Phase II/III Transitional Seamless Trial Designs with Different Objectives and Endpoint Types

Robert A. Tumasian III

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Phase II/III Transitional Seamless Trial Designs with Different Objectives and Endpoint Types

Robert A. Tumasian III

Accelerating regulatory science has become an urgent task due to the ongoing emergence of novel and highly complex diseases, such as coronaviruses and mpox, and the growing real-world evidence and precision medicine paradigms. In research and development, achieving this goal requires the formulation and implementation of innovative clinical trial designs and statistical approaches. Further efforts are needed to devise robust strategies toward expediting the evaluation of drugs and other medical products, which will help to quicken the delivery of safe and effective treatments to individuals in need.

One pioneering concept that has greatly contributed to speeding up product assessment is the application of pre-specified adaptations in clinical trials. Adaptive techniques (like early dose or treatment selection and sample size recalibration) are based on participant data collected during the trial, where investigators/sponsors are able to learn from the study population and adjust trial characteristics accordingly without diminishing the integrity of the study. However, adaptations must be employed cautiously; it is essential to ensure that they are both clinically reasonable and statistically sound (in the sense that they will not inflate the nominal type I error rate).

The evolution of seamless trial designs has also propelled drug/medical product evaluation. For instance, seamless trials that combine a phase II study (e.g., using a biomarker or short-term intermediate endpoint for early dose or treatment selection) and a phase III study (e.g., using a definitive endpoint for confirming efficacy) have become very popular in practice due to their

flexibility and shorter duration compared to running separate studies. In phase II/III seamless trials, a correlation is usually pre-specified between the intermediate and definitive endpoints in order to bridge the phase II and phase III studies and accomplish the overall aims of the intended trial. However, this can be difficult to quantify, especially when the endpoints are non-continuous. An inappropriate or unsubstantiated enumeration of this correlation can yield flawed results and thus misguide approval and labeling decisions, leading to potentially serious consequences.

To avert this issue, a three-stage phase II/III transitional seamless trial design was constructed that does not demand prior knowledge of the correlation between the intermediate and definitive endpoints. The design also allows for interim sample size re-estimation according to the accrued intermediate endpoint data. The utility and validity of the design, with and without sample size adaptation, were considered for different types of endpoints. For each, theoretical proofs establish that the design can control the nominal type I error rate while still reaching the targeted power, and simulations reinforce these findings. Therefore, the design exhibits promise in precipitating the assessment of experimental interventions in a wide range of therapeutic areas.

This dissertation is organized as follows. First, Chapter 1 presents (1) an introduction to the drug/medical product evaluation process, (2) an overview of adaptive techniques used in clinical trials, (3) a description of seamless trial designs, including a real-world example to demonstrate how they can be applied in different data situations, (4) the research goals, and (5) a display of the proposed three-stage phase II/III transitional seamless trial design and its workflow. Chapters 2-5 lay out the models and procedures for the proposed design when using a continuous intermediate endpoint and a continuous definitive endpoint, a binary intermediate endpoint and a continuous definitive endpoint, and a binary intermediate endpoint and a time-to-event definitive endpoint in the presence and absence of censored observations. This is proceeded by statistical justification and supporting simulations for each procedure. As some authors remain skeptical toward the use of adaptations in clinical trials, Chapter 6 counters an argument posed in the literature. In closing, Chapter 7 summarizes the work and its limitations and provides several additional considerations toward enhancing the proposed design and its generalizability.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am forever grateful for the constant love and encouragement that I have received from my family and friends throughout my educational journey. I would not have been able to get to this point without all of their support along the way. I also owe a special thank you to my cohort for their help over the course of my doctoral studies. To my instructors, I have enjoyed learning from you and have gained essential knowledge and skills to succeed in my career and become a leader in the field. Finally, I must express a huge amount of gratitude to my advisor, Dr. Bin Cheng, for his instrumental guidance and for the opportunity to collaborate with him on many exciting projects, and to my other committee members (Drs. Ying-Kuen Cheung, Min Qian, Shing Lee, and Shein-Chung Chow) for their very valuable feedback on my work.

Dedication

This dissertation is dedicated to my family, friends and everyone else who has provided me with opportunities to delve into my interests and work toward achieving my lifelong personal and professional goals. I would also like to dedicate this work to my fellow clinical trial enthusiasts, in the hopes that this material will spark further innovations, and to every student; remember to take chances and never, ever give up.

Preface

My training and research collaborations in biomedical science, clinical trials, mathematics, and statistics have instilled in me a commitment to ensuring the proper evaluation of drugs and alternative treatments and improving quality of life for all people. Through my participation in various learning experiences at the State University of New York (SUNY) at Geneseo, Columbia University, Emory University, Harvard University, the National Institute on Aging (NIA) – National Institutes of Health (NIH), and the U.S. Food and Drug Administration (FDA), I have built the skill sets necessary to contribute to the development of clinical trial designs that can enhance product assessment. This dissertation is motivated by the critical need to make safe and effective therapies available to the community as rapidly as possible. While this work is tailored to clinical trial conductors, regulators and researchers for the purpose of proposing another potential evaluation method and kindling future related efforts, the content of this dissertation is also useful for individuals who are interested in getting a glimpse into the growing world of clinical trials. The proposed design and procedures, and the methodological approaches employed to establish their validity, are original and have not been published prior to the submission of this work.

Chapter 1: Introduction

Over the past several decades, clinical decision-making has become largely data-driven and the clinical trial landscape has improved immensely in terms of strengthening and accelerating the safety and efficacy evaluation of drugs and other medical products [1–9]. The continual appearance of new illnesses and highly mutable viruses requires swift transformations and versatility in the framing and undertaking of clinical trials. Expanding our warehouse of efficient and reliable study designs and computational methods will help us to advance regulatory science and more quickly assess interventions to confront current and future public health emergencies.

1.1 Overview of the Drug/Medical Product Evaluation Pipeline

Randomized controlled trials (RCTs) are considered the gold standard for biomedical research, and they are the preferred method for measuring the performance of a drug/medical product. The lifecycle of a drug/medical product spans over many years and is comprised of five benchmarks. The first is discovery, which entails pinpointing processes or sites within the body (e.g., genes, proteins or receptors) that are associated with the condition at hand and then postulating a therapeutic that could interact with them appropriately. This necessitates interdisciplinary collaboration between biologists, chemists, epidemiologists, physicians, and other health professionals [10].

Once equipped with a compelling and science-driven hypothesis, the second step of pre-clinical testing can begin, which usually involves conducting toxicological experiments on animal models to learn about the functionality of the proposed therapy. Now, the U.S. Food and Drug Administration (FDA) can also accept cell-based assays and computer modeling alternatives per the passing of the FDA Modernization Act 2.0. These studies are essential in determining whether or not the therapy would be safe to test in humans. Likewise, investigators must also explore the therapy's mechanism of action (i.e., how the therapy might operate in the body). Another vital component of pre-clinical testing is to gain insights into dosing that would be suitable for human consumption. Together, the discovery and pre-clinical testing processes can take up to ten years to complete. If promising pre-clinical study results are obtained, then the sponsor may continue to the third step of clinical testing in order to assess the safety and effectiveness of the potential therapy in human populations [10].

Due to the complex nature of clinical testing, this step can also take several years and largely depends on the disease area. Clinical testing involves four main phases. In phase I, a small sample of individuals from the population of interest is enrolled in a first-time-in-humans (FTIH) trial, or first-in-human (FIH) trial, and followed to evaluate safety and side effects and identify the maximum tolerable dose (MTD) that does not induce a harmful outcome. When applicable, the most appropriate route of administration can also be studied during phase I. Once the highest performing dose is selected (e.g., using the conventional "3+3" approach [11, 12] or the continual reassessment method [13, 14]), additional participants are recruited from the target population for a phase II proof-of-concept (PoC) trial to continue assessing safety and start acquiring information on efficacy. If there are no substantial safety concerns (e.g., in terms of adverse or serious adverse events), then the treatment may advance to phase III, which contains a much larger sample size in order to establish safety and effectiveness compared to a control (e.g., placebo and/or standard of care) according to a clinically relevant endpoint. Thus, phase III studies usually include multiple clinical sites to capture a wider portion of the population that is expected to utilize the drug/medical product [10].

Following step four of regulatory approval by the FDA based on clinical judgment and the results of pre-specified statistical analyses, the final step is post-marketing surveillance (i.e., phase IV), in which the long-term effects of the product are scrutinized among several hundreds or thousands of users in the public (e.g., via electronic medical records and/or health registry data) for a period of time no less than required by federal statutes. In phase IV, if there are any severe side effects conceivably related to the treatment that were not seen in prior studies, then the product could be tagged with a black box warning (e.g., an abundance of deleterious events experienced within a subgroup of the target population could result in modified prescription and/or usage guidelines) or be completely removed from the market. The general drug/medical product evaluation pipeline is depicted in Figure 1.1 [15]. If approved, the product can also be examined for use in alternative populations (e.g., among children and other vulnerable or immunocompromised groups) with the condition, or for treating different maladies, by repeating the benchmarks described above [10].

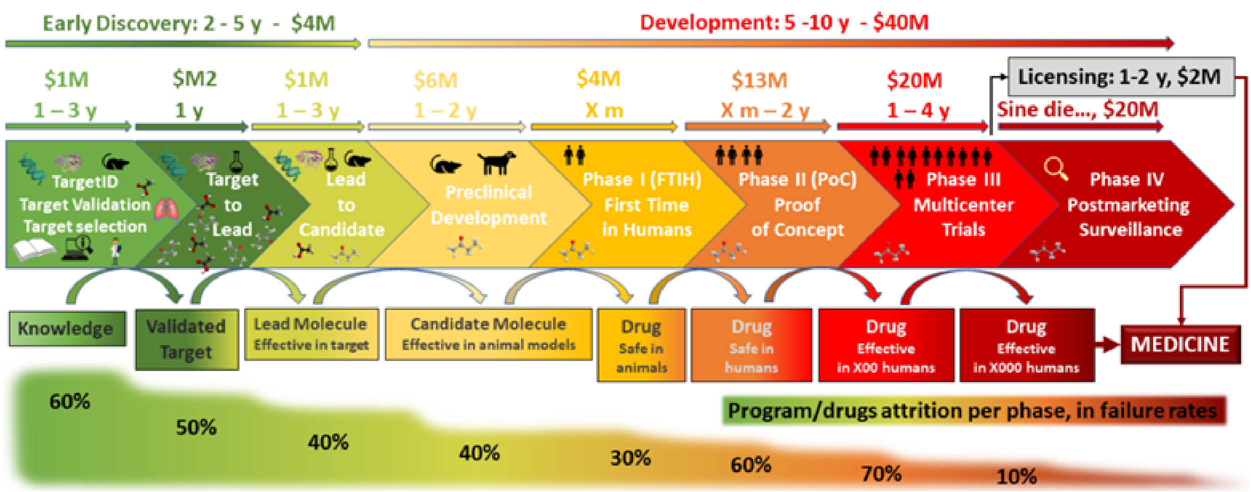

Figure 1.1: Standard lifecycle of a drug/medical product.

1.2 Adaptive Methods in Clinical Trials

In many clinical trial settings, adaptations (e.g., endpoint alteration, sample size re-estimation, and/or the addition or elimination of treatment or dosage groups) can be planned *a priori* and then incorporated into the study based on accrued participant information [16–27]. Adaptive techniques have revolutionized and become quite prevalent in clinical research and development (particularly after the launch of the FDA's Critical Path Initiative geared toward enhancing drug/medical product evaluation practices [28]) due to their ability to adjust the structural properties and statistical procedures of clinical trials after commencement. Numerous computational arguments and large-scale simulation studies have been reported in the literature that demonstrate how pre-specified adaptations can be implemented without biasing the observed treatment effects or jeopardizing the operational characteristics (e.g., inflating the nominal type I error rate) of the trial that is underway [29–40]. A preponderance of published rationales for the use of adaptive methods in clinical trials are cogent and relevant across a vast array of therapeutic areas.

Scientifically and statistically plausible adaptations can play a fundamental role in expediting the pursuit of safe and efficacious interventions. The topic of adaptive techniques continues to be of great interest to regulatory bodies worldwide, including the FDA and the European Medicines Agency (EMA), due to the upward trend of global trials to support simultaneous regulatory filings across multiple countries and geographic regions and the desire to internationally harmonize regulatory practices. This has led to the preparation and release of several guidance documents and reflection papers [41–43] to explicitly define the term "adaptive design," state some of the potential advantages and drawbacks of adaptations, and outline possible methodological issues that may be faced and actions that could be taken to rectify them.

In their earliest guidance [42], the FDA provides nine examples of adaptive techniques: modification to the pre-specified (1) primary endpoint, (2) secondary endpoint(s), (3) patient assessment schedule, (4) patient medication record (e.g., a patient may begin taking another medication during the study), (5) trial duration (e.g., stopping the study early for futility or superiority), (6) sample size, (7) treatment/dosage arms or regimens, (8) randomization scheme, and (9) inclusion criteria. This also encompasses any subsequent analytical changes. The non-exhaustive list above demonstrates the breadth of adaptations that are available and have been employed in previous trials [44]. The EMA and FDA also emphasize that adaptive methods must be pre-specified in the study protocol; they can only be implemented *a posteriori* via an amendment following extensive discussions between all stakeholders involved (e.g., statisticians, medical staff and patient groups).

Adaptive designs can have several positive implications compared to non-adaptive options such as making use of external information that may arise and reducing lead times, total expected sample sizes and duration, all while still preserving the pre-specified type I error rate and achieving the targeted power. Adaptations also reflect real clinical practice and can help to redress incorrect assumptions that were made in the design stage (i.e., before trial start), usually at the discretion of an independent data and safety monitoring committee. Additional advantages of adaptive techniques have been well-recognized by the FDA in its most recent guidance for industry on adaptive designs [41]. For instance, adaptive techniques can be more statistically efficient, or have greater statistical power, in detecting a true treatment effect (e.g., a sample size adjustment typically results in the need for a larger sample size, which would then heighten the power of the study). The benefits of adaptations are profuse and have been discussed at length in both Bayesian [45–47] and frequentist [44, 48, 49] contexts.

Nonetheless, adaptations can lead to ramifications when articulated or applied inappropriately; avoiding this requires meticulous planning with strong clinical and statistical justifications for their use. The general frameworks for an adaptive and non-adaptive (traditional parallel group) design are compared in Figure 1.2 [50]. Although adaptations are highly attractive due to their flexibility,

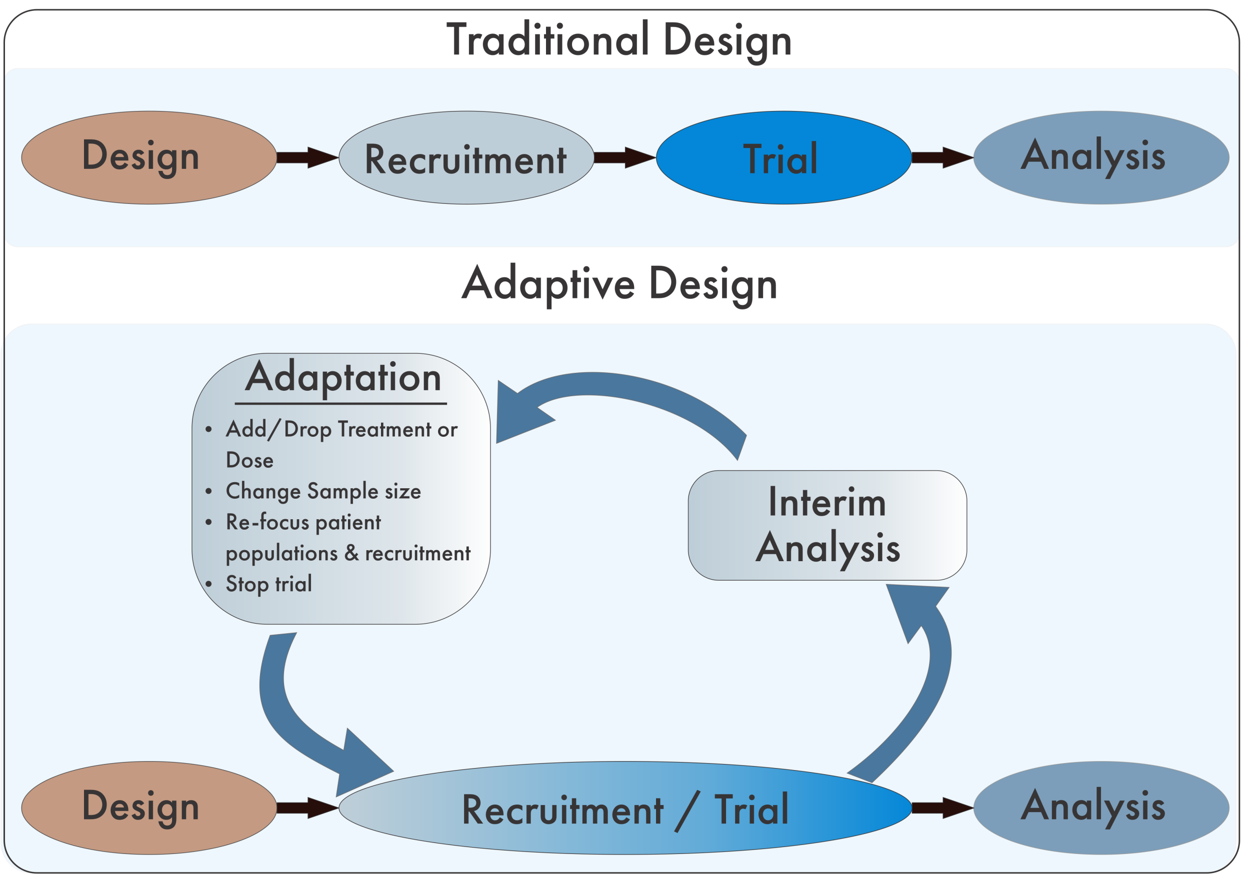

Figure 1.2: Comparison of adaptive and non-adaptive (traditional parallel group) designs.

and as no strict guidelines exist, they should be kept to a minimum in order to reduce trial complexity and bias. The availability of adaptive methods can also detract investigators/sponsors from using the truly best design, and excess adaptations based on insights from study personnel who are analyzing unblinded interim data can increase the likelihood of obtaining positive results that are actually false, thus compromising the integrity and persuasiveness of the trial [51]. Problems can also occur when there are delayed responses (e.g., in oncology trials that use survival or other time-to-event endpoints), since they may cause adaptive trials to be stopped prematurely for futility when in fact the experimental treatment could yield a positive effect later in the course of the trial [44, 52]. Research remains active on utilizing long-term and short-term (surrogate) responses in adaptive settings [53–59]. Further material on the possible disadvantages of adaptations can be found in the literature [60–64].

Some authors have mentioned that adaptive designs can suffer from loss of statistical efficiency and thus they do not recommend them in practice [65, 66]. Specifically, adaptive design test statistics may violate the sufficiency principle if sequentially sampled trial participants are differentially weighted, but this can be resolved by allowing rejection of the null hypothesis only if the conventional (non-adaptive) test statistics would also reject the null at the full α level. Moreover, adding, dropping or changing null hypotheses, or considering multiple null hypotheses simultaneously, can lead to difficulty in interpreting the final hypothesis test. To combat this challenge, the closed testing principle or a hierarchical test could be applied. The closed testing principle controls for the overall type I error rate by rejecting a hypothesis only if each of the intersection hypotheses containing the hypothesis of interest is rejected by the associated local test. For hierarchical tests, endpoints are ranked in descending order of clinical importance, with the most important test conducted first; if a significant p-value is obtained, then the effect is deemed real and the procedure continues down the list until significance is not achieved [44]. Myriad researchers have pointed out that in some scenarios (e.g., when using a long-term outcome) it may be appealing to employ an adaptation (that is both clinically and statistically sound) to improve product assessment [51, 67]. The application of adaptive techniques requires further conversations to increase consensus.

1.3 Seamless Clinical Trial Designs

Many clinical trial design options have been constructed (and justified via statistical rigor and simulations) to condense the clinical testing process described in Section 1.1, including a variety of seamless designs [68–80]. One popularly employed clinical trial design is the two-stage phase II/III seamless design, which merges separate phase II and phase III studies into one single study in order to expedite drug/medical product evaluation. During stage one (i.e., in the phase II study), a common goal is early identification of the most promising treatment or dosage, according to intermediate endpoint data, for further efficacy assessment in stage two (i.e, in the phase III study) using a definitive endpoint. Note that the intermediate endpoint can either be a biomarker, a formal short-term surrogate endpoint or a primary endpoint based on the purpose(s) of the trial, while the definitive endpoint must be a primary endpoint. An illustration of the typical layout and workflow of the two-stage phase II/III seamless design outlined above is provided Figure 1.3. To display how

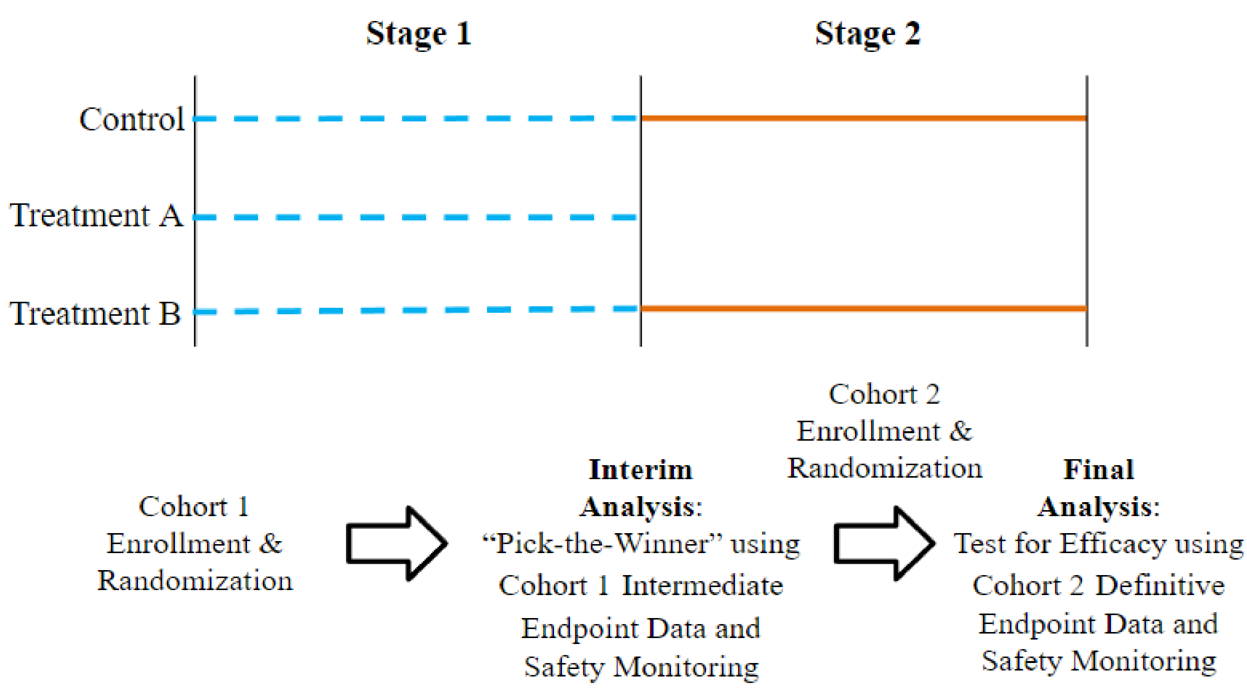

Figure 1.3: A two-stage phase II/III seamless trial design, where treatments A and B denote active interventions being tested against the control. The dashed blue line and solid orange line represent data collection for the intermediate endpoint from the first cohort of participants in stage one and the definitive endpoint from the second cohort of participants in stage two, respectively.

the trial progresses, treatment A is shown to be dropped at the end of stage one and data collection is terminated, since treatment B performed better according to some intermediate endpoint. Therefore, no participants will receive treatment A in stage two (all participants on Treatment A in stage one would be removed from the study, given standard of care and continue to be monitored for safety signals). Thus, the two cohorts are independent and there is no switchover after stage one; cohort one participants cannot be included in the second cohort due to possible carryover effects (i.e., those receiving treatment A in stage one cannot receive treatment B in stage two due to the potential for mixed treatment effects that could undermine the results of the trial). For simplicity of exposition, the example in Figure 1.3 only considers two active treatment arms versus a control, but the design can handle comparison of any number of experimental interventions (or dosages) to a control group. Also, like the majority of seamless studies, the two-stage phase II/III seamless design in Figure 1.3 can be perceived as separate parallel group trials being run at each stage.

There are four scenarios that can be encountered when composing a multi-stage seamless trial: (1) the objectives and endpoints are the same at each stage, (2) the objectives are the same at each stage but the endpoints are different, (3) the endpoints are the same at each stage but the objectives are different, and (4) the objectives and endpoints are different at each stage. For scenario (1), standard group sequential methods that use stage-wise p-values are often employed by considering the end of each stage as a planned interim analysis with a pre-specified set of stopping rules that control the nominal type I error rate, such as Pocock or O'Brien-Fleming boundaries [81–84]. These methods include combination tests via Fisher's product or the weighted inverse normal function and conditional error and power techniques [30, 85–92]. Unfortunately, these procedures cannot be utilized for the other three scenarios when the study objectives and/or endpoints vary at each stage (e.g., the sample size needed to reach a certain power under a group sequential design [87, 93–96] may not be enough to accomplish the study goals under a multi-stage seamless design).

For scenario (2), statistical testing procedures have been developed that use binary, continuous and survival/time-to-event data collected from both stages, but under the often strong assumption that there is a well-defined relationship or correlation between the two endpoints [97–99]. The most common application of scenario (3) is when stage one is used for dose-finding or early treatment selection and stage two is used for efficacy confirmation. However, in the interest of shortening the assessment process, similar but slightly different endpoints are typically considered at each stage (e.g., using a short-term intermediate endpoint in the first stage, such as an abbreviated definitive endpoint, and a comparable definitive endpoint in the second stage), which leads to scenario (4). Each of these scenarios serve their own purpose and require different computational approaches, which depend on the aims of the intended study [70, 71, 100]. Statistical inference for scenario (4) is the most challenging and this research is ongoing in many clinical settings [80, 101–109].

There are several choices that can be made for the stage one intermediate endpoint and the stage two definitive endpoint, including the use of a continuous intermediate endpoint and a continuous definitive endpoint, a binary intermediate endpoint and a continuous definitive endpoint, or a binary intermediate endpoint and a survival/time-to-event definitive endpoint. The endpoint types that are selected can rely on many factors such as disease area, ease of data collection, financial constraints, and other resource limitations. For instance, if the definitive outcome will take longer to observe than the intermediate outcome, then an adaptive procedure (e.g., a seamless design) could help to precipitate the assessment of the drug/medical product.

1.4 Seamless Trial Example

Several years ago, a pharmaceutical company conducted a two-stage, multi-center, randomized, open-label, active-control, double-blind, phase II/III seamless adaptive trial: the Safety and Efficacy Study to Compare Continuous Infusion of Interferon with Standard of Care for Chronic Hepatitis C (also known as the COPE-HCV study; NCT00919633) [110]. This trial evaluated the safety/tolerability and effectiveness of an experimental treatment for subjects 18-65 years of age with hepatitis C virus (HCV) for at least six months who had never received previous interferon or other anti-HCV treatment. The trial integrated independent phase II and phase III studies into one two-stage study, where the first stage was for dose selection and the second stage was for establishing non-inferiority of the dose selected from stage one compared to the control, which was the standard of care at that time. The COPE-HCV study design is provided in Figure 1.4 [100], which included four active arms (Drug B: Doses 1-4) and one control arm (Drug A), a treatment duration of 48 weeks, and a 24-week follow-up period.

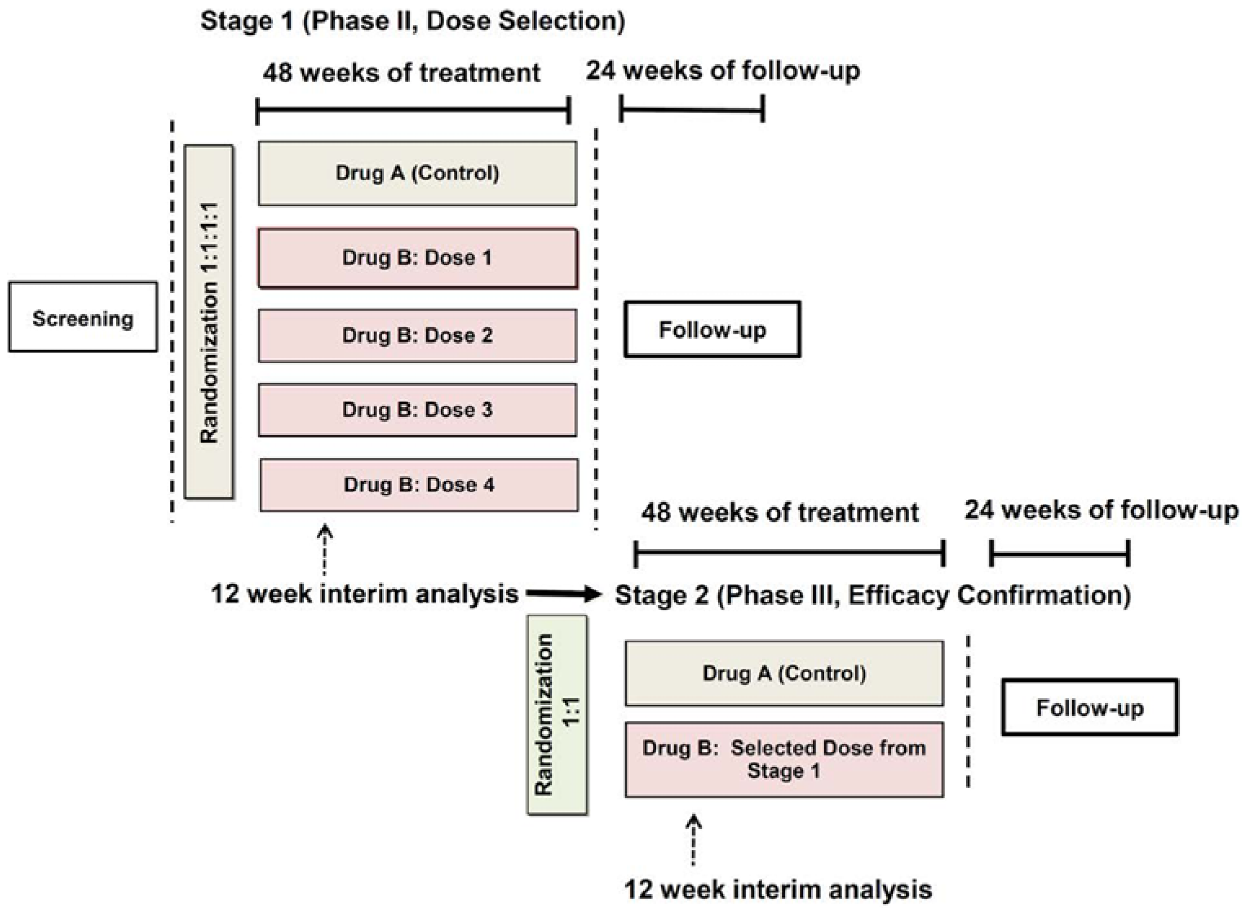

Figure 1.4: COPE-HCV study design.

