Evaluation of Strategies for the Phase II to Phase III Progression in Treatment Discovery

Brittany J. Sanchez

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2014

Committee:
Scott S. Emerson, Chair
Susanne May

Program Authorized to Offer Degree:
School of Public Health, Department of Biostatistics

©Copyright 2014
Brittany J. Sanchez

University of Washington

Abstract

Evaluation of Strategies for the Phase II to Phase III Progression in Treatment Discovery

Brittany J. Sanchez

Chair of the Supervisory Committee:
Professor Scott S. Emerson
Department of Biostatistics

The goal of clinical research is to improve the health of the population through the prevention, diagnosis, and treatment of disease. The drug development process involves a team of scientists trained to identify new treatments, test their efficacy, and evaluate their safety. Beyond the basic science of understanding biologic pathways and molecular mechanisms, treatments with positive health properties are further studied in experiments to identify those that are effective and safe for human use. To enhance this process, we must find a way to increase the number of effective treatments adopted while protecting patients and minimizing the use of time and resources.

Clinical trials are essential for reliably evaluating a proposed treatment to determine whether it should be adopted into clinical practice. Regulatory approval is contingent upon scientifically meaningful and statistically credible results ensuring the safety of the public. Furthermore, results of clinical trials provide the primary information needed for the evidence based medicine that ensures clinical decisions better reflect credible evidence from research. Therefore, current standards involve the evaluation of a new treatment through several phases of investigation. After preliminary evaluations of the safety and ethics of further study, promising treatments are studied in preliminary screening trials and then ultimately large, confirmatory trials. Although well developed, the "treatment discovery process" is lengthy, expensive, and has low success rates for treatments even at confirmatory phases of the investigation. Improvements to trial design and implementation are necessary for better achieving the goals of clinical research.

There has been much discussion in the clinical trials literature about the role of Phase II studies in the treatment discovery process. A major focus of that literature is the Phase II oncology trial

for determining whether further study of a new treatment in confirmatory trials is warranted. Therefore, much of the research on the progression from Phase II studies to Phase III trials tends to be setting-specific and not necessarily generalizable. Few authors have discussed what design parameters may be appropriate more generally for Phase II screening studies when the goal is to screen out useless treatments and to identify useful ones to evaluate in Phase III. Recent interest in making the treatment discovery process more time- and cost-efficient has led to increased focus on adaptive sequential clinical trial designs, which include both the well-studied group sequential designs, as well as more recently proposed approaches. Some aspects of a study design that might be altered in these adaptive designs include the scientific hypotheses, randomization scheme, or sampling plan. There are concerns that the largely neglible gains in efficiency are outweighed by difficulties in the interpretability and, hence, credibility of the results of these trials.

In this research, we consider the progression of studies for investigating a new treatment, and discuss strategies in a framework that encompasses the period from the start of preliminary Phase II studies to the completion of the confirmatory Phase III studies. We first review current practices of clinical investigation including the phasing of clinical trials and the evaluation of a trial design with respect to common frequentist and Bayesian operating characteristics. Using a general notational framework for evaluating new treatments, we discuss the approaches others have taken, and examine optimality criteria for a strategy that best addresses the often competing goals of science, ethics, and efficiency. These optimality criteria include not only the standard frequentist operating characteristics of type I error and power and the standard Bayesian criteria of positive and negative predictive values, but also the efficiency considerations of the number of new treatments identified in a setting with limited resources.

Noting that the current practice of progressing from Phase II to Phase III is in fact a sequential adaptive process, we first focus on a two-stage process consisting of screening Phase II studies and confirmatory Phase III studies. We initially presume a Bayesian probability space in which a population of candidate treatments is to be screened in Phase II studies. We consider both a simplified setting of simple null and alternative hypotheses (a binary prior distribution), with the presumption that the alternative is some minimal clinically important difference (MCID) as well as more general Bayesian priors. We parameterize the Phase II and Phase III designs using frequentist type I error and power in such a way as to attain high Bayesian positive predictive value (PPV). We then explore the impact specific choices of those design parameters have on the number of effective and ineffective treatments identified with constrained resources. We then illustrate how allowing for early trial termination for efficacy or futility with a group sequential design (GSD) within Phase II and/or Phase III improves efficiency in terms of the number of subjects used on average for identifying effective therapies.

Other methods for improving efficiency by eliminating the time spent between Phase II and Phase III have been proposed. A "seamless" Phase II/III trial design is one that combines the Phase II screening stage with the Phase III confirmatory stage. We consider how a single sequential design differs from the optimal approach of independent stages with respect to the timing of the analyses, the frequentist criteria met, and stopping boundaries specified. We explore how the traditional approach of adapting hypotheses at the end of Phase II fits in with the newer adaptive methods. We discuss how powering of Phase III based on Phase II results mimics adaptive sample size re-estimation / re-powering of study and does not offer improvement beyond that of GSDs. Bias in the estimate of the treatment effect is a result of the lack of precision of small samples inherent in Phase II studies and at early interim analyses. We investigate how such bias can be addressed with adjustment methods. We then examine differences between conducting subgroup analyses when there exist homogeneous versus heterogeneous effects and how inflation of the type I error can be controlled in this setting and in the setting of considering multiple summary measures.

We demonstrate that the optimal Phase II to Phase III progression defined by an acceptable PPV and a maximal number of effective treatments can be identified for an anticipated prevalence and hypothesized resources by a parameterization of type I error and power at Phase II. We recognize that several approaches lead to the same optimality criteria, and that the chosen strategy will depend on individual objectives of clinical researchers, trial sponsors, regulatory agencies, patients on study, and those who might benefit from new knowledge about treatments being studied.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Treatment discovery

Clinical trials are essential for evaluating the effectiveness of a new treatment in terms of its superiority, equivalence, or inferiority to treatments used in standard practice. The public health goals of the "treatment discovery process" are to positively modify clinical practice by identifying and adopting as many truly effective treatments as possible while limiting the number of ineffective treatments adopted, conserving resources, limiting exposure of study patients to ineffective or harmful interventions, and providing patients with beneficial treatments as soon as possible.

There has been much discussion in the clinical trials literature about the role of Phase II studies in the treatment discovery process. Holmgren (2007) expressed concerns about the use of screening trials in oncology research for determining whether further study of a new treatment in confirmatory trials is warranted. In particular, the author proposed that the improvement in efficiency in using Phase II trials as a screen for efficacy or as a guide for dose selection is dependent upon the magnitude of the treatment effect. Some authors have discussed what design parameters may be appropriate for Phase II screening studies when the goal is to screen out useless treatments and identify useful ones (Rubenstein, Korn, Freidlin, Hunsberger, 2005). Interest in creating a more time- and cost-efficient treatment discovery process has led to developments in adaptive design where modifications are made to the trial design based on data accrued to the point of observation. It is suggested that adaptive Phase II/III trials have potential gains in time and reduced resources in terms of the numbers of patients needed for finding effective treatments (Orloff, Douglas, Pinheiro, Levinson, 2009; Korn, Freidlin, Abrams, Halabi, 2012). We find that many approaches proposed in the literature are focused on the use of Phase II trials for efficiency gains in specific settings. There still exists a need for research that establishes a comprehensive Phase II to Phase III approach incorporating the concepts of set resources and competing interests, and addressing logistical issues inherent in adaptive design. In section 1.4, we discuss a more general strategy for improving the

treatment discovery process such that the objectives of a particular setting govern the choice of clinical trial design.

1.2 Current phasing of clinical trials

In order for an effective treatment to be adopted, we must ultimately confirm some positive result previously observed for that treatment. This need for exploratory science during which we test new ideas for treatments or strategies targeted toward improving outcomes common to particular diseases and patient populations creates a natural progression of investigations through several phases. We progress from pre-clinical studies to the clinical phases involving human experimentation: Phase I, Phase II, and Phase III clinical trials. Phase I trials are studies of initial drug safety and pharmacology with primary interest in the incidence of major adverse events and the safety and ethics of further investigation. Phase II trials serve in part as screening studies providing a preliminary evaluation of efficacy and further evaluation of the safety of the treatment. In some sense, a major role of Phase II studies is to identify those treatments that are not of interest to study further. Treatments that "pass" Phase II, then progress to Phase III. In Phase III, we conduct large randomized clinical trials with the goal being to reliably confirm whether the treatment is effective and safe when compared to standard of care.

However, how we choose to advance from one phase to another can vary greatly with respect to the timing of progression through phases of the investigation, the design of each phase, and the decisions of how to proceed when preliminary results differ from presumed effects. The progression from Phase II studies to confirmatory Phase III trials is an integral part of the treatment discovery process and a part that has received much attention in the clinical trials literature (see for instance Holmgren, 2007 and Korn, Freidlin, Abrams, Halabi, 2012). Reliable evaluation of the effectiveness of a new treatment involves well-controlled randomized clinical trials with a clearly defined treatment strategy, a clearly defined and ascertained clinical outcome, and a pre-specified primary analysis. In using Phase II studies as screening trials for Phase III, we investigate new treatment ideas with smaller trials and only proceed to larger, confirmatory trials with promising treatments. We also utilize what we learned about the treatment in Phase II when conducting the Phase III study. The results of these investigations depend not only on how the trials are designed and conducted at each phase individually, but also in relation to one another.

The scientific challenges to addressing these goals include the evaluation of promising interventions with clinical trials that are efficient, credible, and ethical. Efficient trials satisfy the economic interests of the sponsor by minimizing the resources and time needed to gain regulatory approval and to adopt treatments into clinical practice when results reflect a truly important benefit. Furthermore, efficient trials conducted on an industry-wide basis promote the public health interests of

rapid discovery of as many effective therapies as possible within the constraints of limited resources (patients, time, money). Credible trials discriminate between relevant hypotheses and provide convincing evidence for or against further study or adoption of the proposed intervention. Ethical trials are conducted in the best interest of the patients, and consider both the individual ethics related to participants enrolled in the trial and group ethics related to future recipients in the wider population of diseased patients who would be administered an effective therapy.

These challenges can be addressed statistically in the design and conduct of clinical trials. As scientific experiments, clinical trials must carefully specify both the scientific hypothesis and the precise statistical hypothesis to be addressed. The statistical hypothesis is typically specified in terms of distributional summary measures (mean, proportion, hazard) that will be used to quantify the effect of an intervention. During the clinical trial, data is collected that will be used to provide estimates of the treatment effect and to quantify the precision with which those estimates could be used to draw inference about the true treatment effect in some target population. In terms of statistical methods and inference, there are two main paradigms that are employed: frequentist and Bayesian. Regardless of which methods are used, it is the strength of the statistical evidence gathered and the precision of our estimate of the treatment effect that guides our ultimate decision for or against the adoption of that intervention into clinical practice.

1.3 Current practice of statistical inference

1.3.1 Setting and notation of the treatment discovery process

Consider the following setting and notation for an asthma trial designed to compare forced expiratory volume (FEV) as a measure of lung function between patients who received an experimental treatment and patients who received placebo. Notationally, we denote potential observations as X_{ki}, the FEV for patient i on treatment k where $k = 0$ for placebo and $k = 1$ for treatment. The random variables, X_{ki} for $i = 1, 2, ...$ are independently distributed with means μ_0 and μ_1 for placebo and treatment respectively, and a common known variance, σ^2. Assuming N subjects are accrued on each treatment arm, we assume the sample mean, $\bar{X}_k$ is approximately normally distributed with mean μ_k and variance σ^2/N. The primary outcome of interest is the difference in mean FEV, $\theta = \mu_1 - \mu_0$ as a measure of treatment effect. Similar methods apply to other statistical models including the comparisons of geometric means, medians, proportions, or constant hazard ratios, as well as when the randomization ratio differs from 1:1 randomization. From the trial data gathered, we estimate this treatment effect as $\hat{\theta} = \bar{X}_1 - \bar{X}_0$. Because a higher FEV reflects better lung function, we would design the trial to test the hypothesis of a greater alternative in mean response. The test would discriminate between the following null and alternative hypotheses

$$H_0 : \ \theta \leq 0$$
$$H_1 : \ \theta \geq \theta_1$$

where $\theta_1 > 0$ is the alternative representing a minimal clinically meaningful difference (MCID) (improvement) in mean response. The data gathered during a trial would be used to compute a point estimate, $\hat{\theta}$ of θ, an interval estimate, (θ_L, θ_U), a range of possible values for which the data might reasonably be observed, and a probability quantifying the evidence for or against our hypothesis (e.g., a frequentist p-value or Bayesian posterior probability).

This notation for a simple trial design could correspond to a study of any phase of the investigation, and under current practices the design of each clinical trial primarily considers each phase independently. We are interested in describing a more general strategy. We suppose that there exists some population of candidate treatments that will be first investigated in a Phase II screening trial with Phase III confirmatory trials conducted only for promising treatments passing early tests. We will first focus on identifying an appropriate Phase II to Phase III progression in the absence of modifications to the study design.

In considering the treatment discovery process, we find it useful to expand our notation slightly. The above notation presumes that we have in mind a specific experimental treatment that has some fixed (but unknown) treatment effect, θ. We find it useful, however, to model the treatment discovery process in which our population of candidate treatments is screened for those that are most promising and worthy of further study. Hence, we can describe a population of θ's across the population of candidate treatments. For our treatment discovery process, then, we can presume a joint distribution, $p(\theta, X)$ of θ (the treatment effect for some treatment selected for Phase II study) and X (the data that might be collected in a Phase II study). The marginal distribution $p(\theta)$ takes on the appearance of a Bayesian prior as will be discussed in section 1.3.3. We note that initially we will consider a marginal distribution in which we examine two hypotheses: the prevalence of an effective therapy at the MCID of θ_1 is π, and θ is 0 (no effect) with probability $1 - \pi$. In later chapters we will relax this assumption by considering a continuous distribution for $p(\theta)$. The average success rate across all therapeutic areas from human experimentation to registration is approximately 11% (Kola, Landis, 2004). This means that one in nine treatments makes it through development and gains approval from European and/or US regulatory agencies. We will consider such success rates when identifying a treatment discovery strategy for a particular setting.

We consider a treatment discovery process in which we use a fixed amount of resources. In particular, we consider a fixed number of patients available for the process by which some number of effective treatments will be identified. We screen K_2 candidate treatments in Phase II, each investigated with a sample size of N_2 subjects. We then conduct K_3 confirmatory Phase III studies,

and initially we suppose that every Phase III study will involve N_3 subjects. The number of Phase III trials to be conducted will depend on how many treatments pass through Phase II. We define β as the proportion of effective therapies and α as the proportion of ineffective therapies that pass a particular phase of the investigation. These parameters are defined for Phase II and Phase III trials as (β_2, α_2) and (β_3, α_3), respectively. Hence, $K_3 = K_2(\pi\beta_2 + (1-\pi)\alpha_2)$. In this notation we can define several quantities among those that are of interest when describing the operating characteristics of a particular treatment discovery strategy, which include:

- The total number of patients used is $N_2K_2 + N_3K_3$. Without loss of generality, we constrain this total to be 1,000,000.
- The overall sample size ($N_2 + N_3$) is the total sample size for each treatment studied in both Phase II and Phase III. This reflects the amount of safety information gathered for each adopted treatment.
- The timing at which we conduct Phase II trials relative to Phase III studies is defined as $N_2/(N_2 + N_3)$.
- The number of effective treatments adopted with fixed resources is $m_{\text{eff}} = K_2\pi\beta_2\beta_3$ and the number of ineffective treatments adopted, $m_{\text{ineff}} = K_2\alpha_3\alpha_2(1-\pi)$.
- The average cost per effective treatment identified and adopted will be defined as the total number of subjects in both Phase II and Phase III divided by the number of effective treatments adopted, $(K_2N_2 + K_3N_3)/(K_2\pi\beta_2\beta_3)$.
- We can further consider an additional "cost" as the number of ineffective therapies adopted at that same time, which we quantify as the positive predictive value (PPV)

$$\text{PPV} = \frac{\beta\pi}{\beta\pi + \alpha(1-\pi)} \tag{1.1}$$

  and discuss further in section 1.4.3.

We note that several of these operating characteristics are clearly based in either frequentist or Bayesian inference paradigms, and we provide further detail on these criteria in the next sections describing typical current practices.

1.3.2 Frequentist inference

In the frequentist setting, we specify a hypothesis about the treatment effect under study. We design and conduct a clinical trial during which data are gathered, and determine the probability of obtaining the observed data under a specified null hypothesis and "design alternative", $\Pr(X|\theta)$. Frequentists address whether we would expect to observe the data gathered if our hypothesis of the

treatment effect were true. The frequentist probability model considers the sampling distribution of the statistic across conceptual replications of the experiment,

$$p(X|\theta) = \frac{p(X,\theta)}{\int p(X,\theta)dX} = \frac{p(X,\theta)}{p(\theta)} \tag{1.2}$$

for a particular value of θ. Operating characteristics are chosen to appropriately discriminate between the null and alternative hypotheses of interest. Most commonly considered are the type I error (the probability of a false-positive), power or one minus the type II error (the probability of a true-positive), and the sample size of the clinical trial required to detect the desired treatment effect. To ensure sufficient precision and trial credibility, frequentist designs specify the type I error probability, α and the power, β as

$$\begin{array}{lll} \Pr(\text{Accept new treatment} \mid H_0) & \leq & \alpha \\ \Pr(\text{Reject new treatment} \mid H_1) & \leq & 1-\beta \end{array}$$

We typically fix the α-level to some suitably low value, find an alternative, θ_1 that achieves that type I error, and compute a power curve. Typical of trials designed in the frequentist paradigm are α-level 0.025 (two-sided 0.05) tests with 80-90% power. Phase II and Phase III trials are generally considered separately and treatments are progressed to Phase III studies after "promising" Phase II results. In our notation described above (α_2, β_2) and (α_3, β_3) represent the type I error and power for Phase II and Phase III trials, respectively. The sample size needed to detect a particular alternative with statistical power, β using a level α trial design can be computed from the standard formula

$$N = \frac{(z_{1-\alpha} + z_\beta)^2 V}{{\theta_1}^2} \tag{1.3}$$

where $z_p = \Phi^{-1}(p)$ is the p-th quantile of a standard normal distribution with cumulative distribution function $\Phi(z)$ and V is the variance contributed by a single sampling unit.

Typical summary measures reported are point estimates (mean, median, mode) and measures of statistical evidence include confidence interval estimates (hypotheses for which the data might reasonably be observed), and the p-value (the probability such extreme data would have been observed under the null hypothesis). We make the decision to reject the null hypothesis when the p-value is below some threshold α or the confidence interval does not include the null hypothesis. For selecting estimators of treatment effect, frequentist optimality criteria such as bias and mean square error are typically used.

The overall probability of obtaining a significant result depends on the prior probability, π of studying a truly beneficial treatment, on the specificity (fixed by α-level of significance), and on the

sensitivity (statistical power chosen for trial design). For making inference, we consider the critical value at which we can reject the null hypothesis (i.e., reject the null hypothesis if $\hat{\theta} \geq \hat{\theta}_c$) and the confidence interval containing the set of hypothesized treatment effects which might reasonably generate data like those observed. For trial designs monitored with group sequential stopping rules, the same operating characteristics are considered in addition to the stopping boundaries, the probability of early stopping at each interim analysis as a function of the true treatment effect, statistical inference to be made at each of the stopping times, and the futility associated with the potential continuation of the trial after each analysis (Emerson, Kittelson, Gillen, 2007a).

1.3.3 Bayesian inference

Although clinical trials are most commonly designed and analyzed in the frequentist paradigm, adherents of Bayesian inference note that frequentist inference fails to address the question of greatest interest: after observing the data, what is the probability that the treatment is truly beneficial? In the Bayesian paradigm, we can address this question by assuming a prior knowledge about the treatment effect parameter and update that distribution using the data we observe (Emerson, Kittelson, Gillen 2007b).

Recall, the joint distribution, $p(\theta, X)$ for the treatment effect parameter, θ and the clinical trial data, X. Our prior distribution, $p(\theta)$ for the treatment effect parameter is representative of the beliefs about the treatment effect in absence of any information on the value of X. The likelihood function for the data is $p(X|\theta)$. The posterior distribution is a conditional distribution for the treatment effect given the data from which we base inference that can be computed using Bayes' Theorem as

$$p(\theta|X = x) = \frac{p(X|\theta)\, p(\theta)}{\int p(X|\theta)\, p(\theta)\, d\theta} \tag{1.4}$$

We are considering the probability distribution for the parameter measuring treatment effect and updating our prior belief about the treatment using information obtained in the clinical trial. The prior distribution might reflect subjective uncertainty at an individual level or a consensus or population average of individual priors. Bayesian inference depends heavily on the choice of prior distribution for the treatment effect parameter and it can similarly be dependent upon any parametric assumptions for the distribution of the data. To relax this latter dependency, we adopt the coarsened Bayesian analysis described in Emerson, Kittelson and Gillen (2007b). In that approach, we condition on $\hat{\theta}_N$, a relatively distribution-free estimate of θ instead of conditioning on the full data. We use approximate normal distributions for our sampling distribution, $p(\hat{\theta}_N|\theta)$.

The point estimates reported are summary measures of the posterior distribution (mean, median, mode), credible interval estimates (e.g., the central 95% of the posterior distribution), posterior probabilities (the probability of a particular hypothesis conditional on the data), and predictive probabilities (the probability of a significant result conditional on the data). The posterior probability can be computed, for some quantile θ_0, as

$$\int_{\theta_0}^{\infty} \hat{p}(\theta|\hat{\theta}_N)d\theta \tag{1.5}$$

For the asthma trial, we might be interested in the posterior probability of the null hypothesis $\Pr(\theta \leq 0|\hat{\theta}_N)$ or the posterior probability of the design alternative $\Pr(\theta \geq \theta_1|\hat{\theta}_N)$. We make the decision to reject the null hypothesis when the posterior probability is low.

The positive predictive value (PPV) is just the posterior probability after seeing $X > c$, where c might be the frequentist critical value $\hat{\theta}_c$.

$$\text{PPV} = \Pr(H_1|X > c) = \frac{\Pr(X|H_1)\Pr(H_1)}{\Pr(X|H_1)\Pr(H_1) + \Pr(X|H_0)\Pr(H_0)} \tag{1.6}$$

With our simple hypothesis, Bayes' Rule is a function of π, α, and β. We can also think about this as prior odds times Bayes factor (BF) equals posterior odds, where Bayes factor is defined as

$$\text{BF} = \frac{\text{PPV}}{1 - \text{PPV}} = \frac{\Pr(X|H_1)\Pr(H_1)}{\Pr(X|H_0)(1 - \Pr(H_1))}, \tag{1.7}$$

equal to the power divided by the type I error. (With continuous priors, the Bayes factor is more complicated and depends on the prior.) In any case, this argues that we can parameterize the Bayesian problem using frequentist criteria. If we do not change anything between Phase II and Phase III, then the posterior from Phase II is the prior to Phase III, and we can write everything as a single process using prior to Phase II and obtaining posterior from Phase III.

1.3.4 Frequentist vs Bayesian

In frequentist analysis of a clinical trial, we gather information about the probability of obtaining the observed data under a specified hypothesis. In Bayesian inference, we are interested in the probability that the treatment is truly beneficial after observing the data. In our development, we note that

- Both paradigms are complementary ways of quantifying the strength of the evidence gathered to convince the scientific community for or against the adoption of a treatment into clinical practice.
- Both paradigms consider conditional distributions derived from the same joint distribution

of the parameter, θ and the estimate of that parameter, $\hat{\theta}$.

- The Bayesian paradigm adheres to the likelihood principle.
- Frequentist criteria can result in tests designed to discriminate between relevant hypotheses such that results are not equally plausible.
- In experimental design, a frequentist can design a test such that the same data will not tend to be obtained under two competing sets of hypotheses.

1.4 Optimality criteria

The optimal treatment discovery process involves identifying an approach that best addresses the often competing goals of science, ethics, and efficiency. Clinical scientists have an interest in the improvement of clinical practice and of the health of the population. Regulatory agencies are concerned with the credibility that a treatment works and is safe to approve for public use. Trial sponsors have an economic interest in the cost and the efficient use of time and resources. Patients on study, patients who might be on study, and those who might benefit from new knowledge about the intervention under study are interested in finding out the truth about treatments specific to their diagnoses. The goal of our research is to consider a formal approach that is comprehensive with respect to both the frequentist and Bayesian paradigms and the goals of the treatment discovery process as a whole. We discuss what other researchers have proposed as an approach that considers both Phase II and Phase III trials.

Rubenstein, Korn, Freidlin, et al. (2005) present the use of Phase II studies for prioritizing new treatments for Phase III evaluations in oncology trials. In particular, the authors focus on the use of a non-definitive randomized comparison of treatment to standard care in Phase II to strengthen the evidence for moving forward with Phase III. In order to target an appropriate treatment effect while restricting the Phase II sample size to something attainable, they consider varying the size and power of Phase II. They recognize the conflicting demands of such trial design parameters and suggest $\alpha = \beta = 0.2$ and $\Delta = 1.5$ as a targeted effect for median progression-free survival (PFS) time in cancer screening trials. Focus is on the design of Phase II in preparation for Phase III and not on the stages collectively. The authors conclude that Phase II screening studies are useful when the importance of a definitive study is recognized and when the design parameters of Phase II lead to an appropriate balance of effective and ineffective treatments identified.

Holmgren (2007) expressed concerns about the efficiency of Phase II studies in oncology with respect to cost and the rate at which active treatments are identified with a positive Phase III trial. His research focuses on optimizing the treatment development process in terms of what might be

considered sponsor goals of overall power and cost (the number of events observed). He concludes that a Phase II/III approach reduces overall power when compared to a single Phase III study and is only more efficient for hazard ratios less than 0.85. When considering the cost as the number of events, the Phase II/III approach is only more efficient over a single Phase III when the costs of bringing a treatment to the point of Phase II/III are not considered. The objectives appear to be the identification of an single effective treatment at low cost while the overall goals of the treatment discovery process are not addressed.

Recent interest in making the treatment discovery process more time- and cost-efficient has led to a focus on adaptive designs. Stallard and Todd (2010) propose a "seamless" Phase II/III approach based on group sequential methods to allow for treatment and dose selection and comparative evaluation of efficacy in the same study while controlling the family-wise error rate (FWER). Other authors have also recognized the interpretability and statistical credibility issues inherent in adaptive designs and have proposed methods for protection (Korn, Freidlin, Abrams, Halabi, 2012). Particularly important is control of the type I error when combining Phase II and Phase III data in confirmatory evaluations of efficacy and safety. There are inconsistencies in the literature in how this error rate is controlled when phases of the investigation are combined.

There still exists a need for research that considers the objective to improve the health of the population when determining how to design Phase II and Phase III trials. In this research, we discuss many of these same objectives and issues collectively when considering the treatment discovery process as a whole.

1.4.1 Optimal Bayes

In order to account for any prior knowledge we have about a candidate treatment as well as information gathered in our experiments, we can consider the Bayesian probability-theoretical framework. Prior to the design, we first consider the purpose and goals of the experiment. Under a Bayesian experimental framework this is known as the utility of the design. Once a utility is specified, we can consider the choice of study design a decision problem and select the design that maximizes the selected utility (Chaloner, 1995).

The purpose of a clinical investigation of a candidate treatment is to determine if that treatment is beneficial and worth adopting into clinical practice. The optimal Bayesian experimental design for identifying new beneficial treatments maximizes the number of truly effective treatments adopted, $m_{\text{eff}} = K_2\pi\beta_2\beta_3$ for an acceptable number of ineffective treatments adopted, $m_{\text{ineff}} = K_2\alpha_3\alpha_2(1-\pi)$ with a fixed sample size for conducting both Phase II and Phase III studies. As previously discussed, this additional "cost" of adopting ineffective therapies is quantified as the positive predictive

value or the probability that the treatments being adopted are beneficial. The optimal Bayesian approach can be defined on either the posterior probability or predictive probability scale.

Although there is strength in numbers, the trade-off between the number of effective and ineffective treatments adopted is appropriately assessed within a particular setting. If we are considering several interventions for the treatment of a single disease, the adoption of ineffective treatments for a greater gain in the number of effective treatments adopted might be considered appropriate. Alternatively, if our setting includes treatments for a number of different diseases then we might be more cautious in adopting an ineffective therapy that may or may not be dismissed or even studied in future trials. We will also want to consider whether we are studying new interventions that will be used to treat a disease for which effective treatments already exist or if the therapies under study are for treating diseases for which beneficial treatments are not currently on the market. If effective treatments exist, are readily available and used in clinical practice, introducing ineffective treatments as a result of adopting additional effective interventions does not seem worthwhile. Although these considerations must be made for each setting specifically, we will define a general strategy for designing studies of particular objectives.

In testing new treatments through a series of experiments, we can consider Phase II trials as screening studies for Phase III trials where the goal is to find a procedure that identifies truly beneficial interventions. A useful application of Bayes' Theorem in the clinical setting is that of investigating the sensitivity and specificity and positive and negative predictive value for screening or diagnostic tests. Applying these same principles to the treatment discovery process, the sensitivity of a clinical trial is defined as the probability of a significant result given that the treatment is truly effective. This is the power or true positive rate. The specificity of a clinical trial is the probability of a non-significant result given that the treatment is truly ineffective. This is the true negative rate, or 1 minus the type I error. Ideal results would include a low probability of adopting ineffective treatments (high specificity, low type I error) and a high probability of adopting truly effective treatments (high sensitivity, low type II error, high power). Important to clinicians and patients though is a high probability that the treatments being adopted are beneficial (high positive predictive value). However, compromise is the best we can do when tradeoffs exist among the various study operating characteristics. We will discuss different sets of optimality criteria that might be considered by various collaborators during the treatment discovery process.

1.4.2 Type I error and power

Regulatory approval is contingent upon scientifically meaningful and statistically credible results ensuring the safety of the public. For confirmatory Phase III trials, regulatory requirements result in testing one-sided hypotheses at a level of significance of $\alpha_3 = 0.025$ and high power, β_3. However,

choice of operating characteristics for Phase II is generally more flexible allowing for the optimal study design to be defined by such parameters.

