Bayesian Adaptive Designs in Phase I/II Clinical Trials

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Wei Zhong

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Bradley P. Carlin

Joseph S. Koopmeiners

September, 2012

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

I would express my deep gratitude to my two advisors, Brad Carlin and Joe Koopmeiners, for their encouragement, support, and invaluable guidance on this work. Brad is the one professor who truly made a difference in my life. His energy, knowledge, and efficiencies helped me through a difficult time when my research topic changed; he also provided many suggestions on future career development. Joe has spent many hours providing me sound advice, modifying my computer programs, and teaching me how to think and work as a good biostatistician.

I am very grateful to my two other committee members, Haitao Chu and Daniel Duprez, for taking the time out of their busy schedules to review and help improve my thesis.

I would also like to thank my intern supervisors, Paul Gallo, Jeff Maca, David Olhssen (at Novartis), Xin Wang and Yifan Huang (at Pfizer), for discussing my intern projects and guiding the early directions of Chapter 4.

My sincere thanks go to Daniel Duprez, Waddah Al-Refaie, Elizabeth Habermann, Tracy Bergemann, Greg Grandits and Wei Pan, for funding the RA projects I was involved in and providing significant assistance during the past few years.

Finally, I am deeply and forever indebted to my parents and my wife Tianli for their love, trust and support throughout my life.

Dedication

This thesis is dedicated to my ever loving companion, Tianli Wang, for her love, encouragement and understanding these years. She has been my inspiration and motivation for continuing to improve my knowledge and pursue my career goal. She is my rock and the sunshine of my day.

Abstract

Recently, many Bayesian methods have been developed for dose-finding when simultaneously modeling both toxicity and efficacy outcomes in a blended phase I/II fashion. A further challenge arises when all the true efficacy data cannot be obtained quickly after the treatment, so that surrogate markers are instead used (e.g, in cancer trials). In this thesis, we first propose a framework to jointly model the probabilities of toxicity, efficacy and surrogate efficacy given a particular dose. The resulting trivariate algorithm utilizes all the available data at any given time point, and can flexibly stop the trial early for either toxicity or efficacy. Our simulation studies demonstrate our proposed method can successfully improve dosage targeting efficiency and guard against excess toxicity over a variety of true model settings and degrees of surrogacy.

Second, we offer a brief catalog of more flexible semiparametric and nonparametric monotone link functions to model the marginal probability of efficacy based on our proposed trivariate binary model. We show via simulation that our flexible link methods can outperform standard parametric CRM approaches in terms of both the probability of correct dose selection and the proportion of patients treated at that dose.

Finally, frequentist sample size determination for binary outcome data usually requires initial guesses of the event probabilities, which may lead to a poor estimate of the necessary sample size. We propose a new two-stage Bayesian design with sample size reestimation at the interim stage. Our design inherits the properties of good interpretation and easy implementation, generalizing an earlier method to a two-sample

setting, and using a fully Bayesian predictive approach to reduce an overly large initial sample size when necessary. Moreover, our design can be extended to allow patient level covariates via logistic regression, now adjusting sample size within each subgroup based on interim analyses. We illustrate the benefits of this approach with a design in non-Hodgkin lymphoma with a simple binary covariate (patient gender), offering an initial step toward within-trial personalized medicine.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

An adaptive design is a clinical trial design that uses accumulating data from the ongoing trial to modify certain aspects of the study without undermining its validity and integrity [1]. In adaptive designs, changes according to what is learned from the accumulating data are made to enhance the trial, either by improved learning or cutting short an unpromising trial. The flexibility of adaptive designs may lead to better treatment of patients, more efficient drug development, and better use of knowledge from other studies. Adaptive trial designs can be applied in a variety of areas, including dose finding, early stopping for efficacy, safety, or futility, seamless phase I/II or II/III designs, adaptive randomization, and sample size reestimation [2, 3]. The procedures of data review, decision making, and decision implementation during adaptive trials should maintain long-run trial integrity (e.g, good Type I error and power), which is very important and is required by regulators in phase III. In February 2010, the Food and Drug Administration (FDA) put out a draft nonspecific guidance on adaptive designs, which is generally viewed as supportive of the proper use of adaptive designs in clinical trials [4].

1.1 Adaptive Dose Finding in Phase I/II Studies

In traditional phase I clinical trials, we seek the maximum tolerated dose (MTD) of an investigational agent, which represents the highest dose with toxicity probability less than a physician-specified acceptable maximum. Based on the estimated MTD from a phase I trial, a phase II trial may be conducted to test the agent's efficacy and possibly refine the optimal dosage for further studies. There are two general classes of phase I clinical trial designs based on dose assignment: rule-based designs, and model-based

designs [3]. Rule-based designs include pharmacologically two-stage designs [5], and the traditional 3+3 design [6] and its variations. Among model-based designs, the continual reassessment method, or CRM [7], is a Bayesian design that has been repeatedly shown to have better average operating characteristics than rule-based designs. The CRM links the true toxicity probabilities and dose levels through a simple one-parameter function such as a one-parameter hyperbolic tangent, logistic or power function, and updates the posterior estimates of the MTD arising from this parameter continuously as patients are enrolled. A variety of popular modifications and refinements, largely aimed at protecting patients in the trial from excessive doses, have been proposed [8, 9, 10].

A limitation of the preceding designs is that efficacy is ignored, and dose-finding is based only on toxicity. This is problematic in settings where a dose exists that is less than the MTD, but further escalation would not result in increased efficacy. Moreover, due to the limited number of available patients in oncology, each patient should be recognized as a valuable resource. Therefore, it would be wise to collect and utilize as much relevant information, including efficacy data in phase I studies, as possible from each patient. Thus a better dose-finding approach is to incorporate both toxicity and efficacy in a blended phase I/II fashion. With this goal in mind, many statistical methods have been developed for simultaneously modeling both toxicity *and* efficacy in clinical trials [11, 12, 13, 14, 15, 16].

We now briefly mention a few recent extensions to the CRM, especially those designed to handle multiple outcomes. One class of approaches actually transform bivariate binary outcomes to univariate trinomial outcomes: no efficacy or toxicity, efficacy without toxicity, and toxicity with or without efficacy [12, 17, 18], then a proportional

odds model or a continuation-ratio model is used for the trinomial outcomes. Other approaches establish the joint model for both binary outcomes in a direct way. Braun (2002) extended univariate CRM to a bivariate CRM (bCRM) design by constructing a conditional probability model for both efficacy and toxicity [14]. Thall and Cook (2004) jointly modeled efficacy and toxicity using a bivariate Gumbel copula [19] and introduced an efficacy-toxicity trade-off contour in two-dimensional space to guide dosage selection [15]. Bekele and Shen (2005) established a probit model with latent variables to jointly investigate a binary and a continuous outcome [20]. Furthermore, to avoid misspecification of the dose-toxicity curve by the use of the rigid parametric link function, Gasparini and Eisele (2000) proposed a curve-free method by introducing a flexible product-of-beta prior (PBP) [21].

1.2 The Use of Surrogate Markers

Phase I designs that consider efficacy and toxicity rely on the timely availability of the efficacy and toxicity outcomes. In practice, however, it is not uncommon for toxicity to be available in a short timeframe, while a relatively long wait is required to observe efficacy. A possible solution to this problem is the use of surrogate markers as end points for efficacy. In a widely read article, Prentice (1989) defines a surrogate marker as, "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint" [22]. Similarly, Temple (1999) defines it as "a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint, that is a direct measure of how a patient feels,

functions, or survives and is expected to predict the effect of the therapy" [23]. Use of such surrogates, when justified, can lead to significant reductions in trial cost and duration. For instance, in cancer trials efficacy may be defined as a survival outcome after some fairly long period of time, say 5 years. In such cases, surrogate markers, such as a seriological measure a physician might check soon after treatment administration, are commonly used for guiding decisions about whether the intervention is promising enough to justify a large definitive trial with clinically meaningful outcomes [24].

A fairly common situation in oncology is when toxicity and surrogate efficacy may be obtained quickly, whereas only part or none of the final efficacy data might be available. Using the previous bivariate models, we can simply replace the missing efficacy data with the surrogate efficacy data in our analysis. However, this could be misleading in the case of a weak surrogate; the mixture of the efficacy and surrogate efficacy data might lead to incorrect conclusions since the primary and surrogate endpoints may react quite differently to the drug. As observed by many authors, the uncritical use of surrogate markers is fraught with danger, since the effect of the intervention on the surrogate may not reliably predict the overall effect on the clinical outcome [24]. Therefore, a possible compromise solution is to include any available data on *both* the exact and surrogate endpoints in the analysis, in the hope that positive (but not perfect) association between the two will lead to better estimation and testing than could be achieved by using either endpoint by itself.

In such settings, more sophisticated biostatistical methods are needed to properly utilize all available information. Development of such methods is the overarching goal of

Chapters 2 and 3 of this dissertation. Specifically, Chapter 2 considers fairly straightforward parametric models, while Chapter 3 describes several more flexible nonparametric models.

1.3 Sample Size Reestimation

Chapter 4 turns to the problem of sample size reestimation, another in-trial adaptive technique. Traditional sample size determination for two-sample binary efficacy data in a frequentist approach is simple, straightforward, and has been implemented in many clinical trials. For example, consider a two-arm clinical trial that compares the effect of two treatments, where we are interested in testing the hypotheses $H_0 : p_1 = p_2$ versus $H_a : p_1 > p_2$, where p_1 and p_2 denote the true event rates in the two treatment groups. To obtain the sample size given some pre-specified significance level α and power β, we must first set some target point estimates of p_1 and p_2 as crude guesses of the event probabilities for two treatments, denoting them as p_1^* and p_2^*, respectively. The designed detectable effect is then $\theta^* = p_1^* - p_2^*$. The sample size can be calculated by the following standard formula [25],

$$n \text{ per group} = \frac{2(Z_{1-\alpha/2} + Z_\beta)^2 \bar{p^*}(1 - \bar{p^*})}{\theta^{*2}}, \tag{1.1}$$

where the average event rate $\bar{p^*} = (p_1^* + p_2^*)/2$, and Z_γ denotes the γ percentile of the standard normal distribution.

Since the selection of the p_1^* and p_2^* are usually based on fairly vague prior knowledge or other studies with small sample sizes, the credibility of the "working alternative hypothesis" that $p_1 = p_1^*$ and $p_2 = p_2^*$ is often questionable [26]. Misspecification of

the event rates may lead to a poor estimate of the necessary sample size [27]. To fix this problem, many sequential designs and adaptive sample size designs incorporating interim analyses have been proposed in recent years [28, 29, 30, 31, 32, 33]. All these methods can provide substantial improvement by adjusting the sample size to achieve the target power while preserving the overall Type I error. However, previous sample size reestimation methods are based on an implicit assumption that estimates of the true unknown treatment effect do not change appreciably over time. In real life situations, this assumption is questionable, especially when more subject-level variability exists in the early recruitment period. A good specification of the expected treatment effect is still required for these frequentist designs.

In contrast, the Bayesian approach considers the treatment effect to be random variable having some distribution, and updates the prior with the data, obtaining a posterior distribution for inference. The interpretation of a credible interval for the treatment effect seems more natural here than that of the traditional frequentist confidence interval. Moreover, the objective of a phase II trial is to accept or reject a new drug for further investigation in a phase III trial, rather than obtain a highly precise estimate of each possible response rate. Generally there are three classes of Bayesian methods for sample size determination. First, a frequentist-Bayesian hybrid approach [34, 35, 36, 37], which considers the predictive probability of achieving the primary study goal based on the available data, but still aiming to control type I error. Second, some Bayesians recommend an interval length-based approach [38, 39, 40], which uses the length of posterior credible intervals as the sample size criterion. Finally, some authors pursue a fully decision-theoretic approach [3, 41, 42, 43], which chooses sample size to maximize

an investigator-selected utility function or minimize a corresponding loss function.

The Bayesian sample size proposed by Whitehead et al. (2008) for exploratory studies on efficacy is an interval length-based approach, but includes an analogy to frequentist Type I and II errors. These authors argue that "the trial should be large enough to ensure that the data collected will provide convincing evidence either that an experimental treatment is better than a control or that it fails to improve upon control by some clinically relevant difference." Like frequentist designs, the expected treatment effect is explicitly set in the design. But the Whitehead et al. sample size does not aim to meet certain power criteria under the alternative hypothesis. Instead, the acceptable minimum sample size N is justified by a "conclusiveness" condition. In the context of a one-sample test for a binary outcome (say, efficacy), it specifies that, regardless of the data, at least one of the two following probability statements should be satisfied at the end of a trial:

$$Pr(p > 0|Y^N) \geq \eta_1 \text{ or } Pr(p < \theta^*|Y^N) \geq \eta_2, \tag{1.2}$$

where $p \in [0, 1]$ denotes the success rate for the treatment, $\theta^* \in [0, 1]$ is the expected (or desired) treatment effect, and Y^N represents any possible dataset of N patients. The threshold probabilities η_1 and η_2 are selected to reflect the degree of certainty we require for convincing evidence, with both values typically close to 1.

One potential problem is that such a sample size might be too conservative. Adding an interim stage to reestimate the sample size might offer a solution, dramatically reducing the sample size where the interim information about the true treatment effect emerges as sufficiently conclusive. Moreover, the corresponding Bayesian approach for comparing two proportions is not discussed by Whitehead et al. (2008) and merits

further exploration.

At the interim stage, one can calculate the predictive power based on the interim posterior estimates of the parameters. The predictive power is actually the "re-estimated" power based on the prior and the data. Thus, a Bayesian approach to sample size estimation seems more sensible and natural here. However, in contrast to the frequentist literature, sample size reestimation has been infrequently discussed in the Bayesian setting. Some Bayesians argue that Bayesian analysis is a naturally sequential procedure, and are thus unconcerned about Type I error inflation resulting from multiple interim looks. Patient recruitment should depend on the data available at that time, and the adequacy of the resulting predictive power for making a final decision. However in practice, the sample size is usually determined before starting the trial and the schedule of interim analyses is also fixed; many trialists feel it is inappropriate to adjust the recruitment plan during the trial. Sample size reestimation, a key factor in interim analysis, is thus relevant in Bayesian design as well. Wang (2007) applies a Bayesian predictive approach to interim sample size reestimation, and compares it to other approaches such as predictive and conditional power approaches [44]. The author recommends its application in exploratory studies, where knowledge about a test drug is still uncertain, and the adaptive sample size is based on the predictive probability of trial success.

"Personalized medicine" is a subject of intense discussion in recent years. The concept refers to the tailoring of treatments to individuals based on personal characteristics, and represents the next step in drug therapy and development toward better understanding of disease and health [45]. The field is closely related to subgroup analysis, a subject of longstanding interest to trialists. For example, a recent study suggested no

improvement in the overall mortality of patients with coronary disease whether treated with percutaneous coronary intervention (PCI) or coronary-artery bypass surgery. But the results also showed age played a key role, with much lower mortality after surgery among patients 65 years or older, while lower mortality after PCI among those 55 years or younger [46]. Although many observational studies and pooled trials have contributed to our understanding of treatment effects at the individual level through subgroup effect analyses and development of prediction rules, a significant obstacle to the implementation of a personalized approach to trials themselves is the lack of appropriately designed studies [47]. Sample size estimation is an important issue for adequate trial design when we seek to study subgroup effects, especially in view of the well-known risk of Type I error inflation resulting from subgroups chosen post-hoc.

The remainder of this thesis proceeds as follows: In Chapter 2, we present a trivariate continual reassessment method for phase I/II trials of toxicity, efficacy and surrogate efficacy in the context of a phase I clinical trial of a novel NK cell treatment for non-Hodgkin lymphoma. In Chapter 3, we describe more flexible semiparametric and nonparametric monotone link functions in the setting of Chapter 2. Chapter 4 discusses Bayesian sample size reestimation in exploratory studies, and explores an extension of this work to the context of personalized medicine where the trial is stratified by an important binary covariate. Finally, we conclude by summarizing our findings and suggesting some areas for future work in Chapter 5.

Chapter 2

A Trivariate Continual Reassessment Method for Phase I/II Trials of Toxicity, Efficacy, and Surrogate Efficacy

Chapter 2 is aimed to handle a situation in an oncology trial when all the true efficacy data cannot be obtained quickly after the treatment, so that surrogate markers are instead used. We propose a framework to jointly model the probabilities of toxicity, efficacy and surrogate efficacy given a specific dose. Our trivariate binary model is specified as a composition of two bivariate binary submodels. In particular, we extend the bCRM approach, as well as utilize the Gumbel copula of Thall and Cook [15]. The full Bayesian design consists of three elements: a trivariate binary model, a set of sensible prior distributions, and a dose-finding algorithm. Given these elements, we can repeatedly generate artificial data from our design, and thus simulate its (Bayesian or frequentist) operating characteristics, notably the empirical probabilities of correct dose selection, and the proportions of trial participants assigned to each dose.