The objectives of the COPE-HCV study were two-fold. During stage one, the goal was to select the optimal dose compared to the standard of care therapy using early virologic response (EVR) as a short-term intermediate endpoint, defined as the change in HCV RNA level from baseline to Week 12, which is on a continuous scale. The goal of the second stage was to demonstrate non-inferiority of the highest performing dose from stage one compared to the standard of care therapy using sustained virologic response (SVR), defined as the change in HCV RNA level from baseline to Week 72, as the long-term definitive endpoint, which is also on a continuous scale. Note that an

HCV RNA level <10 IU/mL was considered undetectable. Because it would take longer to observe the definitive endpoint, the sponsor was interested in considering an intermediate endpoint (in this instance, a definitive endpoint with shorter duration) for early decision-making. Also, the structure of this trial aligns with scenario (4) in Section 1.3, since it had different objectives and employed different endpoints at each stage.

Concisely, the COPE-HCV study procedure was as follows. In the first stage, after obtaining informed consent within 30 days of screening, eligible subjects were randomly assigned to one of the five treatment groups with 1:1:1:1:1 allocation. After the stage one subjects completed Week 12 of the trial, an interim analysis was performed for dose selection based on the accrued safety results (i.e., the number of adverse events experienced in each arm) and EVR data. This study also assumed that EVR at Week 12 was predictive of, or correlated with, SVR at Week 72, where the dosage choice was made via precision analysis using a certain set of pre-specified criteria. In other words, the dose selected to proceed to stage two had the most promise in achieving a clinically and statistically meaningful treatment difference compared to the control that was not due to chance. Stage one subjects who had not yet completed their intermediate endpoints continued with their assigned therapies for the remainder of the planned 48 weeks, with a final follow-up at Week 72.

The second stage carried out a non-inferiority comparison of the selected dose from stage one against the control. A separate cohort of subjects was equally randomized to each treatment group. The second interim analysis was conducted when the stage two subjects completed 12 weeks of the study and 80% of the stage one subjects completed 48 weeks of treatment and 24 weeks of follow-up (72 weeks total). Therefore, the second interim analysis involved verifying the assumption that 12-week EVR was in fact predictive of, or correlated with, 72-week SVR and determining whether or not the *a priori* sample size would be sufficient to accomplish the aims of the trial according to the data collected.

In both stages, subjects not meeting the inclusion criteria for EVR, and subjects who did meet the inclusion criteria for EVR but relapsed or became a breakthrough case at any time up to Week 72, discontinued the study and were each offered standard of care off protocol. All discontinuing subjects also underwent safety monitoring for 60 days after receiving their last dose. For the two planned interim analyses, the participant safety data and the incidence of EVR were reviewed by an independent data and safety monitoring board (DSMB); continuation of the trial was decided by the DSMB based on their clinical judgment and analyses. The type I error rate was controlled at 5% via O'Brien-Fleming boundaries [84] with a targeted power of 85%. Adaptations such as early termination and sample size re-estimation were implemented as recommended by the DSMB.

Alternatively, the investigator/sponsor could have employed a binary intermediate endpoint and a continuous definitive endpoint (e.g., an EVR <10 IU/mL at Week 12 and SVR at Week 72) or a binary intermediate endpoint and a time-to-event definitive endpoint (e.g., an EVR <10 IU/mL at Week 12 and time to first SVR <10 IU/mL or 72 weeks, whichever occurs earlier). These options could have been applied using statistical methods similar to those that will be presented in Chapters 2-5 for the proposed three-stage phase II/III transitional seamless trial design.

1.5 Study Goals

This study sought to contribute to the formulation of new clinical trial designs and corresponding statistical procedures that can foster faster drug/medical product evaluation (i.e., clinical testing) when using (1) a continuous intermediate endpoint and a continuous definitive endpoint, (2) a binary intermediate endpoint and a continuous definitive endpoint, or (3) a binary intermediate endpoint and a time-to-event definitive endpoint with and without censored observations. In particular, the three-stage phase II/III transitional seamless trial design presented in the next section was assembled with different stage-wise objectives and endpoints and the option for sample size recalibration based on accumulated intermediate endpoint data. The validity of the proposed design (i.e., its ability to control the pre-specified type I error rate and achieve the targeted power) was proven through statistical argument and reinforced by simulations to demonstrate its utility. For each pair of endpoint types, the central limit theorem [111] was used to develop the unmodified and modified procedures and Slepian's lemma [112] was applied for validation. The second and third pairs of endpoint types also utilized latent variables [113], where the copula idea [114, 115] and maximum likelihood estimation [116, 117] coupled with the multivariate delta method [118] were employed when handling the time-to-event definitive endpoint in the presence and absence of censoring, respectively. The FDA encourages the exploration of different design options (shown to have reliable operational characteristics via analytical calculations and/or simulations) that can be used to gain accurate conclusions on safety and efficacy to inform regulatory decision-making [28, 41], which has motivated and directly aligns with this work.

1.6 Proposed Design

As indicated in Section 1.3, seamless trials typically merge two separate studies, where a phase II study (e.g., with the goal of early dose or treatment selection using an intermediate endpoint) could be stage one and a phase III study (e.g., an efficacy confirmation trial using a definitive endpoint) could be stage two. However, to expedite the evaluation process, investigators/sponsors may be interested in collecting intermediate endpoint data from the second cohort of subjects and definitive endpoint data from the first cohort of subjects simultaneously, resulting in the proposed three-stage phase II/III transitional seamless design shown in Figure 1.5. The term "transitional" is used to emphasize that there is a smooth transition from stage one to stage two and from stage two to stage three, if the trial advances that far. The format of Figure 1.5 is similar to Figure 1.3 but includes an additional third stage. Note that per the FDA's 2019 guidance on adaptive designs [41], the design proposed in Figure 1.5 is considered an adaptive design since treatment/dosage groups are eliminated after stage one and there is the ability to reassess the total required sample size at the end of stage two (i.e., at the second interim analysis) based on the intermediate and definitive endpoint data from the first cohort of participants.

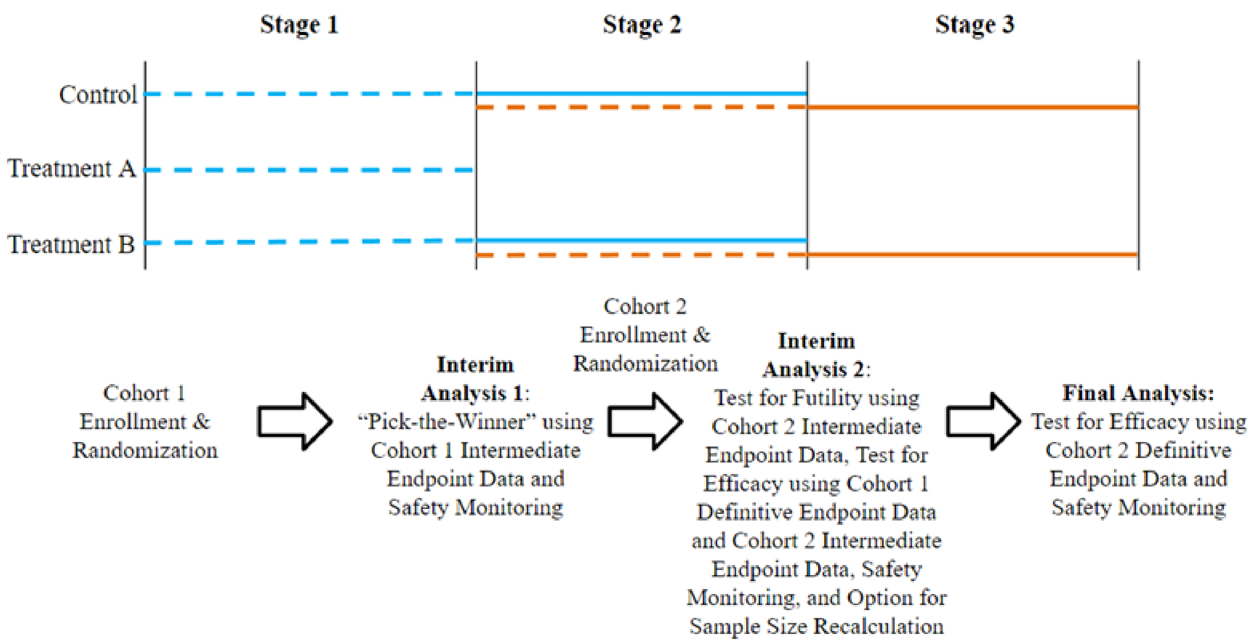

Figure 1.5: Proposed three-stage phase II/III transitional seamless trial design. Treatments A and B denote active interventions being compared to the control arm. Blue and orange lines represent data collection for the first and second cohort of participants, respectively. Dashed and solid lines represent data collection for the intermediate and definitive endpoints, respectively.

Chapter 2: Using a Continuous Intermediate Endpoint and a Continuous Definitive Endpoint

This chapter first outlines the data modeling structure used when employing a continuous intermediate endpoint and a continuous definitive endpoint, and then moves into describing the proposed procedure without interim sample size re-estimation and proving its validity via statistical argument, accompanied by supporting simulations. Afterwards, this is repeated for the modified version of the design, which incorporates a sample size recalibration at interim. Note that the data model presented in the next section is used regardless of whether or not a sample size adjustment is performed, and that the version without sample size modification can be interpreted as a standard group sequential procedure with treatment selection.

2.1 Model

For subject j receiving treatment $i = 0, \ldots, k$, let X_{ij} and Y_{ij} denote the continuous intermediate endpoint and the continuous definitive endpoint, respectively, where $i = 0$ represents that control group C and $i = 1, \ldots, k$ represent the active interventions. For the purposes of this work, since an appreciable portion of biological variables tend to follow approximately normal distributions when the number of observations is large, we make the assumption that for fixed cohort

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \mu_i \\ \nu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right]$$

where μ_i and ν_i denote the mean of the intermediate and definitive endpoint data, respectively, for subjects receiving treatment i, and σ^2 and τ^2 are the variances of the intermediate and definitive endpoint data, respectively. We also assume that the variances of the intermediate and definitive endpoint data in each treatment group are equal (since the variability of the intermediate and definitive outcomes are not often expected to be substantially different across the treatment groups) and that the correlation between the intermediate and definitive endpoints is non-negative ($\rho \geq 0$) to ensure that the intermediate and definitive endpoint data does not yield contradictory conclusions on treatment performance, which is anticipated in practice). Utilizing this data model, pairwise (two-sample) Z-statistics can be constructed via the central limit theorem, which are used in the proposed procedures with and without sample size recalibration. Then, Slepian's lemma is applied to justify both procedures, noting that for fixed treatment i, subject j and cohort, $\mathrm{Cov}(X_{ij}, Y_{ij}) = \rho\sigma\tau$ is a strictly increasing function of ρ since $\sigma^2 > 0$ and $\tau^2 > 0$.

2.2 Design without Sample Size Modification

In this section, details will be provided for the proposed procedure that does not reassess the total required sample size at interim. This is followed by statistical justification that the size of the proposed unmodified procedure can be bounded from above and its power can be bounded from below. Simulation results will also be presented, which support our proofs. Appropriate steps to take for unknown σ^2 and/or τ^2 will be discussed as well.

2.2.1 Procedure

Consider the design proposed in Figure 1.5 for comparing k active treatment groups E_i, $i = 1, \ldots, k$, to a control group C ($i = 0$), where the intermediate endpoint and the subsequent definitive endpoint are both utilized for measuring the treatment effects. Let θ_i and ψ_i be the treatment effect comparing active treatment E_i to the control C measured by the intermediate endpoint and the definitive endpoint, respectively. Assuming that positive values of θ_i and ψ_i correspond to treatment effects that favor the product under study, the hypothesis of interest based on the intermediate endpoint data is

$$H_{0,1} : (\theta_1, \ldots, \theta_k)^{\mathrm{T}} \leq \mathbf{0} \tag{2.1}$$

and the hypothesis of ultimate interest is

$$H_{0,2} : (\psi_1, \ldots, \psi_k)^{\mathrm{T}} \leq \mathbf{0} \tag{2.2}$$

which is based only on the definitive endpoint data, where $\mathbf{0}$ denotes the zero vector. As mentioned in Section 2.1, we presume that the effect based on the intermediate endpoint is in the same direction, or consistent with, the effect based on the definitive endpoint. To do this, we assume $\psi_i = a_i\theta_i$ for some $a_i > 0$, in which case hypotheses (2.1) and (2.2) become equivalent. An outline of the unmodified procedure for known σ^2 and τ^2 and unknown $\rho \geq 0$ is described below.

At stage one of the study, $(k+1)n_1$ subjects are randomized with equal allocation to receive either one of the k active treatments or the control C. As a result, the first cohort will contain n_1 subjects in each group. At the first interim analysis, the most promising treatment is selected based on the intermediate endpoint data and used in the remaining stages. Let

$$\hat{\theta}_{i,1} = \frac{\bar{X}_{i\cdot,1} - \bar{X}_{0\cdot,1}}{\sqrt{\frac{2\sigma^2}{n_1}}} \sim \mathrm{N}(0,1)$$

denote the pairwise Z-statistics comparing active treatment E_i to the control C, where $\bar{X}_{i\cdot,1}$ and $\bar{X}_{0\cdot,1}$ are the average outcomes according to the stage one intermediate endpoint data among subjects in treatment group E_i and the control group C, respectively. Furthermore, let $S = \operatorname*{argmax}_{1\leq i\leq k} \hat{\theta}_{i,1}$ and $T_{S,1} = \hat{\theta}_{S,1}$, where S represents the index of the treatment with the largest observed effect, or the greatest promise, compared to the control C, with $T_{S,1}$ being its observed treatment effect. If $T_{S,1} \leq c_1$ for some $c_1 > 0$, then the trial is stopped early and $H_{0,1}$ (2.1) is accepted. Otherwise, if $T_{S,1} > c_1$, then treatment E_S is recommended as the most promising treatment and proceeds to the next stage for further investigation. Only subjects receiving treatment E_S or the control C will be followed for their definitive endpoint data. Treatment assessment on all other subjects will be terminated; they will be given standard of care off protocol and continue to undergo necessary safety monitoring.

At stage two, $2n_2$ additional subjects are equally randomized to receive either treatment E_S or the control C, where n_2 denotes the sample size in each treatment arm in the second cohort. The stage two interim analysis is conducted when both the intermediate measures for the $2n_2$ cohort two subjects and the definitive measures for the $2n_1$ cohort one subjects who received either treatment E_S or the control C become available. Let

$$W_{S,1} = \hat{\psi}_{S,1} = \frac{\bar{Y}_{S\cdot,1} - \bar{Y}_{0\cdot,1}}{\sqrt{\frac{2\tau^2}{n_1}}} \sim \mathrm{N}(0,1)$$

be the pairwise Z-statistic based on the definitive endpoint data from cohort one, where $\bar{Y}_{S\cdot,1}$ and $\bar{Y}_{0\cdot,1}$ are the average definitive outcomes among all cohort one subjects in treatment group E_S and the control group C, respectively. Also, let

$$\hat{\theta}_{S,2} = \frac{\bar{X}_{S\cdot,2} - \bar{X}_{0\cdot,2}}{\sqrt{\frac{2\sigma^2}{n_2}}} \sim \mathrm{N}(0,1)$$

be the pairwise Z-statistic based on the intermediate endpoint data from cohort two, where $\bar{X}_{S\cdot,2}$ and $\bar{X}_{0\cdot,2}$ are the average intermediate outcomes among all cohort two subjects in treatment group E_S and the control group C, respectively. Define

$$T_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\theta}_{S,2}.$$

If $T_{S,2} \leq c_2$ for some $c_2 > 0$, then the trial is stopped for futility and $H_{0,1}$ (2.1) is accepted. If $T_{S,2} > c_2$ and $W_{S,1} > d_1$ for some $d_1 > 0$, then the trial is stopped for superiority and both $H_{0,1}$ (2.1) and $H_{0,2}$ (2.2) are rejected. If $T_{S,2} > c_2$ but $W_{S,1} \leq d_1$, then the trial advances to the last stage.

At stage three, no new subjects are enrolled. The final analysis is performed when all cohort two subjects complete their definitive endpoints. Let

$$W_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\psi}_{S,2}$$

where

$$\hat{\psi}_{S,2} = \frac{\bar{Y}_{S\cdot,2} - \bar{Y}_{0\cdot,2}}{\sqrt{\frac{2\tau^2}{n_2}}} \sim \mathrm{N}(0, 1)$$

is the pairwise Z-test statistic based on the definitive endpoint data from cohort two and $\bar{Y}_{S\cdot,2}$ and $\bar{Y}_{0\cdot,2}$ are the average definitive outcomes among all cohort two subjects in treatment group E_S and the control group C, respectively. If $W_{S,2} > d_2$ for some $d_2 > 0$, then $H_{0,2}$ (2.2) is rejected and treatment E_S is deemed superior to the control. Otherwise, treatment E_S did not outperform the control and $H_{0,2}$ (2.2) is accepted.

In this procedure, $H_{0,1}$ (2.1) can be interpreted as a gatekeeper, since it must be rejected in stage one in order to move on to testing $H_{0,2}$ (2.2). Moreover, the sample size (n_1, n_2) and critical value (c_1, c_2, d_1, d_2) parameters are chosen via simulations in a manner that controls the pre-specified type I error rate while maintaining a targeted power, as no closed form equations are available or derivable to ascertain these values. The next two sections provide justification for the unmodified procedure and an explanation of the simulation process and findings. Lastly, if σ^2 and τ^2 are not known before trial start and they cannot be approximated in any way (e.g., by using historical data), then hypotheses (2.1) and (2.2) must be redefined to measure relative intermediate and definitive treatment effects, as σ^2 and τ^2 are vital to obtaining the sample size and critical value parameters above that are required to utilize the proposed design appropriately.

2.2.2 Statistical Justification

We now present statistical proofs to verify the unmodified version of the proposed design when employing a continuous intermediate endpoint and a continuous definitive endpoint. To confirm the validity of the unmodified design, we must first ensure that the type I error rate can be controlled (i.e., it will not exceed the nominal level). It can be shown that the size of the testing procedure outlined in the previous section, or the maximum probability of incorrectly rejecting $H_{0,1}$ (2.1) and $H_{0,2}$ (2.2), is achieved when $\theta_1 = \cdots = \theta_k = 0$ and $\psi_1 = \cdots = \psi_k = 0$ [119]. Denoting these two vectors as $\boldsymbol{\theta}_*$ and $\boldsymbol{\psi}_*$, respectively, we can write

$$
\begin{aligned}
\text{Size} &= P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \\
&\quad - kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \{W_{1,1} \le d_1\} \cap \{W_{1,2} \le d_2\}\Big) \\
&\le kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho\le 1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \\
&\quad - kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho\le 1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \{W_{1,1} \le d_1\} \cap \{W_{1,2} \le d_2\}\Big) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \\
&\quad - kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\left(\{S=1\} \cap \left\{c_1 < T_{1,1} \le \frac{\sigma d_1}{\tau}\right\} \cap \left\{c_2 < T_{1,2} \le \frac{\sigma d_2}{\tau}\right\}\right).
\end{aligned}
$$

The first equality is by definition. The second equality is obtained via the unmodified procedure (this is the set of events that lead to an ultimate rejection). Applying the rule $P(A \cap B) = P(A) - P(A \cap B^{\mathrm{c}})$, where B^{c} denotes the complement of B, and De Morgan's law [120] yields the third equality. The second term of the third equality depends on the unknown correlation $\rho \ge 0$, since it contains events related to both the intermediate and definitive endpoint data. The first inequality is obtained via Slepian's lemma, where equality holds when $\rho = 1$ (its maximum). The stage two and stage three test statistics based on the definitive endpoint data can be rescaled under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ as $W_{1,1} = \frac{\tau T_{1,1}}{\sigma}$ and $W_{1,2} = \frac{\tau T_{1,2}}{\sigma}$ since $T_{1,1}$ and $W_{1,1}$, and $T_{1,2}$ and $W_{1,2}$, have perfectly linear relationships when $\rho = 1$, resulting in the final equality. For exposition, $S = 1$ was chosen above, but any other S could be used for the proof since $P(S = i)$ has the same distribution under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ for all i. Therefore, we have identified an upper bound for the size of the unmodified procedure, which is a function of only the intermediate endpoint data. This gives us the type I error constraint

$$
P_{\boldsymbol{\theta}_*}\Big(T_{S,1} > c_1, T_{S,2} > c_2\Big) - P_{\boldsymbol{\theta}_*}\left(c_1 < T_{S,1} \le \frac{\sigma d_1}{\tau}, c_2 < T_{S,2} \le \frac{\sigma d_2}{\tau}\right) = \alpha. \tag{2.3}
$$

Next, the power under some clinically relevant alternative hypotheses should be no less than $1-\beta$ for some small β. Suppose $\delta_1 < \delta_2$ such that $\theta_i \geq \delta_2$ is a clinically meaningful effect, where $\theta_i \in [\delta_1, \delta_2)$ corresponds to a marginal effect measured by the intermediate endpoint. Furthermore, let $\eta_1 < \eta_2$ such that $\psi_i \geq \eta_2$ is a clinically meaningful effect, where $\psi_i \in [\eta_1, \eta_2)$ corresponds to a marginal effect measured by the definitive endpoint. Then it can be shown that the minimal power is achieved when $\theta_2 = \cdots = \theta_k = \delta_1, \theta_1 = \delta_2$, $\psi_2 = \cdots = \psi_k = \eta_1$, and $\psi_1 = \eta_2$. Denote these vectors as $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$, respectively. This alternative is known as the least favorable configuration and indicates that the clinically relevant effect occurs at active treatment $i = 1$ [119, 121]. However, similarly to the preceding proof for the size of the procedure, this is not essential, as the clinically relevant effect could occur in any other treatment group. Therefore, the following proof will still be valid for any other fixed treatment. Now, since the distribution of $P(S = i)$ is different for each i under the alternative, the power can be expressed as

$$\begin{aligned}
\text{Power} &= P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\rho)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= \sum_{i=1}^{k} P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\rho)}\Big(\{S = i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\} \cap (\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\})\Big) \\
&\geq \sum_{i=1}^{k} P_{\boldsymbol{\theta}^*}\Big(\{S = i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\}\Big) \\
&\geq P_{\boldsymbol{\theta}^*}\Big(\{S = 1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\Big).
\end{aligned}$$

The first equality is the definition of power and the second equality mirrors that in the proof for the size of the procedure above, but under the alternative. The first inequality is obtained by Slepian's lemma when $\rho = 0$ (its minimum here) and the last inequality follows by selecting any particular treatment ($i = 1$ is chosen here). Thus, a lower bound for the power has been attained, resulting in the constraint

$$P_{\boldsymbol{\theta}^*}\Big(T_{S,1} > c_1, T_{S,2} > c_2\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{S,1} > d_1\} \cup \{W_{S,2} > d_2\}\Big) = 1 - \beta. \tag{2.4}$$

Unlike condition (2.3) for calculating the type I error, condition (2.4) for calculating the power relies on both the intermediate and definitive endpoint data. Finally, in order to minimize the total expected sample size

$$\mathrm{E}(N) = (k+1)n_1 + 2n_2 P\left(T_{S,1} > c_1\right)$$

where N is the total number of subjects included in the trial, all of the design parameters are chosen such that

$$\frac{1}{2}\mathrm{E}(N|\boldsymbol{\theta}_*, \boldsymbol{\psi}_*) + \frac{1}{2}\mathrm{E}(N|\boldsymbol{\theta}^*, \boldsymbol{\psi}^*) = (k+1)n_1 + n_2\{P_{\boldsymbol{\theta}_*}(T_{S,1} > c_1) + P_{\boldsymbol{\theta}^*}(T_{S,1} > c_1)\} \tag{2.5}$$

is minimized [121]. Altogether, condition (2.3) guarantees that the nominal type I error rate will be preserved, condition (2.4) confirms that the procedure will have a power of at least $1-\beta$ and condition (2.5) ensures that the proposed design is not only valid but also optimal in sample size. Note that the proofs provided above apply Slepian's lemma in order to dodge having to handle ρ, the unknown correlation between the intermediate and definitive endpoints, or the exact sampling distribution of $(S, T_{S,1}, T_{S,2}, W_{S,1}, W_{S,2})^{\mathrm{T}}$, which would be very computationally challenging.

2.2.3 Simulations

To demonstrate the utility of the unmodified procedure proposed in Section 2.2.1, Monte Carlo simulations were conducted as follows. First, $(k+1)(n_1+n_2)$ samples were generated 500 times for $k = \{4, 6\}$ active arms, where $n_1 = \{20, 40, \ldots, 180, 200\}$ and $n_2 = \{20, 40, \ldots, 180, 200\}$. Under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho)$,

$$\begin{pmatrix} X_{0j} \\ Y_{0j} \end{pmatrix} \sim \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}\right]$$

and

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}\right]$$

for active treatments $i = 1, \ldots, k$ and subject j, where $i = 0$ denotes the control group, $\theta_1 = \cdots = \theta_k = 0$ and $\psi_1 = \cdots = \psi_k = 0$. Under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho)$,

$$\begin{pmatrix} X_{0j} \\ Y_{0j} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right]$$

and

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \theta_i \\ r\theta_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right]$$

where $r = 0.5$, $\theta_1 = \delta_2 = 1$, $\theta_2 = \ldots = \theta_k = \delta_1 = \{0.6, 0.8\}$, $\psi_1 = r\delta_2 = 0.5$, and $\psi_2 = \ldots = \psi_k = r\delta_1 = \{0.3, 0.4\}$. The effect size of active treatments $i = 2, \ldots, k$ and active treatment $i = 1$ (the most promising intervention) according to the intermediate endpoint data are represented by δ_1 and δ_2, respectively. We proceed using only conditions (2.3), (2.4) and (2.5), which do not require pre-specification of ρ. Thus, for condition (2.3), $X_{0j} \sim \mathrm{N}(0, \sigma^2)$ and $X_{ij} \sim \mathrm{N}(0, \sigma^2)$, while $X_{0j} \sim \mathrm{N}(0, \sigma^2)$, $X_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$, $Y_{0j} \sim \mathrm{N}(0, \tau^2)$, and $Y_{ij} \sim \mathrm{N}(r\theta_i, \tau^2)$ for condition (2.4). Note that condition (2.5) is used to calculate the average total expected sample size of the trial based on the intermediate endpoint data generated as described above. Here, $(\sigma^2, \tau^2) = \{(1, 1), (1, 2), (2, 1)\}$ were considered.

Condition (2.3) was calculated under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ as follows, using 500 replications. For stage one, which contains $(k + 1)n_1$ samples, intermediate endpoint data was generated for the control arm and the active treatment groups such that $X_{0j} \sim \mathrm{N}(0, \sigma^2)$ and $X_{ij} \sim \mathrm{N}(0, \sigma^2)$. Then, 500 pairwise Z-statistics ($\hat{\theta}_{i,1}$) were obtained comparing each active treatment $i = 1, \ldots, k$ to the control group C ($i = 0$) based on the intermediate endpoint data from the first cohort of subjects. The average of the 500 pairwise Z-statistics was acquired for each comparison in order to get $S = \underset{1 \leq i \leq k}{\operatorname{argmax}}\, \hat{\theta}_{i,1}$ and $T_{S,1} = \hat{\theta}_{S,1}$, where S was the "winning" treatment according to the highest average pairwise Z-statistic. If $T_{S,1} > c_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, then the trial continued to the second stage, and the number of times this occurred out of the 500 replications was recorded. For stage two, which is comprised of $2(n_1 + n_2)$ subjects, 500 pairwise Z-statistics comparing the "winning" treatment to the control were calculated based on the intermediate endpoint data from the second cohort, and then the average was taken to get weighted sum $T_{S,2}$, as defined in Section 2.2.1. The number of instances when $T_{S,1} > c_1$ and $T_{S,2} > c_2$, and when $c_1 < T_{S,1} \leq \frac{\sigma d_1}{\tau}$ and $c_2 < T_{S,2} \leq \frac{\sigma d_2}{\tau}$, were also obtained. For this study, $c_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$, $d_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and $d_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$ were considered. The three aforementioned events that were tracked were used to calculate the estimated type I error rate via condition (2.3) and the first probability in condition (2.5) to get the average total expected sample size under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$.

Condition (2.4) was calculated under the alternative $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 0)$ as follows, again using 500 replications. The first and second stage calculations using the intermediate endpoint data were identical to those described above for condition (2.3), but with the intermediate endpoint data for active treatment i and the control following the distributions $X_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$ and $X_{0j} \sim \mathrm{N}(0, \sigma^2)$, respectively, and setting $S = 1$. For condition (2.4), stage two also involved generating definitive endpoint data for the experimental interventions and the control such that $Y_{0j} \sim \mathrm{N}(0, \tau^2)$ and $Y_{ij} \sim \mathrm{N}(r\theta_i, \tau^2)$, respectively. With this, 500 pairwise Z-statistics ($\hat{\psi}_{i,1}$) comparing each active arm to the control group were obtained according to the definitive endpoint data, and then each set of pairwise Z-statistics was averaged. As was done for the intermediate endpoint data, we set $S = 1$ and then defined $W_{S,1} = \hat{\psi}_{S,1}$. For the third stage, 500 pairwise Z-statistics ($\hat{\psi}_{S,2}$) comparing the highest preforming treatment to the control were calculated based on the definitive endpoint data from the same $2n_2$ subjects from the second stage, with the averages obtained afterwards. The frequencies of $T_{S,1} > c_1$, $T_{S,1} > c_1$ and $T_{S,2} > c_2$, and $W_{S,1} > d_1$ or $W_{S,2} > d_2$ were recorded to determine the estimated power via condition (2.4) and the second probability in condition (2.5) to get the average total expected sample size under the alternative $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 0)$.

