidt test will never exceed the diagonal line between $(p_1, p_2) = (0.025, 0)$ and $(p1, p2) = (0.05/3, 0.05/3)$.

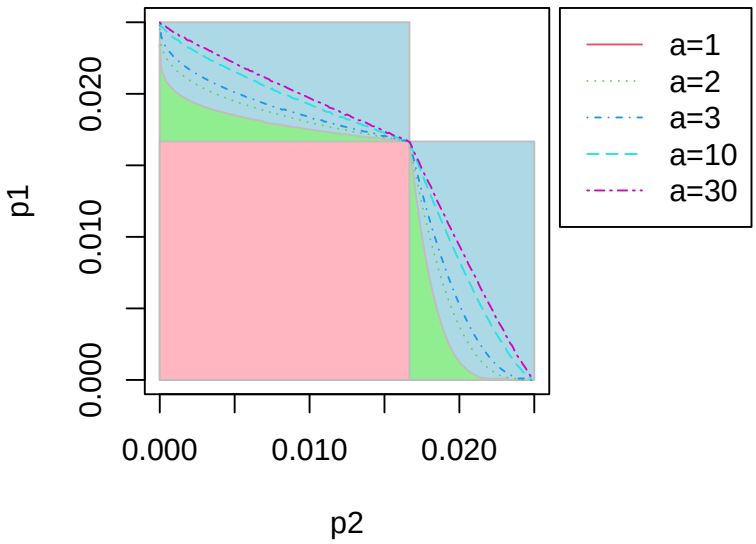


Figure 4.4: Rejection region for the classical Bonferroni test procedure, the Bonferroni-Holm test procedure and the Brannath and Schmidt procedure. The x-axis is the p-value of subgroup 2 under the null scenario. The y-axis is the p-value of subgroup 1 under the null scenario.

Now, we consider the three-subgroup case in the two-stage Magnusson and Turnbull enrichment design. The p-value function of the simultaneous confidence interval for the three-subgroup design is quite similar to the two-subgroup case, except that the p-value is computed conditional on the first-stage results for the other two subgroups. Hence, when computing the upper boundary for the target subgroup when the entire population is selected in the first stage, we must deduct the observed statistics from the two remaining subgroups. Therefore, we define the un-

conditional p-value function for subgroup j as

$$\begin{aligned}\Pr(Y_j \geq u|\mathbf{y}_{1,\ell},\theta_j) = &\int_{\max(u,\tilde{u}_{1,\mathcal{S}}-\mathbf{y}_{1,\ell})}^{\infty} \frac{1}{\sqrt{\delta_{1,j}}}\psi\left(\frac{y_{1,j}-\theta_j\delta_{1,j}}{\sqrt{\delta_{1,j}}}\right)\mathrm{d}y_{1,j} \\ &+ \int_{\tilde{l}_{1,j}}^{\tilde{u}_{1,\mathcal{S}}-\mathbf{y}_{1,\ell}} \frac{1}{\sqrt{\delta_{1,j}}}\psi\left(\frac{y_{1,j}-\theta_j\delta_{1,j}}{\sqrt{\delta_{1,j}}}\right) \\ &\times\left[1-\Phi\left(\frac{u-y_{1,j}-\theta_j\delta_{2,j}}{\sqrt{\delta_{2,j}}}\right)\right]\mathrm{d}y_{1,j}\end{aligned} \tag{4.8}$$

where $\mathbf{y}_{1,\ell} = \sum_{i\in\ell} y_{1,i}$ and $\delta_{k,j} = N_{k,j}/4\sigma^2$ with $k = \{1,2\}$.

To guarantee the null hypothesis is rejected with probability $1-\beta = 0.9$ based on the first criteria we mentioned in Chapter 2 beforehand with $\theta_1 = \theta_2 = \theta_3 = 0.2$, we need to recruit a maximum of 1560 patients in total. However, we round the sample size up to 1600 by considering the drop-off at the first interim analysis. We allocate an equal number of patients to each stage and the prevalence for subgroups are assumed to be $\rho_1 = 0.6$, $\rho_2 = 0.2$, and $\rho_3 = 0.2$ respectively. As described in previous numerical studies, the sample size of each subgroup is randomly generated according to its corresponding prevalence. As for σ^2, again, we estimate it by the pooled sample variance. Moreover, every patient is supposed to be assigned to the experimental or control arm with equivalent probability. If the trial proceeds to the second stage, the proportion randomized to subgroup j is defined as $\rho_j/\sum_{i\in\mathcal{S}^*}\rho_i$, where $\mathcal{S}^*$ is the index set of the selected population. By recursively solving spending error functions given in Chapter 3, we get the standardized boundaries for the three-subgroup design:

$$(l_1,u_1) = (0.7962, 2.7625); (l_2,u_2) = (2.5204, 2.5204).$$

Moreover, when implementing the Brannath and Schmidt method, we choose $\lambda_j(x) = \exp(\max(0,x))$ to be the penalizing function.

Table 4.7 presents a worked example of the three-stage two-subgroup Magnusson and Turnbull (2013) design. The estimation for the pooled sample variance is 1.0061. Since only group 1 and group 2 are selected, the Fisher information scaled upper boundary in the first stage is 35.0528. By subtracting the statistic observed for subgroup 2, the adjusted upper boundary for subgroup 1 is 15.9276. Analogously, the adjusted upper boundary for subgroup 2 only equals 25.4110. Note that subgroup 3 is dropped at the first interim analysis. In this case, we deduct observed values for all selected subgroups when computing the individual upper boundary for the dropped subgroup.

Then we compare the three simultaneous confidence interval approaches by running 10,000 simulated trials. The results of coverage probabilities and overall powers are summarized in

Table 4.7: Result of one simulated trial under scenario $\boldsymbol{\theta} = (0.2, 0.2, 0.2)$.

stage 1	$X_{1,j}$	$N_{1,j}$	$\delta_{1,j}$
Ω_1	9.6419	491	121.9970
Ω_2	19.1252	157	39.0092
Ω_3	-4.9219	152	38.2393
stage 2	$X_{2,j}$	$N_{2,j}$	$\delta_{2,j}$
Ω_1	55.8415	606	150.6136
Ω_2	8.4422	194	48.1595
Ω_3	\	\	\
	Ω_1	Ω_2	Ω_3
adjusted u_1 for individual subgroup j	15.9276	25.4110	6.2858

Table 4.8. Here, overall power refers to the probability of detecting at least one true treatment effect. We notice the Brannath and Schmidt procedure is slightly more conservative than the classic Bonferroni method, which agrees with the conclusion we drew from the two-subgroup case. Meanwhile, the overall power of the new procedure is quite close to the classic Bonferroni method. The Bonferroni-Holm method has greater overall power in general, however, the superiority is not substantial. What stands out in this table is that the Brannath and Schmidt approach has a greater number of rejections in one trial on average compared with the classic Bonferroni method but slightly fewer than the Bonferroni-Holm approach. In Figure 4.5, we illustrate the distribution of the simultaneous confidence interval lower bounds for subgroup 1. Lower bounds in the first row are simulated under the null scenario, while the lower bounds in the second row and third row are simulated under $\boldsymbol{\theta} = (0.2, 0, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2, 0.2)$ respectively. Additionally, the distribution of lower bounds obtained from the classic Bonferroni procedure, the Bonferroni-Holm procedure, and the parameter-dependent weighted approach are presented in the light pink, light blue, and light green histograms respectively. Apparently, Figure 4.5 shows that the lower bounds of Bonferroni-Holm simultaneous confidence intervals are much less informative than the other two methods as there are fewer positive bounds. Taken together, these results suggest that the Brannath and Schmidt simultaneous confidence intervals are more powerful in rejecting hypotheses than the classic Bonferroni simultaneous confidence intervals and provide more information about the true treatment effect compared with the Bonferroni-Holm

simultaneous confidence intervals. However, the classic Bonferroni procedure gives approximately nominal coverage probabilities for all scenarios whereas the Bonferroni-Holm and Brannath and Schmidt procedure clearly become substantially conservative in non-null scenarios. We guess that the extreme conservativeness in the Bonferroni-Holm procedure is owing to the pre-ordering of p-values under the null scenario which is inconsistent with the actual simulation scenario. The conservativeness of the Brannath and Schmidt procedure under non-null scenarios stems from the property that it lowers the positive lower bounds to detect more rejections, which makes the lower bounds less likely to cover the true treatment effect compared with the classic Bonferroni approach.

Table 4.8: The simulation results are presented for the Bonferroni, Holm, and parameter-dependent weighted Bonferroni simultaneous confidence interval lower bounds. The overall power is the probability of rejecting any null hypothesis when they are false. The average number of rejected hypotheses is the mean of the number of rejections in 10,0000 simulation runs.

scenario	classic Bonferroni	Bonferroni-Holm	Brannth and Schmidt
	coverage probability		
$\boldsymbol{\theta} = (0, 0, 0)$	0.9730	0.9730	0.9729
$\boldsymbol{\theta} = (0.2, 0, 0)$	0.9716	0.9767	0.9728
$\boldsymbol{\theta} = (0.2, 0.2, 0.2)$	0.9759	0.9968	0.9792
$\boldsymbol{\theta} = (0.2, 0.3, 0.5)$	0.9761	0.9901	0.9804
	overall power		
$\boldsymbol{\theta} = (0, 0, 0)$	0.027	0.0271	0.0271
$\boldsymbol{\theta} = (0.2, 0, 0)$	0.7835	0.7836	0.7835
$\boldsymbol{\theta} = (0.2, 0.2, 0.2)$	0.7799	0.7800	0.7799
$\boldsymbol{\theta} = (0.2, 0.3, 0.5)$	0.9389	0.9390	0.9390
	average number of rejected hypotheses		
$\boldsymbol{\theta} = (0, 0, 0)$	0.0272	0.0274	0.0273
$\boldsymbol{\theta} = (0.2, 0, 0)$	0.7931	0.7992	0.7934
$\boldsymbol{\theta} = (0.2, 0.2, 0.2)$	0.9621	1.0542	0.9656
$\boldsymbol{\theta} = (0.2, 0.3, 0.5)$	1.5723	1.7970	1.5896

Now we consider the effect of the penalizing function. As we require the penalizing function to be monotonically continuously increasing, so we choose $\lambda_j(x) = \exp(\max(0, ax))$. As can be

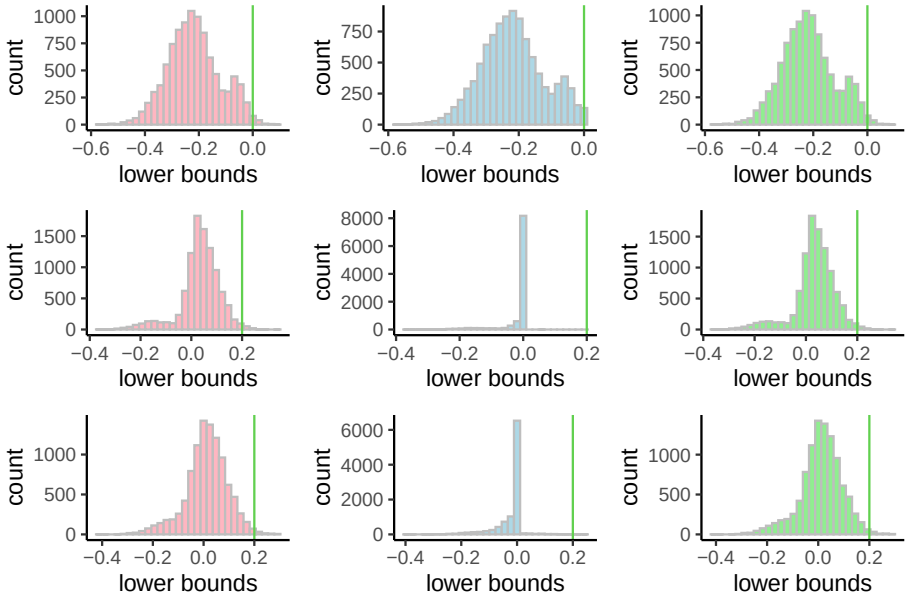


Figure 4.5: Distributions of the classic Bonferroni, Bonferroni-Holm and Brannath and Schmidt simultaneous confidence intervals lower bounds for subgroup 1 under scenario $\boldsymbol{\theta} = (0, 0, 0)$, $\boldsymbol{\theta} = (0.2, 0, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2, 0.2)$. The green vertical lines are the true treatment effects.

seen from Figure 4.6, when a increases, the average number of rejections increases as well. When $a = 0$, the average number of rejections of the weighted approach and the classic Bonferroni are the same, which agrees with the conclusion that when $\lambda_j(x) = 1$, the parameter-dependent weighted approach is exactly the classic method. When a increases, the value of $\lambda_j(x)$ approaches infinity, and the average number of rejections of Brannath and Schmidt (2014) method is close to the average number of rejections of the Holm procedure. This also indicates that the Bonferroni-Holm test is a specific form of the parameter-dependent weighted Bonferroni test.