Chapter 2 is structured as follows: Section 2.1 presents a motivating example for dose finding. In Section 2.2 we introduce a general framework for the trivariate probability model, and our two preferred parametric model specifications. Bayesian prior selection and a dose-finding algorithm are addressed as well. Section 2.3 presents our simulation results under different scenarios, and provides guidance on specifying a future design. Finally, in Section 2.4 we discuss the strengths and limitations of our proposed method, and discuss areas for further investigation.

2.1 Motivating Example

We motivate our design by a Phase I dose-escalation study to evaluate a novel Natural Killer (NK) cell treatment for patients with non-Hodgkin lymphoma. The goal of this study is to evaluate the safety of the treatment, and to identify an optimal dose for

further evaluation in Phase II. Unlike standard chemotherapeutic agents, we anticipate that there is a dose where further escalation would increase the probability of toxicity without a corresponding increase in the probability of efficacy. In this sense, the optimal dose for further investigation in Phase II is likely to be less than the MTD, and we must consider the tradeoff between efficacy and toxicity during dose escalation.

As is typical in phase I, efficacy and toxicity are evaluated as dichotomous outcomes. In our case, toxicity is defined as any grade 3 or higher toxicity during the first 6 weeks, and efficacy is defined as tumor response at week 15. The timing of these outcomes poses an obvious problem. The nearly two month delay between evaluation of the toxicity and efficacy endpoints would delay enrollment of new cohorts, and increase the overall length of our study to a point where the design is no longer practical.

Fortunately, in addition to the efficacy and toxicity outcomes, absolute lymphocyte count will be measured at week 2. Absolute lymphocyte count responds quickly to NK cell treatments and is often used as a surrogate for response to NK cell treatment. An absolute lymphocyte count greater than 1000 cells/μl is thought to predict tumor response at 15 weeks. One approach to overcoming the long delay between the evaluation of toxicity and the evaluation of efficacy is thus to consider the surrogate endpoint in place of the efficacy endpoint, but this could lead to incorrect conclusions about the efficacy of our drug if the surrogate endpoint does not predict true efficacy as well as anticipated. Ideally, we would prefer a dose-escalation study that makes use of the surrogate endpoint but also incorporates efficacy information as it becomes available.

In an adaptive fashion, we design a trial which enrolls a new cohort with sample size $c = 3$ every 6 weeks, this implies that at the enrollment time of the mth ($m \geq 3$)

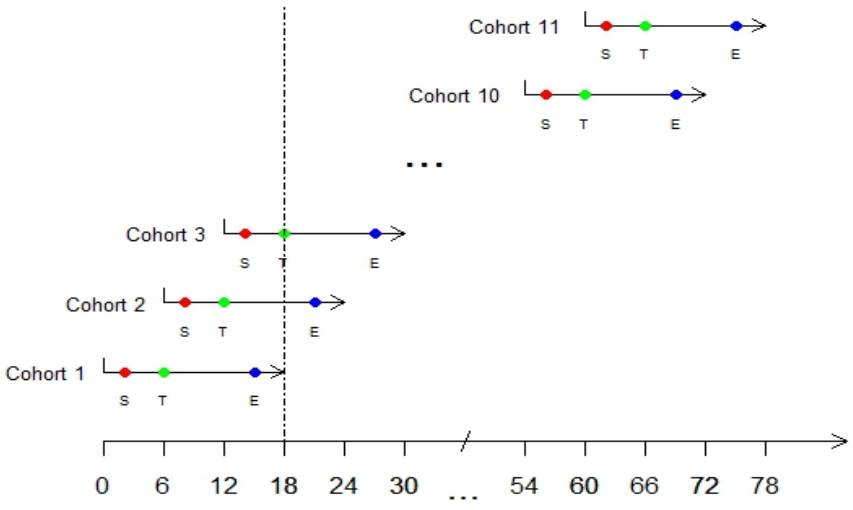

Figure 2.1: The patient enrollment plan and the available toxicity, efficacy and surrogate efficacy data at week 18 when deciding the dose assignment for Cohort 4. At this time point, the toxicity and surrogate efficacy data is available for all the first 3 cohorts of patients while the efficacy data is only available for Cohort 1.

cohort, the efficacy data of the $(m-2)th$ and $(m-1)th$ cohorts are unavailable. The maximum number of cohorts, L, is set to 11. Figure 2.1 shows the patient enrollment plan as well as the available data at a specific enrollment time (e.g, week 18).

2.2 Methods

2.2.1 Trivariate Probability Model

Let $Y_{ij} = (T_{ij}, S_{ij}, E_{ij})$ be the binary indicators of toxicity (T), surrogate efficacy (S), and efficacy (E) for subject i who receives a drug treatment at dose X_j. The joint

trivariate distribution can be decomposed into three parts as

$$Pr(T_{ij} = t, E_{ij} = e, S_{ij} = s)$$
$$= Pr(T_{ij} = t)Pr(E_{ij} = e \mid T_{ij} = t)Pr(S_{ij} = s \mid T_{ij} = t, E_{ij} = e), \quad (2.1)$$

where t, e and $s \in \{0, 1\}$. If we assume conditional independence of S and T given E, the above joint distribution can be simplified to

$$Pr(T_{ij} = t, E_{ij} = e, S_{ij} = s) = Pr(T_{ij} = t)Pr(E_{ij} = e \mid T_{ij} = t)Pr(S_{ij} = s \mid E_{ij} = e). \quad (2.2)$$

Although this conditional independence assumption may initially seem a bit strong, note that we do not assume *marginal* independence of S and T; only that their correlation is accounted for by E. Moreover, our model is consistent with the following latent process that might plausibly generate the trivariate binary outcomes:

$$\text{Toxicity} \Rightarrow \text{Efficacy} \Rightarrow \text{Surrogate efficacy}$$

Biologically, we can think of the data arising from two correlated latent process: a latent process for toxicity and a latent process for efficacy. The efficacy and surrogate efficacy are both generated by the latent process for efficacy. We believe that the correlation between the two latent processes is fully captured by the conditional probability of efficacy given toxicity, in which case it is reasonable to assume that surrogate efficacy is independent of toxicity conditional on efficacy. In this way, the trivariate joint distribution can be represented as a product of one marginal distribution for T and two conditional distributions for E and S respectively. Therefore we can flexibly apply various parametric link functions for the marginal or conditional submodels, as we now describe.

2.2.2 Parametric Functions of Submodels

To monitor the marginal probabilities of T, E and S (p_{Tj}, p_{Ej} and p_{Sj} respectively) given dose X_j, we apply three simple logistic regression models as follows:

$$log\left(\frac{p_{Tj}}{1-p_{Tj}}\right) = \alpha_T + \beta_T X_j, \tag{2.3}$$

$$log\left(\frac{p_{Ej}}{1-p_{Ej}}\right) = \alpha_E + \beta_E X_j, \tag{2.4}$$

$$\text{and } log\left(\frac{p_{Sj}}{1-p_{Sj}}\right) = \alpha_S + \beta_S X_j, \tag{2.5}$$

where α_T, α_S and α_E are assumed to be negative to account for the small probabilities, while β_T, β_S and β_E are assumed to be positive, since both efficacy and toxicity are assumed certain to be increasing with dose. This parametric model is commonly seen in many phase I designs, which assumes a monotonic relationship between dose levels and outcomes.

It is more complicated once we involve the two conditional probabilities ($Pr(E_{ij} = e \mid T_{ij} = t)$ and $Pr(S_{ij} = s \mid E_{ij} = e)$). For a bivariate binary distribution, we investigate approaches endorsed by Braun (2002) and Thall and Cook (2004), respectively [14, 15]. Suppose Z_1 and Z_2 are two binary random variables, or $z_1, z_2 \in \{0, 1\}$. Based on the work of Arnold and Strauss [48], Braun [14] suggests the following copula model for the conditional probability of $Z_1 = 1$ given $Z_2 = z_2$, namely a Bernoulli with success probability

$$Pr(Z_1 = 1|Z_2 = z_2) = \frac{p_1\phi^{z_2}(1-\phi)^{(1-z_2)}}{p_1\phi^{z_2}(1-\phi)^{(1-z_2)} + (1-p_1)(1-\phi)}. \tag{2.6}$$

Note this reduces to p_1 when $Z_2 = 0$ and $\frac{p_1\phi}{p_1\phi+(1-p_1)(1-\phi)}$ when $Z_2 = 1$. Here, ϕ captures the association between Z_1 and Z_2, with $\phi = \frac{1}{2}$ implying independence between Z_1 and Z_2, $\phi > \frac{1}{2}$ indicating positive association, and $\phi < \frac{1}{2}$ indicating negative association. Note that (2.6) equals p_1 for all Z_2 where $\phi = \frac{1}{2}$, clarifying the independence case. A drawback to this specification is that the joint bivariate distribution of Z_1 and Z_2 is not available as a standard family. Still, this model specification can often lead to trial designs with good operating characteristics. To adapt it to our framework, we need only set Z_1 as E and Z_2 as T to establish $Pr(E|T)$, say with association parameter ϕ_1. Similarly, an association parameter ϕ_2 can be used in a conditional model for $Pr(S|E)$. This extension of Braun's method is denoted by "ExB"(extended Braun) method in this paper.

Another approach to studying bivariate binary outcomes is the Gumbel copula utilized by Thall and Cook [15]. The joint bivariate Gumbel copula distribution is specified as

$$Pr(Z_1 = z_1, Z_2 = z_2) =$$
$$p_1^{z_1}(1-p_1)^{(1-z_1)}p_2^{z_2}(1-p_2)^{(1-z_2)} + (-1)^{z_1+z_2}p_1(1-p_1)p_2(1-p_2)\frac{e^\gamma - 1}{e^\gamma + 1}, \quad (2.7)$$

where p_1 and p_2 are the *marginal* probabilities success for Z_1 and Z_2, respectively; note that their interpretations have changed somewhat from p_1 and p_2 in the Braun model. Similar to ϕ, γ captures the association between Z_1 and Z_2, but now where the range of γ covers the whole real line and $\gamma = 0$ corresponds to independence, which is immediately apparent from (3.10). Positive values of γ indicate positive association, while negative values imply negative association. Then the conditional probability of

$Z_1 = 1$ given $Z_2 = z_2$ can be easily calculated as

$$\begin{aligned} Pr(Z_1 = 1|Z_2 = z_2) &= \frac{Pr(Z_1 = 1, Z_2 = z_2)}{Pr(Z_1 = 1, Z_2 = z_2) + Pr(Z_1 = 0, Z_2 = z_2)} \\ &= p_1 + (-1)^{z_2+1} p_1(1-p_1) p_2^{1-z_2}(1-p_2)^{z_2} \frac{e^\gamma - 1}{e^\gamma + 1}, \end{aligned}$$

which is distinct from (2.6); in particular we note the dependence on p_2. In the same fashion as our previous model extension, two association parameters γ_1 and γ_2 can be used in the two required conditional probabilities $Pr(E|T)$ and $Pr(S|E)$. We refer to this extension of the Gumbel copula as the "ExG"(extended Gumbel) method in what follows.

2.2.3 Likelihood and Prior Specification

As with all Bayesian analyses, a full likelihood and a prior distribution for every parameter are required. Following Braun (2002), α_T, α_E and α_S in equations (2.3), (2.4) and (2.5) are all set equal to the constant -3, to reflect the relative rarity of response with the lowest doses of the agent. Suppose that n_j patients are treated at dose X_j $(j = 1, \ldots, k)$, among which n_j^{tes} patients have outcomes $T = t$, $E = e$ and $S = s$. Then the likelihood for the "ExB" model in the complete toxicity, efficacy and surrogate efficacy data is multinomial,

$$L_C(\beta_T, \beta_S, \beta_E, \phi_1, \phi_2 \mid \text{Data}) \propto \prod_{j=1}^{k} {\pi_j^{000}}^{n_j^{000}} {\pi_j^{001}}^{n_j^{001}} {\pi_j^{010}}^{n_j^{010}} {\pi_j^{011}}^{n_j^{011}} {\pi_j^{100}}^{n_j^{100}} {\pi_j^{101}}^{n_j^{101}} {\pi_j^{110}}^{n_j^{110}} {\pi_j^{111}}^{n_j^{111}}, \quad (2.8)$$

where π_j^{tes} represents the joint probability of T, E, and S given dose X_j, and (ϕ_1, ϕ_2) is replaced by (γ_1, γ_2) in the "ExG" model. The joint probabilities π_j^{tes} are computed as

described in the previous two subsections. However, as illustrated in figure 2.1, part or all of the efficacy data are missing at the time of dose assignment so that n_j^{tes} for all the subjects cannot be fully determined. Assuming the efficacy data of subject i treated at dose j is unavailable at a specific time, the likelihood function for subject i is

$$L_i = Pr(T_{ij} = t, S_{ij} = s) = \quad Pr(T_{ij} = t)\ \{Pr(E_{ij} = 1|T_{ij} = t)Pr(S_{ij} = s|E_{ij} = 1) + Pr(E_{ij} = 0|T_{ij} = t)Pr(S_{ij} = s|E_{ij} = 0)\} \tag{2.9}$$

If n_j^{tes} denotes the corresponding number of patients with complete data, then the likelihood function is constructed as

$$L(\beta_T, \beta_S, \beta_E, \phi_1, \phi_2 \mid \text{Data}) \propto L_C \prod_{i=1}^{h} L_i, \tag{2.10}$$

where h represents the number of patients with unobserved efficacy data. Note $n_j^{tes} = 0$ at the early stage, which suggests no efficacy data are available. Imputation for the efficacy data of subject i can be easily handled by `BUGS` software.

Turning to the priors, we assume β_T, β_S and β_E all independently follow exponential distributions with mean 1. In the "ExB" method, ϕ_1 is assumed to follow a $Beta(2, 2)$ distribution and ϕ_2 a $Beta(4, 2)$ distribution, which encourages prior independence between E and T, but positive dependence a priori between E and S. In the "ExG" method, the priors for γ_1 and γ_2 are set as normal distributions ($N(0, 5)$ and $N(0.69, 5)$, respectively), priors designed to match the two beta priors as closely as possible, but on the γ scale.

2.2.4 Dose-Finding Algorithm

Suppose at a specific enrollment time point, Y denotes all the available accumulated data. Let $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$ be the posterior mean probabilities of toxicity and efficacy. After each look at the data, [14] suggests updating $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$, and then selecting the dose by minimizing

$$dist_j = \sqrt{(E[p_{Tj}|Y] - p_T^*)^2 + (E[p_{Ej}|Y] - p_E^*)^2}, \tag{2.11}$$

the Euclidean distance between the current estimates and some physician-specified target rates of toxicity and efficacy, p_T^* and p_E^*. Here we propose a few modifications of this basic approach. First, to discourage excessive toxicity caused by high doses, we put higher weight on the toxicity component contribution to (2.11). Second, we permit different penalties for over and under-dosing by incorporating asymmetry into the distance calculation. Specifically, we let

$$dist_{Tj} = \begin{cases} (E[p_{Tj}|Y] - p_T^*), & \text{if } \ E[p_{Tj}|Y] > p_T^* \\ w_T \ (E[p_{Tj}|Y] - p_T^*), & \text{otherwise} \end{cases} \tag{2.12}$$

$$\text{and } dist_{Ej} = \begin{cases} E[p_{Ej}|Y] - p_E^*), & \text{if } \ E[p_{Ej}|Y] < p_E^*; \\ w_E \ (E[p_{Ej}|Y] - p_E^*), & \text{otherwise} \end{cases} \tag{2.13}$$

where $0 < w_T < 1$ and $0 < w_E < 1$. The full distance for a specific dose X_j is then a modified version of (2.11), namely

$$dist_j = \sqrt{W_{dT} \ dist_{Tj}^2 + W_{dE} \ dist_{Ej}^2}, \tag{2.14}$$

where $W_{dT} > 0$ and $W_{dE} > 0$ are positive weights that can be adjusted to achieve better operating characteristics for the trial. A dose with a smaller value of $dist_j$ is

more desirable since it is closer to the pre-specified probabilities of toxicity and efficacy. Finally, we also employ termination rules to control overtoxicity and to enable an early decision regarding the optimal dosage. First, if for the lowest dose level X_1, the posterior samples of p_{T1} satisfy

$$Pr(p_{T1} < p_T^*|Y) < \tilde{\pi}_T,$$

where $\tilde{\pi}_T$ is some pre-specified small value (say, 0.2), then we will terminate the trial for over-toxicity. Second, if for some dose X_j, the posterior samples of p_{Tj} and p_{Ej} satisfy

$$Pr(p_{Tj} < p_T^*|Y) > \dot{\pi}_T, \text{ and } Pr(p_{Ej} < p_E^*|Y) > \dot{\pi}_E,$$

where $\dot{\pi}_T$ and $\dot{\pi}_E$ are two pre-specified large probabilities (say, 0.8), then we will stop the trial and define dose X_j as the optimal dose. If there are multiple doses which satisfy both probability statements, then the dose with the smallest $dist_j$ in (3.15) is picked as the optimal dose.