Finally, the sextuplets (n_1, n_2, c_1, c_2, d_1, d_2) with an estimated type I error rate ($\hat{\alpha}$) in the interval $(0.045, 0.05)$ and an estimated power $(1 - \hat{\beta})$ in the interval $(0.800, 0.820)$ were extracted. The scenarios chosen for reporting in Table 2.1 yielded an $\hat{\alpha}$ closest to 5% and a $1 - \hat{\beta}$ closest to 80%. If there was no scenario with $\hat{\alpha} = 0.05$ and $1 - \hat{\beta} = 0.800$ or there were multiple scenarios with

$\hat{\alpha} = 0.05$ and $1-\hat{\beta} = 0.800$, then the scenario with the smallest $\hat{N}_{\text{avg}}$, which denotes the average of the total estimated sample sizes under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 0)$, was reported. There were several scenarios in which $\hat{\alpha} < 0.05$ and $1-\hat{\beta} > 0.800$, but the intervals above were chosen to be conservative and to match the thresholds typically used in practice. Since 500 replications were used, an exact value of $\hat{\alpha} = 5\%$ could be achieved but reaching an exact value of $1-\hat{\beta} = 80\%$ would have required making the parameter grid infinitesimally small, as condition (2.4) is a product of probabilities (this was computationally infeasible). Table 2.1 exemplifies that the proposed procedure without interim sample size re-estimation is able to maintain a type I error rate of 5% and achieve a targeted power of at least 80% when utilizing a continuous intermediate endpoint and a continuous definitive endpoint. Future work will need to consider a smaller number of active interventions ($k = 2$ and $k = 3$), different effect sizes (δ_1) and larger variance values ($\sigma^2 > 2$ and $\tau^2 > 2$) in order to enhance the generalizability of the proposed unmodified procedure for continuous endpoints.

Table 2.1: Simulation examples for the proposed design procedure without interim sample size adjustment when using a continuous intermediate endpoint and a continuous definitive endpoint, where $\hat{N}_{\text{avg}}$ is the average estimated sample size required for the trial under the null and alternative hypotheses.

k	δ_1	σ^2	τ^2	n_1	n_2	c_1	c_2	d_1	d_2	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{N}_{\text{avg}}$
4	0.6	1	1	60	100	0.8	1.2	1	2	0.050	0.800	423
4	0.6	2	1	20	80	0.2	1.6	0.2	1.8	0.050	0.800	212
4	0.6	1	2	40	120	0.2	1.6	0.6	1.2	0.050	0.800	369
4	0.8	1	1	80	20	0.6	1.6	1	1.6	0.050	0.800	426
4	0.8	2	1	40	80	0.8	1.4	0.6	1.6	0.050	0.801	295
4	0.8	1	2	60	40	0.8	1.6	0.4	1.2	0.050	0.800	350
6	0.6	1	1	60	20	0.6	1.6	0.8	1.2	0.050	0.800	445
6	0.6	2	1	60	60	0.8	1.2	0.8	2	0.050	0.802	493
6	0.6	1	2	40	120	0.4	1.4	0.6	1.2	0.050	0.800	445
6	0.8	1	1	20	80	0.8	1.2	1	1.2	0.050	0.802	238
6	0.8	2	1	80	60	0.6	1.4	1	2	0.050	0.800	636
6	0.8	1	2	40	120	0.8	1.4	0.6	1.2	0.050	0.800	426

2.2.4 Conclusion

The proposed design in Figure 1.5 demonstrates promise in appropriately and expeditiously evaluating drugs and other medical products, and it has many beneficial features. First, it can save time by merging phase II and phase III studies rather than having to run them separately. It also identifies interventions that are not working early on to enhance patient safety and does not require any information about the correlation between the intermediate and definitive endpoints at trial start, which can be difficult to quantify (e.g., due to a lack of previous studies or adequate data). Therefore, the design helps to avert incorrect decision-making that could result from an inaccurate specification of the correlation. Our three-stage design (Figure 1.5) is also advantageous over the general two-stage design (Figure 1.3) in the sense that it gives the therapy under study more of a chance to demonstrate its efficacy if the stage two intermediate endpoint data is promising but the stage two primary data is not quite sufficient. This can protect against deeming a treatment ineffective when in fact it is truly efficacious. The procedure is also optimal in sample size via condition (2.5), which contributes to minimizing recruitment efforts for sponsors and avoids exposing superfluous subjects to the potential risks associated with participating in the trial.

We have rigorously proven that the size and power of the proposed unmodified procedure can be bounded, which shields patients from unreasonable harm due to an erroneous claim of efficacy. Simulations are also provided in Section 2.2.3 for demonstration to explicitly show that the unmodified procedure can feasibly control the type I error rate while still reaching the desired power. Although the scenarios reported from our numerical study in Table 2.1 are not comprehensive, user-friendly R functions have been developed and are expected to be integrated into an R package for the public, allowing investigators to formulate their own studies that use our design with any parameter values, including those beyond what we considered for our simulations.

2.3 Design with Sample Size Modification

The unmodified procedure in Section 2.2.1 can be reformatted to introduce a sample size re-assessment at the second interim analysis. This section will outline a procedure that modifies the pre-specified total required sample size based on participant information accumulated during the trial in progress. Then, as for the unmodified version of the proposed design, we will describe why the incorporation of this adaptation does not impact the operational characteristics (i.e., the nominal type I error rate and targeted power) of the trial. Lastly, corroborating simulation examples will be presented.

2.3.1 Procedure

The proposed design (Figure 1.5) allows investigators/sponsors to re-evaluate the sample size needed to achieve the goals of their study at the end of stage two based on the data collected from the first cohort of participants. When this adaptation is performed, the procedure in Section 2.2.1 remains the same up until the second interim analysis is completed. If superiority has not yet been demonstrated, then σ^2 and τ^2 can be estimated based on the accrued intermediate and definitive endpoint data, respectively, which we denote as $\hat{\sigma}^2$ and $\hat{\tau}^2$. With these, the trial can be redesigned using $\hat{\sigma}^2$ and $\hat{\tau}^2$ to obtain an updated cohort two sample size per treatment arm, which we call n_2^*. If it is found that $n_2^* > n_2$, where n_2 is the originally planned value, then $2(n_2^* - n_2)$ additional subjects can be recruited, randomly assigned to the highest performing treatment from stage one or the control group, and then followed for their intermediate and definitive endpoint data. Letting

$$\hat{\theta}_{S,2}^* = \frac{\bar{X}_{S\cdot,2} - \bar{X}_{0\cdot,2}}{\sqrt{\frac{2\sigma^2}{n_2^*}}}$$

denote the intermediate treatment effect and

$$\hat{\psi}_{S,2}^* = \frac{\bar{Y}_{S\cdot,2} - \bar{Y}_{0\cdot,2}}{\sqrt{\frac{2\tau^2}{n_2^*}}}$$

denote the definitive treatment effect using the updated sample size data, the weighted sums $T_{S,2}$ and $W_{S,2}$ in Section 2.2.1 can be replaced by

$$T^*_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\theta}^*_{S,2}$$

and

$$W^*_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\psi}^*_{S,2}$$

without compromising the validity (i.e., exceeding the nominal type I error rate) of the procedure. This will be explained further in the next section and then shown by simulations in Section 2.3.3.

The decision rules for the modified version of the proposed design are similar to the unmodified version. Now, if $T^*_{S,2} > c_2$ and $W_{S,1} > d_1$ at stage two, then the trial stops for superiority and $H_{0,1}$ (2.1) and $H_{0,2}$ (2.2) are rejected. If this is this case, then definitive endpoint data would not need to be gathered for the newly enrolled subjects. Otherwise, they must be followed for their definitive endpoint data in stage three. At the final analysis, if $W^*_{S,2} > d_2$, then $H_{0,2}$ (2.2) is rejected and the trial ends for superiority, while the trial ends without evidence of superiority if $W^*_{S,2} \leq d_2$.

2.3.2 Statistical Justification

The sample size recalibration described in the last section must be statistically verified in order to ensure that the modified version of the proposed design can be employed reliably in practice [31, 92, 122]. To do this, observe that under $\boldsymbol{\psi}_*$, as defined in Section 2.2.2, the distribution of $(S, T_{S,1}, T^*_{S,2}, W_{S,1}, W^*_{S,2})^{\mathrm{T}}$ is the same as the unmodified version. Therefore, the unmodified design can be applied to the modified test without inflating the pre-specified type I error rate. Further, for any i under $\boldsymbol{\psi}^*$, as defined in Section 2.2.2, the distribution of $(T_{i,1}, T^*_{i,2}, W_{i,1}, W^*_{i,2})^{\mathrm{T}}$ is stochastically larger than $(T_{i,1}, T_{i,2}, W_{i,1}, W_{i,2})^{\mathrm{T}}$ and so the modified procedure will actually have a power higher than the target. Together, these notions establish the validity of the proposed design with interim sample size reassessment when continuous intermediate and definitive endpoints are used.

2.3.3 Simulations

Monte Carlo simulations were also conducted to exhibit how implementing the sample size recalibration in the proposed design will not jeopardize the integrity of the ongoing trial. These sets of simulations were executed using a very similar process to those described for the unmodified procedure in Section 2.2.3. However, again assuming that the cohort one variances of the intermediate and definitive endpoint data for the active treatment and control groups were equal, conditions (2.3) and (2.4) were used to obtain the sample variances $\hat{\sigma}^2$ and $\hat{\tau}^2$. Then, using those estimates, conditions (2.3) and (2.4) were rerun, looping through $n_2^* = \{n_2, n_2 + 1, n_2 + 2, \ldots, n_2 + 200\}$ in order to obtain the scenario with the smallest average total expected sample size under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 0)$, which we call $\hat{N}^*_{\text{avg}}$, such that the estimated type I error using the updated sample size data ($\hat{\alpha}^*$) was in the interval $(0.045, 0.05)$ and the estimated power using the updated sample size data $(1 - \hat{\beta}^*)$ was in the interval $(0.800, 0.820)$. Our simulation results are provided in Table 2.2, where each row corresponds to the scenarios reported in Table 2.1 (i.e., the same k, δ_1, σ^2, τ^2, n_1, n_2, c_1, c_2, d_1, and d_2 parameters were used). The simulations shown in Table 2.2 demonstrate that the proposed three-stage phase II/III transitional seamless trial design (Figure 1.5) remains valid after recalibrating the total required sample size at the end of stage two when utilizing continuous intermediate and definitive endpoints.

2.3.4 Conclusion

We have provided statistical justification that our proposed design can appropriately incorporate a sample size adaptation at the end of stage two, and we have demonstrated this explicitly through simulations. The positive qualities of our unmodified procedure that were outlined in Section 2.2.4 also hold true for the modified version. But in addition, this sample size reassessment can help to rescue an ongoing trial if the stage two primary endpoint data is shy of an efficacy claim, without diminishing the integrity of the study. Therefore, our proposed sample size re-estimation process can avoid sponsors from having to terminate investigation (possibly prematurely) and reformulate a completely new trial.

Table 2.2: Simulation examples for the proposed design procedure with interim sample size adjustment when using a continuous intermediate endpoint and a continuous definitive endpoint, where $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{N}^*_{\text{avg}}$ are reported as the average under the null and the alternative and a (*) denotes a value that is calculated using n_2^*, the modified cohort two sample size per group.

k	δ_1	σ^2	τ^2	n_1	$\hat{\sigma}^2$	$\hat{\tau}^2$	n_2	n_2^*	$\hat{\alpha}^*$	$1-\hat{\beta}^*$	$\hat{N}^*_{\text{avg}}$
4	0.6	1	1	60	0.987	0.996	100	128	0.050	0.812	453
4	0.6	2	1	20	1.942	0.968	80	89	0.050	0.811	223
4	0.6	1	2	40	0.991	1.949	120	130	0.046	0.814	379
4	0.8	1	1	80	0.987	1.005	20	37	0.048	0.808	448
4	0.8	2	1	40	1.974	0.997	80	91	0.048	0.815	310
4	0.8	1	2	60	0.982	2.008	40	49	0.048	0.812	360
6	0.6	1	1	60	0.989	0.994	20	26	0.046	0.804	453
6	0.6	2	1	60	1.990	0.982	60	81	0.050	0.806	518
6	0.6	1	2	40	0.980	1.990	120	127	0.046	0.804	449
6	0.8	1	1	20	0.983	0.990	80	101	0.048	0.811	261
6	0.8	2	1	80	1.961	0.985	60	93	0.046	0.806	682
6	0.8	1	2	40	0.995	1.990	120	129	0.050	0.802	437

Chapter 3: Using a Binary Intermediate Endpoint and a Continuous Definitive Endpoint

In this chapter, the data modeling approach and the proposed design procedures with and without sample size adjustment are provided when using a binary intermediate endpoint and a continuous definitive endpoint, where the central limit theorem and a latent variable (i.e., an unobserved variable that can be inferred indirectly through other variables that are observed and measurable) are employed. The modified and unmodified procedures are then justified statistically, followed by simulations that concur with our arguments. Like in the previous chapter, the data model presented next is used for both versions of the proposed design and the unmodified version resembles a group sequential procedure that includes treatment selection.

3.1 Model

For subject j receiving active treatment $i = 1, \ldots, k$, define unobserved latent variable U_{ij}. Further, let $X_{ij} = \mathrm{I}(U_{ij} > 0) \sim \text{Bernoulli}(p_i)$ and Y_{ij} represent the binary intermediate endpoint and the continuous definitive endpoint, respectively, where $i = 0$ denotes the control group and $\mathrm{I}(\cdot)$ is the indicator function. Similarly to the last chapter (i.e., when employing continuous intermediate and definitive endpoints), we assume that for fixed cohort

$$\begin{pmatrix} U_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \mu_i \\ \nu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right]$$

for all i, j, in addition to the assumptions outlined in Section 2.1. The parameters in the model above are defined analogously as in Chapter 2, where μ_i and ν_i represent the mean of the latent variable and definitive endpoint data, respectively, for subjects randomly assigned to receive treatment i, and σ^2 and τ^2 are the variance of the latent variable and definitive endpoint data (which are considered equal across the treatment groups), with $\rho \geq 0$ being the correlation between the latent variable and definitive endpoint data. With this model, it can be shown that

$$\begin{aligned}
\mathrm{E}(X_{ij}) &= \mathrm{E}\{\mathrm{I}(U_{ij} > 0)\} \\
&= P(U_{ij} > 0) \\
&= P\left(\frac{U_{ij} - \mu_i}{\sigma} > \frac{0 - \mu_i}{\sigma}\right) \\
&= P\left(Z_{ij} > -\frac{\mu_i}{\sigma}\right) \\
&= P\left(Z_{ij} < \frac{\mu_i}{\sigma}\right) \\
&= \Phi\left(\frac{\mu_i}{\sigma}\right)
\end{aligned}$$

for all i, j, where $Z_{ij} \sim \mathrm{N}(0, 1)$ and Φ denotes the standard normal cumulative distribution function. Notice that $\mathrm{E}(X_{ij})$ does not depend on p_i (the equal probability that each subject allocated to treatment i will experience the intermediate event of interest), which is typically unknown *a priori* and thus not pre-specified in practice.

However, if μ_i and σ^2 are both unknown, then this would lead to an identifiability issue, since it would be impossible to obtain the true parameter values based solely on the data. For instance, for all participants randomly assigned to active treatment $i = 1$, $\mathrm{E}(X_{ij}) = \Phi(0.5)$ when $\mu_1 = 1$ and $\sigma^2 = 4$ and when $\mu_1 = 2$ and $\sigma^2 = 16$. Therefore, for this study, we set $\sigma^2 = 1$ to evade this problem, which is a common tactic in latent variable settings, but note that any other known value for σ^2 could also be used as long as it is carried through the unmodified and modified procedures presented in Section 3.2.1 and Section 3.3.1, respectively. For $\sigma^2 = 1$, $\mathrm{E}(X_{ij}) = \Phi(\mu_i)$. Moreover, since all X_{ij} are independent, asymptotic pairwise (two-sample) Z-statistics can be obtained via the central limit theorem, which will be used to compose the proposed design procedures with and without sample size recalculation. Then, Slepian's lemma will be applied to verify their statistical validity, with supporting simulations that exemplify their utility.

3.2 Design without Sample Size Modification

This section will impart the statistical details of the proposed design procedure without interim sample size reassessment when utilizing a binary intermediate endpoint and a continuous definitive endpoint. This is proceeded by proofs of its ability to control the pre-specified type I error rate and achieve the desired power. Finally, simulations are presented that jibe with our statistical claims and demonstrate the legitimacy of the proposed unmodified procedure in different scenarios.

3.2.1 Procedure

Suppose a sponsor intends to run a clinical trial that implements the proposed three-stage phase II/III transitional seamless design shown in Figure 1.5 (under the assumptions stated in Section 2.1) to compare k active treatment groups E_i, $i = 1, \ldots, k$, to a control group C ($i = 0$), where a binary intermediate endpoint (defined through a latent variable) and a continuous definitive endpoint will be utilized to measure the treatment effects. Like in the last chapter, let θ_i and ψ_i be the treatment effects comparing active treatment E_i to the control C measured via the binary intermediate endpoint and the continuous definitive endpoint, respectively. Thus, with $\mathbf{0}$ denoting the zero vector, the hypothesis of interest based on the intermediate endpoint is

$$H_{0,1} : (\theta_1, \ldots, \theta_k)^{\mathrm{T}} \leq \mathbf{0} \tag{3.1}$$

and the hypothesis of ultimate interest is

$$H_{0,2} : (\psi_1, \ldots, \psi_k)^{\mathrm{T}} \leq \mathbf{0} \tag{3.2}$$

which depends only on the definitive endpoint data. Note that the hypotheses above are identical to hypotheses (2.1) and (2.2) in the previous chapter. The proposed unmodified design procedure with $\sigma^2 = 1$, known τ^2 and unknown correlation $\rho \geq 0$ is described next, with a subsequent remark regarding what must be done if τ^2 is unknown before trial start.

At stage one, $(k+1)n_1$ subjects are randomized and equally allocated to one of the k experimental interventions or the control group C, where n_1 is the sample size per treatment arm for cohort one. At the first interim analysis, the most promising treatment (if any) is selected according to the binary intermediate endpoint data from cohort one. Let

$$\hat{\theta}_{i,1} = \frac{\bar{X}_{i\cdot,1} - \bar{X}_{0\cdot,1}}{\sqrt{\frac{\bar{X}_{i\cdot,1}(1-\bar{X}_{i\cdot,1})+\bar{X}_{0\cdot,1}(1-\bar{X}_{0\cdot,1})}{n_1}}}$$

denote the asymptotic pairwise Z-statistics for active treatment E_i compared to the control C, where $\bar{X}_{i\cdot,1}$ and $\bar{X}_{0\cdot,1}$ are the average intermediate outcomes based on the cohort one data from treatment group E_i and the control group C, respectively. Then by the central limit theorem, $\hat{\theta}_{i,1} \xrightarrow{d} \mathrm{N}(0,1)$ for large n_1. Further, let $S = \underset{1\le i\le k}{\operatorname{argmax}}\, \hat{\theta}_{i,1}$ and $T_{S,1} = \hat{\theta}_{S,1}$, where S denotes the treatment arm with the highest performance compared to the control group C and $T_{S,1}$ is the corresponding observed effect. If $T_{S,1} \le c_1$ for some $c_1 > 0$, then $H_{0,1}$ (3.1) is accepted and the trial is stopped. However, if $T_{S,1} > c_1$, then treatment E_S moves forward as the most promising intervention, advancing to the next stage for further assessment. Subjects receiving treatments other than E_S or the control C in stage one will not be followed for their definitive endpoint data in the second stage; treatment evaluation on these subjects will be terminated, but they will be given standard of care and continue to be monitored for safety signals.

At the second stage, $2n_2$ new subjects are recruited, enrolled and equally randomized to receive either the highest performing treatment from the first stage (E_S) or the control C, where n_2 is the cohort two sample size per arm. The second interim analysis begins after the intermediate endpoint data is collected from the $2n_2$ cohort two participants and the definitive endpoint data is obtained from the $2n_1$ cohort one participants who were randomly assigned to treatment E_S or the control. Let

$$W_{S,1} = \hat{\psi}_{S,1} = \frac{\bar{Y}_{S\cdot,1} - \bar{Y}_{0\cdot,1}}{\sqrt{\frac{2\tau^2}{n_1}}} \sim \mathrm{N}(0,1)$$

be the pairwise Z-statistic at the second stage according to the definitive endpoint data, where $\bar{Y}_{S\cdot,1}$

and $\bar{Y}_{0\cdot,1}$ are the average definitive outcomes among all cohort one participants in active treatment group E_S and the control group C, respectively. Next, let

$$\hat{\theta}_{S,2} = \frac{\bar{X}_{S\cdot,2} - \bar{X}_{0\cdot,2}}{\sqrt{\frac{\bar{X}_{S\cdot,2}(1-\bar{X}_{S\cdot,2})+\bar{X}_{0\cdot,2}(1-\bar{X}_{0\cdot,2})}{n_2}}}$$

be the asymptotic pairwise Z-statistic based on the cohort two intermediate endpoint data, where $\bar{X}_{S\cdot,2}$ and $\bar{X}_{0\cdot,2}$ are the average outcomes among all cohort two subjects receiving active treatment E_S and the control C, respectively. Then $\hat{\theta}_{S,2} \xrightarrow{d} \mathrm{N}(0,1)$ for large n_2 by the central limit theorem. Next, define the sum of the cohort one and cohort two treatment effects of E_S versus the control C, estimated using the accumulated intermediate endpoint data and weighted by the information fractions, as

$$T_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\theta}_{S,2}.$$

If $T_{S,2} \leq c_2$ for some $c_2 > 0$, then $H_{0,1}$ (3.1) is accepted and the trial is stopped for futility. But, if $T_{S,2} > c_2$ and $W_{S,1} > d_1$ for some $d_1 > 0$, then both $H_{0,1}$ (3.1) and $H_{0,2}$ (3.2) are rejected and the trial is stopped for superiority. Otherwise, if $T_{S,2} > c_2$ and $W_{S,1} \leq d_1$, then the trial moves on to stage three in order to gather and assess more definitive endpoint data.

At the third and last stage, no more participants are added to the study. The final analysis can be performed after all of the definitive endpoint data is collected from the second cohort. Like what was done with the treatment effects based on the intermediate endpoint data in the previous stage, define the sum of the cohort one and cohort two effects comparing treatment E_S to the control C, estimated using the definitive endpoint data and weighted by the information fractions, as

$$W_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\psi}_{S,2}$$

where

$$\hat{\psi}_{S,2} = \frac{\bar{Y}_{S\cdot,2} - \bar{Y}_{0\cdot,2}}{\sqrt{\frac{2\tau^2}{n_2}}} \sim \mathrm{N}(0,1)$$

is the pairwise Z-statistic using the cohort two definitive endpoint data, with $\bar{Y}_{S\cdot,2}$ and $\bar{Y}_{0\cdot,2}$ denoting the average definitive outcomes among the cohort two subjects in treatment arm E_S and the control arm C, respectively. If $W_{S,2} > d_2$ for some $d_2 > 0$, then $H_{0,2}$ (3.2) is rejected and intervention E_S is considered superior to the control C. Otherwise, intervention E_S did not outperform the control and $H_{0,2}$ (3.2) is accepted.

In parallel to the unmodified procedure presented in Section 2.2.1 for when a continuous intermediate endpoint and a continuous definitive endpoint are used, $H_{0,1}$ (3.1) serves as a gatekeeper; testing $H_{0,1}$ (3.1) using the binary intermediate endpoint data precedes testing $H_{0,2}$ (3.2) using the continuous definitive data endpoint data. And again, the critical value and sample size parameters c_1, c_2, d_1, d_2, n_1, and n_2 must be obtained via simulations; this cannot be completed analytically. However, this process can be proven to not inflate the nominal type I error rate and still reach the targeted power, which will be explained in the next section. Some simulation results that support our statistical findings will be provided in Section 3.2.3. Similarly to the unmodified procedure proposed in the previous chapter, if the variance of the definitive endpoint data (τ^2) is unknown *a priori* and there are no means to estimate it, then hypotheses (3.1) and (3.2) must be restructured to test for relative treatment effects rather than absolute treatment effects, since a value for τ^2 is essential in ascertaining the required parameters mentioned above. The unmodified procedure detailed here also allows for the sample size to be adapted at interim according to the intermediate and definitive endpoint data from the first cohort without impacting the operational characteristics of the study, which will be discussed in Section 3.3.

3.2.2 Statistical Justification

To verify the unmodified procedure proposed in the last section for when a binary intermediate endpoint and a continuous definitive endpoint are employed, we obtain the sample size and critical value parameters (c_1, c_2, d_1, d_2, n_1, n_2) under three conditions, which will be derived next when $\sigma^2 = 1$, τ^2 is known and correlation $\rho \geq 0$ is unknown. First, since every X_{ij} and Y_{ij} is independent and identically distributed (iid) for fixed treatment group i, subject j and cohort, we know that because $(U_{ij}, Y_{ij})^{\mathrm{T}}$ follows the bivariate normal distribution stated in Section 3.1, $U_{ij} \sim \mathrm{N}(\mu_i, 1)$ and $Y_{ij}|U_{ij} \sim \mathrm{N}(\nu_i + \rho\tau(U_{ij} - \mu_i), 1 - \rho^2)$. Thus, $U_{ij} = \mu_i + Z_{ij}$, where $Z_{ij} \sim \mathrm{N}(0, 1)$. With this, the law of iterated expectations [123] and the fundamental theorem of calculus [124], we can write

$$
\begin{aligned}
\mathrm{Cov}(X_{ij}, Y_{ij}) &= \mathrm{E}(X_{ij}Y_{ij}) - \mathrm{E}(X_{ij})\mathrm{E}(Y_{ij}) \\
&= \mathrm{E}\{\mathrm{I}(U_{ij} > 0)Y_{ij}\} - \Phi(\mu_i)\nu_i \\
&= \mathrm{E}\{\mathrm{I}(U_{ij} > 0)\mathrm{E}(Y_{ij}|U_{ij})\} - \Phi(\mu_i)\nu_i \\
&= \mathrm{E}\{\mathrm{I}(U_{ij} > 0)(\nu_i + \rho\tau(U_{ij} - \mu_i))\} - \Phi(\mu_i)\nu_i \\
&= \mathrm{E}\{\nu_i\mathrm{I}(U_{ij} > 0)\} + \mathrm{E}\{\rho\tau U_{ij}\mathrm{I}(U_{ij} > 0)\} - \mathrm{E}\{\rho\tau\mu_i\mathrm{I}(U_{ij} > 0)\} - \Phi(\mu_i)\nu_i \\
&= \nu_i\mathrm{E}\{\mathrm{I}(U_{ij} > 0)\} + \rho\tau\mathrm{E}\{U_{ij}\mathrm{I}(U_{ij} > 0)\} - \rho\tau\mu_i\mathrm{E}\{\mathrm{I}(U_{ij} > 0)\} - \Phi(\mu_i)\nu_i \\
&= \nu_i\mathrm{E}(X_{ij}) + \rho\tau\mathrm{E}\{(\mu_i + Z_{ij})\mathrm{I}(\mu_i + Z_{ij} > 0)\} - \rho\tau\mu_i\mathrm{E}(X_{ij}) - \Phi(\mu_i)\nu_i \\
&= \nu_i\Phi(\mu_i) + \rho\tau[\mathrm{E}\{\mu_i\mathrm{I}(Z_{ij} > -\mu_i)\} + \mathrm{E}\{Z_{ij}\mathrm{I}(Z_{ij} > -\mu_i)\}] - \rho\tau\mu_i\Phi(\mu_i) - \Phi(\mu_i)\nu_i \\
&= \rho\tau\mu_i\mathrm{E}\{\mathrm{I}(Z_{ij} > -\mu_i)\} + \rho\tau\int_{-\mu_i}^{\infty}\frac{z}{\sqrt{2\pi}}e^{-z^2/2}dz - \rho\tau\mu_i\Phi(\mu_i) \\
&= \rho\tau\mu_i P(Z_{ij} < \mu_i) + \frac{\rho\tau}{\sqrt{2\pi}}e^{-\mu_i^2/2} - \rho\tau\mu_i\Phi(\mu_i) \\
&= \rho\tau\mu_i\Phi(\mu_i) + \frac{\rho\tau}{\sqrt{2\pi}}e^{-\mu_i^2/2} - \rho\tau\mu_i\Phi(\mu_i) \\
&= \frac{\rho\tau}{\sqrt{2\pi}}e^{-\mu_i^2/2}
\end{aligned}
$$

which is a strictly increasing function of ρ. Therefore, by the central limit theorem,

$$
\sqrt{n_i}\begin{pmatrix} \bar{X}_{i\cdot} - \Phi(\mu_i) \\ \bar{Y}_{i\cdot} - \nu_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Phi(\mu_i)(1 - \Phi(\mu_i)) & \frac{\rho\tau}{\sqrt{2\pi}}e^{-\mu_i^2/2} \\ \frac{\rho\tau}{\sqrt{2\pi}}e^{-\mu_i^2/2} & \tau^2 \end{pmatrix}\right]
$$

where $\bar{X}_{i\cdot}$ and $\bar{Y}_{i\cdot}$ are the average intermediate and definitive outcomes, respectively, among the individuals receiving treatment i. Therefore, the joint distribution of $(\bar{X}_{i\cdot}, \bar{Y}_{i\cdot})^{\mathrm{T}}$ can be approximately treated as bivariate normal, similar to the model described in Section 2.1, and so the corresponding justification in Section 2.2.2 can be applied approximately here.