As we mentioned beforehand, there is a trade-off between the power of rejecting a hypothesis and the informativeness in the parameter-dependent weighted Bonferroni approach. Brannath and Schmidt (2014) proposed an utility function to describe this trade-off for subgroup j:

$$\begin{aligned} U_j(\theta_j) =& \pi \times Power + (1 - \pi) \times Informativeness \\ =& \pi Pr(\hat{\theta}_j \geq 0) + (1 - \pi) E_{\theta_j}[\min(\max(\hat{\theta}_j/\theta_j, 0), 1)] \end{aligned}$$

If the number of responsive subgroups is greater than one, the overall utility for the trial is

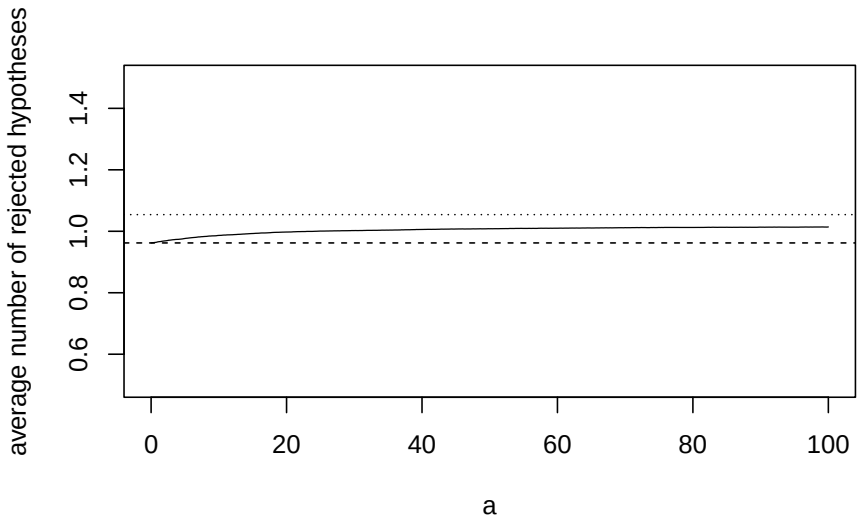


Figure 4.6: Average number of rejected hypotheses when the true treatment effects for all subgroups are 0.2 with $\lambda_j(x) = \exp(\max(0, ax))$ plotted as a function of a. The dotted line is the average number of rejections for the Holm procedure. The dashed line is the average number of rejections for the classic Bonferroni procedure.

defined as:

$$\begin{aligned} U(\boldsymbol{\theta}_\tau) =& \sum_{i\in\tau} U_i(\theta_i) \\ =& \pi \sum_{i\in\tau} Pr(\hat{\theta}_i \geq 0) + (1-\pi) \sum_{i\in\tau} E_{\theta_i}[\min(\max(\hat{\theta}_i/\theta_i, 0), 1)] \end{aligned} \tag{4.9}$$

where $\tau = \{i : \theta_i > 0\}$ and $\sum_{i\in\tau} Pr(\hat{\theta}_i \geq 0) = E_{\boldsymbol{\theta}_\tau}(\text{number of correct rejections})$. In general, the optimal a depends on π, and π is chosen based on the preference for power compared to informativeness. As an illustration, we choose π such that the utilities of the classic Bonferroni and the Bonferroni-Holm procedures are the same. After that, we search for possible a that ensures the Brannath and Schmidt approach outperforms those two methods by substituting the π fixed beforehand.

For instance, considering the 10,000 trials simulated under the scenario $\boldsymbol{\theta} = (0.2, 0.2, 0)$, the powers for the classic Bonferroni and Holm method are 0.9006 and 0.9434. As for informativeness, the expected distances between the lower bound and the true treatment effect are 0.2951 and 0.0018 respectively. By letting the utilities of those two methods be equivalent, the π is supposed to be fixed as 0.8726. As can be seen from Figure 4.7, when a is approximately smaller than 82, the Brannath and Schmidt procedure has greater utility than both the classic Bonferroni and Bonferroni-Holm methods. Therefore, we are supposed to choose a value in $(0, 82)$ for a. When a approximately equals 17, the utility is maximized. Additionally, in Figure 4.7, we can see that there is a downward trend for $a \geq 17$. Meanwhile, Figure 4.6 demonstrates that as a approaches infinity, the power of the Brannath and Schmidt procedure's rejection gradually converges to that of the Bonferroni-Holm procedure, but never quite reaches the same level of power. However, an increase in a results in a decrease in informativeness. Thus, the reduction in utility shown in Figure 4.7 may be due to the fact that the decrease in informativeness outweighs the increase in power when a becomes extremely large.

4.3 Conclusion

In this chapter, the aim is to construct simultaneous confidence intervals for the individual subgroups in the Magnusson and Turnbull (2013) enrichment design. We adopted three approaches: the classic Bonferroni procedure, the Bonferroni-Holm step-down procedure, and the Brannath and Schmidt procedure. We notice that the lower bounds of the Bonferroni simultaneous confidence intervals are informative but lack the power to reject hypotheses. In other words, those lower bounds provide information about how far the true treatment effect is from zero but are

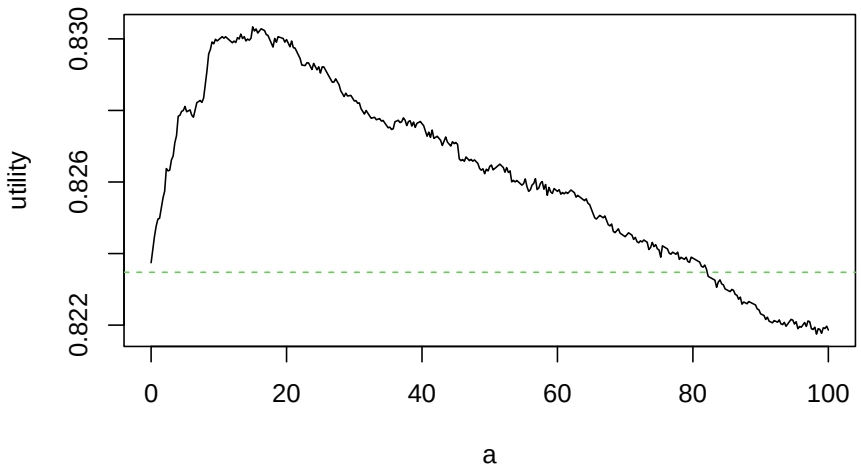


Figure 4.7: Utility of the Brannath and Schmidt procedure with $\pi = 0.8726$ against a. The horizontal dashed green line is the utility of the classic Bonferroni and Bonferroni-Holm methods.

weak in detecting subgroups that respond to the treatment. On the contrary, the Bonferroni-Holm step-down procedure is able to reject more hypotheses when their corresponding subgroups have treatment effects in fact, however, in order to control the FWER, we need to define that the positive lower bounds only exist when all hypotheses are rejected. This means that we have no idea about the scale of the true treatment effect when not all of the hypotheses are rejected. Therefore, Brannath and Schmidt (2014) proposed a new weighted Bonferroni procedure whose weight depends on the parameter. The weighted approach is more powerful in rejecting hypotheses in a single trial while ensuring the simultaneous confidence interval is informative.

An issue that was not addressed in this study is the choice of the penalizing function in the weighted Bonferroni approach. Brannath and Schmidt (2014) propose that we are supposed to choose a that ensures the utility function favors Brannath and Schmidt procedure over the classic Bonferroni and Bonferroni-Holm procedure. However, in order to optimize the expected utility, it would be necessary to specify a prior distribution for the true treatment effects and to estimate (via simulation) the utility for given realizations from the prior distribution, which is very complicated in practice. Therefore, considerably more work will need to be done to determine the optimal choice of the penalizing function.

It is somewhat surprising that the Bonferroni simultaneous confidence intervals give satisfying coverage in all scenarios whereas the other two procedures are conservative in non-null scenarios. Furthermore, as Magnusson and Turnbull (2013) design already make decisions about rejecting and accepting null hypotheses, we are more interested in getting knowledge about the scale of the true treatment effects. When the number of tests in the multiple comparisons is large, the advantage of the Brannath and Schmidt procedure is more substantial. However, the number of subgroups involved in the enrichment design usually is less than five. Hence we prioritize the informativeness of the simultaneous confidence interval construction approach and recommend the classic Bonferroni procedure.

Chapter 5

Generalized method to construct confidence intervals in the enrichment design

5.1 Introduction

In this chapter, we describe the generalized form of our method to construct confidence intervals for a specifically selected subpopulation or an individual subgroup regardless of selection results in the interim analysis. In order to illustrate how the generalized approach works, we apply it to the adaptive enrichment design proposed in Lin et al. (2021)'s paper as an example in simulation studies. An essential assumption in our approach is that the outcomes of patients follow the normal distribution. However, in fact, our approach can also be applied to designs where the outcome of the patient follows non-normal distribution as long as the approximately normal distributed score statistics are available.

5.2 Generalized approach to constructing confidence intervals

In previous chapters, we derived conditional and unconditional p-value functions for Magnusson and Turnbull (2013) design. Given certain subgroups are selected, the conditional confi-

dence intervals are obtained by inverting their corresponding p-value functions. Similarly, the simultaneous confidence intervals at the termination of the trial are obtained by inverting the corresponding unconditional p-value functions based on a certain multiple-testing procedure. In this section, we generalize the approach of confidence interval construction for enrichment design from what we did in Magnusson and Turnbull design and apply it to Lin et al. (2021) design as an example in Section 5.3.

