Dose-Finding Designs for Early-Phase Clinical Trials and Outcome Dependent Sampling for Longitudinal Studies of Gene-Environment Interaction

by

Zhichao Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Thomas M. Braun, Co-Chair
Professor Bhramar Mukherjee, Co-Chair
Research Assistant Professor Philip S. Boonstra
Assistant Professor Sung Kyun Park

To my parents, my husband, and my daughter

ACKNOWLEDGEMENTS

Accomplishing this doctoral dissertation has been one of the most challenging tasks I have ever faced. I am lucky to have many nice people around to help and support me, and make the doctoral life more enjoyable. To them I owe my deepest appreciation.

First of all, I would like to express my sincere gratitude to my co-advisers, Dr. Thomas M. Braun and Dr. Bhramar Mukherjee, for their tremendous mentorship, patient guidance, and continuous inspiration throughout my PhD training. This dissertation would not have been possible without their lasting encouragement and support. I am truly fortunate to have Dr. Braun as my adviser, who has impressed me with his wide range of knowledge, his insightful opinions, his positive attitude towards research, and his superb scientific writing skills. I greatly appreciate Dr. Mukherjee for generously funding me throughout my doctoral study. As a woman professor, she is not only a great adviser but also an excellent mentor/friend, who is full of responsibility and enthusiasm, and always available to discuss my questions both academic and personal related. I am also deeply grateful to my committee members, Dr. Sung Kyun Park and Dr. Philip S. Boonstra, for serving on my dissertation committee and providing valuable input and critical suggestions on my research.

Second, I would like to thank all the faculty members and graduate students in the department of biostatistics. I have learned a lot from professors' brilliant lectures. I enjoyed the free environment of sharing opinions on research and established long-lasting friendships with many graduate students, Yi-An Ko, Yebin Tao, Zhuqing Liu, Yin-Hsiu Chen, and others. I believe this experience will definitely become a precious asset in my future career.

Last but not least, I would like to thank my parents, Zhenming Sun and Min Shen, for their unconditional love. Thanks to my husband Xiao and my daughter Ashlyn, who has been my emotional anchor, for making me strong and making every day of my life meaningful.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Dose-Finding Designs for Early-Phase Clinical Trials and Outcome Dependent Sampling for Longitudinal Studies of Gene-Environment Interaction

by
Zhichao Sun

Co-Chair: Thomas M. Braun

Co-Chair: Bhramar Mukherjee

In this dissertation, we develop study designs for (I) early-phase clinical trials with the goal of identifying the maximum tolerated combination of two agents based on toxicity alone, or the optimal biological dose (OBD) based on bivariate outcomes of toxicity and efficacy, and (II) longitudinal cohort studies with a focus on gene-environment (GxE) interactions.

In the first project, we extend the nonparametric biased coin design (BCD) for studying a single agent to a two-stage adaptive procedure that can be easily implemented for dual-agent Phase I trials. The basic idea of our design is to divide the entire trial into two stages and apply the BCD, with modification, in each stage. Through simulations we show that our design is competitive with four contemporary parametric approaches and promotes patients safety by limiting patient exposure to toxic combinations.

In the second project, we propose two designs for Phase I/II trials when the dose-efficacy curve plateaus within the dose range of interest. We incorporate multiple sets of pre-specified efficacy probabilities and use Bayesian model averaging to address misspecification in the pattern of efficacy with dose. Dose assignment is determined adaptively by maximization of the posterior selection probability among the set of admissible doses. The simulation results demonstrate that our designs identify the OBD effectively and allocate patients around the OBD frequently when compared to a competing approach designed for non-monotonic dose-efficacy curves.

To investigate GxE interaction in longitudinal studies, in the third project, we propose exposure enriched outcome trajectory dependent designs that can inform sample selection by leveraging individual exposure and outcome trajectory, and develop a full conditional likelihood (FCL) analysis that corrects for the biased sampling. We compare the performance of our proposed designs and analysis to combinations of different sampling designs and estimation approaches via simulation. We observe that the FCL provides improved estimates for both interaction and joint exposure effects over uncorrected complete-case analysis, and the exposure enriched outcome trajectory dependent design enhances the estimation efficiency and detection power for the GxE interaction compared to random selection of subjects. We also illustrate the utility of our designs and analysis by an example from the Normative Aging Study, a longitudinal study of Boston area veterans.

CHAPTER I

Introduction

Study design is important to the success of a study because it governs how the data will be collected. A well-designed study can greatly increase the efficiency of the study and strengthen the conclusions from the study. In this dissertation, we will focus on the study design for two specific problems, early-phase dose-finding trials and longitudinal studies of gene-environment interaction. Statistical analysis is closely related to the design of the study, so for each of the problems we consider we will discuss both study design and its corresponding statistical analysis.

1.1 Early-Phase Clinical Trials

Drug development is a continuous process through which the safety and efficacy of a new drug is evaluated. Conventionally, it constitutes four phases: Phase I determines a safe dosage range and identifies side effects of the drug in a small group of people. Phase II assesses the efficacy, in addition to the safety, of the drug in a larger group of people. Phase III provides a thorough examination on the effectiveness of the drug, through a direct comparison to a standard therapy or placebo, using hundreds to thousands of patients. Phase IV monitors the long-term effectiveness and side effects after FDA approval of the drug.

As the first study involving humans, the primary goal of Phase I trials is to identify the maximum tolerated dose (MTD), which is a dose whose rate of dose limiting toxicities (DLT) is closest to a pre-specified target. There are a number of approaches developed for the design of single-agent Phase I clinical trials, which have been well summarized in recent reviews and book chapters (Rosenberger and Haines, 2002; Chevret, 2006; Ting, 2006). The simplest approach is the algorithm-based 3+3 design which relies on the philosophy that MTD is readily identifiable from the data. Patients are enrolled in cohorts of three, with each one in the cohort receiving the same dose. If one patient in the current cohort experiences DLT, and at least one patient among an additional cohort at the same dose experiences DLT, the trial is terminated with the MTD defined as one dose lower than the last assigned dose level (Storer, 1989). In spite of its simplicity in implementation, the 3+3 design has been criticized by being inflexible and over-conservative (Ivanova, 2006).

When considering the MTD as a parameter to be estimated, O'Quigley *et al.* (1990) proposed a model-based design, the Continual Reassessment Method (CRM). The CRM is a Bayesian adaptive procedure which updates the posterior estimates of the dose toxicity probabilities as the trial proceeds and assigns the next cohort to the dose closest to the MTD. Extensions and modifications to the CRM to address patient safety have been proposed (Goodman *et al.*, 1995; Møller, 1995; Leung and Wang, 2002), and a maximum likelihood version of the CRM also exists (O'Quigley and Shen, 1996). Other parametric designs are described in Braun (2014).

In contrast to the model-based approaches, Durham and Flournoy (1995) developed a nonparametric approach they named the Biased Coin Design (BCD). In the BCD, if a patient experiences a DLT on a dose, the next patient will be assigned to a lower dose. If a patient does not experience a DLT on a dose, the next patient will be randomized to the same dose or a higher dose through the flip of a biased coin. This randomization induces a random walk that has a limiting distribution of dose assignment, with its mode peaked around a pre-specified quantile (Durham *et al.*, 1997). Other examples of up-and-down designs, a family of designs of which the BCD is a member, are described in Ivanova *et al.* (2003). Advantages of the BCD include its simplicity of utility, its flexibility of specifying any targeted toxicity probability, its performance comparable to parametric designs like the CRM (Durham *et al.*, 1997), and its asymptotic properties based upon Markov Chain theory (Durham and Flournoy, 1995).

In this dissertation, we focus on study designs of Phase I clinical trials that can be extended into two directions. The first extension is to identify the maximum tolerated combination (MTC) for two agents administered together. In comparison to single agent trials, it is often more complex to develop an approach for dual-agent Phase I trials. The first difficulty involves the potential synergistic or antagonistic effect of the two agents. Provided that an interaction effect exists, the combined risk of toxicity would markedly differ from the sum of the two marginal risks of toxicity estimated separately, leading to the complete order of the joint risk of toxicities not being fully predictable and a number of dose combinations exhibiting similar toxicities (Conaway *et al.*, 2004). A second difficulty relates to the dilemma of limited sample size relative to the number of combinations being studied. As the number of dose levels for each agent increases, the two-dimensional search space expands multiplicatively. For example, in a combination trial in which each agent has 4 dose levels, an exhaustive search of the whole space requires approximately 100 subjects (16 combinations $\times$ 6 subjects/combination), which is unrealistic for a dose-finding trial, as few trials have a sample size larger than 50 in practice. A third difficulty focuses on the selection of the admissible set of combinations and the choice of the dose escalation scheme. Unlike a one-to-one correspondence in a single-agent trial, 3 assignment actions (escalation, stay, and de-escalation) can be associated with 9 possible dose combinations (current combination and 8 adjacent combinations) for subsequent patients in a two-dimensional space. In Chapter II, we aim to develop a simple adaptive algorithm that allows for efficient identification of multiple MTCs, while exerts a control over excessive toxicities based upon the theory of the single agent BCD.