Inflation of the overall type I error affects trial credibility and results in the improper investment of time and resources in treatments that are not truly beneficial. In our notation, the overall type I error is $\alpha_2\alpha_3 = m_{\text{ineff}}/(K_2(1-\pi))$. Increasing the chances of discovering an association is often introduced by multiple comparisons. In interventional experiments, performing many exploratory analyses where analysis methods are modified, multiple endpoints are considered, or subgroups are a focus may result in many apparent associations some of which are simply false positives. To avoid being mislead by spurious associations, we can choose to optimize our study design based on the overall type I error taking into account both Phase II and Phase III studies, minimizing the number of ineffective treatments adopted (a statistically-signicant test in absence of a true treatment effect). It is important to consider such an approach in settings where treatments are costly or the adoption of ineffective treatments will deter patients from truly beneficial interventions. The overall or experiment-wise type I error can be controlled by well-conducted experiments and pre-specified analyses. In this case, the optimal study design would have the lowest overall type I error among all study designs of similar outcomes.

We can alternatively consider approaches based on the overall power, optimizing our study design by minimizing the number of effective treatments that are not adopted (a non-statistically-signicant test in presence of a true treatment effect). In our notation, the overall power is $\beta_2\beta_3$. With efficient study designs we are ensuring that trial results will correspond to a rejection of the null hypothesis when the design alternative is true. In the interest of the study sponsor who necessarily believes that the treatment under study works, with high power we are increasing our chance of regulatory approval and adoption of the new treatment when treatment effect estimates indicate improvements in outcome. In this setting, the optimal study design would have the highest overall power among all study designs of similar outcomes.

In the frequentist setting, we want a low type I error, α and high power, β. To achieve this, we often choose an appropriately fixed α-level and optimize β. The Neyman-Pearson paradigm suggests an appropriate method for choosing a test statistic; in fixing α when performing a test between two hypotheses, the Likelihood Ratio statistic is the uniformly most powerful test for size α. However, this does not prevent us from fixing the power when optimizing the type I error. It is appropriate to consider both parameters simultaneously and the tradeoffs that exist when optimizing the study design.

1.4.3 Positive and negative predictive value

The positive predictive value (PPV) is defined as the probability that a statistically significant trial indicates a truly beneficial treatment. With Bayes' rule, we can parameterize our posterior probability by the type I error, α and power, β at Phase II and Phase III and the prevalence of truly beneficial treatments among all treatments being studied, π. In our notation, the PPV for the entire strategy is

$$\text{PPV} = \frac{\beta_3\beta_2\pi}{\beta_3\beta_2\pi + \alpha_3\alpha_2(1-\pi)} \tag{1.8}$$

We can choose to optimize our study design based on a high PPV achievable by increasing π, increasing $\beta = \beta_3\beta_2$, or reducing $\alpha = \alpha_3\alpha_2$. Again, we can think of this in terms of Bayes Factor (BF) as the ratio

$$\frac{\text{PPV}}{1-\text{PPV}} = \frac{\Pr(X|H_1)\Pr(H_1)}{\Pr(X|H_0)(1-\Pr(H_1))} = \frac{\beta_3\beta_2}{\alpha_2\alpha_3}\frac{\pi}{(1-\pi)} \tag{1.9}$$

In entering Phase II studies, increasing the prevalence of good treatments, π requires careful and thoughtful consideration of which treatments to study. Appropriate scientific questions to be addressed must be specified and supported by prior hypothesis-driven research. Early screening trials dismiss a great majority of ineffective treatments increasing the PPV and resulting in a greater prevalence of beneficial treatments moving forward to Phase III trials. High power, β is a result of successfully implemented clinical trials (minimize missing data, low variation in outcome assessment, adherence) and increased sample size. A low type I error, α can be achieved by pre-specifying all analyses and endpoints and avoiding multiple comparisons, surrogate endpoints, and subgroup analyses.

With the goal being to achieve a high PPV, we should keep in mind that even in just identifying a single effective treatment we can have a PPV of 1, which does not greatly improve clinical practice. We must consider the trade off between a high PPV and adoption of a large number of effective treatments. We can consider how many ineffective treatments, $K_2\alpha_3\alpha_2(1-\pi)$ will result from an acceptable or ideal number of effective treatments adopted, $K_2\pi\beta_2\beta_3$ with high PPV. The optimal study design would have the highest PPV among all study designs of similar outcomes.

The negative predictive value (NPV) is defined as the probability that the treatments not being adopted are truly ineffective. In our notation, the NPV for the entire strategy is

$$NPV = \frac{(1-\alpha_3\alpha_2)(1-\pi)}{(1-\alpha_3\alpha_2)(1-\pi) + (1-\beta_3\beta_2)\pi} \tag{1.10}$$

Because we want to avoid the adoption of ineffective treatments, we want this probability to be

high. In considering the number of effective treatments we want to eventually adopt, we must realize that the cost is the adoption of ineffective treatments, justification for which is setting-specific. However, in using appropriate Phase II trials as screening studies for Phase III, we eliminate after Phase II a great majority of the ineffective treatments initially studied. Therefore, the NPV is hopefully high thereby lessening concern for the adoption of a great number of ineffective therapies into clinical practice.

When conducting experiments, an optimal design is one that is superior to other study designs with respect to some statistical criteria. We often think of optimal experimental designs as those under which bias and variance are minimized, precision and efficiency are maximized, and cost is reduced. To achieve our goal of improving healthcare by positively altering clinical practice, we can consider a variety of study design parameters when optimizing our approach in transitioning from Phase II to Phase III. The parameters that we find of greatest value in optimizing this approach are the PPV, overall power, type I error, and the absolute number of effective and ineffective treatments adopted.

Chapter 2

Comprehensive Strategy for the Phase II to Phase III Transition

Our public health objective is to adopt as many truly effective treatments as possible while avoiding the adoption of ineffective or harmful treatments. This can be achieved by maximizing the number of effective treatments adopted conditional on a high PPV after Phase III. To illustrate this search of an optimal study design, we will first consider an introductory example of two approaches: conducting large confirmatory studies for every treatment or performing small screening trials for a number of candidate treatments and confirmatory trials only for promising treatments passing early tests. This will give an understanding of the usefulness of Phase II screening studies for increasing the efficiency of the treatment discovery process.

2.1 Introductory example

The prevalence of effective treatments, π is defined as the proportion of truly effective treatments that exist among the population of candidate treatments under study. This prevalence depends on our disease setting and on the support of our scientific questions by prior hypothesis-driven research. We will consider that 10% of the treatments being studied are truly effective. For confirmatory Phase III studies, regulatory requirements for one-sided tests of superiority typically result in the use of a level of significance of 0.025 and high power. The same parameters for Phase II screening studies can be varied as a means for obtaining high specificity and sensitivity increasing the PPV. We will consider how altering these parameters used in specifying our Phase II and Phase III study designs changes our outcome in terms of the number of treatments (effective and ineffective) adopted and the overall PPV.

We note that this problem is scalable in the sense that our primary parameters of type I error,

power, PPV, and NPV, are all rates. Hence, when we describe the number of adopted effective treatments (true positives; TP) and ineffective treatments (false positives; FP) based on a fixed number (in our case, 1,000,000) of clinical trial patients for a specified MCID, altering the number of patients and/or the MCID will each have a proportionate effect on our results. For ease of presentation, we chose a setting in which a fixed sample level 0.025 test having 80% power to detect the MCID in a two-sample clinical trial would require 500 subjects. In the case of the asthma example previously mentioned for studying FEV as a measure of lung function, we might be interested in the difference in mean FEV after one second (FEV_1). For the purpose of this example, this corresponds to an MCID of 0.125 liters and standard deviation of 0.5. Using these numbers in the sample size formula (Equation 1.3) results in 500 as desired.

Outlined below in Table 2.1 are outcomes for the approaches previously described. Considering the approach of only conducting large Phase III confirmatory trials for all candidate treatments, we can make the following statements:

- When using a fixed sample level 0.025 one-sided test to detect higher FEV_1 with the candidate treatment, a sample size of 500 subjects provides 80% power to detect the MCID of 0.125 L.
- We can study 2,000 new treatments with 2,000 randomized clinical trials (RCTs) at Phase III. With a prevalence of 10%, there are 200 truly beneficial treatments among those being studied.
- On average, we can expect a significant result for 160 effective treatments (true positives; TP) and 45 ineffective treatments (false positives; FP). These outcomes result in a PPV of 78.0% meaning that 78.0% of the treatments identified as effective actually work.

The second approach considers only Phase III trials for treatments passing Phase II studies where Approach 2a includes a greater number of small screening studies and Approach 2b includes a smaller number of screening studies with a larger sample size. To assess the effect of screening trials on our results, we can first consider using the strategy in Approach 2a and can make the following statements:

- When using a fixed sample level 0.025 one-sided test in Phase II to detect higher FEV_1 with the candidate treatment, a sample size of 100 subjects will provide 23.9% power to detect the MCID of 0.125.
- We begin by studying 7,000 new treatments with 7,000 Phase II trials where 10% or 700 of the treatments under study are truly effective. On average, we will see a positive test for 168 effective treatments and 158 ineffective treatments.

Table 2.1: Examples of various approaches to identifying treatments with Phase II and/or Phase III studies. Approach 1 consists of only performing Phase III studies, while Approaches 2a and 2b consider two different strategies of first conducting Phase II studies. (TP = True positives, FP = False Positives)

		Approach 1	Approach 2a	Approach 2b
Phase II	Number RCT	2,000 (10% eff)	7,000 (10% eff)	2,047 (10% eff)
	Truly Beneficial	200	700	205
	N per RCT	0	100	342
	Type I err; Pwr	–	0.025; 23.9%	0.100; 84.9%
	"Positive" RCT	–	168 eff; 158 not	174 eff; 184 not
Phase III	Number RCT	2,000 (10% eff)	326 (52% eff)	358 (49% eff)
	N per RCT	500	921	829
	Type I err; Pwr	0.025; 80.0%	0.025; 96.7%	0.025; 95.0%
	# Effctve Adopt (TP)	160	162	165
	# Ineff Adopt (FP)	45	4	5
Overall	PPV	78.0%	97.6%	97.1%
	Type I err; Pwr	0.0250; 80.0%	0.0006; 23.1%	0.0025; 80.5%

- In Phase II, we have thus used 700,000 patients, leaving 300,000 patients to study the 326 treatments that progress to Phase III. This allows for 921 subjects per Phase III study.
- When using a fixed sample level 0.025 one-sided test in Phase III to detect higher FEV_1 with the candidate treatment, a sample size of 921 subjects will provide 96.7% power to detect the MCID of 0.125. In moving forward to Phase III with all 326 treatments, 52% of which are truly effective, we can expect to identify 162 effective treatments (true positives; TP) and 4 ineffective treatments (false positives; FP).
- This greatly improves our PPV (97.6% vs 78.0%) and we are identifying about the same number of effective treatments while reducing the number of ineffective treatments passed more than ten-fold. Using the same number of subjects, Phase II studies increase the predictive value of a positive study.

However, we can even further modify our trial designs in terms of the study parameters considered. Using the same number of subjects, we can instead choose to conduct fewer Phase II trials each of a larger sample size as is used in Approach 2b. Using this strategy, we can make the following statements:

- When using a fixed sample level 0.100 one-sided test in Phase II to detect higher FEV_1 with

the candidate treatment, a sample size of 342 subjects will provide 84.9% power to detect the MCID of 0.125.

- We can study 2,047 new treatments 205 of which are truly effective. We can expect 173 effective treatments and 184 ineffective treatments to be identified after Phase II on average.
- Because our resources are restricted to 1,000,000 patients, the 2,047 Phase II studies of 342 subjects each (700,074 subjects total) leaves us with approximately 300,000 subjects for conducting Phase III studies on the 358 treatments that passed Phase II. Therefore, our Phase III studies will have a sample size of 829 subjects each.
- When using a fixed sample level 0.025 one-sided test in Phase III to detect higher FEV_1 with the candidate treatment, a sample size of 829 subjects will provide 95.0% power to detect the MCID of 0.125. When moving forward to Phase III with all 358 treatments of which roughly half are truly beneficial, we would adopt 165 effective treatments (true positives; TP) and 5 ineffective treatments (false positives; FP), results similar to those found in approach 2a, although the PPV is reduced slightly.

While Approach 2b will be used for comparison with other considerations in later chapters, the parameters such as number of treatments to study, the timing of Phase II relative to Phase III, and the statistical power and level of significance used in both screening and confirmatory trials, can all be used to optimize strategies for a particular setting or objective. This is explored in the remaining sections of this chapter.

2.2 Specifying constraints

We now generalize that introductory example to explore a space of possible clinical trial design strategies. To conduct a search for the optimal strategy in progressing from Phase II to Phase III, it is necessary that some constraints be made for addressing the multiple goals of the treatment discovery process. Suppose the following criteria/constraints define the study designs under consideration:

- The desired type I error at Phase III is $\alpha_3 = 0.025$ and overall power is $(\beta_2\beta_3) \geq 0.8$ at a fixed design alternative, $\theta = \Delta$ for both Phase II and Phase III studies. Initially considered are fixed sample size designs with N_2 and N_3 subjects required to meet such operating characteristics.
- The prevalence of truly effective treatments will be fixed at 5, 10, 20 or 50% and the minimal positive predictive value (PPV) will be fixed at 95 or 98%.
- The number of treatments under study, K_2 will define the number of Phase II clinical trials conducted, where K_2 will be between 500 and 2,000 in increments of 100. The type I error

at Phase II, α_2, will be varied between 0.025 and 0.9 in increments of 0.005 to be exhaustive and handle cases with high prevalence.

The primary objective is to find optimal designs meeting these constraints. Optimality is measured by the maximum number of effective treatments passed for what is considered an acceptable PPV after Phase III. This can be achieved by identifying and adopting as many truly effective treatments or true positives (TP) as possible while limiting the number of ineffective treatments or false positives (FP) adopted.

2.3 Comparing designs

Owing to the many design parameters that can be varied, we consider a grid search for the preferred approach over a continuous set of possible approaches. The comparability of different approaches is best among those that achieve similar outcomes of greatest value to the clinical investigators. We make such comparisons within the constraints imposed by the level of granularity specified for the search. We choose to focus attention to designs with the same m_{eff}, and m_{ineff} to highlight ways in which such designs may differ. More specifically, we choose to examine, among all designs with the same PPV, those with different design parameters but with the same number of effective (TP) and ineffective (FP) treatments passed after Phase III. We are interested in determining the parameter values for designs with similar outcomes. This will enable us to examine how we might choose between designs, as well as to demonstrate the parameters that result in a different number of adopted treatments. This in turn will allow us to better our goal of maximizing the number of effective treatments adopted. Therefore, it was necessary to consider a range of adopted treatments for making these comparisons. It should be noted, however, that the chosen examples presented in Tables 2.2 and 2.3 are arbitrary having merely been chosen to be a representative sample of what might be targeted in a general setting.

2.3.1 Thresholds in terms of Phase II errors

Given the above constraints, we first consider a setting within which the prevalence of truly effective treatments among all treatments under study is 10%, and we examine a collection of designs within that setting that result in a PPV of 95% after Phase III.

Consider the design with which we adopt 124 effective and 6 ineffective treatments (Table 2.2). This is attainable by starting with Phase II trials testing 1,300, 1,400, or 1,500 different candidate treatments. Similarities in these three approaches include

- the number of treatments passed (107 effective and 205 ineffective) after Phase II,
- the PPV after Phase II (33%),
- and the power of our Phase III studies.

There are noticeable differences in the operating characteristics of Phase II and overall. Therefore, we look to determine what characteristics define the optimal study for adopting these 124 effective treatments based on the objectives of the individuals involved in the treatment discovery process. We are also interested in optimizing our approach for maximizing the number of effective treatments adopted.

As we go down the table in Table 2.2, we are conducting more Phase II studies and the Phase II and overall type I error and power decrease in order to attain the same number of effective (TP) and ineffective (FP) treatments at the end of Phase III. Although these tradeoffs of type I error (the regulatory objective) and power (the sponsor objective) exist, differences in overall type I error do not appear to be important enough to define the optimal approach. Overall power and the maximum number of treatments adopted relative to the number of treatments tested are possible measures of the use of resources. Therefore, for each collection of study designs defined by the number of treatments adopted and the overall PPV, the optimal approach might be the one with the highest power and the greatest number of effective treatments adopted relative to the number of treatments tested. Overall, the Phase II type I error is consistent at around 0.2 and the PPV after Phase II tends towards 35%. Such consistencies show that Phase II parameter values are predictive of the PPV overall. Examining Table 2.2 overall, there are minor differences within a collection of designs adopting the same number of treatments and major differences across collections of designs where testing more treatments results in the adoption of more treatments.

Now consider a prevalence of truly effective treatments of 10% and a set of designs that result in a PPV of 98% after Phase III. Improvements over studies designed with a PPV of 95% include adopting fewer ineffective treatments (fewer FPs) for the same number of effective treatments (TPs) adopted, a decrease in the overall type I error (0.002 vs 0.005), and an increase in the overall sample size (total subjects per adopted treatment, $N_2 + N_3$), information useful for assessing treatment safety (Table 2.3). The Phase II type I error is mostly consistent at around 0.07 and the PPV after Phase II tends towards 60%. This increase in the PPV after Phase II is a direct result of a lower α_2 causing fewer ineffective treatments to be passed along to Phase III. It is intuitive to think that a low α_2 is optimal, however, because a low α_2 requires a larger Phase II sample size, N_2 when powered at an appropriate level to detect the desired design alternative, sending α_2 to 0 limits the number of treatments adopted after Phase III when holding resources constant.

By fixing overall operating characteristics (type I error, power, PPV) at an acceptable level, we can optimize our approach by selecting Phase II operating characteristics that maximize the number of treatments adopted. Within either the 95% or 98% PPV setting, the optimal approach is one that requires the fewest number of tested treatments necessary for adopting the maximum number of effective treatments after Phase III. Across these settings, the approach with a higher PPV is preferred. Adopting more treatments results in a loss of overall power compared to study designs where fewer treatments are adopted. We will discuss the cost of high power in section 2.3.2. The Phase II to Phase III approach at the bottom of Table 2.3 is very similar to Approach 2b in Table 2.1 and represents the design with the most effective treatments identified among the constraints considered including an overall PPV of 98%. Because of the similarity of this design with Approach 2b and the idea that it is a result of our search for the optimal design, we will also use this design for some comparisons in later chapters.

We further explore what is required beyond the specifications of our constraints to achieve an even greater number of effective treatments adopted. Because of the relationship between overall power and the number of effective treatments adopted, it is apparent that the study design which truly maximizes the number of effective treatments adopted after Phase III with constant resources, 98% PPV, and $\alpha_3 = 0.025$ requires a lower power at Phase II and overall. With a prevalence of effective treatments of 10%, this approach is one that begins with testing 4,000 candidate treatments in Phase II with α_2=0.055 and β_2=0.52. Of the 400 truly effective treatments, 208 pass through to Phase III along with 198 ineffective treatments. After Phase III with β_3=0.93, 194 effective treatments and 4.9 ineffective treatments are adopted. Although this approach meets many of the constraints outlined in section 2.2, the overall power ($\beta_2\beta_3 = 0.48$) is probably unacceptably low from the viewpoint of the sponsor. This illustrates the trade off between overall power and the number of effective treatments adopted after Phase III.

If we decrease the prevalence of truly effective treatments to only 5%, similarities exist with respect to the trends of our operating characteristics. However, a lower α_2 is required to address the issue of there existing fewer effective treatments and more ineffective treatments to identify (Appendix). On the other hand, if instead the prevalence of truly effective treatments is 50%, we are able to tolerate a larger α_2 while still maintaining other desirable operating characteristics including a minimum PPV of 97%. With high prevalence ($\geq$ 20%), the optimal approach meeting the constraints outlined in section 2.2 is defined by a single study design where the number of treatments studied directly corresponds to the number adopted. For example, if we seek to adopt 296 effective treatments in a setting where the prevalence is 20%, approaches resulting in either 95 and 98% PPV require that we study 1,800 new treatments (Appendix).

Table 2.2: Examples of various Phase II to Phase III approaches with 10% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 95% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N_3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
600	509	0.220; 98%	59; 119	33%	178	3,912	0.025; 100%	59; 3.0	4,421	0.12	0.0056; 98%
700	223	0.190; 84%	59; 120	33%	179	4,727	0.025; 100%	59; 3.0	4,950	0.05	0.0048; 84%
900	342	0.200; 93%	84; 162	34%	246	2,817	0.025; 100%	84; 4.0	3,159	0.11	0.0049; 93%
1,000	232	0.180; 84%	84; 162	34%	246	3,121	0.025; 100%	84; 4.0	3,353	0.07	0.0044; 84%
1,100	466	0.205; 97%	107; 203	34%	310	1,574	0.025; 100%	107; 5.1	2,040	0.23	0.0052; 97%
1,200	282	0.190; 89%	107; 205	34%	312	2,120	0.025; 100%	107; 5.1	2,402	0.12	0.0047; 89%
1,300	218	0.175; 82%	107; 205	34%	312	2,302	0.025; 100%	107; 5.1	2,520	0.09	0.0044; 82%
1,300	428	0.200; 96%	125; 234	35%	359	1,236	0.025; 99%	124; 5.9	1,664	0.26	0.0050; 95%
1,400	282	0.190; 89%	125; 239	34%	364	1,662	0.025; 100%	124; 6.0	1,944	0.15	0.0048; 89%
1,500	227	0.175; 83%	125; 236	35%	361	1,827	0.025; 100%	124; 5.9	2,054	0.11	0.0044; 83%
1,600	322	0.200; 92%	147; 288	34%	435	1,116	0.025; 99%	145; 7.2	1,438	0.22	0.0050; 91%
1,700	249	0.185; 86%	146; 283	34%	429	1,344	0.025; 100%	146; 7.1	1,593	0.16	0.0046; 86%
1,800	209	0.175; 81%	146; 284	34%	430	1,452	0.025; 100%	145; 7.1	1,661	0.13	0.0044; 81%
1,900	287	0.185; 89%	169; 316	35%	485	936	0.025; 97%	164; 7.9	1,223	0.23	0.0046; 86%
2,000	223	0.180; 83%	166; 324	34%	490	1,132	0.025; 99%	164; 8.1	1,355	0.16	0.0045; 82%

Table 2.3: Examples of various Phase II to Phase III approaches with 10% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 98% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N_3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
600	793	0.070; 98%	59; 38	61%	97	5,424	0.025; 100%	59; 0.9	6,217	0.13	0.0017; 98%
700	414	0.060; 84%	59; 39	61%	98	7,353	0.025; 100%	59; 0.9	7,767	0.05	0.0014; 84%
900	541	0.075; 93%	84; 61	58%	145	3,550	0.025; 100%	84; 1.5	4,091	0.13	0.0019; 93%
1,000	401	0.065; 84%	84; 59	59%	143	4,205	0.025; 100%	84; 1.5	4,606	0.09	0.0017; 84%
1,200	453	0.075; 89%	107; 81	57%	188	2,432	0.025; 100%	107; 2.0	2,885	0.16	0.0019; 89%
1,300	364	0.070; 82%	107; 82	57%	189	2,793	0.025; 100%	107; 2.0	3,157	0.12	0.0017; 82%
1,400	465	0.070; 89%	125; 88	59%	213	1,639	0.025; 100%	124; 2.2	2,104	0.22	0.0017; 89%
1,500	388	0.065; 83%	125; 88	59%	213	1,969	0.025; 100%	124; 2.2	2,357	0.16	0.0016; 83%
1,700	404	0.075; 86%	146; 115	56%	262	1,197	0.025; 99%	145; 2.9	1,601	0.25	0.0019; 86%
1,800	342	0.075; 81%	146; 122	55%	268	1,438	0.025; 100%	145; 3.0	1,780	0.19	0.0019; 81%
1,800	372	0.090; 86%	155; 146	51%	301	1,097	0.025; 100%	153; 3.6	1,464	0.25	0.0022; 86%
1,900	321	0.085; 81%	154; 145	51%	299	1,303	0.025; 99%	153; 3.6	1,616	0.20	0.0021; 81%
2,000	345	0.085; 83%	166; 153	52%	319	974	0.025; 99%	162; 3.8	1,319	0.26	0.0021; 81%

2.3.2 The cost of high power

Consider each collection of study designs defined by the number of treatments adopted and the overall PPV (collection of rows in Tables 2.2 and 2.3). The ratio of ineffective (FP) to effective (TP) treatments is approximately 1/20 across all collections when holding the PPV to be at least 95% and approximately 1/50 when holding the PPV to be at least 98%. This suggests that implementing an approach adopting a large number of treatments will have results similar to repeating an approach where fewer treatments are adopted. For example, using an additional 1,000,000 patients to repeat the approach where 84 effective treatments and 1.5 ineffective treatments are adopted would double our outcome to 168 effective and 3 ineffective treatments (Table 2.3). This is comparable to starting with 2,000 Phase II trials and adopting 162 effective and 3.8 ineffective treatments. It appears as though the latter approach is inferior with fewer effective and more ineffective treatments adopted, however; we can consider the average cost per effective treatment defined by the total number of subjects in both Phase II and Phase III divided by the number of effective treatments adopted as defined in section 1.3.1.

Conducting 900 Phase II trials of 541 subjects each and 145 Phase III trials of 3,550 subjects each to approve 84 effective treatments costs, on average, 11,924 subjects per effective treatment adopted. Instead, conducting 2,000 Phase II trials of 345 subjects each and 319 Phase III trials of 974 subjects each to approve 162 effective treatments costs, on average, 6,177 subjects per effective treatment adopted. At the cost of overall power, starting with more than twice as many treatments to study reduces the number of subjects needed to approve each beneficial treatment by about half as well as doubles the number of approved effective treatments. These relationships are illustrated by the plot in Figure 2.1. The designs featured in this plot are those with 98% PPV overall and the highest overall power, $(\beta_2\beta_3) \geq 0.80$ for identifying a particular number of effective treatments. For example, if two approaches result in the adoption of 50 treatments after Phase III, one with 83% and the other with 99% overall power, we chose the approach with more power to represent 50 treatments in Figure 2.1.

For study designs that pass $\leq$120 effective treatments there is a range of overall powers that can be achieved but as we adopt more treatments, our power is limited to the lower end of the constraint that $(\beta_2\beta_3) \geq 0.8$. This average cost in terms of subjects per effective treatment decreases as we adopt more treatments. The point on the plot in the lower left-hand corner represents a Phase II to Phase III approach adopting 40 effective treatments with 81% overall power costing 24,861 subjects per treatment. The point on the plot in the lower right-hand corner represents a Phase II to Phase III approach adopting 160 effective treatments with 80% overall power costing 6,250 subjects per treatment.

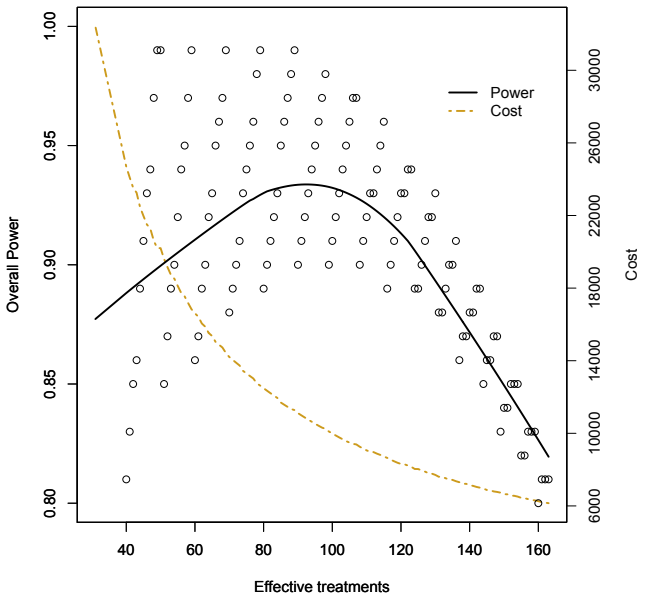


Figure 2.1: Overall power and cost against the number of effective treatments adopted. Cost is defined as the total number of subjects in both Phase II and Phase III divided by the number of effective treatments adopted. The curve shown for power is a LOWESS (locally-weighted polynomial regression) smoother summarizing the central tendency of the distribution of overall power (y-axis) at different locations within the distribution for the number of effective treatments adopted (x-axis).

The optimal approach in terms of the number of effective treatments adopted and the most efficient approach in terms of cost is one that maximizes the number of treatments studied with a high overall PPV. This comes at the cost of sponsor power but the approach of simply increasing the number of treatments taken through the Phase II process is convincing in terms of the efficient use of resources for finding more treatments. Figure 2.2 illustrates this tradeoff between overall power and the number of treatments tested for the same number of effective treatments adopted. This demonstrates the need for only a slight increase in the number of treatments tested in order for a lower overall power to be considered acceptable.

Considering the impact of this power/cost tradeoff, a small biotech company with only a single treatment to study, for example, will naturally desire to have high power. However, this requires a great amount of resources. On the other hand, a large pharmaceutical company or investment community that is funding drug development can consider the averages. For instance, they can fund 110 treatment trials with high overall power (e.g., $> 90\%$) and expect to have 10 "successes", or they can fund 125 treatment trials with lower overall power (e.g., 80%) and expect to have 10 "successes" as illustrated by Figure 2.2. Fewer subjects are needed per trial to detect a treatment

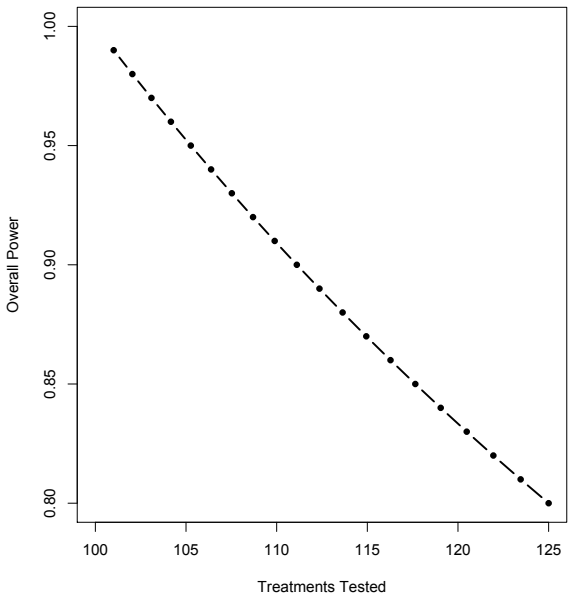


Figure 2.2: Overall power against the number of treatments tested to expect to pass 10 effective treatments

effect with lower overall power. Hence, by studying more treatments with lower power and fewer subjects per trial, we can adopt the same number of treatments at a lower cost or total number of subjects per effective treatment adopted. (Lowering power even more could achieve greater gains, at the cost of missing more truly effective treatments.)