5.2.1 Conditional confidence intervals given a certain subgroup selected

In this section, we propose a generalized approach to constructing confidence for those subpopulations that are retained after the interim analysis. The general approach could be applied in various adaptive enrichment designs as long as the selection rule is specified beforehand. In this manuscript, we concentrate on the two-stage adaptive enrichment design with two disjoint subgroups. The selection rule varies in different enrichment designs. However, all of the entire sample spaces in the first stage Ω_0 should be able to be partitioned into the following disjoint subspaces for conditional confidence intervals construction:

$$\Omega_1^c = \{\text{all subgroups stop for futility}\},$$

$$\Omega_2^c = \{\text{only enrich subgroup 1 at stage 2}\},$$

$$\Omega_3^c = \{\text{only enrich subgroup 2 at stage 2}\},$$

$$\Omega_4^c = \{\text{enrich both subgroups at stage 2}\},$$

$$\Omega_5^c = \{\text{subgroup 1 stops for efficacy and subgroup 2 stops futility at stage 1}\},$$

$$\Omega_6^c = \{\text{subgroup 2 stops for efficacy and subgroup 1 stops futility at stage 1}\},$$

$$\Omega_7^c = \{\text{the overall group stops for efficacy at stage 1}\} .$$

Again, the term "conditional" refers to conditioning on the information that certain subgroups are chosen in the first interim analysis. The boundaries of these disjoint subspaces are determined by the selection rule and FWER restrictions. As seen in Figure 2.2 for the Magnusson and Turnbull (2013) design, region 1 corresponds to the sample space Ω_1^c in this case. Regions 2 and 3 correspond to the sample spaces Ω_2^c and Ω_3^c, respectively. If the observed statistics fall within regions 4 or 7, it indicates that the whole group has been chosen and may be enriched

in the second stage or discontinued for efficacy in the first. Finally, sample space Ω_5^c and Ω_6^c corresponds to region 5 and region 6.

Let $\mathcal{S}^*$ be the possible index set of the selected subpopulation after the interim analysis. ρ_j is the prevalence of subgroup j, which determines the sample size of subgroup j. Assuming that N patients in total will be recruited in the trial, the number of patients allocated to stage 1 and stage 2 will be N_1 and N_2 respectively. Among the N_1 patients, $N_{1,j} = \rho_j N_1$ of them are expected to be recruited from subpopulation j. Let $Y_{k,\mathcal{S}^*}$ be the accumulated and combined statistic for chosen subset $\mathcal{S}^*$ at stage K. Additionally, we denote the statistic increment for the selected subset at stage k as $X_{k,\mathcal{S}^*}$ ($k = \{1, 2\}$), then the accumulated statistic is defined as $Y_{1,\mathcal{S}^*} = X_{1,\mathcal{S}^*}$ at stage 1. When the trial proceeds to stage 2, the statistic increment for the enriched subset $\mathcal{S}$ is defined as $X_{2,\mathcal{S}^*} = Y_{2,\mathcal{S}^*} - Y_{1,\mathcal{S}^*}$. Note that in general the combined statistic $Y_{k,\mathcal{S}^*}$ depends on the stage 2 sample sizes which may themselves be a function of the stage 1 statistics. Therefore we use a function to describe the combined statistic $Y_{k,\mathcal{S}^*} = \tau(y_{k,\forall j\in\mathcal{S}^*}, N_{k,\forall j\in\mathcal{S}^*})$. Since we concentrate on the two-stage adaptive enrichment design with two disjoint subgroups, $Y_{k,0}$, the combined statistic for the entire group, can be explicated as $Y_{k,0} = \tau(y_{k,1}, y_{k,2}, N_{k,1}, N_{k,2})$. For example, we construct score statistics in the Magnusson and Turnbull (2013) design, thereby the combined statistic is defined as $Y_{k,\mathcal{S}^*} = \sum_{j\in\mathcal{S}^*} Y_{k,j}$ and $X_{k,\mathcal{S}^*} = \sum_{j\in\mathcal{S}^*} X_{k,j}$. However, in Lin et al. (2021) design, the definition is different and we will show this in Section 5.3.1. Let u be the observed statistic value. We denote the set $\Omega_1^{obs} = \{(y_{1,1}, y_{1,2}) : [y_{1,1} \in (u, \infty), y_{1,2} \in \mathbb{R}]\}$ as the sample space for observed value given subgroup 1 is selected, $\Omega_2^{obs} = \{(y_{1,1}, y_{1,2}) : [y_{1,1} \in \mathbb{R}, y_{1,2} \in (u, \infty)]\}$ as the sample space for observed value given subgroup 2 is selected and $\Omega_0^{obs} = \{(y_{1,1}, y_{1,1}) : [\tau(y_{1,1}, y_{1,2}, N_{1,1}, N_{1,2}) \in (u, \infty)]\}$ as the observed value region given that the entire population is retained after the interim analysis. The maximum total sample size usually is determined by the power constraints. But in the second stage, the sample size depends on the first interim analysis results and the observed statistic values in the first stage. Therefore, we define the second stage sample size as $N_{2,\mathcal{S}^*} = \mathcal{N}_2(\mathbf{y}_1)$, where $\mathcal{N}_2(\cdot)$ is the second stage sample size function and $\mathbf{y}_1 = (y_{1,1}, y_{1,2})$.

We are supposed to test the null hypothesis $H_{0,\mathcal{S}} : \theta_{\mathcal{S}} = 0$ given that subpopulation $\mathcal{S}$ is selected at the first interim analysis. The one-sided alternative hypothesis is $H_{a,\mathcal{S}} = \theta_{\mathcal{S}} > 0$, which implies that the experimental treatment is effective for subpopulation $\mathcal{S}$. The following setups of the generalized approach are quite similar to what we proposed in Chapters 3 and 4. We have two arms in our trials: experimental and control. If the ith patient in subgroup j belongs to the experimental arm at stage k, his or her outcome $Z_{k,j,i}^E$ will follow a normal

distribution with the mean equals to μ_j^E and variance σ_0^2. On the other hand, if the ith patient is from subgroup j and assigned to the control arm, then the outcome follows the distribution $Z_{k,j,i}^C \sim N(\mu_j^C, \sigma_0^2)$. We denote the treatment effect for subgroup j as $\theta_j = \mu_j^E - \mu_j^C$. The overall treatment effect θ_0 will be a linear combination of the treatment effects for individual subgroups.

Let $f_{k|\mathcal{S}}(\cdot)$ and $F_{k|\mathcal{S}}$ be the conditional density and cumulative distribution functions of the cumulative score statistics for group $\mathcal{S}$ at stage k, where $k = \{1, 2\}$. Suppose that σ_0^2 is known, the p-value function for subgroup 1 can be defined as follows conditioning on the event that only subgroup 1 is selected at stage 1 by using the score ordering method:

$$\begin{aligned}\Pr(Y_{k,1} > u|\theta_1, \mathbf{y}_1 \in \Omega_2^c \cup \Omega_5^c) = &\int_{\Omega_1^{obs} \cap \Omega_5^c} f_{1|1}(y_{1,1}|\theta_1, \mathbf{y}_1 \in \Omega_2^c \cup \Omega_5^c)\mathrm{d}y_{1,1} \\ &+ \int_{\Omega_2^c} f_{1|1}(y_{1,1}|\theta_1, \mathbf{y}_1 \in \Omega_2^c) \\ &\times F_{2|1}(u|y_{1,1}, \theta_1)\mathrm{d}y_{1,1}.\end{aligned} \tag{5.1}$$

We can derive the conditional p-value function for subgroup 2 in a similar manner. For the overall group, we map the combined statistic $\tau(y_{1,1}, y_{1,2}, N_{1,1}, N_{1,2})$ to the new sample spaces Ω_4^* (i.e. $\Omega_4^* = \{(y_{1,1}, y_{1,2}) : \tau(y_{1,1}, y_{1,2}, N_{1,1}, N_{1,2}) \in \Omega_4^c\}$) and Ω_7^* (i.e. $\Omega_7^* = \{(y_{1,1}, y_{1,2}) : \tau(y_{1,1}, y_{1,2}, N_{1,1}, N_{1,2}) \in \Omega_7^c\}$). Below is the probability of observing a value more extreme than u in the entire population given that no subgroup is dropped in the interim analysis:

$$\begin{aligned}\mathbb{P}(Y_{k,0} > u|\theta_0, \mathbf{y}_1 \in \Omega_4^* \cup \Omega_7^*) = &\int_{\Omega_0^{obs} \cap \Omega_7^*} f_{1|0}(y_{1,0}|\theta_0, \mathbf{y}_1 \in \Omega_4^* \cup \Omega_7^*)\mathrm{d}y_{1,0} \\ &+ \int_{\Omega_4^*} f_{1|0}(y_{1,0}|\theta_0, \mathbf{y}_1 \in \Omega_4^*) \\ &\times F_{2|0}(u|y_{1,0}, \theta_0)\mathrm{d}y_{1,0}\end{aligned} \tag{5.2}$$

where $N_{2,0} = N_2$.

As we are testing the one-sided null hypothesis $H_{0,\mathcal{S}} : \theta_{\theta_\mathcal{S}} = 0$ given that subpopulation $\mathcal{S}$ is chosen at the interim analysis, the one-sided conditional confidence interval for selected subpopulation $\mathcal{S}$ will be $CI_{con} = (\hat{\theta}_\mathcal{S}, \infty)$, where $\hat{\theta}_\mathcal{S}$ is obtained by searching for possible zero roots of the equation $\Pr(Y_\mathcal{S} > u|\theta_\mathcal{S}, \mathcal{S}^* = \mathcal{S}, \boldsymbol{y}_1) = \alpha$. $1 - \alpha$ is the coverage probability we specified beforehand. As for the two-sided test, such as testing $H_{0,\mathcal{S}} = 0$ vs. $H_{a,\mathcal{S}} \neq 0$, the two-sided conditional confidence interval construction approach is quite similar to the one-sided case. Because the conditional p-value function is monotonically increasing on the treatment effect, we search for $\hat{\theta}_{1-\alpha/2,\mathcal{S}}$ and $\hat{\theta}_{\alpha/2,\mathcal{S}}$ that satisfy $\Pr(Y_\mathcal{S} > u|\theta_{1-\alpha/2,\mathcal{S}}, \mathcal{S}^* = \mathcal{S}, \boldsymbol{y}_1) = 1 - \alpha/2$ and $\Pr(Y_\mathcal{S} > u|\theta_{\alpha/2,\mathcal{S}}, \mathcal{S}^* = \mathcal{S}, \boldsymbol{y}_1) = \alpha/2$ correspondingly. Then the two-sided confidence interval will be $(\hat{\theta}_{\alpha/2,\mathcal{S}}\hat{\theta}_{1-\alpha/2,\mathcal{S}})$.