In summary, our proposed trivariate dose-finding algorithm is as follows:

Trivariate Dose-finding Algorithm

1. Treat the first cohort patients at the lowest dose level.

2. Update the posterior distributions of the probabilities of toxicity and efficacy at all dose levels.

3. Calculate the criteria to check for early trial termination.

4. If not terminated, calculate the distances $dist_j$ for $j = 1, \ldots, 5$.

5. Treat the next patient cohort at the dose having the minimum distance (3.15) under the restrictions of no dosage shift of more than one level of escalation or

deescalation.

6. Repeat from Step 2 until the trial is terminated early or maximum sample size is achieved.

2.3 Simulation Results

We now present the result of several simulation studies based on the motivating example in Section 2.1. The quality of surrogate marker in our models can be evaluated in two ways: the difference between S and E in *marginal* posterior probability, or via the association parameters (ϕ_2 or γ_2) between S and E. We define a "good" surrogate as one that has a strong association and a marginal probability close to that for true efficacy, while a "bad" surrogate has a weak association and a dissimilar marginal probability. To model the false positivity of the surrogate in a real situation, we set $Pr(S = 1) = 1.1 * Pr(E = 1)$ for a "good" surrogate, and $Pr(S = 1) = 1.5 * Pr(E = 1)$ for a "bad" surrogate. Note all of our ϕ_1 and γ_1 settings imply modest positive association between E and T, whereas the "bad" surrogate choices for ϕ_2 and γ_2 assume independence of E and S. We set $w_E = w_T = \frac{1}{3}$ as our penalty reduction for undershooting toxicity and overshooting efficacy in (3.16) and (3.17), and set $W_{dT} = 2$ but $W_{dE} = 1$ in (3.15), thus making toxicity twice as important as efficacy in the distance calculation.

For our simulation, we considered both our trivariate joint models (ExB and ExG) and the corresponding bivariate models (Braun and Gumbel) that simply replaced the efficacy data with the surrogate efficacy data, assuming the latter to be without error (as is sometimes done in practice). Due to the different meanings of ϕ_1 and γ_1 (and ϕ_2 and γ_2), we primarily seek to compare ExB to Braun and ExG to Gumbel, respectively, to

evaluate the benefits of our trivariate model. Each of our simulation studies used 1000 simulated trials, each analyzed by generating two MCMC chains in the `WinBUGS` software (`www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`). We ran 1000 MCMC iterations after a 1000-iteration burn-in period for each chain. Standard convergence diagnostics [49] did not reveal significant MCMC convergence issues.

Table 2.1 presents the parameter settings for three different scenarios. Scenario 1 and Scenario 2 assume the optimal doses are Dose 4 and Dose 2, respectively, with their corresponding true distances in equation (3.15) as 0. The third scenario describes a situation where all doses are over-toxic based on the the physician-supplied target value, to say, $p_T > p_T^*$. In Scenario 3, none of the five doses meets the condition that its true p_T and p_E values respectively equal p_T^* and p_E^*, and thus correspond to a true distance of 0.

Table 2.2 shows the empirical selection probabilities and percents of patients treated at each dose for the competing methods under "good" surrogacy across three different scenarios, while Table 2.3 does the same for "bad" surrogacy. In general, our trivariate models perform better than the corresponding bivariate models, especially under "bad" surrogacy. In addition, using a "good" surrogate have more chance in identifying and assigning more patients to the optimal dose than using a "bad" one.

Specifically in Scenario 1, ExB and ExG identify the correct dose (4) slightly more often than the Braun and Gumbel models under "good" surrogacy, respectively. Both trivariate models are also substantially better under "bad" surrogacy than the corresponding bivariate models. The standard Braun and Gumbel models select the correct dose in only 21% and 28% of the simulated studies, compared to 42% and 49% for ExB

and ExG, respectively. The proportions of patients treated at the correct dose increase for ExB and ExG are higher than Braun and Gumbel model under "bad" surrogacy as well (26% vs 18% and 27% vs 20%, respectively). No early terminations due to over-toxicity were detected in this scenario.

In Scenario 2, comparing ExB and ExG to their corresponding bivariate models (Braun and Gumbel), our proposed trivariate models always have better operating characteristics under the "good" and "bad" surrogacy scenarios. Under "bad" surrogacy, the probability of selecting the correct dose increases from 62% with the Braun model to 76% with ExB and 64% with the Gumbel model to 78% with ExG. The assignment of patients at the optimal dose is also improved from Braun (48%) and Gumbel (47%) to ExB (59%) and ExG (60%), respectively under "bad" surrogacy. The stopping probabilities due to over-toxicity were very close to 0 (less than 2%).

In Scenario 3, the probability of toxicity for all doses exceeds the physician-specified target ($p_T^* = 0.4$) and the correct decision is to terminate early. The results suggest that our trivariate models can do as well as bivariate models in stopping the trial early due to over-toxicity with a high probability. All the four models can successfully stop the trial early at least 79% of the time due to over-toxicity. In the "good" surrogacy scenario, the average numbers of patients treated under the ExB, Braun, ExG and Gumbel models are 15.1, 15.1, 14.7 and 16.1, respectively, far below the number of the patients in the initial enrollment plan (3*11=33). Note that the average number of patients treated at each dose can also be calculated. For example, the "good" surrogacy results reveal there were $15.1 \times 0.14 \approx 2.1$ patients actually treated at Dose level 2 on average by the ExB model. Under "bad" surrogacy, the total average numbers of patients treated

under the four methods changes only slightly to 15.1, 14.8, 14.7 and 16.2, respectively.

We also considered varying the weights w_T, w_E, W_{dT} and W_{dE}. For example, if we set all the weights equal under scenario 1, the targeting efficiency still improves for our trivariate models (results not shown). But we might expect this may lead to a slightly higher probability of overdose (Dose 5) than that of our weighted version. In practice, we suggest calibrating the weights to the desired level of overdose control. For example, an investigator most interested in controlling toxicity for the trial's patients would take $W_{dT} >> W_{dE}$.

2.4 Discussion

Our proposed method can successfully improve phase I/II dosage targeting efficiency by jointly modeling toxicity, efficacy and surrogate efficacy. Firstly, whether under "good" or "bad" surrogacy scenarios, the targeting performance is improved by adding some efficacy data, as opposed to using only the surrogate data. The quality of surrogate markers is an important factor in finding an optimal dosage. In the above simulation studies, we assume a higher marginal probability for the surrogate marker, which reflects reality in the use of surrogate markers. When we use only the surrogate efficacy data, as we expect, this makes the final dose more variable, hence a poorer estimate of the optimal dose. Especially under "bad" surrogacy, we modeled a large probability of false positive efficacy, leading to a downward effect in the dose selected. However, with some efficacy data, our joint models can eliminate part of the mean squared error and improve targeting accuracy. We also made the bad surrogate "more bad", by using a 2x multiplier (instead of merely 1.5x) in the true efficacy probabilities. As we expected, we did observe

an even bigger detriment to using a "very bad" surrogate. The gains resulting from the change from bivariate models to our trivariate models are even greater.

Secondly, we want to point out that direct comparison of the performance of the ExB and ExG methods might not be sensible, since the interpretation of association parameters ϕ or γ is different in each model and it is thus unclear how fair comparisons can be made. Our results indicates ExG performs quite similar to ExB except that under bad surrogacy in Scenario 1, ExG beats ExG (49% vs 42%). Since ExG and Gumbel models explicitly specify a joint bivariate distribution between outcomes, we would recommend the use of ExG rather ExB.

Thirdly, the purpose of our trivariate model is actually to borrow some strength from the surrogate efficacy data so that we can learn more on the missing efficacy data. Our dose finding methods perform slightly better in a trial with a good surrogate than with a bad surrogate, which is consistent with intuition as well as efforts to find a good surrogate in clinical research. It is reasonable that this improvement is not very strong because the missing efficacy data is not so much. In addition, penalty weights w_T, w_E, W_{dT} and W_{dT} in our models can be flexibly adjusted to obtain an optimal dose under different conditions; we suggest putting more weight on w_T and/or W_{dT} to control over-toxicity.

An alternate approach to overcoming the delay between the measurement of toxicity and the measurement of efficacy is to extend the TITE-CRM (Cheung and Chappel reference) to accomodate multiple outcomes. A disadvantage of this approach is that it would ignore the information present in the surrogate outcome. Our trivariate model provides a (slight) improvement in our ability to correctly identify the optimal dose

compared to the bivariate model even when the surrogate is a good predictor of efficacy. This suggests that including both the surrogate and efficacy endpoints provides additional benefit beyond allowing us to complete the trial in a more reasonable time-frame.

Our setting is just an idealization of actual practice, and could be modified. For example, the use of the rigid logistic function of form with intercept fixed at -3 could be replaced with other parametric or nonparametric forms for the toxicity, efficacy, or surrogate efficacy probabilities [50, 51, 52]. Adding an upper bound $\theta < 1$ on the probability of efficacy (or surrogate efficacy) may also be sensible. We experimented with the addition of a "plateau" parameter that provides an upper bound on efficacy, but found this parameter hard to estimate (since relevant information about it is confined to the rarely-visited highest doses) and led to only very small gains in the performance measures reported in our tables. A quadratic model is also a possibility, but again with our small initial datasets, this approach is not sensible without overly informative priors.

Another extension of our method could be to jointly model several surrogate markers for toxicity and efficacy. All the surrogate markers could be assumed to be operate in "parallel", meaning that all their inter-connections are captured conditional on efficacy. As mentioned in Subsection 2.2.1, the assumption of conditional independence of toxicity and surrogate efficacy might be too strong. Exploration of the relationship between T, E and S (or multiple Ss) may shed light on this issue. We also tried to detect the association between toxicity, efficacy and surrogate efficacy by posterior samples of related association parameters. Unfortunately, this estimation was not satisfactory, with large posterior variance. The reason might be that we are trying to estimate a

fairly data-insensitive parameter with too small a sample size of binary outcomes that themselves contain little information about the association parameter. Finally, our definition of a "good" or "bad" surrogate is a little arbitrary. A better quantification of the quality of surrogacy and its effect on our dose-finding are another subject for future investigation.

Dose	1	2	3	4	5
Scenario 1: Dose 4 optimal					
True Pr(Tox=1)	0.08	0.12	0.18	**0.27**	0.38
True Pr(Eff=1)	0.08	0.14	0.23	**0.35**	0.50
True Pr(Sur=1) good	0.088	0.154	0.253	**0.385**	0.55
True Pr(Sur=1) bad	0.12	0.21	0.345	**0.525**	0.75
True distance	0.28	0.22	0.13	**0**	0.16
Scenario 2: Dose 2 optimal					
True Pr(Tox=1)	0.12	**0.27**	0.50	0.73	0.88
True Pr(Eff=1)	0.20	**0.35**	0.40	0.48	0.60
True Pr(Sur=1) good	0.22	**0.385**	0.44	0.528	0.66
True Pr(Sur=1) bad	0.30	**0.525**	0.60	0.72	0.9
True distance	0.17	**0**	0.33	0.65	0.87
Scenario 3: All the doses over-toxic					
True Pr(Tox=1)	0.60	0.75	0.80	0.85	0.90
True Pr(Eff=1)	0.08	0.14	0.23	0.35	0.50
True Pr(Sur=1) good	0.088	0.154	0.253	0.385	0.55
True Pr(Sur=1) bad	0.12	0.21	0.345	0.525	0.75
True distance	0.36	0.52	0.57	0.64	0.71
Target Probabilities	p_T^* **=0.4**	p_E^***=0.3**			
Association parameters					
ExB (good surrogate)	$\phi_1 = 0.7$	$\phi_2 = 0.9$			
ExB (bad surrogate)	$\phi_1 = 0.7$	$\phi_2 = 0.5$			
ExG (good surrogate)	$\gamma_1 = 1$	$\gamma_2 = 2$			
ExG (bad surrogate)	$\gamma_1 = 1$	$\gamma_2 = 0$			

Table 2.1: Simulation parameter settings in three different scenarios

Scenario	Method	Operating characteristics	Dose 1	2	3	4	5	over-toxic
Scenario 1: Dose 4 optimal	ExB	selection probability	0.01	0.07	0.26	**0.51**	0.15	0
		proportion of patients treated	0.11	0.20	0.31	0.26	0.12	
	Braun	selection probability	0.03	0.05	0.37	**0.47**	0.08	0
		proportion of patients treated	0.13	0.18	0.33	0.28	0.08	
Scenario 2: Dose 2 optimal	ExB	selection probability	0.13	**0.81**	0.06	0	0	0
		proportion of patients treated	0.24	0.62	0.13	0.01	0	
	Braun	selection probability	0.21	**0.75**	0.03	0	0	0.01
		proportion of patients treated	0.28	0.61	0.10	0.01	0	
Scenario 3: Over-toxic	ExB	selection probability	0.16	0	0	0	0	**0.84**
		proportion of patients treated	0.85	0.14	0.01	0	0	
	Braun	selection probability	0.21	0	0	0	0	**0.79**
		proportion of patients treated	0.86	0.13	0.01	0	0	
Scenario 1: Dose 4 optimal	ExG	selection probability	0	0.04	0.28	**0.52**	0.16	0
		proportion of patients treated	0.11	0.16	0.30	0.31	0.12	
	Gumbel	selection probability	0	0.04	0.29	**0.50**	0.17	0
		proportion of patients treated	0.12	0.16	0.29	0.30	0.13	
Scenario 2: Dose 2 optimal	ExG	selection probability	0.14	**0.80**	0.06	0	0	0
		proportion of patients treated	0.26	0.59	0.14	0.01	0	
	Gumbel	selection probability	0.17	**0.77**	0.05	0	0	0.01
		proportion of patients treated	0.28	0.58	0.13	0.01	0	
Scenario 3: Over-toxic	ExG	selection probability	0.15	0	0	0	0	**0.85**
		proportion of patients treated	0.87	0.13	0	0	0	
	Gumbel	selection probability	0.15	0	0	0	0	**0.85**
		proportion of patients treated	0.87	0.12	0.01	0	0	

Table 2.2: Operating characteristics of the four methods under good surrogacy

Scenario	Method	Operating characteristics	Dose 1	2	3	4	5	over-toxic
Scenario 1: Dose 4 optimal	ExB	selection probability	0.01	0.06	0.32	**0.42**	0.19	0
		proportion of patients treated	0.12	0.20	0.33	0.26	0.11	
	Braun	selection probability	0.05	0.14	0.59	**0.21**	0.01	0
		proportion of patients treated	0.16	0.22	0.41	0.18	0.03	
Scenario 2: Dose 2 optimal	ExB	selection probability	0.17	**0.75**	0.07	0	0	0.01
		proportion of patients treated	0.26	0.59	0.14	0.01	0	
	Braun	selection probability	0.35	**0.62**	0.02	0	0	0.01
		proportion of patients treated	0.47	0.48	0.05	0	0	
Scenario 3: Over-toxic	ExB	selection probability	0.16	0	0	0	0	**0.84**
		proportion of patients treated	0.85	0.14	0.01	0	0	
	Braun	selection probability	0.20	0	0	0	0	**0.80**
		proportion of patients treated	0.87	0.12	0.01	0	0	
Scenario 1: Dose 4 optimal	ExG	selection probability	0	0.05	0.28	**0.49**	0.18	0
		proportion of patients treated	0.11	0.20	0.27	0.27	0.15	
	Gumbel	selection probability	0	0.09	0.62	**0.28**	0.01	0
		proportion of patients treated	0.12	0.24	0.41	0.20	0.03	
Scenario 2: Dose 2 optimal	ExG	selection probability	0.14	**0.78**	0.07	0	0	0.01
		proportion of patients treated	0.24	0.60	0.15	0.01	0	
	Gumbel	selection probability	0.34	**0.64**	0.01	0	0	0.01
		proportion of patients treated	0.49	0.47	0.04	0	0	
Scenario 3: Over-toxic	ExG	selection probability	0.15	0	0	0	0	**0.85**
		proportion of patients treated	0.89	0.11	0	0	0	
	Gumbel	selection probability	0.16	0	0	0	0	**0.84**
		proportion of patients treated	0.87	0.12	0.01	0	0	

Table 2.3: Operating characteristics of the four methods under bad surrogacy

Chapter 3

Flexible Link Continual Reassessment Methods for Trivariate Binary Outcome Phase I/II Trials

Chapter 3 presents a further extension of the work in Chapter 2, using the same motivating example. As discussed in Section 2.4, the use of the standard CRM fixed-intercept logistic link function for the marginal probabilities of toxicity, efficacy and surrogate efficacy might not be adequate in practice. Even with the intercept allowed to vary, parametric functions may be too smooth to accurately model the relationships between these probabilities and the dose. Generalizing to include an upper bound $\phi < 1$ on the probability of efficacy (and surrogate efficacy) may also be sensible.