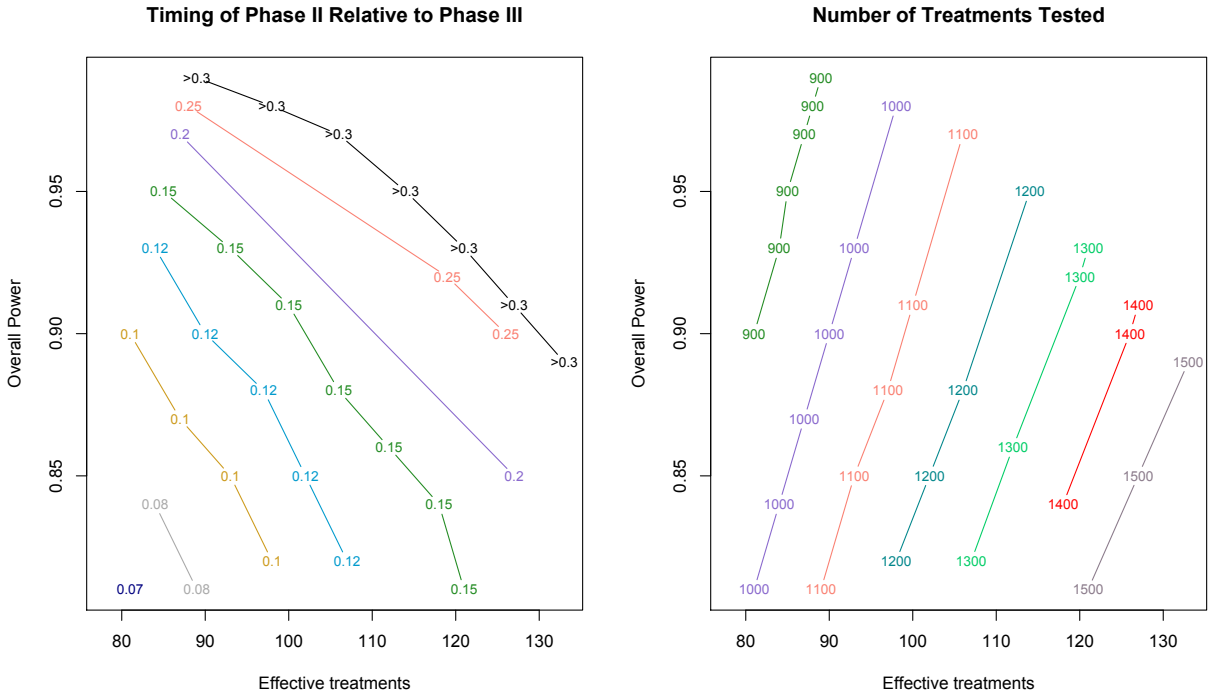
2.3.3 Timing of Phase II relative to Phase III

Determining when and how to conduct Phase II screening trials involves choosing from an infinite number of possibilities. We can choose not to conduct Phase II trials prior to confirmatory studies, to conduct a large number of small Phase II trials, to conduct a small number of large Phase II trials, or to be somewhere in between. When optimizing the treatment discovery process we seek to determine the most efficient relative sizes of the phases of investigation.

For all adopted treatments, the number of subjects involved in testing each treatment is the sum of the Phase II and Phase III sample sizes, $N_2 + N_3$. The amount of information used for the Phase II assessment is defined by $N_2/(N_2 + N_3)$ where smaller values indicate a preliminary assessment of the treatment at an earlier stopping point with less information. In comparing the various designs in Tables 2.2 and 2.3, we can discuss the trends that exist with respect to the timing of our investigations. We defined the optimal approach among all designs with the same number of effective and ineffective treatments passed after Phase III as one with the most effective treat-

ments adopted relative to the number of treatments tested. As the number of treatments adopted increases, the amount of information at which Phase II is conducted increases. This suggests that more efficient designs are a result of conducting Phase II studies at later stopping points with more information. It is also true that as we move from having 95% to 98% PPV, the timing of the end of Phase II relative to the end of Phase III shifts further out.

Holding constant the number of ineffective treatments adopted, plotted in Figure 2.3 is the number of effective treatments adopted against the overall power of the study (the proportion of effective treatments adopted out of all truly effective treatments tested at the start of Phase II).



(a) (b)

Figure 2.3: Overall power against the number of effective treatments passed with an approach that passes 2 ineffective treatments with 98% positive predictive value (PPV) overall. Contour lines in figure (a) on the left show the relationship with the timing of Phase II relative to Phase III as a proportion of the number of subjects used in the overall approach, and contour lines in figure (b) on the right show the relationship with the number of treatments initially tested in Phase II.

The plot on the left illustrates the relationship of overall power and the number of effective treatments adopted with constant contours of the timing of Phase II relative to Phase III. A point in the lower left-hand corner represents a design with a low overall power, fewer effective treatments

identified, and a low proportion of the maximal combined sample size ($N_2 + N_3$) used in Phase II. Although not simultaneously, we can achieve highest power and the greatest number of effective treatments when conducting our Phase II trials at more than 30% of our maximal combined sample size ($N_2 + N_3$). The plot on the left illustrates the relationship of overall power and the number of effective treatments adopted with constant contours noting the number of treatments initially tested. A point in the upper left-hand corner represents a design with a high overall power, fewer effective treatments identified, and fewer treatments tested in Phase II. This plot is representative of study designs starting with a range of treatments tested at Phase II (900-1500) where more effective treatments are adopted when more new treatments are tested, with a resulting loss of overall power.

Progressing from Phase II to Phase III is in fact a sequential process. We can think of the timing of Phase II relative to Phase III as the spacing of an interim analysis in a group sequential design (GSD). We can search for the optimal design based on a number of parameters including the spacing of our analyses. Sequential sampling will be discussed in chapter 3.

2.4 Exploring a continuous prior distribution

Previously, we considered a marginal distribution in which we examined two hypotheses: the prevalence of an effective therapy at the MCID of θ_1 is π, and θ is 0 (no effect) with probability $1 - \pi$. However, more realistic is a setting within which the prior distribution for the treatment effect is continuous. In comparison to our binary prior, this represents a smearing of the effect centered at the null or design alternative. The candidate treatments under study may correspond to an array of θ's different from the specified hypotheses. Therefore, we will relax the assumption on the marginal distribution $p(\theta)$ by considering a mixture of normal distributions with means centered at no effect and the effect that is the MCID, θ_1.

We define the treatment effect parameter, θ as an independently and identically distributed (iid) random variable drawn from two different normal distributions and specify the model as

$$\theta \sim N(\mu_0 = 0, \sigma_0^2) \text{ with probability } 1 - \pi$$
$$\theta \sim N(\mu_1 = \theta_1, \sigma_1^2) \text{ with probability } \pi$$

Then the mixture of distributions can be defined as

$$f(\theta; \pi, \mu_0, \mu_1, \sigma_0, \sigma_1) = (1 - \pi)\phi\left(\frac{(\theta - \mu_0)}{\sigma_0}\right) + \pi\phi\left(\frac{(\theta - \mu_1)}{\sigma_1}\right)$$

We imagine a population of treatments that includes $K_2(1-\pi)$ truly ineffective treatments with absolutely no effect ($\theta = 0$) and $K_2\pi$ truly effective treatments identified in pre-clinical studies with

some positive effect that is not necessarily the design alternative. To model this relationship, we will consider $\sigma_0 = 0$ and $\sigma_1 = 0.04$. The choice of σ_1 is arbitrary and was chosen to provide only positive treatment effects. In order to extend results of simulations already presented, we can simplify this approach by considering a discretized normal distribution. After generating a random normal with mean, μ_1 and standard deviation, σ_1, we identified quantiles of the distribution at 1/18, 3/18, 5/18, 7/18, 9/18,..., 17/18. We then placed 1/9 of the mass at each quantile to represent a "smeared" distribution for the effect of beneficial treatments. The true weights assigned to this weighted average of the positive treatment effect distribution would be defined as the prevalence of treatments exhibiting a particular treatment effect in a multinomial prior distribution. This distribution was chosen for demonstrating the effects of a smeared treatment effect on our outcomes and we recognize that other distributions may be considered. Tables 2.2 and 2.3 showing outcomes for the Phase II to Phase III approach were re-created with the application of this new smeared prior distribution.

The new prior distribution alters the power with which we can detect beneficial treatment effects with a trial powered to detect the design alternative. We are more powered to detect more extreme treatment effects but much less powered to detect small effects reducing our power at each phase and overall. Hence, the major change in outcome is the number of effective treatments passed after each phase of the investigation but this only differs by 10-20% of the treatments we obtained before with a binary prior distribution (Tables 2.4 and 2.5). We still have approximately 33% and 50-60% PPV after Phase II when achieving a PPV after Phase III of 95% and 98%, respectively. Otherwise, our results for the approach defined as most efficient with respect to the various parameters of greatest value remains the same. Thus, our findings suggest that considering an average treatment effect is suitable when identifying an optimal treatment discovery strategy defined by various operating characteristics.

In considering this prior distribution we recognize that the effects of some of the "positive" treatments are not greater than or equal to the MCID. We can think about whether we should consider a positive result for a treatment with an effect less than the MCID as a false positive or a true positive. To examine this further, in Figure 2.4 is the posterior distribution for the 137 adopted treatments with a positive effect when studying 2,000 (10% truly effective) treatments in Phase II and 309 treatments in Phase III (Table 2.5). We show that 89 out of the 137 (65%) positive treatments passed after Phase III have treatment effects greater than or equal to the MCID of 0.125. Overall, estimates of the average number of adopted treatments might be slightly overstated but the methods for choosing the optimal approach still apply.

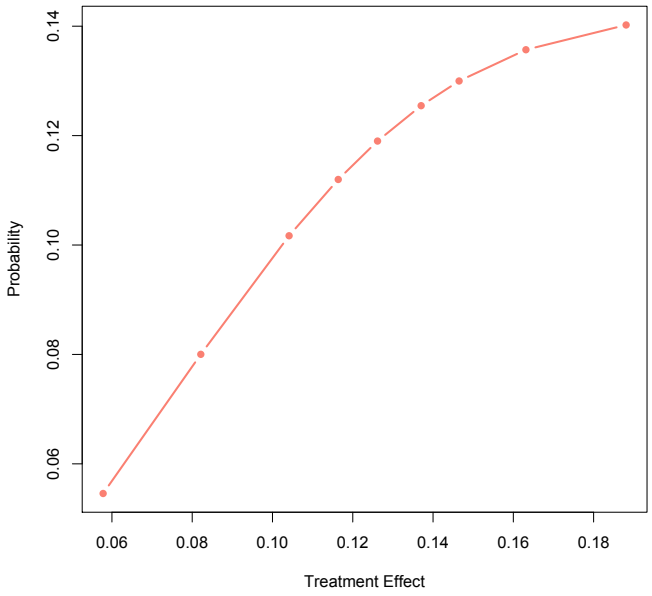


Figure 2.4: Posterior distribution for adopted treatments with a "positive" or significant effect when the minimal clinically important difference (MCID) is 0.125

Table 2.4: Examples of various Phase II to Phase III approaches with a multinomial prior distribution, 10% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 95% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
600	509	0.220; 94%	56; 119	32%	175	3,912	0.025; 100%	56; 3.0	4,421	0.12	0.0055; 93%
700	223	0.190; 82%	57; 120	32%	177	4,727	0.025; 100%	57; 3.0	4,950	0.05	0.0048; 82%
900	342	0.200; 87%	78; 162	32%	240	2,817	0.025; 96%	75; 4.0	3,159	0.11	0.0050; 83%
1,000	232	0.180; 79%	79; 162	33%	241	3,121	0.025; 98%	78; 4.0	3,344	0.07	0.0045; 78%
1,100	466	0.205; 92%	101; 203	33%	304	1,574	0.025; 93%	94; 5.1	2,040	0.23	0.0051; 85%
1,200	282	0.190; 85%	102; 205	33%	308	2,120	0.025; 96%	99; 5.1	2,402	0.12	0.0047; 82%
1,300	218	0.175; 81%	105; 205	34%	310	2,302	0.025; 99%	104; 5.1	2,520	0.09	0.0044; 80%
1,300	428	0.200; 91%	118; 234	34%	352	1,236	0.025; 89%	105; 5.8	1,664	0.26	0.0050; 81%
1,400	282	0.190; 86%	121; 239	34%	360	1,662	0.025; 96%	116; 6.0	1,944	0.15	0.0048; 83%
1,500	227	0.175; 78%	117; 236	33%	353	1,827	0.025; 95%	110; 5.9	2,054	0.11	0.0044; 74%
1,600	322	0.200; 87%	139; 288	33%	427	1,116	0.025; 87%	121; 7.2	1,438	0.22	0.0050; 76%
1,700	249	0.185; 82%	139; 283	33%	422	1,344	0.025; 91%	127; 7.1	1,593	0.16	0.0046; 75%
1,800	209	0.175; 78%	141; 284	33%	425	1,452	0.025; 94%	133; 7.1	1,661	0.13	0.0044; 74%
1,900	287	0.185; 86%	163; 316	34%	479	936	0.025; 86%	140; 7.9	1,223	0.23	0.0046; 74%
2,000	223	0.180; 77%	153; 324	32%	477	1,132	0.025; 87%	133; 8.1	1,355	0.16	0.0045; 66%

Table 2.5: Examples of various Phase II to Phase III approaches with a multinomial prior distribution, 10% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 98% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
600	793	0.070; 92%	55; 38	59%	93	5,424	0.025; 100%	55; 0.9	6,217	0.13	0.0017; 83%
700	414	0.060; 79%	55; 38	59%	93	7,353	0.025; 100%	55; 0.9	7,767	0.05	0.0015; 68%
900	541	0.075; 85%	77; 61	56%	137	3,550	0.025; 100%	76; 1.5	4,091	0.13	0.0019; 75%
1,000	401	0.065; 78%	78; 59	57%	136	4,205	0.025; 100%	77; 1.5	4,606	0.09	0.0016; 66%
1,200	453	0.075; 81%	97; 81	55%	178	2,432	0.025; 97%	95; 2.0	2,885	0.16	0.0019; 74%
1,300	364	0.070; 79%	103; 82	56%	185	2,793	0.025; 99%	102; 2.0	3,157	0.12	0.0018; 62%
1,400	465	0.070; 85%	119; 88	57%	207	1,639	0.025; 96%	114; 2.2	2,104	0.22	0.0017; 71%
1,500	388	0.065; 80%	120; 88	58%	207	1,969	0.025; 98%	117; 2.2	2,357	0.16	0.0016; 60%
1,700	404	0.075; 80%	136; 115	54%	250	1,197	0.025; 91%	124; 2.9	1,601	0.25	0.0019; 69%
1,800	342	0.075; 77%	138; 122	53%	260	1,438	0.025; 95%	131; 3.0	1,780	0.19	0.0019; 65%
1,800	372	0.090; 82%	147; 146	50%	293	1,097	0.025; 91%	134; 3.6	1,469	0.25	0.0022; 68%
1,900	321	0.085; 78%	148; 145	50%	293	1,303	0.025; 94%	139; 3.6	1,624	0.20	0.0021; 62%
2,000	345	0.085; 78%	156; 153	50%	309	974	0.025; 88%	137; 3.8	1,319	0.26	0.0021; 66%

Chapter 3

Sequential Sampling

In previous sections, we considered only fixed sample-size designs for Phase II and Phase III trials. We extend the Phase II to Phase III approach to the sequential setting to demonstrate how sequential monitoring in either phase can reduce our average sample size, thereby increasing the number of effective treatments that can be identified with fixed resources.

Sequential monitoring of a clinical trial involves using data accrued to the point of observation to make decisions about future sampling. We can think of the progression from Phase II studies to Phase III trials as sequential: after Phase II, with information from only a subset of the maximal sample size ($N_2 + N_3$), we analyze the data to determine if the experimental treatment is worth studying in a confirmatory Phase III trial. Methods for sequential sampling include "pre-specified" maximal information plans (e.g. group sequential designs) and "adaptive" plans (e.g. consideration of subgroups, population enrichment, re-powering a trial for a different treatment effect). With "pre-specified" plans, we control the type I error to lessen our chances of finding spurious associations or making invalid inference. We pre-specify conditions under which the trial will be modified as well as the modifications to be made. In this section, we will examine the pre-specified sequential sampling setting and the methods that can address the scientific, statistical, and ethical concerns characteristic of clinical trials. Adaptive designs allowing more extensive modifications of the trial design will be discussed in chapter 5.

Pre-specified sequential sampling plans allow for interim monitoring of data for efficacy and futility to address ethical and efficiency concerns. Using interim estimates of the treatment effect, early termination of a clinical trial may result from concerns about patient safety, evidence of efficacy of the new treatment, or evidence of such a lack of efficacy that further study of the intervention is no longer warranted or desired. At each analysis (or during each phase of investigation) we consider parameters such as the significance level, the power to detect the alternative as a function of the true treatment effect, the amount of information necessary for making inference at any stage of the

investigation, and the probability of obtaining a significant result. The goal of a sequential sampling plan is to proceed to the maximal sample size only when we are not yet certain of treatment benefit and when potential remains for the results of the trial to lead to modification of standard clinical practice. We are interested in accruing a minimal number of subjects when there is evidence that the treatment is harmful or ineffective, only minimally effective, or extremely beneficial.

3.1 Setting and notation

We will describe a generic sequential sampling plan with four boundaries using the following notation. Let $X_1, X_2, X_3, ...$ denote the potential observations. We will perform up to J analyses at sample sizes $N_1, N_2, N_3, ..., N_J$, which can be randomly determined independent of treatment effect. At each analysis, we will specify stopping boundaries $a_j \leq b_j \leq c_j \leq d_j$ and compute a test statistics $T_j = T(X_1, ..., X_{N_j})$. The trial will stop under the following conditions

- if $T_j \leq a_j$, stop for inferiority of the new treatment relative to control;
- if $b_j \leq T_j \leq c_j$, stop for approximate equivalence between the new treatment and control;
- if $T_j \geq d_j$, stop for superiority of the new treatment relative to control;
- otherwise, continue the trial.

Under this notation, we choose $a_J = b_J$ and $c_J = d_J$ to guarantee early trial termination by the final analysis. The interim analysis at which the data first meet the conditions of early stopping will be denoted by M where T_M is the value of the statistic at that stopping time. It is important to note that the desired power, β and type I error, α of a fixed sample-size design can be achieved with a sequential sampling plan (O'Brien, Fleming, 1979).

A number of sequential sampling plans are commonly implemented in clinical trials defined by the test statistic used in the definition of the stopping rule. Group sequential designs (GSDs) are sampling plans with pre-specified boundary shape functions that define the conservatism of each boundary and achieve a desired type I error and power (Pocock, 1977, 1982; O'Brien, Fleming, 1979; DeMets, Ware, 1980, 1982). A stopping rule defined for one test statistic is automatically defined on other scales, including the partial sum, crude estimate of treatment effect, the normalized Z statistic, fixed sample p-value, error spending, Bayesian posterior probability, and conditional/predictive power. There exists families of GSD boundary functions (e.g., O'Brien-Fleming, 1979, Pocock, 1977, Wang & Tsiatis, 1989, triangular (Whitehead, Stratton, 1983)) parameterized by the proportion of statistical information accrued and the hypothesis rejected by the boundary (Emerson, Kittelson, Gillen, 2007a). Of importance is appropriate consideration of the operating

characteristics of a particular stopping rule such that statistical credibility of the study is achieved. The O'Brien-Fleming (OBF) boundary is the most commonly implemented stopping rule in clinical trials and the Pocock is a common design used in exploring group sequential methods. The OBF of the unified family methods (Kittelson, Emerson, 1999) is more conservative in early stages and tends to stop trials early only for extreme treatment effects that are highly significant. The Pocock sets the same critical boundary at each analysis, is constant on the standardized Z scale, and tends to be highly efficient in terms of the average sample size used under the hypotheses being tested.

Repeated analysis of accruing data involves multiple comparisons. In order to make use of the advantages of sequential monitoring, we are forced to employ methods of controlling the false positive rate at an acceptable level. On the error spending scale, we can define the amount of error spent at each interim analysis such that we ensure a particular cumulative α-level at the end of the trial. For example, a Pocock α-level 0.05 stopping rule with three interim analyses specifies three equally spaced fixed sample level 0.022 analyses with cumulative error spent at each analysis of 0.022 (46%), 0.038 (77%), and 0.050 (100%). We might also consider an OBF α-level 0.10 stopping rule with three interim analyses, which specifies equally spaced fixed sample level 0.003, 0.036, and 0.087 analyses with cumulative error spent at each analysis of 0.003 (3%), 0.038 (38%), and 0.100 (100%).

When inspecting data at interim looks, we can no longer analyze the data in the same way as fixed sample-size studies. New methods for calculating point estimates, confidence intervals, and p-values are required. Otherwise, if traditional methods are applied at the end of a trial that stops early for efficacy of the new treatment, the point estimate calculated will be too extreme, the confidence interval too narrow, and the p-value too small (Todd, Whitehead, Stallard, 2001). Well developed methods for statistical analyses following a sequential trial have been comprehensively evaluated and documented (Whitehead, 1997; Jennison, Turnbull, 2000). Trials are typically monitored by a data safety monitoring board (DSMB) to ensure that scientific and ethical decisions are made with as little bias as possible. It is important that interim looks are implemented appropriately to ensure statistical credibility of trial results.

The Phase II to Phase III progression as a whole is similar to this concept of testing a treatment at multiple stages of our investigation. Phase II and Phase III are typically unequally spaced analyses and the timing of those evaluations makes a difference when considering the error spent. The major difference between sequential sampling and the Phase II/Phase III progression involves time delay between the end of Phase II and the start of Phase III studies, and the decisions that are made between Phase II and Phase III. Because this approach is already sequential in nature, we can consider introducing interim analyses into each phase to allow for early stopping for efficacy or futility. Early stopping reduces the average sample number (ASN) making the design more time-

and cost-efficient on average and addresses the ethical concerns of continuing a trial with sufficient evidence for or against the treatment under study.

3.2 Sequential sampling in Phase II to Phase III approach

Phase II studies are hypothesis-generating research and exploratory in nature. Results are primarily used for determining the course for further investigation of a new treatment rather than for providing definitive evidence of efficacy (Stallard, Whitehead, Todd, 2001). This requires that we be necessarily conservative when applying methods and interpreting results during such preliminary investigations. The concerns of bias in our Phase II estimates and the detriment of believing Phase II results or designing Phase III studies based on Phase II are discussed further in sections 5.1.1 and 5.1.2.

It is important to consider a conservative stopping rule for efficacy in Phase II if early stopping for efficacy is allowed at all. If a Phase II trial is stopped early for efficacy, next steps are to move forward with studying the treatment in Phase III, but now with less information because the trial ended with data on fewer patients. It might be more appropriate to continue the trial to obtain more information about the treatment prior to further evaluations in future studies. We choose to consider either not stopping early or applying a conservative OBF boundary for efficacy in Phase II.

Trials are also monitored for treatment-related adverse events as part of the investigation to determine if further study is warranted. Such safety monitoring rarely involves a formal stopping rule. However, it is also appropriate to consider discontinuing a trial for lack of efficacy. We will use either Pocock or OBF boundaries as a "futility" boundary to allow for early stopping when there is sufficient evidence of a lack of efficacy.

Confirmatory Phase III studies are the most rigorous evaluations of the treatment discovery process and are required to be well-controlled, randomized trials with a clearly defined treatment strategy, a clearly defined and ascertained clinical outcome, and a pre-specified primary analysis. Many authors have discussed the importance of sequential stopping rules in Phase III studies because they allow for definitive evidence of the effect of a treatment to lead to a more rapid and economically and ethically desirable treatment discovery process (Todd, Whitehead, Stallard, 2001; Emerson, Kittelson, Gillen, 2007a). It is typical to consider stopping early for both efficacy and futility in Phase III trials for achieving the goals of a definitive investigation of a new treatment. Earliest analyses tend to be more conservative when allowing a trial to stop for efficacy such that decisions to accept a new treatment are not made based on little evidence.

To demonstrate the impact of group sequential methods on the ASN of our Phase II to Phase

III approach, we will introduce interim analyses to the phases of a single design in Table 2.3. The approach that will be considered is one which starts with 2,000 Phase II studies leading to 319 Phase III trials and resulting in 162 effective and 3.8 ineffective treatments adopted after Phase III. In order to cover a range of designs in our evaluation of group sequential methods, we will consider the following stopping rules applied to both Phase II and Phase III trials with either two or four equally spaced interim analyses:

1. *Pocock Futility boundary*: Stopping rule with J equally spaced analyses having a Pocock futility boundary.
2. *OBF Futility boundary*: Stopping rule with J equally spaced analyses having an OBF futility boundary.
3. *OBF Efficacy, Pocock Futility boundary*: Stopping rule with J equally spaced analyses having OBF efficacy and Pocock futility boundaries.
4. *OBF Efficacy, OBF Futility boundary*: Stopping rule with J equally spaced analyses having OBF efficacy and OBF futility boundaries.

The stopping boundaries for the clinical trial designs presented above can also be presented in tabular form as in Table A6 (Appendix). While there are a number of scales on which the boundaries can be defined, we present the scales of the crude estimate of the treatment effect and the cumulative error spent, where the cumulative error spending function refers to the type I error spending function for the efficacy boundary and to the type II error spending function for the futility boundary.

3.2.1 Improvement in average efficiency

With the addition of interim analyses, an inflation of the maximal sample size is necessary to maintain power under the design alternative (Table 3.1). Among the designs considered, the greatest increases in maximal sample size for Phase II trials result from the introduction of a conservative OBF efficacy boundary and a Pocock futility boundary (14% with $J = 2$; 28% with $J = 4$ over fixed sample) (Table 3.1). The same is true for Phase III (10% with $J = 2$; 20% with $J = 4$ over fixed sample). This increase in the maximal sample size is requiring that more resources be used when we are not certain of treatment benefit at any given interim analysis but need the maximum amount of information to make a decision. The impact of this increase is minimal when we examine the improvement in average efficiency.

Allowing for early stopping with any of the stopping rules above results in an improvement in the ASN under the null (no effect) and design alternative. Shown in Table 3.1 are sample sizes for the fixed Phase II to Phase III approach and the approach that utilizes sequential monitoring at each stage of the investigation. We choose to focus on stopping rules that may be appropriate

Table 3.1: The sample size for a fixed sample-size study and the maximum sample size, average sample number (ASN) and total sample sizes at Phase II and Phase III when introducing 2 or 4 Pocock or O'Brien-Fleming (OBF) interim analyses. In all Phase II designs, the type I error is 0.10 and the power is 0.85. In all Phase III designs, the type 1 error is 0.025 and the power is 0.97.

Fixed Phase II			Phase II GSD							
Eff; Ineff	N_2	N_2K_2	J	Efficacy	Futility	Max N_2	$\text{ASN}_{\theta=0}$	$\text{ASN}_{\theta=\Delta}$	Max Total	Total ASN
200; 1,800	345	690,000	2	-	Pocock	378	252	360	756,000	525,600
			2	-	OBF	352	268	344	704,000	551,200
			2	OBF	Pocock	393	253	287	786,000	512,800
			2	OBF	OBF	362	267	285	724,000	537,600
			4	-	Pocock	406	216	375	812,000	463,800
			4	-	OBF	358	243	346	716,000	506,600
			4	OBF	Pocock	443	213	263	886,000	436,000
			4	OBF	OBF	381	242	260	762,000	487,600
Fixed Phase III			Phase III GSD							
Eff; Ineff	N_3	N_3K_3	J	Efficacy	Futility	Max N_3	$\text{ASN}_{\theta=0}$	$\text{ASN}_{\theta=\Delta}$	Max Total	Total ASN
166; 153	974	310,706	2	-	Pocock	1,059	655	1,051	337,821	275,345
			2	-	OBF	980	738	979	312,620	275,594
			2	OBF	Pocock	1,068	657	758	340,692	217,717
			2	OBF	OBF	987	739	740	314,853	227,441
			4	-	Pocock	1,128	544	1,115	359,832	269,318
			4	-	OBF	993	656	990	316,767	265,040
			4	OBF	Pocock	1,168	545	694	372,592	191,617
			4	OBF	OBF	1,017	658	659	324,423	202,930

for each phase of the treatment discovery process. A typical approach to sequential sampling in Phase II would be to include only a futility boundary such that the trial continues to the maximum amount of information if the treatment is beneficial. As mentioned previously, this ensures that more information is available for moving forward to Phase III. It is important, though to consider stopping Phase II trials early for futility such that the trial is stopped as soon as there is evidence of a lack of efficacy, thus saving time and resources. By introducing four interim analyses and a Pocock futility boundary in Phase II, we can reduce our total ASN (under the null and alternative) to 67% of the fixed sample size ($N_2 = 232$ vs 345 subjects). In Phase III, we are interested in making a decision for or against the new treatment as soon as we have sufficient evidence to confirm a result. However, we still tend to be necessarily conservative at very early analyses for efficacy. Four interim analyses, an OBF efficacy boundary and a Pocock futility boundary reduces our total ASN to 64% of the fixed sample design ($N_3 = 622$ vs 974 subjects).

Phase II studies are typically small and relatively short-term with a focus on a more rapid determination of how to proceed with a new treatment. While this setting facilitates monitoring data

while it is accrued to ascertain results more quickly, there may not be enough time or information at Phase II to consider many interim analyses. Phase III studies are larger, involve longer-term endpoints with a purpose to confirm evidence of the effects of an experimental treatment. With a larger scale study, we have more information earlier in the trial and may consider more analyses throughout the study. While various combinations of such designs for Phase II and Phase III can be considered, a reasonable approach might be to implement a sequential stopping rule with two analyses having a Pocock futility boundary at Phase II and a stopping rule with four analyses having OBF (conservative) efficacy and Pocock futility boundaries at Phase III. With this approach, we can improve in average efficiency with a total ASN of 525,600 subjects for Phase II studies and a total ASN of 191,617 subjects for our Phase III trials (Table 3.1). This results in a total of 717,217 subjects for the entire approach compared to the 1,000,000 subjects used with a fixed sample-size design. By allowing the phases of our investigation to stop early when sufficient evidence is obtained, we can find the same number of treatments with fewer resources, on average.