Another problem with current Phase I clinical trials is that, dose-finding on the basis of toxicity, while ignoring efficacy, may no longer be appropriate for further investigation when the dose-efficacy curve reaches a plateau at nontoxic dose levels. With cytotoxic agents in oncology trials, the general belief is that the more toxic dose level is, the more effective it will be. In contrast, with cytostatic molecularity targeted agents (MTAs), efficacy may not continuously increase with dose escalation, but level off at certain dose due to their distinct biological mechanism (Korn *et al.*, 2001). In an effort to address this problem, in Chapter III, we aim to develop dose-finding strategies for MTAs by incorporating both toxicity and efficacy as endpoints in seamless Phase I/II trials. The primary objective of such a trial is to identify the optimal biological dose (OBD) that is safe and most efficacious by imposing pre-specified constraints on the toxicity and efficacy probabilities.

1.2 Longitudinal Studies of Gene-Environment Interaction

Investigation on the etiology of complex human diseases has led to increased interest in the genetic and environment risk factors. Accumulating evidence has suggested the presence of gene-environment (GxE) interaction in that the association between an environmental exposure and an outcome can be changed for a subgroup with a certain genotype (Hunter, 2005). For instance, physical activity was found to attenuate the effect of *FTO* variant on the waist size in an Indian health study (Moore *et al.*, 2012). However, it is often difficult to detect a GxE interaction due to the large sample required, as compared to a marginal genetic or environmental effect. In case-control studies, for example, sample size in the thousands of subjects is needed for GxE interaction using candidate genes, while tens of thousands are needed in a genome-wide search. Consequently, data collection for studies of GxE interaction would become prohibitive given a high genotyping or exposure assay cost.

While there has been extensive literature on statistical analysis for estimating or testing multiplicative GxE interactions (Mukherjee *et al.*, 2012b), in Chapter IV of this dissertation, we address the problem from a different perspective, by considering study designs that can improve the detection power and estimation efficiency of a GxE interaction.

Two-phase sampling design is known to be of great use when the cost of measuring a specific covariate of interest is high relative to others. In a two-phase design for GxE interaction, suppose inexpensive data including environmental exposure, response, and other covariates are collected in a Phase I sample. These data are utilized to help inform the selection of subsample in Phase II where subjects are genotyped. In studies with budgetary constraints, two-phase design, commonly used for case-control studies, can enhance the precision of estimates (Chatterjee and Mukherjee, 2008).

Longitudinal cohort design has been recommended for GxE interactions over case-control study mainly because of its ability to characterize time-varying genetic susceptibility, and potentially time-dependent GxE interactions, both of which are important to describe the complex genetic architecture of a disease-related quantitative trait and its implications on public health. In Chapter IV, we aim to develop study designs/sampling schemes that combine features of two-phase sampling and longitudinal cohort study in order to better detect/evaluate the GxE interactions.

CHAPTER II

A Two-Dimensional Biased Coin Design for Dual-Agent Dose-Finding Trials

2.1 Introduction

In a Phase I clinical trial, the primary goal is to determine the highest dose with a dose limiting toxicity (DLT) rate closest to a pre-specified target; this dose is named the maximum tolerated dose (MTD). Phase I clinical trials typically have small sample sizes and are driven by the ethical constraint of minimizing patients exposed to toxic doses while controlling the likelihood that patients are treated at ineffective doses (Rosenberger and Haines, 2002). Given the limited efficacy observed with single agents, there is growing interest in identifying the maximum tolerated combination (MTC) for two agents, such as two agents with different biological targets that may show enhanced treatment efficacy when used together. Paller *et al.* (2014), as part of an NCI task force examining combination trials in oncology, published a consensus of their findings, and they produced an excellent survey of previous combination trials in oncology and a series of recommendations for the design future of combination trials.

Beyond the examples provided by Paller et al., we provide two more recent examples of combination trials in oncology. Berenson *et al.* (2009) conducted a combination trial examining the safety, tolerability and initial efficacy of samarium lexidronam and bortezomib for patients with relapsed or refractory multiple myeloma. Gandhi *et al.* (2014) presented a Phase I trial that combined neratinib and temsirolimus, which were found in preclinical data to be synergistic with respect to inhibiting tumor growth, suggesting their use for treating patients with advanced solid tumors. Unfortunately, these studies and many of their counterparts have tended to use simplistic designs and have motivated a body of methodology focused on statistically sound designs for combination trials that belong to two general classes of parametric designs. We emphasize that we consider designs that are based solely on toxicity and do not incorporate efficacy data.

One strategy simplifies the two-dimensional search by splitting the trial into several subtrials and confining exploration of dose combinations within each subtrial. An example is the independent design employed by Kuzuya *et al.* (2001), in which the dose of one agent is fixed in each subtrial and the dose of the other agent varies. Yuan and Yin (2008) suggested conducting a series of subtrials in sequential order to reduce the overall sample size and eliminate suboptimal combinations examined in the independent design.

An alternative parametric design allows simultaneous modification of doses for both agents. Thall *et al.* (2003) described a Bayesian two-stage design that models the joint toxicity risk of dose combinations via a six-parameter logistic regression model. Wang and Ivanova (2005) proposed a linear model for the dose-toxicity relationship on the basis of standardized and log transformed dose levels, with two parameters characterizing the marginal effects and one parameter for the interaction effect. Yin and Yuan (2009a) and Yin and Yuan (2009c) developed copula-type models in which the joint toxicity risk is modeled as a function of the marginal risk of each agent, as well as correlation parameter to quantify any interaction. Braun and Wang (2010) proposed a hierarchical design that uses a Beta-Binomial model and links the hyperparameters to dose levels through log-linear regression models, and a proportional odds logistic regression model for the joint toxicity risk was used by Braun and Jia (2013).

Despite the empirical evidence supporting the utility of parametric approaches, there are some concerns over the assumed models, either the over-parameterization of complex models relative to the limited sample size used in dose-finding trials, or the failure of a parsimonious model to account for interactions or other complicated relationships between the two agents (Gasparini, 2013). The idea of partial ordering was developed as an attempt to address both of these limitations and was first proposed by Kramar *et al.* (1999), who introduced the idea of specifying a prior ordering of toxicity risks for a selected subset of dose combinations. Conaway *et al.* (2004) and Wages *et al.* (2011) recently extended this idea to multiple-agent trials by evaluating all possible orders of combinations.

Nonparametric methods have also been suggested as a remedy to the perceived concerns with parametric models. The most common nonparametric approach is often some variant of the standard 3+3 algorithm for single agent trials (Fan *et al.*, 2009; Hamberg *et al.*, 2010; Braun and Alonzo, 2011). In a slightly different vein, Ivanova and Wang (2004) proposed a design in which the direction of dose escalation of either agent depends jointly upon the number of DLTs in the most recent cohort and the cumulative toxicity probability of the current dose combination . Their approach is nonparametric in the sense that no formal model is used to explain how the probability of DLT is related to doses of both agents and is an extension of the Narayana design developed for single-agent designs (Ivanova *et al.*, 2003).

Another member of nonparametric designs for single agent trials is the Biased Coin Design (BCD), which is a simple up-and-down algorithm that determines dose assignments for future patients using the outcomes of enrolled patients without the use of a formal model (Durham and Flournoy, 1995; Durham *et al.*, 1997). We propose a two-stage BCD that would be appropriate for combination trials such as those of Berenson *et al.* (2009) and Gandhi *et al.* (2014) that we described earlier. Our design inherits the favorable properties of the BCD and enables adaptive learning from sequentially conducted stages. In Section 2.2, we describe the dose-finding algorithm and estimation procedure of our proposed design. Section 2.3 describes the steps necessary for implementing our design and presents the dose assignments that might occur in an actual trial to demonstrate how our design works. Section 2.4 presents operating characteristics of our proposed design in comparison with four existing designs under different simulation settings, and Section 2.5 concludes our work with a discussion.

2.2 Methods

Let us consider a dual-agent dose finding trial with J discrete dose levels of agent A, $A_1 < A_2 < \cdots < A_J$, and K discrete dose levels of agent B, $B_1 < B_2 < \cdots < B_K$, under investigation. The maximum number of patients to be enrolled in the trial is N. Let A_jB_k represent the combination of agent A at dose level A_j, $j = 1, ..., J$, and agent B at dose level B_k, $k = 1, ..., K$, and let Y_{ijk} denote the binary outcome of experiencing DLT for patient i, $i = 1, ..., N$, treated at combination A_jB_k. The probability of DLT at this combination is $\pi_{jk} = Pr\{Y_{ijk} = 1\}$; we assume the probability of DLT increases with dose escalation for one or both agents, i.e. $\pi_{jk} \leq \pi_{j'k'}$ for $j \leq j'$ and $k \leq k'$. If the target probability of DLT is pre-specified at γ ($\gamma \leq 0.5$), our goal is to identify a set of combinations whose estimated probabilities of DLT are closest to the target, similar to the rule used by the Continual Reassessment Method (CRM) for single-agent trials (O'Quigley *et al.*, 1990).

2.2.1 BCD for a Single Agent

The original BCD was created for identifying the MTD of a single agent, say agent A, using up-and-down rules (Durham and Flournoy, 1995). If dose A_j has been given to patient i, the assignment of patient $i+1$ is based upon the outcome of patient i. If patient i experiences DLT, patient $i+1$ receives dose level A_{j-1}. If patient i does not experience DLT, we flip a biased coin with probability of heads $b = \gamma/(1-\gamma)$. If the coin shows tails, patient $i+1$ receives dose A_j, and if the coin shows heads, patient $i+1$ receives dose A_{j+1}. Naturally, as our target DLT rate γ increases, the probability of heads also increases, so that we are more likely to escalate in the absence of DLTs.