For proving that the unmodified procedure can maintain the pre-specified type I error rate, we again use the fact that the probability of erroneously rejecting $H_{0,1}$ (3.1) and $H_{0,2}$ (3.2), which is also known as the size of the procedure, is maximal when $\theta_1 = \cdots = \theta_k = 0$ and $\psi_1 = \cdots = \psi_k = 0$ [119]. Specifically, letting these vectors be represented by $\boldsymbol{\theta}_*$ and $\boldsymbol{\psi}_*$, respectively,

$$\begin{aligned}
\text{Size} &= P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho)}\Big(\{S=1\}\cap\{T_{1,1}>c_1\}\cap\{T_{1,2}>c_2\}\cap\big(\{W_{1,1}>d_1\}\cup\{W_{1,2}>d_2\}\big)\Big) \\
&\lesssim kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho\le 1)}\Big(\{S=1\}\cap\{T_{1,1}>c_1\}\cap\{T_{1,2}>c_2\}\cap\big(\{W_{1,1}>d_1\}\cup\{W_{1,2}>d_2\}\big)\Big) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{S=1\}\cap\{T_{1,1}>c_1\}\cap\{T_{1,2}>c_2\}\cap\big(\{W_{1,1}>d_1\}\cup\{W_{1,2}>d_2\}\big)\Big).
\end{aligned}$$

The first, second and third lines are similar to the proof in Section 2.2.2 when using a continuous intermediate endpoint and a continuous definitive endpoint. However, an asymptotic inequality ($\lesssim$) follows from applying Slepian's lemma asymptotically, assuming the limiting distribution described above, where equality holds when the correlation ρ is maximal (i.e., when $\rho = 1$), which yields the last equality. For simplicity, $S = 1$ was selected as the most promising treatment for the proof. Since the distribution of $P(S = i)$ is equivalent for every treatment under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho)$, any of the treatments could have be used with the same steps. Thus, an asymptotic upper bound for the size of the unmodified procedure has been identified, yielding our first design constraint:

$$P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{T_{S,1}>c_1\}\cap\{T_{S,2}>c_2\}\cap\big(\{W_{S,1}>d_1\}\cup\{W_{S,2}>d_2\}\big)\Big)=\alpha. \tag{3.3}$$

Contrary to condition (2.3) derived in the previous chapter, condition (3.3) relies on both the intermediate and definitive endpoint data.

For establishing a condition to ensure that the power of the unmodified version of the proposed design will be no smaller than the desired $1 - \beta$ under the alternative, we define vectors $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$ as in Section 2.2.2, which represent the least favorable configuration for any fixed treatment E_i and achieve the minimum power among all other possibilities [119, 121]. With this, we can express the power of the proposed unmodified procedure as

$$
\begin{aligned}
\text{Power} &= P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\rho)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= \sum_{i=1}^{k} P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\rho)}\Big(\{S=i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\} \cap (\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\})\Big) \\
&\gtrsim \sum_{i=1}^{k} P_{\boldsymbol{\theta}^*}\Big(\{S=i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\}\Big) \\
&\geq P_{\boldsymbol{\theta}^*}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\Big).
\end{aligned}
$$

The steps above are identical to those shown in Section 2.2.2 when using a continuous intermediate endpoint and a continuous definitive endpoint, except that the first inequality is now replaced by an asymptotic inequality ($\gtrsim$). Therefore, we have identified a lower bound for the power to obtain the constraint

$$P_{\boldsymbol{\theta}^*}\Big(T_{S,1} > c_1, T_{S,2} > c_2\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{S,1} > d_1\} \cup \{W_{S,2} > d_2\}\Big) = 1 - \beta. \tag{3.4}$$

Note that condition (3.4) is the same as condition (2.4), the power constraint obtained in the last chapter. Again, to minimize the total expected sample size of the trial, which is

$$\mathrm{E}(N) = (k+1)n_1 + 2n_2 P\left(T_{S,1} > c_1\right)$$

where the total number of participants is denoted by N, the design parameters $(c_1, c_2, d_1, d_2, n_1, n_2)$ are chosen as in Section 2.2.2 such that

$$\frac{1}{2}\mathrm{E}(N|\boldsymbol{\theta}_*, \boldsymbol{\psi}_*) + \frac{1}{2}\mathrm{E}(N|\boldsymbol{\theta}^*, \boldsymbol{\psi}^*) = (k+1)n_1 + n_2\{P_{\boldsymbol{\theta}_*}(T_{S,1} > c_1) + P_{\boldsymbol{\theta}^*}(T_{S,1} > c_1)\} \tag{3.5}$$

is minimized [121].

The three conditions presented here assure that the proposed unmodified procedure will have robust operational characteristics and an optimal/minimal sample size. Like in the previous chapter, the size and power proofs above avoid having to deal with the unknown correlation ρ or derive the complicated joint distribution of $(S, T_{S,1}, T_{S,2}, W_{S,1}, W_{S,2})^{\mathrm{T}}$. These proofs were also completed under the assumption that τ^2 is known, in which case the required parameter values can be obtained to employ the design reliably. Hypotheses (3.1) and (3.2) must be reoriented to evaluate treatment effects relative to a particular level of variability in the definitive endpoint data if τ^2 is unknown before trial commencement.

3.2.3 Simulations

To exemplify the capability of the unmodified version of the proposed design when utilizing a binary intermediate endpoint and a continuous definitive endpoint, Monte Carlo simulations were conducted that followed the procedure outlined in Section 3.2.1. Here, $k = \{4, 6\}$ active treatment groups were considered, with each of the $k + 1$ treatment groups having cohort one and cohort two sample sizes of $n_1 = \{20, 40, \ldots, 180, 200\}$ and $n_2 = \{20, 40, \ldots, 180, 200\}$ per arm, respectively. Trial data for $(k+1)(n_1+n_2)$ subjects was generated 500 times for conditions (3.3), (3.4) and (3.5), which do not rely on the unknown correlation ρ, where $\sigma^2 = 1$ and $\tau^2 = \{1, 2\}$. The following holds for fixed cohort. For participant j randomly assigned to the control group ($i = 0$) or active treatment $i = 1, \ldots, k$,

$$\begin{pmatrix} X_{0j} \\ Y_{0j} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \Phi(\mu_0) \\ \nu_0 \end{pmatrix}, \begin{pmatrix} \Phi(\mu_0)(1 - \Phi(\mu_0)) & \frac{\tau}{\sqrt{2\pi}} e^{-\mu_0^2/2} \\ \frac{\tau}{\sqrt{2\pi}} e^{-\mu_0^2/2} & \tau^2 \end{pmatrix} \right]$$

and

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \Phi(\mu_i) \\ \nu_i \end{pmatrix}, \begin{pmatrix} \Phi(\mu_i)(1 - \Phi(\mu_i)) & \frac{\tau}{\sqrt{2\pi}} e^{-\mu_i^2/2} \\ \frac{\tau}{\sqrt{2\pi}} e^{-\mu_i^2/2} & \tau^2 \end{pmatrix} \right]$$

under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$. For the purposes of this study, we set $\mu_i = 0$ and $\nu_i = 0$ for all i, which in turn makes $\theta_i = \Phi(\mu_i) - \Phi(\mu_0) = 0$ and $\psi_i = \nu_i - \nu_0 = 0$ for all i. Then under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$,

$$\begin{pmatrix} X_{0j} \\ Y_{0j} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \Phi(\mu_0) \\ \nu_0 \end{pmatrix}, \begin{pmatrix} \Phi(\mu_0)(1 - \Phi(\mu_0)) & 0 \\ 0 & \tau^2 \end{pmatrix} \right]$$

and

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} \Phi(\mu_i) \\ \nu_i \end{pmatrix}, \begin{pmatrix} \Phi(\mu_i)(1 - \Phi(\mu_i)) & 0 \\ 0 & \tau^2 \end{pmatrix} \right]$$

where we considered $r = 0.5$, $\mu_0 = 0$, $\theta_1 = 0.4 = \delta_2$, $\theta_2 = \cdots = \theta_k = \{0.1, 0.2, 0.3\} = \delta_1$, $\nu_0 = 0$, $\nu_i = r\Phi(\mu_i)$ for $i = 1, \ldots, k$, $\psi_1 = 0.45$, and $\psi_2 = \cdots = \psi_k = \{0.3, 0.35, 0.4\}$. Note that for condition (3.3), the intermediate and definitive endpoint data must be generated jointly, while $X_{ij} \sim \mathrm{N}(\Phi(\mu_i), \Phi(\mu_i)(1 - \Phi(\mu_i)))$ and $Y_{ij} \sim \mathrm{N}(\nu_i, \tau^2)$ for all i for condition (3.4). As in the last chapter, the average total expected sample size of the study is obtained via condition (3.5), which is a function of only the intermediate endpoint data under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$.

To begin the simulation process, condition (3.3) was calculated under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and replicated 500 times. For the first stage of the study, comprised of $(k + 1)n_1$ individuals, intermediate endpoint data was generated for the active treatment and control groups as described above. Afterwards, 500 asymptotic pairwise Z-statistics $(\hat{\theta}_{i,1})$ were obtained comparing each active arm $i = 1, \ldots, k$ to the control group ($i = 0$) based on the intermediate endpoint data from the first cohort of trial participants. Then the average of the 500 asymptotic pairwise Z-statistics was acquired for each comparison to get $S = \underset{1 \le i \le k}{\operatorname{argmax}}\, \hat{\theta}_{i,1}$ and $T_{S,1} = \hat{\theta}_{S,1}$, where S denotes the highest performing treatment according to the largest average Z-statistic and $T_{S,1}$ is the respective value. If $T_{S,1} > c_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, then the trial advanced to the next stage and the frequency that this occurred out of the 500 replications was obtained. In the second stage of the study, which contains $2(n_1 + n_2)$ participants, 500 asymptotic pairwise Z-statistics comparing the most promising treatment from stage one to the control were calculated based on the intermediate endpoint data from the second cohort of participants, and then the average was taken to get weighted sum $T_{S,2}$, as defined in Section 3.2.1. The number of occurrences of $T_{S,1} > c_1$ and $T_{S,2} > c_2$ were collected. For this study, $c_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$, $d_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and $d_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$ were considered. Each of the events mentioned in this paragraph had to be tracked throughout the 500 replications in order to calculate the estimated type I error rate using condition (3.3) and the average total expected sample size under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ via condition (3.5).

Then, condition (3.4) was calculated under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ and was also replicated 500 times. The calculations for the first two stages of the trial using the intermediate endpoint data were the same as those detailed above for condition (3.3); however, we chose $S = 1$ and the data for the control and active arms followed the distributions $X_{0j} \sim \mathrm{N}(\Phi(\mu_0), \Phi(\mu_0)(1-\Phi(\mu_0)))$ and $X_{ij} \sim \mathrm{N}(\Phi(\mu_i), \Phi(\mu_i)(1-\Phi(\mu_i)))$, respectively. Stage two also involved generating definitive endpoint data for the control and experimental groups, where $Y_{0j} \sim \mathrm{N}(\nu_0, \tau^2)$ and $Y_{ij} \sim \mathrm{N}(\nu_i, \tau^2)$, for condition (3.4). With this, 500 pairwise Z-statistics ($\hat{\psi}_{i,1}$) comparing each active therapy to the control arm were calculated based on the definitive endpoint data from the first cohort of participants, and then each set of Z-statistics was averaged. Similarly to what was completed with the intermediate endpoint data in stage one, $W_{S,1} = \hat{\psi}_{S,1}$ was acquired using the definitive endpoint data. Finally, for the last stage of the trial, 500 pairwise Z-statistics ($W_{S,2} = \hat{\psi}_{S,2}$) comparing the highest preforming treatment from stage one and the control were obtained according to the definitive endpoint data from the $2n_2$ cohort one participants in the second stage, with the average calculated subsequently. The number of occurrences of $T_{S,1} > c_1$, $T_{S,1} > c_1$ and $T_{S,2} > c_2$, and $W_{S,1} \leq d_1$ and $W_{S,2} \leq d_2$ were recorded for determining the estimated power of the procedure via condition (2.4) and the average total expected sample size under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ using condition (3.5).

For tabulating the simulation results, the scenarios in which the estimated type I error rate ($\hat{\alpha}$) was in the interval $(0.045, 0.05)$ and the estimated power $(1-\hat{\beta})$ was in the interval $(0.800, 0.820)$ were extracted. The scenarios reported in Table 3.1 produced an $\hat{\alpha}$ closest to 5%, a $1-\hat{\beta}$ closest to

80% and the smallest $\hat{N}_{\text{avg}}$, which represents the average of the total estimated sample sizes under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$. In this numerical study, many scenarios yielded better estimated type I error and power values (i.e., $\hat{\alpha} < 0.045$ and $1 - \hat{\beta} > 0.820$), but the intervals above were utilized to reflect the common targets of $\alpha = 0.05$ and $1-\beta = 0.80$ in practice. Again, due to the parameter grid that was considered, an exact value of $1 - \hat{\beta} = 80\%$ could not be achieved. Shrinking the parameter gradations would have made the simulation process extremely computationally intense. Altogether, Table 3.1 demonstrates that the proposed unmodified design can preserve a nominal type I error rate of 5% and reach a target power of 80% when employing a binary intermediate endpoint and a continuous definitive endpoint. Assessing the robustness of our proposed unmodified procedure for different values of k, δ_1 and τ^2 will help to solidify its validity and increase confidence in its utility.

Table 3.1: Simulation examples for the proposed design procedure without interim sample size adjustment when using a binary intermediate endpoint and a continuous definitive endpoint, where $\hat{N}_{\text{avg}}$ is the average estimated sample size required for the trial under the null and alternative hypotheses.

k	δ_1	τ^2	n_1	n_2	c_1	c_2	d_1	d_2	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{N}_{\text{avg}}$
4	0.1	1	20	20	0.2	1.4	0.8	1.6	0.046	0.813	129
4	0.1	2	60	40	0.2	1.4	1	2	0.046	0.806	356
4	0.2	1	20	40	0.4	1.4	0.8	2	0.048	0.815	155
4	0.2	2	40	80	0.8	1.2	1	1.8	0.046	0.802	298
4	0.3	1	20	20	0.4	1.4	0.8	1.6	0.046	0.811	127
4	0.3	2	40	60	0.6	1.2	1	1.6	0.046	0.812	276
6	0.1	1	20	40	0.8	1.2	1	1.8	0.046	0.813	189
6	0.1	2	60	20	0.2	1.6	1	2	0.046	0.814	448
6	0.2	1	20	20	0.4	1.4	0.6	1.8	0.050	0.803	166
6	0.2	2	40	40	0.6	1.4	0.8	1.8	0.050	0.808	332
6	0.3	1	20	40	0.6	1.2	1	1.8	0.048	0.818	192
6	0.3	2	40	20	1.0	1.2	0.8	1.4	0.050	0.802	303

3.2.4 Conclusion

In Chapter 2, we focused on when the intermediate and definitive endpoints are on a continuous scale. However, different types of endpoints cater to different needs and disease areas, and selection can depend on many intrinsic and extrinsic factors, like financial resources and the availability of certain devices or tools. There are various clinical settings in which a binary intermediate endpoint is commonly employed, such as oncology. Although binary endpoints are less informative than continuous endpoints, we were still able to prove that the unmodified procedure proposed in Section 3.2.1 is asymptotically valid in terms of its ability to preserve the nominal type I error rate and achieve the desired power with a minimal sample size. Again, simulation examples were reported for demonstration purposes and do not represent every possible scenario that could be encountered in practice, but R functions similar to those that were created for the continuous-continuous data scenario in Chapter 2 have also been built for the binary-continuous data scenario addressed in this chapter. When utilizing a binary intermediate endpoint and a continuous definitive endpoint, the benefits of our design outlined in Section 2.2.4 remain the same but in an asymptotic sense due to our application of the central limit theorem.

3.3 Design with Sample Size Modification

The unmodified version of our proposed design procedure presented in Section 3.2.1, when using a binary intermediate endpoint and a continuous definitive endpoint, can be reconstructed to introduce a sample size recalibration at the end of stage two without weakening the persuasiveness of the trial. This section starts by describing what alterations are made for the modified procedure and then illustrates how these changes do not influence the integrity (i.e., inflate the pre-specified type I error rate) of the study via statistical rigor and simulations.

3.3.1 Procedure

As in the previous chapter, our proposed design (Figure 1.5) permits the total required sample size of the trial to be reassessed during the second interim analysis based on the definitive endpoint data accumulated from the first cohort of subjects. For the modified version of the proposed design, a few important steps are added to the unmodified procedure furnished in Section 3.2.1. In particular, if the product has not been shown to outperform the control based on the definitive endpoint data at the end of stage two, then τ^2 can be estimated according to the accrued cohort one data and signified by $\hat{\tau}^2$. Equipped with this estimate, the trial can be rerun to obtain an adjusted cohort two sample size per arm (referred to as n_2^*) that is required to accomplish the objectives of the trial in progress. In the event that n_2^* is larger than the original n_2, then $2(n_2^* - n_2)$ new participants can be enrolled and randomly allocated to the most promising active therapy from stage one or the control to gather their intermediate and definitive endpoint data. The proposed modified procedure for $\sigma^2 = 1$, known τ^2 and unknown correlation $\rho \geq 0$ is as follows.

Again, when sample size modification is introduced into the procedure, the trial process stays the same until the second interim analysis is conducted. Now, let

$$\hat{\theta}^*_{S,2} = \frac{\bar{X}_{S\cdot,2} - \bar{X}_{0\cdot,2}}{\sqrt{\frac{\bar{X}_{S\cdot,2}(1-\bar{X}_{S\cdot,2})+\bar{X}_{0\cdot,2}(1-\bar{X}_{0\cdot,2})}{n_2^*}}}$$

and

$$\hat{\psi}^*_{S,2} = \frac{\bar{Y}_{S\cdot,2} - \bar{Y}_{0\cdot,2}}{\sqrt{\frac{2\tau^2}{n_2^*}}}$$

be the updated intermediate and definitive treatment effects comparing the most promising treatment from stage one (E_S) to the control group C, respectively. Then, the weighted sums $T_{S,2}$ and $W_{S,2}$ defined in Section 3.2.1 can be exchanged with

$$T^*_{S,2} = \sqrt{\frac{n_1}{n_1 + n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1 + n_2}}\hat{\theta}^*_{S,2}$$

and

$$W_{S,2}^{*} = \sqrt{\frac{n_1}{n_1 + n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1 + n_2}}\hat{\psi}_{S,2}^{*}$$

without impacting the nominal type I error rate of the procedure. Of course, the desired power can also still be achieved after recalculating the sample size. The statistical rationale for these points will be discussed in Section 3.3.2 and then demonstrated via simulations in Section 3.3.3.

After recalibrating the sample size during the second interim analysis, the decision rules stated in Section 3.2.1 change slightly. Namely, if $T_{S,2}^{*} > c_2$ and $W_{S,1} > d_1$, then $H_{0,1}$ (3.1) and $H_{0,2}$ (3.2) are rejected and the trial ends with superiority of active treatment E_S compared to the control. If this occurs, the newly recruited participants would no longer need to be followed for their definitive endpoint data. Otherwise, the study proceeds and they must be followed up with in the third stage. At the final analysis, if $W_{S,2}^{*} > d_2$, then $H_{0,2}$ (3.2) is rejected and the trial ends with superiority of treatment E_S over the control C. However, if $W_{S,2}^{*} \leq d_2$, then the definitive endpoint data does not support treatment E_S over the control and $H_{0,2}$ (3.2) cannot be rejected.

3.3.2 Statistical Justification

It is imperative to ensure that adapting the sample size at interim (using the variance of the cohort one definitive endpoint data) will not compromise the results of the trial. First, since the distributions of $(S, T_{S,1}, T_{S,2}, W_{S,1}, W_{S,2})^{\mathrm{T}}$ and $(S, T_{S,1}, T_{S,2}^{*}, W_{S,1}, W_{S,2}^{*})^{\mathrm{T}}$ are asymptotically equivalent under the null $(\boldsymbol{\theta}_{*}, \boldsymbol{\psi}_{*}, \rho = 1)$, the pre-specified type I error rate will not be surpassed. Furthermore, since $(T_{i,1}, T_{i,2}^{*}, W_{i,1}, W_{i,2}^{*})^{\mathrm{T}}$ is asymptotically stochastically larger than $(T_{i,1}, T_{i,2}, W_{i,1}, W_{i,2})^{\mathrm{T}}$ for any i under the alternative $(\boldsymbol{\theta}^{*}, \boldsymbol{\psi}^{*}, \rho = 0)$, the updated sample size will yield a power that is in fact greater than the targeted power. Therefore, our modified procedure proposed in Section 3.3.1 is statistically valid via asymptotic reasoning when using a binary intermediate endpoint and a continuous definitive endpoint.

3.3.3 Simulations

Monte Carlo simulations were conducted in order to explicitly show that the operational characteristics of an underway study will not be affected after incorporating the proposed sample size reassessment procedure into our design. These simulations were conducted in a manner similar to those run for the unmodified procedure, but included the additional steps outlined in Section 3.3.1. Specifically, this involved applying conditions (3.3) and (3.4) to calculate $\hat{\tau}^2$ and then running conditions (3.3) and (3.4) again for $n_2^* = \{n_2, n_2+1, n_2+2, \ldots, n_2+200\}$. This allowed us to identify the scenario that had the lowest average expected sample size under the null ($\boldsymbol{\theta}_*, \boldsymbol{\psi}_*,\ \rho = 1$) and the alternative ($\boldsymbol{\theta}^*, \boldsymbol{\psi}^*,\ \rho = 0$), where $\hat{\alpha}^* \in (0.045, 0.05)$ and $1-\hat{\beta}^* \in (0.800, 0.820)$. These results are provided in Table 3.2. As in the last chapter, each scenario in Table 3.2 matches the unmodified scenario in Table 3.1; in other words, for simplicity, the set of parameters (k, δ_1, τ^2, n_1, n_2, c_1, c_2, d_1, d_2) associated with the scenarios reported in Table 3.1 were used for the scenarios reported in Table 3.2, which indeed indicate that the modified procedure is able to control the type I error rate and reach the desired power.

Table 3.2: Simulation examples for the proposed design procedure with interim sample size adjustment when using a binary intermediate endpoint and a continuous definitive endpoint, where $\hat{\tau}^2$ and $\hat{N}^*_{\text{avg}}$ are reported as the average under the null and the alternative and a (*) denotes a value that is calculated using n_2^*, the modified cohort two sample size per group.

k	δ_1	τ^2	n_1	$\hat{\tau}^2$	n_2	n_2^*	$\hat{\alpha}^*$	$1-\hat{\beta}^*$	$\hat{N}^*_{\text{avg}}$
4	0.1	1	20	1.011	20	26	0.048	0.816	137
4	0.1	2	60	1.998	40	52	0.048	0.809	371
4	0.2	1	20	0.982	40	50	0.050	0.816	168
4	0.2	2	40	2.005	80	85	0.050	0.807	307
4	0.3	1	20	1.002	20	33	0.050	0.813	147
4	0.3	2	40	1.986	60	76	0.046	0.816	305
6	0.1	1	20	1.022	40	47	0.048	0.816	202
6	0.1	2	60	1.993	20	29	0.050	0.818	458
6	0.2	1	20	0.994	20	31	0.048	0.805	180
6	0.2	2	40	1.988	40	44	0.050	0.810	338
6	0.3	1	20	1.016	40	46	0.050	0.820	203
6	0.3	2	40	2.007	20	24	0.046	0.808	310

3.3.4 Conclusion

The remarks stated in Section 2.3.4 for the modified version of the proposed design procedure in the continuous-continuous data scenario carry over to the binary-continuous data scenario here, but asymptotically. Our statistical arguments for the binary-continuous scenario demonstrate that a sample size recalibration can be appropriately conducted at interim, and we have also exemplified this through simulations. Again, this data-driven sample size adaptation can help to prevent a trial from being incorrectly stopped, which would in turn keep an intervention that is actually effective out of the hands of populations that are in urgent need of care.

Chapter 4: Using a Binary Intermediate Endpoint and a Time-to-Event Definitive Endpoint without Censoring

This chapter begins with a presentation of our data modeling approach when a binary intermediate endpoint and a time-to-event definitive endpoint are used in the absence of censored observations, which utilizes the central limit theorem, latent variables, and the copula technique [114, 115]. Next, the proposed design procedure (without interim sample size re-estimation) is explained, followed by justification of its validity (i.e., its preservation of the nominal type I error rate and ability to attain the targeted power) and simulations that underscore its utility and reinforce our statistical claims. Like in the previous two chapters, the procedure described in Section 4.2 can be perceived as a general group sequential process that implements treatment selection.

4.1 Model

Consider a clinical trial that will use the proposed three-stage phase II/III transitional seamless design (Figure 1.5) to compare k active treatments to a single control. For each subject j randomly assigned to treatment group $i = 0, \ldots, k$, where $i = 0$ represents the control arm and each group is the same size, let X_{ij} denote the binary intermediate endpoint and Y_{ij} denote the time since the observation of X_{ij} to the event of interest (i.e., the time-to-event definitive endpoint) for fixed cohort. Furthermore, we assume that marginally $X_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_i)$, where p_i is the equal probability of each subject in treatment group i experiencing the intermediate event of interest. We also assume that marginally $Y_{ij} \overset{iid}{\sim} \text{Exponential}(\xi_i) = \frac{1}{\xi_i} \exp\left\{-\frac{Y_{ij}}{\xi_i}\right\}$, where ξ_i is the mean of the time-to-event definitive endpoint data among all subjects receiving treatment i.

To introduce a correlation between the intermediate and definitive endpoint data, we apply the copula idea, which enables us to decompose the joint distribution of $(X_{ij}, Y_{ij})^{\mathrm{T}}$ into the marginal distributions of X_{ij} and Y_{ij} and a function that merges them together (i.e., the copula) in order to separately specify their dependency structure. Suppose there are unobserved latent random vectors $(U_{ij}, V_{ij})^{\mathrm{T}}$ such that

$$\begin{pmatrix} U_{ij} \\ V_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

for all i, j. Here, $\rho > 0$ denotes the unknown correlation between latent variables U_{ij} and V_{ij},

$$X_{ij} = \mathrm{I}(U_{ij} > -\Phi^{-1}(p_i))$$

and

$$Y_{ij} = -\xi_i \log(1 - \Phi(V_{ij}))$$

where $\mathrm{I}(\cdot)$ is the indicator function and Φ is the standard normal cumulative distribution function with inverse Φ^{-1}. Defining X_{ij} and Y_{ij} in this manner (i.e., using the probability integral transform [125]) ensures that their marginal distributions are indeed as stated in the last paragraph, and we again proceed under the set of assumptions outlined in Section 2.1.

Since $X_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_i)$ and $Y_{ij} \overset{iid}{\sim} \text{Exponential}(\xi_i)$, $\mathrm{E}(X_{ij}) = p_i$, $\mathrm{E}(Y_{ij}) = \xi_i$, $\mathrm{Var}(X_{ij}) = p_i(1 - p_i)$, and $\mathrm{Var}(Y_{ij}) = \xi_i^2$. Therefore, by the central limit theorem,

$$\sqrt{n_i} \begin{pmatrix} \bar{X}_{i\cdot} - p_i \\ \bar{Y}_{i\cdot} - \xi_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_i \right]$$

where

$$\Sigma_i = \begin{pmatrix} p_i(1 - p_i) & \gamma\sqrt{p_i(1 - p_i)}\xi_i \\ \gamma\sqrt{p_i(1 - p_i)}\xi_i & \xi_i^2 \end{pmatrix}.$$

The average of the intermediate and definitive outcomes among subjects randomized to treatment i are represented by $\bar{X}_{i\cdot}$ and $\bar{Y}_{i\cdot}$, respectively, and γ is the unknown correlation between the endpoints.

Although γ is unknown, it can be shown that

$$\begin{aligned}\gamma &= \frac{\mathrm{Cov}(X_{ij}, Y_{ij})}{\sqrt{\mathrm{Var}(X_{ij})\mathrm{Var}(Y_{ij})}} \\ &= \frac{\mathrm{E}(X_{ij}Y_{ij}) - p_i\xi_i}{\xi_i\sqrt{p_i(1-p_i)}}\end{aligned}$$

is an increasing function of ρ. Because $(U_{ij}, V_{ij})^{\mathrm{T}}$ follows the bivariate normal distribution specified above, $V_{ij} \sim \mathrm{N}(0, 1)$ and $U_{ij}|V_{ij} \sim \mathrm{N}(\rho V_{ij}, 1-\rho^2)$. Applying these facts in conjunction with the law of iterated expectations [123], we can write

$$\begin{aligned}\zeta = \mathrm{E}(X_{ij}Y_{ij}) &= \mathrm{E}\left\{-\mathrm{I}(U_{ij} > -\Phi^{-1}(p_i))\xi_i \log(1-\Phi(V_{ij}))\right\} \\ &= \mathrm{E}\left\{-\xi_i \log(1-\Phi(V_{ij}))\mathrm{E}\{\mathrm{I}(U_{ij} > -\Phi^{-1}(p_i))|V_{ij}\}\right\} \\ &= \mathrm{E}\left\{-\xi_i \log(1-\Phi(V_{ij}))P(U_{ij} > -\Phi^{-1}(p_i)|V_{ij})\right\} \\ &= \mathrm{E}\left\{-\xi_i \log(1-\Phi(V_{ij}))P\left(\frac{U_{ij}-\rho V_{ij}}{\sqrt{1-\rho^2}} > \frac{-\Phi^{-1}(p_i)-\rho V_{ij}}{\sqrt{1-\rho^2}}\middle| V_{ij}\right)\right\} \\ &= \mathrm{E}\left\{-\xi_i \log(1-\Phi(V_{ij}))\Phi\left(\frac{\Phi^{-1}(p_i)+\rho V_{ij}}{\sqrt{1-\rho^2}}\right)\right\}\end{aligned}$$

which is demonstrated to be a strictly increasing function of the correlation ρ via simulations in Figure A.1.

Next, letting $\theta_i = p_i - p_0$ and $\psi_i = \xi_i - \xi_0$ represent the treatment effects comparing active intervention i to the control according to the intermediate and definitive endpoint data, respectively, we can again use the central limit theorem to obtain

$$\sqrt{n_i}\begin{pmatrix} (\bar{X}_{i\cdot} - \bar{X}_{0\cdot}) - \theta_i \\ (\bar{Y}_{i\cdot} - \bar{Y}_{0\cdot}) - \psi_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_i + \Sigma_0\right]$$

where $\bar{X}_{0\cdot}$ and $\bar{Y}_{0\cdot}$ are the average intermediate and definitive outcomes, respectively, among subjects in the control group and

$$\Sigma_0 = \begin{pmatrix} p_0(1-p_0) & \gamma\sqrt{p_0(1-p_0)}\xi_0 \\ \gamma\sqrt{p_0(1-p_0)}\xi_0 & \xi_0^2 \end{pmatrix}.$$

Utilizing this data model, we will construct asymptotic pairwise (two-sample) Z-statistics in our proposed procedure, which is outlined in the next section. Similarly to the previous chapter when employing a binary intermediate endpoint and a continuous definitive endpoint, Slepian's lemma can be applied asymptotically to statistically validate our procedure, showing that it does not compromise the nominal type I error rate and is able to reach a targeted level of power, and thus can be employed soundly in practice. This is proceeded by simulations in Section 4.4 that coincide with our statistical justifications and demonstrate that the design functions well in several settings.