Chapter 3 is organized as follows. In Section 3.1 we offer a brief catalog of more flexible semiparametric and nonparametric monotone link functions to model the marginal probability of efficacy. The joint probability model for toxicity, efficacy and surrogate efficacy, Bayesian prior selection, and the details of our dose-finding algorithm are addressed as well. Section 3.2 presents simulation results comparing the operating characteristics of various models and provides guidance on specifying a future design. Finally, Section 3.3 discusses our findings and proposes areas for future work, including a "limited memory" modification, designed to improve performance by reducing trial-to-trial variability.

3.1 Methods

3.1.1 Parametric and More Flexible Monotone Link Functions

Although dose-toxicity or dose-efficacy link functions are not necessarily monotone, as studied by Berry et al. [53] and Bekele and Shen [20], assuming monotonicity in the design is reasonable in most cases and can greatly improve efficiency when searching for the MTD. Here, we provide a series of parametric, semiparametric and nonparametric

monotone link functions. One simple and popular parametric link function is the logistic, expressed for our three marginal endpoint probabilities as

$$p_{Tj} = \frac{exp(\alpha_T + \beta_T X_j)}{1 + exp(\alpha_T + \beta_T X_j)} \ , \tag{3.1}$$

$$p_{Ej} = \frac{exp(\alpha_E + \beta_E X_j)}{1 + exp(\alpha_E + \beta_E X_j)} \ , \tag{3.2}$$

$$\text{and } \ p_{Sj} = \frac{exp(\alpha_S + \beta_S X_j)}{1 + exp(\alpha_S + \beta_S X_j)} \ , \ \ j = 1, \ldots, k \ , \tag{3.3}$$

where j indexes the k possible dose levels X_j, and α_T, α_S and α_E are assumed to be negative to ensure appropriately sized probabilities, while β_T, β_S and β_E are assumed to be positive to ensure the probabilities increase with dose. In the sequel we follow CRM convention and fix the intercepts α_T, α_E and α_S at a particular value (we use -3); this model is referred as the *base model*.

Figure 3.1 illustrates the dose-response curve in the base model for various values of β_E. The dose-response curve increases smoothly in all cases and will provide an adequate fit if the true dose-response relationship is similarly smooth. On the other hand, the logistic curve may not provide an adequate fit if the true dose-response relationship includes dramatic increases in efficacy as dose increases. To simplify computation and avoid over-parameterization, the suggested parametric, semiparametic and nonparametric link functions will be applied only to the marginal probabilities of efficacy, retaining the above parametric forms for the other two endpoints.

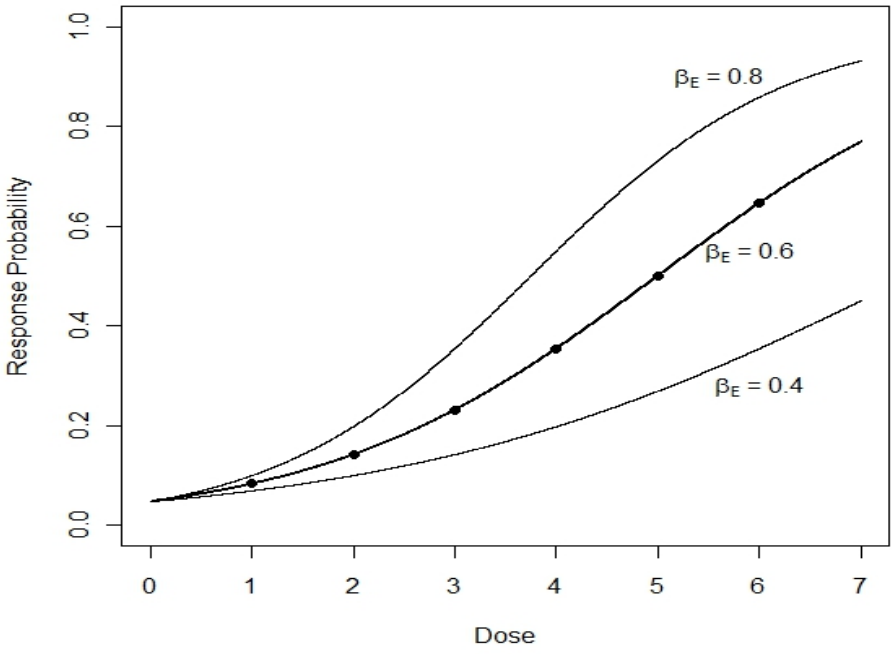

Figure 3.1: Dose-response curves in the base model for various values of β_E. Given $\alpha_E = -3$, the curves of all simple logistic functions cross the point $(0, 0.047)$, and is fully determined by β_E. We show three different curves corresponding to different slopes ($\beta_E = 0.4, 0.6$ and 0.8), respectively.

Mixture of Logistic Functions (MLF)

An obvious first extension of the logistic link is to use a *weighted mixture* of L logistic functions with different fixed intercept terms. Here, the the marginal probability of efficacy can be expressed as

$$p_{Ej} = \sum_{l=1}^{L} v_l \, q_l(X_j), \tag{3.4}$$

where $q_l(X_j) = \frac{exp(\alpha_{El}+\beta_E X_j)}{1+exp(\alpha_{El}+\beta_E X_j)}$ represents the marginal probability of efficacy calculated using the logistic link with fixed intercept α_{El}, and v_l is the positive weight of the

lth logistic function such that $\sum_{l=1}^{L} v_l = 1$. Given our previous choice of $\alpha_E = -3$, it may be reasonable to take $L = 7$ and set $\alpha_{E1} = -6$, $\alpha_{E2} = -5$, $\alpha_{E3} = -4$, $\alpha_{E4} = -3$, $\alpha_{E5} = -2$, $\alpha_{E6} = -1$, and $\alpha_{E7} = 0$ so that our mixture is "centered" at the standard link. When we have little prior information regarding the α_{El}, it may be appropriate to set the prior for $\boldsymbol{v}$ as a vague Dirichlet(0.5,0.5,0.5,0.5,0.5,0.5,0.5) distribution, where $\boldsymbol{v}$'=$(v_1, v_2, \ldots, v_L)$. The Dirichlet is available in the `WinBUGS` (`www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`) or `JAGS` (`mcmc-jags.sourceforge.net`) languages, so modification of standard parametric code to handle (3.4) is routine.

Figure 3.2 provides an illustration of this link function. The individual curves represent logistic dose-response curves with $\beta_E = 0.6$ and various values of α_E, while the dashed curve represents the MLF model with $\boldsymbol{v}$'=(0.1, 0.4, 0.2, 0.1, 0.1, 0.05, 0.05).

Mixture of Incomplete Beta Functions (MIB)

A closely related but slightly more sophisticated approach is motivated by the work of Gelfand and Mallick [54], and adapted to hierarchical Cox model analysis by Carlin and Hodges [55]. Define $\tilde{J}_0(X_j) = \frac{exp(-3+\beta_E X_j)}{1+exp(-3+\beta_E X_j)}$, which conveniently transforms the dose level to the interval (0,1), as in (3.1)-(3.3). Given $\tilde{J}_0(X_j)$, we model the true marginal probability of efficacy p_{Ej} as a mixture of incomplete beta (beta cumulative distribution) functions "centered" around $\tilde{J}_0(X_j)$,

$$p_{Ej} = \sum_{l=1}^{L} v_l \; IB(\tilde{J}_0(X_j); r_l, u_l), \tag{3.5}$$

where $IB(\cdot\,; a, b)$ denotes the incomplete beta function with parameters a and b, and the v_l's are again positive weights such that $\sum_{l=1}^{L} v_l = 1$. Since any distribution function on [0,1] can be approximated arbitrarily well by a finite mixture of Beta cdf's [56], the use

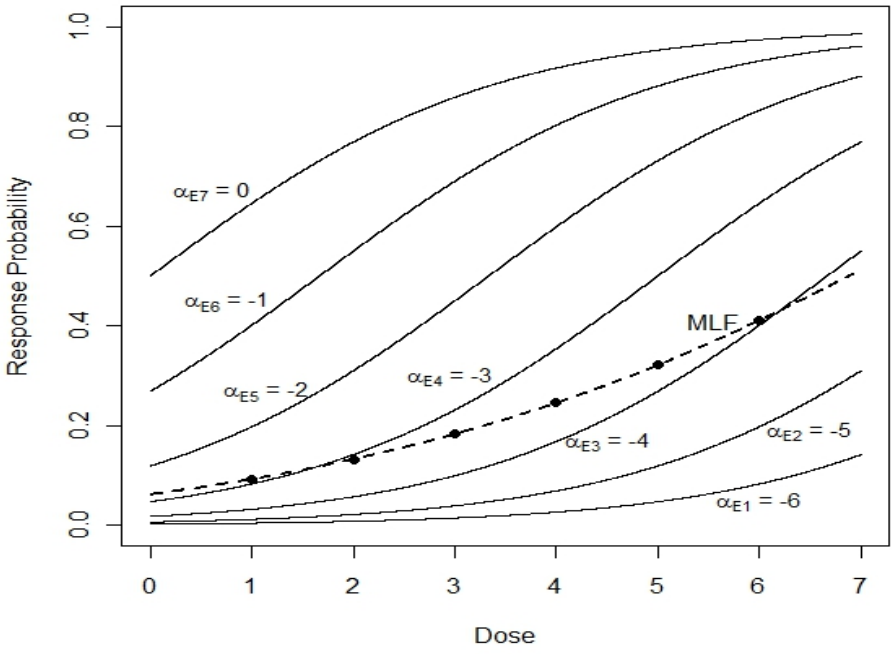

Figure 3.2: Logistic dose-response curves with $\beta_E = 0.6$ and various values of α_E. The dashed curve represents the dose-response in the MLF model with $\boldsymbol{v}' = (0.1, 0.4, 0.2, 0.1, 0.1, 0.05, 0.05)$.

of (3.5) brings substantial flexibility. If we fix L, now the number of beta cdf's, at some integer and use the "evenly spaced" choices $r_l = \psi l$ and $u_l = \psi(L-l+1)$ for some $\psi > 0$, then the resulting beta densities effectively "cover" the interval [0,1]. In our setting, we might take $L = 7$ and $\psi = 1$ to get $\mathbf{r}' = (1, 2, 3, 4, 5, 6, 7)$ and $\mathbf{u}' = (7, 6, 5, 4, 3, 2, 1)$. See Figure 3.3 for illustration. Similar to Subsection 3.1.1, we can choose a suitable Dirichlet prior distribution for $\mathbf{v}$ and again use `WinBUGS` or `JAGS`.

Before moving on to fully nonparametric models, we remark that it is often the case that the probability of efficacy will plateau at some level $\phi \in (0, 1)$, representing

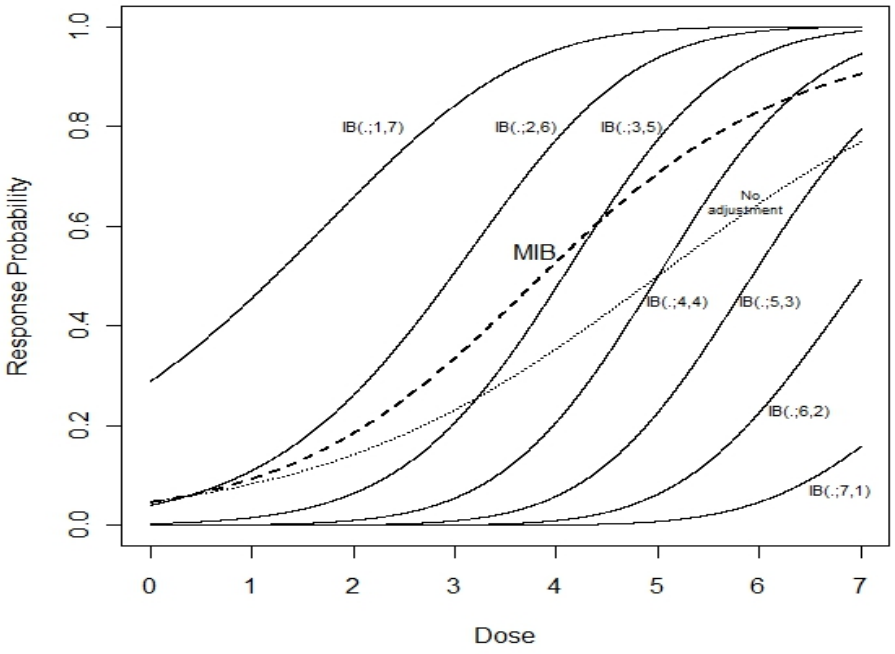

Figure 3.3: IB adjustment of the logistic dose-response curves with $\alpha_E = -3$ and $\beta_E = 0.6$. The dashed curve represents the adjusted dose-response probability in the MIB model with $\boldsymbol{v}$'=(0.1, 0.4, 0.2, 0.1, 0.1, 0.05, 0.05). Other solid curves indicates the dose-response after different IB adjustments. The dotted curve suggests no adjustment for the simple logistic function.

an efficacy level that cannot be exceeded even if the dose is increased. Such a plateau parameter ϕ is easily added to the right hand sides of equations (3.2)- (3.5), modifying the three link functions to

$$p_{Ej} = \phi \frac{exp(\alpha_E + \beta_E X_j)}{1 + exp(\alpha_E + \beta_E X_j)} , \tag{3.6}$$

$$p_{Ej} = \phi \sum_{l=1}^{L} v_l \frac{exp(\alpha_{El} + \beta_E X_j)}{1 + exp(\alpha_{El} + \beta_E X_j)} , \tag{3.7}$$

$$\text{and} \;\; p_{Ej} = \phi \sum_{l=1}^{L} v_l \, IB(\tilde{J}_0(X_j); r_l, u_l) \; . \tag{3.8}$$

Isotonic Nonparametric Model with Uniform Priors (NUP)

Our previous two link functions are semiparametric in that they use parametric weights with a fairly arbitrary collection of smoothly increasing functions. To even further relax our assumptions, we might avoid any smooth baseline link entirely, and assume only that $0 \leq p_{E1} < p_{E2} < \cdots < p_{Ek} \leq 1$. To implement this nonparametric isotonic regression in a Bayesian fashion, a simple solution is to add these constraints to an otherwise vague prior. For example, in our six-dose setting we can set $p_{E1} \sim Unif(0, p_{E2})$, $p_{E2} \sim Unif(p_{E1}, p_{E3})$, $p_{E3} \sim Unif(p_{E2}, p_{E4})$, $p_{E4} \sim Unif(p_{E3}, p_{E5})$, $p_{E5} \sim Unif(p_{E4}, p_{E6})$, and $p_{E6} \sim Unif(p_{E5}, 1)$, where $Unif(a, b)$ denotes the uniform distribution on the interval (a, b). Once again, such a prior is readily implemented `WinBUGS` and may perform well when the number of dose levels is not too large. In `JAGS`, prior ordering of top-level parameters like this can be achieved using the `sort` function, which sorts a vector into ascending order.