Durham and Flournoy (1995) showed that the random walk induced by the BCD is irreducible and recurrent. Thus, there is a guarantee that a stationary distribution of dose assignments exists, with a unique mode peaked around the targeted quantile. Furthermore, under the sequential allocation of dose levels to patients, the fast convergence rate of the random walk makes it possible for the favorable properties of the asymptotic distribution to hold in a finite sample size. It has been suggested that within the class of up-and-down designs, the BCD is the optimal one with regard to the "peakedness" of its stationary distribution around the mode (Giovagnoli and Pintacuda, 1998). With a sample size of 25 patients under numerous simulation configurations, similar operating characteristics between the CRM and the BCD have been seen (Durham *et al.*, 1997).

2.2.2 BCD-2d: Stage 1

Description of how the BCD can be applied to a dose combination study first requires a visual schematic of the dose combinations under study. Figure 2.1 visually presents the 16 dose combinations examined in a hypothetical trial of four doses of each agent. The first stage of our design examines the set of ordered combinations $\mathbb{S}^1 = \{A_1B_1$, A_2B_2, ..., $A_\ell B_\ell\}$ in which $\ell = min\{J, K\}$. In reference to Figure 2.1, $\ell = 4$ and $\mathbb{S}^1$ is simply all combinations lying on the diagonal of the grid. Since the DLT rates of the combinations in $\mathbb{S}^1$ are ordered, it is straightforward to use the BCD to determine the combination in $\mathbb{S}^1$

assigned to each patient in Stage 1. Specifically, we assign the first patient to combination A_1B_1 and then determine the assignments of future patients using the BCD as described in Section 2.2.1.

Our idea of first searching "diagonally" among the grid of combinations is a discretized approach to the method used by Thall *et al.* (2003) in their parametric design. The approach of Thall et al. is motivated by the belief that there is a contour that connects combinations from the upper left corner of the grid to the lower right corner of the grid such that all combinations on the contour have the desired DLT rate γ. By searching along the diagonal starting at combination A_1B_1, we expect to eventually intersect with the contour and examine a combination that also lies on, or is close to, this contour. It is this combination that would ideally be identified as the MTC by the end of Stage 1 with our design. And certainly the concept of using a "lead-in" subset of combinations to direct the focus of a later stage examining remaining combinations has been suggested by others as well (Yuan and Yin, 2008; Fan *et al.*, 2009).

Given that Stage 1 is really an exploratory stage among a small fraction of all combinations under study, we suggest that the number of patients allocated to Stage 1, which we denote as N_1, be approximately $1/3$ of the total sample size available to the entire study N. Thus, with a total sample size of $N = 50$, which is not uncommon for combination studies, this allocates 16-17 patients to Stage 1, which is close to the size of a traditional single-agent study. It was suggested by a reviewer that a more equal allocation of patients to Stages 1 and 2 might be possible as well. However, in simulations described in Section 2.4, we found that a 50/50 allocation of patients did not improve the operating characteristics of the BCD-2d and actually increased the rate of early termination in settings where termination was not necessary. The idea of an imbalanced allocation scheme is also proposed by Thall *et al.* (2003).

At the end of Stage 1, we smooth the observed DLT rates for the combinations in $\mathbb{S}^1$ using isotonic regression and select as the MTC the dose combination whose estimated DLT rate is closest to the target γ; we let $A_m B_m$ denote this combination. One issue with isotonic regression is that often two or more combinations will have the same smoothed DLT rate. Continuing with our example in Figure 2.1, suppose we assign one, two, four, and three subjects, respectively, to combinations $A_1B_1, A_2B_2, A_3B_3,$ and A_4B_4, and we observe 0/1, 1/3, 1/4, and 1/3 DLTs, respectively, at each combination. The corresponding smoothed DLT rates from isotonic regression would be $0, 2/7, 2/7,$ and $1/3$. If the target DLT rate were $\gamma = 0.3$, then both combinations A_2B_2 an A_3B_3 would qualify as the MTC after Stage 1. In general, to break such ties for the MTC, we choose the combination of lowest doses when the DLT rates are equal to or above γ, and we choose the combination of highest doses when the DLT rates are below γ. It is also possible that two combinations will have smoothed DLT rates in which one DLT rate is less than γ by the same amount that the other DLT rate is above γ. If this occurs, we move forward to Stage 2 with the combination of lower doses.

We note that in Stage 1, we also incorporate a stopping rule for early termination of the study when combination A_1B_1 appears to be overly toxic. Specifically, if at any time in the trial the number of patients assigned to combination A_1B_1 who experience DLT reaches a threshold of C_1, the trial will be terminated. Otherwise, enrollment will continue until all N_1 patients have been enrolled in Stage 1.

2.2.3 BCD-2d: Stages 2a and 2b

In Stage 2, we take the remaining $N_2 = N - N_1$ patients and allocate half of them to Stage 2a, which examines all combinations in $\mathbb{S}^{2a} = \{A_jB_k, j \geq m, k \leq m\}$, i.e. the combinations “above and to the left” of A_mB_m. The other $N_2/2$ patients are allocated to

Stage 2b, which examines all combinations in $\mathbb{S}^{2b} = \{A_jB_k, j \leq m, k \geq m\}$ i.e. the combinations "below and to the right" of A_mB_m. Dose-finding in Stages 2a and 2b will be conducted in parallel in order to minimize the length of the trial. By the assumption of a monotonic dose-toxicity relationship for both agents, DLT rates of combinations in $\mathbb{S}^{2a}$ and $\mathbb{S}^{2b}$ are bounded by the DLT rates of combinations $A_{m-1}B_{m-1}$ and $A_{m+1}B_{m+1}$, and it is possible to find combinations in $\mathbb{S}^{2a}$ and $\mathbb{S}^{2b}$ that have DLT rates close to γ. We now assign two patients to combination A_mB_m, one to direct dose assignments in Stage 2a and one to direct dose assignments in Stage 2b and observe both patients for occurrence of DLT.

Unlike the traditional BCD for doses of a single agent, the decisions to "escalate," "de-escalate," and "remain" at the current dose with the BCD-2d do not have unique actions associated with them. If the current patient has been assigned to combination A_jB_k, there are a maximum of nine combinations to consider; the number of combinations is less than nine whenever the current combination contains the highest or lowest dose of one of the agents, which we refer to as a "boundary" condition. "Escalation" occurs when the dose of either or both agents is increased, i.e. combinations $A_jB_{k+1}, A_{j+1}B_k$, and $A_{j+1}B_{k+1}$. "De-escalation" occurs when the dose of either or both agents is decreased, i.e. combinations $A_jB_{k-1}, A_{j-1}B_k$, and $A_{j-1}B_{k-1}$. The other three combinations A_jB_k, $A_{j-1}B_{k+1}$, and $A_{j+1}B_{k-1}$ are considered "remain" because the ordering of the DLT rates of $A_{j-1}B_{k+1}$ and $A_{j+1}B_{k-1}$ relative to that of A_jB_k is unknown. Within each of these decisions, we emphasize that we prefer to move as far from the current assignment as possible to promote exploration through the entire space of combinations.

We now focus upon Stage 2a, as the same decision rules apply directly to Stage 2b. If the current patient in Stage 2a experiences a DLT, the BCD states de-escalation is necessary for the next patient. For this de-escalation, we will simultaneously decrease the doses of both agents, if that combination is a part of $\mathbb{S}^{2a}$. If that combination is not possible, then we consider dose de-escalation of only one agent. If both of these single-agent dose reductions are possible, we randomly choose one of them with equal probability.

If the current patient in Stage 2a does not experience a DLT, we flip a biased coin with probability of heads equal to $b = \gamma/(1-\gamma)$. If the coin shows heads, the BCD states that escalation is necessary for the next patient. We will simultaneously increase the doses of both agents, if that combination is a part of $\mathbb{S}^{2a}$. If that combination is not possible, then we consider dose escalation of only one agent. If both of these single-agent dose increases are possible, we randomly choose one of them with equal probability. If the coin instead shows tails, the BCD states that we should "remain." Since our goal is to explore as much as possible, we first consider the two combinations in which the dose of one agent is increased and the dose of the other agent is decreased. We select the combination that is a member of $\mathbb{S}^{2a}$ and randomly choose one if both are members. If neither is a member, then we remain at the current combination. Lastly, at any point in the trial, if none of the possible assignments are members of $\mathbb{S}^{2a}$, then assignment should remain at the current combination.

We highlight the fact that boundary conditions on $\mathbb{S}^{2a}$ and $\mathbb{S}^{2b}$ will restrict the possible assignments for future patients. Using our example presented in Figure 1, if the current patient in Stage 2a were assigned to combination A_4B_1 and experienced a DLT, further dose de-escalation of Agent B is not possible, and combination A_3B_1 would be the only assignment option for the next patient. As another example, if the current patient in Stage 2a were assigned to combination A_3B_1 and experienced a DLT, we could not further de-escalate to combination A_2B_1 because that combination is not an element of $\mathbb{S}^{2a}$. As a result, we would assign the next patient to combination A_3B_1 as well.