4.2 Procedure

Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)^{\mathrm{T}}$ are the vectors of treatment effects comparing active treatment $i = 1, \ldots, k$ to the control arm ($i = 0$) based on the intermediate and definitive endpoint data, respectively, where $\theta_i = p_i - p_0$ and $\psi_i = \xi_i - \xi_0$ for all i. We are interested in testing the hypotheses

$$H_{0,1} : \boldsymbol{\theta} \leq \mathbf{0} \tag{4.1}$$

and

$$H_{0,2} : \boldsymbol{\psi} \leq \mathbf{0} \tag{4.2}$$

where $\mathbf{0}$ is the zero vector. Recall that these hypotheses are the same as those that were considered in the last two chapters. The proposed procedure when a binary intermediate endpoint and a time-to-event definitive endpoint are employed, without conducting an interim sample size recalibration, is described below in the absence of censoring.

Starting at the first stage, $(k + 1)n_1$ subjects are equally and randomly allocated to receive one of the k active therapies or the control, where n_1 denotes the sample size per group for cohort one.

Define asymptotic pairwise Z-statistics

$$\hat{\theta}_{i,1} = \frac{\bar{X}_{i\cdot,1} - \bar{X}_{0\cdot,1}}{\sqrt{\frac{\bar{X}_{i\cdot,1}(1-\bar{X}_{i\cdot,1})+\bar{X}_{0\cdot,1}(1-\bar{X}_{0\cdot,1})}{n_1}}}$$

comparing each active intervention i to the control, where $\bar{X}_{i\cdot,1}$ and $\bar{X}_{0\cdot,1}$ are the average intermediate outcomes according to the data from the stage one subjects in active treatment arm i and the control group, respectively. Therefore, for large n_1, $\hat{\theta}_{i,1} \xrightarrow{d} \mathrm{N}(0,1)$ by the central limit theorem. These asymptotic pairwise Z-statistics, calculated at the first interim analysis, are used to select the most promising treatment to examine further, if one exists, based on the binary intermediate endpoint data. Then, $S = \operatorname*{argmax}_{1 \le i \le k} \hat{\theta}_{i,1}$ is defined as the highest performing treatment at the end of the first stage, where $T_{S,1} = \hat{\theta}_{S,1}$ is the corresponding effect. With the data accrued during stage one, a go/no-go decision can be made as follows. The trial is stopped early and $H_{0,1}$ (4.1) is accepted if $T_{S,1} \le c_1$ for some $c_1 > 0$. However, $H_{0,1}$ (4.1) is rejected and the trial proceeds to the second stage if $T_{S,1} > c_1$, in which case the most promising therapy continues to be evaluated compared to the control and all other arms are eliminated. Thus, only subjects randomly allocated to the highest performing intervention or the control will be followed for their definitive endpoint measures. All subjects from the removed treatment groups will continue necessary safety monitoring and be given the standard of care therapy.

For stage two, $2n_2$ additional subjects will enter the trial and be randomized with 1:1 allocation to receive either the most promising treatment chosen in the previous stage or the control, where n_2 denotes the sample size in each group for cohort two. After gathering the binary intermediate endpoint data from subjects in the first and second cohorts and the definitive endpoint data from subjects in the first cohort, the second interim analysis can be run. Define the asymptotic pairwise Z-statistic

$$\hat{\theta}_{S,2} = \frac{\bar{X}_{S\cdot,2} - \bar{X}_{0\cdot,2}}{\sqrt{\frac{\bar{X}_{S\cdot,2}(1-\bar{X}_{S\cdot,2})+\bar{X}_{0\cdot,2}(1-\bar{X}_{0\cdot,2})}{n_2}}}$$

based on the cohort two intermediate endpoint data, where $\bar{X}_{S\cdot,2}$ and $\bar{X}_{0\cdot,2}$ are the average intermediate outcomes among the subjects in cohort two receiving the most promising treatment from stage one and the control, respectively. Also, let

$$W_{S,1} = \hat{\psi}_{S,1} = \frac{\bar{Y}_{S\cdot,1} - \bar{Y}_{0\cdot,1}}{\sqrt{\frac{\bar{Y}^2_{S\cdot,1} + \bar{Y}^2_{0\cdot,1}}{n_1}}}$$

be the asymptotic pairwise Z-statistic according to the cohort two definitive endpoint data, where $\bar{Y}_{S\cdot,1}$ and $\bar{Y}_{0\cdot,1}$ denote the average definitive outcomes among the subjects in cohort one receiving the highest performing treatment from the first stage and the control, respectively. Like in the prior stage, $\hat{\theta}_{S,2} \xrightarrow{d} \mathrm{N}(0,1)$ and $\hat{\psi}_{S,1} \xrightarrow{d} \mathrm{N}(0,1)$ by the central limit theorem. Finally, define

$$T_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\theta}_{S,2}$$

as the weighted sum of the cohort one and cohort two intermediate treatment effects comparing the most promising therapy to the control. We employ $T_{S,2}$ and $W_{S,1}$ to make the following go/no-go decision. The trial is stopped for futility and $H_{0,1}$ (4.1) is accepted if $T_{S,2} \le c_2$ for some $c_2 > 0$. The trial is stopped for superiority and $H_{0,1}$ (4.1) and $H_{0,2}$ (4.2) are rejected when $T_{S,2} > c_2$ and $W_{S,1} > d_1$ for some $d_1 > 0$. The trial advances to the final stage if the intermediate endpoint data from the first and second cohorts is promising but the definitive endpoint data from the first cohort is not yet compelling (i.e., $T_{S,2} > c_2$ but $W_{S,1} \le d_1$).

Lastly, at stage three, new subjects do not join the trial. After the cohort two subjects complete their definitive endpoints, the final analysis can be conducted. Define the asymptotic pairwise Z-statistic

$$\hat{\psi}_{S,2} = \frac{\bar{Y}_{S\cdot,2} - \bar{Y}_{0\cdot,2}}{\sqrt{\frac{\bar{Y}^2_{S\cdot,2} + \bar{Y}^2_{0\cdot,2}}{n_2}}}$$

comparing the top performing intervention to the control according to the definitive endpoint data from the second cohort of subjects, where $\bar{Y}_{S\cdot,2}$ and $\bar{Y}_{0\cdot,2}$ denote the average of the definitive outcomes from the cohort two subjects receiving the most promising treatment and the control, respectively. By the central limit theorem, $\hat{\psi}_{S,2} \xrightarrow{d} \mathrm{N}(0,1)$. Similarly to the last stage, let

$$W_{S,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\psi}_{S,2}$$

be the weighted sum of the cohort one and cohort two definitive treatment effects. The experimental therapy being studied is considered superior to the control when $H_{0,2}$ (4.2) is rejected (i.e., when $W_{S,2} > d_2$ for some $d_2 > 0$) and is not considered superior to the control when $H_{0,2}$ (4.2) is accepted (i.e., when $W_{S,2} \leq d_2$).

As in the other data scenarios that have been discussed in the previous two chapters, $H_{0,1}$ (4.1), which is based solely on the intermediate endpoint data, can be seen as a gatekeeper for testing $H_{0,2}$ (4.2), which depends solely on the definitive endpoint data. Also, simulations are still required to determine the design parameters (c_1, c_2, d_1, d_2, n_1, n_2), since there are no explicit equations to calculate them directly, but obtaining the parameters in this way will not inflate the pre-specified type I error rate of the study or prevent a certain power from being achieved. This will be proven in the next section and illustrated by simulations in Section 4.4.

4.3 Statistical Justification

The first step in verifying our proposed procedure when using a binary intermediate endpoint and a time-to-event definitive endpoint without censoring events (outlined in the previous section) is to ensure that the nominal type I error rate will not be exceeded. This can be proven identically as in the previous chapter by utilizing Slepian's lemma asymptotically, since $(\bar{X}_{i\cdot}, \bar{Y}_{i\cdot})^{\mathrm{T}}$ follows the approximately bivariate normal distribution stated in Section 4.1; namely,

$$\sqrt{n_i}\begin{pmatrix} \bar{X}_{i\cdot} - p_i \\ \bar{Y}_{i\cdot} - \xi_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} p_i(1-p_i) & \gamma\sqrt{p_i(1-p_i)}\xi_i \\ \gamma\sqrt{p_i(1-p_i)}\xi_i & \xi_i^2 \end{pmatrix}\right]$$

for all $i = 0, \ldots, k$. Still, the maximum probability of wrongly rejecting $H_{0,1}$ (4.1) and $H_{0,2}$ (4.2) can be shown to occur when $\theta_1 = \cdots = \theta_k = 0$ and $\psi_1 = \cdots = \psi_k = 0$ [119]. Again, letting $\boldsymbol{\theta}_*$ and

$\boldsymbol{\psi}_*$ denote these zero vectors,

$$\begin{aligned}
\text{Size} &= P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\gamma)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\gamma)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big) \\
&\lesssim kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho\le 1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big).
\end{aligned}$$

Recall that the first inequality above is an asymptotic inequality ($\lesssim$), which is obtained by applying Slepian's lemma asymptotically. This can be done since we can show that γ is a strictly increasing function of ρ in our model in Section 4.1. Thus, we have found an upper bound for the size, which gives us the constraint

$$P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{T_{S,1} > c_1\} \cap \{T_{S,2} > c_2\} \cap \big(\{W_{S,1} > d_1\} \cup \{W_{S,2} > d_2\}\big)\Big) = \alpha. \tag{4.3}$$

Note that this condition is the same as condition (3.3), the type I error constraint obtained when using a binary intermediate endpoint and a continuous definitive endpoint. Thus, the type I error condition when using a binary intermediate endpoint and a time-to-event definitive endpoint, when no censoring occurs, also depends on both the intermediate and definitive endpoint data. Under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$, $U_{ij} = V_{ij}$ and thus

$$\begin{aligned}
\gamma &= \frac{\mathrm{E}(X_{ij}Y_{ij}) - \mathrm{E}(X_{ij})\mathrm{E}(Y_{ij})}{\xi_i\sqrt{p_i(1-p_i)}} \\
&= \frac{\mathrm{E}\left\{-\mathrm{I}(U_{ij} > -\Phi^{-1}(p_i))\xi_i \log(1-\Phi(U_{ij}))\right\} - p_i\xi_i}{\xi_i\sqrt{p_i(1-p_i)}} \\
&= \frac{\mathrm{E}\left\{-\mathrm{I}(W_{ij} > 1-p_i)\xi_i \log(1-W_{ij})\right\} - p_i\xi_i}{\xi_i\sqrt{p_i(1-p_i)}} \\
&= \frac{-\xi_i\int_{1-p}^{1}\log(1-w)dw - p_i\xi_i}{\xi_i\sqrt{p_i(1-p_i)}} \\
&= \frac{-\sqrt{p_i}\log p_i}{\sqrt{1-p_i}}
\end{aligned}$$

where $W_{ij} \overset{iid}{\sim} \text{Uniform}(0,1)$ since $U_{ij} \overset{iid}{\sim} \text{N}(0,1)$ and $V_{ij} \overset{iid}{\sim} \text{N}(0,1)$.

We also must make sure that the procedure proposed in the previous section can reach a power that is no less than the target of $1-\beta$. To do this, we define $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$ to be the same least favorable configuration as for the last two data scenarios in Section 2.2.2 and Section 3.2.2, which yield the smallest possible power compared to every other available choice [119, 121]. We can then write the power of the procedure as

$$\begin{aligned}
\text{Power} &= P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\gamma)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= \sum_{i=1}^{k} P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\gamma)}\Big(\{S=i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\} \cap (\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\})\Big) \\
&\gtrsim \sum_{i=1}^{k} P_{\boldsymbol{\theta}^*}\Big(\{S=i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\}\Big) \\
&\geq P_{\boldsymbol{\theta}^*}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\Big)
\end{aligned}$$

where again Slepian's lemma is used asymptotically, resulting in the first inequality becoming an asymptotic inequality ($\gtrsim$). These steps provide us with a lower bound for the power, which gives us the constraint

$$P_{\boldsymbol{\theta}^*}\Big(T_{S,1} > c_1, T_{S,2} > c_2\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{S,1} > d_1\} \cup \{W_{S,2} > d_2\}\Big) = 1-\beta. \tag{4.4}$$

This condition is equivalent to those obtained for the prior two pairs of endpoint types discussed in Chapters 2-3. Finally, with N being the total number of trial subjects, we again minimize the total expected sample size of the study

$$\text{E}(N) = (k+1)n_1 + 2n_2 P\left(T_{S,1} > c_1\right)$$

by selecting the design parameters c_1, c_2, d_1, d_2, n_1, and n_2 such that we minimize

$$\frac{1}{2}\text{E}\left[N|(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*)\right] + \frac{1}{2}\text{E}\left[N|(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*)\right] = (k+1)n_1 + n_2\left\{P_{\boldsymbol{\theta}_*}(T_{S,1} > c_1) + P_{\boldsymbol{\theta}^*}(T_{S,1} > c_1)\right\}. \tag{4.5}$$

Conditions (4.3), (4.4) and (4.5) establish that the proposed design procedure maintains the prespecified type I error rate, achieves a power no smaller than the target, and is optimal in sample size, respectively. Therefore, the procedure presented in Section 4.2.1 is anticipated to operate well in practice. As in the previous two chapters, Slepian's lemma was employed so that we did not have to manage the unknown correlation between the intermediate and definitive endpoints or determine the elaborate joint distribution of $(S, T_{S,1}, T_{S,2}, W_{S,1}, W_{S,2})^{\mathrm{T}}$, which would be very computationally intensive.

4.4 Simulations

With the constraints stated in the last section, Monte Carlo simulations were used to demonstrate the utility of the proposed design described in Section 4.2 when utilizing a binary intermediate endpoint and a time-to-event definitive endpoint when there are no censored observations. For this study, $k = \{4, 6\}$ equally sized investigational arms were considered, where the first and second cohorts had sample sizes of $n_1 = \{20, 40, \ldots, 180, 200\}$ and $n_2 = \{20, 40, \ldots, 180, 200\}$ in each group, respectively. First, conditions (4.3), (4.4) and (4.5) were applied to generate intermediate and definitive endpoint data for $(k+1)(n_1+n_2)$ subjects, since these constraints do not depend on the correlation, which we want to view as unknown. Under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$, for fixed cohort,

$$\begin{pmatrix} X_{0j} \\ Y_{0j} \end{pmatrix} \approx \mathrm{N}_2 \left[\begin{pmatrix} p_0 \\ \xi_0 \end{pmatrix}, \begin{pmatrix} p_0(1-p_0) & -\xi_0 p_0 \log p_0 \\ -\xi_0 p_0 \log p_0 & \xi_0^2 \end{pmatrix} \right]$$

and

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \approx \mathrm{N}_2 \left[\begin{pmatrix} p_i \\ \xi_i \end{pmatrix}, \begin{pmatrix} p_i(1-p_i) & -\xi_i p_i \log p_i \\ -\xi_i p_i \log p_i & \xi_i^2 \end{pmatrix} \right]$$

for every subject j randomly allocated to the control group ($i = 0$) or active treatment $i = 1, \ldots, k$, where we set $p_0 = p_i$ and $\xi_0 = \xi_i$ for all i. Similarly, under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$, for fixed cohort,

$$\begin{pmatrix} X_{0j} \\ Y_{0j} \end{pmatrix} \approx \mathrm{N}_2 \left[\begin{pmatrix} p_0 \\ \xi_0 \end{pmatrix}, \begin{pmatrix} p_0(1-p_0) & 0 \\ 0 & \xi_0^2 \end{pmatrix} \right]$$

and

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \approx \mathrm{N}_2 \left[\begin{pmatrix} p_i \\ \xi_i \end{pmatrix}, \begin{pmatrix} p_i(1-p_i) & 0 \\ 0 & \xi_i^2 \end{pmatrix} \right]$$

where $p_0 = 0.5$, $p_1 = 0.75$, $p_2 = \cdots = p_k = \{0.6, 0.65, 0.7\}$, $\xi_0 = 1$, $\xi_1 = 3$, and $\xi_2 = \cdots = \xi_k = \{1.5, 2, 2.5\}$. Thus, $\theta_1 = p_1 - p_0 = 0.25$, $\theta_2 = \cdots = \theta_k = \{0.1, 0.15, 0.2\} = \delta_1$, $\psi_1 = \xi_1 - \xi_0 = 2$, and $\psi_2 = \cdots = \psi_k = \{0.5, 1, 1.5\} = \delta_2$. As in the last chapter, the intermediate and definitive endpoint data must be produced jointly under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ due to the nature of condition (4.3). However, for condition (4.4), the intermediate and definitive endpoint data can be generated separately for all i under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$, where $X_{ij} \sim \mathrm{N}(p_i, p_i(1-p_i))$ and $Y_{ij} \sim \mathrm{N}(\xi_i, \xi_i^2)$. Additionally, condition (4.5) is for calculating the total expected sample size of the trial, averaged under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$, and is based only on the intermediate endpoint data.

The simulations were conducted as follows. For condition (4.3), all intermediate endpoint data for the investigational and control groups was generated 500 times under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ distributions above for each scenario considered in our parameter grid. For stage one, which has $(k+1)n_1$ subjects, 500 asymptotic pairwise Z-statistics ($\hat{\theta}_{i,1}$) were calculated, which compared each active therapy $i = 1, \ldots, k$ to the control arm according to the intermediate endpoint data that was produced for subjects in the first cohort. Then, the 500 asymptotic pairwise Z-statistics were averaged for each treatment group. Afterwards, $S = \underset{1 \leq i \leq k}{\operatorname{argmax}} \hat{\theta}_{i,1}$ and $T_{S,1} = \hat{\theta}_{S,1}$ were defined and obtained, where S signifies the most promising intervention based on which had the largest average asymptotic pairwise Z-statistic and $T_{S,1}$ denotes the average for that particular intervention. This was used for the go/no-go decision at the end of the first stage. In the event that $T_{S,1} > c_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, then the study proceeded to the second stage and the number of times that this happened out of the 500 replications was collected. Then, the second stage of the trial, containing $2(n_1 + n_2)$ subjects, involved obtaining 500 asymptotic pairwise Z-statistics that compared the highest performing treatment from the first stage to the control according to the intermediate endpoint data generated for subjects in the second cohort. Similarly to stage one, the average was acquired in order to get the weighted sum $T_{S,2}$ defined in Section 4.2. The frequency that $T_{S,1} > c_1$ and $T_{S,2} > c_2$ occurred was recorded. In this numerical study, we considered $c_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$, $d_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and $d_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$. Keeping track of these events was essential for determining both the estimated type I error rate via condition (4.3) and the average total expected sample size of the study under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ using condition (4.5).

Condition (4.4) was also replicated 500 times for each scenario that we examined. Each of the steps mentioned above for producing and using the intermediate endpoint data at stage one and stage two for condition (4.3) were executed in the same way here, but with the active treatment and control data following the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ distributions and setting $S = 1$. Definitive endpoint data was also generated for the second stage under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ in order to calculate condition (4.3), where 500 asymptotic pairwise Z-statistics were obtained to compare every treatment $i = 1, \ldots, k$ to the control according to the definitive endpoint data from subjects in cohort one. The averages for each arm were then acquired. Lastly, for stage three, 500 asymptotic pairwise Z-statistics were obtained that compared the most promising active intervention to the control based on the definitive endpoint data from the second cohort of subjects, and then they were averaged. The number of instances of $T_{S,1} > c_1$, of $T_{S,1} > c_1$ and $T_{S,2} > c_2$, and of $W_{S,1} \leq d_1$ and $W_{S,2} \leq d_2$ were recorded for estimating both the power of the proposed procedure via condition (4.4) and the average total expected sample size under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ via condition (4.5).

For our simulation findings reported in Table 4.1, we focused on the scenarios that yielded an estimated type I error rate ($\hat{\alpha}$) in the interval $(0.045, 0.05)$ and an estimated power $(1-\hat{\beta})$ in the interval $(0.800, 0.820)$, where $\hat{N}_{\text{avg}}$ is the total expected sample size of the study averaged under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$. For this numerical study, there were several scenarios that had even smaller estimated type I error rates ($\hat{\alpha} < 0.045$) and even larger estimated powers ($1-\hat{\beta} > 0.82$), many more than were found for the two prior data scenarios in Chapters 2-3 (this is likely due to our choices of δ_1 and δ_2), but we directed our attention to the typical targets of $\alpha = 5\%$ and $1-\beta = 80\%$. As mentioned previously, achieving an exact power of 80% would have required using a much finer parameter grid. Altogether, Table 4.1 shows that the proposed design has the ability to preserve the pre-specified type I error rate while still reaching a desired amount of power across the scenarios that we investigated when employing a binary intermediate endpoint and a time-to-event definitive endpoint without censoring events.

4.5 Conclusion

In Chapter 4, we considered when a binary intermediate endpoint and a time-to-event definitive endpoint are used in the absence of censoring. Asymptotically, we were again able to demonstrate that the proposed procedure is valid through direct proofs and simulations. For this data scenario, a copula was employed in order to preserve the marginal distributions of X_{ij} and Y_{ij} while separately defining a correlation structure between them. Additionally, for the time-to-event definitive endpoint (Y_{ij}), an exponential distribution was used, since this is the simplest model typically used in practice for sample size calculations with time-to-event endpoints. However, Y_{ij} could be assumed to follow another relevant distribution, like the Weibull distribution. This would still be permissible in our proposed procedure, but the derivation of the limiting distribution for the endpoints would become more complicated. This work also built the foundation for assembling our data model that accounts for censored observations, which is presented in the next chapter.

Table 4.1: Simulation examples for the proposed design procedure without interim sample size adjustment when using a binary intermediate endpoint and a time-to-event definitive endpoint without censoring events, where $\hat{N}_{\text{avg}}$ is the average estimated sample size required for the trial under the null and alternative hypotheses.

k	δ_1	δ_2	n_1	n_2	c_1	c_2	d_1	d_2	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{N}_{\text{avg}}$
4	0.1	0.5	20	160	0.8	1.6	0.4	1.8	0.050	0.808	268
4	0.1	1	80	180	0.2	1.4	0.8	1.4	0.046	0.820	641
4	0.1	1.5	20	40	0.6	1.4	1	1.2	0.050	0.805	145
4	0.15	0.5	100	60	0.8	2	0.6	2	0.050	0.802	568
4	0.15	1	80	60	0.2	2	0.8	1.2	0.048	0.806	485
4	0.15	1.5	80	80	1	1.6	0.8	1.6	0.048	0.808	476
4	0.2	0.5	20	60	0.8	1.6	0.2	2	0.050	0.804	161
4	0.2	1	40	80	1	1.8	0.2	1.2	0.048	0.810	277
4	0.2	1.5	60	20	0.4	2	0.2	1.4	0.050	0.820	326
6	0.1	0.5	20	20	0.4	1.6	0.8	1.4	0.050	0.804	165
6	0.1	1	20	100	0.2	1.2	0.8	2	0.046	0.814	270
6	0.1	1.5	100	180	0.2	1.6	1	1.6	0.050	0.810	942
6	0.15	0.5	80	40	0.8	1.8	1	1.8	0.046	0.806	606
6	0.15	1	60	60	0.8	1.2	0.8	2	0.046	0.818	481
6	0.15	1.5	120	20	0.2	2	0.2	1.4	0.046	0.814	868
6	0.2	0.5	20	40	0.8	1.6	0.4	1.8	0.050	0.804	182
6	0.2	1	40	60	1	1.4	0.8	1.6	0.046	0.800	338
6	0.2	1.5	40	40	0.6	1.8	0.6	2	0.050	0.817	326

Chapter 5: Using a Binary Intermediate Endpoint and a Time-to-Event Definitive Endpoint with Censoring

In practice, it is very common for censoring to occur over the course of a clinical trial. Therefore, this chapter mimics the previous chapter but considers a time-to-event definitive endpoint that includes censored observations. First, the data model will be derived through utilizing maximum likelihood estimation and the multivariate delta method [118, 126]. Then, the proposed procedure will be described, followed by our statistical rationale and simulation examples that concur with our statistical claims.

5.1 Model

Suppose a clinical trial intends to evaluate $i = 1, \ldots, k$ experimental interventions compared to a control and/or standard of care (denoted by $i = 0$) using our design proposed in Figure 1.5, where every treatment group has an equal sample size. The following holds for fixed subject, treatment and cohort. Like what was done in Chapter 4, let $X_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_i)$ and $Y_{ij} \overset{iid}{\sim} \text{Exponential}(\xi_i)$ for subjects $j = 1, \ldots, n_i$ with their correlation structure represented by the bivariate normal copula

$$\begin{pmatrix} U_{ij} \\ V_{ij} \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

where

$$X_{ij} = \mathrm{I}(U_{ij} > -\Phi^{-1}(p_i))$$

and

$$Y_{ij} = -\xi_i \log(1 - \Phi(V_{ij})).$$

In the case of censoring, $Y_{ij}^* = \min\{Y_{ij}, C\}$ is observed, where $C > 0$ is a constant censoring time. The distribution of Y_{ij}^* can be expressed as

$$Y_{ij}^* = \min\{Y_{ij}, C\} \sim \begin{cases} \frac{1}{\xi_i} \exp\left\{-\frac{Y_{ij}}{\xi_i}\right\} & \text{if } Y_{ij} < C \\ \exp\left\{-\frac{C}{\xi_i}\right\} & \text{if } Y_{ij} > C \end{cases}$$

since

$$\int_C^{\infty} \frac{1}{\xi_i} \exp\left\{-\frac{y}{\xi_i}\right\} dy = \exp\left\{-\frac{C}{\xi_i}\right\}.$$

Therefore, the marginal likelihood of ξ_i based on Y_{ij}^* is

$$\begin{aligned} L(\xi_i) &= \prod_{j=1}^{n_i} \left(\frac{1}{\xi_i} \exp\left\{-\frac{Y_{ij}}{\xi_i}\right\}\right)^{\mathrm{I}(Y_{ij}<C)} \exp\left\{-\frac{C}{\xi_i}\right\}^{\mathrm{I}(Y_{ij}>C)} \\ &= \xi_i^{-\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij}<C)} \exp\left\{-\frac{\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\xi_i}\right\} \end{aligned}$$

where $\mathrm{I}(\cdot)$ is the indicator function. The log-likelihood is then

$$\ell(\xi_i) = -\log(\xi_i) \sum_{j=1}^{n_i} \mathrm{I}(Y_{ij} < C) - \frac{\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\xi_i}.$$

Setting the score function

$$\frac{d\ell}{d\xi_i} = -\frac{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij} < C)}{\xi_i} + \frac{\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\xi_i^2}$$

equal to zero yields

$$\hat{\xi}_i = \frac{\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij} < C)}.$$

In fact, $\hat{\xi}_i$ is the maximum likelihood estimator for ξ_i because

$$\left.\frac{d^2\ell}{d\xi_i^2}\right|_{\xi_i=\hat{\xi}_i} = \frac{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij} < C)}{\hat{\xi}_i^2} - \frac{2\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\hat{\xi}_i^3}$$

$$
\begin{aligned}
&= \frac{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij} < C)}{\left(\frac{\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij}<C)+C\mathrm{I}(Y_{ij}>C)\right]}{\sum_{j=1}^{n_i}\mathrm{I}(Y_{ij}<C)}\right)^2} - \frac{2\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\left(\frac{\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij}<C)+C\mathrm{I}(Y_{ij}>C)\right]}{\sum_{j=1}^{n_i}\mathrm{I}(Y_{ij}<C)}\right)^3} \\
&= \frac{\left(\sum_{j=1}^{n_i}\mathrm{I}(Y_{ij} < C)\right)^3}{\left(\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]\right)^2} - \frac{2\left(\sum_{j=1}^{n_i}\mathrm{I}(Y_{ij} < C)\right)^3}{\left(\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]\right)^2} \\
&= -\frac{\left(\sum_{j=1}^{n_i}\mathrm{I}(Y_{ij} < C)\right)^3}{\left(\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]\right)^2} \\
&< 0.
\end{aligned}
$$

The Fisher information for ξ_i is

$$
\begin{aligned}
\mathcal{I}(\xi_i) &= -\mathrm{E}\left(\frac{d^2\ell}{d\xi_i^2}\middle|\xi_i\right) \\
&= -\mathrm{E}\left(\frac{\sum_{j=1}^{n_i}\mathrm{I}(Y_{ij} < C)}{\xi_i^2} - \frac{2\sum_{j=1}^{n_i}\left[Y_{ij}\mathrm{I}(Y_{ij} < C) + C\mathrm{I}(Y_{ij} > C)\right]}{\xi_i^3}\middle|\xi_i\right) \\
&= -\frac{n_i}{\xi_i^2}\mathrm{E}\{\mathrm{I}(Y_{ij} < C)\} + \frac{2n_i}{\xi_i^3}\mathrm{E}\{Y_{ij}\mathrm{I}(Y_{ij} < C)\} + \frac{2n_iC}{\xi_i^3}\mathrm{E}\{\mathrm{I}(Y_{ij} > C)\} \\
&= \frac{-n_i}{\xi_i^2}\int_0^C \frac{1}{\xi_i}\exp\left\{-\frac{y}{\xi_i}\right\}dy + \frac{2n_i}{\xi_i^3}\int_0^C y\frac{1}{\xi_i}\exp\left\{-\frac{y}{\xi_i}\right\}dy + \frac{2n_iC}{\xi_i^3}\int_C^\infty \frac{1}{\xi_i}\exp\left\{-\frac{y}{\xi_i}\right\}dy \\
&= \frac{-n_i}{\xi_i^2}\left(1 - \exp\left\{\frac{-C}{\xi_i}\right\}\right) + \frac{2n_i}{\xi_i^3}\left(\xi_i - (\xi_i + C)\exp\left\{-\frac{C}{\xi_i}\right\}\right) + \frac{2n_iC}{\xi_i^3}\exp\left\{-\frac{C}{\xi_i}\right\} \\
&= -\frac{n_i}{\xi_i^2} + \frac{n_i\exp\left\{-\frac{C}{\xi_i}\right\}}{\xi_i^2} + \frac{2n_i}{\xi_i^2} - \frac{2n_i\exp\left\{-\frac{C}{\xi_i}\right\}}{\xi_i^2} \\
&= \frac{n_i\left(1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)}{\xi_i^2}.
\end{aligned}
$$

Since the distribution of Y_{ij}^* satisfies sufficient regularity conditions [127],

$$
\sqrt{n_i}(\hat{\xi}_i - \xi_i) \xrightarrow{d} \mathrm{N}\left(0, \frac{\xi_i^2}{1 - \exp\left\{-\frac{C}{\xi_i}\right\}}\right).
$$