In a search for the optimal design of equally spaced analyses where optimal is defined as the best average efficiency, we vary the P parameter of the boundary shape function that alters the difficulty of stopping early for futility within the unified family (Emerson, Kittelson, Gillen 2007a). Our findings suggest that the optimal Phase II design (ASN of 67% of the fixed sample size) with four interim analyses has a futility boundary specified by P = 0.495 that makes stopping early slightly less difficult than a traditional Pocock boundary. Note that P = 1 is an OBF boundary and P = 0.5 is a Pocock boundary. The optimal Phase III design (ASN of 64% of the fixed sample size) features a futility boundary where early stopping is slightly more difficult than a traditional Pocock boundary (P = 0.575). We recognize that optimality of these boundaries is dependent on the prevalence of truly beneficial treatments at the start of Phase II and Phase III. We are simply demonstrating that optimality of our approach does not cease at a fixed sample design but that we can be even more efficient with the introduction of an optimal stopping rule.

In previous sections, we considered optimizing the Phase II to Phase III approach with 1,000,000 subjects by searching for a design with a high PPV and a maximal number of effective treatments adopted. With a sequential monitoring plan, we improve the average efficiency of our Phase II to Phase III approach. We can now identify 162 effective and 3.8 ineffective treatments with an overall PPV of 98% with approximately 2/3 the number of subjects used in a fixed sample-size trial. By allowing clinical trials to stop early for efficacy or futility, we conserve resources necessary to conduct studies with the same degree of optimality and improve the treatment discovery process by more rapidly discovering beneficial treatments.

3.2.2 Bayesian posterior probability scale

As previously mentioned, the parameters that we find of greatest value in optimizing our Phase II to Phase III approach are the PPV, overall power, type I error and the absolute number of effective and ineffective treatments adopted. We can specify our type I and type II errors at the levels desired in a fixed sample-size design when defining stopping rules for a sequential design. Therefore, we would obtain the same number of adopted treatments at the end of the trial. What might be of interest to us are the negative and positive predictive values (NPV and PPV) after each stopping point of our GSD. If we were targeting a specific predictive value but our trial stops early, we might want to know how those rates differed from what we would expect from a fixed sample-size trial. In the Bayesian paradigm, we might even base our decision to stop early on the NPV or PPV scales. We will consider the previously discussed optimal stopping rules on such scales.

In Phase II, we found an optimal boundary that makes stopping early for futility slightly less difficult than a traditional Pocock boundary with no boundary for efficacy. Without an efficacy boundary, we cannot stop a trial early for benefit to adopt new treatments, therefore there is no PPV at each interim. We do achieve the overall PPV and the absolute number of treatments adopted after Phase II as obtained by the fixed sample-size design (Table 3.2). Because trials are allowed to stop early for futility, we can calculate the number of treatments (effective and ineffective) considered ineffective (false negatives; FN and true negatives; TN) at each interim analysis based on the stopping probabilities. The probability of stopping at each interim analysis is a direct specification of the error spending function. The amount of error spent at each analysis can be specified in the trial design. The NPV at each interim analysis is the probability that the treatments for which trials stop early for futility are truly ineffective (Equation 1.10). Shown in Table 3.2, the NPV at each interim analysis is high because most of the trials stopping early for futility are for treatments that are truly ineffective (TN). The overall NPV after Phase II is 98%.

We found that in Phase III, the most efficient boundary is one which makes stopping early for futility slightly more difficult that a traditional Pocock boundary with an OBF boundary for efficacy. The PPV at each interim analysis is the probability that the treatments for which trials stop early for efficacy are truly effective (Equation 1.8). At the first three interim analyses, the PPV is quite high ($\geq 97\%$). At the fourth interim analysis, the probability of stopping for efficacy under the null is the highest resulting in more ineffective treatments (false positive; FP) being adopted and a much lower PPV (84%). The overall PPV achieved is the same that was obtained in the fixed sample-size design. The NPV is high for each interim analysis and overall.

We note that the NPV and PPV rates are not constant but were simply a result of a search for the optimal Phase II and Phase III sequential designs. Examination of the stopping rules applied

Table 3.2: Stopping boundaries on the NPV and PPV scales for optimal Phase II (4 interim analyses, futility boundary with P = 0.495) and Phase III (4 interim analyses, futility boundary with P = 0.575, O'Brien-Fleming (OBF) efficacy boundary) designs

	Phase II						Phase III					
Analysis	"Positive"			"Negative"			"Positive"			"Negative"		
Time	TP	FP	PPV	FN	TN	NPV	TP	FP	PPV	FN	TN	NPV
1	0	0	-	12.0	693	0.983	6	0	0.998	1.1	53	0.980
2	0	0	-	8.9	507	0.983	90	0.4	0.995	1.2	60	0.981
3	0	0	-	7.1	292	0.976	56	1.5	0.974	1.1	27	0.961
4	166	153	0.520	6.0	156	0.963	10	1.9	0.843	0.9	8	0.903
Total	166	153	0.520	34	1,647	0.980	162	3.8	0.977	4.3	149	0.972

to a trial design on such scales is recommended recognizing that optimal boundaries are not going to be constant on the posterior probability scale. In the Bayesian paradigm, we might choose to define our stopping rule on the posterior probability scale. To do this, we would modify our design by specifying the error spending function such that the stopping probabilities under the null and design alternative at each interim analysis result in the desired NPV and PPV rates. In our Phase II to Phase III approach, the PPV after Phase II was a direct result of the error spent at Phase II; when less error was spent in Phase II, we achieved a higher PPV after each phase. We might choose to formulate our sequential design such that an acceptable NPV or PPV is achieved at each interim analysis. The same methods and considerations for finding an optimal, more efficient design apply when we divide our two phase approach into multiple stages with a sequential monitoring plan.

Chapter 4

Seamless Phase II/III

We have thus far considered a two-stage process consisting of screening Phase II studies and confirmatory Phase III studies by quantifying the power and type I error for each phase and overall. We also examined how introducing sequential sampling to each phase can improve the efficiency of the treatment discovery process. Other methods for improving efficiency by eliminating the time spent between Phase II and Phase III have been proposed. A "seamless" Phase II/III trial design is one that combines the Phase II screening stage with the Phase III confirmatory stage. The traditional objectives of Phase II, evaluation of dose, efficacy and safety, are assessed at an interim analysis while comparative evaluation of treatment efficacy and safety with placebo or standard of care is performed in the same study.

Eliminating the time delay or "white space" between the two phases shortens the time to market or timeline needed to make a decision about a new treatment. This time delay between phases is typically used for regulatory review, safety assessments, logistical and operational preparation for Phase III such as the operation of study sites, patient recruitment, IRB (Institutional Review Board) approval, sponsor buy-in, licensing of the treatment to larger corporations, and consideration of results from the investigation of other agents. Independent review of results from preliminary phases and of plans for future trials is important for assessing the appropriateness of the clinical trial protocol as well as the risks and benefits to study participants. Maintaining a non-adaptive approach, we can consider how our strategy would differ if we were to conduct the investigation as a single trial. For a design to be truly seamless without being adaptive, all considerations must be made in advance such as choosing the initial sample size and study endpoints, criteria for futility termination, and how to control the overall type I error. We will demonstrate that although seamless trials can be pre-planned, disregarding issues of safety, study logistics, and regulatory review that are often addressed during "white space", it remains most important to control the type I error for the entire process.

If no modifications are made between Phase II and Phase III of a seamless design, it is analogous to a GSD with a futility stopping rule where testing of unpromising treatments is discontinued after Phase II. Therefore, similar to the design of GSD, the design of seamless Phase II/III trials is different from the design of Phase II and Phase III studies independently. A major difference is the inclusion of all data in the final analysis when combining phases. Considering independent studies, Phase II trials might be non-randomized (this is more true of oncology trials than in other disease settings), have a smaller sample size, use a surrogate endpoint, and different operating characteristics (e.g. higher type I and II errors than Phase III). Design parameters for seamless Phase II/III studies are selected based on both the Phase II and Phase III trial characteristics. The overall power is the probability of a positive trial under the alternative hypothesis, which requires a positive Phase II and positive Phase III result, approximately the product of the individual powers (Korn, Freidlin, Abrams, Halabi, 2012).

Determining the overall type I error requires consideration of the error spent at each analysis. In analyzing Phase II data with data from Phase III, the Phase II trial is essentially an interim analysis traditional of GSDs. Korn, Freidlin and Halabi (2012) propose that the overall type I error of a Phase II/III design is equal to the type I error of its Phase III component alone, however; interim looks at the data introduce the multiple comparison issue inflating the type I error. Similar to methods for GSDs, this requires that we control the overall type I error for the seamless Phase II/III design to avoid inflation. Some authors ignore this inflation of the overall type I error or propose a Bonferroni correction when ultimately designing each phase separately in terms of the operating characteristics while combining results for making inference (Hunsberger, Zhao, Simon, 2009; Orloff, Douglas, Pinheiro, 2009). Other authors propose group sequential method for seamless Phase II/III studies (Stallard, Todd, 2003; Chow, Lu, Tse, 2007). At the first stage of the Phase II/III trial, patients are randomized between several experimental treatments and a control. Based on the interim analysis, only a single experimental treatment is selected to continue along with control to one or more subsequent stages. The selection of a stopping boundary is similar to GSD methods where the cumulative error spent is controlled to a desirable level. There are relatively few papers that discuss what the desirable level is for the overall type I error of a Phase II/III design. We worry that a seamless approach might not specify or achieve a level as low as what is attainable with our Phase II to Phase III approach previously considered ($\alpha = 0.0021$).

Proposed by Bauer and Kohne (1994), others have considered combining evidence from both phases using combination tests such as Fisher's combination test or the weighted inverse normal to obtain a single p-value at the end of the investigation (Bretz, Schmidli, Konig, 2006; Kimani, Stallard, Hutton, 2009). They note that the weighted combination test corresponds to a classical two-stage group sequential test without adaptation. If we recognize that controlling the overall type

I error is necessary and no modifications are made between Phase II and Phase III then we can consider a seamless Phase II/III design a group sequential test with the number of analyses equal to the sum of the analyses in each phase designed with desired overall operating characteristics.

Consider a single GSD combining the optimal Phase II and Phase III designs identified previously in section 3.2.1.

- This corresponds to a GSD with 8 interim analyses of size $\alpha = 0.0021$ with power, $\beta = 81\%$ to detect the design alternative with a stopping rule that includes an OBF boundary for efficacy and some futility boundary that defines the optimal design.
- An interim analysis would occur at the end of what would be considered Phase II with approximately 345 subjects and the last analysis at the end of Phase III with 1,319 subjects.
- We again search for the optimal design of equally spaced analyses where optimal is defined as the best average efficiency by varying the P parameter of the boundary shape function within the unified family. Our findings suggest that P = 0.14 provides the optimal design with a total ASN = 318 subjects when combining Phase II and Phase III together in a single GSD meeting our desired operating characteristics.

In Table 4.1, we examine the stopping boundaries on the posterior probability scale. What differs here from considering each phase independently is the inclusion of an OBF efficacy boundary in what would be considered Phase II (the first two interim analyses based on sample size). While we prefer not to allow stopping for efficacy in very early analyses, there is a low stopping probability (0.1) with this design. We demonstrate that the optimal Phase II/III approach might have different spacing compared to the independent design of each phase. The design presented here has two analyses for what would be considered Phase II and six analyses for the confirmatory stage. Posterior probabilities are not constant but instead, overall rates start high and average out over the course of the trial. The number of treatments considered "positive" by our studies (true positives; TP and false positives; FP) after this combined trial design is the same (162 effective and 3.8 ineffective) when trials were considered independently because the overall operating characteristics were achieved. Treatments not adopted by our studies here are considered "negative" (true negatives; TN and false negatives; FN).

However, this design is only comparable to the separate designs considered previously if no changes are made based on early results. A purpose of Phase II screening studies is to identify those treatments that are not of interest to study further. This preliminary phase is considered a learning stage; information about candidate therapies can be used in the design and implementation of future, confirmatory studies. Therefore, the idea of there not existing differences between Phase

Table 4.1: Seamless Phase II/III design parameterized as a GSD with stopping boundaries on NPV and PPV scales

	Seamless Phase II/III						
Analysis		"Positive"			"Negative"		
Time	N_j	TP	FP	PPV	FN	TN	NPV
1	155	0	0	1.000	18.7	1,069	0.983
2	309	0.1	0	1.000	6.5	385	0.983
3	464	8	0.01	0.999	3.8	181	0.979
4	618	38	0.1	0.997	2.7	88	0.971
5	773	54	0.4	0.992	2.1	42	0.953
6	927	39	0.9	0.977	1.7	20	0.919
7	1,082	18	1.3	0.932	1.5	9	0.857
8	1,237	5	1.1	0.827	1.0	3	0.762
Total	1,237	162	3.82	0.977	38.0	1,796	0.979

II and Phase III study designs, at least in some settings, is unrealistic. Hence, what we do not gain from considering a trial that combines phases into a single study is the ability to apply adaptive methods. Such methods will be discussed in chapter 5. We can also consider pre-planned differences between phases of our investigation mainly referring to the endpoints used in either stage, which will be discussed in section 4.1.

4.1 Non-adaptive seamless Phase II/III

A non-adaptive seamless Phase II/III approach can allow differences to exist between Phase II and Phase III while the intentions and the pre-specified plans for Phase II and Phase III do not change based on observed outcomes. For example, we can plan in advance to screen treatments with an immediate surrogate endpoint (ISE) such as measured progression while confirming results with a delayed clinical endpoint (DCE) such as overall survival. The advantages or disadvantages of using surrogate endpoints in early stages will depend on whether we are using a valid surrogate such that it can accurately diagnose the ability of a treatment to treat or cure the disease.

4.1.1 Surrogate endpoints

A common approach taken for seamless Phase II/III designs is to use different outcomes and alternative hypotheses for each of the screening and confirmatory stages. Unlike the primary outcome in confirmatory Phase III trials, which is typically representative of direct clinical benefit to the patient, immediate surrogate endpoints (ISE) are generally used in Phase II (Korn, Freidlin, Abrams,

Halabi, 2012). Surrogate endpoints are often physical symptoms of disease, biological markers, and results from radiological tests that can be measured precisely and in a shorter timeframe than clinical outcomes such as mortality. A reduction in power and inflation of the type I error can result when considering a particular effect of the treatment on a surrogate (e.g., progression-free survival [PFS]) as evidence for continuing the study to evaluate the treatment effect on a more clinically meaningful endpoint (e.g., survival). However, some authors propose that the use of non-survival endpoints in Phase II creates a natural progression for conducting definitive Phase III trials with delayed clinical endpoints (DCE) (Rubenstein, Korn, Freidlin, Hunsberger, 2005). We can consider the following situations where issues from such an approach might arise:

- The effect on the outcome of interest, e.g. survival, is not as strong as the effect on a surrogate due to deaths from other causes
- The treatment has some effect that is predictive of survival and some effect that is not predictive of survival, thereby contaminating and attenuating the effect
- There exists a combination of treatments - some treatments for which the surrogate is valid and some for which the surrogate is invalid

Because of such occurrences, a treatment that truly works will have a positive effect on the surrogate endpoint and an attenuated effect on survival. The first situation describes the issue of competing risks. Patients on either arm of the trial may experience death from a variety of causes due to comorbidities interfering with the ability to observe disease-specific events of primary interest. Power estimates are dependent on the appropriate derivation of treatment effect and incidence estimates based on components predictive of disease-specific and competing risk outcomes (Mell, Jeong, 2010). Ignoring competing risks will result in reduced power for the treatment effect that is considered a minimal clinically meaningful improvement.

A similar effect is true of treatments with some effect predictive of survival and some effect not predictive of survival. A trial designed for a treatment effect that is greater than the truth will be underpowered to detect the effect of interest. A valid surrogate endpoint is correlated with and predictive of the outcome of interest. Invalid surrogate markers may be correlated with the outcome of interest but often fail to capture the effect of the treatment on the outcome. The use of invalid surrogates leads to further study of treatments that show benefit for the surrogate but do not improve the outcome causing inflation of the type I error and further study of treatments that are ineffective. Surrogate endpoints can be validated with reliable trials that use both the surrogate and the DCE.

Illustrated in Figure 4.1 is a causal diagram for the relationships between the disease, the treatment, and the various outcomes with which we can assess the effect of the treatment. Our

ideal outcome is perfectly diagnostic of whether treatments are effective in treating or curing the disease. However, we can consider that there may exist treatments for which the outcome either lacks specificity, sensitivity, or both. Within the causal pathway between the disease and our DCE, survival is an intermediate stage such as true progression that is somewhat diagnostic of the effect of the treatment on survival. Treatments that act on the pathway between the disease and true progression are true positives (effective) because true progression is in the causal pathway between the disease and survival. Treatments that act on survival but not on true progression are false negatives because a test of the treatment on the intermediate endpoint would declare the treatment ineffective. There also exists an ISE such as measured progression that is not in the causal pathway. Therefore a treatment can have a positive effect on this outcome but not affect survival. In other words, some treatments only treat the symptom (Phase II endpoint of measured progression) and not the disease (Phase III endpoint of survival). True negatives (ineffective) are not represented here because they do not act on any of the outcomes considered. We acknowledge that primary endpoints such as overall survival can, like surrogate endpoints, lack sensitivity and specificity for diagnosing the effectiveness of a therapy to treat or cure a disease.

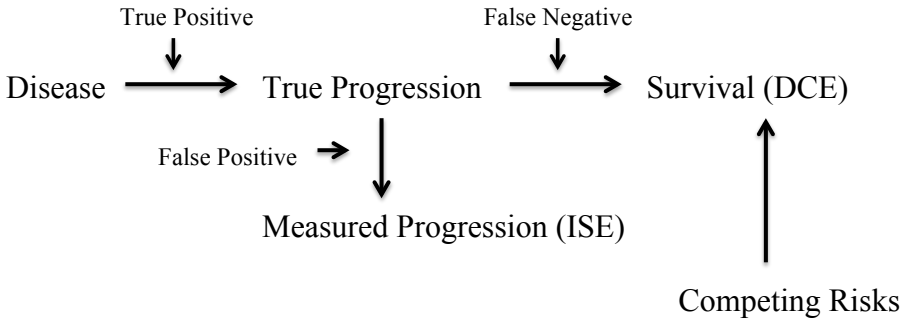


Figure 4.1: Causal diagram for the detected effect of the treatment on immediate and delayed clinical endpoints

To examine these effects, we present notation in Table 4.2 for the prevalence of ineffective treatments (true negative; TN and false positive; FP) and effective treatments (true positive; TP and false negative; FN) and the probabilities that treatments will pass each phase of the investigation under the primary endpoint, overall survival (OS) or a surrogate endpoint such as progression-free survival (PFS). We denote the prevalence of false positive treatments that affect the ISE but not the DCE as γ and the prevalence of false negative treatments that affect the DCE but not the ISE as δ. In designs previously considered, we assumed that the primary endpoint was used in both phases of the investigation such that ineffective and effective treatments were correctly identified according to the operating characteristics of the study design. In those cases, the endpoint did not lack sensitivity or specificity ($\gamma = \delta = 0$). Thus, the probabilities that ineffective and ef-

fective treatments pass Phase II and Phase III are α_2 and β_2 and α_3 and β_3, respectively (Table 4.2).

When using a surrogate endpoint in Phase II, the outcome can be non-specific for some treatments such that they are considered effective when they are truly ineffective (FP) because they have a positive effect on PFS but not on survival. For all treatments considered effective by the surrogate endpoint in Phase II (TP and FP treatments), we have additional power to detect the outcome than we do for the OS endpoint due to competing risks. Therefore, we will arbitrarily define the probability that such treatments pass Phase II as $1.1\beta_2$, a relative 10% increase in power with the surrogate endpoint chosen for purposes of illustration. In the same setting, the outcome can be insensitive for some treatments such that they are considered ineffective when they are truly effective (FN). For all treatments considered ineffective by the surrogate endpoint in Phase II (TN and FN treatments), the probability that they pass Phase II is the type I error rate of α_2. With the use of the OS endpoint in Phase III, the probabilities that effective and ineffective treatments pass are defined by the Phase III operating characteristics (α_3, β_3). We will examine selected values for $\gamma = (0, 0.05, 0.1, 0.2)$ to represent the proportion of ineffective treatments that treat the symptoms and not the disease and for $\delta = (0, 0.1\pi)$ to represent the proportion of effective treatments that do not affect the surrogate outcome. We will illustrate the effect of the use of surrogate outcomes in Phase II on PPV and efficiency of our overall treatment discovery process.

We will introduce the use of a surrogate endpoint in Phase II to an approach previously presented in Table 2.3. The design that will be considered is one which uses 1,000,000 subjects, starting with 2,000 Phase II studies, leading to 319 Phase III trials, and resulting in 162 effective and 3.8 ineffective treatments adopted after Phase III. Again, we assumed that the primary endpoint was used in both phases of the investigation such that ineffective and effective treatments were correctly identified according to the operating characteristics of the study design ($\gamma = \delta = 0$).

- If we consider a perfectly valid surrogate such that it does not lack specificity or sensitivity for correctly diagnosing the effect of treatments on the DCE, we have more power ($1.1\beta_2 = 1.1 \cdot 83 = 91\%$) at Phase II with the surrogate endpoint and $\gamma = \delta = 0$ remains to be true.
- With more power at Phase II, we pass more effective treatments on to Phase III and adopt 180 effective and 3.8 ineffective treatments with an overall PPV of 98% (Table 4.3, 4.4, 4.5).

Therefore, with a perfectly valid surrogate, we have more power to detect effective therapies that will go on to improve our Phase II survival outcome. Higher power at Phase II requires that we conduct more Phase III studies increasing the number of subjects necessary for the entire process to more than 1,000,000. Tables 4.3 – 4.5 also present expectations when the resources are restricted to 1,000,000 patients.

Table 4.2: Effect of the use of surrogate endpoints in Phase II on the operating characteristics of both phases of the investigation

	Using OS endpoint for entire approach		
Treatments	Prev	Pr(pass Phase II)	Pr(pass Phase III)
Ineffective Treatments			
True Negative	$1-\gamma-\pi$	α_2	α_3
False Positive	γ	α_2	α_3
Effective Treatments			
True Positive	$\pi-\delta$	β_2	β_3
False Negative	δ	β_2	β_3
	Using Surrogate at Phase II		
Treatments	Prev	Pr(pass Phase II)	Pr(pass Phase III)
Ineffective Treatments			
True Negative	$1-\gamma-\pi$	α_2	α_3
False Positive	γ	$1.1\beta_2$	α_3
Effective Treatments			
True Positive	$\pi-\delta$	$1.1\beta_2$	β_3
False Negative	δ	α_2	β_3

Now considering $\gamma \neq 0$, we can examine the effects of having a surrogate outcome that is non-specific to some proportion of the ineffective treatments that exist. With $\pi = 10\%$ prevalence of truly effective treatments and γ equal to some proportion of the 90% of truly ineffective treatments, the number of ineffective treatments that exist decreases. However, with 91% power to detect treatments that are efficacious on our surrogate and not effective for survival, we pass more truly ineffective treatments from Phase II to Phase III with higher values of γ.

We can also consider that our surrogate is insensitive for some proportion of the effective treatments such that the treatments are effective in improving survival but do not have a positive effect on the surrogate endpoint. This decreases the number of effective treatments identified because a proportion of them are not capable of being captured by the surrogate endpoint. With surrogate endpoints that are valid for identifying effective treatments, we can increase the number identified by as much as 11%. If our surrogate lacks specificity then we may end up adopting many more ineffective treatments.

Scaling these approaches back down to our fixed amount of resources of 1,000,000 subjects, we identify an approach that is comparable to our reference design adopting 162 effective and 3.8 ineffective treatments.

- Starting with 1,821 treatments in Phase II, we adopt 164 effective (TP), 3.3 ineffective (TN) and 2.1 false positive treatments when $\gamma = 0.05$.
- We increase the number of effective treatments identified by about 1% but also the number of ineffective treatments identified by 42%.
- As the number of treatments that our surrogate is invalid for increases, our PPV and the number of effective treatments adopted decrease.

The considerations presented here are just a few ways in which immediate surrogate endpoints can fail to provide reliable evidence about the benefit of interventions on clinically meaning endpoints. The separate Phase II to Phase III approach discussed in chapter 2 achieves a high PPV, a high number of effective treatments adopted, high power, low type I error, and credible statistical evidence, providing the clinically relevant endpoint is used at each phase. We also demonstrated an improvement in efficiency through the use of sequential sampling methods in each phase. If we instead choose to combine these phases, we are obscuring the roles of the screening and confirmatory phases of the investigation and potentially compromising our results by incorporating possibly biased data from the interim stage into our final analysis. This bias of our treatment effect estimate common to sequential sampling can be addressed with adjustment methods, however; we should be necessarily cautious when reporting estimates or powering subsequent studies for estimates obtained from a trial stopped early. The bias of preliminary analyses is discussed in section 5.1.1 and

adjustment methods are discussed in section 5.1.3.

With respect to surrogate endpoints, many authors propose validation of surrogates to assess if they reliably predict effects on clinically meaningful endpoints. Fleming and Powers (2012) present several clinical trial examples of the use of biomarkers as replacement endpoints and discuss where they fail to provide reliable evidence of the efficacy and safety of a treatment. Although surrogate endpoints can reduce the size and duration of our experiments, this might come at the risk of a valid and reliable evaluation of the intervention. When a surrogate is used as the primary endpoint in Phase II screening studies, it may provide evidence about the effect of a treatment on the pathway that hopefully leads to a true clinical benefit (Fleming, 2005). However, we have demonstrated that with invalid surrogates, either more resources are necessary for identifying treatments or fewer treatments can be studied at the cost of a lower PPV when such endpoints lack specificity or sensitivity to the effects of particular treatments. Of importance is to obtain reliable evidence for the effect of an intervention requiring that evidence be gathered on outcome measures that are representative of true benefit to the population.

Table 4.3: Examples of Phase II outcomes for screening trials using surrogate endpoints

γ	δ	Num of RCTs	N_2	Type I err; Pwr	TP + FN = Effective	TN + FP = Ineffective	PPV
0	0	2,000	345	0.085; 83%	166 + 0 = 166	153 + 0 = 153	52%
Starting with 2,000 treatments							
0	0	2,000	345	0.085; 91%	182 + 0 = 182	153 + 0 = 153	54%
0.05	0	2,000	345	0.085; 91%	182 + 0 = 182	145 + 91 = 236	44%
0.1	0	2,000	345	0.085; 91%	182 + 0 = 182	136 + 182 = 318	36%
0.2	0	2,000	345	0.085; 91%	182 + 0 = 182	119 + 364 = 483	27%
0	0.1π	2,000	345	0.085; 91%	164 + 1.7 = 165.7	153 + 0 = 153	52%
0.05	0.1π	2,000	345	0.085; 91%	164 + 1.7 = 165.7	145 + 91 = 236	41%
0.1	0.1π	2,000	345	0.085; 91%	164 + 1.7 = 165.7	136 + 182 = 318	34%
0.2	0.1π	2,000	345	0.085; 91%	164 + 1.7 = 165.7	119 + 364 = 483	26%
Using 1,000,000 subjects							
0	0	1,966	345	0.085; 91%	179 + 0 = 179	150 + 0 = 150	54%
0.05	0	1,821	345	0.085; 91%	166 + 0 = 166	132 + 83 = 215	44%
0.1	0	1,697	345	0.085; 91%	154 + 0 = 154	115 + 154 = 269	36%
0.2	0	1,494	345	0.085; 91%	136 + 0 = 136	89 + 272 = 361	29%
0	0.1π	2,000	345	0.085; 91%	164 + 1.7 = 165.7	153 + 0 = 153	52%
0.05	0.1π	1,847	345	0.085; 91%	151 + 1.6 = 153.6	133 + 84 = 217	41%
0.1	0.1π	1,720	345	0.085; 91%	141 + 1.5 = 142.5	117 + 157 = 274	34%
0.2	0.1π	1,511	345	0.085; 91%	124 + 1.3 = 125.3	90 + 275 = 365	25%

Table 4.4: Examples of Phase III outcomes for screening trials using surrogate endpoints

γ	δ	Num of RCTs	N_3	Type I err; Pwr	TP + FN = Effective	TN + FP = Ineffective
0	0	319	974	0.025; 99%	162 + 0 = 162	3.8 + 0 = 3.8
Starting with 2,000 treatments						
0	0	335	974	0.025; 99%	180 + 0 = 180	3.8 + 0 = 3.8
0.05	0	418	974	0.025; 99%	180 + 0 = 180	3.6 + 2.3 = 5.9
0.1	0	500	974	0.025; 99%	180 + 0 = 180	3.4 + 4.6 = 8.0
0.2	0	665	974	0.025; 99%	180 + 0 = 180	3.0 + 9.1 = 12.1
0	0.1π	319	974	0.025; 99%	162 + 1.7 = 163.7	3.8 + 0 = 3.8
0.05	0.1π	402	974	0.025; 99%	162 + 1.7 = 163.7	3.6 + 2.3 = 5.9
0.1	0.1π	484	974	0.025; 99%	162 + 1.7 = 163.7	3.4 + 4.6 = 8.0
0.2	0.1π	649	974	0.025; 99%	162 + 1.7 = 163.7	2.7 + 8.2 = 10.9
Using 1,000,000 subjects						
0	0	329	974	0.025; 99%	177 + 0 = 177	3.8 + 0 = 3.8
0.05	0	381	974	0.025; 99%	164 + 0 = 164	3.3 + 2.1 = 5.4
0.1	0	423	974	0.025; 99%	152 + 0 = 152	2.9 + 3.9 = 6.8
0.2	0	497	974	0.025; 99%	124 + 0 = 124	2.2 + 6.8 = 9.0
0	0.1π	319	974	0.025; 99%	162 + 1.7 = 163.7	3.8 + 0 = 3.8
0.05	0.1π	371	974	0.025; 99%	149 + 1.6 = 151.6	3.3 + 2.1 = 5.4
0.1	0.1π	417	974	0.025; 99%	140 + 1.4 = 141.4	2.9 + 3.9 = 6.8
0.2	0.1π	491	974	0.025; 99%	123 + 1.3 = 124.3	2.3 + 6.9 = 9.2

Table 4.5: Examples of overall outcomes for screening trials using surrogate endpoints

γ	δ	Effective	Ineffective	PPV	Type I err; Pwr	N_{total}
0	0	162	3.8	98%	0.0021; 81%	1,000,000
Starting with 2,000 treatments						
0	0	180	3.8	98%	0.0021; 90%	1,016,290
0.05	0	180	5.9	97%	0.0033; 90%	1,097,132
0.1	0	180	8.0	96%	0.0044; 90%	1,177,000
0.2	0	180	12.1	94%	0.0067; 90%	1,337,710
0	0.1π	162	3.8	98%	0.0031; 82%	1,000,706
0.05	0.1π	162	5.9	97%	0.0033; 82%	1,081,548
0.1	0.1π	162	8.0	95%	0.0044; 82%	1,161,416
0.2	0.1π	162	10.9	94%	0.0061; 82%	1,322,126
Using 1,000,000 subjects						
0	0	177	3.8	98%	0.0021; 90%	1,000,000
0.05	0	164	5.4	97%	0.0033; 90%	1,000,000
0.1	0	152	6.8	96%	0.0044; 90%	1,000,000
0.2	0	124	9.0	93%	0.0067; 90%	1,000,000
0	0.1π	163.7	3.8	98%	0.0031; 82%	1,000,000
0.05	0.1π	151.6	5.4	96%	0.0033; 82%	1,000,000
0.1	0.1π	141.4	6.8	95%	0.0044; 82%	1,000,000
0.2	0.1π	124.3	9.2	93%	0.0067; 82%	1,000,000

Chapter 5

Adaptations to the Trial Design

In previous chapters, we examined optimality criteria parameterized by the operating characteristics at Phase II emphasizing the relationships between frequentist and Bayesian criteria in the treatment discovery process. Our goal is to identify an approach to adopt a maximal number of effective treatments in a setting with limited resources such that a high Bayesian positive predictive value is attained. In chapter 3, we introduced non-adaptive sequential designs within each phase for improved efficiency in our approach. We considered non-adaptive seamless Phase II/III designs in chapter 4, where pre-planned analyses involve the switching of outcomes between phases. We also examined the characterization of the Phase II/III progression as a single sequential design. We now turn our attention to exploring how the traditional approach of adapting hypotheses at the end of Phase II fits in with the newer adaptive methods.