Each of the rules above is also subject to a stopping rule based upon the cumulative observed number of DLTs seen at each combination. We select a threshold of C_2, and if at least C_2 patients, all assigned to combination A_jB_k, experience DLTs, combination A_jB_k, and any combination with a higher dose of either agent, can no longer be assigned to future patients. For example, using our example described in Figure 1, if an excessive number of DLTs were seen for combination A_3B_2, then this combination, as well as combinations A_3B_3, A_3B_4, A_4B_2, A_4B_3, and A_4B_4 could no longer be assigned to future patients.

Our design allocates an equal number of patients to each of Stage 2a and Stage 2b and those sample sizes are fixed at the beginning of the trial. In settings where there are more toxic combinations in Stage 2a relative to 2b, or vice versa, we might want to adaptively alter the number of patients allocated to each stage. To that end, we propose modification to the BCD-2d that allows for adaptive allocation of patients to Stages 2a and 2b; we call this design BCD-2da. We let p_{2a} and p_{2b} denote the collective probabilities of experiencing DLT for all combinations in Stage 2a and 2b, respectively, and we place a prior Beta (θ, θ) distribution on both p_{2a} and p_{2b}, with θ being a fixed value determined before Stage 2 commences. At any point in Stage 2, we will have enrolled n_{2a} patients in Stage 2a and n_{2b} patients in Stage 2b, and we let m_{2a} and m_{2a} denote the accumulated number of DLTs observed among all combinations in Stage 2a and 2b, respectively. The respective posterior means for p_{2a} and p_{2b} would be $(m_{2a} + \theta)/(n_{2a} + 2\theta)$ and $(m_{2b} + \theta)/(n_{2b} + 2\theta)$, and we would assign the next patient to the stage with the lower posterior mean, as it suggests that stage has a lower expectation of DLTs among its combinations. If the two posterior means are tied, we would randomly assign the next patient to one of the stages with equal probability.

When all N patients are completely followed to the end of the trial, we use bivariate isotonic regression (Bril *et al.*, 1984) to estimate the DLT probabilities for $J \times K$ combinations, and identify dose combinations with their estimates closest to the target γ as the MTCs. In settings where multiple estimated DLT probabilities are equidistant from γ, recommendation as MTCs will be given to: (a) combinations with a smaller sum of discrete dose levels $(j + k)$ if $\widehat{\pi}_{jk} > \gamma$; (b) combinations with a larger sum of dose levels if $\widehat{\pi}_{jk} < \gamma$; or (c) all tied combinations if $\widehat{\pi}_{jk} = \gamma$.

2.3 Guidelines for Practical Implementation

In order to use the BCD-2d in practice, investigators must first define quantities that are a part of all two-agent designs: (a) the number of doses of Agent A (J) and doses of Agent B to study (K), (b) the targeted DLT rate (γ), and (c) the maximum sample size (N). As we stated earlier, a total sample size of $N = 50$ seems to be common for most designs, and the sample size in Stage 1 should be approximately $N_1 = N/3$, which is 16 or 17 patients. These sample sizes are simply guidelines; a sensitivity analysis of operating characteristics examining a small set of potential sample sizes should certainly be done when designing an actual trial.

Beyond these universal quantities, investigators must further consider two additional quantities specific to the stopping rules in the BCD-2d. The first quantity is C_1, the cumulative number of DLTs in Stage 1 observed in patients assigned to combination A_1B_1. The second quantity is C_2, the cumulative number of DLTs in Stage 2 observed on any combination that then makes that combination and all combinations consisting of higher doses, no longer available for assignment to future patients. Certainly the values for C_1 and C_2 will be context-dependent and different values of both quantities should be examined in a sensitivity analysis to determine the trade-off between limiting patient exposure to possibly toxic combinations (smaller value of C_1 and/or C_2) and continuing with the trial and increasing the likelihood of identifying the MTC (larger value of C_1 and/or C_2).

To illustrate how assignments are made for each patient by the BCD-2d without adaptive assignments, we consider the setting examined by Gandhi *et al.* (2014), which was a clinical trial to assess combinations of four doses of Neratinib (120, 160, 200, and 240 mg), which in our notation is Agent A, and four doses of Temsirolimus (15, 25, 50, and 75 mg), which in our notation is Agent B. We enroll a total sample size of $N = 50$, and we allocate the first $N_1 = 16$ patients to Stage 1. Our targeted DLT rate is $\gamma = 0.20$.

Table 2.1 contains three pieces of information for each patient in this trial: (i) the assigned combination, (ii) an indicator of DLT, and (3) the result of the coin flip, if necessary. The first patient is assigned to combination A_1B_1 and did not have a DLT. The resulting coin toss produced a head, so no escalation occurs and the second patient is also assigned to combination A_1B_1. This pattern continues until the fifth patient who does not have a DLT and the coin toss now results in a head. Thus, escalation occurs with the sixth patient. It is not until the eleventh patient that another head results in absence of DLT. Thus, escalation occurs again with patient 12, who then experiences DLT. Thus, no coin toss is necessary and de-escalation occurs with patient 13. Patients 14 and 15 are also assigned to combination A_2B_2, and Patient 16 is assigned to combination A_1B_1 because patient 15 experienced a DLT.

After all $N_1 = 16$ patients have been enrolled and observed for DLT in Stage 1, the observed DLT rates for combinations $A_1B_1, A_2B_2, A_3B_3, A_4B_4$, are 1/6, 1/9, 1/1, and 0/0, respectively. Isotonic regression leads to combination A_2B_2 having an estimated DLT rate closest to our target so that the first patient in each of Stages 2a and 2b is assigned to combination A_2B_2. We note that the stopping rule threshold for Stage 1 was $C_1 = 3$. Since the number of DLTs at combination A_1B_1 was below C_1, Stage 1 was not terminated early.

In Stage 2a, assignment is restricted within $\mathbb{S}^{2a} = \{A_2B_1, A_2B_2, A_3B_1, A_3B_2, A_4B_1, A_4B_2\}$. Patient 17 does not experience DLT, and the coin toss results in a tail, leading to a decision to "remain." Thus, patient 18 is assigned to combination A_3B_1 (A_1B_3 was not considered because it is not a member of $\mathbb{S}^{2a}$). Since patient 18 goes on to experience DLT, de-escalation is warranted. Thus, patient 19 is assigned to combination A_2B_1 since no lower dose of Agent B is available. The next four successive patients (20, 21, 22, and 23) are also assigned to combination A_2B_1 in the absence of further DLTs and coin tosses all showing tails. At this point, patient 23 experiences DLT, but further de-escalation is not possible because combination A_1B_1 is not a member of $\mathbb{S}^{2a}$. Thus, at this point in Stage 2a, the only way to move from combination A_2B_1 is for a patient to not have a DLT and a coin toss producing a head, which is exactly what happens with patient 26. We leave the reader to work through the remaining assignments in Stage 2a.

In Stage 2b, assignment is restricted within $\mathbb{S}^{2b} = \{A_1B_2, A_2B_2, A_1B_3, A_2B_3, A_1B_4, A_2B_4\}$. We leave the reader to reason through the assignments of the first seven patients, at which point patient 40 has been assigned to combination A_2B_4. Due to boundary conditions, future assignments cannot deviate from this combination until a DLT is observed. Thus, even though we do toss a coin after the next several patients, all of whom do not experience DLT, the outcomes of those coin tosses are irrelevant. It is not until patient 48 experiences a DLT that we decide to de-escalate the assignment for patient 49 to combination A_1B_3.

Once all 50 patients have been assigned a combination and observed for DLT, we use bivariate isotonic regression to compute smoothed DLT rates for all combinations and find that combinations A_2B_2, A_2B_3, and A_2B_4 all have an estimated DLT rate of 0.17 that is closest to the target γ. We select combination A_2B_4 as the MTC because the sum of its dose levels (2+4) is the largest among the three combinations. Our threshold value in Stage 2 was $C_2 = 3$, so no combinations were eliminated in Stage 2a nor Stage 2b because the cumulative number of DLTs at combinations in either stage did not exceed the threshold value.

User-friendly code written for the statistical package R that can be used for implementing the BCD-2d (as well as the adaptive version of the BCD-2d) in an actual trial can be found at http://www-personal.umich.edu/$\sim$tombraun/software.html.

2.4 Simulations

2.4.1 Description of Settings and Methods

We sought to evaluate the performance of the BCD-2d, both with and without adaptive assignment in Stage 2, in a variety of plausible settings for two-agent trials. These settings varied not only in the pattern of DLT rates, but also in the number of dose combinations being examined and the targeted DLT rate. We present the results for ten of these settings, the DLT rates of which are given in Table 2.2. Scenarios A-F are trials studying combinations of four doses of each of two agents, Scenarios G and H are trials studying combinations of five doses of each agent, and Scenarios I and J are trials studying combinations of six doses of each agent. The targeted DLT rate is $\gamma = 0.2$ for Scenarios A-F and $\gamma = 0.3$ for Scenarios G-J.

Regardless of the setting, we set the maximum sample size to $N = 50$ patients, of whom $N_1 = 16$ patients will be enrolled in Stage 1. We also examined maximum sample sizes of $N = 40$ and $N = 60$, and found that increasing the sample size from $N = 40$ to $N = 50$ led to better performance in terms of both MTC identification and patient assignment, while further increasing the sample size from $N = 50$ to $N = 60$ did not appreciably change the results, supporting $N = 50$ as an appropriate sample size in our trials. We also tried using a balanced sample size distribution with $N/2$ patients allocated to Stage 1 and $N/2$ patients allocated to Stage 2 ($N_1 = N_2 = N/2$), but found no improvement in operating characteristics and an increased rate of early termination when termination was unwarranted.