Furthermore, by the central limit theorem,

$$\sqrt{n_i}\begin{pmatrix} \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} - p_i \\ \frac{\sum_{j=1}^{n_i} Y_{ij}\mathrm{I}(Y_{ij}<C)}{n_i} - \vartheta_i \\ \frac{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij}>C)}{n_i} - \exp\left\{-\frac{C}{\xi_i}\right\} \end{pmatrix} \xrightarrow{d} \mathrm{N}_3\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Lambda_i^*\right]$$

where $\vartheta_i = \mathrm{E}(Y_{i1}\mathrm{I}(Y_{i1} < C)) = \xi_i - (\xi_i + C)\exp\left\{-\frac{C}{\xi_i}\right\}$ and

$$\Lambda_i^* = \begin{pmatrix} p_i(1-p_i) & \mathrm{Cov}(X_{i1}, Y_{i1}\mathrm{I}(Y_{i1} < C)) & \mathrm{Cov}(X_{i1}, \mathrm{I}(Y_{i1} > C)) \\ \mathrm{Cov}(X_{i1}, Y_{i1}\mathrm{I}(Y_{i1} < C)) & \mathrm{Var}(Y_{i1}\mathrm{I}(Y_{i1} < C)) & \mathrm{Cov}(Y_{i1}\mathrm{I}(Y_{i1} < C), \mathrm{I}(Y_{i1} > C)) \\ \mathrm{Cov}(X_{i1}, \mathrm{I}(Y_{i1} > C)) & \mathrm{Cov}(Y_{i1}\mathrm{I}(Y_{i1} < C), \mathrm{I}(Y_{i1} > C)) & \exp\left\{-\frac{C}{\xi_i}\right\}\left(1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right) \end{pmatrix}.$$

Rewriting the maximum likelihood estimator as

$$\hat{\xi}_i = \frac{\frac{\sum_{j=1}^{n_i} Y_{ij}\mathrm{I}(Y_{ij}<C)}{n_i} + C\frac{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij}>C)}{n_i}}{1 - \frac{\sum_{j=1}^{n_i} \mathrm{I}(Y_{ij}>C)}{n_i}}$$

and applying the multivariate delta method with $g_1(q_1, q_2, q_3) = q_1$ and $g_2(q_1, q_2, q_3) = \frac{q_2 + Cq_3}{1 - q_3}$ for $q_1 = p_i$, $q_2 = \vartheta_i$ and $q_3 = \exp\left\{-\frac{C}{\xi_i}\right\}$, we obtain

$$\sqrt{n_i}\begin{pmatrix} \bar{X}_i - p_i \\ \hat{\xi}_i - \xi_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_i^*\right]$$

where

$$\Sigma_i^* = \begin{pmatrix} \frac{\partial g_1}{\partial q_1} & \frac{\partial g_1}{\partial q_2} & \frac{\partial g_1}{\partial q_3} \\ \frac{\partial g_2}{\partial q_1} & \frac{\partial g_2}{\partial q_2} & \frac{\partial g_2}{\partial q_3} \end{pmatrix} \Lambda_i^* \begin{pmatrix} \frac{\partial g_1}{\partial q_1} & \frac{\partial g_2}{\partial q_1} \\ \frac{\partial g_1}{\partial q_2} & \frac{\partial g_2}{\partial q_2} \\ \frac{\partial g_1}{\partial q_3} & \frac{\partial g_2}{\partial q_3} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} & \frac{C+\vartheta_i}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2} \end{pmatrix} \Lambda_i^* \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} \\ 0 & \frac{C+\vartheta_i}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2} \end{pmatrix}$$

$$= \begin{pmatrix} p_i(1-p_i) & \mathrm{Cov}\left(X_{i1}, \frac{Y_{i1}\mathrm{I}(Y_{i1}<C)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C+\vartheta_i)\mathrm{I}(Y_{i1}>C)}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2}\right) \\ \mathrm{Cov}\left(X_{i1}, \frac{Y_{i1}\mathrm{I}(Y_{i1}<C)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C+\vartheta_i)\mathrm{I}(Y_{i1}>C)}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2}\right) & \frac{\xi_i^2}{1-\exp\left\{\frac{-C}{\xi_i}\right\}} \end{pmatrix}$$

for all i. We know that the asymptotic correlation

$$\eta = \mathrm{Cov}\left(X_{i1}, \frac{Y_{i1}\mathrm{I}(Y_{i1} < C)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C+\vartheta_i)\mathrm{I}(Y_{i1} > C)}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2}\right) \Bigg/ \left(\xi_i \sqrt{\frac{p_i(1-p_i)}{1-\exp\left\{\frac{-C}{\xi_i}\right\}}}\right)$$

is an increasing function of ρ if we can show that

$$\zeta = \mathrm{E}\left\{X_{i1}\left(\frac{Y_{i1}\mathrm{I}(Y_{i1} < C)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C+\vartheta_i)\mathrm{I}(Y_{i1} > C)}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2}\right)\right\}$$

is an increasing function of ρ. Similar to Chapter 4, by the law of iterated expectations [123] and because $V_{ij} \sim \mathrm{N}(0,1)$ and $U_{ij}|V_{ij} \sim \mathrm{N}(\rho V_{ij}, 1-\rho^2)$ according to the copula above,

$$\mathrm{E}\left\{X_{i1}\left(\frac{Y_{i1}\mathrm{I}(Y_{i1} < C)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C+\vartheta_i)\mathrm{I}(Y_{i1} > C)}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2}\right)\right\} =$$

$$\mathrm{E}\left\{\Phi\left(\frac{\Phi^{-1}(p_i)+\rho V_{i1}}{\sqrt{1-\rho^2}}\right) \cdot \frac{-\xi_i \log(1-\Phi(V_{i1}))\mathrm{I}(-\xi_i \log(1-\Phi(V_{i1})) < C) + (C+\xi_i)\mathrm{I}(-\xi_i \log(1-\Phi(V_{i1})) > C)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}}\right\}$$

which we demonstrate to indeed be strictly increasing in the correlation ρ using simulations (see Figure B.1).

Like in the last chapter, suppose the observed treatment effects based on the intermediate and definitive endpoint data are $\theta_i = p_i - p_0$ and $\psi_i = \xi_i - \xi_0$, respectively, for all $i = 1, \ldots, k$. Then by the central limit theorem,

$$\sqrt{n_i}\begin{pmatrix} (\bar{X}_i - \bar{X}_0) - \theta_i \\ (\hat{\xi}_i - \hat{\xi}_0) - \psi_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_i^* + \Sigma_0^*\right].$$

As in previous chapters that considered different pairs of endpoint types, this data model allows us to define asymptotic pairwise Z-statistics throughout our procedure proposed in the next section for when a binary intermediate endpoint and a time-to-event definitive endpoint are employed with cases of censoring. In Section 5.3, this data model also enables us to again apply Slepian's lemma asymptotically for verification (i.e., for ensuring that the pre-specified type I error rate is preserved while achieving a desired power). Simulations follow in Section 5.4, which support our statistical arguments and indicate that the proposed procedure can be used reliably.

5.2 Procedure

We use the same general procedure described for the absence of censoring events in Chapter 4. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)^{\mathrm{T}}$ be the vectors of treatment effects comparing active intervention $i = 1, \ldots, k$ to the control ($i = 0$) based on the intermediate and definitive endpoint data, respectively, where $\theta_i = p_i - p_0$ and $\psi_i = \xi_i - \xi_0$ for all i. With $\mathbf{0}$ representing the zero vector, the hypotheses of interest are

$$H_{0,1} : \boldsymbol{\theta} \leq \mathbf{0} \tag{5.1}$$

and

$$H_{0,2} : \boldsymbol{\psi} \leq \mathbf{0}. \tag{5.2}$$

Hypotheses (5.1) and (5.2) are equivalent to those stated in Chapters 2-4. The layout of the proposed procedure when using a binary intermediate endpoint and a time-to-event definitive endpoint, without reassessing the sample size at interim, is provided below when there is censoring.

At the first stage, $(k + 1)n_1$ participants are randomly and equally allocated to each of the k active arms or the control, with n_1 being the cohort one sample size per arm. Define asymptotic pairwise Z-statistics

$$\hat{\theta}_{i,1} = \frac{\bar{X}_{i\cdot,1} - \bar{X}_{0\cdot,1}}{\sqrt{\frac{\bar{X}_{i\cdot,1}(1-\bar{X}_{i\cdot,1})+\bar{X}_{0\cdot,1}(1-\bar{X}_{0\cdot,1})}{n_1}}}$$

comparing active treatment i to the control group, with $\bar{X}_{i\cdot,1}$ and $\bar{X}_{0\cdot,1}$ being the average intermediate outcomes based on the cohort one data from participants in active arm i and the control, respectively. Thus, we know that $\hat{\theta}_{i,1} \xrightarrow{d} \mathrm{N}(0,1)$ for large n_1 by the central limit theorem. These asymptotic pairwise Z-statistics that are calculated at the first interim analysis are used to select the highest performing treatment for continued study, if there is one, based on the accrued intermediate endpoint data. Define $S = \underset{1\le i\le k}{\operatorname{argmax}}\, \hat{\theta}_{i,1}$ as the most promising intervention at the end of stage one and $T_{S,1} = \hat{\theta}_{S,1}$ as its observed effect. With the data accumulated during the first stage, a stop/go decision is made. The trial is stopped early and $H_{0,1}$ (5.1) is accepted if $T_{S,1} \le c_1$ for some $c_1 > 0$. Otherwise, $H_{0,1}$ (5.1) is rejected and the trial proceeds to the second stage if $T_{S,1} > c_1$, where the highest performing treatment undergoes further evaluation compared to the control group and all other arms are removed from the study. Therefore, only participants receiving the most promising therapy or the control in stage one will be followed for their definitive endpoint data. All participants from the eliminated arms will continue appropriate safety monitoring and be offered standard of care.

If the study advances to the second stage, $2n_2$ new participants are recruited and enrolled into the trial with 1:1 allocation to randomly receive either the highest performing treatment from the prior stage or the control, with n_2 being the cohort two sample size per group. After collecting the cohort one and cohort two intermediate endpoint data and the cohort one definitive endpoint data, the second interim analysis can be conducted. Define the asymptotic pairwise Z-statistic

$$\hat{\theta}_{S,2} = \frac{\bar{X}_{S\cdot,2} - \bar{X}_{0\cdot,2}}{\sqrt{\frac{\bar{X}_{S\cdot,2}(1-\bar{X}_{S\cdot,2})+\bar{X}_{0\cdot,2}(1-\bar{X}_{0\cdot,2})}{n_2}}}$$

according to the intermediate endpoint data from the second cohort, with $\bar{X}_{S\cdot,2}$ and $\bar{X}_{0\cdot,2}$ being the average intermediate outcomes among cohort two participants receiving the highest performing therapy from the first stage and the control, respectively. Additionally, let

$$W_{S,1} = \hat{\psi}_{S,1} = \frac{\hat{\xi}_{i,1} - \hat{\xi}_{0,1}}{\sqrt{\left(\frac{\hat{\xi}_{i,1}^2}{1-\exp\left\{-\frac{C}{\hat{\xi}_{i,1}}\right\}} + \frac{\hat{\xi}_{0,1}^2}{1-\exp\left\{-\frac{C}{\hat{\xi}_{0,1}}\right\}}\right) \Big/ n_1}}$$

be the asymptotic pairwise Z-statistic according to the definitive endpoint data from the second stage, with $\hat{\xi}_{i,1}$ and $\hat{\xi}_{0,1}$ being the maximum likelihood estimator (as defined in Section 5.1) among cohort one participants receiving the most promising treatment from stage one and the control, respectively. Similarly to the previous stage, $\hat{\theta}_{S,2} \xrightarrow{d} \mathrm{N}(0,1)$ and $\hat{\psi}_{S,1} \xrightarrow{d} \mathrm{N}(0,1)$ by the central limit theorem. Finally, let

$$T_{S,2} = \sqrt{\frac{n_1}{n_1 + n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1 + n_2}}\hat{\theta}_{S,2}$$

be the weighted sum of the cohort one and cohort two intermediate treatment effects comparing the highest performing intervention to the control. These $T_{S,2}$ and $W_{S,1}$ values will be used to make another stop/go decision. The trial is stopped for futility and $H_{0,1}$ (5.1) is accepted if $T_{S,2} \leq c_2$ for some $c_2 > 0$. However, the trial is stopped for superiority and $H_{0,1}$ (5.1) and $H_{0,2}$ (5.2) are rejected when $T_{S,2} > c_2$ and $W_{S,1} > d_1$ for some $d_1 > 0$. Otherwise, the trial proceeds to the last stage if the intermediate endpoint data from the first and second cohorts is compelling but the definitive endpoint data from the first cohort is not sufficiently convincing (i.e., $T_{S,2} > c_2$ but $W_{S,1} \leq d_1$).

For the third stage, no additional participants enter the trial. After the cohort two participants complete their definitive endpoints, the final analysis can be done. Define the asymptotic pairwise Z-statistic

$$\hat{\psi}_{S,2} = \frac{\hat{\xi}_{i,2} - \hat{\xi}_{0,2}}{\sqrt{\left(\frac{\hat{\xi}_{i,2}^2}{1-\exp\left\{-\frac{C}{\hat{\xi}_{i,2}}\right\}} + \frac{\hat{\xi}_{0,2}^2}{1-\exp\left\{-\frac{C}{\hat{\xi}_{0,2}}\right\}}\right) \Big/ n_2}}$$

comparing the most promising treatment to the control based on the cohort two definitive endpoint data, with $\hat{\xi}_{i,2}$ and $\hat{\xi}_{0,2}$ being the maximum likelihood estimator (as defined in Section 5.1) among participants in the second cohort receiving the highest performing treatment and the control, respectively. Thus, $\hat{\psi}_{S,2} \xrightarrow{d} \mathrm{N}(0, 1)$. Similarly to stage two, let

$$W_{S,2} = \sqrt{\frac{n_1}{n_1 + n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1 + n_2}}\hat{\psi}_{S,2}$$

be the weighted sum of the cohort one and cohort two definitive treatment effects. The treatment under study is considered superior to the control when $H_{0,2}$ (5.2) is rejected (i.e., when $W_{S,2} > d_2$ for some $d_2 > 0$) and is not considered superior to the control when $H_{0,2}$ (5.2) is accepted (i.e., when $W_{S,2} \leq d_2$).

As in Chapters 2-4, $H_{0,1}$ (5.1) is based only on the intermediate endpoint data and can be seen as a gatekeeper for testing $H_{0,2}$ (5.2), which relies only on the definitive endpoint data. Like each of the previous data scenarios, simulations are necessary to ascertain the required design parameters (c_1, c_2, d_1, d_2, n_1, and n_2), as they cannot be obtained analytically using closed form equations. However, identifying the parameters via simulations will not compromise the nominal type I error rate of the trial or prevent a particular power from being reached. This will be proven explicitly in the next section and illustrated by simulations in Section 5.4.

5.3 Statistical Justification

The procedure proposed above for utilizing a binary intermediate endpoint and a time-to-event definitive endpoint with censored observations must be validated. First, the procedure must preserve the pre-specified type I error rate. This can be proven in the same manner as in Chapter 4 since

$$\sqrt{n_i}\begin{pmatrix} \bar{X}_i - p_i \\ \hat{\xi}_i - \xi_i \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_i^*\right]$$

for all i, as shown in Section 5.1. Again, the maximum probability of incorrectly rejecting $H_{0,1}$ (5.1) and $H_{0,2}$ (5.2) occurs when $\theta_1 = \cdots = \theta_k = 0$ and $\psi_1 = \cdots = \psi_k = 0$ [119]. Denoting these vectors by $\boldsymbol{\theta}_*$ and $\boldsymbol{\psi}_*$, respectively, the size of the procedure is

$$\begin{aligned}
\text{Size} &= P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\eta)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\eta)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big) \\
&\leq kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho\leq 1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big) \\
&= kP_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\} \cap \big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\big)\Big).
\end{aligned}$$

The first inequality above remains an asymptotic inequality ($\leq$), since Slepian's lemma is applied asymptotically. We can employ Slepian's lemma because η is a strictly increasing function of ρ in our model outlined in Section 5.1. Therefore, a constraint for the type I error rate of the procedure can been obtained:

$$P_{(\boldsymbol{\theta}_*,\boldsymbol{\psi}_*,\rho=1)}\Big(\{T_{S,1} > c_1\} \cap \{T_{S,2} > c_2\} \cap \big(\{W_{S,1} > d_1\} \cup \{W_{S,2} > d_2\}\big)\Big) = \alpha. \tag{5.3}$$

Condition (5.3) is identical to condition (4.3) derived for the non-censoring case, where the type I error condition when using a binary intermediate endpoint and a time-to-event definitive endpoint (when censoring is present) also relies on both the intermediate and definitive endpoint data.

In order to calculate the covariance term in η (call it υ) under the null ($\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1$), there are two cases to consider. For both, $U_{ij} = V_{ij}$ and let $W_{ij} \sim \text{Uniform}(0,1)$. The first case is when $p_i < \exp\{-\frac{C}{\xi_i}\}$, where

$$\begin{aligned}
\mathrm{E}\{X_{ij}Y_{ij}\mathrm{I}(Y_{ij} < C)\} &= \mathrm{E}\{-\mathrm{I}(\Phi(U_{ij}) > 1 - p_i)\xi_i \log(1 - \Phi(U_{ij}))\mathrm{I}(-\xi_i \log(1 - \Phi(U_{ij})) < C)\} \\
&= \mathrm{E}\left\{-\mathrm{I}(\Phi(U_{ij}) > 1 - p_i)\xi_i \log(1 - \Phi(U_{ij}))\mathrm{I}\left(\Phi(U_{ij}) < 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\
&= \mathrm{E}\left\{-\mathrm{I}(W_{ij} > 1 - p_i)\xi_i \log(1 - W_{ij})\mathrm{I}\left(W_{ij} < 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\
&= 0
\end{aligned}$$

and thus

$$\begin{aligned}\operatorname{Cov}\{X_{ij}, Y_{ij}\mathrm{I}(Y_{ij} < C)\} &= \mathrm{E}\{X_{ij}Y_{ij}\mathrm{I}(Y_{ij} < C)\} - \mathrm{E}(X_{ij})\mathrm{E}\{Y_{ij}\mathrm{I}(Y_{ij} < C)\} \\ &= -p_i\left(\xi_i - (C+\xi_i)\exp\left\{-\frac{C}{\xi_i}\right\}\right).\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathrm{E}\{X_{ij}\mathrm{I}(Y_{ij} > C)\} &= \mathrm{E}\{\mathrm{I}(\Phi(U_{ij}) > 1 - p_i)\mathrm{I}(-\xi_i \log(1-\Phi(U_{ij})) > C)\} \\ &= \mathrm{E}\left\{\mathrm{I}(\Phi(U_{ij}) > 1 - p_i)\mathrm{I}\left(\Phi(U_{ij}) > 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\ &= \mathrm{E}\left\{\mathrm{I}(W_{ij} > 1 - p_i)\mathrm{I}\left(W_{ij} > 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\ &= \mathrm{E}\{\mathrm{I}(W_{ij} > 1 - p_i)\} \\ &= P(W_{ij} > 1 - p_i) \\ &= p_i\end{aligned}$$

and so

$$\begin{aligned}\operatorname{Cov}\{X_{ij}, \mathrm{I}(Y_{ij} > C)\} &= \mathrm{E}\{X_{ij}\mathrm{I}(Y_{ij} > C)\} - \mathrm{E}(X_{ij})\mathrm{E}\{\mathrm{I}(Y_{ij} > C)\} \\ &= p_i\left(1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right).\end{aligned}$$

This yields

$$\begin{aligned}\upsilon &= \frac{-p_i\left(\xi_i - C\exp\left\{-\frac{C}{\xi_i}\right\} - \xi_i\exp\left\{-\frac{C}{\xi_i}\right\}\right)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C+\vartheta_i)p_i\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)}{\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)^2} \\ &= \frac{p_i C}{1-\exp\left\{-\frac{C}{\xi_i}\right\}}.\end{aligned}$$

The other case is when $p_i \geq \exp\{-\frac{C}{\xi_i}\}$, where

$$\begin{aligned}
\mathrm{E}\{X_{ij}Y_{ij}\mathrm{I}(Y_{ij} < C)\} &= \mathrm{E}\left\{-\mathrm{I}(W_{ij} > 1 - p_i)\xi_i \log(1 - W_{ij})\mathrm{I}\left(W_{ij} < 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\
&= -\xi_i \int_{1-p_i}^{1-\exp\left\{-\frac{C}{\xi_i}\right\}} \log(1 - w)dw \\
&= -\xi_i \exp\left\{-\frac{C}{\xi_i}\right\} - C\exp\left\{-\frac{C}{\xi_i}\right\} + \xi_i p_i - \xi_i p_i \log(p_i)
\end{aligned}$$

and thus

$$\begin{aligned}
\mathrm{Cov}\{X_{ij}, Y_{ij}\mathrm{I}(Y_{ij} < C)\} &= \mathrm{E}\{X_{ij}Y_{ij}\mathrm{I}(Y_{ij} < C)\} - \mathrm{E}(X_{ij})\mathrm{E}\{Y_{ij}\mathrm{I}(Y_{ij} < C)\} \\
&= -\xi_i(1 - p_i)\exp\left\{-\frac{C}{\xi_i}\right\} - C(1 - p_i)\exp\left\{-\frac{C}{\xi_i}\right\} - \xi_i p_i \log(p_i).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathrm{E}\{X_{ij}\mathrm{I}(Y_{ij} > C)\} &= \mathrm{E}\left\{\mathrm{I}(W_{ij} > 1 - p_i)\mathrm{I}\left(W_{ij} > 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\
&= \mathrm{E}\left\{\mathrm{I}\left(W_{ij} > 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)\right\} \\
&= P\left(W_{ij} > 1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right) \\
&= \exp\left\{-\frac{C}{\xi_i}\right\}
\end{aligned}$$

and thus

$$\mathrm{Cov}\{X_{ij}, \mathrm{I}(Y_{ij} > C)\} = (1 - p_i)\exp\left\{-\frac{C}{\xi_i}\right\}.$$

This results in

$$\upsilon = -\frac{\xi_i(1 - p_i)\exp\left\{-\frac{C}{\xi_i}\right\} + \xi_i p_i \log(p_i) + C(1 - p_i)\exp\left\{-\frac{C}{\xi_i}\right\}}{1 - \exp\left\{-\frac{C}{\xi_i}\right\}} + \frac{(C + \vartheta_i)(1 - p_i)\exp\left\{-\frac{C}{\xi_i}\right\}}{\left(1 - \exp\left\{-\frac{C}{\xi_i}\right\}\right)^2}$$

$$= \frac{-\xi_i p_i \log(p_i)}{1 - \exp\left\{-\frac{C}{\xi_i}\right\}}.$$

Putting this all together, we get

$$\upsilon = \begin{cases} \frac{p_i C}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} & \text{if } p_i < \exp\left\{-\frac{C}{\xi_i}\right\} \\ \frac{-\xi_i p_i \log(p_i)}{1-\exp\left\{-\frac{C}{\xi_i}\right\}} & \text{if } p_i \geq \exp\left\{-\frac{C}{\xi_i}\right\}. \end{cases}$$

The procedure laid out in Section 5.2 also needs to be able to achieve a certain targeted power for it to be valid. To prove this, let $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$ denote the least favorable configuration explained in Section 2.2.2. These selections yield the lowest power compared to any other choices [119, 121]. The power of the proposed procedure can be expressed as

$$\begin{aligned}
\text{Power} &= P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\eta)}(\text{reject } H_{0,1} \text{ and } H_{0,2}) \\
&= \sum_{i=1}^{k} P_{(\boldsymbol{\theta}^*,\boldsymbol{\psi}^*,\eta)}\Big(\{S=i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\} \cap (\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\})\Big) \\
&\gtrsim \sum_{i=1}^{k} P_{\boldsymbol{\theta}^*}\Big(\{S=i\} \cap \{T_{i,1} > c_1\} \cap \{T_{i,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{i,1} > d_1\} \cup \{W_{i,2} > d_2\}\Big) \\
&\geq P_{\boldsymbol{\theta}^*}\Big(\{S=1\} \cap \{T_{1,1} > c_1\} \cap \{T_{1,2} > c_2\}\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{1,1} > d_1\} \cup \{W_{1,2} > d_2\}\Big).
\end{aligned}$$

Note that the first inequality is again an asymptotic inequality ($\gtrsim$), because Slepian's lemma is applied asymptotically. This gives us the following constraint for the power:

$$P_{\boldsymbol{\theta}^*}\Big(T_{S,1} > c_1, T_{S,2} > c_2\Big) \cdot P_{\boldsymbol{\psi}^*}\Big(\{W_{S,1} > d_1\} \cup \{W_{S,2} > d_2\}\Big) = 1 - \beta. \tag{5.4}$$

This condition is the same as those found for all of the other pairs of endpoint types considered in Chapters 2-4. Lastly, we also minimize the total expected sample size of the study

$$\mathrm{E}(N) = (k+1)n_1 + 2n_2 P\left(T_{S,1} > c_1\right)$$

by selecting the design parameters c_1, c_2, d_1, d_2, n_1, and n_2 such that

$$\frac{1}{2}\mathrm{E}\left[N|(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*)\right] + \frac{1}{2}\mathrm{E}\left[N|(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*)\right] = (k+1)n_1 + n_2\left\{P_{\boldsymbol{\theta}_*}(T_{S,1} > c_1) + P_{\boldsymbol{\theta}^*}(T_{S,1} > c_1)\right\} \quad (5.5)$$

is minimized [121].

The three constraints above, namely conditions (5.3), (5.4) and (5.5), establish the validity of the proposed procedure, as they provide an upper bound for the size of the procedure, a lower bound for the power so that the target can be reached, and a sample size that is optimal. Therefore, this procedure is expected to function well in practice. Like in the previous three chapters, Slepian's lemma was employed to avoid having to manage the unknown correlation between the intermediate and definitive endpoints or specify the complicated joint distribution of $(S, T_{S,1}, T_{S,2}, W_{S,1}, W_{S,2})^{\mathrm{T}}$, which would be very computationally complex.

5.4 Simulations

Using the three conditions obtained in Section 5.3, Monte Carlo simulations were conducted to exemplify the utility and validity of the proposed procedure when a binary intermediate endpoint and a time-to-event definitive endpoint are employed when censoring is present. In this numerical study, $k = \{4, 6\}$ experimental groups of equal size and a constant censoring time of $C = 1$ were considered, with first and second cohort sample sizes of $n_1 = \{20, 40, \ldots, 180, 200\}$ and $n_2 = \{20, 40, \ldots, 180, 200\}$ per arm, respectively. To start, conditions (5.3), (5.4) and (5.5) were used to simulate intermediate and definitive endpoint data for $(k+1)(n_1+n_2)$ individuals, as they do not rely on the unknown correlation. For fixed cohort l,

$$\begin{pmatrix} \bar{X}_{i\cdot,l} \\ \hat{\xi}_{i,l} \end{pmatrix} \approx \mathrm{N}_2\left[\begin{pmatrix} p_i \\ \xi_i \end{pmatrix}, \begin{pmatrix} \frac{p_i(1-p_i)}{n_l} & \frac{\upsilon}{n_l} \\ \frac{\upsilon}{n_l} & \frac{\xi_i^2}{n_l\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)} \end{pmatrix}\right]$$

for all $i = 0, \ldots, k$ under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$, with υ defined as in Section 5.3, $p_0 = p_i$ and $\xi_0 = \xi_i$. Under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$,

$$\begin{pmatrix} \bar{X}_{i\cdot,l} \\ \hat{\xi}_{i,l} \end{pmatrix} \approx \mathrm{N}_2 \left[\begin{pmatrix} p_i \\ \xi_i \end{pmatrix}, \begin{pmatrix} \frac{p_i(1-p_i)}{n_l} & 0 \\ 0 & \frac{\xi_i^2}{n_l\left(1-\exp\left\{-\frac{C}{\xi_i}\right\}\right)} \end{pmatrix} \right]$$

for all $i = 0, \ldots, k$. For consistency with Chapter 4, we again considered $p_0 = 0.5$, $p_1 = 0.75$, $p_2 = \cdots = p_k = \{0.6, 0.65, 0.7\}$, $\xi_0 = 1$, $\xi_1 = 3$, and $\xi_2 = \cdots = \xi_k = \{1.5, 2, 2.5\}$. Therefore, $\theta_1 = p_1 - p_0 = 0.25$, $\theta_2 = \cdots = \theta_k = \{0.1, 0.15, 0.2\} = \delta_1$, $\psi_1 = \xi_1 - \xi_0 = 2$, and $\psi_2 = \cdots = \psi_k = \{0.5, 1, 1.5\} = \delta_2$. Because of condition (5.3), the intermediate and definitive endpoint data has to be generated jointly under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$, but this is not the case for condition (5.4), which is calculated under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$. Recall that condition (5.5) is used to obtain the average total expected sample size of the trial under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$, which is based solely on the intermediate endpoint data.

The steps taken for this numerical study are explained next. For condition (5.3), intermediate endpoint data for each active treatment arm and the control group was produced 500 times under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ distributions for each scenario. For stage one, which contains $(k+1)n_1$ participants, 500 asymptotic pairwise Z-statistics $(\hat{\theta}_{i,1})$ were obtained, comparing each investigational therapy to the control based on the intermediate endpoint data from the cohort one participants, and then they were averaged for each arm. After, $S = \underset{1 \le i \le k}{\operatorname{argmax}}\, \hat{\theta}_{i,1}$ and $T_{S,1} = \hat{\theta}_{S,1}$ were obtained, with S being the highest performing experimental treatment according to which had the smallest average Z-statistic and $T_{S,1}$ representing that average. The stop/go decision at the first interim analysis was made for each of the 500 replications. If $T_{S,1} > c_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$, then the trial advanced to stage two, and the number of instances of this was recorded. For stage two, which has $2(n_1 + n_2)$ subjects, 500 asymptotic pairwise Z-statistics were computed to compare the most promising intervention from stage one to the control based on the intermediate endpoint data from the cohort two participants. Then, the average was calculated to get $T_{S,2}$, as defined in Section 5.2. The number of times that $T_{S,1} > c_1$ and $T_{S,2} > c_2$ occurred was collected. For these simulations,

$c_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$, $d_1 = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $d_2 = \{1.2, 1.4, 1.6, 1.8, 2.0\}$ were considered. Tracking these events was needed in order to get both the estimated type I error rate via condition (5.3) and the average total expected sample size of the trial under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ using condition (5.5).