Isotonic Nonparametric Model with Product-of-Beta Priors (NPBP)

An alternative is to model p_E using the curve-free method of Gasparini and Eisele [21]. These authors recommend a monotonicity-preserving reparameterizaton of the p_E's. Specifically, our six marginal efficacy probabilities are reparameterized as

$$\theta_1 = 1 - p_{E1}, \; \theta_2 = \frac{1 - p_{E2}}{1 - p_{E1}}, \; \ldots, \; \text{and} \; \theta_6 = \frac{1 - p_{E6}}{1 - p_{E5}} \; . \tag{3.9}$$

It is easy to see that $p_{Ej} = 1 - \theta_1\theta_2\cdots\theta_j$, preserving monotonicity while avoiding explicit inequality constraints. Placing independent beta priors on $\theta_1, \ldots, \theta_6$ completes the hierarchical specification.

3.1.2 The Gumbel Copula Setting for Two Submodels

The semiparametric and nonparametric links described above can be incorporated into our trivariate model by specifying the conditional probabilities $Pr(E|T)$ and $Pr(S|T)$ within a copula model. Specifically, if $Z_1, Z_2 \in \{0, 1\}$ are two binary random variables, the joint bivariate Gumbel copula is

$$Pr(Z_1 = z_1, Z_2 = z_2) = p_1^{z_1}(1-p_1)^{(1-z_1)}p_2^{z_2}(1-p_2)^{(1-z_2)} + (-1)^{z_1+z_2}p_1(1-p_1)p_2(1-p_2)\kappa \,, \tag{3.10}$$

where p_1 and p_2 are the marginal probabilities success for Z_1 and Z_2, respectively, and $\kappa \in (-1, 1)$ is a real-valued parameter to capture the association between Z_1 and Z_2. As is obvious from equation (3.10), $\kappa = 0$ implies independence, positive values of κ imply positive association, and negative values indicate negative association. The conditional probability of Z_1 given Z_2 can then be derived as

$$Pr(Z_1 = 1|Z_2) = p_1 + (-1)^{Z_2+1}p_1(1-p_1)p_2^{1-Z_2}(1-p_2)^{Z_2}\kappa \,. \tag{3.11}$$

The two association parameters κ_1 and κ_2 can be used to model the two conditional probabilities we require, $Pr(E|T)$ and $Pr(S|E)$.

3.1.3 Likelihood and Prior Specification

As with all Bayesian analyses, we must specify a full likelihood, and a prior distribution for every parameter. Suppose that n_j patients are treated at dose X_j $(j = 1, \ldots, k)$,

among which n_j^{tes} patients have outcomes $T = t$, $E = e$ and $S = s$. Then the likelihood for the joint model is multinomial,

$$L(\beta_T, \beta_E, \kappa_1, \kappa_2, \boldsymbol{\zeta} \mid \text{Data}) \propto \prod_{j=1}^{k} {\pi_j^{000}}^{n_j^{000}} {\pi_j^{001}}^{n_j^{001}} {\pi_j^{010}}^{n_j^{010}} {\pi_j^{011}}^{n_j^{011}} {\pi_j^{100}}^{n_j^{100}} {\pi_j^{101}}^{n_j^{101}} {\pi_j^{110}}^{n_j^{110}} {\pi_j^{111}}^{n_j^{111}}, \quad (3.12)$$

where π_j^{tes} represents the joint probability of T, E, and S given dose X_j, and $\boldsymbol{\zeta}$ represents the related parameter vector in the link function for the probability of efficacy. For example, $\boldsymbol{\zeta} = \beta_E$ in the base model while $\boldsymbol{\zeta} = (\beta_E, \boldsymbol{v}, \phi)'$ in the MLF model (3.7). The joint probabilities π_j^{tes} are computed as described in the previous two subsections.

However, as illustrated in Figure 2.1, not all of the efficacy data are available when deciding dose assignment. Thus, some n_j^{tes} cannot be fully determined in time. Assuming the efficacy data of subject i treated at dose j is unavailable at a specific time, the likelihood function for subject i is

$$\begin{aligned} L_i = Pr(T_{ij} = t) \quad & \{Pr(E_{ij} = 1 | T_{ij} = t) Pr(S_{ij} = s | E_{ij} = 1) + \\ & Pr(E_{ij} = 0 | T_{ij} = t) Pr(S_{ij} = s | E_{ij} = 0)\} \end{aligned} \quad (3.13)$$

If n_j^{tes} denotes the corresponding number of patients with complete data, then the likelihood function is constructed as

$$L(\beta_T, \beta_S, \beta_E, \phi_1, \phi_2 \mid \text{Data}) \propto L_C \prod_{i=1}^{h} L_i, \quad (3.14)$$

where h represents the number of patients with unobserved efficacy data. Note $n_j^{tes} = 0$ at the early stage, which suggests no efficacy data are available. Imputation for the efficacy data of subject i can be easily handled by `BUGS` software.

As for the priors, we assume both β_T and β_S independently follow exponential distributions with mean 1. To set priors for κ_1 and κ_2 in the range $(-1, 1)$, we simply transform them to $\kappa_1 = 2\psi_1 - 1$ and $\kappa_2 = 2\psi_2 - 1$, and set the priors for ψ_1 and ψ_2 as Beta(4,2) and Beta(2,2) distributions, respectively, corresponding to vague prior beliefs in positive association between efficacy and toxicity, but prior independence between the two efficacy measures (i.e., a poor surrogate). We set the priors for β_E and ϕ as an exponential distribution with mean 1 and a Beta(7,3) distribution, respectively, the latter suggesting a prior upper efficacy bound of 0.7. Finally, $\theta_1, \ldots, \theta_6$ are assumed independent and identically distributed according to a Beta(4,1) distribution.

3.1.4 Dose-Finding Algorithm

Define Y as all the available accumulated data at a specific enrollment time point, so that $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$ are the posterior mean probabilities of toxicity and efficacy. After each look at the data, we update $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$, and then select the next dose by minimizing an asymmetric distance

$$dist_j = \sqrt{W_{dT}\ dist_{Tj}^2 + W_{dE}\ dist_{Ej}^2}\ , \tag{3.15}$$

where $dist_{Tj}$ and $dist_{Ej}$ represent the probability distances of toxicity and efficacy, respectively (see below) where $W_{dT} > 0$ and $W_{dE} > 0$ are positive weights that can be adjusted either subjectively or to achieve better (frequentist or Bayesian) operating characteristics for the trial. Different penalties for over- and under-dosing can be utilized

for the calculation of the two component distances. Specifically, we let

$$dist_{Tj} = \begin{cases} (E[p_{Tj}|Y] - p_T^*), & \text{if } \; E[p_{Tj}|Y] > p_T^* \\ w_T \; (E[p_{Tj}|Y] - p_T^*), & \text{otherwise} \end{cases} \tag{3.16}$$

$$\text{and } \; dist_{Ej} = \begin{cases} E[p_{Ej}|Y] - p_E^*), & \text{if } \; E[p_{Ej}|Y] < p_E^*; \\ w_E \; (E[p_{Ej}|Y] - p_E^*), & \text{otherwise} \end{cases} \tag{3.17}$$

where $0 < w_E, w_T < 1$, and p_T^* and p_E^* are physician-specified target rates of toxicity and efficacy. In practice, we recommend putting higher weight on the toxicity component contribution (i.e., choose $W_{dT} > W_{dE}$) to discourage excessive toxicity caused by high doses.

We also employ early termination rules to control overtoxicity and to enable an early decision regarding the optimal dosage. First, if for the lowest dose level X_1, the posterior samples of p_{T1} satisfy

$$Pr(p_{T1} < p_T^*|Y) < \bar{\pi}_T \; ,$$

where $\bar{\pi}_T$ is some pre-specified small value (say, 0.2), then the trial terminates for excessive toxicity. Alternatively, if for some dose X_j the posterior samples of p_{Tj} and p_{Ej} satisfy

$$Pr(p_{Tj} < p_T^*|Y) > \dot{\pi}_T, \text{ and } Pr(p_{Ej} < p_E^*|Y) > \dot{\pi}_E \; ,$$

where $\dot{\pi}_T$ and $\dot{\pi}_E$ are two pre-specified large probabilities (say, 0.95), then the trial terminates for efficacy, and we declare dose X_j to be the optimal dose. If more than one dose satisfies *both* probability statements, the dose with the smallest $dist_j$ in (3.15) is deemed optimal.

In summary, our proposed trivariate dose-finding algorithm is as follows:

Trivariate Dose-finding Algorithm:

1. Assign the first cohort patients to the lowest dose level.
2. Update the posterior distributions of the probabilities of toxicity and efficacy at all dose levels based on the data available.
3. Check if the early termination rules apply.
4. If not terminated, calculate the distances $dist_j$ for $j = 1, \ldots, 5$.
5. Treat the next patient cohort at the dose with the minimum distance (3.15) having no more than one level of escalation or deescalation from the current dose.
6. Repeat from Step 2 until the trial is terminated early or maximum sample size is achieved.

3.2 Simulation Studies

The design of our simulation studies is based on the motivating example of Section 2.1, just with different numerical setting. Suppose the times to obtain the surrogate efficacy, toxicity and efficacy information for each patient are 4 weeks, 6 weeks and 20 weeks, respectively. We also assume that the patient recruitment plan is to enroll a new cohort of size $c = 3$ every 6 days, implying that at the enrollment time of the mth ($m \geq 4$) cohort, the efficacy data of the $(m-3)$th, $(m-2)th$ and $(m-1)th$ cohorts are unavailable. The maximum number of cohorts, L, is set to 15.

Our primary purpose here is to use the semiparametric and nonparametric monotone link functions listed in Subsection 3.1.1 to model the marginal probability of efficacy, and compare their operating characteristics to those of the parametric base model within the framework of our trivariate Gumbel joint model. As discussed in Chapter 2, correct dose selection can benefit from the joint model more under "bad" surrogacy than under "good" surrogacy. We thus only perform simulation studies in the scenario of "bad" surrogacy, which we capture by setting $Pr(S = 1) = 1.5 * Pr(E = 1)$. We set the true κ_1 and κ_2 equal to 0.5 and 0, respectively, indicating positive association between toxicity and efficacy, but independence between efficacy and surrogate efficacy (again, consistent with "bad" surrogacy). We take $w_E = w_T = \frac{1}{3}$ as our reduced penalty weights for undershooting toxicity and overshooting efficacy in (3.16) and (3.17), and set $W_{dT} = 2$ but $W_{dE} = 1$ in (3.15), thus making toxicity twice as important as efficacy in the overall distance calculation.

Below we present simulation results under the assumption that the optimal dose is Dose 4; entries for this dose are highlighted in boldface in the tables for easier identification. To check the effects of the plateau parameter ϕ of Subsection 3.1.1, we investigate two base models, in which one excludes ϕ (i.e., fixing $\phi = 1$) while the other includes it; we refer to these base models as Base A and Base B, respectively. We used `R2jags` (`cran.r-project.org/web/packages/R2jags`) to call `JAGS` from `R` version 2.12.2. Each of our simulation studies used 1000 simulated trials, each analyzed by generating two MCMC chains. We retained 2000 iterations for inference following a burn-in period of between 2000 and 12000 iterations, as needed to decrease the $\widehat{R}$ convergence diagnostic to less than 1.1 (the default value in `R2jags`). Random manual checks revealed no

significant MCMC convergence issues using standard convergence diagnostics [49].

Dose	1	2	3	4*	5	6
True Pr(Tox=1)	0.08	0.14	0.23	**0.35**	0.5	0.65
True Pr(Eff=1)	0.07	0.13	0.23	**0.36**	0.5	0.61
True Pr(Sur=1)	0.12	0.21	0.345	**0.525**	0.75	0.915
True distance	0.317	0.25	0.142	**0**	0.217	0.432
Association	$\kappa_1 = 0.5$	$\kappa_2 = 0$				

Table 3.1: Simulation parameter settings when Dose 4 is optimal and a simple logistic function with a plateau parameter can readily fit the true probability of efficacy.

		Dose					
Model	Operating characteristics	1	2	3	4*	5	6
Base A	selection probability	0	0.008	0.197	**0.673**	0.122	0
	proportion of patients treated	0.074	0.112	0.276	0.394	0.123	0.021
Base B	selection probability	0	0.011	0.208	**0.668**	0.11	0.003
	proportion of patients treated	0.073	0.104	0.264	0.396	0.136	0.027
MLF	selection probability	0	0.014	0.238	**0.652**	0.096	0
	proportion of patients treated	0.07	0.104	0.276	0.397	0.13	0.023
MLB	selection probability	0	0.003	0.204	**0.679**	0.112	0.002
	proportion of patients treated	0.07	0.104	0.276	0.397	0.13	0.023
NUP	selection probability	0.001	0.01	0.246	**0.691**	0.052	0
	proportion of patients treated	0.072	0.11	0.382	0.418	0.018	0
NPBP	selection probability	0	0.009	0.232	**0.667**	0.092	0
	proportion of patients treated	0.073	0.123	0.344	0.394	0.066	0

Table 3.2: Operating characteristics of the six models when Dose 4 is optimal and a simple logistic function with a plateau parameter can readily fit the true probability of efficacy.

Table 3.1 shows the parameter settings in the first scenario, for which a simple logistic function (possibly with plateau $\phi = 0.8$) can describe the true probability of efficacy reasonably well. Table 3.2 then lists the empirical selection probabilities and percents of patients treated at each dose using the competing parametric, semiparametric and nonparametric link functions. From Table 3.2, we see that in general all six models have very similar performance regarding both the selection probability of the correct

Dose	1	2	3	4*	5	6
True Pr(Tox=1)	0.08	0.14	0.23	**0.35**	0.5	0.65
True Pr(Eff=1)	0.1	0.12	0.2	**0.55**	0.6	0.62
True Pr(Sur=1)	0.15	0.18	0.3	**0.825**	0.9	0.93
True distance	0.468	0.441	0.355	**0**	0.213	0.425
Association	$\kappa_1 = 0.5$	$\kappa_2 = 0$				

Table 3.3: Simulation parameter settings when Dose 4 is optimal and a parametric function cannot readily fit the true probability of efficacy.

		Dose					
Model	Operating characteristics	1	2	3	4*	5	6
Base A	selection probability	0	0.006	0.069	**0.648**	0.272	0.005
	proportion of patients treated	0.072	0.092	0.166	0.407	0.222	0.042
Base B	selection probability	0	0.006	0.058	**0.684**	0.24	0.012
	proportion of patients treated	0.072	0.089	0.175	0.408	0.213	0.044
MLF	selection probability	0.001	0.008	0.096	**0.682**	0.208	0.005
	proportion of patients treated	0.073	0.083	0.164	0.402	0.227	0.051
MIB	selection probability	0	0.004	0.05	**0.694**	0.25	0.002
	proportion of patients treated	0.073	0.083	0.164	0.402	0.227	0.051
NUP	selection probability	0	0.002	0.038	**0.778**	0.182	0
	proportion of patients treated	0.072	0.088	0.162	0.494	0.183	0.001
NPBP	selection probability	0	0	0.069	**0.758**	0.17	0.003
	proportion of patients treated	0.069	0.091	0.188	0.457	0.17	0.025

Table 3.4: Operating characteristics of the six models when Dose 4 is optimal and a parametric function cannot readily fit the true probability of efficacy.

dose and the proportion of patients treated at all dose levels. Among them, NUP and MLB perform the best, making the correct choice in about 68% of the simulated studies, closely followed by the other methods, all of which have correct selection percentages at least 65%. The percents of patients treated at the correct dose are all about 40%. Another interesting finding is that NUP has lowest chance of overdose, assigning less than 2% of the patients to Dose 5 and no patients to Dose 6. All early stopping percentages due to over-toxicity were very close to 0 (less than 0.2%), and were therefore omitted from the tables. These results suggest that when the marginal probability

can be easily described by a parametric function, our proposed semiparametric and nonparametric link functions still obtain equally good results.

Table 3.3 describes a second, unsmooth scenario, in which no logistic function can readily approximate the true probabilities of efficacy, due to the large jump in efficacy (and surrogate efficacy) between Doses 3 and 4. Results for all the link functions in this scenario are given in Table 3.4. The semiparametric and nonparametric models perform better than the two parametric base models, with the exception of MLF. NUP and NPBP are the overall winners; these two methods select the correct dose with highest probabilities (both over 0.75), and treat more patients at the optimal dose compared to the other methods. MIB is next best with correct selection probability of 0.694, followed by Base B, MLF and Base A, all with correct selection probabilities over 0.64. As expected, when the true efficacy probability is hard to fit using a parametric model, our more sophisticated semiparametric and nonparametric models can offer worthwhile improvements in dose targeting accuracy.