We compare the operating characteristics of the BCD-2d and BCD-2da with four existing approaches: the copula-based approaches of Yin and Yuan (2009a) and Yin and Yuan (2009c), the proportional odds logistic regression model of Braun and Jia (2013), as well as the continual reassessment method for partial ordering (POCRM) (Wages *et al.*, 2011). We used the same priors, skeleton DLT rates and stopping rules as described in Braun and Jia (2013) for the first three approaches. For the POCRM, we generated skeletons by the algorithm of Lee and Cheung (2009) with a indifference interval half-width of 0.03, enforced an initial escalation scheme that simultaneously increases the doses of both agents before occurrence of the first DLT, and utilized the R package *pocrm* to run simulated trials. No early stopping rule is specified in the POCRM, so we used the same one as in the BCD-2d.

In the BCD-2d, the sample size in Stage 2a and Stage 2b is fixed at 17 patients, while in the BCD-2da we specified $\theta = 3$ as the hyperparameter for the prior distribution of the collective probabilities p_{2a} and p_{2b}. We examined other values of θ, but found that patient allocation is relatively insensitive to value of θ, and we felt that $\theta = 3$ led to a prior that was not overly informative and allowed the posterior means of the p_{2a} and p_{2b} to adequately alter allocation when the number of DLTs varied greatly between Stages 2a and 2b.

For Stage 1, we set our stopping rule threshold to $C_1 = 3$ as the maximum number of allowable DLTs for combination A_1B_1 before terminating the entire trial. Likewise, for Stage 2, we set our stopping rule threshold $C_2 = 3$ as the maximum number of allowable DLTs for a combination in Stages 2 that would remove that combination and more toxic combinations from further study in that stage.

For each scenario and each of the six designs examined, we simulated 2,000 hypothetical trials and summarized the performance of the designs with three quantities: the proportion of simulations in which specific combinations were identified as the MTC after Stage 2, the average percentage of patients that were assigned to specific combinations in both Stages 1 and 2, and the average number of DLTs observed in the trial.

2.4.2 Summary of Operating Characteristics

Table 2.3 presents the simulation results for Scenarios A-F in which sixteen combinations are examined and only one of those combinations is the MTC. Overall, the operating characteristics of the BCD-2d and BCD-2da are extremely similar to each other in all the scenarios, suggesting that the ability of the BCD-2d to identify the MTC and assign patients is fairly insensitive to incorporation of our suggested adaptive allocation scheme. The BCD-2d and BCD-2da also have the lowest average number of observed DLTs among all the designs, which provides strong evidence that our design is at least as safe, and perhaps safer, than other published designs. This finding is impressive, especially for a setting such as Scenario E in which Stage 2a will examine mostly safe combinations, whereas Stage 2b will examine mostly toxic combinations. Nonetheless, we see in Table 2.3 that the BCD-2d and BCD-2da assign patients within a 10-point neighborhood (in terms of DLT rate) of the MTC as well as the other designs.

In Scenario A where all sixteen combinations are safe, we see all designs do very well in terms of identifying the MTC, with a slight edge to the competing designs over the BCD-2d and BCD-2da. The conservative nature of the BCD-2d and BCD-2da is also obvious in this setting with a greater proportion of patients assigned to combinations with DLT rates less than 0.10. Scenario B is a setting in which most combinations are toxic and toxicity increases faster with dose increases of Agent A than with Agent B. Although we see that all designs identify the MTC within a 10-point neighborhood of the true MTC in a majority of simulations, the BCD-2d and BCD-2da is more likely than its competitors to identify the

MTC at a combination whose true DLT is more than 10 points beyond the target. Likewise, the BCD-2d and BCD-2da assign combinations outside the 10-point neighborhood of the correct MTC more often than the other designs. This result occurs because combinations along the diagonal have DLT rates that increase rapidly and combination A_1B_1 has a DLT rate closest to the target γ and should be recommended at the end of Stage 1. However, in our simulations, we found that the BCD-2d selected combinations A_1B_1 and A_2B_2 in 53% and 36% of simulations, respectively, as the MTC at the end of Stage 1.

Scenario C represents a situation in which all combinations are excessively toxic and early termination of the trial is desired and demonstrates that all the designs examined tend to have equivalent safety profiles. In Scenario D, where the DLT rates increase greatly with increases of Agent B and the DLT rates of all 16 combinations span over a reasonable range (0.08-0.41), we see better performance of the BCD-2d and BCD-2da relative to the competing designs, as reflected by the appreciable increase in the likelihood of identifying the MTC in a 10-point neighborhood of the correct MTC. We also see competitive operating characteristics among the designs in Scenarios E and F that suggest that the BCD-2d and BCD-2da work as well as the other designs, especially in settings where a simple additive model is not able to account for the dose-toxicity pattern.

Table 2.4 contains the operating characteristics of the designs in settings where greater numbers of dose combinations are examined and more than one combination is the MTC. The priors used by the Bayesian designs in Scenarios G and H were calibrated using the DLT rates of Scenario G in Table 2.2 and for Scenarios I and J, the priors were calibrated using the DLT rates of Scenario I in Table 2.2. In all four settings, we see that our designs identify the MTC at combinations whose actual DLT rates are within 10 points of the target more often than the competing designs, although the ability to identify the MTC at a combination whose DLT rate is exactly equal to the target varies among the designs and among the settings.

Figure 2.2 is a plot of the average DLT rate of the combinations assigned to each patient in Scenarios A to J using the BCD-2d as a way to assess how well assignments converge over time toward a combination with the targeted DLT rate. Scenario C is not presented because all the combinations under study are toxic and there is no possible convergence toward the targeted DLT rate. In Scenario A, we see that all combinations have DLT rates below the target; as a result, convergence remains below the targeted DLT rate, but gets as close to the targeted DLT rate as possible. In the other eight scenarios, we see excellent convergence by the end of Stage 1, regardless of the number of combinations being examined.

Convergence in Stage 2 is much more variable from setting to setting and is highly dependent upon the actual DLT rates of the combinations that are examined in Stages 2a and 2b. For example, in Scenarios B, D, F, G, and J, we see that patients near the end of Stages 2a and 2b are very likely to be assigned to combinations with DLT rates close to the target, but in Scenario E, patients in Stage 2a are much more likely to be assigned to combinations with DLT rates above the target simply because Stage 2a has little chance of containing either combination A_1B_1 or A_2B_2, whose DLT rates are below the target. We also see slightly poorer convergence in Stage 2 for Scenarios H and I, but the positive side of these results is that the BCD-2d tends to err toward combinations with DLT rates below the target whenever possible.

All of the scenarios presented to this point have been in settings with equal numbers of doses of both agents. However, the BCD-2d can be applied to combination trials in which the two agents have unequal numbers of doses. We investigated the operating characteristics of the BCD-2d in such settings by taking Scenarios B, D, and E displayed in Table 2.2 and removing the highest dose of Agent B. This results in three scenarios, referred to as Scenarios B$'$, D$'$, and E$'$, with four dose levels of Agent A and three dose levels of Agent B. We again simulated 2,000 trials under each of these scenarios, and the operating characteristics of the BCD-2d are summarized in Table 2.5. In all three scenarios, we see that the BCD-2d identifies combinations with similar probabilities and assigns similar proportions of patients in the neighborhood of the MTC as it did in Scenarios B, D and E.

2.5 Discussion

In this study, we have proposed a procedure for dual-agent dose-finding trials that is an extension of the BCD proposed for single agent trials. In the numerous simulation scenarios we have examined, our method performs well in terms of identification of MTC and allocation of patients relative to the performance of existing parametric methods. Since the BCD-2d makes no parametric assumptions regarding how DLT rates vary with doses of both agents, our proposed design is robust to model misspecification. The BCD-2d also inherits the desirable properties of the original BCD, including its flexibility of targeting any DLT rate and its fast convergence of assignments to a stationary distribution. We feel a strength of our design is conducting the trial in two sequential stages, which uses the information collected on the combinations in Stage 1 to inform the study as to which combinations to examine in Stage 2. By doing so, we expect to increase the likelihood of patients being treated to combinations in a neighborhood of the MTC and improve the precision of the estimated DLT rates for those combinations.

We stated in Section 2.2.2 that our intent is to promote exploration over the entire grid of combinations so that each future assignment differs from the previous assignment as much as possible. However, as we see in the example presented in Section 2.3, several patients can end up with the same assignment due to boundary restrictions. An important future area of research would be developing methods that increase the rate of exploration of the BCD-2d. However, this increased exploration must also be balanced with the desire to treat as many patients as possible to combinations with DLT rates at or close to the targeted DLT rate. Related to this concept is exploring other options to our adaptive allocation rule suggested for the BCD-2da, which, although reasonable, is quite simplistic.

Save for one exception (Cheung, 2013), sample size determination is an imprecise science for most Phase I clinical trial designs, which as we state, is done through a sensitivity analysis comparing the operating characteristics that result from several candidate sample sizes. However, because a closed-form expression for the expected number of assignments per dose in the original BCD is available (Durham *et al.*, 1995), we are pursuing work that will provide practical guidance toward a sample size calculation for dual-agent trials, both in terms of the total sample size, as well as how much of that sample size is distributed among Stages 1, 2a, and 2b.