For every scenario, condition (5.4) also used 500 replications. All of the steps described above for generating and using the intermediate endpoint data at stage one and stage two for condition (5.3) were executed very similarly, but with the active intervention and control data following the alternative distributions and setting $S = 1$. Additionally, definitive endpoint data was produced for stage two under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ for computing condition (5.3), where 500 asymptotic pairwise Z-statistics were calculated for comparing each active therapy to the control based on the definitive endpoint data from the cohort one subjects. The averages for each group were acquired. Lastly, for the third stage, 500 asymptotic pairwise Z-statistics were obtained for comparing the highest performing treatment to the control based on the cohort two definitive endpoint data, and then they were averaged. The frequency of $T_{S,1} > c_1$, of $T_{S,1} > c_1$ and $T_{S,2} > c_2$, and of $W_{S,1} \leq d_1$ and $W_{S,2} \leq d_2$ were tracked for estimating the power of the proposed procedure via condition (5.4) and the average total expected sample size of the trial under the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$ using condition (4.5).

For reporting, we honed in on the scenarios that resulted in an estimated type I error ($\hat{\alpha}$) in the interval $(0.045, 0.05)$ and an estimated power $(1 - \hat{\beta})$ in the interval $(0.800, 0.820)$, with $\hat{N}_{\text{avg}}$ denoting the average total expected sample size under the null $(\boldsymbol{\theta}_*, \boldsymbol{\psi}_*, \rho = 1)$ and the alternative $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*, \rho = 0)$. In our study, there were numerous scenarios with lower estimated type I error rates and higher estimated powers than the 5% and 80% targets, but we focused on those falling within the ranges above to be conservative. In order to reach an exact power of 80%, we would need to use smaller parameter gradations, requiring much more computational time. Overall, Table 5.1 shows that our proposed procedure in Section 5.2 is able to maintain the nominal type I error rate while still achieving a targeted power among the scenarios that we investigated when employing a binary intermediate endpoint and a time-to-event definitive endpoint with censored observations.

Table 5.1: Simulation examples for the proposed design procedure without interim sample size adjustment when using a binary intermediate endpoint and a time-to-event definitive endpoint with censoring events, where $\hat{N}_{\text{avg}}$ is the average estimated sample size required for the trial under the null and alternative hypotheses.

k	δ_1	δ_2	n_1	n_2	c_1	c_2	d_1	d_2	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{N}_{\text{avg}}$
4	0.1	0.5	60	80	0.6	1.4	0.4	1.4	0.048	0.812	438
4	0.1	1	40	40	0.8	1.8	0.6	2	0.050	0.814	261
4	0.1	1.5	80	60	0.2	1.2	0.2	1.4	0.050	0.807	492
4	0.15	0.5	40	80	0.6	1.8	1	1.6	0.050	0.803	344
4	0.15	1	100	60	0.4	1.2	0.6	1.8	0.048	0.820	603
4	0.15	1.5	60	20	0.8	2	0.8	1.2	0.046	0.811	325
4	0.2	0.5	20	100	1	1.8	0.4	2	0.048	0.813	268
4	0.2	1	40	80	0.8	1.4	1	1.8	0.050	0.801	338
4	0.2	1.5	60	60	0.4	1.4	0.8	1.2	0.048	0.802	394
6	0.1	0.5	40	60	0.4	1.6	0.2	1.6	0.050	0.814	377
6	0.1	1	80	40	0.6	1.8	0.4	1.8	0.050	0.806	616
6	0.1	1.5	100	80	0.2	1.2	0.4	1.6	0.050	0.820	831
6	0.15	0.5	20	60	1	2	1	1.4	0.046	0.816	242
6	0.15	1	20	20	0.6	1.6	0.2	1.8	0.050	0.812	160
6	0.15	1.5	60	40	0.8	1.4	1	1.8	0.048	0.810	469
6	0.2	0.5	40	20	0.8	1.8	0.6	1.4	0.048	0.808	292
6	0.2	1	80	120	0.4	1.2	0.2	1.4	0.046	0.808	776
6	0.2	1.5	60	60	0.6	1.6	0.8	1.2	0.046	0.812	510

5.5 Conclusion

In Chapter 5, we expanded upon the material in Chapter 4 to address the occurrence of censoring events, which are typically encountered in real-world trials. Employing our data model in Section 5.1, we were again able to prove directly, via Slepian's lemma, that the procedure proposed in Section 5.2 can asymptotically control the nominal type I error rate and reach the desired power, and this was also shown through simulation examples in Section 5.4. Incorporating censored observations necessitated defining a much more involved joint distribution than the one needed in Chapter 4. Furthermore, we also considered a constant censoring time for all subjects; however, in practice, subject censoring times can surely vary. With some computational changes, our data model can be altered to account for this as long as the censoring times are independent of the time to event. Finally, since the log-rank test is common in survival settings (e.g., in oncology), it could be of interest to utilize the log-rank test statistics based on the definitive endpoint data rather than the asymptotic pairwise Z-statistics described in Section 5.2, but this would require a restructuring of our proposed procedure.

Chapter 6: Countering an Argument against Adaptive Methods

In this chapter, we promote the appropriate evaluation and use of pre-specified clinical trial adaptations that are scientifically and statistically justifiable and based on accumulated participant information. Widespread studies have demonstrated that adaptive methods can be advantageous and are applicable in almost every therapeutic area. For instance, they can contribute to improved decision-making throughout a clinical trial without impacting its operational characteristics (i.e., without inflating the nominal type I error rate). However, there are a few individuals who have conveyed opposition toward utilizing adaptations in clinical trials. One reason for this is that test statistics from adaptive designs can be insufficient, but many strategies have been developed over the years to overcome this obstacle (e.g., the closed testing principle and hierarchical tests). In the following sections, we discuss an article that presents a dispute against the utility of clinical trial adaptations and then highlight some of its defects in order to encourage the proper assessment and employment of adaptive techniques in practice.

6.1 Argument

Tsiatis and Mehta (2003) [66] assert that although adaptive design methods seem highly appealing, they can be shown to be inefficient. Additionally, they suggest that all adaptive designs can be improved upon via group sequential tests based on the sequential likelihood ratio test (LRT) statistic and that for every adaptive test, a uniformly better group sequential test can always be identified using a generalization of the Neyman-Pearson lemma [128]. These claims are quite strong and the methods that they used to verify them are unconvincing. Before explaining in what ways their approaches were debatable, the authors' notation, examples, definitions, and theorems will be introduced for clarity.

6.1.1 Notation and Examples

In a clinical trial, suppose sequential decisions will be made at time points $j = 1, \ldots, K$, where $X_1, \ldots, X_K$ are the independent random vectors of participant data. Specifically, X_1 is the data accrued between baseline and the first time point, X_2 is the data accrued between the first and second time points, and so on. For all j, the probability density function of X_j can be expressed as $p_{0j}(x_j)$ and $p_{1j}(x_j)$ under the null and alternative hypotheses, respectively. The interim trial decisions are made according to $Y_j = (X_1, \ldots, X_j)$, which denotes all of the data collected up to time point j. The first example provided by Tsiatis and Mehta (2003) [66] outlines a group sequential procedure, described below.

Suppose a sponsor/investigator is interested in comparing a single active treatment to a placebo control and/or standard of care, where up to n pairs of participants will be enrolled and each pair will be randomly split and assigned to one of the two groups. Let $Z_j \stackrel{iid}{\sim} \mathrm{N}(\mu, 1)$, where Z_j is the difference in the normally distributed responses between the j pairs of participants and μ is the average treatment difference. Decision-making occurs at the interim analyses after $n_1 < n_2 < \cdots < n_K = n$ observations are gathered at each stage or time point, where n denotes the total sample size of the trial. The null hypothesis of interest is $H_0 : \mu = 0$, which will be tested against the alternative hypothesis $H_1 : \mu = \mu_1 > 0$. Letting $n_0 = 0$, $X_j = (Z_{n_{j-1}+1}, \ldots, Z_{n_j})$ and $Y_j = (Z_1, \ldots, Z_{n_j})$, the probability density function of each X_j under the null ($q = 0$) and the alternative ($q = 1$) can be written as

$$p_{qj}(x_j) = (2\pi)^{\frac{n_{j-1}-n_j}{2}} \exp\left\{-\sum_{l=n_{j-1}+1}^{n_j} \frac{(z_l - q\mu_1)^2}{2}\right\}.$$

In this group sequential test, H_0 will be rejected at time point j if

$$S_j = \sum_{l=1}^{n_j} Z_l > u_j.$$

Otherwise, if $S_j < l_j$, where $l_j < u_j$ for $j = 1, \ldots, K-1$ and $l_K = u_K$, then H_0 is accepted. Note that the l_j and u_j values are chosen prior to trial start such that the test has level α under H_0.

The next example that the authors present is an adaptive procedure, which used the same design framework as the group sequential example above but now allowing for an adaptation to take place at interim if $S_j \in [l_j, u_j]$, or when neither the futility boundary nor the superiority boundary is crossed at time point j. For instance, for $j = 1$, the interval $[l_1, u_1]$ can be partitioned into $K - 1$ disjoint sets, say Γ_j for $j = 2, \ldots, K$. If the test statistic $S_1 \in \Gamma_j$, then the second interim analysis will be conducted when n_j observations are obtained. Otherwise, the trial is terminated early for futility or superiority after the first interim analysis. For $j = 2, \ldots, K$, the test statistic

$$T_j = S_1 + \sqrt{\frac{n_2 - n_1}{n_j - n_1}}(S_j - S_1)$$

can be constructed. If the trial is comprised of two stages, then at the second interim analysis, H_0 is rejected if $T_j > u_2$ and H_0 is accepted if $T_j \leq u_2$. The two-stage adaptive test here will have the same level α as the two-stage version of the group sequential test detailed earlier.

6.1.2 Definitions and Theorems

The authors go on to illustrate how any group sequential test can be represented by a sequence of disjoint and Y_j-measurable rejection and acceptance regions ($\mathcal{R}_j$ and $\mathcal{A}_j$, respectively) at each time point $j = 1, \ldots, K$. For the group sequential example in the previous section, $\mathcal{R}_1 = (S_1 > u_1)$, $\mathcal{A}_1 = (S_1 < l_1)$ and for $j = 2, \ldots, K$,

$$\mathcal{R}_j = (l_1 \leq S_1 \leq u_1, \ldots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j > u_j)$$

and

$$\mathcal{A}_j = (l_1 \leq S_1 \leq u_1, \ldots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j < l_j).$$

Namely, the rejection region at stage one ($\mathcal{R}_1$) is where the first stage test statistic (S_1) is larger than the stage one upper boundary (u_1). Similarly, the stage one acceptance region ($\mathcal{A}_1$) is where the first stage test statistic (S_1) is less than the stage one lower boundary (l_1). Additionally, for

$j = 2, \ldots, K$, the j-th rejection region is the set of continuation regions until the stage at which the rejection occurs, and the j-th acceptance region is the set of continuation regions up until the acceptance stage. Note that this is due to the closed testing principle; all hypothesis tests must be carried throughout the procedure or substantial bias/error can be incurred, which could result in inaccurate efficacy conclusions.

For the adaptive design example in the previous section, $\mathcal{R}_1 = (S_1 > u_1)$, $\mathcal{A}_1 = (S_1 < l_1)$ and for $j = 2, \ldots, K$,

$$\mathcal{R}_j = (S_1 \in \Gamma_j, T_j > u_2)$$

and

$$\mathcal{A}_j = (S_1 \in \Gamma_j, T_j \leq u_2).$$

Thus, the first stage acceptance and rejection regions are the same as the group sequential test above, but the j-th rejection and acceptance regions are slightly different. Specifically, rejection occurs in the region where the stage one test statistic is contained in the j-th partition of the stage one boundary, in addition to the stage two test statistic being greater than the stage two upper boundary. The acceptance region only differs in the fact that T_j must be less than or equal to the stage two upper boundary.

Next, for all j, the authors define $\bar{\mathcal{R}}_j = \mathcal{R}_1 \cup \cdots \cup \mathcal{R}_j$, $\bar{\mathcal{A}}_j = \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_j$ and $C_j = (\bar{\mathcal{R}}_j \cup \bar{\mathcal{A}}_j) = \bar{\mathcal{R}}_j^c \cap \bar{\mathcal{A}}_j^c$ by De Morgan's law [120], where "c" denotes the complement. Thus, $\bar{\mathcal{R}}_j$ is the union of the rejection regions at all j time points, $\bar{\mathcal{A}}_j$ is the union of the acceptance regions at all j time points and C_j is the continuation region at time point j. The doublet $(\bar{\mathcal{R}}_K, \bar{\mathcal{A}}_K)$ was used to represent the overall acceptance and rejection region, where the acceptance region is equal to the complement of the rejection region and vice versa (i.e., $\bar{\mathcal{A}}_K = \bar{\mathcal{R}}_K^c$ and $\bar{\mathcal{R}}_K = \bar{\mathcal{A}}_K^c$). Every test with level α satisfies $P_0(\bar{\mathcal{R}}_K) = \alpha$ or equivalently $P_0(\bar{\mathcal{A}}_K) = 1 - \alpha$, where P_0 is the probability calculated under H_0. Therefore, the probability of rejecting H_0 at or before time point j is $P_0(\bar{\mathcal{R}}_j) = \alpha_j$, where $\alpha_1 \leq \cdots \leq \alpha_K = \alpha$ characterize the non-decreasing α-spending function of the test. In the same way, the probability of accepting H_0 at or before time point j can be expressed as $P_0(\bar{\mathcal{A}}_j) = \theta_j$, where $\theta_1 \leq \cdots \leq \theta_K = 1 - \alpha$ characterize the θ-spending function. The authors then define what it means for a test in this class to be optimal, as shown below.

Definition. *Among all group sequential tests* $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, ..., K\}$ *with a pre-specified* (α, θ)*-spending function such that* $P_0(\bar{\mathcal{R}}_j) = \alpha_j$ *and* $P_0(\bar{\mathcal{A}}_j) = \theta_j$*, the test* $\{(\mathcal{R}_j^{opt}, \mathcal{A}_j^{opt}), j = 1, \ldots, K\}$ *is optimal within this class if* $P_1(\bar{\mathcal{R}}_j^{opt}) \geq P_1(\bar{\mathcal{R}}_j)$ *for all* j*.*

Thus, if the optimal test exists, it is the most powerful test among all group sequential tests with a pre-specified (α, θ)-spending function and, under H_1, has a greater chance of rejecting H_0 at or before any time point j. This does not suggest that the optimal test will surely stop the study on average earlier under H_1, since the study could also be terminated due to accepting H_0. Rather, this implies that the optimal test will stop the trial more often earlier to make the correct decision.

Next, the authors show that for a pre-specified (α, θ)-spending function, their group sequential test based on the sequential LRT statistic is optimal, where the likelihood ratio was denoted by

$$L_j = L_j(x_j) = \frac{p_{1j}(x_j)}{p_{0j}(x_j)}$$

and the sequential LRT statistic was denoted by

$$\bar{L}_j = \bar{L}_j(y_j) = \prod_{l=1}^{j} L_l(x_l)$$

for all j, with $\bar{L}_j$ being assumed absolutely continuous under H_0. Similarly to the group sequential example in the last section, the level α group sequential test based on the sequential LRT statistic, with a pre-specified (α, θ)-spending function, was defined as $\{(\mathcal{R}_j^{\mathrm{LR}}, \mathcal{A}_j^{\mathrm{LR}}), j = 1, \ldots, K\}$, where $\mathcal{R}_1^{\mathrm{LR}} = (\bar{L}_1 > u_1^{\mathrm{LR}})$, $\mathcal{A}_1^{\mathrm{LR}} = (\bar{L}_1 < l_1^{\mathrm{LR}})$ and for $j = 2, \ldots, K$,

$$\mathcal{R}_j^{\mathrm{LR}} = (l_1^{\mathrm{LR}} \leq \bar{L}_1 \leq u_1^{\mathrm{LR}}, \ldots, l_{j-1}^{\mathrm{LR}} \leq \bar{L}_{j-1} \leq u_{j-1}^{\mathrm{LR}}, \bar{L}_j > u_j^{\mathrm{LR}})$$

and

$$\mathcal{A}_j^{\mathrm{LR}} = (l_1^{\mathrm{LR}} \leq \bar{L}_1 \leq u_1^{\mathrm{LR}}, \ldots, l_{j-1}^{\mathrm{LR}} \leq \bar{L}_{j-1} \leq u_{j-1}^{\mathrm{LR}}, \bar{L}_j < l_j^{\mathrm{LR}}).$$

Here, the stage one rejection region based on the likelihood ratio is where the stage one LRT statistic is larger than the stage one upper boundary, and the stage one acceptance region based on the likelihood ratio is where the stage one LRT statistic is less than the stage one lower boundary. For $j = 2, \ldots, K$, the j-th acceptance/rejection region based on the likelihood ratio is again comprised of the continuation regions up to the stage j acceptance/rejection, and each of the stage-wise upper and lower boundaries are constructed in a way that ensures that the test will have level α. In particular, $P_0(\mathcal{R}_1^{\mathrm{LR}}) = \alpha_1$, $P_0(\mathcal{A}_1^{\mathrm{LR}}) = \theta_1$ and for $j = 2, \ldots, K$, $P_0(\mathcal{R}_j^{\mathrm{LR}}) = \alpha_j - \alpha_{j-1}$ and $P_0(\mathcal{A}_j^{\mathrm{LR}}) = \theta_j - \theta_{j-1}$.

Finally, the authors produce the theorem below to formalize their optimal group sequential test based on the sequential LRT statistic.

Theorem 1. *The optimal group sequential test with a pre-specified (α, θ)-spending function is the group sequential test $\{(\mathcal{R}_j^{LR}, \mathcal{A}_j^{LR}), j = 1, \ldots, K\}$, derived using the sequential LRT statistics defined above.*

The authors used the following general process to begin proving their theorem. First, it can be shown that $P_1(\bar{\mathcal{R}}_j) = \mathrm{E}_0(\bar{L}_j \mathrm{I}_{\bar{\mathcal{R}}_j})$ using a Borel measurability argument and the definition of expectation, where $\mathrm{I}_{\bar{\mathcal{R}}_j} = 1$ when $\bar{\mathcal{R}}_j$ is true and $\mathrm{I}_{\bar{\mathcal{R}}_j} = 0$ otherwise. In other words, $\mathrm{I}_{\bar{\mathcal{R}}_j} = 1$ if a rejection occurs at stage j. Therefore, among the class of group sequential tests $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \ldots, K\}$ with a pre-specified (α, θ)-spending function, the optimal group sequential test $\{(\mathcal{R}_j^{\mathrm{opt}}, \mathcal{A}_j^{\mathrm{opt}}), j = 1, \ldots, K\}$, if it exists, must satisfy $\mathrm{E}_0(\bar{L}_j \mathrm{I}_{\bar{\mathcal{R}}_j^{\mathrm{opt}}}) \geq \mathrm{E}_0(\bar{L}_j \mathrm{I}_{\bar{\mathcal{R}}_j})$. Thus, the expectation using the optimal is always larger than or equal to the expectation using any other $\bar{\mathcal{R}}_j$. This is quite complicated to prove directly, so it suffices to show that $\bar{L}_j \mathrm{I}_{\bar{\mathcal{R}}_j^{\mathrm{opt}}}$ is stochastically larger than or equal to $\bar{L}_j \mathrm{I}_{\bar{\mathcal{R}}_j}$ under H_0 using induction and a generalized Neyman-Pearson lemma. More details and the remaining steps of the proof are provided in the Appendix of Tsiatis and Mehta (2003) [66].

According to the authors' theorem, and using the normality approximation in the group sequential example in the previous section, the LRT statistics can be written as $\bar{L}_j = \exp(\mu_1 S_1 - n_j \mu_1^2/2)$, where $S_j = Z_1 + \cdots + Z_{n_j}$. Since $\bar{L}_j$ is a strictly monotone increasing function of S_j, the optimal group sequential test is equivalent to the test that rejects H_0 at time j if

$$\mathcal{R}_j^{\mathrm{LR}} = (l_1 \leq S_1 \leq u_1, \ldots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j > u_j)$$

and accepts H_0 at time j if

$$\mathcal{A}_j^{\mathrm{LR}} = (l_1 \leq S_1 \leq u_1, \ldots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j < l_j).$$

Because the normal distribution has the monotone likelihood ratio (MLR) property, this test is independent of μ_1 and thus is uniformly optimal for testing H_1. Although all of these results may seem reasonable on the surface, the next section will discuss some fundamental issues with the authors' paper.

6.2 Criticism

There are various aspects of Tsiatis and Mehta (2003) [66] that raise concerns. In their article, they state that "a consequence of Theorem 1 is that any adaptive test will be less efficient ... or dominated by the corresponding standard group sequential test based on the likelihood ratio test statistic with the same (α, θ)-spending function." The authors also insist that "for any adaptive design, one can always construct a standard group sequential test based on the sequential likelihood ratio test statistic that, for any parameter value in the space of alternatives, will reject the null hypothesis earlier with higher probability, and, for any parameter value not in the space of alternatives, will accept the null hypothesis earlier with higher probability." Unfortunately, both claims are poorly grounded, as no rigorous proofs of them were supplied, and the conditions under which they arrived at their conclusions are questionable.

The authors obtained their optimality results under the special case where (1) there was a fixed set of null and alternative hypotheses, (2) there were fixed scheduled interim analysis times, that is, at times when n_j subjects completed the study, and a fixed set of (α, θ)-spending functions, (3) their efficacy was measured solely by the statistical power, (4) the generalized Neyman-Pearson lemma led to optimality only when the alternative is a finite set (recall that the traditional Neyman-Pearson lemma assumes a single alternative), and (5) they assumed that data were collected independently in each interim analysis.

Item (1) can be too restrictive when sponsors would like to have the flexibility to add or drop treatments based on interim findings. Items (2) and (3) are related in the sense that both are assuming a Neyman-Pearson paradigm in which the target is to control the type I error and maximize the power. However, when the adaptive design becomes more seamless and interim decisions are made via trial data and other external information, the benefit of the study should not be measured only by power and must be evaluated more comprehensively. Item (4) is almost always violated, because hypotheses that are tested in practice are usually compositional and multi-dimensional, which render the generalized Neyman-Pearson lemma inapplicable. Finally, item (5) is too restrictive for multi-stage seamless designs, like the ones proposed in this dissertation, where patients contribute longitudinal, and hence correlated, data at different stages.

To summarize, the benefits of the adaptive design over the group sequential design must be evaluated more fairly and in a broader framework. For example, instead of just relying on power for assessing efficacy, optimality could be redefined through a Bayesian approach by introducing a more appropriate loss function (typical power-based designs use a 0-1 loss function) that accounts for other important features, such as trial cost and duration [129]. Further research on this is needed and will be a part of my ongoing work.

Chapter 7: Discussion

This study directly contributes to the development and evaluation of different clinical trial design options and analytical techniques that can be used to advance regulatory science. In several common data situations, our proposed three-stage phase II/III transitional seamless design (with and without interim sample size recalibration) exhibits potential in accelerating assessment in order to more quickly bring drugs and other medical products to populations in need, which is applicable to all therapeutic areas. Our proposed design is also attractive due to its efficiency (e.g., reducing the lead time between the phase II study and the phase III study) and flexibility (e.g., permitting early treatment/dosage selection and trial termination).

Clinical trial designs with bivariate endpoints were first considered at the turn of the century [130] and designs with endpoint switch were developed soon after [119, 131], but some of these assumed that the correlation between the intermediate and definitive endpoint data was known and used simulations to guide its choice. While such an assumption is computationally convenient, it may be practically difficult, especially when there is limited knowledge about their relationship. As a contrast, our proposed design eliminates the need for information regarding the correlation between the phase II and phase III endpoints, which can be difficult to quantify in practice; this feature avoids downstream consequences that could be caused by misspecifying the correlation. Yet, once intermediate and definitive endpoint data is collected, their correlation could be analyzed to ensure that the intermediate endpoint is indicative of the definitive endpoint using similar methods to those from the COPE-HCV study outlined in Section 1.4.

Compared to the standard two-stage design, an advantage of the proposed three-stage design is that it gives the treatment under study more of an opportunity to prove itself. In other words, at the second interim analysis, if the intermediate endpoint data is promising but the definitive endpoint data is not, then the third stage allows for more definitive endpoint data to be collected and analyzed, which can help to protect against missing out on a therapy that is actually effective. Also, as with all seamless designs that involve treatment selection, the proposed design allows for early identification of ineffective or unsafe therapies at the first interim analysis, thereby improving patient safety. Although we chose the highest performing treatment to be the one with the largest treatment effect via (asymptotic) pairwise Z-statistics, there are also other techniques such as using the predictive probability of success (PPS) or the probability of being the best (PBB) [132].

The PPS for treatment $i = 1, \ldots, k$ at interim can be defined as $p_i^s = P(d_i > d)$, where d_i is the difference in means between treatment i and the control based on the interim data and d is the minimum clinically meaningful effect. Denoting the observed interim difference for treatment i as $\bar{X}_i$ and the pooled sample standard deviation as S_i, the PPS for treatment i can be expressed as

$$p_i^s = \Phi\left(\frac{\bar{X}_i - \bar{X}_0 - d}{S_i}\right)$$

where Φ is the standard normal cumulative distribution function and the most promising treatment at interim will have the largest p_i^s. Alternatively, the PBB for treatment $i = 1, \ldots, k$ at interim can be defined as $p_i^b = P(\max_i\{d_i\} = d_i)$, where d_i, $\bar{X}_i$ and S_i represent the same quantities as above. Since it can be shown that $(\bar{X}_{i_1}, S_{i_1}^2)$ and $(\bar{X}_{i_2}, S_{i_2}^2)$ are independent for all $i_1 > i_2$, the PBB for treatment i can be expressed as

$$p_i^b = \int \frac{1}{S_i}\phi\left(\frac{z - (\bar{X}_i - \bar{X}_0)}{S_i}\right) \prod_{g=1,\, i\neq g}^{k} \Phi\left(\frac{z - (\bar{X}_g - \bar{X}_0)}{S_g}\right) dz$$

where ϕ is the standard normal probability density function and the best treatment at interim will have the greatest p_i^b. Both of these approaches could be incorporated into our proposed procedures but different statistical methods may be required for validation. Lastly, although it would be very unlikely in practice for a tie to occur at interim based on how our proposed procedures are constructed, a sponsor could use the PBB first and then apply the PPS if there is somehow a tie.

One major concern when applying adaptations in clinical trials is how to control the overall type I error rate at a pre-specified level of significance. In Chapters 2-5, the proposed unmodified and modified design procedures were described and statistically justified with supporting simulations when employing (1) a continuous intermediate endpoint and a continuous definitive endpoint, (2) a binary intermediate endpoint and a continuous definitive endpoint, (3) a binary intermediate endpoint and a time-to-event definitive endpoint without censoring, and (4) a binary intermediate endpoint and a time-to-event definitive endpoint with censoring, respectively. Chapter 4 considered a time-to-event definitive endpoint in the absence of censoring, which can be quite rare, but this material was used for Chapter 5 to construct a framework that accounts for censoring events. Note that Chapters 4-5 do not include modified versions; our research is ongoing on this topic, including the introduction of a cohort one sample size re-estimation procedure at the end of stage one. Finally, in Chapter 6, we promote the proper appraisal and implementation of adaptive methods in clinical trials to enhance regulatory decision-making based on data-driven evidence.

While both versions of the proposed design have many benefits, there are also some limitations. First, we have verified that the proposed design is statistically sound for the data scenarios mentioned above under particular assumptions and using specific data models, but there are other pairs of endpoint types that can be selected based on the study objectives (e.g., a binary intermediate endpoint and a binary definitive endpoint or a time-to-event intermediate endpoint and a time-to-event definitive endpoint) and the endpoints may not follow our stated assumptions or bivariate normal distributions. This necessitates further work to increase the generalizability of the proposed design by investigating how our assumptions could be relaxed or removed, and what happens when the endpoints have different joint distributions. For example, the intermediate and definitive endpoint data may yield opposing conclusions on some occasions, though they are usually expected to align, and they may not bear a linear or monotonic relationship. This is also unexpected in practice, but the correlation between the endpoints could be negative, which may affect the statistical properties of our design. Furthermore, not having to specify an exact joint distribution for the intermediate and definitive endpoint data would be ideal, but would likely involve more complex processes such as the Markov chain Monte Carlo (MCMC) algorithm [133] or the expectation-maximization (EM) algorithm [134].

Along with this, for the first two data scenarios discussed in Chapters 2-3, the variance of the intermediate and definitive endpoint data must be known before trial start, so that the design parameters needed to maintain a certain type I error rate and reach a certain power can be obtained. Determining if the parameters can be acquired without knowledge of the data variability could be explored so that direct treatment effects rather than relative treatment effects can always be measured, or seeing if the variances can be adapted at interim if they are misspecified. Our design also uses only the cohort one and cohort two intermediate endpoint data (no definitive endpoint data) for the futility test at the end of stage two, and uses only the cohort one and cohort two definitive endpoint data (no intermediate endpoint data) at the final analysis. The former feature is pertinent when there are only a few treatment options available (e.g., there are currently only two FDA-approved drugs, nintedanib and pirfenidone, available for treating idiopathic pulmonary fibrosis); it is important to further explore an experimental intervention as long as the intermediate endpoint data is compelling. The second and final analyses could include the cohort one definitive endpoint data and the cohort one and cohort two intermediate endpoint data, respectively, by adding additional parameters to the unmodified and modified procedures and slightly altering our statistical steps. However, the superiority tests are of most interest, where a claim of efficacy can only be made based on the definitive endpoint data. It is likely that doing this would also require higher sample sizes, which is undesirable.

There are various additional future directions to consider that could expand or strengthen our proposed design. First, the design combines phase II and phase III studies, but our structure may be transferable to merging phase I and phase II studies. Nonetheless, since phase I/II studies typically have smaller sample sizes than phase II/III studies, confirming the statistical validity of our procedures in these settings may require alternative approaches. Second, our design does not permit cohort one subjects who are removed from the trial at the end of stage one to enter the second cohort due to potential carryover effects and loss of independence; however, this could facilitate recruitment and lessen some of the burden on study personnel. One possible route would be to incorporate response-adaptive randomization [135], but this would complicate how to ensure that the nominal type I error rate will not be exceeded [71]. Specifically, subjects from treatment groups that were removed after stage one could be re-enrolled based on their reaction to their assigned intervention and its relation to the highest performing therapy that was identified. Moreover, carryover effects could be overcome by using a washout period with a duration chosen according to the therapy's mechanism of action, though this would still not address the independence violation.