3.3 Discussion

Although model-based designs for dose finding are attractive, selecting a good model remains challenging. If the available data are not sufficient to fully evaluate the model, we must investigate operating characteristics under model misspecification. In this situation, semiparametric and nonparametric models usually offer greater flexibility and robustness. Our simulation results show that the MIB, NUP and NPBP links have good performance regardless of the underlying smoothness of the true efficacy probabilities over the dose levels. For NPBP, we specified only moderately informative priors for θ's,

as in Subsection 3.1.3. If more informative priors are used, the probability of finding the correct dose in the above two scenarios can rise to over 90% (results not shown). But if the priors are set too ambiguously, NPBP can lead to undesirable operating characteristics [57]. Therefore, we recommend the use of MIB or NUP if little to no prior information about the treatment exists.

For the plateau parameter ϕ in the first scenario, we expected Base B to beat Base A, since the true probabilities of efficacy satisfied a $\phi = 0.8 < 1$ plateau. The results cannot show any improvement by adding ϕ. Considering that very few to none of the patients were assigned to Doses 5 or 6 by Base B, (i.e., most of the data are confined to the other dose levels), it perhaps should not be surprising that a simple logistic function without ϕ suffices to describe the efficacy probabilities. However, when a parametric model is a poor fit for the true efficacy probabilities, as in our second scenario, the results indicate some benefit of adding the plateau parameter ϕ, increasing the selection probability about 4%. Adding ϕ may provide some flexibility in fitting an unsmooth scenario.

We were also surprised by the poor performance of MLF, which we thought might rival MIB. Upon further reflection, this method may not be as general as it appears. Suppose we replace the link function by a *probit* instead of a logit, i.e., denoting the standard normal cdf by Φ we set

$$q_l(X_j) = \Phi(\alpha_{El} + \beta_E X_j) \ , \ \text{ where } \alpha_{El} \overset{iid}{\sim} N(\alpha_E, \delta_l^2) \ .$$

If we then assume a *continuous* (instead of discrete finite) mixture in (3.4) and take each weight v_l as the $N(\alpha_E, \delta_l^2)$ density evaluated at α_{El}, it is straightforward to show that p_{Ej} emerges as another probit function, thus providing no additional modeling

flexibility. While no similar closed form exists in the case of a logit link, it is plausible that the apparent flexibility offered by MLF may also be a mirage.

Future work in this area might begin with tools for Bayesian model selection. For example, we can investigate use of the Deviance Information Criterion (DIC), a Bayesian hierarchical models generalization of the Akaike information criterion (AIC) based on the posterior distribution of the deviance statistic [58]. Model selection is related to model complexity and data fitting, which we might expect would be consistent with our operating characteristic evaluation. We might also seek to evaluate and compare trial-to-trial variability across models. Recent work by Oron and Hoff criticizes the CRM for its "long memory," i.e., the tendency for its dose-finding algorithm to become "stagnant" (stuck on a particular dose) as data accumulate [59]. This is a consequence of CRM's use of the full posterior distribution as a guide for dose selection; each additional observation comprises a declining proportion of the total information available. The result is potentially large trial-to-trial variability in the recommended MTD, something the adjustments of Goodman et al. [9] and others were designed to avoid. One might investigate "limited memory" adjustments to our CRM algorithm by utilizing only the K most recent observations, Y_K^*, when determining the next dose. That is, we would replace $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$ in (3.16) and (3.17) above by $E[p_{Tj}|Y_K^*]$ and $E[p_{Ej}|Y_K^*]$, respectively.

Chapter 4

A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials

In Chapter 4, we propose a new two-stage Bayesian two-arm phase II trial design with sample size reestimation by implementing a predictive approach using the Whitehead et al. stopping rule [60]. Then, we extend it to a four-subgroup trial design that considers an important binary covariate (gender) crossed with the treatment effect. Chapter 4 is organized as follows. In Section 4.1, we introduce our proposed two-stage Bayesian design in the case of a binary endpoint for a drug treatment trial. Section 4.2 presents the application of our design for a sample cancer trial with gender stratification, and compares its operating characteristics with those of the Whitehead et al. design. Finally, Section 4.3 discusses the advantages and limitations of our design, other applications, and suggests areas for further research.

4.1 Method

4.1.1 Initial Sample Size Calculation

In an equal-size two-arm phase II trial to test the efficacy difference between a drug and a placebo for a binary endpoint, we generalize (1.2) to provide an initial sample size N per group determined to satisfy at least one of the following two probability statements for any dataset of this size,

$$Pr(p_T - p_C > 0 | s_T, s_C) \geq \eta_1 \text{ or } Pr(p_T - p_C < \theta^* | s_T, s_C) \geq \eta_2, \tag{4.1}$$

where θ^* is the desired level of improvement in treatment response rate, p_T and p_C denote the success rates in the treatment and placebo groups, respectively, and s_T and s_C represent the numbers of efficacy events among N subjects in each group. Note that s_T and s_C can take any values in $\{0, 1, 2, \ldots, N\}$. Suppose we place vague conjugate

$Beta(\alpha_T, \beta_T)$ and $Beta(\alpha_C, \beta_C)$ priors on p_T and p_C. The posterior for p_T then follows a $Beta(\alpha_T+s_T, \beta_T+N-s_T)$ distribution. In similar notation, the posterior for p_C follows a $Beta(\alpha_C+s_C, \beta_C+N-s_C)$. Then, the exact posterior distribution of (p_T-p_C) is denoted as the beta difference distribution, denoted as $BDI(\alpha_T+s_T, \beta_T+N-s_T; \alpha_C+s_C, \beta_C+N-s_C)$ [61]. Although the probability density function of this distribution is complex, Monte Carlo sampling from it is straightforward by generating from two independent beta distributions. In reality it is usually unacceptable to choose a sample size of less than 10 subjects per group, and thus we assume the smallest possible N to be 10. Thus, starting from $N = 10$, we check the criteria (4.1). If both $Pr(p_T - p_C > 0|s_T, s_C) < \eta_1$ and $Pr(p_T - p_C < \theta^*|s_T, s_C) < \eta_2$ hold for any possible values of s_T and s_C, we increase N to $N+1$ and continue to check until a minimum N is obtained that satisfies (4.1) for all s_T and s_C. Considering not much information will be available at the beginning of the trial, it is reasonable to begin with this rather conservative sample size. Note that this sample size only depends on the two threshold probabilities η_1 and η_2, a selected prior on tumor response rates in each arm, and the target rate difference between two treatments we seek to detect.

4.1.2 Sample Size Reestimation

To add in a sample size reestimation step, we first assume all patients enroll simultaneously across groups, so that the numbers of patients at the interim look are the same for both groups. Suppose the efficacy data for N_1 patients are available for both groups at the interim look, with s_{T1} positive responses in the treatment group, and s_{C1} positive responses in the control group. N_1 could be determined as a proportion

of the initial sample size; for example, $N_1 = \frac{1}{3}N$. Define the new sample size after recruiting N_1 patients to be M, and assume there are s_{T2} and s_{C2} successful events among these M additional patients in the treatment and control groups, respectively. From the previous discussion, the interim posterior distribution of $(p_T - p_C)$ is then $BDI(\alpha_T + s_{T1} + s_{T2}, \beta_T + N_1 + M - s_{T1} - s_{T2}; \alpha_C + s_{C1} + s_{C2}, \beta_C + N_1 + M - s_{C1} - s_{C2})$, and can be easily sampled.

Our sample size reestimation is still based on a conclusiveness condition. Similar to (4.1), we are particularly interested in ensuring at least one of the following two conditions:

$$Pr(p_T - p_C > 0 | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_1$$
$$\text{or } Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_2. \quad (4.2)$$

Note that both M, s_{T2} and s_{C2} are unknown random variables at the interim stage, while N_1, s_{T1} and s_{C1} are observed data. At this stage, since knowledge has been gained from the data of the first N_1 patients, it is reasonable to assume that the future data will be generated based on the interim posterior distributions of p_T and p_C. Unlike the consideration of all possible data for the initial sample size, this assumption may greatly decrease the potential expected variability of the future data, thus substantially decreasing the sample size.

Given M, s_{T2} and s_{C2} can be predicted using beta-binomial marginal distributions having pmfs,

$$P(s_{T2}) = \frac{B(\alpha_T + s_{T1} + s_{T2}, \beta_T + N_1 + M - s_{T1} - s_{T2})}{(M+1)B(M - s_{T2} + 1, s_{T2} + 1)B(\alpha_T + s_{T1}, \beta_T + N_1 - s_{T1})}, \quad (4.3)$$

$$\text{and } P(s_{C2}) = \frac{B(\alpha_C + s_{C1} + s_{C2}, \beta_C + N_1 + M - s_{C1} - s_{C2})}{(M+1)B(M - s_{C2} + 1, s_{C2} + 1)B(\alpha_C + s_{C1}, \beta_C + N_1 - s_{C1})}, \quad (4.4)$$

where B is the beta function, $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, and s_{T2}, $s_{C2} \in \{0, \ldots, M\}$. Following Sec 4.2 of [3], the predictive probabilities we require can be obtained as

$$Pr_{pred1} = \sum_{s_{T2}=0}^{M} \sum_{s_{C2}=0}^{M} I\{Pr(p_T - p_C > 0 | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_1\} P(s_{T2}) P(s_{C2}), \quad (4.5)$$

and

$$Pr_{pred2} = \sum_{s_{T2}=0}^{M} \sum_{s_{C2}=0}^{M} I\{Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_2\} P(s_{T2}) P(s_{C2}), \quad (4.6)$$

where $I(.)$ is the indicator function, taking the value 0 or 1, depending on whether the condition is false or true. We select the minimum sample size M so that at least one of the two predictive probabilities is no less than a desired level γ. A special case is that M could be 0, in which case $Pr(p_T - p_C > 0 | N_1, s_{T1}, s_{C1}) \geq \eta_1$ or $Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}) \geq \eta_2$. This situation suggests the interim data are already strong enough to reach a decision to stop the trial at that point due to either efficacy or futility, respectively.

When M is large, the computation in (4.5) and (4.6) is extensive. The sampling time to get the BDI distribution of $p_T - p_C$ and the calculation of $P(S_{T2})$ and $P(S_{C2})$ for all possible values of S_{T1} and S_{C1} are the two main factors. The use of the normal approximation arising from the so-called Bayesian Central Limit Theorem (BCLT) can speed the computation; see Section 3.2 of [62] for details. As $(N_1 + M) \to \infty$, the posterior distribution of $(p_T - p_C)$ at the interim stage can be approximated by a

Normal distribution,

$$(p_T - p_C | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \rightsquigarrow N\left(\hat{p}_T - \hat{p}_C, \frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{N_1 + M}\right), \quad (4.7)$$

where $\hat{p}_T = \frac{s_{T1}+s_{T2}}{N_1+M}$ and $\hat{p}_C = \frac{s_{C1}+s_{C2}}{N_1+M}$ are the MLE estimators of p_T and p_E in a frequentist version. Thus (4.5) and (4.6) can be approximated as

$$Pr_{pred1} \approx \sum_{s_{T2}=0}^{M} \sum_{s_{C2}=0}^{M} I\left\{\Phi\left(\frac{\hat{p}_T - \hat{p}_C}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)+\hat{p}_C(1-\hat{p}_C)}{N_1+M}}}\right) \geq \eta_1\right\} P(s_{T2})\, P(s_{C2}), \quad (4.8)$$

$$\text{and } Pr_{pred2} \approx \sum_{s_{T2}=0}^{M} \sum_{s_{C2}=0}^{M} I\left\{\Phi\left(\frac{\theta^* - (\hat{p}_T - \hat{p}_C)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)+\hat{p}_C(1-\hat{p}_C)}{N_1+M}}}\right) \geq \eta_2\right\} P(s_{T2})\, P(s_{C2}), \quad (4.9)$$

where $\Phi(x)$ represents the cumulative density function of a standard normal distribution. Alternatively, to also avoid calculating $P(s_{T2})$ and $P(s_{C2})$ in (4.8) and (4.9) for all possible values of s_{T1} and s_{C1}, we can approximate (4.8) and (4.9) by directly sampling s_{T2} and s_{C2} from their beta-binomial distributions (4.3) and (4.4), obtaining $\{(s^*_{T2j}, s^*_{C2j}), j = 1, \ldots, J\}$. Then (4.8) and (4.9) can be approximated as

$$Pr_{pred1} \approx \frac{1}{J}\sum_{j=1}^{J} I\left\{\Phi\left(\frac{\hat{p}^*_{Tj} - \hat{p}^*_{Cj}}{\sqrt{\frac{\hat{p}^*_{Tj}(1-\hat{p}^*_{Tj})+\hat{p}^*_{Cj}(1-\hat{p}^*_{Cj})}{N_1+M}}}\right) \geq \eta_1\right\}, \quad (4.10)$$

$$\text{and } Pr_{pred2} \approx \frac{1}{J}\sum_{j=1}^{J} I\left\{\Phi\left(\frac{\theta^* - (\hat{p}^*_{Tj} - \hat{p}^*_{Cj})}{\sqrt{\frac{\hat{p}^*_{Tj}(1-\hat{p}^*_{Tj})+\hat{p}^*_{Cj}(1-\hat{p}^*_{Cj})}{N_1+M}}}\right) \geq \eta_1\right\}, \quad (4.11)$$

where $\hat{p}^*_{Tj} = \frac{s_{T1}+s^*_{T2j}}{N_1+M}$ and $\hat{p}^*_{Cj} = \frac{s_{C1}+s^*_{C2j}}{N_1+M}$. See Section 3.3 of Carlin and Louis (2009) for a review of noninterative Monte Carlo methods. This sampling approach avoids extensive computation on $(M+1)^2$ different (s_{T2}, s_{C2}) pairs, potentially valuable when M is large.

4.1.3 Final Trial Conclusion

At the end of the trial, if the final results show $Pr(p_T - p_C > 0|all\ data) \geq \eta_1$ or $Pr(p_T - p_C < \theta^*|all\ data) \geq \eta_2$, then we conclude that the the test drug is minimally effective, or not as effective as we had hoped, respectively. Note in particular that our Whitehead formulation means that both conclusions may be drawn simultaneously. This is clearly an odd situation, but one that is inherent in Whitehead et al.'s definition of "conclusiveness". Of course, despite our best efforts, it may be that neither of the previous two probability statements holds, suggesting either bad luck or perhaps a data pattern for later patients that is different from that of the interim data. For temporally homogeneous data, this situation should happen rarely provided we choose the predictive probability of conclusiveness γ reasonably close to 1. Generally our proposed two-stage Bayesian sample size algorithm for each stratum is as follows:

Sample Size Estimation Algorithm:

1. Calculate the initial sample size based on the pre-specified expected effect.
2. Choose a proportion of the initial sample size, N_1, to determine the interim time for sample size reestimation.
3. Begin the trial, and estimate the interim posterior distribution of the treatment effect at the interim time (i.e., when N_1 patients have reported their data in each treatment group). Check the conclusiveness condition, thus making a decision on whether to stop the trial there.
4. If not terminated, find the minimum sample size M so that at least one of the two predictive probabilities (4.5) and (4.6) is over a desired level γ; this is our reestimated Bayesian sample size.

5. Resume the trial, and make a conclusion on the efficacy of the drug at the end of the trial, after all $N_1 + M$ patients in each group have reported.

4.1.4 Extension to Personalized Medicine

As mentioned in Section 1.3, patient demographics such as gender and age may be highly correlated with treatment effect. Thus, these confounding factors should be considered when testing the treatment effect and constitute a first step toward within-trial consideration of personalized medicine. Without loss of generality, let $Y_i \in \{0, 1\}$ be the tumor response for the ith patient $(i = 1, \ldots, n)$, and assume the Y_i follow a Bernoulli distribution with success rate p_i. Then we apply a logistic regression model for p_i as follows,

$$p_i = \frac{exp\{x_i'\beta\}}{1 + exp\{x_i'\beta\}}, \tag{4.12}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ is a vector of covariate values for the ith patient, and β is a $k \times 1$ vector of regression coefficients $(k \geq 3)$. We set $x_{i1} = 1$ for the estimation of the intercept term, and let $x_{i2} = 1$ or 0 for the treatment and placebo groups, respectively. Interaction terms, such as the multiplicative interaction between drug and gender, may also be desired in β. Suppose there are L different $(x_{i3}, \ldots, x_{ik})$ covariate vector values in the study cohort, and all patients can be classified into one of these L strata, with a particular $(x_{i3}, \ldots, x_{ik})$ for each stratum. We also assume equal sample size for both the drug and placebo arms within each stratum. If continuous covariates such as age and weight are included in the logistic model, then we can create a set of ranges based on their values, and use dummy variables to control the number of strata L.