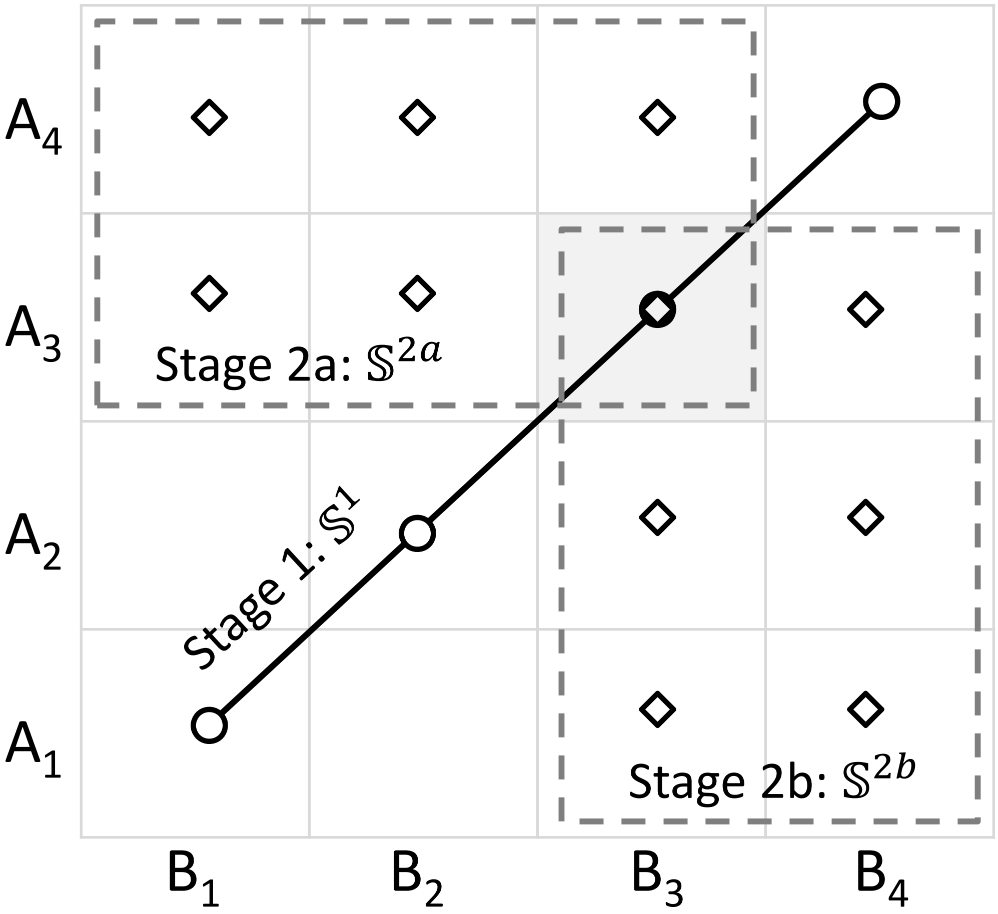
A4
A3
A2
A1
B1
B2
B3
B4
Stage 2a: $\mathbb{S}^{2a}$
Stage 1: $\mathbb{S}^{1}$
Stage 2b: $\mathbb{S}^{2b}$

Figure 2.1: Illustration of three stages divided in the BCD-2d with four dose levels for each agent. Dose combinations to be tested in Stage 1 are represented by circles. Assume combination A_3B_3 is identified as the MTC after Stage 1 (shaded in grey), dashed rectangles indicate refined search spaces of Stages 2a and 2b, and diamonds indicate combinations to be examined in Stages 2a and 2b.

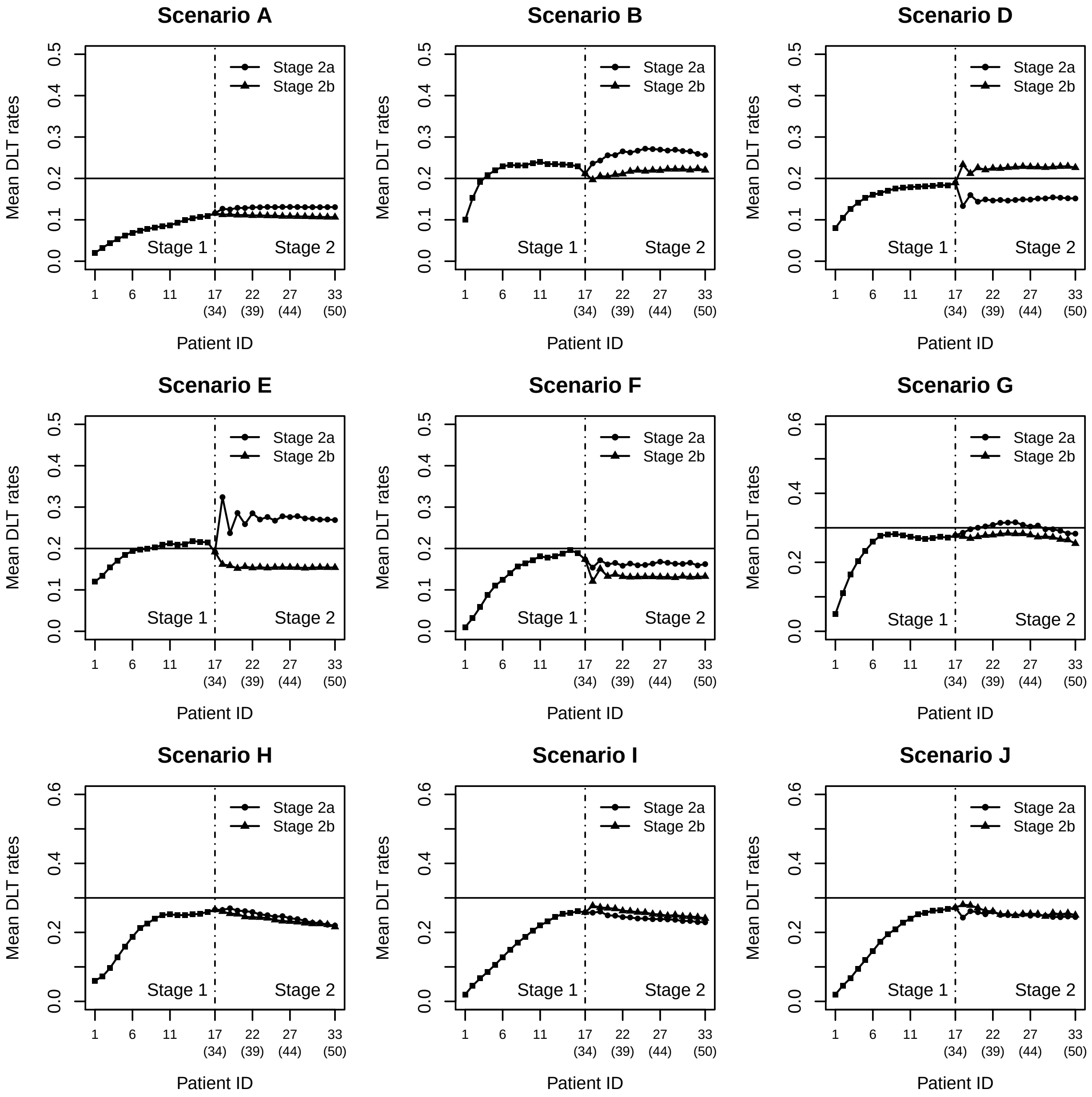

Figure 2.2: Average DLT rates of combinations sequentially assigned to each patient using the BCD-2d. The horizontal line shows the targeted DLT rate, and the vertical dashed line represents the start of Stage 2a and Stage 2b.

Table 2.1: Data resulting from a hypothetical trial using the BCD-2d. For each patient, 0=no DLT and 1=DLT; H=heads, T=tails, and ●=no coin toss necessary.

Stage 1				Stage 2a				Stage 2b			
	Combo		Coin		Combo		Coin		Combo		Coin
Patient	Assigned	DLT	Toss	Patient	Assigned	DLT	Toss	Patient	Assigned	DLT	Toss
1	A_1B_1	0	T	17	A_2B_2	0	T	34	A_2B_2	0	T
2	A_1B_1	0	T	18	A_3B_1	1	●	35	A_1B_3	0	T
3	A_1B_1	0	T	19	A_2B_1	0	T	36	A_2B_2	1	●
4	A_1B_1	0	T	20	A_2B_1	0	T	37	A_1B_2	0	T
5	A_1B_1	0	H	21	A_2B_1	0	T	38	A_1B_2	0	H
6	A_2B_2	0	T	22	A_2B_1	0	T	39	A_2B_3	0	H
7	A_2B_2	0	T	23	A_2B_1	1	●	40	A_2B_4	0	T
8	A_2B_2	0	T	24	A_2B_1	0	T	41	A_2B_4	0	T
9	A_2B_2	0	T	25	A_2B_1	0	T	42	A_2B_4	0	T
10	A_2B_2	0	T	26	A_2B_1	0	H	43	A_2B_4	0	H
11	A_2B_2	0	H	27	A_3B_2	0	T	44	A_2B_4	0	H
12	A_3B_3	1	●	28	A_4B_1	1	●	45	A_2B_4	0	T
13	A_2B_2	0	T	29	A_3B_1	0	T	46	A_2B_4	0	T
14	A_2B_2	0	T	30	A_4B_2	0	H	47	A_2B_4	0	H
15	A_2B_2	1	●	31	A_4B_2	1	●	48	A_2B_4	1	●
16	A_1B_1	1	●	32	A_3B_1	1	●	49	A_1B_3	0	T
				33	A_2B_1	0	●	50	A_2B_2	0	●

Table 2.2: True DLT rates for ten scenarios in a combination trial, with the targeted DLT probability shown in bold.