5.2.2 Unconditional confidence intervals for individual subgroups

Recall that the term "unconditional" here refers to the case that we construct confidence intervals for an individual subgroup regardless of the interim analysis result. If the target subgroup is chosen, we condition on the statistics observed in other selected subgroups to construct the p-value function. Otherwise, if the subgroup is dropped, we use all the information we observed from selected subgroups. For the individual p-value function for subgroup j, we need to adjust decision boundaries based on the statistic values of the rest of chosen subgroups $\{Y_{1,\mathcal{P}^\dagger} : \mathcal{P}^\dagger = \mathcal{S}^* \backslash j\}$. Let $\mathcal{P}_j$ be the index of the subset that subgroup j is included (i.e. $\mathcal{P}_1 = \{0, 1\}$). When constructing a confidence interval for subgroup j, Ω_0 will be re-partitioned in an adjusted way for the two-stage two-subgroup design:

$$\Omega_1^{uc} = \{\text{subgroup } j \text{ stops for futility at stage 1}\},$$

$$\Omega_2^{uc} = \{\text{only enrich subgroup } j \text{ at stage 2}\},$$

$$\Omega_3^{uc} = \{\text{the overall group is proceeded to stage 2}\},$$

$$\Omega_4^{uc} = \{\text{subgroup } j \text{ stops for efficacy at stage 1}\}.$$

Note that the above sets also depend on the statistics observed in other selected subgroups. Let $f_{k,j}^{uc}(\cdot)$ and $F_{k,j}^{uc}(\cdot)$ be the unconditional density and cumulative distribution function for subgroup j at stage k. We can write the general form of subgroup j's individual p-value function as below:

$$\begin{aligned} \Pr[Y_j > u|\theta_1, \mathbf{y}_{1,\mathcal{P}^\dagger}] = &\int_{\Omega_1^{obs} \cap (\Omega_4^{uc} \cup \Omega_1^{uc})} f_{1,j}^{uc}(y_{1,j}|\theta_j, \mathbf{y}_{1,\mathcal{P}^\dagger}) \mathrm{d}y_{1,j} \\ &+ \int_{\Omega_2^{uc} \cup \Omega_3^{uc}} f_{1,j}^{uc}(y_{1,j}|\theta_j, \mathbf{y}_{1,\mathcal{P}^\dagger}) \\ &\times F_{2,j}^{uc}(u|y_{1,j}, \mathbf{y}_{1,\mathcal{P}^\dagger}, \theta_j) \mathrm{d}y_{1,j} \end{aligned} \tag{5.3}$$

where again, Ω_{obs} is the sample space for which the statistic for group j exceeds the value, u, observed. In general, there are two possible cases where subgroup j is involved after the interim analysis: exits at stage 1 or proceeds to stage 2. The first term in Equation (5.3) describes the case where subgroup j exits in the first stage. The second line describes the probability of observing the statistic $y_{2,j}$ greater than u when subgroup j proceeds to the second stage for cases where only subgroup j or the entire population is enriched. However, the cumulative distribution function might change when the interim decision reveals that the overall subgroup selected instead of only subgroup j is chosen. Therefore, in order to take all possible selection

results into account, the second term probably will expand to several terms to correspond to different results.

Since the unconditional p-value functions are monotonically increasing according to θ_j and we are testing one-sided hypotheses, the confidence interval for group j will be $(\hat{\theta}_j, \infty)$ and $\hat{\theta}_j$ satisfy $\{\hat{\theta}_j : \Pr(Y_j > u|\boldsymbol{\theta}) = \alpha_{SCI}\}$ that ensures an exact coverage probability at level $1 - \alpha_{SCI}$. Let $p_j(\theta_j) = \Pr[Y_j > u|\theta_1, \mathbf{y}_{1,\mathcal{P}^\dagger}]$ be the individual p-value function. By plugging in the individual p-value function into a specific multiple-testing procedure, we are able to construct simultaneous confidence intervals for all subgroups at the end of the trial.

5.3 Example: Sample size re-estimated adaptive enrichment design

In the subsequent sections, we are going to use the above adaptive enrichment design which re-estimates the sample size in the second stage to illustrate that our confidence interval construction approach could be generally applied in various types of enrichment designs.

5.3.1 Conditional confidence intervals in the Lin et al. design

As the confidence interval is obtained by inverting the corresponding p-value function, first of all, we derive the probability of observing a statistic greater than u, the observed threshold. Let $\boldsymbol{y} = (y_{1,1}, y_{1,2})$ and $\mu_j^C = 0$ in the entire population. According to the statistic distribution we stated above, the density function and cumulative distribution function in the Equation (5.1) can be specified as below:

$$f_{k|\mathcal{S}^*}(y_{1,j}|\theta_j) = \frac{\psi(y_{1,j} - \theta_j/\sqrt{4\sigma^2/N_{1,j}})}{1 - \Phi\left(l_1 - \theta_j/\sqrt{4\sigma^2/N_{1,j}}\right)},$$

$$F_{2|j}(u|y_{1,j}, \theta_j) = 1 - \Phi\left(\frac{u - \sqrt{\omega}y_{1,j}}{\sqrt{1-\omega}} - \frac{\sqrt{N_{2,j}}\theta_j}{\sqrt{4\sigma^2}}\right).$$

Under the score ordering method, the probability of observing a statistic greater than u conditioning on the event that only subgroup j is selected can be written as:

$$\begin{aligned} Pr(Y_j > u|\theta_j, \mathcal{S}^* = \{j\}, \boldsymbol{y}_1) = & \int_{\max(u,u_1)}^{\infty} \frac{\psi(y_{1,j} - \theta_j/\sqrt{4\sigma^2/N_{1,j}})}{1 - \Phi\left(l_1 - \theta_j/\sqrt{4\sigma^2/N_{1,j}}\right)} \mathrm{d}y_{1,j} \\ & + \int_{l_1}^{u_1} \frac{\psi(y_{1,j} - \theta_j/\sqrt{4\sigma^2/N_{1,j}})}{1 - \Phi\left(l_1 - \theta_j/\sqrt{4\sigma^2/N_{1,j}}\right)} \\ & \times \left[1 - \Phi\left(\frac{u - \sqrt{\omega}y_{1,j}}{\sqrt{1-\omega}} - \frac{\sqrt{N_{2,j}}\theta_j}{\sqrt{4\sigma^2}}\right)\right] \mathrm{d}y_{1,j} \end{aligned} \tag{5.4}$$

where $\psi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions respectively, and $N_{2,j} = \mathcal{N}_2(y_{1,1}, y_{1,2})$.

When the entire population is selected in the first interim analysis, we consider two possible occasions. Since only when $y_{1,g_{\min}} > u_1$ and $l_1 < y_{1,g_{\min}} \leq y_{1,g_{\max}} < u_1$ the overall group will either be claimed to have treatment efficacy or enriched in the second stage, the conditional p-value function for the overall subgroup can be separated into two components:

$$\begin{aligned} & Pr(Y_0 > u|\boldsymbol{\theta}, \mathcal{S}^* = \{1,2\}, \boldsymbol{y}) \\ & = \frac{\Pr(y_{1,0} > u, y_{1,g_{\min}} > u_1|\boldsymbol{\theta})}{\Pr(y_{1,g_{\min}} > u_1|\boldsymbol{\theta}) + \Pr(l_1 < y_{1,g_{\min}} \leq y_{1,g_{\max}} < u_1|\boldsymbol{\theta})} \\ & + \frac{\Pr(y_{2,0} > u, l_1 < y_{1,g_{\min}} \leq y_{1,g_{\max}} < u_1|\boldsymbol{\theta})}{\Pr(y_{1,g_{\min}} > u_1|\boldsymbol{\theta}) + \Pr(l_1 < y_{1,g_{\min}} \leq y_{1,g_{\max}} < u_1|\boldsymbol{\theta})}. \end{aligned} \tag{5.5}$$

The denominator in the above equation corresponds to the probability of observing the first stage statistics belonging to $\Omega_4^* \cup \Omega_7^*$. Moreover, according to the distribution of $y_{1,1}$ and $y_{1,2}$ we defined in Section 5.3, the overall statistic combination function can be explicated as $y_{1,0} = \tau(y_{1,1}, y_{1,2}, N_{1,2}, N_{1,2}) = \sqrt{\rho_1}y_{1,1} + \sqrt{\rho_2}y_{1,2}$ with $\rho_1 = N_{1,1}/N_1$ and $\rho_2 = N_{1,2}/N_1$.

For the second component, since the overall group proceeds to the second stage only when both subgroup statistics locate between boundaries at stage 1, the density function and the cumulative distribution function for the overall group in the first stage can be explicated as:

$$\begin{aligned} & f_{1|\mathcal{S}^*=0}(y_{1,0}|\boldsymbol{\theta}) \\ & = \int_{\max(l_1, \frac{y_{1,0} - \sqrt{1-\rho}l_1}{\sqrt{\rho}})}^{\min(u_1, \frac{y_{1,0} - \sqrt{1-\rho}l_1}{\sqrt{\rho}})} \frac{\psi(y_{1,1} - \theta_1/\sqrt{4\sigma^2/N_{1,1}})}{\Phi(u_1 - \theta_1/\sqrt{4\sigma^2/N_{1,1}}) - \Phi(l_1 - \theta_1/\sqrt{4\sigma^2/N_{1,1}})} \\ & \times \frac{1}{\sqrt{1-\rho}} \frac{\psi(\frac{y_{1,0} - \sqrt{\rho}y_1^1}{\sqrt{1-\rho}} - \theta_2/\sqrt{4\sigma^2/N_{1,2}})}{\Phi(u_1 - \theta_2/\sqrt{4\sigma^2/N_{1,2}}) - \Phi(l_1 - \theta_2/\sqrt{4\sigma^2/N_{1,2}})} \mathrm{d}y_{1,1} \end{aligned} \tag{5.6}$$

and

$$
\begin{aligned}
&F_{2|\mathcal{S}^*=0}(y_{1,0}|\boldsymbol{\theta}) \\
=&1-\Phi\left(\frac{u-\sqrt{\omega}y_{1,0}}{\sqrt{1-\omega}}-\frac{\sqrt{N_{2,0}}\theta_0}{\sqrt{4\sigma^2}}\right)
\end{aligned} \tag{5.7}
$$

By plugging in Equation (5.6) and (5.7), terms in Equation (5.5) can be expressed as following:

$$
\begin{aligned}
&\Pr(y_{1,0}>u, y_{1,g_{\min}}>u_1|\boldsymbol{\theta}) \\
=&\int_{\max(u,u_1(\sqrt{\rho}+\sqrt{1-\rho}))}^{\infty}\frac{\psi(y_{1,0}-\theta_1/\sqrt{4\sigma^2/N_{1,1}})}{1-\Phi(u_1-\theta_1/\sqrt{4\sigma^2/N_{1,1}})} \\
\times&\frac{1-\Phi(\frac{u-\sqrt{\rho}t_1}{\sqrt{1-\rho}}-\theta_2/\sqrt{4\sigma^2/N_{1,2}})}{1-\Phi(u_1-\theta_2/\sqrt{4\sigma^2/N_{1,2}})}\mathrm{d}y_{1,0} \\
\times&\Pr(y_{1,g_{\min}}>u_1|\boldsymbol{\theta});
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(y_{2,0}>u, l_1<y_{1,g_{\min}}\leq y_{1,g_{\max}}<u_1|\boldsymbol{\theta}) \\
=&\int_{l_1(\sqrt{\rho}+\sqrt{1-\rho})}^{u_1(\sqrt{\rho}+\sqrt{1-\rho})} f_{1|\mathcal{S}^*=0}(y_{1,0}|\boldsymbol{\theta})\times\left(1-\Phi\left(\frac{u-\sqrt{\omega}y_{1,0}}{\sqrt{1-\omega}}-\frac{\sqrt{N_{2,0}}\theta_0}{\sqrt{4\sigma^2}}\right)\right)\mathrm{d}y_{1,0} \\
\times&\Pr(l_1<y_{1,g_{\min}}\leq y_{g_{1,\max}}<u_1|\boldsymbol{\theta});
\end{aligned}
$$

where

$$
\begin{aligned}
&\Pr(y_{1,g_{\min}}>u_1|\boldsymbol{\theta}) \\
=&(1-\Phi(u_1-\theta_1/\sqrt{4\sigma^2/N_{1,1}}))\times(1-\Phi(u_1-\theta_2/\sqrt{4\sigma^2/N_{1,2}})),
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(l_1<y_{1,g_{\min}}\leq y_{1,g_{\max}}<u_1|\boldsymbol{\theta}) \\
=&(\Phi(u_1-\theta_1/\sqrt{4\sigma^2/N_{1,1}})-\Phi(l_1-\theta_1/\sqrt{4\sigma^2/N_{1,1}})) \\
\times&(\Phi(u_1-\theta_2/\sqrt{4\sigma^2/N_{1,2}})-\Phi(l_1-\theta_2/\sqrt{4\sigma^2/N_{1,2}})).
\end{aligned}
$$

Let the nominal coverage probability level be fixed at $1-\alpha$. Recall that we are testing the one-sided hypotheses:$H_{0,\mathcal{S}}=\theta_{\mathcal{S}}=0$ versus $H_{a,\mathcal{S}}:\theta_{\mathcal{S}}>0$ conditioning on subset $\mathcal{S}$ is chosen. By inverting $\Pr(Y_{\mathcal{S}}|\hat{\theta}_{\mathcal{S}},\mathcal{S}^*=\mathcal{S},\boldsymbol{y}_1)=1-\alpha$ based on the observed selection results and Wald statistic values in the trials, we get the one-sided confidence intervals $CI_{lin,con}=(\hat{\theta}_{\mathcal{S}},\infty)$.