Traditionally, the adaptive process is discrete. The Phase II study is conducted according to a pre-specified plan (perhaps sequential, but non-adaptive). During the "white space", Phase II data is thoroughly analyzed and reviewed by regulatory authorities and the scientific community, and it is during this white space that the hypotheses are traditionally modified. A Phase III study is then planned, which may have different eligibility criteria, treatment regimens, and outcomes than were used in the Phase II study. The change from a surrogate outcome to a more clinically relevant outcome may have been pre-planned, however, changes in the eligibility criteria, treatment, and outcomes are most often driven by the Phase II results. Typically, there is no formal means of controlling the type I error during such "data dredging." The Phase III trial design is then reviewed by regulatory authorities and IRBs prior to trial conduct (which might involve a sequential stopping rule).

Adaptive clinical trial design, which by our definition involves non-GSD adaptations that are based on unblinded analyses of the estimated treatment effect, has been proposed as an approach that can improve the efficiency of the treatment discovery process. Modifications of particular

interest involve adaptive changes to the

- sample size (a re-powering of the study),
- eligibility criteria ("enrichment" of the sample to include only those types of patients who appear to be getting the most benefit),
- treatment regimens (dose finding), and
- changing of outcomes (e.g., changing the clinical endpoint from OS to PFS, or changing the summarization of the distribution from 5 year survival to a hazard ratio [HR]).

Whether these adaptations are made following Phase II to determine how to proceed with Phase III or whether they are introduced at each interim analysis of a single study, we must consider how these revisions affect future results and inference. From the results of our previous chapters, it should be clear that consideration of type I error and power will be important. Rich literature exists on how to protect the type I error while making such adaptations within a trial, though it is not clear how they should be implemented to best gain efficiency. Furthermore, there are concerns that unblinded adaptation during a trial may lead to less credible results, because the true impact of such methods is not well-understood.

In this chapter, we explore the impact of adaptations from Phase II to Phase III in the traditional setting. While this approach does not capture the seamless Phase II/III approach exactly, the fact that it relies on data from completed trials means that we ought to have more reliable behavior. The relevance of the results of this chapter to the seamless Phase II/III will then be discussed in chapter 6 along with ideas on how the methods for adaptation within a trial might be implemented to achieve better efficiency.

5.1 Adapting sample size

5.1.1 Bias of Phase II results and the need for confirmatory studies

As previously discussed, the primary objective of Phase II studies is to identify those treatments that are not of interest to study further in large, confirmatory Phase III trials. In general, screening studies are smaller and results tend to vary from what is subsequently observed in larger trials. Within cancer research, Phase II studies are often non-randomized. In other areas of research, RCTs predominate. With a wider use of randomized Phase II trials for selecting treatments, the reliability of such studies increases, however, results cannot stand alone without a definitive experiment. The planning of the requisite follow-on study is then influenced by any bias that results from the selection of the most promising results.

Bias in the estimate of the treatment effect is a result of the lack of precision of small samples inherent in Phase II studies (Liu, LeBlanc, Desai, 1999). If Phase II results were unbiased and sufficiently precise, there would be no need for Phase III comparisons. However, that is not the case.

Publication and selection biases are unavoidable when there is a delay in the publication of unfavorable results or when promising treatments show benefit for only a selection of the broader population of interest. Many investigators have classified, quantified, and documented the bias in clinical research through reports of the number of studies approved versus the number ultimately published. A review of 218 studies analyzed with tests of significance showed that trials with positive results were 2.32 times more likely to be published than negative studies (95% CI 1.47–3.66), with a significantly shorter time to publication (median 4.8 vs 8.0 years) (Stern 1997). The absence of results from non-significant trials in the clinical literature leads to an unrealistic confidence in experimental interventions. To address this issue, the International Committee of Medical Journal Editors (ICMJE) now requires clinical trial registration as a condition of consideration for publication. Selection bias is inherent of treatment discovery: for those invested in a particular agent, finding the population for which the treatment works is a major goal of the investigation. Searching through subgroups is a form of selection bias that affects the planning and implementation of future studies. Issues with the implementation of subgroup analyses is discussed further in section 5.2. In this section, we focus on the role that "random high bias" and restricted attention to highly selected trials might play in the accurate prediction of truly beneficial treatments.

In the absence of a true treatment effect, all treatments under study in Phase II have equal chance of being the "most promising." As Liu, LeBlanc, and Desai (1999) discussed, the chance that a candidate treatment will appear superior to control is $(A_2-1)/A_2$, or 0.50 when $A_2 = 2$, 0.67 when $A_2 = 3$, and 0.75 when $A_2 = 4$ where A_2 is the number of treatment arms under study. To prevent the scientific community from considering Phase II results conclusive evidence, the authors propose a simple Phase II trial design that consists of using non-randomized Phase II studies without comparing treatments to a control arm, performing hypothesis tests, or presenting p-values. A simple Phase II trial design will limit the ability to draw erroneous inferences requiring the conduct of a confirmatory Phase III comparison where statistical error rates are properly controlled (Liu, LeBlanc, Desai, 1999). Therefore, it is important that we consider how to design the next stage of the investigation based on potentially biased preliminary findings. We will examine how believing Phase II results and re-powering Phase III for the treatment effect observed in Phase II changes the outcomes of the treatment discovery process.

5.1.2 Powering Phase III based on Phase II results

Given that we plan to conduct a Phase III study, we must determine the sample size necessary for detecting the treatment effect of interest. It seems logical to base the design of Phase III on what we observed in Phase II however, as we discuss below, this can lead to inefficiencies in our approach. There are various strategies to consider. In chapter 2, we took an approach that considered the MCID, and we powered our Phase II trial for that important difference. Then, despite the estimates of treatment effect obtained in Phase II, we continued to power for that effect when studying promising treatments in Phase III. An alternative strategy would be to power Phase III studies for the treatment effect estimate (either frequentist or Bayesian) observed in Phase II.

We will consider alternative strategies to Approach 2b previously presented in Table 2.1. Starting with 2,047 Phase II studies of 342 subjects each, after Phase II we pass 174 effective treatments when powered at 84.9% and 184 ineffective treatments with a type I error rate of 0.1. After 358 Phase III trials powered at 95.0% with a type I error rate of 0.025, we adopt 165 effective treatments and 5 ineffective treatments. We start with our simple hypothesis for conducting Phase II trials and adjust the sample size in Phase III to retain our desired power of 95% to detect the estimated treatment effect obtained in Phase II.

Based on the results of 100,000 simulations, powering Phase III for our Phase II treatment effect estimate results in a sample size for Phase III ranging from 126 to 2,712 subjects.

- Under the null, a significant Phase II result suggests that the point estimate is tending towards the critical value of 0.069 when the MCID is 0.125 requiring larger Phase III sample sizes to detect small treatment effects. The sample size required to detect an effect of 0.069 with 95% power is 2,712 subjects. The average sample size required at Phase III for studying treatments with a true effect of 0 is 1,644 subjects.
- Under the alternative, the expected effect among the significant trials of truly effective treatments is 0.140. The sample size required to detect an effect of 0.140 with 95% power is 660 subjects. The sample size required to detect the largest estimated effect of 0.321 with 95% power is 126 subjects.
- This approach reduces the number of Phase II trials conducted (1,772 vs 2,047) because more of the fixed resources are required at Phase III with an average Phase III sample size of 1,272 subjects.
- Because fewer treatments are studied initially, fewer are studied further in Phase III and fewer treatments are adopted overall (127 vs 165) (Table 5.1).
- A PPV overall of 96.6% is attained but overall power is reduced (81 vs 72%).

We thus consider re-powering for an effect obtained in Phase II an "optimistic" approach for both truly effective and ineffective treatments for the following reasons. At the end of Phase II, 49% of the treatments moving forward are effective with a true treatment effect that is the MCID ($\mu = 0.125$) and 51% are truly ineffective ($\mu = 0$). However, all have an estimated effect at Phase II above the critical value ($\mu \geq 0.069$). Under the null, the estimate of the effect for significant treatments will be close to the critical value. Of the 159 ineffective (false positive; FP) treatments passed after Phase II, 142 (89.6%) of the treatments were passed with an estimated effect between the critical value and the MCID. This means that 89.6% of the ineffective treatments passed after Phase II will require a Phase III sample size larger than the 829 subjects we planned for when continuing to power for the MCID. Choosing to power our Phase III study for an effect obtained in Phase II that is close to the critical value requires a much larger sample size at Phase III making our approach inefficient. We are being optimistic in believing that the effects of truly ineffective treatments will be significant if we simply increase our sample size. In truth, we are using more resources to find that only 2.5% will be significant after Phase III.

Under the alternative, we will obtain an effect that is biased upward. Of the 150 effective treatments (true positive; TP) passed after Phase II, 89 (59%) of the treatments were passed with an estimated effect greater than the MCID. Therefore, 59% of the effective treatments passed after Phase II will require a Phase III sample size smaller than the 829 subjects we planned for when continuing to power for the MCID. The median sample size required at Phase III for studying treatments with a true effect of 0.125 is 708 subjects. In choosing to power our Phase III study for an estimated effect greater than the MCID obtained in Phase II when the true effect is the MCID, we will have too small a sample size, and we will fail to achieve the 95% power of our Phase III trial design. For both truly effective and ineffective interventions, the results of our Phase II studies are suggestive of larger treatments effects than those that truly exist. Therefore, we are optimistically moving forward with the treatments to the next phase of the investigation. When we consider the amount of resources spent on identifying treatments, too few subjects are used for studying truly effective treatments and too many for studying ineffective treatments.

Scientifically, it is important to consider the MCID when powering trials because detecting differences less than the MCID does not result in a meaningful improvement in the health of the population. Statistically, we should consider the treatment effect necessary to obtain a significant result and may choose to power studies such that the MCID is the critical value for significance. If we power Phase III studies for the effect that was considered an MCID at the start of Phase II, we adopt more effective treatments with the same amount of resources. However, if it is desirable to power Phase III for the effect obtained in Phase II, it seems logical to consider adjusted methods to account for the bias of early estimates. Adjustment methods will be discussed in section 5.1.3.

Table 5.1: Examples of various approaches to powering Phase III studies including the approach of powering for the effect used to power the Phase II study and an "optimistic" approach of re-powering for an effect obtain in Phase II for both a simple (binary) prior and a continuous (mixture) prior

	Hypothesis	Approach 2b Binary Prior	Optimistic Binary Prior	Approach 2b Continuous Prior	Optimistic Continuous Prior
Phase II	Number RCT	2,047 (10% eff)	1,772 (10% eff)	2,014 (10% eff)	1,754 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 85%	0.100; 85%	0.100; 78%	0.100; 80%
	"Positive" RCT	174 eff; 184 not	150 eff; 159 not	157 eff; 183 not	141 eff; 157 not
Phase III	Number RCT	358 (49% eff)	309 (49% eff)	340 (46% eff)	298 (47% eff)
	N per RCT	829	126 – 2,712	829	111 – 2,776
	Type I err; Pwr	0.025; 95%	0.025; 85%	0.025; 92%	0.025; 84%
	# Effctve Adopt	165	127	145	118
	# Ineff Adopt	5	4	5	4
Overall	PPV	97.3%	96.6%	96.9%	96.6%
	Type I err; Pwr	0.0025; 81%	0.0025; 72%	0.0025; 72%	0.0026; 67%

We recognize that many investigators do not power the study adequately to detect the MCID, but their constraints on resources should argue that the same principles hold relative to the effect that was feasible to detect. Sequential analyses can ensure that futile trials are not continued needlessly.

Now we will consider our continuous (mixture) prior to assess differences in the outcome of our approach. We can study approximately 2,000 candidate treatments as in Approach 2b (Table 2.1) where there exists a range of positive effects among the treatments that truly work. We design our Phase II trials to be powered at 85% to detect the design alternative meaning that we have less power to detect treatment effects smaller than the MCID and more powered to detect larger differences. However, there is a loss of power overall (78 vs 85%). This results in a slight reduction in the number of effective treatments identified after Phase II. Again, our power at Phase III is reduced for finding effective treatments among those with some positive effect. Overall, we adopt fewer treatments (145 vs 165) but our PPV and overall type I error is comparable. We can again consider powering Phase III for the treatment effect estimate obtained in Phase II. We see the same trends of fewer treatments studied in Phase II and Phase III, more resources used in Phase III and fewer treatments adopted overall. With this multinomial treatment effect distribution, we should consider the number of treatments that are considered significant with effects less than the MCID. The loss of power is not worth powering Phase III trials for the estimated treatment effect obtained in Phase II.

Similar trends exist among these various approaches however, powering for the MCID in all phases of the investigation seems preferable and results in a more statistically reliable and scientifically meaningful approach. A key issue is the bias of our Phase II results and whether that causes issues with our overall approach. In the next section, we consider how bias adjusted estimates of the treatment effect in preliminary analyses might ameliorate such concerns.

5.1.3 Solutions for biased estimates

We previously demonstrated that powering our Phase III studies based on Phase II estimates leads to less efficient identification of treatments. This is largely a function of the bias that results from screening trials – an issue closely related to publication bias. To further document this concern, we consider explorations of the bias from estimates obtained from GSDs stopped early in this section.

Several authors have proposed that the upward bias of treatment effect estimates from GSDs stopped early exaggerate the true benefit of the treatment misleading clinicians and patients (Fan, DeMets, & Gordon, 2007; Bassler et al., 2008; Freidlin & Korn, 2009; Bassler et al., 2010; Zhang et al., 2012). To assess this bias, Freidlin and Korn (2009) compared this inflation of the treatment effect estimator when a trial is stopped early with the inflation seen in a comparable set of trials with fixed sample sizes via simulation. They discuss several different approaches for defining a fixed sample-size trial estimator for making comparisons. The comparator they chose considered, for each possible interim analysis, the probability of stopping at that point and compared that upper quantile of results to the results in the same quantile that would have been obtained if that trial were allowed to go to completion. They concluded that trials stopped early show greater deviation from the truth, but only at very early interim analyses ($\leq$ 25% information) is this inflation of concern. Although statistical methods exist to adjust for this bias, they are not typically used and were not considered by Freidlin and Korn.

Similarly, in order to explore factors associated with overestimates of effect, Bassler et al. compared treatment effects from truncated randomized clinical trials (RCTs) with that from a meta-analysis of RCTs addressing the same question but not stopped early. They compared the relative risk (RR) generated by truncated RCTs with a pooled RR from matching non-truncated RCTs. They found that truncated RCTs were associated with greater effect sizes. Meta-regression analysis showed the effect was independent of the presence of statistical stopping rules but greatest in smaller studies. Our concerns with this analysis include their retrospective placement of trials into two groups and comparisons of trials with dissimilar effect sizes. They did not consider statistical methods of adjustment.

Fan, DeMets, and Lan investigate conditional and marginal biases and propose new conditional estimators to significantly reduce the conditional bias from unconditional estimators. They discuss the differences between conditional and marginal biases and show that the overall bias is a weighted average of all the conditional biases with weights being the probability of stopping at each interim analysis (Fan, DeMets, and Lan, 2004). The three conditional estimators proposed are the maximum conditional likelihood estimate (MLCE), the conditional moment estimate (CME), and the conditional bias reduced estimate (CBRE), all shown to be equivalent. The comparisons made are between unconditional estimators such as the maximum likelihood estimate (MLE), the bias adjusted mean (BAM) proposed by Whitehead in 1986, and the newly proposed conditional estimators. However, they do not consider comparisons between fixed sample-size trial designs and GSDs, nor do they consider inference under any other hypothesis than the alternative.

To assess these concerns, and to determine where methods of adjustment should be applied under a spectrum of true treatment effects, we will compare the bias of the treatment effect estimator of a trial stopped early for efficacy to the bias of the treatment effect estimator of a fixed sample-size trial with the same sample size and significance level as the interim analysis at which the GSD stopped early.

A Pocock trial design with one treatment arm and an interim monitoring plan of four equally spaced analyses is used to test the hypothesis of a greater alternative in mean response. We consider trials designed to detect a mean response greater than or equal to 0.23 with 97.5% power (one-sided type I error of 0.025) and monitoring boundaries for efficacy and futility at N = 100, 200, 300, and 400 subjects. We then consider four separate fixed sample-size trial designs with N = 100, 200, 300, and 400 subjects and significance levels that correspond to that at each interim analysis of their GSD comparators, $\alpha = 0.010$, 0.018, 0.023, and 0.025.

We will first focus on the bias adjusted mean (BAM) as an adjusted estimate for treatment effect obtained by solving the following in θ

$$\theta = \hat{\theta} - b(\hat{\theta}), \tag{5.1}$$

where $\hat{\theta}$ is the MLE and $b(\hat{\theta})$ is the marginal bias discussed previously as a measure involving the conditional biases and the probability of stopping at each interim analysis.

Presented in Table 5.2 below are the expected estimated bias adjusted mean responses conditional on stopping at a given interim analysis along with fixed sample-size trial estimators both as a function of the true mean response. The expected estimated BAMs were calculated via a total of 100,000 simulation runs. The stopping probabilities at each interim analysis are also listed in

the table. We observe that the conditional estimators from both trial designs are biased however, an inflation of the true treatment effect is not true for estimates at every analysis. Conditional on stopping at the first analysis, as the true treatment effect becomes more extremely positive, the conditional estimate for that effect is less biased. The fixed sample-size estimator from a study with that same sample size is more biased than the conditional estimate. As we move to stopping at future analyses, the conditional estimators of the GSD trials become negatively biased for extremely positive true treatment effects while those from the fixed sample-size trials converge to the true effect size. There exists some point at which the estimators switch roles in terms of which is more biased.

To better reflect these relationships, plotted in Figure 5.1 is the true treatment effect against the bias of the BAM estimators from both the GSD trials and the fixed sample-size designed trials. Here, bias is calculated for the fixed sample design as the difference between our estimate $\hat{\theta}$ and the true treatment effect θ,

$$b(\hat{\theta}) = E(\hat{\theta}|p \leq \alpha) - \theta \tag{5.2}$$

where $\alpha = 0.010$, 0.018, 0.023, or 0.025. And for the GSD as

$$b(\hat{\theta}) = E(\hat{\theta}|\eta = j) - \theta \tag{5.3}$$

for stopping stage η where j = 1, 2, 3, or 4.

Estimates from a fixed sample-size trial with N = 100 and $\alpha = 0.01$, as is for the first interim analysis of the GSD trial, are always more biased than the conditional estimator comparators. The same is true for the second interim analysis comparison until we get to large treatment effects. The difference between biases of the two estimators is greatest at the third and fourth interim analyses. For extremely positive treatment effects, conditional estimators do not do as well as their fixed sample-size comparators. However, conditional estimates are conservative and do not exaggerate the true treatment effect at those later analyses. If we were going to regard a smaller study with a moderate treatment effect compelling, we do better when that smaller sample size is a result of a well-planned sequential stopping rule than when the small sample size was fixed in advance. While smaller sample sizes always provide less information than larger sample sizes, group sequential designs are no more misleading in terms of their estimated treatment effect than fixed sample-size designs.

Thus far, we have only considered the BAM. We can also consider the maximum likelihood estimate (MLE), the Fan, DeMets and Lan (2004) maximum conditional likelihood estimate (MCLE),

Table 5.2: Expected estimated bias adjusted mean responses (BAM) and confidence intervals (CI) for trials stopped early for efficacy and for fixed sample-size trials of the same sample size for various true treatment effects

		Expected estimated BAM (97.5% CI)	
True mean response	Fraction of trials stopped (%)	1st interim	Fixed (N=100)
0.05	3.4	0.250 (0.2501, 0.2506)	0.273 (0.2724, 0.2728)
0.10	9.0	0.258 (0.2576, 0.2581)	0.279 (0.2789, 0.2794)
0.15	20.5	0.268 (0.2680, 0.2686)	0.288 (0.2876, 0.2882)
0.20	36.9	0.283 (0.2824, 0.2831)	0.301 (0.3010, 0.3016)
0.23	48.6	0.294 (0.2931, 0.2940)	0.311 (0.3108, 0.3115)
		2nd interim	Fixed (N=200)
0.05	3.5	0.171 (0.1709, 0.1711)	0.181 (0.1808, 0.1812)
0.10	11.3	0.176 (0.1757, 0.1760)	0.190 (0.1903, 0.1907)
0.15	24.2	0.183 (0.1827, 0.1830)	0.205 (0.2051, 0.2056)
0.20	33.9	0.193 (0.1926, 0.1930)	0.228 (0.2276, 0.2283)
0.23	34.8	0.201 (0.2005, 0.2009)	0.246 (0.2460, 0.2467)
		3rd interim	Fixed (N=300)
0.05	3.4	0.142 (0.1420, 0.1422)	0.144 (0.1440, 0.1443)
0.10	11.0	0.146 (0.1457, 0.1459)	0.156 (0.1560, 0.1564)
0.15	18.4	0.151 (0.1506, 0.1508)	0.177 (0.1763, 0.1769)
0.20	16.4	0.157 (0.1570, 0.1573)	0.208 (0.2080, 0.2086)
0.23	11.3	0.162 (0.1622, 0.1625)	0.233 (0.2329, 0.2335)
		4th interim	Fixed (N=400)
0.05	2.1	0.127 (0.1273, 0.1274)	0.124 (0.1242, 0.1245)
0.10	6.7	0.130 (0.1299, 0.1300)	0.138 (0.1380, 0.1383)
0.15	8.5	0.134 (0.1339, 0.1340)	0.163 (0.1627, 0.1632)
0.20	4.7	0.139 (0.1392, 0.1394)	0.202 (0.2013, 0.2019)
0.23	2.1	0.143 (0.1428, 0.1430)	0.230 (0.2292, 0.2300)

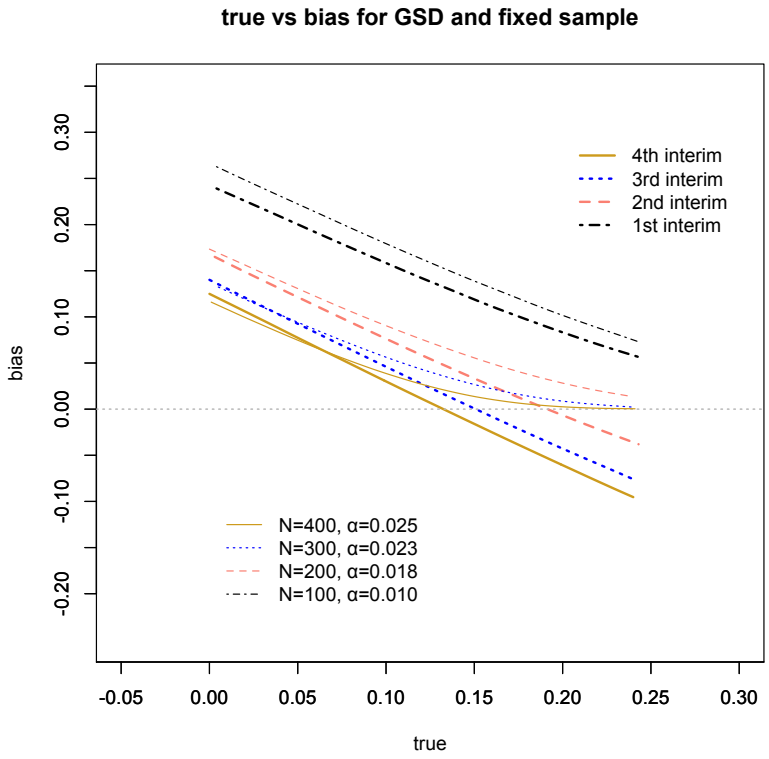


Figure 5.1: True treatment effect against the bias of the estimated effect at each interim and for fixed sample-size studies of the same sample size

and our adjustment to the BAM that we call the conditional bias reduced mean (cBRM) estimate. Fan, DeMets and Lan (2004) also present estimators equivalent to the MCLE that are the conditional moment estimate (CME) and the conditional bias reduction estimate (CBRE).

The MCLE, denoted by $\check{\theta}_{MCLE}$ maximizes the likelihood of statistic $(\eta, X_1, ..., X_n)$ conditioning on the stopping stage, η,

$$\check{\theta}_{MCLE} = \arg\max_{\theta} f_{\theta}(\eta, X_1, ..., X_n | \eta = j) \tag{5.4}$$

Denoted by $\check{\theta}_{CME}$, the CME is the solution of θ to the following conditional moment estimate equation,

$$E[(t_n, S_n)|\eta = j)] = (t_n, S_n) \tag{5.5}$$

The CBRE, $\check{\theta}_{CBRE}$ is the solution in θ to the following equation,

$$\theta = \hat{\theta} - b(\hat{\theta}|\eta) \tag{5.6}$$

where $\hat{\theta}$ is the MLE and $b(\hat{\theta}|\eta)$ is the bias conditional on stopping at stage η. Fan, DeMets, and Lan (2004) show that $\check{\theta}_{MCLE} = \check{\theta}_{CME} = \check{\theta}_{CBRE}$. For the purposes of comparison, we will refer to their conditional estimator as the MCLE. Our ad-hoc estimator, the cBRM $\ddot{\theta}$ is a function of the BAM $\tilde{\theta}$. The intent is to reduce the conditional bias of an estimator given the stopping stage, η. The cBRM is defined as

$$\ddot{\theta} = \begin{cases} a + \frac{\tilde{\theta}_b - a}{b - d_j}(\hat{\theta} - d_j), & \text{if } \hat{\theta} < b. \\ \tilde{\theta}, & \text{otherwise.} \end{cases} \tag{5.7}$$

where d_j is the stopping boundary at the j^{th} analysis, a is the boundary at the final interim analysis, b is the sum of a and d_j, and $\tilde{\theta}_b$ is the BAM evaluated at $\theta = b$. To better adjust our estimate as a result of stopping at a given interim analysis, we are taking into consideration the level of conservatism of the stopping boundary at that interim analysis as well as the boundary at the final interim analysis.

Shown in Figure 5.2 below is a plot of the true treatment effect against the bias of the MLE, BAM, MCLE, and cBRM from the GSD trials and the BAM from the fixed sample-size trials. The cBRM and MCLE do better in terms of bias compared to all other estimators at the first interim analysis. As a function of the BAM, we see the same pattern in terms of the cBRM being more biased than the fixed sample-size estimator for extremely positive treatment effects in later analyses. We continue to see similarly exaggerated effects for all estimators for modest true treatment effects and conservative estimates for conditional estimators for extremely positive true treatment effects.

It is inference at the first interim analysis that is of most concern with regards to inflating the treatment effect and stopping early for efficacy only to approve a treatment that is not truly beneficial. However, when comparing the conditional estimators to a fixed sample-size estimator from a small trial of the size of this early interim analysis, the conditional estimators inflate the true treatment effect much less on average. It is the lack of information at that stage that results in that amount of bias. If it is inappropriate to conduct such a small study, then it is also inappropriate to design a trial with a stopping rule at that amount of information. Although use of adjustment methods is advisable when reporting a point estimate reflective of the estimated treatment effect from a group sequential trial, caution must be taken when using that treatment effect to power a subsequent confirmatory trial as discussed in section 5.1.2. In conservatively adjusting a potentially exaggerated effect towards the null hypothesis, there is the possibility of adjusting such that the critical value for statistical significance is set outside the range of values that are considered clinically meaningful, which may be of concern.