	Agent B							Agent B					
Agent A	k=1	k=2	k=3	k=4	k=5	k=6	Agent A	k=1	k=2	k=3	k=4	k=5	k=6
	Scenario A							*Scenario B*					
j=4	0.11	0.13	0.15	0.17			j=4	0.55	0.65	0.75	0.85		
j=3	0.08	0.10	0.12	0.14			j=3	0.40	0.50	0.60	0.70		
j=2	0.05	0.07	0.09	0.11			j=2	0.25	0.35	0.45	0.55		
j=1	0.02	0.04	0.06	0.08			j=1	0.10	**0.20**	0.30	0.40		
	Scenario C							*Scenario D*					
j=4	0.62	0.66	0.70	0.74			j=4	0.11	0.21	0.31	0.41		
j=3	0.56	0.60	0.64	0.68			j=3	0.10	**0.20**	0.30	0.31		
j=2	0.50	0.54	0.58	0.62			j=2	0.09	0.19	0.29	0.30		
j=1	0.44	0.48	0.52	0.56			j=1	0.08	0.18	0.28	0.29		
	Scenario E							*Scenario F*					
j=4	0.50	0.52	0.54	0.55			j=4	0.10	0.30	0.50	0.80		
j=3	0.44	0.45	0.46	0.47			j=3	0.06	0.15	0.30	0.45		
j=2	0.16	0.18	**0.20**	0.22			j=2	0.04	0.10	0.15	**0.20**		
j=1	0.12	0.13	0.14	0.15			j=1	0.01	0.02	0.03	0.04		
	Scenario G							*Scenario H*					
j=5	0.45	0.56	0.64	0.72	0.80		j=5	0.19	**0.30**	0.42	0.50	0.60	
j=4	**0.30**	0.42	0.52	0.62	0.70		j=4	0.10	0.19	**0.30**	0.41	0.50	
j=3	0.20	**0.30**	0.40	0.50	0.60		j=3	0.09	0.12	0.20	**0.30**	0.40	
j=2	0.12	0.20	**0.30**	0.40	0.50		j=2	0.08	0.09	0.12	0.20	**0.30**	
j=1	0.05	0.10	0.20	**0.30**	0.40		j=1	0.06	0.08	0.10	0.12	0.20	
	Scenario I							*Scenario J*					
j=6	0.10	0.20	**0.30**	0.40	0.50	0.60	j=6	0.10	**0.30**	0.40	0.50	0.60	0.65
j=5	0.08	0.14	0.20	**0.30**	0.40	0.60	j=5	0.08	0.14	0.20	**0.30**	0.41	0.60
j=4	0.06	0.12	0.15	0.20	**0.30**	0.40	j=4	0.06	0.12	0.15	**0.30**	0.40	0.60
j=3	0.04	0.10	0.13	0.16	0.20	**0.30**	j=3	0.04	0.10	0.13	0.16	0.20	**0.30**
j=2	0.03	0.08	0.12	0.14	0.18	0.20	j=2	0.03	0.08	0.12	0.14	0.18	0.20
j=1	0.02	0.06	0.10	0.12	0.15	0.18	j=1	0.02	0.06	0.10	0.12	0.15	0.18

Table 2.3: Summary of simulations from Scenarios A to F: average probability (%) of selection as the MTC and the mean percentage of allocation for combinations whose DLT rates are at target $\gamma = 0.2$, within 10 points of γ, and beyond 10 percentages from γ, and the average number of DLTs. The five designs (described in Section 2.4.1) are the proportional odds logistic regression design (gCRM), two copula-based designs (YY09a & YY09b), the CRM for partial ordering (POCRM), and the proposed BCD-2d and BCD-2da. No selection or allocation to further patients will be made when a early stopping condition is met.

	Selection				Allocation				Number
Design	At γ	1-10 pts of γ	>10 pts of γ	None	At γ	1-10 pts of γ	>10 pts of γ	None	of DLTs
					Scenario A				
gCRM	0	94	3	3	0	87	13	0	8
YY09a	0	99	1	0	0	86	14	0	8
YY09b	0	96	4	0	0	85	15	0	8
POCRM	0	95	5	0	0	85	15	0	7
BCD-2d	0	88	12	0	0	63	37	0	5
BCD-2da	0	89	11	0	0	64	36	0	5
					Scenario B				
gCRM	45	39	5	11	30	41	18	11	17
YY09a	41	50	5	4	27	54	16	3	16
YY09b	42	47	5	6	29	55	11	5	16
POCRM	30	52	9	9	20	49	22	9	12
BCD-2d	22	48	21	9	12	51	24	13	10
BCD-2da	23	51	18	8	13	52	26	9	10
					Scenario C				
gCRM	0	0	4	96	0	0	22	78	12
YY09a	0	0	1	99	0	0	20	80	10
YY09b	0	0	1	99	0	0	16	84	12
POCRM	0	0	9	91	0	0	19	81	7
BCD-2d	0	0	9	91	0	0	19	81	4
BCD-2da	0	0	9	91	0	0	19	81	1
					Scenario D				
gCRM	9	70	14	7	5	56	32	7	15
YY09a	6	65	27	2	9	55	24	2	14
YY09b	7	67	25	1	6	54	38	2	15
POCRM	14	69	15	3	9	61	27	3	11
BCD-2d	9	81	7	3	3	60	32	5	9
BCD-2da	11	80	6	3	5	62	31	2	9
					Scenario E				
gCRM	13	70	6	11	10	64	16	10	16
YY09a	14	76	6	4	7	75	14	4	15
YY09b	12	74	7	7	7	77	9	7	16
POCRM	14	64	13	9	10	60	21	9	11
BCD-2d	11	67	13	9	3	69	17	11	10
BCD-2da	12	70	9	9	4	73	14	9	9
					Scenario F				
gCRM	25	68	5	2	18	57	24	1	15
YY09a	12	76	12	0	3	71	26	0	15
YY09b	15	72	13	0	7	61	32	0	15
POCRM	34	59	7	0	20	58	22	0	10
BCD-2d	21	70	9	0	8	59	33	0	7
BCD-2da	21	71	8	0	9	59	32	0	7

Table 2.4: Summary of simulations from Scenarios G to J: average probability (%) of selection as the MTC and the mean percentage of allocation for combinations whose DLT rates are at target $\gamma = 0.3$, within 10 points of γ, and beyond 10 percentages from γ, and the average number of DLTs. The five designs (described in Section 2.4.1) are the proportional odds logistic regression design (gCRM), two copula-based designs (YY09a & YY09b), the CRM for partial ordering (POCRM), and the proposed BCD-2d and BCD-2da. No selection or allocation to further patients will be made when a early stopping condition is met.

	Selection				Allocation				Number
Design	At γ	1-10 pts of γ	>10 pts of γ	None	At γ	1-10 pts of γ	>10 pts of γ	None	of DLTs
Scenario G									
gCRM	52	19	29	0	34	20	42	4	16
YY09a	55	17	26	2	42	20	36	2	16
YY09b	53	23	23	1	41	25	33	1	14
POCRM	52	12	31	5	38	17	40	5	16
BCD-2d	38	40	17	5	16	40	36	8	12
BCD-2da	37	43	16	4	17	43	36	4	12
Scenario H									
gCRM	49	17	30	4	32	16	47	5	15
YY09a	51	7	41	1	38	9	51	2	13
YY09b	47	14	39	0	31	13	55	1	12
POCRM	47	15	35	3	33	14	50	3	14
BCD-2d	35	26	36	3	15	23	58	4	11
BCD-2da	34	29	34	3	16	22	59	3	11
Scenario I									
gCRM	41	28	31	0	27	22	50	1	15
YY09a	32	26	40	2	21	20	58	1	15
YY09b	34	27	39	0	22	21	55	2	15
POCRM	40	18	42	0	27	18	55	0	13
BCD-2d	33	47	20	0	14	37	49	0	11
BCD-2da	31	46	23	0	14	34	52	0	10
Scenario J									
gCRM	45	18	35	2	30	15	55	0	15
YY09a	30	19	49	2	22	13	62	3	15
YY09b	31	20	48	1	22	12	64	2	15
POCRM	28	22	50	0	24	15	61	0	13
BCD-2d	35	30	35	0	19	17	64	0	12
BCD-2da	37	28	35	0	18	17	65	0	10

Table 2.5: Summary of simulations from Scenarios B$'$, D$'$ and E$'$ using the BCD-2d for combination trials with unequal dose levels of two agents: average probability (%) of selection as the MTC and the mean percentage of allocation for combinations whose DLT rates are at target γ, within 10 points of $\gamma = 0.2$, and beyond 10 percentages from γ, and the average number of DLTs. No selection or allocation to further patients will be made when a early stopping condition is met.