5.3.2 Numerical study

In this section, we are going to run 10,000 simulated trials to test the performance of the confidence intervals under various scenarios. Given that the prevalence of subgroup 1 and subgroup 2 are $\rho_1 = \rho_2 = 0.5$, we randomize the number of patients allocated to each subgroup by utilizing the binomial distribution with probability equaling 0.5. We also assume that each patient has an equivalent probability of being assigned to the experimental arm and control arm. A total of 620 patients are enrolled at stage 1. The futility and efficacy stopping boundaries in the first stage are 1.036 and 2.212 respectively. Sample sizes in the second stage are updated based on the value of test statistics observed in the first stage.

We present two of the simulated trials in Table 5.1 to illustrate how the Lin et al. (2021) design works. In both trials, the statistic of subgroup 1 exceeds l_1 but the statistic of subgroup 2 does not, therefore only subgroup 1 will be enriched in the second stage. What stands out is that the statistic of subgroup 1 in Trial 1 is smaller than the statistic observed in Trial 2 while the sample size required in the second stage is much larger for Trial 1 than Trial 2. This is because the design aims to preserve a constant conditional power and when the statistics observed in the first stage are closer to the lower boundary, we need more information in the second stage to maintain the power of rejection.

Table 5.1: Two worked examples for the Lin et al. design with only subgroup 1 is chosen.

	Trial 1		Trial 2	
stage 1	$X_{1,j}$	$N_{1,j}$	$X_{1,j}$	$N_{1,j}$
Ω_1	1.3261	319	1.5916	304
Ω_2	0.8416	301	-1.1963	316
stage 2	$X_{2,j}$	$N_{2,j}$	$X_{2,j}$	$N_{2,j}$
Ω_1	2.2498	1571	2.3842	479
Ω_2	\	\	\	\
Total	$Y_{2,\mathcal{S}}$	$N_{\mathcal{S}}$	$Y_{2,\mathcal{S}}$	$N_{\mathcal{S}}$
Ω_1	3.5759	1890	3.9759	783

Also, given that only subgroup 1 is chosen in the first interim analysis, we present the distribution of lower bounds obtained under three different scenarios in Figure 5.1 based on 10,000 simulated trials in each case. All lower bounds in the light blue histogram are calculated from the null scenario. The light pink and light green histogram present lower bounds from scenario

$\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ respectively. The red vertical line is the 97.5% quantile of the distribution. Obviously, our p-value inversion approach still has nominal coverage probabilities in Lin et al. design as the quantile lines locate around the true treatment effect in all scenarios.

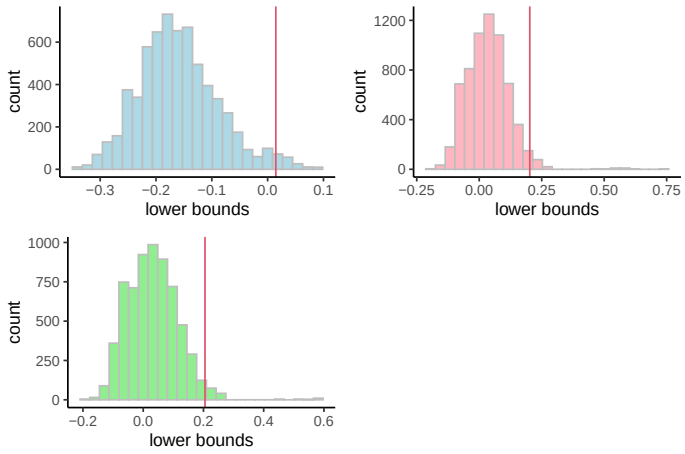


Figure 5.1: Distribution of lower bounds under scenario $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ given subgroup 1 is selected.

Our simulation also contains the case where the entire population is retained at the interim analysis. However, we again need to address the limitation of our approach that is we are not able to obtain conditional confidence intervals for the overall group that account for the possibility that the treatment effect is heterogeneous among the disjoint subgroups. Although, we still run simulations under scenario $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$, unsurprisingly, the performance under the second scenario is poor. Figure 5.2 provides the distribution of lower bounds for the overall group conditioning on the event that it is selected. The red vertical lines are the 97.5% quantile. It is apparent from the figure that the 97.5% percentile of the lower bounds lies below the true treatment effect which implies that our approach has a coverage probability that is close to the nominal level for scenario $\boldsymbol{\theta} = (0, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$.

We present coverage probabilities and powers for selected subpopulations in Table 5.2. In cases where only subgroup 1 is selected, again we find that both the coverage probabilities and powers are close to the nominal level. However, when the entire population is chosen in the interim analysis, we notice that the coverage probability is conservative under the null scenario. This is due to the fact that the number of trials in which both subgroups are selected is small (236 out of 10,000 trials). Under the scenario where the true treatment effects are heterogeneous among subgroups, our p-value function assumes that treatment effects are homogeneous and this

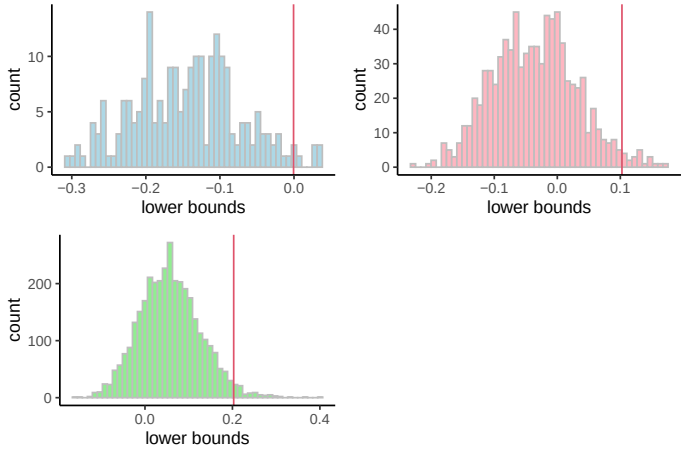


Figure 5.2: Distribution of lower bounds under scenario $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$ given the entire population is selected.

also leads to the conservation of coverage probability and poor power (as shown in Figure 5.2).

Table 5.2: The probability of covering the true treatment effect (coverage probability) and rejecting at least one null hypothesis (power) given the subgroup 1 and the overall group is chosen.

scenario	$S = \{1\}$		$S = \{1, 2\}$	
	coverage probability	power	coverage probability	power
$\boldsymbol{\theta} = (0, 0)$	0.9667	0.0333	0.9775	0.0225
$\boldsymbol{\theta} = (0.2, 0)$	0.9739	0.6400	1	0.2936
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9710	0.6313	0.9740	0.7946

5.3.3 Simultaneous confidence intervals in the Lin et al. design

Recall that the outcome of each patient recruited follows the normal distribution, thereby Lin et al. (2021) proposes to construct a Wald statistic,

$$X_{k,j} = \frac{\hat{\mu}_j^E - \hat{\mu}_j^C}{\sqrt{4\sigma^2/N_{k,j}}} \tag{5.8}$$

where $\hat{\mu}_j^E$ and $\hat{\mu}_j^C$ are the sample mean for subgroup j at stage k. Apparently, Equation (5.8) is normally distributed with mean $(\theta_j - \theta_0)/\sqrt{4\sigma^2/N_{k,j}}$ and variance 1. Note that σ^2 is the pooled sample variance.

For the unconditional p-value function for an individual subgroup j, the probability depends on statistics of subpopulation $\mathcal{P}^{\dagger}$. Let $l_1^{\star} = l_1 \times I(y_{1,\mathcal{P}^{\dagger}} < u_1) + u_1 \times I(y_{1,\mathcal{P}^{\dagger}} \geq u_1)$, the probability of observing a statistic greater than u for an individual subgroup j is

$$
\begin{aligned}
&\Pr(Y_j > u | \theta_j, y_{1,\mathcal{P}^{\dagger}}) \\
=&[1 - \Phi(\max(u, u_1) - \frac{\theta_j}{\sqrt{4\sigma^2/N_{1,j}}}] \\
&+\Phi(l_1^{\star} - \frac{\theta_j}{\sqrt{4\sigma^2/N_{1,j}}}) - \Phi(\min(u, l_1^{\star}) - \frac{\theta_j}{\sqrt{4\sigma^2/N_{1,j}}}) \\
&+ \int_{l_1}^{u_1} \psi(y_{1,j} - \frac{\theta_j}{\sqrt{4\sigma^2/N_{1,j}}})[1 - \Phi(\frac{u - \sqrt{\omega} y_{1,j}}{\sqrt{1-\omega}} - \frac{\sqrt{N_{2,j}}\theta_j}{4\sigma^2})]\mathrm{d}y_{1,j} \\
&\times I(y_{1,\mathcal{P}^{\dagger}} < l_1) \\
&+ \int_{l_1}^{u_1} \psi(y_{1,j} - \frac{\theta_j}{\sqrt{4\sigma^2/N_{1,j}}})[1 - \Phi(\frac{u - \sqrt{\omega} y_{1,j}}{\sqrt{1-\omega}} - \frac{\sqrt{N'_{2,j}}\theta_j}{4\sigma^2})]\mathrm{d}y_{1,j} \\
&\times I(l_1 < y_{1,\mathcal{P}^{\dagger}} < u_1)
\end{aligned}
\tag{5.9}
$$

where $I(\cdot)$ is the indicator function and $\Phi(\cdot)$ is the cumulative distribution function for standard normal distribution. Compared to the unconditional p-value function we derived for Magnusson and Turnbull design, we utilize the information of the remaining subgroup in the selected subpopulation in a quite different way. We divide the p-value function into three parts; (1) subgroup j stops at stage 1 (line 1 and line 2 in Equation (5.9)); (2) subgroup j terminates at stage 2 but only subgroup j is enriched (line 3 and line 4); (3) the overall group is enriched in the second stage (line 5 and line 6). This is owing to the fact that the sample re-estimation in the second stage depends on the statistic we observed in the first stage. Since the statistics are different in cases where only subgroup j is chosen and the overall group is selected, the number of patients we are required to recruit varies. Therefore, the term $N_{2,j}$ in line 3 and $N'_{2,j}$ in line 5 might have different values according to the different statistics observed at stage 1.