Our two-stage Bayesian design can be applied to the multiple strata case, where we

now aim to make a conclusiveness statement on the drug effect in each stratum. When L is small, it is easy to directly use beta priors for the probabilities of success events in each subgroup to adjust sample size within each stratum. We illustrate such subgroup analysis with a numerical example in Section 4.2. The sample sizes for both treatment groups are assumed equal, but we permit different sample sizes between groups with different covariate values. The four-subgroup trial includes two sub-trials proposed by our method for the two covariate groups, with final conclusions made for both groups; we simply apply the algorithm in the previous subsection to every stratum.

However, when L is large, priors are often set for β in the logistic model, rather than directly for the stratum-specific p_i in (4.12) . Note that it is not important here to find the initial sample size so as to meet the conclusiveness condition regardless of the data, since logistic models are usually less than ideal in their handling of extreme data (i.e., when the number of success events is low even though the total number of patients is large). Our method enables the borrowing of strength from the interim data in sample size reestimation to make a conclusion on drug treatment effects.

Under either a uniform or a normal prior specification, the posterior distribution for β is complicated and does not have an analytic closed form. Although MCMC software may be used to obtain the exact distribution, the computation is extensive if using the predictive approach to test all possible datasets with the new sample size. Therefore, we again suggest use of the BCLT to simply computation.

Suppose a uniform (improper) prior is specified for β, i.e., $P(\beta) \propto 1$. Then the posterior distribution of β follows a multivariate normal distribution as $n \to \infty$,

$$\beta|Y \rightsquigarrow MVN_k(\hat{\beta}, (X'\widehat{V}X)^{-1}), \qquad \text{as } n \to \infty, \tag{4.13}$$

where $Y = (Y_1, Y_2, \ldots, Y_n)'$ is the binary outcome data, $\hat{\beta}$ is the MLE of β, and $X = (X_1, X_2, \ldots, X_n)'$ is an $n \times k$ covariate matrix. Then $\widehat{V}$ is an $n \times n$ diagonal matrix with ith diagonal element:

$$V_{ii} = \frac{exp\{x_i'\hat{\beta}\}}{(1 + exp\{x_i'\hat{\beta}\})^2}.$$

Suppose the ith and jth patients are in the drug and placebo arms within the same lth stratum, respectively ($1 \leq l \leq L$). The difference in the probabilities of tumor response between the new drug and placebo within this stratum is

$$p_i - p_j = \frac{exp\{x_i^T\beta\}}{1 + exp\{x_i^T\beta\}} - \frac{exp\{x_j^T\beta\}}{1 + exp\{x_j^T\beta\}}\ . \tag{4.14}$$

Let $x_i = (1, 1, \ldots)'$ and $x_j = (1, 0, \ldots)'$ be the covariate vectors for the drug and placebo arms, respectively. By the delta method, the posterior distribution of $p_i - p_j$ can be approximated by a normal distribution,

$$p_i - p_j | \text{ data } \rightsquigarrow N\left(\frac{exp\{x_i'\hat{\beta}\}}{1 + exp\{x_i'\hat{\beta}\}} - \frac{exp\{x_j'\hat{\beta}\}}{1 + exp\{x_j'\hat{\beta}\}},\ (X_{i\beta} + X_{j\beta})'(X'\hat{V}X)^{-1}(X_{i\beta} + X_{j\beta})\right), \tag{4.15}$$

$$\text{where } X_{i\beta} = \left(\frac{exp\{x_i'\hat{\beta}\}}{(1 + exp\{x_i'\hat{\beta}\})^2}x_{i1}, \frac{exp\{x_i'\hat{\beta}\}}{(1 + exp\{x_i'\hat{\beta}\})^2}x_{i2}, \ldots, \frac{exp\{x_i'\hat{\beta}\}}{(1 + exp\{x_i'\hat{\beta}\})^2}x_{ik}\right)',$$

$$\text{and } X_{j\beta} = \left(\frac{exp\{x_j'\hat{\beta}\}}{(1 + exp\{x_j'\hat{\beta}\})^2}x_{j1}, \frac{exp\{x_j'\hat{\beta}\}}{(1 + exp\{x_j'\hat{\beta}\})^2}x_{j2}, \ldots, \frac{exp\{x_j'\hat{\beta}\}}{(1 + exp\{x_j'\hat{\beta}\})^2}x_{jk}\right)'.$$

Similarly, we can check the probability criteria as in (4.5) and (4.6) in a predictive approach to reestimate sample size at the interim stage. In the presence of covariates leading to L strata, we can design a multi-subgroup trial by combining multiple sub-trials as proposed above. Note that we would likely power the study at the interim look for each stratum separately; Section 4.2 offers an illustration. Alternatively, unpromising strata might be abandoned at the interim look as a cost-saving measure.

4.2 Application

In this section, we give a simple but representative example, which concerns a design of phase II trial for patients with non-Hodgkin lymphoma. The primary goal of this study is to design a trial with an efficient sample size to assess the efficacy of a novel natural killer (NK) cell treatment compared to placebo. A decision as to whether the new regimen deserves a test in a large confirmatory phase III trial must be made at the conclusion of this study. The decrease in the expression level for polychlorinated biphenyl (PCB) is considered as the tumor response. Little information is available for both treatments at the planning stage, but it is expected that the treatment effects may be different in women and men. Therefore, a two-arm randomized trial stratified by gender is preferred. As in the previous section, this four-subgroup trial design reestimates the sample size after one interim analysis, to improve the chance of a conclusive decision regarding the treatment effect for both women and men. We also permit different sample sizes in the two groups.

In this study, the target difference in tumor response rates between the two treatments is set as $\theta^* = 0.2$. To retain the exploratory nature of our design, the threshold probabilities η_1 and η_2 are both fixed at 90%, fairly modest values. The proportion of the initial sample size for the interim time is fixed at $N_1 = N/3$. The desired conclusiveness level γ, or the threshold of predictive probability of conclusiveness, is set as 0.9. We apply very weak independent Beta(0.5,0.5) priors to the tumor response rates in all four subgroups: Trt+Women, Control+Women, Trt+Men and Control+Men. Figure 4.1 illustrates the timing of the proposed design and a potential outcome of sample size reestimation in the two subgroups.

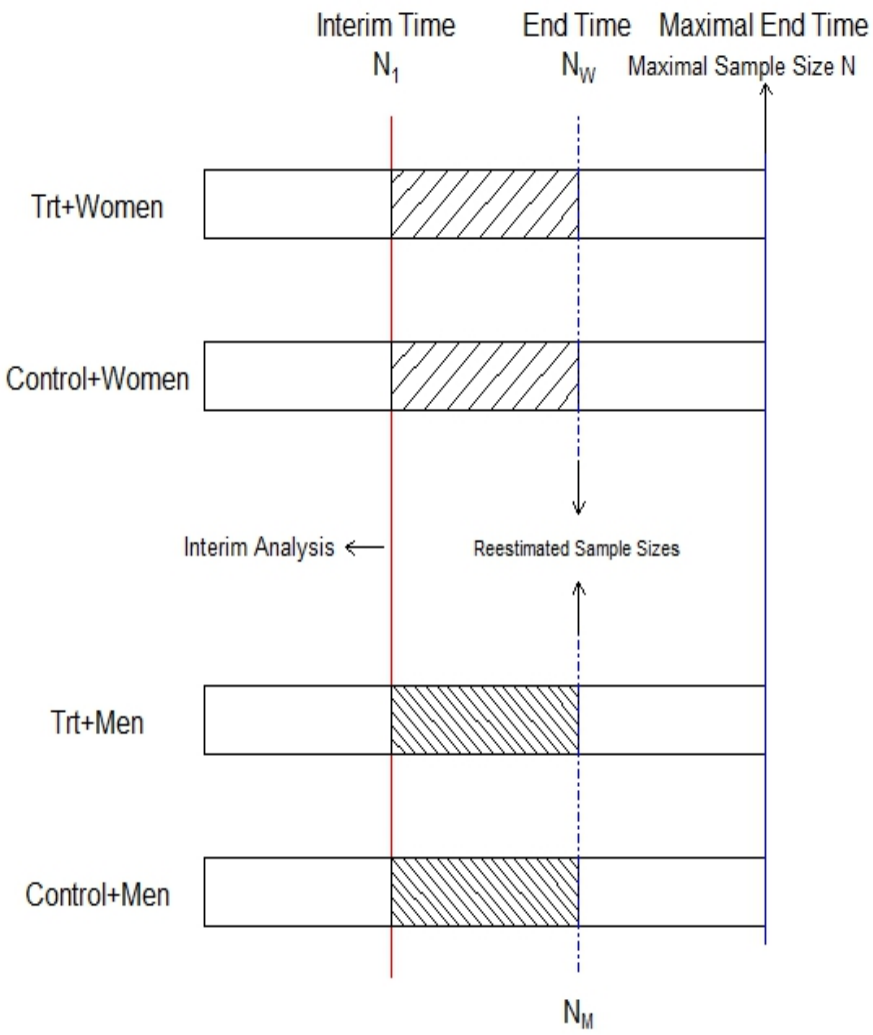

Figure 4.1: The application of our two-stage design with sample size reestimation in a II trial with a gender binary covariate.

We assume the binary outcome data is independently generated from binomial distributions with different tumor response rates in all four subgroups. To evaluate our new design, we consider the six different simulation scenarios given in Table 4.1. The

true tumor response rates for all four subgroups in each scenario are listed as p_{TW}, p_{CW}, p_{TM} and p_{CM}. The Δp_W and Δp_M represent the true effect differences between the treatments for women and men, respectively. The first scenario assumes that the new drug has a better treatment effect than placebo in both women and men (0.3 and 0.2, respectively). Scenario 2 instead supposes that the new treatment is superior to placebo only in women, but not in men (the differences are 0.2 and 0, respectively). In Scenario 3, no treatment difference exists between the two treatments across both gender strata. Scenario 4 indicates that the new drug is marginally better than placebo in both women and men, with the improvement in success rate as 0.1, less than the hoped-for value of $\theta^* = 0.2$. The next scenario shows the improvement in men is 0.3, while that in women is only 0.1, marginally better. The last scenario assumes that the new treatment is much more effective than placebo in women, but even *worse* than placebo in men.

Scenario	p_{TW}	p_{CW}	p_{TM}	p_{CM}	Δp_W	Δp_M
1	0.5	0.2	0.5	0.3	0.3	0.2
2	0.4	0.2	0.3	0.3	0.2	0
3	0.3	0.3	0.4	0.4	0	0
4	0.4	0.3	0.3	0.2	0.1	0.1
5	0.4	0.3	0.5	0.2	0.1	0.3
6	0.6	0.2	0.3	0.4	0.4	-0.1

Table 4.1: The numeric settings for all four subgroups in six different scenarios. The p_{TW} represents the true tumor response rate in women with the new drug treatment. The p_{CW} is for women with placebo treatment, p_{TM} is for men with new drug treatment, and p_{CM} is for men with placebo treatment. The Δp_W and Δp_M represent the treatment differences in women and men, respectively.

First, we calculate the initial sample size for each subgroup as in Subsection 4.1.1. The minimum sample size that is sufficient to make a conclusive statement on the treatment effect is 80 in each of the four subgroups. At the interim stage ($N_1 = 80 \times$

$1/3 \approx 27$), our design reestimates the sample size using the interim data by the predictive probability approach described in Section 4.1. Then, the conclusiveness condition is examined predictively using the new sample size. We simulated 1000 trials to investigate the operating characteristics of this approach using the `R` software. As a comparison, the performance of simply using the initial sample size generalized from Whitehead et al. (2008) without any sample size reestimation is also investigated.

Table 4.2 presents the simulated operating characteristics of our design, including the average sample size per subgroup, probabilities of early termination (PET) at the interim time, and conclusiveness probabilities at the end of the trial for women and men, respectively. Operating characteristics for the fixed-sample design using the generalized Whitehead et al. sample size can be found in Table 4.3. Both tables further break down the conclusiveness probabilities into their "minimal efficacy" and "futility" components.

Comparing the results in Tables 4.2 and 4.3, our proposed design results in a substantial decrease in sample size for all scenarios. Whereas the fixed-sample design requires 80 men and 80 women, our design requires an average of 28-45 women and 32-44 men, depending on the scenario. Furthermore, we achieve conclusiveness in all cases, even though we require only half the sample size on average.

One drawback to our proposed design is that we observe an increased probability of reaching an incorrect conclusion. For example, in Scenario 1 where $\Delta p_M = 0.2$, we conclude that $\Delta p_M < 0.2$ 18% of time, compared to only 11% for the fixed-sample design. A similar phenomenon is observed in Scenarios 2 and 3. These results should not be surprising. Although our model is estimated using the Bayesian paradigm, the results reported in Table 4.2 represent the frequentist operating characteristics of our

Scenario	$E(N_W)$	PET_W	Women P(efficacy) ($\Delta p_W > 0$)	P(futility) ($\Delta p_W < 0.2$)	P(conclusive) (W overall)
1	32.88	0.89	0.97	0.03	1.00
2	41.15	0.73	0.86	0.16	1.00
3	40.67	0.74	0.15	0.86	1.00
4	45.44	0.65	0.49	0.53	1.00
5	45.39	0.65	0.52	0.50	1.00
6	28.17	0.98	1.00	0.00	1.00
Scenario	$E(N_M)$	PET_M	Men P(efficacy) ($\Delta p_M > 0$)	P(futility) ($\Delta p_M < 0.2$)	P(conclusive) (M overall)
1	42.53	0.71	0.83	0.18	1.00
2	40.09	0.75	0.15	0.87	1.00
3	43.80	0.68	0.16	0.85	1.00
4	43.32	0.69	0.49	0.56	1.00
5	33.15	0.88	0.98	0.02	1.00
6	32.46	0.90	0.02	0.98	1.00

Table 4.2: Operating characteristics of our design. The $E(N_W)$ and $E(N_M)$ denote the average numbers in each treatment group for women and men, respectively. The PET_W and PET_M represent the probabilities of early termination due to conclusiveness at the interim time for women and men, respectively ($N_{Wmax} = N_{Mmax} = 80, \eta_1 = \eta_2 = 0.9, \gamma = 0.9$).

Scenario	Women P(efficacy) ($\Delta p_W > 0$)	P(futility) ($\Delta p_W < 0.2$)	P(conclusive) (W overall)	Men P(efficacy) ($\Delta p_M > 0$)	P(futility) ($\Delta p_M < 0.2$)	P(conclusive) (M overall)
1	1.00	0.00	1.00	0.91	0.11	1.00
2	0.92	0.12	1.00	0.11	0.92	1.00
3	0.09	0.94	1.00	0.12	0.90	1.00
4	0.51	0.56	1.00	0.58	0.59	1.00
5	0.53	0.53	1.00	1.00	0.00	1.00
6	1.00	0.00	1.00	0.00	1.00	1.00

Table 4.3: Operating characteristics of the design generalized from Whitehead et al. without any sample size reestimation ($N_W = N_M = 80$).

proposed design. As in any sequential procedure, allowing early termination at the interim look increases the Type I and Type II error rates.

Scenario	$E(N_W)$	PET_W	Women P(efficacy) ($\Delta p_W > 0$)	P(futility) ($\Delta p_W < 0.2$)	P(conclusive) (W overall)
1	51.48	0.91	1.00	0.00	1.00
2	68.10	0.72	0.93	0.08	1.00
3	66.53	0.74	0.06	0.95	1.00
4	84.72	0.53	0.52	0.52	1.00
5	86.80	0.51	0.52	0.52	1.00
6	44.87	0.99	1.00	0.00	1.00
Scenario	$E(N_M)$	PET_M	Men P(efficacy) ($\Delta p_M > 0$)	P(futility) ($\Delta p_M < 0.2$)	P(conclusive) (M overall)
1	72.45	0.67	0.92	0.08	1.00
2	65.49	0.75	0.07	0.94	1.00
3	71.75	0.68	0.07	0.94	1.00
4	80.02	0.59	0.55	0.55	1.00
5	51.13	0.92	0.99	0.01	1.00
6	50.00	0.93	0.01	0.99	1.00

Table 4.4: Operating characteristics of our design ($N_{Wmax} = N_{Wmax} = 131, \eta_1 = \eta_2 = 0.95, \gamma = 0.9$).