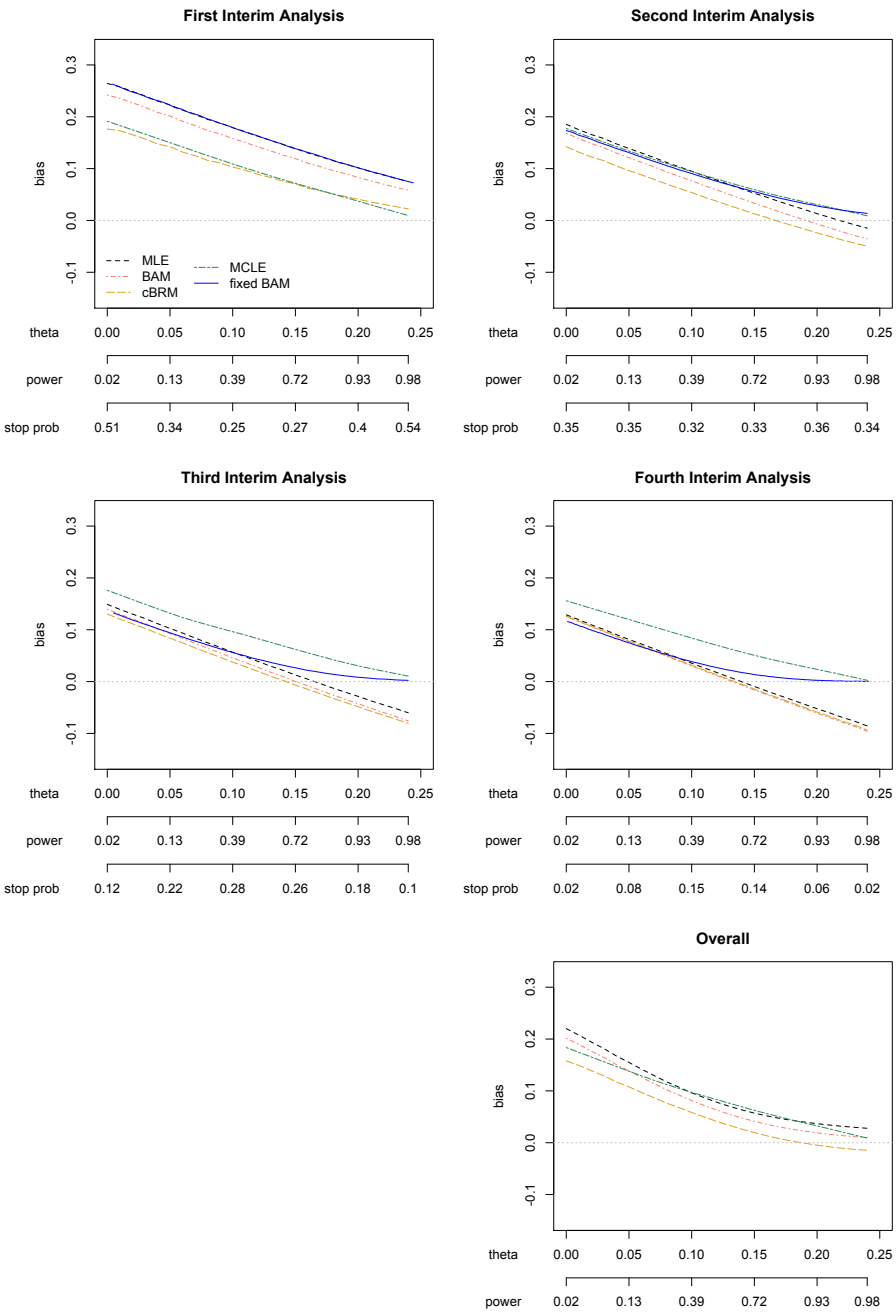


Figure 5.2: Pocock – True treatment effect and power to detect against the bias of the estimated effect of the conditional estimators including the maximum likelihood estimate (MLE), the bias adjusted mean (BAM), the maximum conditional likelihood estimate (MCLE), and the conditional bias reduced mean (cBRM). The dashed line indicates zero bias.

The previously considered trial design with one treatment arm and an interim monitoring plan of four equally spaced analyses was modified to instead reflect an O'Brien-Fleming sequential stopping approach. We consider trials designed to detect a mean response greater than or equal to 0.20 with 97.5% power (one-sided type I error of 0.025) and monitoring boundaries for efficacy and futility at N = 100, 200, 300, and 400 subjects. We then consider four separate fixed sample-size trial designs with N = 100, 200, 300, and 400 subjects and significance levels that correspond to that at each interim analysis of their GSD comparators, $\alpha = 0.010$, 0.018, 0.023, and 0.025.

Shown in Figure 5.3 below is a plot of the true treatment effect against the bias of the previously referred to conditional estimators from the GSD trials and the BAM from the fixed sample-size trials. Due to the more conservative nature of the OBF stopping rule, especially during the first interim analysis, stopping early for efficacy requires that we see an extreme treatment effect. This results in greater bias for the MLE and BAM during the first interim analysis compared to the Pocock trial designs. Bias is largest during the third and fourth interim analyses for instances where the stopping probability is essentially equal to zero. Thus, we should be most worried about biased estimates when we are not stopping to obtain estimates very often. The adjusted conditional estimators do quite well in terms of bias when considering occurrences that are frequent and expected to happen such as stopping during the third interim analysis with an estimated treatment effect close to the design alternative.

Bias adjusted estimators accommodate early tests and interim analyses and are typically more precise and more accurate than other estimators. An RCT design with an interim monitoring plan to allow for early stopping can be an ethically acceptable, statistically credible, and adequately efficient clinical trial design. Some authors discuss this bias of early estimates as a warning against early trial termination (Freidlin, Korn, 2009). However, as noted above, sequential trials with adjusted estimators are actually more robust than fixed sample-size trials of the same sample size. Furthermore, ethical concerns quickly arise, for instance in a situation where an identified treatment is shown to be superior but patients are continued to be randomized for the purpose of obtaining a more precise or longer-term estimate of the degree of its effect. This issue of a moderate upward bias of the treatment effect that can be ameliorated with existing methodology should not prohibit the use of sequential monitoring plans. Clinical trials should be designed such that results are credible and accepted by the scientific community as evidence for or against a candidate treatment. Prespecified sequential methods that preserve desired operating characteristics (type I error, power, PPV) remain viable means of investigation at either Phase II or Phase III.

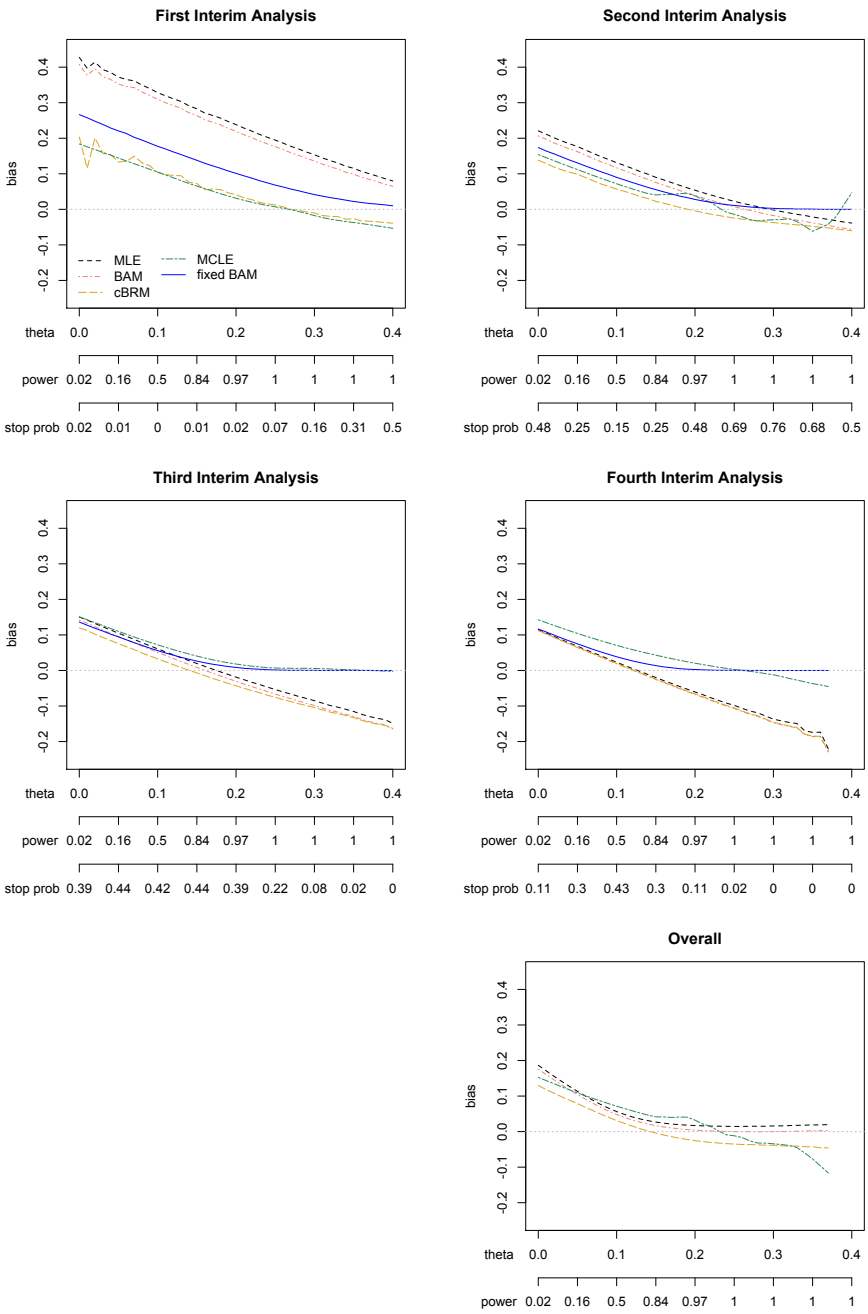


Figure 5.3: O'Brien-Fleming (OBF) – True treatment effect and power to detect against the bias of the estimated effect of the conditional estimators including the maximum likelihood estimate (MLE), the bias adjusted mean (BAM), the maximum conditional likelihood estimate (MCLE), and the conditional bias reduced mean (cBRM). The dashed line indicates zero bias.

5.1.4 Relevance to adaptive sample size

Powering of the Phase III study based on Phase II trial results mimics adaptive sample size re-estimation / re-powering of the study. We have examined the bias of preliminary results to consider the effects of such bias on further investigation of the treatment. Having demonstrated that proper bias adjustment can be beneficial in the settings of screening or publication bias, we still want to consider whether such adjusted analyses can be advantageous when re-powering a study between Phase II and Phase III (or at an interim analysis in a seamless Phase II/III design).

We chose design characteristics that were, in some sense, well-behaved for our binary prior distribution. At the end of Phase II, a PPV of approximately 50% and Figures 5.2 and 5.3 suggest that it will be impossible to provide an adaptive bias adjustment that would be appropriate for both a true effect of 0 and a true effect of 0.125. With continuous priors, the adjustment to consider would need to be more complicated and would depend on the prior. Overall progression of studies chosen at the start of Phase II should take into account the prior belief about the treatment effect, and the data we observed in conjunction with that prior is more or less pre-determined. Therefore, adaptive sample size re-estimation / re-powering of the study changes the progression and ultimately, the outcome of our treatment discovery process that we initially planned for based on prior belief about the treatment effect. It is instead appropriate to consider that, at a given stopping point of the investigation, a mixture of treatment effects will still exist as a subset of those at the start. The best approach is one that was, at the start of Phase II, believed to be appropriately chosen for that particular setting.

With a seamless Phase II/III design, re-powering our study after what is considered Phase II might be comparable to stopping for an adaptation at the second interim analysis. As we have seen, the estimate of the treatment effect at the second interim analysis continues to be extremely biased. Removing bias from estimates of the effect for ineffective treatments at such an early stage of the investigation would cause us to abandon the follow-on study for treatments with effects adjusted below the critical value. This is essentially altering the type I error and power of our Phase II trial design. Effects above the critical value after a downward adjustment would require an even larger sample size at Phase III. Removing bias from estimates of the effect for effective treatments would cause us to retain a larger sample size at Phase III, which is what was initially recommended. While it is of interest to consider the advantages of bias adjustment, it remains true that powering for the MCID in all phases of the investigation seems preferable and results in a more efficient use of resources.

Levin, S.C. Emerson, and S.S. Emerson (2012) found that optimal pre-specified adaptive designs with inference based on the minimal sufficient statistic lead to negligible efficiency gains over optimal

GSDs with the same number of analyses. With an RCT design with a maximum of two analyses, the authors found adaptive designs that attained an ASN at the design alternatives of only 0.5% less than an efficient GSD however, they also showed that the addition of interim analyses in a GSD provides more substantial efficiency gains over adding analyses for adapting the sample size.

5.2 Adapting focus to subgroups

Defining the study population is an integral part of the treatment discovery process. When testing a new treatment targeted towards improving outcomes common to a particular disease, in current practice we tend to identify the study population as a subset of the overall population with the condition or characteristics of interest. We then gather study participants meeting eligibility for an intervention-control comparison and most often design the study for what is presumed to be the treatment effect in the overall group. We generally recognize that no treatment works for the entire population, and thus, we might be interested in subdividing the enrolled participants such that the intervention-control comparison is made within one or more subgroups rather than or in addition to the overall comparison. This avoids potential harm in subjects who do not benefit and increases precision for study of the treatment in a more ideal population. The intent is to determine for which group, if any, use of the treatment should be ultimately recommended. The exact approach taken will tend to balance any prior beliefs about the likelihood of a homogeneous effect in our enrolled population: we might be just trying to protect ourselves against a possibility that one of a pre-specified number of subgroups might exhibit stronger effects, or we might be trying to protect ourselves against the possibility that previously observed tendencies toward stronger effects in some subgroups were spurious. Modifying eligibility criteria to include only those patients who appear to benefit most is known as "enrichment" in the adaptive clinical trial setting. We choose to explore this concept in the traditional Phase II to Phase III progression. Preliminary Phase II studies of a candidate treatment can be used to identify a subset of the population of interest that appears to respond positively to the intervention more so than other patients such that the study population in future confirmatory trials can more appropriately reflect those who will benefit.

When early trial results suggest that a distinguishable subgroup (e.g., particular patient demographic or patients of a particular disease progression) responds more favorably, it is appealing to focus on that sub-population for the remainder of the investigation in hopes of targeting the group that will benefit from the new intervention. For instance, if we are interested in proceeding to Phase III studies only with populations with favorable Phase II outcomes, we can plan to identify those subpopulations in our Phase II evaluations. There can be two overlapping motivations for such an approach. Because the variability within subgroups is less than the variations associated with an entire population, we achieve greater precision when enrollment is restricted to the group having greatest efficacy. Nevertheless, the subjectivity of our prior belief about the effect of the

treatment is required to meet the objectivity of trial regulations. Therefore, we must consider the trade-offs between the advantages of identifying and testing the narrowest indication precisely with the smallest enrolled sample size and the potential regulatory difficulties of identifying broader indications. A balance can be achieved through the evaluation of frequentist operating characteristics at levels required by regulatory authorities and the consideration of the impact of different Bayesian priors about the existence of treatments that are only (or mostly) effective in subgroups on design precision and statistical inference.

The effects of our analyses on our treatment discovery process will depend on what true effect exists within the population that we are examining and which analyses we plan to conduct (or conduct without planning). Subgroup findings should be interpreted with care. Post-hoc exploratory subgroup analyses involve multiple comparisons, which without correction or control results in inflation of the type I error, thereby increasing the potential for chance findings. Subgroup hypotheses specified in advance have the greatest credibility, but there is likely a loss of power for detecting differences in subgroups of smaller sample sizes. An increase in the false positive rate of our design affects the outcome of our Phase II to Phase III approach.

As presented previously, with our simple hypothesis, Bayes' Rule is a function of the prevalence of truly effective treatments, π, the type I error, α, and the power, β. Again, we can think about this as prior odds times Bayes factor equals posterior odds, with Bayes factor being the power divided by type I error in our simple setting with a binary prior. (With continuous priors, the Bayes factor is more complicated and depends on the prior.) We demonstrated that the optimal Phase II to Phase III progression defined by an acceptable PPV and a maximal number of effective treatments can be parameterized by type I error and power at Phase II. Hence, increasing our type I error and changing the power of our study design will simply result in a different outcome in terms of the number of treatments identified and the overall PPV. In current practice Phase II studies are typically designed presuming homogeneous effects however, that does not eliminate the possibility of data dredging and exploratory analyses such as a search for subgroups. When this occurs the Phase II trial ultimately has a higher type I error due to the exploratory nature of the learning phase of the investigation. Without control of that type I error, we are simply allowing for a decrease in our overall PPV under the same strategy presented in chapter 2. The concern is not necessarily with the analyses being conducted at Phase II but with proper documentation of what is being done such that correct methods are applied and appropriate considerations are made for future phases of the investigation.

Subgroup analyses may be specified in the study protocol, implied but not explicitly stated, identified by other, similar trials, or may simply emerge during the course of the trial. How

subgroups are selected will affect our overall false positive rate and the populations to consider for future analyses. This will in turn determine for which subgroups treatments are ultimately adopted. Typically, trials are powered to detect an effect overall, thus with a small number of participants in each group, it is unlikely that significant differences will be correctly identified. We can consider the power to detect an effect in a subgroup and in the overall population when moving forward with various strategies for identifying important differences. We will consider three such strategies that we believe are reflective of at least some current practice for identifying a subgroup with a positive treatment effect:

- *Prefer All*: We imagine that a sponsor who is primarily interested in obtaining the broadest indication will consider a significant result overall enough evidence without the need to search further for subgroup effects. This approach might appeal to the study sponsor with a vested interest in the economic aspect of finding a population and use for a particular treatment. The sponsor might be hopeful of a homogeneous effect. If there exists an effect in the overall population, the treatment can be adopted for a larger population. If the treatment is not statistically significant in the overall population, we search for a subpopulation that might benefit. The intent is to find a use for the candidate treatment where use in the overall population is preferred. Therefore, with this strategy, if the treatment effect is significant overall, we proceed with the entire study population, otherwise we choose the most significant subgroup. This strategy requires careful consideration because we are choosing to look for subgroups when the overall effect is non-significant, a characteristic of data dredging when interpretation of findings can be problematic.

- *Choice of Significant Subgroup*: Alternatively, we imagine that a sponsor who is primarily interested in having the smallest sample size needed to demonstrate efficacy will want to eliminate subgroups that appear to have no effect. Investigators might choose this strategy to increase power by identifying and enrolling a particular group at later phases. With a pre-specified analysis outlining what comparisons are to be made, this strategy may agree with regulatory requirements. This might be so when the search for subgroups is based on reasonable expectations from well-established scientific or medical differences or other, similar trial findings and assessments of safety within particular subgroups is important. The intent is to determine if an imbalance of subgroup effects is contributing to our estimate of the overall effect. Hence, with this strategy, we choose to assess the intervention-control comparison in the overall population. If the overall effect is significant, we still proceed with subgroup analyses and choose the most significant subgroup when the opposite subgroup has an estimated nil effect. The idea is that the effect in the overall group might only be due to an effect in a particular subgroup and a possible nil or even harmful effect in the opposite group. With this strategy, we can assess safety concerns within subgroups that might not be

apparent from effects in the overall study population. In the case of non-significant subgroups, we choose the overall study population.

- *Smallest P-value*: A third approach that might be yet another alternative is identifying the subgroup in which efficacy might be most easily demonstrated. Motivation for choosing this strategy might be the use of Phase II studies as hypothesis generating research with the intent that any result obtained will be confirmed in later phases. The intent here might be to identify the subgroup with the largest treatment effect suggesting the treatment is "most beneficial" in that group. This might appeal to those with the naive ambition to find a population in which a treatment works "best". In this strategy, we choose the subgroup with the analysis that results in the smallest p-value.The focus is placed on subgroups with the largest treatment effect or intervention-control difference. Even with only a few subgroups, as we will demonstrate, the chances of spurious findings can be substantial.

Each of these strategies involves multiple comparisons, which without control causes an inflation of our false discovery rate. The magnitude of the inflation depends on the number of comparisons, the relative sample size in each subpopulation, and degree of overlap between subgroups defined using different variables. For independent comparisons, the family-wise error rate (FWER) is defined by

$$\alpha_{\text{fw}} = 1 - (1 - \alpha_{\text{pc}})^s \tag{5.8}$$

where α_{fw} is the family-wise error rate, α_{pc} is the per-comparison error rate and s is the number of comparisons. If comparisons are positively correlated, the FWER is less than $1 - (1 - \alpha_{\text{pc}})^s$. Subgroups defined by the same variable are independent. Hence, if we considered 2 independent comparisons (e.g., within males and females separately) each at an α-level of 0.05, the probability that at least one comparison would result in a type I error is $1 - (1 - 0.05)^2 = 0.0975$. However, if we also chose to test for an effect in the overall study population, our comparisons would no longer be independent considering each subgroup is correlated with the overall group. Three such comparisons would result in a type I error of 0.113.

Nonetheless, Phase II considerations are often not limited in the number of subgroups tested when subgroup analyses are performed. The impact of introducing yet another group of interest (e.g., old and young) in addition to the overall comparison and the test of males and females each at a level 0.05 is an increase our type I error rate to 0.161 (this presumes that old vs young is defined by the median age and that age and sex are independent). Increasing our chance of spurious findings in Phase II without proper adjustment of the entire Phase II to Phase III strategy leads to an increase in the number of Phase III studies devoted to ineffective treatments, thus misusing

time and resources. In chapter 2, we used tradeoffs between type I error and power to ensure a high PPV. If subgroup analyses are pre-specified, we can choose to control our type I error rate to a desired level. But the consequence of a higher type I error without a proportionate increase in the power at Phase II may result in suboptimal outcomes for our Phase II to Phase III treatment discovery approach.

The Phase II to Phase III progression is a sequential adaptive process making it reasonable to assume that preliminary findings might lead to a certain belief about particular subgroups in any stage of the investigation. In investigating what the effects of subgroup analyses are on our approach, we will consider two cases, truly homogeneous and truly heterogeneous treatment effects. We will then briefly discuss the impact of mixtures of these among the population of ideas we have at the start of Phase II.

5.2.1 Subgroup analyses under homogeneous effects

We explore the possibility that three variables are used to explore six subgroups. For ease of exposition, we choose to presume interest in subgroups defined by sex, age, and weight, though our results are truly generalizable to any variables having the same distribution. We further restrict attention to a setting in which the eight possible subgroups defined by each combination of male vs female, young vs old, heavy vs light are equally likely. Without loss of generality, we choose to divide our study population equally among males and females and in the homogeneous case, the treatment is equally effective in both subgroups. We will also examine subgroups dichotomized by age (old vs young) and weight (light vs heavy) within which the ratio of males to females is 1:1 and the treatment is equally effective in all subgroups. We choose the subgroups to be of equal sizes. Our results may not generalize directly to settings with unequal subgroup sizes owing to the greater impact statistical noise might have on choosing suboptimal subgroups and the decrease in power to detect a truly effective treatment that works only in a small subgroup.

We again consider our comparison design, Approach 2b (Table 2.1) starting with 2,047 candidate treatments in Phase II with 84.9% power and a type I error rate of 0.1. With this approach, we adopt 165 truly effective treatments (true positives; TP) with 95.0% power at Phase III and 5 truly ineffective treatments (false positives; FP) with a type I error rate of 0.025. With this design, the PPV overall is 97.1%, which does not achieve a PPV of 98%, a constraint in our search for an optimal design in chapter 2. With a Phase II type I error less than 0.10, the overall PPV achieved is no higher than 97.9% and can be as low as 94.9%, 92.4%, 90.0% as the type I error at Phase II doubles, triples, quadruples while only negligibly increasing power.

Data were simulated for 100,000 trials of normally distributed outcomes with no differential

subgroup effects to determine the extent to which such strategies for subgroup analyses incorrectly identify a subgroup effect in a homogeneous setting. Performing the 7 partially correlated comparisons (overall population, males, females, old, young, light, heavy) each at the Phase II α-level of 0.1 increases our type I error rate to 0.337. This inflation of our type I error also increases our power to 94.8% for detecting the design alternative in at least some subgroup with 342 subjects in Phase II (Table 5.3). We pass a greater number of ineffective treatments to Phase III resulting in a 24% prevalence after Phase II. This requires that more resources be used in Phase III reducing the number of treatments initially studied in Phase II (1,488 vs 2,047).

The following results (Table 5.3) apply to the progression of treatments to Phase III in at least some population. It is instructive to consider how often the Phase III (and presumably the ultimate indication) is restricted to a subgroup. If we decide to move forward with the overall population with a significant overall effect (prefer all strategy), at the end of Phase III, we now only adopt 134 effective treatments (vs 165) with 11 ineffective treatments (vs 5) and our PPV is reduced to 92.2%. Of the 134 effective treatments adopted, 120 (89.6%) would be adopted for the overall population and 2.3 (1.72%) treatments would be recommended for each of the subgroups. When we choose to proceed with the most significant subgroup when the opposite subgroup has an estimate nil effect, we shift some of the treatments adopted for the overall population to the subgroups and even more so when choosing to proceed with the group with the smallest p-value.

Hence, by indiscriminately performing subgroup analyses, we are inefficiently using our resources; more subjects are needed in later confirmatory phases for studying more treatments passing Phase II, 76% of which are truly ineffective. We must also consider that the treatment truly works for the overall population but we are restricting use of the treatment to only a particular subset for 10% of the effective treatments using the prefer all strategy. Even fewer treatments would be recommended for the entire population if we were to choose a significant subgroup over a significant overall effect (81.3%) or if we were to choose the group with the smallest p-value (32.1%). Overall, we are adopting fewer effective treatments (TP) and a greater number of ineffective treatments (FP) making this an inappropriate adaptation.

To account for the multiple comparisons, we can plan to control our type I error at Phase II by performing all 7 subgroup analyses using a significance level of 0.0226 yielding a FWER of 0.1. While we could assign different p-values across subgroups, for simplicity, we have chosen a common p-value that attains the overall FWER desired. For example, we could have chosen to test the overall population at a significance level of 0.05 and each subgroup at a level of 0.0185. A naive investigator who chooses to conduct several analyses without controlling or at least acknowledging the inflation of the overall type I error might choose a significance level of 0.025 for each of the

Table 5.3: Examples of various strategies to subgroup analyses when there exist homogeneous effects

		No Subgroups	Prefer All Inflate error	Choose Subgroup Inflate error	Smallest P-value Inflate error
Phase II	Number RCT	2,047 (10% eff)	1,488 (10% eff)	1,488 (10% eff)	1,488 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 84.9%	0.337; 94.8%	0.337; 94.8%	0.337; 94.8%
	"Positive" RCT	174 eff; 184 not	141 eff; 451 not	141 eff; 451 not	141 eff; 451 not
Phase III	Number RCT	358 (49% eff)	592 (24% eff)	592 (24% eff)	592 (24% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 95.0%	0.025; 95.0%	0.025; 95.0%	0.025; 95.0%
	# Effctve Adopt	165	134	134	134
	For Overall	165	120	109	43
	For Males	0	2.3	4.3	15
	For Females	0	2.3	4.3	15
	For All Others	0	9.2	16.8	60
	# Ineff Adopt	5	11	11	11
Overall	PPV	97.3%	92.2%	92.2%	92.2%
	Type I err; Pwr	0.0025; 80.6%	0.0082; 90.0%	0.0082; 90.0%	0.0082; 90.0%

analyses. This is quite close to the 0.0262 necessary to achieve a FWER of 0.1 and would result in an overall type I error of 0.113.

Each analysis now requires a more significant effect altering the probability of moving forward with a particular group in Phase III. In Table 5.4 are outcomes for Phase II when there exists homogeneous effects such that the treatment does not work ($\mu = 0$ in all subgroups). This table represents the error spent on each group under the null. For each strategy for identifying significant subgroups, we present what can be thought of as the boundaries of our design on the error spending scale. For example, when we choose to control our FWER to 0.1 with the prefer all strategy, we spend 0.0226 on the overall group and 0.0131 on each of the subgroups. How these errors are spent depend on the strategy chosen. We present three out of a multitude of strategies that might be considered.

When controlling our type I error at Phase II, we move forward to Phase III with fewer ineffective treatments (fewer false positives; FP) allowing for the use of more resources at Phase II such that we can study more treatments (2,081 vs 1,488) (Table 5.5). This control of our type I error also decreases our power to 77.2% for detecting the design alternative with 342 subjects in Phase II. If we decide to move forward with the overall population with a significant overall effect (prefer all strategy), at the end of Phase III, we now adopt 153 effective treatments (vs 165) with only 5 ineffective treatments with a PPV of 97.0%.

Of the 153 effective treatments adopted, 122 (79.7%) would be adopted for the overall population and 5.1 (3.33%) treatments would be recommended for each of the subgroups. Again, when we choose to proceed with the most significant subgroup when the opposite subgroup has an estimate nil effect, we shift some of the treatments adopted for the overall population to the subgroups and even more so when choosing to proceed with the group with the smallest p-value. With this strategy, we lose 12 effective treatments (153 vs 165) and fail to adopt the interventions for 100% of the population that they truly benefit: for only 54 of the treatments, we appropriately identify that the treatments benefit the overall population. For the other 99 interventions, we fail to adopt the treatments for 50% of the population that they truly benefit.

Thus, we can improve in our approach of inappropriately considering subgroup analyses under homogeneous effects by controlling our type I error. However, there is still a cost of performing such analyses in the presence of homogeneity. In Table 5.6 are outcomes for Phase II when there exists homogeneous effects such that the treatment works in all subgroups ($\mu = 0.125$ in all subgroups) and we choose to perform subgroup analyses according to the three different strategies considered, prefer all, choose the significant subgroup, or proceed with the group with the smallest p-value.