In Chapter 4, we discussed three multiple-test procedures to construct simultaneous confidence intervals for Magnusson and Turnbull design and concluded that the classic Bonferroni correction approach is more appropriate than the two others. We recommended the classic Bonferroni test due to the fact that although the classic Bonferroni approach rejects fewer null hypotheses, it provides the most information about the scale of the true treatment effect, and the testing procedure in Magnusson and Turnbull design itself already made decisions about whether accept the null hypotheses or not. Here, we again construct simultaneous confidence intervals based on those three multiple test procedures and check whether the same conclusion holds in

the Lin et al. design via numerical studies we present in Section 5.3.4.

5.3.4 Numerical study

In this section, we compare three multiple testing procedures: the classic Bonferroni, Bonferroni Holm and the new parameter-dependent weighted Bonferroni approach proposed by Brannath and Schmidt (2014) applied in Lin et al. design. Again, we focus on the two-stage Lin et al. design with two subgroups. Being similar to the numerical study we performed in the conditional confidence interval session, we run simulations under three scenarios: $\boldsymbol{\theta} = (0, 0)$, $\boldsymbol{\theta} = (0.2, 0)$ and $\boldsymbol{\theta} = (0.2, 0.2)$. In the first stage, we recruit 620 patients in total and randomly assign them to two subgroups by utilizing binominal distribution with $\rho_1 = \rho_2 = 0.5$. Also, each patient is equally likely to be involved in the experimental arm or the control arm (i.e. we adopt block randomization here). According to the design procedure we described in Section 5.3, the sample size in the second stage depends on the statistic value we observed in the first stage.

In Table 5.3 we again find that both the Bonferroni-Holm procedure and Brannath and Schmidt procedure are more conservative than the classic Bonferroni procedure as their coverage probabilities are greater than the classic Bonferroni's, especially under the non-null scenarios. However, the Bonferroni-Holm test and the Brannath and Schmidt test detect more rejections. These conclusions all agree with what we found for the confidence intervals for the Magnusson and Turnbull design in Chapter 4. In Figure 5.3, we display the distributions of the lower bounds in the simultaneous confidence intervals based on the three multiple testing procedures. We assume that the true treatment effects are 0.2 for subgroup 1 and 0 for subgroup 2. The light blue histograms represent the distribution of lower bounds for subgroup 1 and the light pink histograms represent the distribution of lower bounds for subgroup 2. In the first row, all lower bounds are obtained by the classic Bonferroni test. We notice that around 98.75% of them (i.e. the red vertical line) locate below 0.2 and 0 for subgroup 1 and subgroup 2 respectively, which again implies that the one-sided classic Bonferroni simultaneous confidence intervals have coverage probabilities close to the nominal level. The lower bounds in the second row are obtained by utilizing the Bonferroni-Holm procedure. We find that a large proportion of lower bounds are 0 for subgroup 1. This is due to the fact that when only part of the null hypotheses is rejected in a single trial, the lower bounds of the rejected hypotheses are defined to be 0 in the Holm test procedure. As for the lower bounds derived from the Brannath and Schmidt procedure, we observe that their distribution is quite similar to the classic Bonferroni procedure except that Brannath and Schmidt procedure rejects slightly more null hypotheses than the classic Bonferroni

test.

Again, the simulation results above indicate that the Bonferroni-Holm procedure is very powerful in rejecting null hypotheses but worse in providing information about the scale of the true treatment effect. On the contrary, the classic Bonferroni procedure gives the most information but is relatively weak in detecting false hypotheses. Brannath and Schmidt (2014) procedure trades off between informativeness and the power of rejecting non-null hypotheses, however, it still sacrifices informativeness of the positive lower bounds to detect more non-null hypotheses. As Lin et al. design already determined to reject which hypothesis, we prefer to use the multiple test procedure that provides more information to construct simultaneous confidence intervals for individual subgroups. Therefore, the classic Bonferroni procedure might be the most appropriate approach in Lin et al. design's case.

Table 5.3: The coverage probability, power of rejecting the null hypotheses, and the average number of rejections in each trial for different scenarios.

scenario	classic Bonferroni			Holm			Brannath and Schmidt		
	coverage probability	power	average rejection	coverage probability	power	average rejection	coverage probability	power	average rejection
$\boldsymbol{\theta} = (0, 0)$	0.9703	0.0297	0.03	0.9703	0.0297	0.0302	0.9703	0.0297	0.03
$\boldsymbol{\theta} = (0.2, 0)$	0.9737	0.668	0.6728	0.9799	0.668	0.6819	0.9747	0.668	0.6734
$\boldsymbol{\theta} = (0.2, 0.2)$	0.9732	0.8424	1.0325	0.9839	0.8425	1.1474	0.9747	0.8425	1.0383

5.3.5 Comparison between Magnussona-Turnbull design and Lin et al. design

In this section, we will compare the confidence intervals obtained for the Magnussona and Turnbull design to the Lin et al. design in situations where the data collected is the same. First of all, we assume that the prevalences of subgroup 1 and subgroup 2 are 0.6 and 0.4 respectively. Note that whereas in the Magnusson and Turnbull design, the stagewise sample sizes are fixed, Lin et al. (2021) set the stage 2 sample size to maintain a level of conditional power. Specifically, recall that Lin et al. (2021) proposed a conditional power function for selected subset $\mathcal{S}$

$$\Pr(Y_{2,\mathcal{S}} > y|\theta_{\mathcal{S}}, y_{1,\mathcal{S}}) = 1 - \Phi\left(\frac{y - \sqrt{\omega}y_{1,\mathcal{S}}}{\sqrt{1-\omega}} - \frac{\theta_{\mathcal{S}}}{\sqrt{4\sigma^2/N_2}}\right) \tag{5.10}$$

to calculate the sample size in the second stage. By setting $\Pr(Y_{2,\mathcal{S}} > y|\theta_{\mathcal{S}}, y_{1,\mathcal{S}}) = 1 - \beta_{\mathcal{S}}$ (i.e. $1 - \beta_{\mathcal{S}}$ is the required conditional power level), we can get a unique solution for N_2. Note that

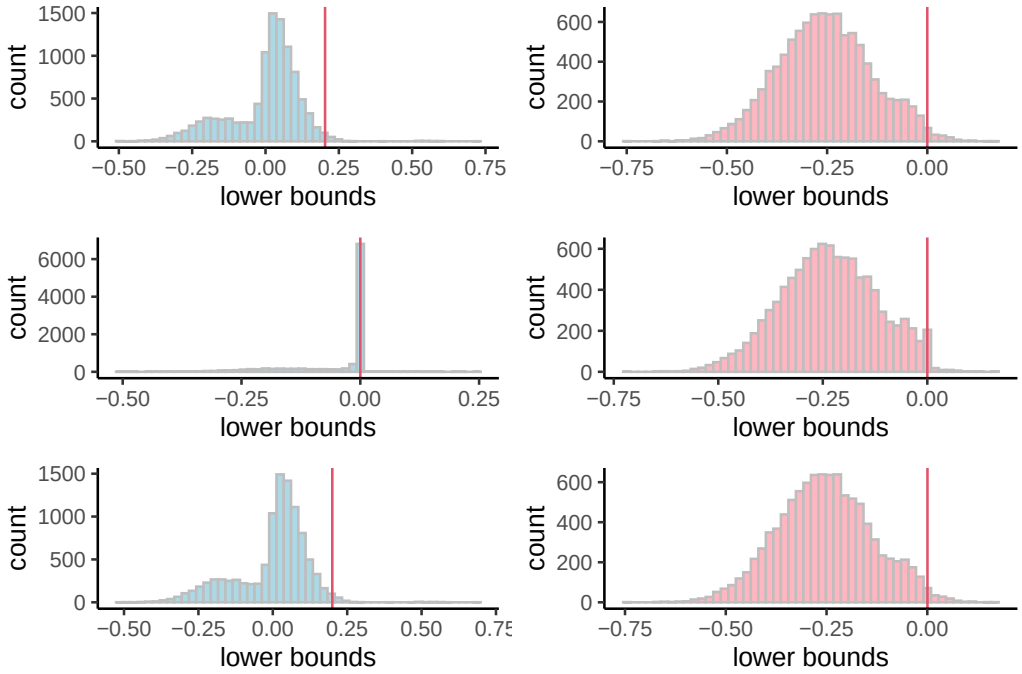


Figure 5.3: Distribution of lower bounds of the classic Bonferroni (top row), Bonferroni-Holm (middle row) and Brannath and Schmidt (bottom row) simultaneous confidence intervals under scenario $\boldsymbol{\theta} = (0.2, 0)$.

we replace $\theta_{\mathcal{S}}$ in Equation (5.10) by the maximum likelihood estimate $\hat{\theta}_{\mathcal{S}} = y_{1,\mathcal{S}}/\sqrt{4\sigma^2/N_{1,\mathcal{S}}}$ when computing the conditional power. In the simulation study of the Magnusson and Turnbull design, we assume that the sample sizes in the first and second stages are fixed to be 625. To compare these two designs, we calculate the corresponding first stage statistics in the Lin et al. design by setting N_2 to 625 in Equation (5.10) and solving for $y_{1,\mathcal{S}}$, which yields a value of 1.8235. Hence, we can directly compare results from the Magnusson and Turnbull's design with Lin et al.'s design in the case where the stage 1 statistic is $y_{1,\mathcal{S}} = 1.8235$ since both designs will yield a stage 2 sample size of $N_2 = 625$. Suppose that only subgroup 1 is chosen at the first interim analysis, the lower bounds for conditional confidence intervals in the two designs are presented in Table 5.4 and Figure 5.4 where we vary the observed data from subgroup 1 at stage 2. It is evident from the results that the lower bounds for the Magnusson and Turnbull design are lower than the lower bounds of the Lin et al. design when the trial intends to terminate for futility (i.e. when $X_{2,1}$ is negative). Nonetheless, in case the trial is expected to progress to the second stage or discontinued due to effectiveness, the confidence intervals resulting from Magnusson and Turnbull design have a smaller range than those acquired from Lin et al. design. At the point when the standard statistic increment is around 1.8, it is clear that the lower limit slopes rise. Based on the data presented in Table 5.4, we observe that at the transition point, the standardized accumulated statistic reaches the first stage upper boundary for both designs.

Table 5.4: The lower bounds of the one-sided confidence interval given subgroup 1 is selected in the Magnusson and Turnbull and Lin et al. design.

		lower bounds of the one-sided conditional confidence interval	
$X_{2,1}$(standardized)	Y_1 (standardized)	Magnusson and Turnbull $((l_1, u_1) = (0.5192, 2.5529))$	Lin et al. $((l_1, u_1) = (1.0364, 2.5197))$
-3	-1.2551	-0.3629	-0.2904
0	1.1166	-0.1420	0.1455
1.8	2.5396	-0.0471	-0.0514
2.1	2.7768	-0.0082	-0.0109
2.2	2.8559	0.0030	0
5	5.0694	0.2511	0.1687

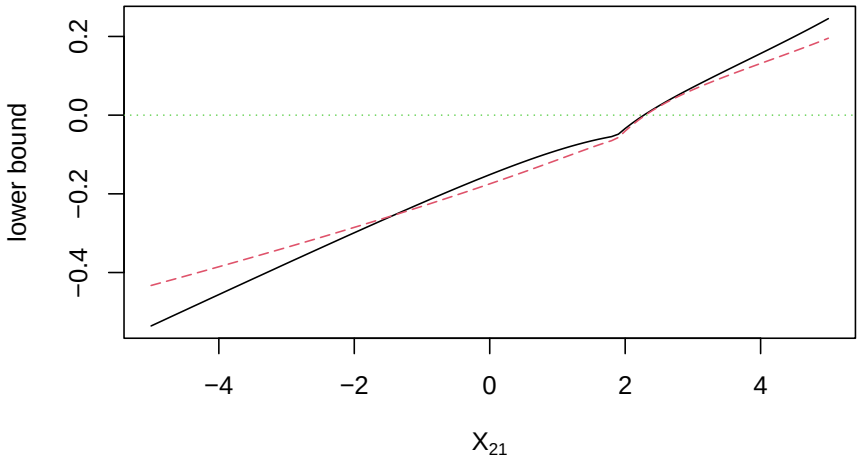


Figure 5.4: The lower bounds for Magnusson and Turnbull design are depicted by the black solid line, whereas the red dotted line represents the lower bounds for Lin et al. design. Meanwhile, the green horizontal line represents that the lower bound equals 0. X_{21} is the standardized statistic increment at stage 2.