One approach to improving the percentage of correct conclusions is to adopt stricter stopping criteria. For example, we may raise the probability thresholds η_1 and η_2 from 0.9 to 0.95, which also increases the maximum sample size from 80 to 131 per subgroup for both women and men. The operating characteristics of our design under this setting are shown in Table 4.4. This change decreases the probability of reaching an incorrect conclusion to that observed for the fixed-sample case (Table 4.3). Increasing η_1 and η_2 also increases the expected sample size compared to Table 4.2, but the expected sample sizes are still less than those in the fixed-sample case in all but the most difficult scenarios, namely when Δp_M or $\Delta p_W = 0.1$ (exactly halfway between 0 and θ^*).

In addition, to check if our normal approximation method in Subsection 4.1.2 works

Scenario	$E(N_W)$	PET_W	Women P(efficacy) ($\Delta p_W > 0$)	P(futility) ($\Delta p_W < 0.2$)	P(conclusive) (W overall)
1	51.19	0.92	1.00	0.00	1.00
2	68.21	0.73	0.95	0.07	1.00
3	68.66	0.73	0.09	0.92	1.00
4	85.76	0.52	0.53	0.51	1.00
5	89.05	0.50	0.50	0.54	1.00
6	44.44	0.99	1.00	0.00	1.00
Scenario	$E(N_M)$	PET_M	Men P(efficacy) ($\Delta p_M > 0$)	P(futility) ($\Delta p_M < 0.2$)	P(conclusive) (M overall)
1	74.67	0.66	0.94	0.07	1.00
2	72.21	0.69	0.07	0.94	1.00
3	73.30	0.68	0.09	0.92	1.00
4	78.71	0.60	0.53	0.56	1.00
5	53.01	0.90	1.00	0.00	1.00
6	53.22	0.89	0.00	1.00	1.00

Table 4.5: Operating characteristics of our design by approximation method ($N_{Wmax} = N_{Wmax} = 131, \eta_1 = \eta_2 = 95\%, \gamma = 90\%$).

well, we compared to the approximate method in Equations (4.10)- (4.11), which utilizes both the BCLT approximation and Monte Carlo sampling from the predictive distribution. The results in Table 4.5 are very close to those in Table 4.4 by the computation of the exact distribution, while the computation speed of the former method is much faster (about 10 times faster by our simulation programs). To look into more details, we compared Equations (4.5)- (4.6) with Equations (4.10)- (4.11) on computation procedure. When computing the exact distribution, suppose the number of Monte Carlo samples to calculate $Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2})$ given particular M, s_{T2} and s_{C2} is W (W usually should be large enough to approximate the exact BDI distribution), then the total number of simulaiton samples in Equations (4.5) or (4.6) is

$W \times (M+1)^2$. In addition, there is extensive computation for the mass probabilities of beta-binomial distributions $P(s_{T2})$ and $P(s_{C2})$. In comparison, the total number of sampling in the BCLT approximation method is fixed as J. In our simulation program, we let $W = J = 4000$. Consider M varies from 0 to $N - N_1$ conditioning on the interim data, 10 times faster computing speed in our approximation method is not surprising.

4.3 Discussion

Our design inherits the properties of good interpretation and easy implementation from Whitehead et al. (2008), generalizes their method to a two-sample setting, and uses a fully Bayesian predictive approach to reduce an unnecessarily large sample size and save patients in exploratory studies. Moreover, we extend our method to multiple subgroups with varying categorical covariates, and allow flexible sample size within each subgroup based on interim analyses. With these merits, our design might be applied to many early phase studies, with consequent advances for personalized medicine.

Due to the different context between frequentists and Bayesians, when evaluating operating characteristics of a Bayesian approach by simulated data from fixed treatment effects, there is always an issue of how to best select control parameters such as η_1, η_2 and γ. Our results suggest that using stricter criteria in our method can reduce some errors and improve operating characteristics while not increasing sample size beyond Whitehead et al. levels. We also tried increasing γ from 0.9 to 0.95, but the corresponding impacts on the sample sizes and the percentages of making correct conclusions were not as great.

Of course, there are also some limitations to our method. First, although we use

logistic models for the case of many strata, the probability criteria are hard to determine because the interpretation for the transformation from the odds ratio to the absolute probability difference is not straightforward and depends on the baseline probability. Still more assumptions are required for the use of logistic regression models in clinical trials. For example, logistic regression often ignores some interaction terms between different covariates to decrease the number of model coefficients. In addition, in this paper we assumed independent binomial models for all subgroups, without considering any correlations between them. How to incorporate correlations into our design is a topic for future study. Finally, although the selection of the interim time and how it influences operating characteristics can be investigated by simulation studies, our future work also includes finding more general guidance for choosing N_1, the interim sample size, a subject that is also often discussed in frequentist sequential analysis.

Chapter 5

Conclusion

5.1 Summary of Major Findings

In this thesis, we have developed Bayesian methods for adaptive dose finding in phase I clinical trials, and sample size reestimation in exploratory studies. The first topic was covered in Chapters 2 and 3. The question addressed there was how to design a Bayesian trial when only part or none of the final efficacy data were available, and surrogate markers must instead be used. In other words, we proposed a new design that adaptively incorporates the toxicity, surrogate efficacy and partial efficacy data. In Chapter 2, our simulation studies indicated that compared to bivariate designs that simply replaced the efficacy data with the surrogate efficacy data, our proposed trivariate designs can successfully improve the empirical probability of correct dose selection and the proportion of trial participants assigned to the optimal dose. This improvement can be significant, especially when the quality of surrogate markers is poor. In addition, our designs can do as well as bivariate models in stopping the trial early due to over-toxicity with a high probability. Some design parameters (i.e., penalty weights) can also be flexibly adjusted to obtain the desired operating characteristics under different conditions.

Chapter 3 continued the work in Chapter 2, adding a plateau parameter and using a series of more flexible semiparametric and nonparametric monotone link functions to model the marginal probability of efficacy. We showed via simulation that our flexible link methods can outperform standard parametric CRM approaches in terms of both the probability of correct dose selection and the proportion of patients treated at the optimal dose.

In Chapter 4, we focused on the area of sample size reestimation in bivariate outcome

phase II trials. We adopted the concept of the "conclusiveness" condition proposed by Whitehead et al. [60] due to its good interpretation and easy implementation. We then generalized their method to a two-sample setting, and used a fully Bayesian predictive approach to reduce an unnecessarily large sample size. Moreover, we extended our proposed two-stage Bayesian design with sample size reestimation to subgroup analysis, thus allowing flexible sample size within each subgroup based on concurrent interim analyses. Therefore, our design might be applied to many early phase studies, with consequent advances for personalized medicine.

5.2 Extensions and Future Work

In adaptive dose finding, as discussed in Chapter 2, one extension of our design is to jointly model multiple surrogate markers for toxicity and efficacy. We could assume all the surrogate markers operate in "parallel", and investigate if the new model can borrow more strength from multiple surrogate markers to learn about the missing efficacy data. Another extension would be to check our trivariate model incorporating the surrogacy data for each marker separately. Then, we may pick the best surrogate marker based on the DIC criterion, which may be important for other studies.

One limitation of our trivariate probability model is that our conditional independence of S and T given E assumption may be too strong. The purpose here was to break a trivariate model into two bivariate submodels, where copulas have been well-studied. Our future work may include constructing schemes for higher dimensional copulas and applying them directly to toxicity, efficacy and (multiple) surrogate efficacy in adaptive dose finding. The association parameters in multivariate copulas might then be better

estimated and studied. In addition, a better quantification of the quality of surrogacy and its effect on dose-finding is also of interest.

In Chapter 3, we merely mentioned the "long memory" property of the CRM, and suggested the revision of decision rules when the "limited memory" property is important under some circumstances. The idea is that, while CRM maximizes the probability of selecting the correct dose, the trial-to-trial variability in this selection can be severe. Although we have obtained some preliminary simulation results (results not shown), further studies are required in this area.

In sample size reestimation, as discussed in Chapter 4, our future work should include how to incorporate correlation between subgroups into our design, and to find more general guidance for choosing the interim sample size. More challenges exist for how to use logistic regression models in sample size reestimation for continuous covariates without any stratification, which is also an interesting topic for our future investigation.

References

[1] P. Gallo, C. Chuang-Stein, V. Dragalin, B. Gaydos, M. Krams, and J. Pinheiro. Adaptive designs in clinical drug development??an executive summary of the phrma working group. *Journal of Biopharmaceutical Statistics*, 16(3):275–283, 2006.

[2] K.M. Anderson, D. Berry, N. Burnham, C. Chuang-Stein, J. Dudinak, P. Fardipour, P. Gallo, and S. Givens. Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal*, 43:539–556, 2009.

[3] S.M. Berry, B.P. Carlin, J.J. Lee, and P. Muller. *Bayesian Adaptive Methods for Clinical Trials*. CRC Press, 2010.

[4] FDA Draft Guidance for Industry - Adaptive Design Clinical Trials for Drugs and Biologics, The United State Food and Drug Administration, Rockville, Maryland., 2010.

[5] J.M. Collins, C.K. Grieshaber, and B.A. Chabner. Pharmacologically guided phase I clinical trials based upon preclinical drug development. *Journal of the National Cancer Institute*, 82(16):1321, 1990.

[6] B.E. Storer. Design and analysis of phase I clinical trials. *Biometrics*, 45(3):925–937, 1989.

[7] J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46(1):33–48, 1990.

[8] D. Faries. Practical modifications of the continual reassessment method for phase I cancer clinical trials. *Journal of Biopharmaceutical Statistics*, 4(2):147, 1994.

[9] S.N. Goodman, M.L. Zahurak, and S. Piantadosi. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine*, 14(11):1149–1161, 1995.

[10] J. O'Quigley and L.Z. Shen. Continual reassessment method: a likelihood approach. *Biometrics*, 52(2):673–684, 1996.

[11] T.A. Gooley, P.J. Martin, L.D. Fisher, and M. Pettinger. Simulation as a design tool for phase I/II clinical trials: an example from bone marrow transplantation. *Controlled Clinical Trials*, 15(6):450–462, 1994.

[12] P.F. Thall and K.E. Russell. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, 54(1):251–264, 1998.

[13] J. O'Quigley, M.D. Hughes, and T. Fenton. Dose-finding designs for HIV studies. *Biometrics*, 57(4):1018–1029, 2001.

[14] T.M. Braun. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials*, 23(3):240–256, 2002.

[15] P.F. Thall and J.D. Cook. Dose-finding based on efficacy–toxicity trade-offs. *Biometrics*, 60(3):684–693, 2004.

[16] G. Yin, Y. Li, and Y. Ji. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics*, 62(3):777–787, 2006.

[17] S. Fan and K. Chaloner. Optimal designs and limiting optimal designs for a trinomial response. *Journal of Statistical Planning and Inference*, 126(1):347–360, 2004.

[18] W. Zhang, D.J. Sargent, and S. Mandrekar. An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine*, 25(14):2365–2383, 2006.

[19] P.A. Murtaugh and L.D. Fisher. Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Communications in Statistics-Theory and Methods*, 19(6):2003–2020, 1990.

[20] N.B. Bekele and Y. Shen. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics*, 61(2):343–354, 2005.

[21] M. Gasparini and J. Eisele. A curve-free method for phase I clinical trials. *Biometrics*, 56(2):609–615, 2000.

[22] R.L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989.

[23] R. Temple. Are surrogate markers adequate to assess cardiovascular disease drugs? *JAMA*, 282(8):790–795, 1999.

[24] T.R. Fleming and D.L. DeMets. Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*, 125(7):605, 1996.

[25] J.M. Lachin. Sample size determinations for r x c comparative trials. *Biometrics*, 33:315–324, 1977.

[26] DJ Spiegelhalter and LS Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13, 1986.

[27] W.J. Shih, P.L. Zhao, et al. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine*, 16(17):1913–1923, 1997.

[28] E.A. Gehan. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, 13(4):346–353, 1961.

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

[30] C. Jennison and B.W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press, 2000.

[31] A.L. Gould. Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine*, 20(17-18):2625–2643, 2001.

[32] J.S. Denne. Sample size recalculation using conditional power. *Statistics in Medicine*, 20(17-18):2645–2660, 2001.

[33] T. Friede and M. Kieser. Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics*, 3(4):269–279, 2004.

[34] B.W. Brown, J. Herson, E. Neely Atkinson, and M. Elizabeth Rozell. Projection from previous studies: a bayesian and frequentist compromise. *Controlled Clinical Trials*, 8(1):29–44, 1987.

[35] D.J. Spiegelhalter, L.S. Freedman, and M.K.B. Parmar. Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine*, 12(15-16):1501–1511, 1993.

[36] B. Lecoutre. Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*, 79(1):93–105, 1999.

[37] S.J. Lee and M. Zelen. Clinical trials and sample size considerations: another perspective. *Statistical Science*, 15(2):95–110, 2000.

[38] T. Pham-Gia and N. Turkkan. Sample size determination in Bayesian analysis. *The Statistician*, 41:389–397, 1992.

[39] L. Joseph, D.B. Wolfson, and R. Du Berger. Some comments on Bayesian sample size determination. *The Statistician*, 44(2):167–171, 1995.

[40] H. Pezeshk. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research*, 12(6):489–504, 2003.

[41] N. Stallard. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics*, 54:279–294, 1998.

[42] K. Claxton, L.F. Lacey, and S.G. Walker. Selecting treatments: a decision theoretic approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):211–225, 2000.

[43] SK Sahu and TMF Smith. A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):235–253, 2006.

[44] M.D. Wang. Sample size reestimation by Bayesian prediction. *Biometrical Journal*, 49(3):365–377, 2007.

[45] J. Woodcock. The prospects for "personalized medicine" in drug development and drug therapy. *Clinical Pharmacology & Therapeutics*, 81(2):164–169, 2007.

[46] M.A. Hlatky, D.B. Boothroyd, D.M. Bravata, E. Boersma, J. Booth, M.M. Brooks, D. Carrié, T.C. Clayton, N. Danchin, M. Flather, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. *The Lancet*, 373(9670):1190–1197, 2009.

[47] A.M. Garber and S.R. Tunis. Does comparative-effectiveness research threaten personalized medicine? *New England Journal of Medicine*, 360(19):1925–1927, 2009.

[48] BC Arnold and DJ Strauss. Bivariate distributions with conditionals in prescribed exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):365–375, 1991.

[49] M.K. Cowles and B.P. Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[50] Y. Li, B.N. Bekele, Y. Ji, and J.D. Cook. Dose-schedule finding in phase I/II clinical trials using a Bayesian isotonic transformation. *Statistics in Medicine*, 27(24):4895–4913, 2008.

[51] G. Yin and Y. Yuan. A latent contingency table approach to dose finding for combinations of two agents. *Biometrics*, 65(3):866–875, 2009.

[52] S.J. Mandrekar, R. Qin, and D.J. Sargent. Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: Methods and challenges. *Statistics in Medicine*, 29(10):1077–1083, 2010.

[53] D.A. Berry, P. Mueller, A.P. Grieve, M. Smith, T. Parke, R. Blazek, N. Mitchard, and M. Krams. Adaptive bayesian designs for dose-ranging drug trials. *Case studies in Bayesian statistics*, 5:99–181, 2001.

[54] A.E. Gelfand and B.K. Mallick. Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics*, 51:843–852, 1995.

[55] B.P. Carlin and J.S. Hodges. Hierarchical proportional hazards regression models for highly stratified data. *Biometrics*, 55(4):1162–1170, 1999.

[56] P. Diaconis and D. Ylvisaker. Quantifying prior opinion. In *Bayesian Statistics*, Volume 2, J. M. Bernardo, M. H. De Groot, D. V. Lindley, and A. F. M. Smith. (eds), 133-156. Amsterdam: North Holland , 1985.

[57] Y.K. Cheung. On the use of nonparametric curves in phase I trials with low toxicity tolerance. *Biometrics*, 58(1):237–240, 2002.

[58] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[59] A.P. Oron and P.D. Hoff. Small-sample behavior of novel phase I designs. Technical report, Department of Statistics, University of Washington. 2011.

[60] J. Whitehead, E. Valdés-Márquez, P. Johnson, and G. Graham. Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*, 27(13):2307–2327, 2008.

[61] T. Pham-Gia and N. Turkkan. Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods*, 22(6):1755–1771, 1993.

[62] B.P. Carlin and T.A. Louis. *Bayesian methods for data analysis*, volume 78. Chapman & Hall/CRC, 2009.