Table 5.4: Probability of choosing a particular group to move forward with in Phase III among various strategies when there exist homogeneous effects under the null such that the treatment is truly ineffective for the entire study population

		Subgroups moving to Phase III				
Homogeneous	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.101	0.101	0.075	0.017	0.025
μ=0	Males	0.101	0.040	0.044	0.053	0.025
σ^2=0.25	Females	0.101	0.039	0.043	0.054	0.025
	Old	0.102	0.040	0.044	0.054	0.025
Males	Young	0.101	0.039	0.044	0.054	0.025
μ=0	Light	0.101	0.039	0.043	0.053	0.025
σ^2=0.25	Heavy	0.101	0.039	0.043	0.053	0.025
	Any	–	0.337	0.336	0.338	–

		Subgroups moving to Phase III				
Control error Homogeneous	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.023	0.023	0.021	0.007	0.025
μ=0	Males	0.023	0.013	0.013	0.015	0.025
σ^2=0.25	Females	0.023	0.013	0.013	0.016	0.025
	Old	0.023	0.013	0.014	0.016	0.025
Males	Young	0.023	0.013	0.013	0.016	0.025
μ=0	Light	0.023	0.013	0.013	0.015	0.025
σ^2=0.25	Heavy	0.023	0.013	0.014	0.016	0.025
	Any	–	0.100	0.101	0.101	–

Table 5.5: Examples of various strategies to subgroup analyses when controlling the type I error and there truly exist homogeneous effects

		No Subgroups	Prefer All Control error	Choose Subgroup Control error	Smallest P-value Control error
Phase II	Number RCT	2,047 (10% eff)	2,081 (10% eff)	2,081 (10% eff)	2,081 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 84.9%	0.100; 77.2%	0.100; 77.2%	0.100; 77.2%
	"Positive" RCT	174 eff; 184 not	161 eff; 187 not	161 eff; 187 not	161 eff; 187 not
Phase III	Number RCT	358 (49% eff)	348 (46% eff)	348 (46% eff)	348 (46% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 95.0%	0.025; 95.0%	0.025; 95.0%	0.025; 95.0%
	# Effctve Adopt	165	153	153	153
	For Overall	165	122	119	54
	For Males	0	5.1	5.7	16.4
	For Females	0	5.1	5.7	16.4
	For All Others	0	20.4	22.8	65.6
	# Ineff Adopt	5	5	5	5
Overall	PPV	97.3%	97.0%	97.0%	97.0%
	Type I err; Pwr	0.0025; 80.6%	0.0025; 73.3%	0.0025; 73.3%	0.0025; 73.3%

Examining the probability of a significant effect at Phase II within each group and overall, we have study power, 85.1% to detect an effect in the overall population of 342 subjects and approximately 64% to detect an effect in each subgroup due to the smaller sample sizes. When controlling our FWER to 0.1, our power reduces to 61.6% for overall and approximately 35% for each subgroup.

- *Prefer All*: If we decide to look for subgroups only when the treatment effect is non-significant overall, we will choose any of the subgroups with probability 0.0163 (0.0259 when controlling our type I error) and the overall population with probability 0.851.
- *Choice of Significant Subgroup*: If we decide to move forward with a subgroup when the effect is highly significant and the opposite group has an estimated nil effect regardless of a significant effect overall, we reduce how often we choose the overall study population to 77.0% (60.0% when we control our type I error) and increase the probability of moving forward with a subgroup to 0.0295.
- *Smallest P-value*: And the last strategy to consider is choosing the group with the smallest p-value where the overall study population is chosen with probability 0.306 (0.274 with controlled type I error) and each of the subgroups with probability 0.106 (0.0827 with controlled type I error).

If prior beliefs suggest that subgroup analyses are important to the scientific relevance of a trial, the optimal approach is a pre-specified analysis at Phase II controlling the FWER and an increase in the sample size such that the study is more appropriately powered. However, in the case of true homogeneous effects, there is still a loss of the number of adopted treatments and a loss of the number of treatments appropriately recommended for the subgroups within which they are truly beneficial. Trial sponsors have an economic interest in the cost and the efficient use of time and resources. From that perspective, subgroup analyses are economically inefficient limiting the population for which the treatment is ultimately recommended when it is truly effective in the overall population.

To some extent, how we choose to proceed in Phase III is a direct result of the accuracy of preliminary results and scientific findings that lead to the consideration of subgroups initially. In this particular setting, we are unnecessarily searching for subgroups resulting in a less efficient treatment discovery process in terms of the number of treatments identified and the subgroups for which use of such treatments are recommended. When controlling the false positive rate, we improve our approach in terms of the number of effective treatments identified but not relative to the number of treatments studied or in terms of the subgroups for which the treatments are ultimately adopted.

Table 5.6: Probability of choosing a particular group to move forward with in Phase III among various strategies when there exist homogeneous effects under the alternative such that the treatment is truly effective for the entire study population

		Subgroups moving to Phase III				
Homogeneous	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.851	0.851	0.770	0.306	0.950
μ=0.125	Males	0.642	0.016	0.030	0.106	0.950
σ^2=0.25	Females	0.639	0.016	0.030	0.106	0.950
	Old	0.642	0.017	0.030	0.109	0.950
Males	Young	0.640	0.017	0.030	0.107	0.950
μ=0.125	Light	0.638	0.015	0.028	0.106	0.950
σ^2=0.25	Heavy	0.642	0.016	0.030	0.108	0.950
	Any	–	0.948	0.948	0.948	–

		Subgroups moving to Phase III				
Control error Homogeneous	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.616	0.616	0.600	0.274	0.950
μ=0.125	Males	0.354	0.026	0.029	0.082	0.950
σ^2=0.25	Females	0.351	0.026	0.029	0.082	0.950
	Old	0.352	0.027	0.030	0.084	0.950
Males	Young	0.352	0.025	0.028	0.083	0.950
μ=0.125	Light	0.351	0.026	0.029	0.083	0.950
σ^2=0.25	Heavy	0.353	0.026	0.029	0.084	0.950
	Any	–	0.772	0.774	0.772	–

5.2.2 Subgroup analyses under heterogeneous effects

However, subgroup analyses are not intended for settings of homogeneity. With heterogeneous effects, when the overall group includes a mixture of a population within which the treatment works and a population with an attenuated or nil effect, we can plan to evaluate the overall population and subpopulations using the strategies previously mentioned in hopes that we will correctly identify the subgroup for which the treatment works.

We can describe the overall effect by its mean and variance. Consider subgroup 1 made up of n_1 subjects distributed with mean, μ_1 and variance τ_1^2 and subgroup 2 with n_2 subjects distributed with mean, μ_2 and variance τ_2^2. We can define the mean, μ and variance, σ^2 of the overall group as

$$\mu = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2} \tag{5.9}$$

$$\sigma^2 = \frac{(\tau_1^2 + \mu_1^2)n_1 + (\tau_2^2 + \mu_2^2)n_2}{(n_1 + n_2)} - \mu^2 \tag{5.10}$$

We will consider each strategy in a setting with homogeneous effects among all subgroups, and in settings with heterogeneous effects defined by the treatment effect in a subgroup within which the treatment works and overall. The results will depend on how many subgroups are considered and on the relative sizes of the subgroups. We will first introduce notation for the settings to be considered.

Among the many scenarios to consider, we choose to examine those defined by the treatment effects that truly exist in the overall population and in subgroups, as well as by the treatment effects presumed for the study design. Without loss of generality, we choose to divide our study population equally among males and females such that in the heterogeneous case, the treatment will only be effective in males or half of the population. We will also examine subgroups dichotomized by age (old vs young) and weight (light vs heavy) within each of which the ratio of males to females is 1:1. To determine the number of patients needed to detect the MCID, we can use the sample size formula (Equation 1.3) presented in section 1.3.1 parameterized by the operating characteristics (type I error, power) and the variability of the statistic used. We choose a setting in which a fixed sample level 0.1 test having 85% power to detect the MCID in a two-sample clinical trial would require 342 subjects in Phase II. This corresponds to an MCID, $\mu = 0.125$ and standard deviation, $\sigma = 0.5$. The following settings will be considered:

- We previously considered *homogeneous effects*: the treatment is equally effective in all subgroups with $\mu = 0.125$ or $\mu = 0$.
- We now consider *heterogeneous effects*: the treatment is only effective in males. We choose three scenarios that result in the true effect in males reflective of double the effect when males

and females are combined. This is achievable in a number of ways. We explore three settings that vary with respect to the size of the overall effect. We believe these settings are among many that might be indicative of some sponsor beliefs as they are considering a search for subgroup effects.

A: We plan the study presuming homogeneous effects such that we design the study to detect an effect, μ in the overall study population. We perform subgroup analyses of the intervention-control comparison (males, females, old, young, light, heavy) for protection such that we will identify important subgroups if they exist. The treatment is truly only effective in males.

Considering the setting in which the treatment has an effect on subgroup 1 (males), $\mu_1 = \mu = 0.125$, no effect on subgroup 2 (females), $\mu_2 = 0$, equal variances among subgroups ($\tau_1^2 = \tau_2^2 = 0.25$), and the subgroups are of equal size ($n_1 = n_2$), the formulas in Equations 5.11 and 5.12 reduce to

$$\mu = \mu_1/2 \tag{5.11}$$

$$\sigma^2 = \tau^2 + (\mu_1/2)^2 \tag{5.12}$$

Therefore the outcomes for each group are distributed,

Males:	$Y \sim (\mu, \tau^2)$
Females:	$Y \sim (0, \tau^2)$
Overall:	$Y \sim (\mu/2, \sigma^2)$

An example of this case might be that we obtained accurate estimates of the treatment effect from some preliminary study and unknowingly expanded the population to a group within which it does not work.

B: We plan the study presuming homogeneous effects such that we design the study to detect an effect, μ in the overall study population. The treatment is truly effective in males with a mean of 0.174. This scenario is considered an intermediate between scenarios A described above (treatment is effective with an effect of the MCID in males) and C described below (treatment is effective with an effect of the MCID in the overall population and at twice that effect in males). In this setting, the outcomes for each group are distributed,

Males:	$Y \sim (0.174, \tau^2)$
Females:	$Y \sim (0, \tau^2)$

Overall: $Y \sim (0.174/2 = 0.087, \sigma^2)$ where $\sigma^2 = \tau^2 + (0.174/2)^2$

C: We plan the study presuming homogeneous effects such that we design the study to detect an effect, μ in the overall study population. We perform subgroup analyses of the intervention-control comparison (males, females, old, young, light, heavy) for protection such that we will identify important subgroups if they exist. The treatment is truly only effective in males with an effect that is twice the effect we powered the study to detect, 2μ. The effect in the overall group averages to be the MCID, μ. In this setting, the outcomes for each group are distributed,

Males: $Y \sim (2\mu, \tau^2)$
Females: $Y \sim (0, \tau^2)$
Overall: $Y \sim (\mu, \sigma^2)$ where $\sigma^2 = \tau^2 + \mu^2$

An example of this case might be that we obtained an accurate estimate of the treatment effect for the overall population but in truth there exists some subgroup that possesses all of the effect.

To consider how these scenarios will affect our treatment discovery process, we can examine how often we will choose to proceed with a subgroup rather than continuing to study the population as a whole beyond our Phase II evaluation. Each of the scenarios were simulated for 100,000 trials of normally distributed outcomes as described above. Presented in the following tables are the probabilities for detecting a significant effect and moving forward with a particular group in Phase III when we choose to study the treatment according to the strategies previously described.

In Table 5.7 are outcomes for Phase II when we have heterogeneous treatment effects and choose to perform subgroup analyses according to the three different strategies (prefer all, choose significant subgroup, or smallest p-value) considered in each case A (worst case), B (intermediate case), or C (best case).

In Case A, the treatment is effective in males with an effect of the MCID, μ and therefore an attenuated effect in the overall population of $\mu/2$. If we are just looking for a significant effect at Phase II within a particular group, we are powered at 64.2% to detect an effect in males due to the smaller sample size of $N_2/2$ of the subgroup. We would find a significant effect for the overall group with probability 0.451, females with the false positive rate of 0.102 and each of the other subgroups (old, young, light, and heavy) with probability 0.322 within which the treatment is effective half of those subgroup populations.

- *Prefer All*: If we decide to look for subgroups only when the treatment effect is non-significant overall, we will choose males with probability 0.196 and each of the stratified subgroups with probability 0.0344 while continuing with the overall population with probability 0.451.

- *Choice of Significant Subgroup*: If we decide to move forward with a subgroup when the effect is most significant among some subgroup and the opposite group has an estimated nil effect regardless of a significant effect overall, we reduce how often we choose the overall study population to 30.9% and instead move forward with the significant male subgroup with probability 0.291.

- *Smallest P-value*: And the last strategy to consider is choosing the group with the smallest p-value where the overall study population is chosen with probability 0.0639 and males with probability 0.407.

- Our overall power for moving forward with any group in studying treatments in Phase III for each of these strategies is the same. The probability of studying a group in a confirmatory study depends on how we choose to move forward and therefore, the probabilities are simply shifted between subgroups among the different strategies.

The same examinations can be made for Cases B and C where similar trends exist; we are more likely to move forward with the overall population than the significant subgroup if we prefer all and more likely to move forward with studying the treatment only in males when we choose to continue searching for a significant subgroup or choose the group with the smallest p-value. Comparing these three cases, we consider Case A to be an example of being underpowered, Case C an example of being overpowered such that the true effect in males is twice the MCID, and Case B to be an intermediate scenario. In Table 5.8 are outcomes for Phase II when we have heterogeneous treatment effects and choose to control our false positive rate at Phase II when performing subgroup analyses and moving forward to Phase III with particular subgroups. Among all cases, we are less powered to detect the MCID with 342 subjects at Phase II when controlling our type I error.

We can again consider Approach 2b (Table 2.1) for introducing subgroup analyses when there exists heterogeneous effects. We will first demonstrate the effects of such considerations using the "worst case scenario" defined by Case A where the treatment works in males with an effect μ and therefore there is an attenuated effect in the overall population of $\mu/2$. Hence, our study is underpowered. In Table 5.9 is Approach 2b with reduced power at Phase II of 45.1% due to the treatment only being effective in males. Therefore, we pass fewer treatments on to Phase III allowing for more treatments to be studied in Phase II (2,203 vs 2,047). If we were to proceed with the overall population in Phase III (no subgroup analyses), we would only have 43.7% power to detect an effect when the treatment only works in half the population at the MCID. Overall, we adopt 44

effective and 5 ineffective treatments with this approach (89.8% PPV).

When there truly exists an effect only in a subgroup but we choose not to perform subgroup analyses, we do not identify as many effective treatments and suggest that those identified are beneficial for the overall population including females and males. With Phase II results that suggest that a treatment is promising for a particular subgroup, we can choose to study the treatment in 829 subjects from the group of interest resulting in a more powerful study design at Phase III. Which subgroups we choose to study the treatment with will depend on the chosen strategy.

- *Prefer All*: Choosing to proceed with the entire study population when the effect is significant overall and with subgroups when there does not exist a significant overall effect inflates our type I error and increases our power to 79%. We identify 67 effective treatments when we choose the prefer all strategy such that 30 (45%) of the treatments will be recommended for the overall population, 28 (42%) for males, and 2.3 (3.3%) for each of the non-gender subgroups recognizing that a true effect only exists in males.
- *Choice of Significant Subgroup*: With the choice of a significant subgroup over a significant overall effect, we adopt 75 effective treatments, 21 (28%) for the overall population and 42 (56%) for males and 3.1 (4.1%) for each of the other subgroups.
- *Smallest P-value*: With the choice of the group with the smallest p-value, we identify 83 effective treatments, 4 (4.8%) for the overall population and 59 (71%) for males and 5.1 (6.1%) for each of the other subgroups.

We have more power to detect an effect in males when searching for significant subgroups despite significant effects in the overall population. However, we adopt more effective treatments, with a majority in the correct subgroup, at the cost of suggesting that some treatments are effective in subgroups in which they truly only benefit half of the population. Inflation of our type I error through multiple comparisons results in the adoption of 12 ineffective therapies.

With a controlled type I error rate, fewer ineffective treatments are passed to Phase III for evaluation allowing for more treatments to be studied in Phase II (Table 5.10). However, when controlling the false positive rate with such an attenuated effect in the overall population as in Case A, the power to detect the MCID at Phase II is reduced to approximately 48%. Therefore, we initially study more treatments while ultimately adopting fewer treatments. With heterogeneous effects, the strategies of choosing the most significant subgroup or choosing the subgroup with the smallest p-value result in identifying more treatments for males but also for other subgroups. An important improvement with this strategy is reducing the number of ineffective treatments adopted to 5 (vs 12).

Table 5.7: Probability of choosing a particular group to move forward with in Phase III among various strategies when there exist heterogeneous effects described by Case A, B, or C such that the treatment is truly only effective in males

Heterogeneous Case A	Group	Subgroups moving to Phase III: Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.451	0.451	0.309	0.064	0.437
μ=0.125/2	Males	0.642	0.196	0.291	0.407	0.950
σ^2=0.25+(0.125/2)^2	Females	0.102	0.004	0.005	0.010	0.025
	Old	0.322	0.035	0.046	0.078	0.437
Males	Young	0.322	0.034	0.046	0.077	0.437
μ=0.125	Light	0.321	0.034	0.045	0.076	0.437
τ^2=0.25	Heavy	0.323	0.034	0.046	0.077	0.437
	Any	–	0.788	0.788	0.789	–

Heterogeneous Case B	Group	Subgroups moving to Phase III: Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.630	0.630	0.398	0.072	0.709
μ=0.174/2	Males	0.844	0.204	0.391	0.592	0.999
σ^2=0.25+(0.174/2)^2	Females	0.101	0.001	0.001	0.003	0.025
	Old	0.444	0.020	0.032	0.062	0.709
Males	Young	0.444	0.020	0.031	0.062	0.709
μ=0.174	Light	0.443	0.019	0.031	0.060	0.709
τ^2=0.25	Heavy	0.446	0.020	0.030	0.061	0.709
	Any	–	0.913	0.914	0.912	–

Heterogeneous Case C	Group	Subgroups moving to Phase III: Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.851	0.851	0.473	0.061	0.950
μ=0.125	Males	0.978	0.119	0.474	0.806	0.999
σ^2=0.25+0.125^2	Females	0.101	0.000	0.000	0.000	0.025
	Old	0.642	0.005	0.011	0.031	0.950
Males	Young	0.640	0.005	0.011	0.031	0.950
μ=2·(0.125)	Light	0.638	0.005	0.011	0.029	0.950
τ^2=0.25	Heavy	0.642	0.005	0.011	0.030	0.950
	Any	–	0.990	0.989	0.988	–

Table 5.8: Probability of choosing a particular group to move forward with in Phase III among various strategies when controlling the type I error and there exist heterogeneous effects described by Case A, B, or C such that the treatment is truly only effective in males

		Subgroups moving to Phase III				
Control error Heterogeneous Case A	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.194	0.194	0.170	0.045	0.437
μ=0.125/2	Males	0.354	0.178	0.198	0.262	0.950
σ^2=0.25+(0.125/2)^2	Females	0.022	0.002	0.002	0.004	0.025
	Old	0.118	0.025	0.027	0.042	0.437
Males	Young	0.115	0.026	0.027	0.042	0.437
μ=0.125	Light	0.115	0.025	0.027	0.041	0.437
τ^2=0.25	Heavy	0.116	0.026	0.027	0.041	0.437
	Any	–	0.476	0.478	0.477	–
		Subgroups moving to Phase III				
Control error Heterogeneous Case B	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.339	0.339	0.277	0.059	0.709
μ=0.174/2	Males	0.604	0.271	0.329	0.469	0.999
σ^2=0.25+(0.174/2)^2	Females	0.022	0.001	0.001	0.001	0.025
	Old	0.190	0.021	0.023	0.041	0.709
Males	Young	0.188	0.022	0.023	0.042	0.709
μ=0.174	Light	0.189	0.021	0.023	0.041	0.709
τ^2=0.25	Heavy	0.191	0.021	0.023	0.041	0.709
	Any	–	0.696	0.699	0.694	–
		Subgroups moving to Phase III				
Control error Heterogeneous Case C	Group	Probability of Significance at Phase II	Prefer All	Choice	Smallest P-value	Probability of Significance at Phase III
Overall	All	0.610	0.610	0.422	0.057	0.950
μ=0.125	Males	0.895	0.279	0.464	0.759	0.999
σ^2=0.25+0.125^2	Females	0.022	0.000	0.000	0.000	0.025
	Old	0.346	0.009	0.010	0.027	0.950
Males	Young	0.346	0.009	0.011	0.027	0.950
μ=2·(0.125)	Light	0.346	0.008	0.010	0.026	0.950
τ^2=0.25	Heavy	0.347	0.008	0.010	0.027	0.950
	Any	–	0.923	0.927	0.923	–

- *Prefer All*: When choosing to proceed with the entire study population when the effect is significant overall we identify 65 effective treatments when we choose the prefer all strategy such that 19 (29.2%) of the treatments will be recommended for the overall population, 37 (56.9%) for males, and 2.5 (3.85%) for each of the non-gender subgroups recognizing that a true effect only exists in males.

- *Choice of Significant Subgroup*: With the choice of a significant subgroup over a significant overall effect, we adopt 68 effective treatments, 16 (23.5%) for the overall population and 41 (60.3%) for males and 2.5 (3.68%) for each of the other subgroups.

- *Smallest P-value*: With the choice of the group with the smallest p-value, we identify 75 effective treatments, 4 (5.33%) for the overall population and 55 (73.3%) for males and 4 (5.33%) for each of the other subgroups.

We obtain a higher PPV when controlling the false positive rate but have much lower power overall. For particular settings, it may be of importance to assess both the risk of patients being denied an effective intervention and the risk of patients being treated with an ineffective or even harmful intervention. In this case, the strategy to control of our type I error and choosing to progress to Phase III with the group with the smallest p-value dominates in terms of the number of treatments identified. Although we identify more effective interventions with higher power overall when inflating our false positive rate, we also adopt an undesirable number of ineffective treatments resulting in an unacceptably high PPV. Even by controlling our error and selecting the most significant subgroup, we still only obtain a PPV of 93.8% and overall power of 34.1%.

This is the best we can do with this design when we believe that the estimate of the overall treatment effect is half the MCID because of a subgroup. With the smallest p-value strategy we adopt more effective and fewer ineffective interventions and 73.3% of the treatments are identified for the appropriate subgroup. If prior belief about the treatment effect was this case, this study design is inappropriate. A better method of protection would be to power the study higher and to introduce an acceptable futility rule, which would result in early stopping for a large number of studies.

We have thus far considered homogeneous and heterogeneous treatment effects separately but can conceptualize a setting within which there exists some proportion of each among the population of candidate treatments. This allows us to examine the "tipping point", the setting in which moving forward with either single strategy produces similar results when there exists a mixture of homogeneous and heterogeneous treatment effects. That is we look for when the number of treatments lost when the treatment effect is truly homogeneous, but we searched for subgroups, is equal to the number of effective treatments gained when the treatment effect is truly heterogeneous.

Table 5.9: Examples of various approaches to subgroup analyses when there exist heterogeneous effects described by Case A (overall $\mu = 0.125/2$; males $\mu = 0.125$)

		No Subgroups	Prefer All Inflate error	Choose Subgroup Inflate error	Smallest P-value Inflate error
Phase II	Number RCT	2,203 (10% eff)	1,518 (10% eff)	1,518 (10% eff)	1,518 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 45.1%	0.337; 78.8%	0.337; 78.8%	0.337; 78.9%
	"Positive" RCT	99 eff; 198 not	120 eff; 460 not	120 eff; 460 not	120 eff; 460 not
Phase III	Number RCT	297 (33% eff)	580 (21% eff)	580 (21% eff)	580 (21% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 43.7%	0.025; 56.3%	0.025; 62.4%	0.025; 70.0%
	# Effctve Adopt	44	67	75	83
	For Overall	44	30	21	4
	For Males	0	28	42	59
	For Females	0	0	0	0
	For All Others	0	9.1	12.2	20.4
	# Ineff Adopt	5	12	12	12
Overall	PPV	89.8%	85.4%	86.6%	87.9%
	Type I err; Pwr	0.0025; 20.0%	0.0084; 44.3%	0.0084; 49.2%	0.0084; 55.0%

For example, when applying the smallest p-value strategy under homogeneous effects (Table 5.5), we lose 12 effective treatments (165 vs 153) that would have been identified if we did not perform subgroup analyses. When applying the same strategy under heterogeneous effects in Case A (Table 5.10), we gain 31 effective treatments (44 vs 75). Therefore, if we believe the mixture of treatment effects is at most 72% homogeneous and at least 28% heterogeneous, choosing the smallest p-value strategy would allow for the gain in treatments with subgroup analyses for treatments with heterogeneous effects to make up for the loss of treatments with subgroup analyses for treatments with homogeneous effects. An imbalance in the proportion of treatments with homogeneous or heterogeneous effects will result in either an overall loss or overall gain in the number of effective treatments identified. This does not consider the number of treatments identified for the appropriate subgroup.

We will now consider the effects of moving to Phase III with different subgroups using the intermediate scenario defined by Case B where the treatment works in males with an effect 0.174

Table 5.10: Examples of various approaches to subgroup analyses when controlling the type I error and there exist heterogeneous effects described by Case A (overall $\mu = 0.125/2$; males $\mu = 0.125$)

		No Subgroups	Prefer All Control error	Choose Subgroup Control error	Smallest P-value Control error
Phase II	Number RCT	2,203 (10% eff)	2,192 (10% eff)	2,192 (10% eff)	2,192 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 45.1%	0.100; 47.6%	0.100; 47.8%	0.100; 47.7%
	"Positive" RCT	99 eff; 198 not	104 eff; 197 not	105 eff; 197 not	105 eff; 197 not
Phase III	Number RCT	297 (33% eff)	301 (35% eff)	302 (35% eff)	302 (35% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 43.7%	0.025; 62.7%	0.025; 64.8%	0.025; 71.5%
	# Effctve Adopt	44	65	68	75
	For Overall	44	19	16	4
	For Males	0	37	41	55
	For Females	0	0	0	0
	For All Others	0	9.8	10.0	15.9
	# Ineff Adopt	5	5	5	5
Overall	PPV	89.8%	93.0%	93.2%	93.8%
	Type I err; Pwr	0.0025; 20.0%	0.0025; 29.9%	0.0025; 31.0%	0.0025; 34.1%

and in the overall population with the effect of 0.174/2. In Table 5.11 is Approach 2b with reduced power at Phase II to 63.8% due to the treatment only being effective in males. Our power at Phase II is higher than in Case A because the true treatment effect overall (and in males) is greater (0.174/2 vs 0.13/2). Again, we pass fewer treatments on to Phase III allowing for more treatments to be studied in Phase II (2,130 vs 2,047). If we were to proceed with the overall population in Phase III (no subgroup analyses), we would have 70.9% power to detect an effect when the treatment only works in half the population with an effect that is slightly greater than the MCID. Overall, we adopt 97 effective and 5 ineffective treatments with this approach (95.3% PPV).

Similar to Case A, choosing to proceed with subgroups in Case B increases the number of effective treatments adopted; when controlling our false positive rate, we adopt 121, 125, and 132 effective treatments (vs 97) with the prefer all, choose significant subgroup, and smallest p-value strategies, respectively. The results obtained with an inflated type I error at Phase II can be found in the Appendix (Table A7). In this case, we are not quite as underpowered for detecting an effect in the overall population but there still remains an advantage to searching for subgroups. The strategy to choose based on the smallest p-value again dominates in terms of the number of effective treatments identified and the number identified for the appropriate subgroup. If we believe the mixture of treatment effects is at most 74.5% homogeneous and at least 25.5% heterogeneous, choosing the smallest p-value strategy would allow for the gain in treatments with subgroup analyses for treatments with heterogeneous effects to make up for the loss of treatments with subgroup analyses for treatments with homogeneous effects.

We will now consider the best case scenario defined by Case C where the treatment works in males with an effect that is twice the MCID, 2μ and in the overall population with the effect of the MCID, μ. In Table 5.12 is Approach 2b with 84.9% power at Phase II due to the treatment being effective in males at twice the effect are were looking for. This results in the same number of treatments being passed as in the heterogeneous case, 165 effective and 5 ineffective. In this case, are studies are overpowered to detect an effect in males making it advantageous to conduct subgroup analyses. We identify 15-20 more effective interventions and increase our power overall to approximately 90% when searching for significant subgroups with a controlled type I error. The results obtained with an inflated type I error at Phase II can be found in the Appendix (Table A8).

If we truly believe that the effect is twice the MCID in a particular subgroup, the advantage of moving forward with the group with the smallest p-value is the correct identification of the subgroup that the treatment truly benefits. In this case, we identify 186 effective interventions with the smallest p-value strategy 11 (5.91%) of which would be adopted for the overall population, 154 (82.85%) for males, and 5.2 (2.80%) for each of the other subgroups. An overwhelming majority

Table 5.11: Examples of various approaches to subgroup analyses when controlling the type I error and there exist heterogeneous effects described by Case B (overall $\mu = 0.174/2$; males $\mu = 0.174$)

		No Subgroups	Prefer All Control error	Choose Subgroup Control error	Smallest P-value Control error
Phase II	Number RCT	2,130 (10% eff)	2,110 (10% eff)	2,110 (10% eff)	2,110 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 63.8%	0.100; 69.6%	0.100; 69.9%	0.100; 69.4%
	"Positive" RCT	136 eff; 192 not	147 eff; 190 not	147 eff; 190 not	146 eff; 190 not
Phase III	Number RCT	328 (41% eff)	337 (44% eff)	337 (44% eff)	336 (43% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 70.9%	0.025; 82.1%	0.025; 84.4%	0.025; 90.4%
	# Effctve Adopt	97	121	125	132
	For Overall	97	51	41	9
	For Males	0	57	69	99
	For Females	0	0	0	0
	For All Others	0	12.7	13.8	24.7
	# Ineff Adopt	5	5	5	5
Overall	PPV	95.3%	96.2%	96.3%	96.5%
	Type I err; Pwr	0.0024; 45.5%	0.0025; 57.1%	0.0025; 59.0%	0.0025; 62.7%

of the treatments are identified for the appropriate population. Again, it may be of importance to assess both the risk of patients being denied an effective intervention and the risk of patients being treated with an ineffective or even harmful intervention. With such high power in this case, it proves useful to test treatments for heterogeneity to identify which subgroups to move forward in future studies in terms of being much closer to achieving the desired 98% PPV. If we believe the mixture of treatment effects is at most 63.6% homogeneous and at least 36.4% heterogeneous, choosing the smallest p-value strategy would allow for the gain in treatments with subgroup analyses for treatments with heterogeneous effects to make up for the loss of treatments with subgroup analyses for treatments with homogeneous effects.

If we decide not to look for subgroup effects and there exists a subgroup for which the treatment works well when the opposite subgroup has an estimated nil effect, we will have less power in the total population for detecting a difference between subgroups. In this setting, we might choose to increase our sample size to accommodate the multiple comparisons such that we have more subjects in each group for detecting effects.