5.4 Conclusion

A main factor that distinguishes various types of adaptive enrichment design is the specific decision rules utilized by the conductor. As long as the decision rule in the interim analysis is fixed in advance, we are able to partition the entire sample space in the first stage into a few disjoint subspaces. By integrating over those subspaces, we can obtain p-values conditioning on certain subpopulations chosen or for an individual subgroup. Then we search for a unique solution for the true treatment effect that makes the p-value function equal to the significance level. Since we assume that we are testing the one-sided null hypothesis, the unique solution will be the lower bound of the conditional or unconditional confidence interval. Note that the way we divide the entire sample space for the conditional confidence interval might be different from the way we partition it for the unconditional confidence interval. Furthermore, although we focus on the two-stage two-subgroup design in this chapter, the generalized method can be extended to the k-stage enrichment design with m subgroups where $k \geq 2$ and $m \geq 2$.

When constructing conditional confidence intervals for the Magnusson and Turnbull (2013) design, we adopted three sample space ordering approaches: stagewise, score, and MLE. The score ordering method outperforms the other two approaches as it treats all stages equally. Hence when generalizing the method of constructing confidence intervals given the selection result, we merely utilize the score ordering method. If desired, it would be relatively straightforward to adapt the procedure to MLE ordering by changing the definitions of f and F. The conditional p-value functions given a single subgroup is selected are quite similar across different designs as the only difference is the density function for the single subgroup statistics. However, if more than two subgroups are chosen, the conditional p-value function will rely on both the decision rule and the statistic combination rule of the specific design. When treatment effects vary among subgroups, constructing a conditional p-value function using multiple chosen subgroups becomes more intricate because it requires determining the joint distribution and correlation among the subgroups.

Chapter 6

Discussion

Based on the p-value function inversion approach and through appropriate conditioning on stage statistics, we are able to construct confidence intervals for selected subsets along with individual subgroups in a more direct way. Magnusson and Turnbull (2013) proposed an iterative bootstrap method for the purpose of constructing confidence intervals, however, the coverage probabilities of the resulting intervals are often poor. Compared to the bootstrap method, our method has closer to nominal coverage probabilities while controlling the FWER at the desired level. At the end of the trial, we are supposed to construct the simultaneous confidence intervals for all subgroups ensuring the simultaneous coverage is close to the nominal level. In this thesis, we employed the classic Bonferroni, Bonferroni-Holm and parameter-dependent weighted Bonferroni procedures for the simultaneous confidence interval construction. Numerical studies show that the Bonferroni simultaneous confidence intervals are informative but at the expense of some power to reject false null hypotheses. Conversely, the Bonferroni-Holm simultaneous confidence intervals are most powerful in detecting rejected hypotheses but provide no information for rejected subgroups when only part of the hypotheses are rejected. The parameter-dependent weighted Bonferroni procedure obtained through the procedure proposed by Brannath and Schmidt (2014) trades off the informativeness and the power of rejecting null hypotheses and it can asymptotically approach either the classic Bonferroni procedure and the Bonferroni-Holm procedure by modifying the penalizing function. Nonetheless, our simulation studies show that the gain in power from the weighted Bonferroni intervals is relatively small in the usual situation where there are two or three subgroups, whilst they lead to intervals that have conservative coverage when the null hypothesis is false. Furthermore, as we can observe the decision of each subgroup (i.e. stop for

futility or exit for efficacy) at the termination of the trial, we prioritize the informativeness and consider the classic Bonferroni procedure most appropriate.

Our method has a limitation in that there is no guarantee of agreement between the rejection or acceptance of a hypothesis in the testing procedure proposed in the design, and the inclusion or exclusion of the corresponding null value to the constructed confidence interval. This usually occurs when the one-sided confidence interval includes 0 but the corresponding statistic observed crosses the upper boundary. This is due to the selection made during the interim analysis. The design's decisions are based on all data observed in the trial, while the conditional confidence intervals mostly rely on data from the retained subgroups. To address this inconsistency between design decisions and confidence intervals for Magnusson and Turnbull's design, we proposed a new ordering method that pulls the adjusted boundaries of each stage to the same level by transforming the cumulative score statistic at stage 2 by an appropriate power of the Fisher Information. However, this method requires adjusting the significance level based on the selection result, making it difficult to extend to other designs.

When constructing a confidence interval for the combined treatment effect of all selected subgroups, we assume homogeneity of treatment effects across all selected subgroups. However, when only one of the two subgroups had a true treatment effect, but both were chosen (i.e. $\boldsymbol{\theta} = (0.2, 0)$), the conditional confidence interval for the overall group was quite conservative. To address this issue a general approach would be to first construct a joint confidence region for (θ_1, θ_2) and then find the smallest rectangle containing the joint confidence region. To find the confidence region, we could construct a joint p-value function of the form

$$p(t_1, t_2; \theta_1, \theta_2) = \Pr(Y_1 > t_1, Y_2 > t_2 | \theta_1, \theta_2)$$

where the idea would be to only include values of (θ_1, θ_2) for which $p(t_1, t_2; \theta_1, \theta_2)$ is not "small". However, unlike in the univariate case where the probability integral transform ensures the p-value function has a standard uniform distribution for the true value of θ, in higher dimensions, this is not the case (Genest and Rivest, 2001). Instead $p(t_1, t_2; \theta_1, \theta_2)$ has a different distribution depending on (θ_1, θ_2). Nevertheless, in principle it is possible to find the $c_\alpha(\theta_1, \theta_2)$ that satisfies $\Pr\{p(Y_1, Y_2 | \theta_1, \theta_2) > c_\alpha(\theta_1, \theta_2)\} = \alpha$ for each pair (θ_1, θ_2). Hence the confidence region would contain all possible values of the treatment effects (θ_1, θ_2) which lead to a joint p-value greater than $c_\alpha(\theta_1, \theta_2)$. As an alternative to the conditional moment approach presented in Chapter 4, a similar method can be used to obtain an unconditional simultaneous confidence region for θ_1 and θ_2. However, this approach may be very computationally intensive and would become even

worse for more than two subgroups.

In this thesis, we have only focused on the problem of interval estimation. However, having constructed p-value functions it should be straightforward to construct median unbiased estimators (see Robertson et al., 2023, Section 5.2) of the treatment effects. Specifically, given the p-value defined in the earlier chapters, which is a function of the true treatment effect, and conditioned on the decision made in the interim analysis and statistic values of complementary subgroups, a median unbiased estimator can be obtained by taking the value $\hat{\theta}$ that satisfies $p(\boldsymbol{x}; \hat{\theta}) = 0.5$ where $\boldsymbol{x}$ is the observed data set. Therefore, our future work might include evaluating the performance of this point estimator.

The main focus of this thesis is on designs that assume consistency between the endpoints used in the interim and final analyses. Nevertheless, some trials adopt diverse endpoints for these analyses, as exemplified in the Jenkins et al. (2011) design that utilizes progression-free survival (PFS) to establish population enrichment in the second stage, while collecting overall survival (OS) data in both stages for the final analysis. Since there is scarce literature on constructing confidence intervals for this category of design, it could be a prospective field for future research.

Lastly, in the calculations performed in the thesis we have assumed the individual patient responses are normally distributed. However, the same approach can be applied to binary (Simon and Simon, 2013) and time-to-event data (Kimani et al., 2020) to give confidence intervals by assuming that the stage-wise score statistics are approximately normally distributed.

Bibliography

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, 44(1/2):9–26.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, 18(14):1833–1848.

Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(14):1029–1041.

Brannath, W. and Schmidt, S. (2014). A new class of powerful and informative simultaneous confidence intervals. *Statistics in Medicine*, 33(19):3365–3386.

Brent, R. P. (2013). *Algorithms for Minimization without Derivatives*. Courier Corporation.

Burnett, T. and Jennison, C. (2021). Adaptive enrichment trials: what are the benefits? *Statistics in Medicine*, 40(3):690–711.

Cohen, A. and Sackrowitz, H. B. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters*, 8(3):273–278.

CONSORT (2010). Extensions of the consort statement. `http://www.consort-statement.org/extensions`.

EMA (1998). ICH E9 statistical principles for clinical trials - scientific guideline European Medicines Agency. `https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline`.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):875–892.

Fairbanks, K. and Madsen, R. (1982). P values for tests using a repeated significance test design. *Biometrika*, 69(1):69–74.

FDA (2007). Reporting results from studies evaluating diagnostic tests - guidance. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda.

FDA (2018). Precision medicine. https://www.fda.gov/medical-devices/in-vitro-diagnostics/precision-medicine.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., and Fawcett, J. (2010). Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA: The Journal of the American Medical Association*, 303(1):47–53.

Genest, C. and Rivest, L.-P. (2001). On the multivariate probability integral transformation. *Statistics & probability letters*, 53(4):391–399.

Hodson, R. (2016). Precision medicine. *Nature*, 537(7619):S49–S49.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Guilford School Practitioner. Taylor & Francis.

Jenkins, M., Stone, A., and Jennison, C. (2011). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10(4):347–356.

Kimani, P. K., Todd, S., Renfro, L. A., Glimm, E., Khan, J. N., Kairalla, J. A., and Stallard, N. (2020). Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection. *Statistics in Medicine*, 39(19):2568–2586.

Kimani, P. K., Todd, S., and Stallard, N. (2013). Conditionally unbiased estimation in phase ii/iii clinical trials with early stopping for futility. *Statistics in Medicine*, 32(17):2893–2910.

Kimani, P. K., Todd, S., and Stallard, N. (2014). A comparison of methods for constructing confidence intervals after phase ii/iii clinical trials. *Biometrical Journal*, 56(1):107–128.

Kjaersgaard, P., Reiertsen, O., Trondsen, E., Rosseland, A., and Larsen, S. (1994). Comparison of sequential and fixed-sample designs in a controlled clinical trial with laparoscopic versus conventional cholecystectomy. *Scandinavian Journal of Gastroenterology*, 29(9):854–858.

Knottnerus, J. A. and Tugwell, P. (2013). Heterogeneity and clinical reality. *Journal of Clinical Epidemiology*, 66(8):809–811.

Kosorok, M. R. and Laber, E. B. (2019). Precision medicine. *Annual Review of Statistics and its Application*, 6:263–286.

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

Lin, R., Yang, Z., Yuan, Y., and Yin, G. (2021). Sample size re-estimation in adaptive enrichment design. *Contemporary Clinical Trials*, 100:106216.