It is also worth looking into if our proposed design and statistical procedures remain valid when the sample sizes in each treatment arm are not exactly equal, which is often the case in real-world trials due to dropouts and loss to follow-up, or when there is unequal allocation at baseline, which can be used when a therapy is being evaluated for a condition that has no available interventions. When the groups are closely equal in size, there should be no major issues with the functioning of our design but if the group sizes substantially differ, this would need to be examined further. The methods presented here could be extended to seamless designs with more than three stages, but adding more stages would demand clinical and statistical justification. Another consideration is how to adjust for important variables on top of treatment arm, like age and sex. As this would go beyond the bivariate case, the multidimensional version of Slepian's lemma [136] could perhaps be used for proving that the operational characteristics of the design would not be compromised. Finally, only positive critical value parameters ($c_1, c_2, d_1, d_2 > 0$) were considered in this study, but negative boundaries could be more suitable in some scenarios, and more simulations are required to verify our proposed design procedures for different values of k, σ^2, τ^2, δ_1, and δ_2.

References

[1] S. C. Chow and M. Chang, *Adaptive Design Methods in Clinical Trials*. Chapman and Hall/CRC, 2006.

[2] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger, *Fundamentals of Clinical Trials*. Springer, 2015.

[3] A. Granholm *et al.*, "Randomised clinical trials in critical care: Past, present and future," *Intensive care medicine*, pp. 1–15, 2022.

[4] D. M. Harmon, "The digitization and decentralization of clinical trials," *Mayo Clinic Proceedings*, 2023.

[5] S. Harrer, P. Shah, B. Antony, and J. Hu, "Artificial intelligence for clinical trial design," *Trends in pharmacological sciences*, vol. 40, 8 2019.

[6] D. Machin, P. M. Fayers, and B. C. Tai, *Randomized Clinical Trials: Design, Practice and Reporting*. John Wiley & Sons, 2021.

[7] P. S. Mishra-Kalyani *et al.*, "External control arms in oncology: Current use and future directions," *Annals of Oncology*, 2022.

[8] S. Piantadosi, *Clinical Trials: A Methodologic Perspective*. John Wiley & Sons, 2017.

[9] K. D. Wick *et al.*, "Opportunities for improved clinical trial designs in acute respiratory distress syndrome," *The Lancet Respiratory Medicine*, vol. 10, pp. 916–924, 9 2022.

[10] Biotechnology Innovation Organization, *Drug development, review & lifecycle management*, `https://www.bio.org/policy/human-health/drug-development-review-lifecycle-management`, Accessed: 2022-12-07, 2022.

[11] B. E. Storer, "An evaluation of phase I clinical trial designs in the continuous dose–response setting," *Statistics in Medicine*, vol. 20, pp. 2399–2408, 16 2001.

[12] ——, "Design and analysis of phase I clinical trials," *Biometrics*, pp. 925–937, 1989.

[13] E. Garrett-Mayer, "The continual reassessment method for dose-finding studies: A tutorial," *Clinical Trials*, vol. 3, pp. 57–71, 1 2006.

[14] J. O'Quigley, M. Pepe, and L. Fisher, "Continual reassessment method: A practical design for phase 1 clinical trials in cancer," *Biometrics*, pp. 33–48, 1990.

[15] S. Senar, *Machine learning applied to drug discovery*, `https://doctortarget.com/machine-learning-applied-drug-discovery/`, Accessed: 2022-12-07, 2022.

[16] D. C. Angus *et al.*, "The remap-cap (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. rationale and design," *Annals of the American Thoracic Society*, vol. 17, pp. 879–891, 7 2020.

[17] S. E. Carlson *et al.*, "Higher dose docosahexaenoic acid supplementation during pregnancy and early preterm birth: A randomised, double-blind, adaptive-design superiority trial," *EClinicalMedicine*, vol. 36, p. 100 905, 2021.

[18] J. F. Donohue, C. Fogarty, and J. Lötvall, "Once-daily bronchodilators for chronic obstructive pulmonary disease: Indacaterol versus tiotropium," *American Journal of Respiratory and Critical Care Medicine*, vol. 182, pp. 155–162, 2 2010.

[19] D. N. Hager *et al.*, "The vitamin C, thiamine and steroids in sepsis (victas) protocol: A prospective, multi-center, double-blind, adaptive sample size, randomized, placebo-controlled, clinical trial," *Trials*, vol. 20, pp. 1–16, 1 2019.

[20] C. Hünseler *et al.*, "Continuous infusion of clonidine in ventilated newborns and infants: A randomized controlled trial," *Pediatric Critical Care Medicine*, vol. 15, pp. 511–522, 6 2014.

[21] E. Israel *et al.*, "Precise: Precision medicine in severe asthma: An adaptive platform trial with biomarker ascertainment," *Journal of Allergy and Clinical Immunology*, vol. 147, pp. 1594–1601, 5 2021.

[22] V. Li *et al.*, "Designing multi-arm multistage adaptive trials for neuroprotection in progressive multiple sclerosis," *Neurology*, vol. 98, pp. 754–764, 18 2022.

[23] E. M. McCreedy *et al.*, "Measuring the effects of a personalized music intervention on agitated behaviors among nursing home residents with dementia: Design features for cluster-randomized adaptive trial," *Trials*, vol. 22, pp. 1–12, 1 2021.

[24] C. Morgan *et al.*, "Adaptive design: Results of 2012 survey on perception and use," *Therapeutic Innovation & Regulatory Science*, vol. 48, pp. 473–481, 4 2012.

[25] S. Paganoni *et al.*, "Adaptive platform trials to transform amyotrophic lateral sclerosis therapy development," *Annals of neurology*, vol. 91, pp. 165–175, 2 2022.

[26] J. P. Zajicek, J. C. Hobart, A. Slade, D. Barnes, P. G. Mattison, and MUSEC Research Group, "Multiple sclerosis and extract of cannabis: Results of the MUSEC trial," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 83, pp. 1125–1132, 11 2012.

[27] U. Zeymer *et al.*, "The Na+/H+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. results of the evaluation of the safety and cardioprotective effects of eniporide in acute myocardial infarction (ESCAMI) trial," *Journal of the American College of Cardiology*, vol. 38, E1644–E1650, 6 2001.

[28] R. Mahajan and K. Gupta, "Food and drug administration's critical path initiative and innovations in drug development paradigm: Challenges, progress, and controversies," *Journal of Pharmacy And Bioallied Sciences*, vol. 2, p. 307, 4 2010.

[29] P. R. Chaturvedi, Z. Antonijevic, and C. Mehta, "Practical considerations for a two-stage confirmatory adaptive clinical trial design and its implementation: ADVENT trial," *Practical considerations for adaptive trial design and implementation*, pp. 383–411, 2014.

[30] F. Bretz, F. Koenig, W. Brannath, E. Glimm, and M. Posch, "Adaptive designs for confirmatory clinical trials," *Statistics in Medicine*, vol. 28, pp. 1181–1217, 8 2009.

[31] L. Cui, H. M. J. Hung, and S. J. Wang, "Modification of sample size in group sequential clinical trials," *Biometrics*, vol. 55, pp. 853–857, 3 1999.

[32] L. Flight, S. Julious, A. Brennan, and S. Todd, "Expected value of sample information to guide the design of group sequential clinical trials," *Medical Decision Making*, vol. 42, pp. 461–473, 4 2022.

[33] T. Friede and N. Stallard, "A comparison of methods for adaptive treatment selection," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 50, pp. 767–781, 5 2008.

[34] C. Jennison and B. W. Turnbull, "Mid-course sample size modification in clinical trials based on the observed treatment effect," *Statistics in Medicine*, vol. 22, pp. 971–993, 6 2003.

[35] M. Kieser, P. Bauer, and W. Lehmacher, "Inference on multiple endpoints in clinical trials with adaptive interim analyses," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 41, pp. 261–277, 3 1999.

[36] M. Kieser and T. Friede, "Simple procedures for blinded sample size adjustment that do not affect the type I error rate," *Statistics in medicine*, vol. 22, pp. 3571–3581, 23 2003.

[37] F. Koenig, W. Brannath, F. Bretz, and M. Posch, "Adaptive dunnett tests for treatment selection," *Statistics in medicine*, vol. 27, pp. 1612–1625, 10 2008.

[38] Q. Liu, M. A. Proschan, and G. W. Pledger, "A unified theory of two-stage adaptive designs," *Journal of the American Statistical Association*, vol. 97, pp. 1034–1041, 460 2002.

[39] M. Posch, F. Koenig, M. Branson, W. Brannath, C. Dunger-Baldauf, and P. Bauer, "Testing and estimation in flexible group sequential designs with adaptive treatment selection," *Statistics in medicine*, vol. 24, pp. 3697–3714, 24 2005.

[40] S. J. Wang, H. J. Hung, and R. T. O'Neill, "Adapting the sample size planning of a phase III trial based on phase II data," *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, vol. 5, pp. 85–97, 2 2006.

[41] FDA, *Adaptive design clinical trials for drugs and biologics guidance for industry*, `https://www.fda.gov/media/78495/download`, Accessed: 2022-11-06, 2019.

[42] ——, *Adaptive design clinical trials for drugs and biologics*, `https://www.emwa.org/Documents/Freelancer/adaptivestudydesign/FDA%20adaptive%20design_Aug%202010.pdf`, Accessed: 2023-01-16, 2010.

[43] EMA, *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*, `https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf`, Accessed: 2023-01-16, 2007.

[44] P. Bauer, F. Bretz, V. Dragalin, F. König, and G. Wassmer, "Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls," *Statistics in Medicine*, vol. 35, pp. 325–347, 3 2016.

[45] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller, *Bayesian adaptive methods for clinical trials*. CRC press, 2010.

[46] W. Brannath *et al.*, "Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology," *Statistics in medicine*, vol. 28, pp. 1445–1463, 10 2009.

[47] A. C. Graf, M. Posch, and F. Koenig, "Adaptive designs for subpopulation analysis optimizing utility functions," *Biometrical Journal*, vol. 57, pp. 76–89, 1 2015.

[48] S. C. Chow and R. Corey, "Benefits, challenges and obstacles of adaptive clinical trial designs," *Orphanet Journal of Rare Diseases*, vol. 6, pp. 1–10, 1 2011.

[49] J. A. Kairalla, C. S. Coffey, M. A. Thomann, and K. E. Muller, "Adaptive trial designs: A review of barriers and opportunities," *Trials*, vol. 13, pp. 1–9, 1 2012.

[50] Vanderbilt Institute for Infection, Immunology and Inflammation, *A covid-19 vaccine: Why we need it and how we get there yesterday*, `https://www.vumc.org/viiii/immuknow/covid-vaccine-pandemic-speed`, Accessed: 2022-11-07, 2020.

[51] M. Proschan, G. K. K. Lan, and J. T. Wittes, *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer Science & Business Media, 2006.

[52] H. Schäfer and H. H. Müller, "Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections," *Statistics in medicine*, vol. 20, pp. 3741–3751, 24 2001.

[53] N. W. Bean, J. G. Ibrahim, and M. A. Psioda, "Bayesian design of multi-regional clinical trials with time-to-event endpoints," *Biometrics*, 2023.

[54] C. Chen and E. H. Rubin, "Adaptive phase 2/3 designs for oncology drug development–time to hedge," *Contemporary Clinical Trials*, vol. 125, p. 107 047, 2023.

[55] Q. Li, J. Lin, M. Liu, L. Wu, and Y. Liu, "Using surrogate endpoints in adaptive designs with delayed treatment effect," *Statistics in Biopharmaceutical Research*, vol. 14, pp. 661–670, 4 2022.

[56] A. Mukherjee, D. S. Coad, and S. Jana, "Covariate-adjusted response-adaptive designs for censored survival responses," *Journal of Statistical Planning and Inference*, 2023.

[57] L. Z. Sun, W. Li, C. Chen, and J. Zhao, "Advanced utilization of intermediate endpoints for making optimized cost-effective decisions in seamless phase II/III oncology trials," *Statistics in Biopharmaceutical Research*, vol. 12, pp. 224–233, 2 2020.

[58] X. Wang, T. Cai, L. Tian, F. Bourgeois, and L. Parast, "Quantifying the feasibility of shortening clinical trial duration using surrogate markers," *Statistics in medicine*, vol. 40, pp. 6321–6343, 28 2021.

[59] L. Wu, Q. Li, M. Liu, and J. Lin, "Incorporating surrogate information for adaptive subgroup enrichment design with sample size re-estimation," *Statistics in Biopharmaceutical Research*, vol. 14, pp. 493–504, 4 2022.

[60] C. F. Burman and C. Sonesson, "Are flexible designs sound?" *Biometrics*, vol. 62, pp. 664–669, 3 2006.

[61] P. Gallo, "Operational challenges in adaptive design implementation," *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, vol. 5, pp. 119–124, 2 2006.

[62] E. L. Korn and B. Friedlin, "Outcome-adaptive randomization: Is it useful?" *Journal of Clinical Oncology*, vol. 29, p. 771, 6 2011.

[63] R. van der Graaf, K. C. B. Roes, and J. J. M. van Delden, "Adaptive trials in clinical research: Scientific and ethical issues to consider," *JAMA*, vol. 307, pp. 2379–2380, 22 2012.

[64] J. Wason, P. Brocklehurst, and C. Yap, "When to keep it simple–adaptive designs are not always useful," *BMC Medicine*, vol. 17, pp. 1–7, 1 2019.

[65] C. Jennison and B. W. Turnbull, "Adaptive and nonadaptive group sequential tests," *Biometrika*, vol. 93, pp. 1–21, 1 2006.

[66] A. A. Tsiatis and C. Mehta, "On the inefficiency of the adaptive design for monitoring clinical trials," *Biometrika*, vol. 90, pp. 367–378, 2 2003.

[67] A. Anderer, H. Bastani, and J. Silberholz, "Adaptive clinical trial designs with surrogates: When should we bother?" *Management Science*, vol. 68, pp. 1982–2002, 3 2022.

[68] F. Bretz, H. Schmidli, F. König, A. Racine, and W. Maurer, "Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 48, pp. 623–634, 4 2006.

[69] Y. J. Chen, R. Gesser, and A. Luxembourg, "A seamless phase IIB/III adaptive outcome trial: Design rationale and implementation challenges," *Clinical Trials*, vol. 12, pp. 84–90, 1 2015.

[70] S. C. Chow, *Controversial Issues in Clinical Trials*. Chapman & Hall/CRC, 2011.

[71] S. C. Chow and Y. H. Tu, "On two-stage seamless adaptive design in clinical trials," *Journal of the Formosan Medical Association*, vol. 107, S52–S60, 12 2008.

[72] T. Conroy *et al.*, "Folfirinox versus gemcitabine for metastatic pancreatic cancer," *New England journal of medicine*, vol. 364, pp. 1817–1825, 19 2011.

[73] T. Friede, N. Stallard, and N. Parsons, "Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R," *Biometrical Journal*, vol. 62, pp. 1264–1283, 5 2020.

[74] A. Hoering, M. LeBlanc, and J. Crowley, "Seamless phase I-II trial design for assessing toxicity and efficacy for targeted agents," *Clinical Cancer Research*, vol. 17, pp. 640–646, 4 2011.

[75] C. Jennison and B. W. Turnbull, "Adaptive seamless designs: Selection and prospective testing of hypotheses," *Journal of biopharmaceutical statistics*, vol. 17, pp. 1135–1161, 6 2007.

[76] C. Léauté-Labrèze *et al.*, "A randomized controlled trial of oral propranolol in infantile hemangioma," *New England Journal of Medicine*, vol. 72, pp. 735–146, 8 2015.

[77] J. Maca, S. Bhattacharya, V. Dragalin, P. Gallo, and M. Krams, “Adaptive seamless phase II/III designs—background, operational aspects, and examples,” *Drug Information Journal*, vol. 40, pp. 463–473, 4 2006.

[78] D. Palermos *et al.*, “Lung cancer clinical trials with a seamless phase II/III design: Systematic review,” *Journal of Clinical Medicine*, vol. 11, p. 7176, 23 2022.

[79] N. Stallard and S. Todd, “Seamless phase II/III designs,” *Statistical methods in medical research*, vol. 20, pp. 623–634, 6 2011.

[80] N. Stallard, “A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information,” *Statistics in medicine*, vol. 29, pp. 959–971, 9 2010.

[81] J. M. Dambrosia and S. W. Greenhouse, “Early stopping for sequential restricted tests of binomial distributions,” *Biometrics*, pp. 695–710, 1983.

[82] D. L. DeMets and J. H. Ware, “Symmetric group sequential boundaries for monitoring clinical trials,” *Biometrika*, pp. 661–663, 1982.

[83] ——, “Group sequential methods for clinical trials with a one-sided hypothesis,” *Biometrika*, pp. 651–660, 1980.

[84] P. C. O’Brien and T. R. Fleming, “A multiple testing procedure for clinical trials,” *Biometrics*, pp. 549–556, 1979.

[85] P. Bauer and K. Kohne, “Evaluation of experiments with adaptive interim analyses,” *Biometrics*, pp. 1029–1041, 1994.

[86] P. Bauer and J. Röhmel, “An adaptive method for establishing a dose-response relationship,” *Statistics in Medicine*, vol. 14, pp. 1595–1607, 14 1995.

[87] C. Jennison and B. W. Turnbull, *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, 2000.

[88] W. Lehmacher and G. Wassmer, “Adaptive sample size calculations in group sequential trials,” *Biometrics*, vol. 55, pp. 1286–1290, 4 1999.

[89] G. Li, W. J. Shih, and Y. Wang, “Two-stage adaptive design for clinical trials with survival data,” *Journal of Biopharmaceutical Statistics*, vol. 15, pp. 707–718, 4 2005.

[90] Q. Liu and G. Y. H. Chi, “On sample size and inference for two-stage adaptive designs,” *Biometrics*, vol. 57, pp. 172–177, 1 2001.

[91] M. Posch and P. Bauer, “Interim analysis and sample size reassessment,” *Biometrics*, vol. 56, pp. 1170–1176, 4 2000.

[92] M. A. Proschan and S. A. Hunsberger, “Designed extension of studies based on conditional power,” *Biometrics*, pp. 1315–1324, 1995.

[93] N. R. Parsons *et al.*, “Group sequential designs in pragmatic trials: Feasibility and assessment of utility using data from a number of recent surgical RCTs,” *BMC Medical Research Methodology*, vol. 22, pp. 1–18, 1 2022.

[94] S. J. Pocock, “Group sequential methods in the design and analysis of clinical trials,” *Biometrika*, vol. 64, pp. 191–199, 2 1977.

[95] S. K. Wang and A. A. Tsiatis, “Approximately optimal one-parameter boundaries for group sequential trials,” *Biometrics*, pp. 193–199, 1987.

[96] J. Whitehead and I. Stratton, “Group sequential clinical trials with triangular continuation regions,” *Biometrics*, pp. 227–236, 1983.

[97] S. C. Chow, Q. Lu, and S. K. Tse, “Statistical analysis for two-stage seamless design with different study endpoints,” *Journal of Biopharmaceutical Statistics*, vol. 17, pp. 1163–1176, 6 2007.

[98] Q. Lu, S. K. Tse, and S. C. Chow, “Analysis of time-to-event data under a two-stage survival adaptive design in clinical trials,” *Journal of Biopharmaceutical Statistics*, vol. 20, pp. 705–719, 4 2010.

[99] Q. Lu, S. C. Chow, S. K. Tse, Y. Chi, and L. Y. Yang, “Sample size estimation based on event data for a two-stage survival adaptive trial with different durations,” *Journal of Biopharmaceutical Statistics*, vol. 19, pp. 311–323, 2 2009.

[100] S. C. Chow and M Lin, “Analysis of two-stage adaptive seamless trial design,” *Pharmaceutica Analytica Acta*, vol. 6, p. 2, 341 2015.

[101] L. Cui, T. Zhan, L. Zhang, Z. Geng, Y. Gu, and I. S. Chan, “An automation-based adaptive seamless design for dose selection and confirmation with improved power and efficiency,” *Statistical Methods in Medical Research*, vol. 30, pp. 1013–1025, 4 2021.

[102] T. Friede *et al.*, “Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis,” *Statistics in medicine*, vol. 30, pp. 1528–1540, 13 2011.

[103] L. Jiang, R. Li, F. Yan, T. A. Yap, and Y. Yuan, “Shotgun: A bayesian seamless phase I-II design to accelerate the development of targeted therapies and immunotherapy,” *Contemporary clinical trials*, vol. 104, p. 106 338, 2021.

[104] M. Jin and P. Zhang, "A seamless adaptive 2-in-1 design expanding a phase 2 trial for treatment or dose selection into a phase 3 trial," *Statistics in Biopharmaceutical Research*, vol. 14, pp. 334–341, 3 2022.

[105] ——, "An adaptive seamless phase 2-3 design with multiple endpoints," *Statistical Methods in Medical Research*, vol. 30, pp. 1143–1151, 4 2021.

[106] W. Ma, M. Wang, and H. Zhu, "Seamless phase II/III clinical trials with covariate adaptive randomization," *Statistica Sinica*, 2020.

[107] K. Takahashi, R. Ishii, K. Maruo, and M. Gosho, "Statistical tests for two-stage adaptive seamless design using short-and long-term binary outcomes," *Statistics in medicine*, vol. 41, pp. 4130–4142, 21 2022.

[108] P. Zhang, X. N. Li, K. Lu, and C. Wu, "A 2-in-1 adaptive design to seamlessly expand a selected dose from a phase 2 trial to a phase 3 trial for oncology drug development," *Contemporary Clinical Trials*, vol. 122, p. 106 931, 2022.

[109] P. Zhang, X. N. Li, K. Lu, C. Wu, and M. Ge, "A variation of a 2-in-1 adaptive design to seamlessly expand a selected dose from a phase 2 trial to a phase 3 trial for oncology drug development," *Contemporary Clinical Trials*, vol. 127, p. 107 119, 2023.

[110] National Library of Medicine, *Safety and efficacy study to compare continuous infusion of interferon with standard of care for chronic hepatitis C (COPE-HCV)*, https://clinicaltrials.gov/ct2/show/study/NCT00919633?term=COPE-HCV&draw=2&rank=1, Accessed: 2022-11-17, 2019.

[111] H. Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, 2011.

[112] D. Slepian, "The one-sided barrier problem for gaussian noise," *Bell System Technical Journal*, vol. 41, pp. 463–501, 2 1962.

[113] D. Borsboom and G. J. Mellenbergh, "The theoretical status of latent variables," *Psychological review*, vol. 110, p. 203, 2 2003.

[114] P. Jaworski, F. Durante, W. K. Hardle, and T. Rychlik, *Copula theory and its applications*. Springer, 2010.

[115] R. B. Nelsen, *An introduction to copulas*. Springer Science & Business Media, 2007.

[116] S. R. Eliason, *Maximum likelihood estimation: Logic and practice*. Sage, 1993.

[117] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of Mathematical Psychology*, vol. 47, pp. 90–100, 1 2003.

[118] C. Cox, “Delta method,” *Encyclopedia of Biostatistics*, vol. 2, 2005.

[119] N. Stallard and S. Todd, “Point estimates and confidence regions for sequential trials involving selection,” *Journal of Statistical Planning and Inference*, vol. 135, pp. 402–419, 2 2005.

[120] A. De Morgan, *Formal logic: or, the calculus of inference, necessary and probable*. Taylor and Walton, 1847.

[121] P. F. Thall, R. Simon, and S. S. Ellenberg, “Two-stage selection and testing designs for comparative clinical trials,” *Biometrika*, vol. 75, pp. 303–310, 2 1988.

[122] Y. Shen and L. Fisher, “Statistical inference for self-designing clinical trials with a one-sided hypothesis,” *Biometrics*, vol. 55, pp. 190–197, 1 1999.

[123] N. A. Weiss, *A Course in Probability*. Boston: Addison–Wesley, 2005, pp. 380–383.

[124] G. B. Thomas, M. D. Weir, J. Hass, F. R. Giordano, and R. Korkmaz, *Thomas' Calculus (Vol. 12)*. Boston: Pearson, 2010.

[125] J. E. Angus, “The probability integral transform and related results,” *SIAM review*, vol. 36, pp. 652–654, 4 1994.

[126] A. Papanicolaou, *Taylor approximation and the delta method*, `https://www.academia.edu/download/52915304/TaylorAppDeltaMethod.pdf`, Accessed: 2023-01-06, 2009.

[127] A. W. Van der Vaart, *Asymptotic Statistics (Vol. 3)*. Cambridge University Press, 2000.

[128] J. Neyman and E. S. Pearson, “IX. On the problem of the most efficient tests of statistical hypotheses,” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 694–706 1933.

[129] J. Aouni, J. Bacro, G. Toulemonde, P. Colin, L. Darchy, and B. Sébastien, “On the use of utility functions for optimizing phase II/phase III seamless trial designs,” *Journal of clinical trials*, vol. 10, 415 2021.

[130] S. Todd, “An adaptive approach to implementing bivariate group sequential clinical trial designs,” *Journal of biopharmaceutical statistics*, vol. 13, pp. 605–619, 4 2003.

[131] Q. Liu and G. W. Pledger, “Phase 2 and 3 combination designs to accelerate drug development,” *Journal of the American Statistical Association*, vol. 100, pp. 493–502, 470 2005.

[132] J. Zheng and S. C. Chow, “Criteria for dose-finding in two-stage seamless adaptive design,” *Journal of Biopharmaceutical Statistics*, vol. 29, pp. 908–919, 5 2019.

[133] A. E. Raftery and S. M. Lewis, *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996, pp. 115–130.

[134] S. K. Ng, T. Krishnan, and G. J. McLachlan, *Handbook of Computational Statistics*. Springer, 2012, pp. 139–172.

[135] H. Zhu, J. Piao, J. J. Lee, F. Hu, and L. Zhang, “Response adaptive randomization procedures in seamless phase II/III clinical trials,” *Journal of biopharmaceutical statistics*, vol. 30, pp. 3–17, 1 2020.

[136] S. D. Gupta, M. L. Eaton, I. Olkin, M. Perlman, L. J. Savage, and M. Sobel, “Inequalities on the probability content of convex regions for elliptically contoured distributions,” *Sixth Berkeley Symposium*, pp. 241–265, 1971.

Appendix A: Plot of the Expectation ζ Defined in Chapter 4

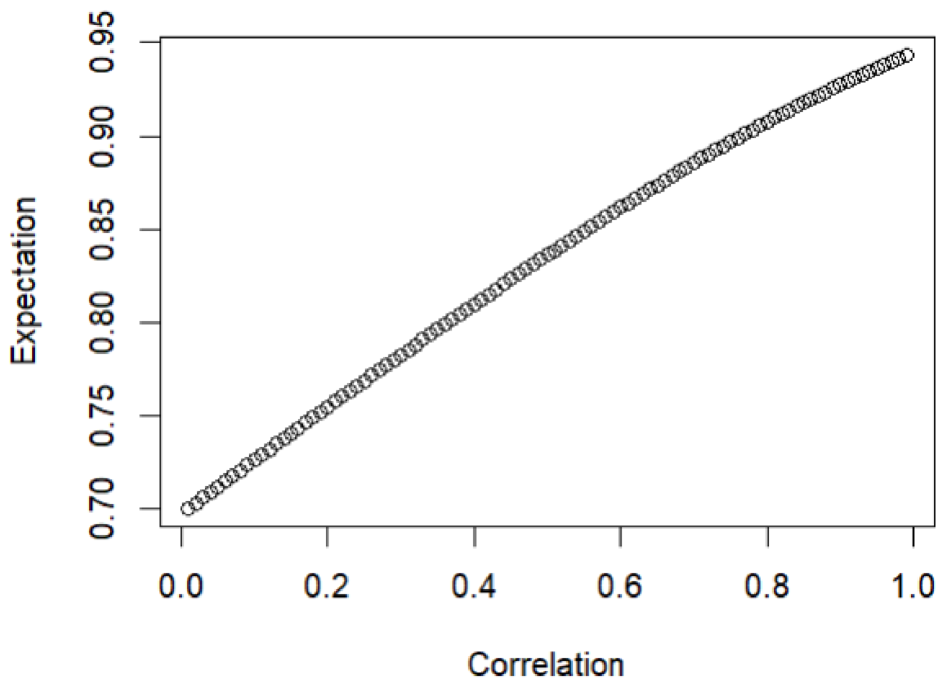

Figure A.1: Plot of the expectation ζ defined in Chapter 4 via 10000 replications with $p_i = 0.7$, $\xi_i = 1$ and V_{ij} simulated from a standard normal distribution 1000 times, demonstrating its strictly increasing nature. Note that the curve will remain strictly increasing irrespective of the choice of $p_i > 0$ and $\xi_i > 0$.

Appendix B: Plot of the Expectation ζ Defined in Chapter 5

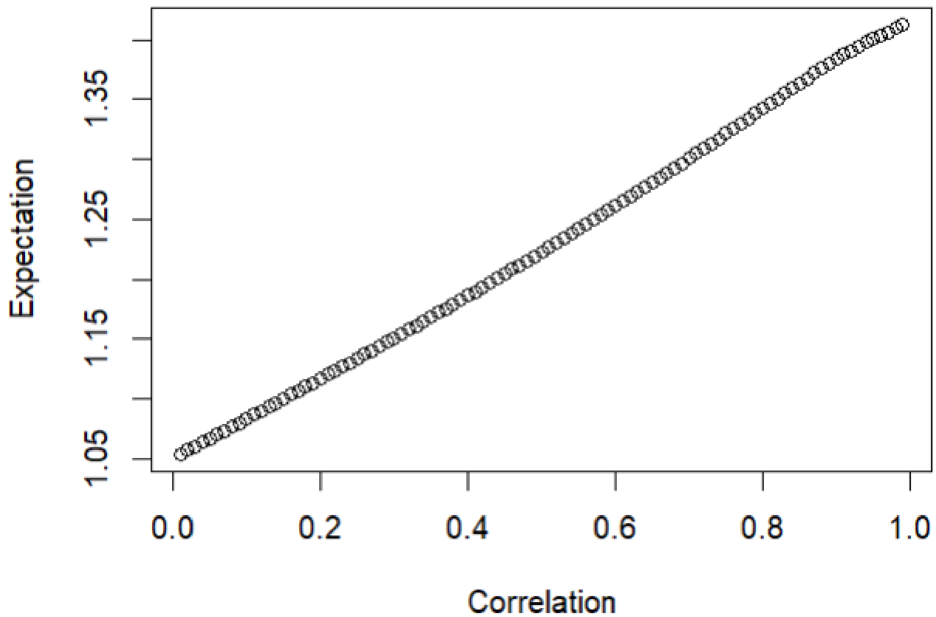

Figure B.1: Plot of the expectation ζ defined in Chapter 5 via 10000 replications with $C = 1$, $p_i = 0.7$, $\xi_i = 1$, and V_{ij} simulated from a standard normal distribution 1000 times, demonstrating its strictly increasing nature. Note that the curve will remain strictly increasing irrespective of the choice of $C > 0$, $p_i > 0$ and $\xi_i > 0$.