Table 5.12: Examples of various approaches to subgroup analyses when controlling type I error and there exist heterogeneous effects described by Case C (overall $\mu = 0.125$; males $\mu = 0.250$)

		No Subgroups	Prefer All Control error	Choose Subgroup Control error	Smallest P-value Control error
Phase II	Number RCT	2,047 (10% eff)	2,028 (10% eff)	2,028 (10% eff)	2,028 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 84.9%	0.100; 92.3%	0.100; 92.7%	0.100; 92.3%
	"Positive" RCT	174 eff; 184 not	187 eff; 183 not	188 eff; 183 not	187 eff; 183 not
Phase III	Number RCT	358 (49% eff)	370 (51% eff)	371 (51% eff)	370 (51% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 95.0%	0.025; 96.5%	0.025; 97.5%	0.025; 99.1%
	# Effctve Adopt	165	181	183	186
	For Overall	165	118	81	11
	For Males	0	57	94	154
	For Females	0	0	0	0
	For All Others	0	6.6	7.9	20.6
	# Ineff Adopt	5	5	5	5
Overall	PPV	97.3%	97.5%	97.6%	97.6%
	Type I err; Pwr	0.0025; 80.6%	0.0025; 89.1%	0.0025; 90.4%	0.0025; 91.5%

With homogeneous effects, we are most successful in identifying effective interventions when considering only the population as a whole. However, if reason stands for considering a subset of those enrolled, we can still identify a number of effective treatments when non-adaptive methods are applied and pre-specified analyses are performed such that our type I error is controlled. We must consider how many treatments are ultimately adopted for inappropriate subgroups and the safety risks associated with particular treatments.

With heterogeneous effects, only considering the treatment effect on the population overall affects the efficiency of our approach because we fail to identify the true population that would benefit from the treatment. However, we demonstrate the risk of subgroup analyses by considering how often treatments are adopted for an inappropriate subgroup. Looking for subgroups while controlling our errors does improve our outcomes when heterogeneous effects truly exist and does not result in a great loss when a treatment works in the entire population.

We demonstrate the idea of determining the "tipping point" at which there is a balance between the number of treatments gained or lost when moving forward with subgroups under both homogeneous and heterogeneous treatment effect settings. We stress the importance of controlling the type I error when considering multiple comparisons and other adaptive methods that affect the credibility and reliability of trial results. Without the evidence necessary to truly improve clinical practice and the health of the population, trial results are not necessarily useful.

5.3 Adapting analysis model or endpoint

We have thus far discussed a sequential adaptive Phase II to Phase III approach in which we test new ideas for treatments or strategies targeted toward improving outcomes common to particular diseases and patient populations. We design our studies such that our hypotheses reflect a particular indication in terms of the disease setting, patient population targeted, and treatment outcome of interest. To determine what treatment outcomes are best measures of safety, efficacy, and effectiveness, we consider clinically relevant and attainable endpoints that we believe the treatment might plausibly affect.

In current practice, in the absence of a treatment effect, we choose to decide a benefit exists with some low probability, 0.025 in a one-sided study, for example. This rate of obtaining a false positive is true when addressing a single question of the effect of a treatment on a single outcome. When trying to determine the effect of a treatment on multiple outcomes, the chance of deciding an ineffective treatment should be adopted for any one of the outcomes considered increases. Similar

to the consideration of subgroup analyses, the actual inflation of our type I error depends on the number of comparisons made and the correlation between the endpoints.

To avoid this problem, we often pre-specify a primary endpoint. That primary endpoint is one that is most relevant clinically, that the treatment is most likely to affect, and that can be assessed most accurately and precisely. Of most importance is the ability of the outcome to address the scientific question of interest. Common to current practice is the consideration of "secondary" endpoints as supportive and confirmatory of the effect obtained for the primary endpoint. For example, we might choose to consider intent to treat (ITT) as our primary analysis and a secondary measure that is per protocol and compliance based for confirmation of the effect. If the primary analysis shows that the treatment is beneficial, the effect should be even stronger in patients who are compliant.

Sometimes we choose to define our endpoint based on multiple outcomes. This can be the consideration of a composite endpoint where we are looking for measures that reflect a combination of outcomes such as

- satisfaction of all endpoints specified,
- satisfaction of at least one endpoint specified,
- a particular average or sum of individual scores, or
- the time of the occurrence of the earliest event among several considered

A composite endpoint involves the evaluation of every individual with respect to all outcomes. A co-primary endpoint involves the improvement in a treatment group on all endpoints. With a co-primary outcome, it is possible that for one treatment group, more subjects experience a symptomatic improvement and a physiologic improvement without the same subjects experiencing both successes. These approaches address the goals of determining if a treatment positively affects not only the symptoms but the disease mechanism as well.

When choosing to consider multiple endpoints, we can discuss the choice of statistical summary measures or the choice of clinical endpoints. Choice of statistical summary measure might be among the mean, median, geometric mean, proportion above some threshold, among others. Defining our endpoints clinically, we might choose to consider the differences between cause specific survival and overall survival or changing the period of time we sample over.

Adaptively choosing endpoints can be, similar to subgroup analyses, reflected as an inflation of our type I error increasing our chance of passing ineffective treatments on to Phase III. Moving

forward to Phase III with treatments with false positive effects, we have a lower PPV and we decrease the number of treatments we can identify when misusing resources on ineffective therapies passed at Phase II. The same conclusions are drawn when we consider the losses that are a result of data dredging. Again, we can control our false discovery rate by pre-specifying our analyses and performing each test using a much lower level of significance. Although a comprehensive exploration of these issues is not included, a presentation of the problems that exist and a discussion of the similarities between these issues and others that result in an inflation of the false discovery rate is useful and suggestive of more appropriate approaches to the Phase II to Phase III progression.

Chapter 6

Overall conclusions

In this research, we considered the progression of studies for investigating a new treatment, and discussed strategies in a framework that encompasses the period from the start of preliminary Phase II studies to the completion of confirmatory Phase III studies. Using a general notational framework for evaluating new treatments, we examined optimality criteria including not only the standard frequentist operating characteristics of type I error and power and the standard Bayesian criteria of positive and negative predictive values, but also the efficiency considerations of the number of new treatments identified in a setting with limited resources. We demonstrated that Phase II type I error and power are merely parameterizations for our treatment discovery approach. We explored process parameters for a binary prior to identify a general, comprehensive approach, emphasizing the relationship between frequentist and Bayesian criteria, and we illustrated how design parameters change when more general Bayesian priors are adopted.

Early trial termination for efficacy or futility with a group sequential approach within Phase II or Phase III improves efficiency in terms of the average number of subjects used for identifying effective therapies. Seamless Phase II/III approaches are proposed for eliminating the time spent between Phase II and Phase III. We considered how a single sequential design differs from the optimal approach of independent stages with respect to the timing of the analyses, the frequentist criteria met, and stopping boundaries specified. Pre-planned use of surrogate endpoints in Phase II causes issues with our approach in terms of the interpretability and applicability of results.

For addressing recent interest in adaptive clinical trial design, we explored how the traditional Phase II to Phase III approach fits in with the newer adaptive methods. Our traditional approach is similar to adaptive seamless Phase II/III with the time spent between phases (white space) used for adaptation being the only difference. The major issues of such approaches are the type I error / power tradeoffs and how they affect our PPV and efficiency. We discussed how powering of Phase III based on Phase II results mimics adaptive sample size re-estimation / re-powering of study. This

results in using too few subjects in Phase III for studying effective treatments with estimates that are biased upward after Phase II and using too many subjects in Phase III for studying ineffective treatments with estimates above the critical value for significance in Phase II.

In addition, the bias in our estimate of the treatment effect as a result of the lack of precision of small samples can be addressed with adjustment methods. The use of conditional estimators such as the biased adjusted mean (BAM) can be beneficial in the settings of screening or publication bias. However, we demonstrated that such adjusted analyses are not necessarily advantageous when re-powering a study between Phase II and Phase III (or at an interim analysis in a seamless Phase II/III design). For a binary prior, at the end of Phase II, a PPV of approximately 50% suggests that it will be impossible to provide an adaptive bias adjustment that would be appropriate for both a true effect of 0 and a true effect of 0.125. Instead, powering for the MCID in all phases of the investigation seems preferable and results in a more efficient use of resources.

In adapting to subgroups, we considered doing so unnecessarily (treatment effect is truly homogeneous) and doing so to good advantage (treatment only works in a particular subgroup). We considered three strategies for proceeding with subgroup analyses based on Phase II results. In the presence of homogeneous effects, we identify fewer effective treatments when searching for subgroup effects that do no exist. We considered three cases for describing heterogeneous effects. Among all cases, the best approach to consider for maximizing the number of effective treatments adopted is one which proceeds to Phase III with the subgroup with the smallest p-value in Phase II. However, when considering a prior belief about the effects of a treatment, properly powering Phase II studies based on that prior belief is important. We discussed the "tipping point" of our approach at which there is a balance between the number of treatments gained or lost when moving forward with subgroups under both homogeneous and heterogeneous treatment effect settings. Key to our findings, however, is the assumption that any subgroup effect was within one of the six subgroups defined by three variables. We have not explored how greatly increasing the number of subgroups or having subgroups of very small size might change our results. We stressed the importance of controlling the type I error when considering multiple comparisons and other adaptive methods that affect the credibility and reliability of trial results. Finally, we discussed how adapting to different summary measures or clinical outcomes also inflates our type I error but can be controlled.

Our research has some important limitations. We focused on the use of Phase II studies as screening trials for Phase III, a setting in which we investigate new treatment ideas with smaller trials and only proceed to larger, confirmatory trials with promising treatments. This is only a single setting among many used in practice. For most of our considerations, we primarily considered a binary prior. In implementing a continuous prior, we chose to examine a mixture of normal

distributions with means centered at no effect and the effect that is the MCID. To be more comprehensive, we would have to more carefully consider the MCID, which is generally poorly quantified. The cost of adopting effective therapies was defined based on the number of ineffective treatments adopted and the number of subjects used. Other definitions of cost were not considered. While considering a fixed set of resources for all approaches, we did not differentiate between the cost of a Phase II patient and the cost of a Phase III patient, but in practice differences can be substantial.

In our examination of seamless designs, we have opined that current standards of the Phase II to Phase III progression only differs from continuous/seamless adaptive approaches in the handling of the white space. We did not truly examine the use of Bauer and Koehne or Cui Hung and Wang etc. but note that our methods considered pre-specified adaptation, which allows use of sufficient statistics. Seamless adaptive methods that control type I error and power can prove valuable providing that a seamless design mimics type I error / power of the phased studies approach without a great loss of ethical oversight, regulatory oversight, or efficiency. Although we challenge whether the use of adaptations in the Phase II to Phase III process achieves the goals of efficiency where other methods do not, we recognize that newer adaptive methods may prove beneficial in formalizing the data dredging that often occurs during white space. We have not at all explicitly addressed safety issues (beyond noting that this is for what white space is used).

The settings and strategies considered when choosing to conduct subgroup analyses are not exhaustive. We chose to explore equal sized subgroups of independent variables (sex, age, weight) and acknowledge that the correlation between variables will affect outcomes. When controlling our errors, we did not explore other methods for splitting the type I error across subgroups. The settings and examples presented are simply a proposal of considerations to be made when progressing from Phase II to Phase III during the treatment discovery process. We did not simulate data for examining the consideration of different summary measures or clinical outcomes. Instead we noted that such analyses would result in an inflation of the type I error at Phase II, comparable to searching for subgroups, and indicated that similar concerns and methods apply.

Overall, we demonstrated that the optimal Phase II to Phase III progression defined by an acceptable positive predictive value (PPV) and a maximal number of effective treatments can be identified for an anticipated prevalence and hypothesized resources by a parameterization of type I error and power at Phase II. We recognize that several approaches lead to the same optimality criteria and that the chosen strategy will depend on the individual objectives of clinical researchers, trial sponsors, regulatory agencies, patients on study, and those who might benefit from new knowledge about the treatments being studied.

Appendix

The following tables supplement the results that were presented in previous chapters.

Table A.1: Examples of various Phase II to Phase III approaches with 5% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 95% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
500	512	0.100; 94%	24; 48	33%	72	10,476	0.025; 100%	24; 1.2	10,988	0.05	0.0025; 94%
600	322	0.085; 81%	24; 48	33%	72	11,086	0.025; 100%	24; 1.2	11,408	0.03	0.0021; 81%
800	829	0.100; 99%	40; 76	34%	116	2,913	0.025; 100%	40; 1.9	3,742	0.22	0.0025; 99%
900	403	0.090; 88%	40; 77	34%	117	5,467	0.025; 100%	40; 1.9	5,870	0.07	0.0022; 88%
1,000	322	0.085; 81%	40; 81	33%	121	5,588	0.025; 100%	40; 2.0	5,910	0.05	0.0021; 81%
1,100	709	0.100; 98%	54;105	34%	159	1,393	0.025; 100%	54; 2.6	2,102	0.34	0.0025; 98%
1,200	438	0.090; 90%	54;103	34%	157	3,029	0.025; 100%	54; 2.6	3,467	0.13	0.0023; 90%
1,300	345	0.085; 83%	54;105	34%	159	3,473	0.025; 100%	54; 2.6	3,818	0.09	0.0021; 83%
1,500	470	0.095; 92%	69; 135	34%	204	1,445	0.025; 100%	69; 3.4	1,915	0.25	0.0024; 92%
1,600	373	0.090; 86%	69; 137	33%	206	1,958	0.025; 100%	69; 3.4	2,331	0.16	0.0023; 86%
1,700	322	0.085; 81%	69; 137	33%	206	2,192	0.025; 100%	69; 3.4	2,514	0.13	0.0021; 81%

Table A.2: Examples of various Phase II to Phase III approaches with 5% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 98% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
500	696	0.040; 94%	24; 19	55%	43	15,342	0.025; 100%	24; 0.5	16,038	0.04	0.0011; 94%
600	461	0.035; 81%	24; 20	55%	43	16,350	0.025; 100%	24; 0.5	16,811	0.03	0.0009; 81%
800	1,091	0.035; 99%	40; 27	60%	67	1,924	0.025; 100%	40; 0.7	3,015	0.36	0.0009; 99%
900	568	0.035; 88%	40; 30	57%	70	7,027	0.025; 100%	40; 0.7	7,595	0.07	0.0008; 88%
1,200	637	0.030; 90%	54;34	61%	88	2,672	0.025; 100%	54; 0.9	3,309	0.19	0.0008; 90%
1,300	512	0.030; 83%	54;37	59%	91	3,676	0.025; 100%	54; 0.9	4,188	0.12	0.0007; 83%
1,600	510	0.040; 86%	69; 61	53%	130	1,414	0.025; 100%	69; 1.5	1,924	0.27	0.0010; 86%
1,700	461	0.035; 81%	69; 57	55%	126	1,728	0.025; 100%	69; 1.4	2,189	0.21	0.0009; 81%

Table A.3: Examples of various Phase II to Phase III approaches with 20% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 95% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N_3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
500	104	0.405; 85%	85; 162	34%	247	3,838	0.025; 100%	85; 4.0	3,942	0.03	0.0100; 85%
600	130	0.420; 89%	107; 202	35%	309	2,990	0.025; 100%	107; 5.0	3,120	0.04	0.0104; 89%
700	147	0.430; 91%	127; 241	35%	368	2,437	0.025; 100%	127; 6.0	2,584	0.06	0.0107; 91%
800	169	0.440; 93%	149; 282	35%	431	2,010	0.025; 100%	149; 7.0	2,179	0.08	0.0109; 93%
900	183	0.445; 94%	169; 320	35%	489	1,707	0.025; 100%	169; 8.0	1,890	0.10	0.0111; 94%
1,100	117	0.410; 87%	191; 361	35%	552	1,578	0.025; 100%	191; 9.0	1,695	0.07	0.0102; 87%
1,100	299	0.455; 98%	216; 400	35%	616	1,089	0.025; 99%	212;10.0	1,388	0.22	0.0114; 98%
1,400	174	0.430; 93%	260; 482	35%	742	1,020	0.025; 98%	255; 12.0	1,194	0.15	0.0107; 93%
1,800	110	0.390; 85%	306; 562	35%	868	924	0.025; 97%	296; 14.0	1,034	0.11	0.0097; 85%

Table A.4: Examples of various Phase II to Phase III approaches with 20% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 98% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N_3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
500	298	0.130; 85%	85; 52	62%	137	6,212	0.025; 100%	85; 1.3	6,510	0.05	0.2125; 98%
600	339	0.140; 89%	107; 67	61%	174	4,578	0.025; 100%	107; 1.7	4,917	0.07	0.2229; 98%
700	373	0.140; 91%	127; 78	62%	205	3,589	0.025; 100%	127; 2.0	3,962	0.09	0.2268; 98%
800	409	0.145; 93%	149; 93	62%	242	2,785	0.025; 100%	149; 2.3	3,194	0.13	0.2328; 98%
900	435	0.145; 94%	169; 104	62%	273	2,225	0.025; 100%	169; 2.6	2,660	0.16	0.2328; 98%
1,100	317	0.135; 87%	191; 119	62%	310	2,101	0.025; 100%	191; 3.0	2,418	0.13	0.2347; 98%
1,200	626	0.140; 98%	235; 134	64%	369	674	0.025; 90%	212; 3.4	1,300	0.48	0.2170; 98%
1,400	416	0.140; 93%	260; 157	62%	417	1,000	0.025; 98%	255; 3.9	1,416	0.29	0.2208; 98%
1,800	346	0.135; 89%	320; 194	62%	514	734	0.025; 92%	296; 4.9	346	0.32	0.2056; 98%

Table A.5: Examples of various Phase II to Phase III approaches with 50% prevalence of truly effective treatments, overall power, $\beta_2\beta_3 \geq 0.80$, and approximately 98% positive predictive value (PPV) overall

Phase II					Phase III				Overall		
Num of RCTs	N_2	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Pass	PPV	Num of RCTs	N_3	Type I err; Pwr	Eff; Ineff (TP); (FP) Drugs Adopt	$N_2 + N_3$	$\frac{N_2}{(N_2+N_3)}$	Type I err; Pwr
500	53	0.500; 82%	205; 125	62%	330	2,949	0.025; 100%	205; 3.1	3,002	0.02	0.8200; 82%
600	70	0.530; 87%	261; 159	62%	420	2,280	0.025; 100%	261; 4.0	2,350	0.03	0.8700; 87%
700	119	0.575; 94%	329; 201	62%	530	1,729	0.025; 100%	329; 5.0	1,848	0.06	0.9400; 94%
900	75	0.535; 88%	396; 241	62%	637	1,464	0.025; 100%	395; 6.0	1,539	0.05	0.8778; 88%
1,100	60	0.510; 84%	462; 281	62%	743	1,258	0.025; 99%	459; 7.0	1,318	0.05	0.8345; 83%

Table A.6: Stopping boundaries introduced at Phase II and/or Phase III on crude estimate and error spending scales

		Phase II Stopping boundaries			Phase III Stopping boundaries	
Analysis time	N_{2j}	Crude estimate of treatment effect	Cumulative type I or II error spent	N_{3j}	Crude estimate of treatment effect	Cumulative type I or II error spent
Pocock Futility boundary, $J = 2$						
Fut 1	189	0.0317	0.05328	529	0.0309	0.01469
Fut 2	378	0.0661	0.08500	1,059	0.0587	0.02500
OBF Futility boundary, $J = 2$						
Fut 1	176	-0.0040	0.02191	490	-0.0007	0.00258
Fut 2	352	0.0720	0.08500	980	0.0624	0.02500
OBF Efficacy, Pocock Futility boundary, $J = 2$						
Eff 1	197	0.1365	0.02781	534	0.1182	0.00316
Eff 2	393	0.0682	0.08500	1,068	0.0591	0.02500
Fut 1	197	0.0349	0.05498	534	0.0315	0.01482
Fut 2	393	0.0682	0.08500	1,068	0.0591	0.02500
OBF Efficacy, OBF Futility boundary, $J = 2$						
Eff 1	181	0.1479	0.02330	493	0.1256	0.00264
Eff 2	362	0.0739	0.08500	987	0.0628	0.02500
Fut 1	181	0.0000	0.02330	493	0.0000	0.00264
Fut 2	362	0.0739	0.08500	987	0.0628	0.02500
Pocock Futility boundary, $J = 4$						
Fut 1	101	-0.0299	0.03484	282	-0.0149	0.00911
Fut 2	203	0.0228	0.05695	564	0.0262	0.01577
Fut 3	304	0.0462	0.07276	846	0.0445	0.02088
Fut 4	406	0.0601	0.08500	1,128	0.0554	0.02500
OBF Futility boundary, $J = 4$						
Fut 1	90	-0.1662	0.00147	248	-0.1313	0.00003
Fut 2	179	-0.0090	0.01830	497	-0.0028	0.00211
Fut 3	269	0.0433	0.04915	745	0.0400	0.01046
Fut 4	358	0.0695	0.08500	993	0.0614	0.02500
OBF Efficacy, Pocock Futility boundary, $J = 4$						
Eff 1	111	0.2596	0.00316	292	0.2274	0.00005
Eff 2	221	0.1298	0.02780	584	0.1137	0.00302
Eff 3	332	0.0865	0.06259	876	0.0758	0.01315
Eff 4	443	0.0649	0.08500	1,168	0.0568	0.02500
Fut 1	111	-0.0201	0.03690	292	-0.0120	0.00935
Fut 2	221	0.0297	0.06013	584	0.0283	0.01618
Fut 3	332	0.0518	0.07630	876	0.0462	0.02140
Fut 4	443	0.0649	0.08500	1,168	0.0568	0.02500
OBF Efficacy, OBF Futility boundary, $J = 4$						
Eff 1	95	0.2958	0.00194	254	0.2512	0.00003
Eff 2	191	0.1479	0.02133	509	0.1256	0.00232
Eff 3	286	0.0986	0.05472	763	0.0837	0.01118
Eff 4	381	0.0739	0.08500	1,017	0.0628	0.02500
Fut 1	95	-0.1479	0.00194	254	-0.1256	0.00003
Fut 2	191	0.0000	0.02133	509	0.0000	0.00232
Fut 3	286	0.0493	0.05472	763	0.0419	0.01118
Fut 4	381	0.0739	0.08500	1,017	0.0628	0.02500

Table A.7: Examples of various approaches to subgroup analyses when there exist heterogeneous effects described by Case B (overall $\mu = 0.174/2$; males $\mu = 0.174$)

		No Subgroups	Prefer All Inflate error	Choose Subgroup Inflate error	Smallest P-value Inflate error
Phase II	Number RCT	2,130 (10% eff)	1,493 (10% eff)	1,493 (10% eff)	1,493 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 63.8%	0.337; 91.4%	0.337; 91.4%	0.337; 91.2%
	"Positive" RCT	136 eff; 192 not	136 eff; 453 not	136 eff; 453 not	136 eff; 453 not
Phase III	Number RCT	328 (41% eff)	591 (23% eff)	589 (23% eff)	589 (23% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 70.9%	0.025; 77.3%	0.025; 83.2%	0.025; 89.5%
	# Effctve Adopt	97	105	114	122
	For Overall	97	67	42	8
	For Males	0	30	59	88
	For Females	0	0	0	0
	For All Others	0	8.4	13.3	26.0
	# Ineff Adopt	5	11	11	11
Overall	PPV	95%	90%	91%	91%
	Type I err; Pwr	0.0026; 45.5%	0.0084; 70.6%	0.0084; 76.0%	0.0084; 81.6%

Table A.8: Examples of various approaches to subgroup analyses when there exist heterogeneous effects described by Case C (overall $\mu = 0.125$; males $\mu = 0.26$)

		No Subgroups	Prefer All Inflate error	Choose Subgroup Inflate error	Smallest P-value Inflate error
Phase II	Number RCT	2,047 (10% eff)	1,481 (10% eff)	1,481 (10% eff)	1,481 (10% eff)
	N per RCT	342	342	342	342
	Type I err; Pwr	0.100; 84.9%	0.337; 99.0%	0.337; 98.9%	0.337; 98.8%
	"Positive" RCT	174 eff; 184 not	147 eff; 449 not	146 eff; 449 not	146 eff; 449 not
Phase III	Number RCT	358 (49% eff)	596 (25% eff)	595 (25% eff)	595 (25% eff)
	N per RCT	829	829	829	829
	Type I err; Pwr	0.025; 95.0%	0.025; 95.6%	0.025; 97.4%	0.025; 99.1%
	# Effctve Adopt	165	140	143	145
	For Overall	165	120	67	9
	For Males	0	18	70	119
	For Females	0	0	0	0
	For All Others	0	2.8	6.2	17.0
	# Ineff Adopt	5	11	11	11
Overall	PPV	97%	93%	93%	93%
	Type I err; Pwr	0.0025; 80.6%	0.0084; 94.6%	0.0084; 96.3%	0.0084; 97.9%

References

Dirk Bassler, Victor M. Montori, Matthias Briel, Paul Glasziou, Gordon Guyatt (2008). Early stopping of randomized clinical trials for overt efficacy is problematic. Journal of Clinical Epidemiology 61: 241–246.

Dirk Bassler, Victor M. Montori, Matthias Briel, Melanie Lane, Paul Glasziou, Qi Zhou, Diane Heels-Ansdell, Stephen D. Walter, Gordon Guyatt, the STOPIT-2 Study Group (2010). Stopping randomized trials early for benefit and estimation of treatment effects. JAMA 303: 1180–1187.

Frank Bretz, Heinz Schmidli, Franz Konig, Amy Racine, Willi Maurer (2006). Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts. Biometrical Journal 48(4): 623–634.

Kathryn Chaloner, Isabella Verdinelli (1995). Bayesian Experimental Design: A Review. Statistical Science 10(3): 273–304.

John Chan, Stefanie Ueda, Valerie Sugiyama, Christopher Stave, et. al. (2008). The price of innovation: new estimates of drug development costs. Journal of Clinical Oncology 26(9): 1511–1518.

Shein-Chung Chow, Qingshu Lu, Siu-Keung Tse (2007). Statistical analysis for two-stage seamless design with different study endpoints. Journal of Biopharmaceutical Statistics 17(6): 1163–76.

Catherine De Angelis, Jeffrey M Drazen, Frank A. Frizelle (2004). Clinical trial registration: a statement from the International Committee of Medical Journal Editors. The Lancet 364(2004): 911–912.

David L. DeMets, James H. Ware (1980). Group sequential methods in clinical trials with a one-sided hypothesis. Biometrika 67(3): 651–660.

David L. DeMets, James H. Ware (1982). Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika 69(3): 661–663.

Scott S. Emerson, John M. Kittelson, Daniel L. Gillen (2007a). Frequentist evaluation of group sequential clinical trial designs. Statistics in Medicine, 26: 5047–5080.

Scott S. Emerson, John M. Kittelson, Daniel L. Gillen (2007b). Bayesian evaluation of group sequential designs. Statistics in Medicine. 26(7): 1431–1449.

Thomas Fleming (2005). Surrogate Endpoints And FDAs Accelerated Approval Process. Health Affairs 24(1): 67–78.

Thomas Fleming, John Powers (2012). Biomarkers and surrogate endpoints in clinical trials. Statistics in Medicine 31: 2973–2984.

Xiaoyin (Frank) Fan, David L. DeMets, K. K. Gordon Lan (2004). Conditional Bias of Point Estimates Following a Group Sequential Test. Journal of Biopharmaceutical Statistics 14(2): 505–530.

Boris Freidlin, Edward L. Korn (2009). Stopping clinical trials early for benefit: impact on estimation. Clinical Trials 6: 119–125.

Eric Holmgren (2007). Are Phase 2 screening trials in oncology obsolete? Statistics in Medicine 27; 556–567.

Sally Hunsberger, Yingdong Zhao, Richard Simon (2009). A Comparison of Phase II Study Strategies. Clinical cancer research : an official journal of the American Association for Cancer Research 15(19): 5950–5955.

Christopher Jennison, Bruce W. Turnbull. Group Sequential Methods with Applications to Clinical Trials. Boca Raton, USA: Chapman and Hall/CRC, 2000.

Peter K. Kimani, Nigel Stallard, Jane L. Hutton (2009). Dose selection in seamless phase II/III clinical trials based on efficacy and safety. Statistics in Medicine 28: 917–936.

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

Ismail Kola, John Landis (2004). Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery 3(8): 711–716.

Edward L. Korn, Boris Freidlin, Jeffrey S. Abrams, Susan Halabi (2012). Design Issues in Randomized Phase II/III Trials. Journal of Clinical Oncology 30(6): 667–671.

Gregory P. Levin, Sarah C. Emerson, Scott S. Emerson (2012). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. Statistics in Medicine 32(8): 1259–1275.

P.Y Liu, Michael LeBlanc, Manisha Desai (1999). False Positive Rates of Randomized Phase II Designs. Controlled Clinical Trials 20(4): 343–352.

Loren Mell, Jong-Hyeon Jeong (2010). Pitfalls of Using Composite Primary End Points in the Presence of Competing Risks. Journal of Clinical Oncology 28(28): 4297–4299.

Peter C. O'Brien, Thomas R. Fleming (1979). A Multiple Testing Procedure for Clinical Trials. Biometrics 35(3): 549–556.

John Orloff, Frank Douglas, Jose Pinheiro, Susan Levinson, et. al. (2009). The future of drug development: advancing clinical trial design. Nature Reviews Drug Discovery 8; 949–957.

Stuart. J. Pocock (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika 64(2): 191–199.

Stuart. J. Pocock (1982). Interim analyses for randomized clinical trials: The group sequential approach. Biometrics 38: 153–162.

Lawrence V. Rubinstein, Edward L. Korn, Boris Freidlin, Sally Hunsberger, S. Percy Ivy, Malcom A. Smith (2005). Design Issues of Randomized Phase II Trials and a Proposal for Phase II Screening Trials. Journal of Clinical Oncology 28(28); 7199–7206.

Kenneth F. Schulz, David A. Grimes (2005). Multiplicity in randomised trials II: subgroup and interim analyses. The Lancet 365(9471): 1657–1661.

Nigel Stallard, John Whitehead, Susan Todd, Anne Whitehead (2001). Stopping rules for phase II studies. British Journal of Clinical Pharmacology 51(6): 523–529.

Nigel Stallard, Sue Todd (2010). Seamless phase II/III designs. Statistical Methods in Medical Research 20(6): 623–634.

Jerome M. Stern, R John Simes (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ. 315: 640–645.

Susan Todd, Anne Whitehead, Nigel Stallard, John Whitehead (2001). Interim analyses and sequential designs in phase III studies. British Journal of Clinical Pharmacology 5(5): 394–399.

John Whitehead. The Design and Analysis of Sequential Clinical Trials. Chichester, UK: John Wiley & Sons Ltd, 1997.

Jenny Zhang, Gideon Blumenthal, Kun He, Shenghui Tang, Patricia Cortazar, Rajeshwari Sridhara (2012). Overestimation of the effect size in group sequential trials. Clinical Cancer Research 18(18): 4872–4876.