Magirr, D., Jaki, T., Posch, M., and Klinglmueller, F. (2013). Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika*, 100(4):985–996.

Magnusson, B. P. and Turnbull, B. W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine*, 32(16):2695–2714.

Ondra, T., Jobjörnsson, S., Beckman, R. A., Burman, C.-F., König, F., Stallard, N., and Posch, M. (2019). Optimized adaptive enrichment designs. *Statistical Methods in Medical Research*, 28(7):2096–2111.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1):1–15.

Polyak, K. et al. (2011). Heterogeneity in breast cancer. *The Journal of Clinical Investigation*, 121(10):3786–3788.

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

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51(4):1315–1324.

Robertson, D. S., Choodari-Oskooei, B., Dimairo, M., Flight, L., Pallmann, P., and Jaki, T. (2023). Point estimation for adaptive trial designs i: a methodological review. *Statistics in Medicine*, 42(2):122–145.

Rosenblum, M., Fang, E. X., and Liu, H. (2020). Optimal, two-stage, adaptive enrichment designs for randomized trials, using sparse linear programming. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):749–772.

Rosenblum, M., Luber, B., Thompson, R. E., and Hanley, D. (2016). Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine*, 35(21):3776–3791.

Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika*, 75(4):723–729.

Simon, N. and Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625.

Stallard, N. and Todd, S. (2005). Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference*, 135(2):402–419.

Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine*, 27(24):4914–4927.

Wagner, M., Balk, E. M., Kent, D. M., Kasiske, B. L., and Ekberg, H. (2009). Subgroup analyses in randomized controlled trials: the need for risk stratification in kidney transplantation. *American Journal of Transplantation*, 9(10):2217–2222.

Wang, S.-J., James Hung, H., and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):358–374.

Wang, S.-J., O'Neill, R. T., and Hung, H. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*, 6(3):227–244.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. John Wiley & Sons.

Appendix A

P-value functions under stage-wise and MLE ordering in the 3-stage Magnusson and Turnbull design

A.1 Stage-wise ordering

Stage-wise ordering defines that the probability of observing a value at stage $k-1$ should be considered more extreme than stage k. Therefore, the conditional p-value function is defined as:

$$
\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\Pr(Y_{1,\mathcal{S}} > u) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \times I(Y_{1,\mathcal{S}} > \tilde{u}_{1,\mathcal{S}}) \\
&+[\Pr(Y_{1,\mathcal{S}} > \tilde{u}_{1,\mathcal{S}}) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) + \Pr(Y_{2,\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})] I(Y_{1,\mathcal{S}} < \tilde{u}_{1,\mathcal{S}}) \\
&\times I(Y_{2,\mathcal{S}} > \tilde{u}_{2,\mathcal{S}} \text{ or } Y_{2,\mathcal{S}} < \tilde{l}_{2,\mathcal{S}}) \\
&+[\Pr(Y_{1,\mathcal{S}} > \tilde{u}_{1,\mathcal{S}}) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) + \Pr(Y_{2,\mathcal{S}} > \tilde{u}_{2,\mathcal{S}} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) + \Pr(Y_{3,\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})] \\
&\times I(Y_{1,\mathcal{S}} < \tilde{u}_{1,\mathcal{S}}) I(\tilde{l}_{2,\mathcal{S}} < Y_{2,\mathcal{S}} < \tilde{u}_{2,\mathcal{S}})
\end{aligned}
$$

where $I(\cdot)$ is the indication function.

Assume that $X_{i,j} \sim N(\theta_j \delta_{i,j}, \delta_{i,j})$ and $X_{i,j}$ only depends on $X_{i-1,j}$, if the subpopulation $\mathcal{S}$ is selected, we can explicate the p-value function for three-stage design as following by applying

stage-wise ordering

$$
\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\int_u^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S}} \times I(y_{1,\mathcal{S}} > \tilde{u}_{1,\mathcal{S}}) \\
+&\Big\{ \int_{\tilde{u}_{1,\mathcal{S}}}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S}} \\
+&\int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \times \left[1 - \Phi\left(\frac{u - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}y_{1,\mathcal{S}} \Big\} \\
&\times I(y_{1,\mathcal{S}} \leq \tilde{u}_{1,\mathcal{S}}) I(y_{2,\mathcal{S}} < \tilde{l}_{2,\mathcal{S}} | y_{2,\mathcal{S}} > \tilde{u}_{2,\mathcal{S}}) \\
+&\Big\{ \int_{\tilde{u}_{1,\mathcal{S}}}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S}} \\
+&\int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) \times \left[1 - \Phi\left(\frac{\tilde{u}_{2,\mathcal{S}} - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}y_{1,\mathcal{S}} \Big\} \\
&\times I(y_{1,\mathcal{S}} \leq \tilde{u}_{1,\mathcal{S}}) I(\tilde{l}_{2,\mathcal{S}} \leq y_{2,\mathcal{S}} \leq \tilde{u}_{2,\mathcal{S}}) \\
+&\int_{\tilde{l}_{1,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{2,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) f_{2|\mathcal{S}}(y_{2,\mathcal{S}}|y_{1,\mathcal{S}}, \theta_{\mathcal{S}}) \\
&\times \left[1 - \Phi\left(\frac{u - y_{2,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{3,\mathcal{S}}}{\sqrt{\delta_{3,\mathcal{S}}}} \right) \right] \mathrm{d}y_{2,\mathcal{S}} \mathrm{d}y_{1,\mathcal{S}} \\
&\times I(y_{1,\mathcal{S}} \leq \tilde{u}_{1,\mathcal{S}}) I(\tilde{l}_{2,\mathcal{S}} \leq y_{2,\mathcal{S}} \leq \tilde{u}_{2,\mathcal{S}})
\end{aligned}
$$

where

$$
f_{1|\mathcal{S}}(y_{1,\mathcal{S}}|\theta_{\mathcal{S}}) = \begin{cases} \dfrac{1}{\sqrt{N_{1,\mathcal{S}}/4\sigma^2}} \psi\left(\dfrac{y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{1,\mathcal{S}}}{\sqrt{\delta_{1,\mathcal{S}}}} \right) \Phi\left[\dfrac{\tilde{l}_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{1,\mathcal{S}}}{\sqrt{\delta_{1,\mathcal{S}}}} \right]^{-1} & \mathcal{S} \in \{1, 2\} \\ \displaystyle\int_{\tilde{l}_{1,1}}^{\tilde{u}_{1,\mathcal{S}} - \tilde{l}_{1,2}} f_{1|1}(y_{1,1}|\theta_1) f_{1|2}(y_{1,\mathcal{S}} - y_{1,1}|\theta_2) \mathrm{d}y_{1,1} & \mathcal{S} = \{0\} \end{cases} .
$$

A.2 MLE ordering

For MLE ordering method, since it needs to convert the standardized score statistic to MLE scaled statistics corresponding to different stages, we denote the converted statistics in three stages as below:

1. If the trial stops at stage 1:

$$
\begin{aligned}
Y_{1,mle,\mathcal{S}} &= Z_{\mathcal{S}} / \mathcal{I}_1^{-0.5} = Y_{1,\mathcal{S}}, \\
Y_{2,mle,\mathcal{S}} &= Z_{\mathcal{S}} / \mathcal{I}_2^{-0.5} = Y_{1,\mathcal{S}} I_1^{-0.5} / I_2^{-0.5}, \\
Y_{3,mle,\mathcal{S}} &= Z_{\mathcal{S}} / \mathcal{I}_3^{-0.5} = Y_{1,\mathcal{S}} I_1^{-0.5} / I_3^{-0.5};
\end{aligned}
$$

2. If the trial stops at stage 2:

$$
\begin{aligned}
Y_{1,mle,\mathcal{S}} &= Z_{\mathcal{S}}/\mathcal{I}_1^{-0.5} = Y_{2,\mathcal{S}} I_2^{-0.5}/I_1^{-0.5}, \\
Y_{2,mle,\mathcal{S}} &= Z_{\mathcal{S}}/\mathcal{I}_2^{-0.5} = Y_{2,\mathcal{S}}, \\
Y_{3,mle,\mathcal{S}} &= Z_{\mathcal{S}}/\mathcal{I}_3^{-0.5} = Y_{2,\mathcal{S}} I_2^{-0.5}/I_3^{-0.5};
\end{aligned}
$$

3. If the trial stops at stage 3:

$$
\begin{aligned}
Y_{1,mle,\mathcal{S}} &= Z_{\mathcal{S}}/\mathcal{I}_1^{-0.5} = Y_{3,\mathcal{S}} I_3^{-0.5}/I_1^{-0.5}, \\
Y_{2,mle,\mathcal{S}} &= Z_{\mathcal{S}}/\mathcal{I}_2^{-0.5} = Y_{3,\mathcal{S}} I_3^{-0.5}/I_2^{-0.5}, \\
Y_{3,mle,\mathcal{S}} &= Z_{\mathcal{S}}/\mathcal{I}_3^{-0.5} = Y_{3,\mathcal{S}};
\end{aligned}
$$

where $Z_{\mathcal{S}}$ is the standardized statistic and $Y_{k,\mathcal{S}}$ is the raw statistic. Similarly, let $c_{1,mle}$, $c_{2,mle}$ and $c_{3,mle}$ be the transformed observed value at each stage. Then we define the conditional p-value function in terms of the converted statistics:

$$
\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\Pr(Y_{1,mle,\mathcal{S}} > \max(c_{1,mle}, \tilde{u}_{1,\mathcal{S}})) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
+&\Pr(Y_{2,mle,\mathcal{S}} > \max(c_{2,mle}, \tilde{u}_{2,\mathcal{S}}) | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})] \\
+&\Pr(Y_{3,mle,\mathcal{S}} > c_{3,mle} | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}})]
\end{aligned}
\tag{A.1}
$$

Equation (A.1) can also be expressed as:

$$
\begin{aligned}
&\Pr(Y_{\mathcal{S}} > u | \mathcal{S}^* = \mathcal{S}, \theta_{\mathcal{S}}) \\
=&\int_{\max(c_{1,mle}, \tilde{u}_{1,\mathcal{S}})}^{\infty} f_{1|\mathcal{S}}(y_{1,\mathcal{S}} | \theta_{\mathcal{S}}) \mathrm{d}y_{1,\mathcal{S}} \\
+&\int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}} | \theta_{\mathcal{S}}) \left[1 - \Phi \left(\frac{\max(c_{2,mle}, \tilde{u}_{2,\mathcal{S}}) - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{2,\mathcal{S}}}{\sqrt{\delta_{2,\mathcal{S}}}} \right) \right] \mathrm{d}y_{1,\mathcal{S}} \\
+&\int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{1,\mathcal{S}}} \int_{\tilde{l}_{2,\mathcal{S}}}^{\tilde{u}_{2,\mathcal{S}}} f_{1|\mathcal{S}}(y_{1,\mathcal{S}} | \theta_{\mathcal{S}}) f_{2|\mathcal{S}}(y_{2,\mathcal{S}} | y_{1,\mathcal{S}}, \theta_{\mathcal{S}}) \\
\times&\left[1 - \Phi \left(\frac{c_{3,mle} - y_{1,\mathcal{S}} - \theta_{\mathcal{S}} \delta_{3,\mathcal{S}}}{\sqrt{\delta_{3,\mathcal{S}}}} \right) \right] \mathrm{d}y_{2,\mathcal{S}} \mathrm{d}y_{1,\mathcal{S}}.
\end{aligned}
$$