	Selection				Allocation				
Scenario	At γ	1-10 pts of γ	>10 pts of γ	None	At γ	1-10 pts of γ	>10 pts of γ	None	DLTs
B$'$	23	51	16	10	12	51	23	14	10
D$'$	10	81	6	4	3	63	29	5	8
E$'$	19	59	13	9	8	65	15	11	9

CHAPTER III

A dose-Finding Design for Phase I/II Trials When the Dose-Efficacy Curve Plateaus

3.1 Introduction

Toxicity and efficacy of chemotherapeutic agents in oncology are often evaluated in two separate phases, termed Phase I and Phase II. A Phase I trial examines the toxicity profile of several dose levels of an agent and identifies the maximum tolerated dose (MTD) that has a dose limiting toxicity rate closest to and no greater than a pre-specified target. A Phase II trial assesses the efficacy of the agent at the established MTD with the assumption that both toxicity and efficacy increase monotonically with dose escalation so that the MTD has the greatest likelihood of efficacy while being safe. However, the emergence of molecularly targeted agents (MTAs) has challenged such a sequential paradigm (Korn *et al.*, 2001). By interacting specifically with molecules involved in tumor growth, MTAs are expected to be less toxic than chemotherapeutic agents, and their efficacy may no longer follow the monotonicity assumption but plateau at certain dose level beyond which no higher efficacy can be achieved (Jain *et al.*, 2010; LeTourneau *et al.*, 2010; LoRusso *et al.*, 2010). As a result, the MTD based solely on toxicity in Phase I is no longer appropriate for further investigation in Phase II trials. For these agents, there is a need to consider a dose-finding design for Phase I/II trials that can incorporate both toxicity and efficacy outcomes in order to identify the optimal biological dose (OBD), which is defined as the lowest dose with the highest efficacy probability and toxicity probability at or below a desired threshold.

Recently, there has been increasing research into the development of Phase I/II trial designs. One strategy collapses binary outcomes of toxicity and efficacy into a ordinal trinary outcome (no efficacy and no toxicity, efficacy and no toxicity, and toxicity regardless of efficacy), and utilizes either a proportional odds model (Thall and Russell, 1998) or a continuation ratio model (Zhang *et al.*, 2006) to identify the OBD. An alternate approach extends the Continual Reassessment Method (CRM) by joint modeling toxicity and efficacy through copula models with a correlation parameter, and adaptively allocates patients depending on a weighted Euclidean distance (Braun, 2002) or a pre-specified efficacy-toxicity tradeoff (Thall and Cook, 2004). A further approach assumes independence between toxicity and efficacy, models both outcomes separately, and selects the OBD that has the highest efficacy among the set of admissible doses. The efficacy probability estimates can be obtained by a wide variety of possible models, including fully parametric logistic regression with a quadratic term, nonparametric isotonic regression, and other smooth functions of the dose (Cunanan and Koopmeiners, 2014; Zang *et al.*, 2014).

Despite various nonmonotonic efficacy patterns considered, there are few dose-finding designs suited for Phase I/II trials when the OBD is on a plateau of the dose-efficacy curve. Wages and Tait (2015) developed a design that combines features of the CRM for toxicity and order restricted inference for efficacy in the determination of the OBD. Asakawa *et al.* (2014) extended the idea of Bayesian model averaging CRM (BMA-CRM), initially introduced by Yin and Yuan (2009b) for Phase I trials, to accommodate the bivariate binary outcomes of toxicity and efficacy. Both of these designs attempt to address misspecification in the pattern of efficacy with dose by incorporating multiple sets of prespecified efficacy probabilities (also known as the skeleton). However, the methods differ because Wages and Tait (2015) adopted a Bayesian model selection technique and chose the skeleton with the largest posterior model probability to guide dose assignments, while Asakawa *et al.* (2014) allocated patients to the dose that minimizes the Euclidean distance of the estimated toxicity and efficacy probabilities from 0 and 1, respectively. One issue with this Euclidean distance is that a dose whose toxicity probability is u points higher may be equivalent to another dose whose efficacy probability is u points lower. For example, when comparing two doses, one with estimated toxicity and efficacy probabilities of 0.2 and 0.7, respectively, and the other with probabilities of 0.3 and 0.8, respectively, both doses are equivalent estimates of the OBD by this criterion, whereas the latter dose would be preferred if 0.3 is an acceptable rate of toxicity.

In this manuscript, we propose a design that would be appropriate for identification of the OBD when the dose-efficacy curve plateaus within the range of therapeutic interest. Specifically, we take into account different sets of skeletons for the dose-efficacy relationship, each representing a prior guess of where the efficacy probability starts to plateau. The decision of dose assignment is made adaptively through the computation of a posterior selection probability that is applied to the set of doses whose toxicity probabilities are bounded by a maximum acceptable rate. In Section 3.2, we describe the dose-finding algorithm of our proposed design, which is implemented using both Bayesian and maximum likelihood methods. In Section 3.3, we conduct simulation studies to examine the operating characteristics of our proposed design and compare it to those of Wages and Tait (2015). We conclude with a brief discussion in Section 3.4.

3.2 Methods

Consider a trial investigating J discrete dose levels of a targeted agent $\{d_j : j = 1, ..., J\}$. Let T_j and E_j denote the binary indicators of toxicity and efficacy for patients treated at dose d_j. The marginal probabilities of toxicity and efficacy at d_j are $\pi_{Tj} = Pr(T_j = 1)$ and $\pi_{Ej} = Pr(E_j = 1)$, respectively. We consider an independence model in which the joint probability of toxicity and efficacy is the product of their marginal probabilities because Cunanan and Koopmeiners (2014) demonstrated that dose-selection is relatively unaffected if the joint distribution is modeled through the use of a copula, and the correlation parameter is estimated. We assume the probability of toxicity is strictly increasing with dose escalation ($\pi_{Tj} < \pi_{Tj'}$ for $j < j'$), while the probability of efficacy increases at low dose levels and plateaus at higher dose levels ($\pi_{Ej} \leq \pi_{Ej'}$ for $j < j'$). Provided that the maximum number of patients to be enrolled in the trial is N, we now describe our dose-finding design for identifying the OBD, which is the lowest dose with the largest efficacy that maintains a toxicity probability no greater than a pre-specified maximum acceptable rate Γ.

3.2.1 Bayesian Approach

Modeling Probability of Toxicity

Following the CRM proposed by O'Quigley *et al.* (1990), we model the true toxicity probability at dose d_j by the power model, $\pi_{Tj}(\beta) = p_j^{\exp(\beta)}$, which has been suggested to work well in practice (Paoletti and Kramar, 2009). Here $(p_1, ..., p_J)$ is the set of elicited toxicity probabilities (skeleton) for all dose levels under investigation, and β is the unknown parameter that links the skeleton to the true toxicity probabilities. Suppose that among N_j patients treated at dose d_j, T_j and E_j patients have experienced toxicity and efficacy, respectively. Let $\mathscr{D} = \{(N_j, T_j, E_j), j = 1, ..., J\}$ denote the observed data. The marginal likelihood for β is

$$L(\beta|\mathscr{D}) \propto \prod_{j=1}^{J} \{\pi_{Tj}(\beta)\}^{T_j}\{1 - \pi_{Tj}(\beta)\}^{N_j - T_j}. \tag{3.1}$$

Using Bayes' theorem, dose toxicity probabilities can be estimated based upon their posterior means

$$\hat{\pi}_{Tj} = \int p_j^{\exp(\beta)} \frac{L(\beta|\mathscr{D})f(\beta)}{\int L(\beta|\mathscr{D})f(\beta)d\beta} d\beta, \tag{3.2}$$

where $f(\beta)$ is the prior density for the parameter β. An approximate 95% credible interval of the toxicity probability for dose d_j can be constructed by

$$[LB(\hat{\pi}_{Tj}), UB(\hat{\pi}_{Tj})] = [p_j^{\exp\left(\hat{\beta} - Z_{.975}\sqrt{Var(\hat{\beta})}\right)}, p_j^{\exp\left(\hat{\beta} + Z_{.975}\sqrt{Var(\hat{\beta})}\right)}], \tag{3.3}$$

where $Z_{.975}$ is the 97.5^{th} percentile of a standard normal distribution, and $\hat{\beta}$ and $Var(\hat{\beta})$ are the posterior mean and variance of β, respectively. This interval will be used in dose allocation as described shortly.

Moreover, since the OBD has a prespecified maximum acceptable toxicity probability, we can also compute for each dose level the posterior probability of the toxicity rate being no greater than Γ as:

$$P(\pi_{Tj} \leq \Gamma|\mathscr{D}) = \int I\{p_j^{\exp(\beta)} \leq \Gamma\} \frac{L(\beta|\mathscr{D})f(\beta)}{\int L(\beta|\mathscr{D})f(\beta)d\beta} d\beta, \tag{3.4}$$

where $I\{.\}$ is an indicator function of the dose toxicity probability. The estimated toxicity probability $\hat{\pi}_{Tj}$ and posterior probability $P(\pi_{Tj} \leq \Gamma|\mathscr{D})$ at each dose level are updated as data from each successive patient become available. Note that the posterior distribution of β given $\mathscr{D}$ does not have a closed form expression. Thus, the integration in (3.2) and (3.4) is approximated numerically using adaptive quadrature via the *integrate* function in the R package 'stats' (Piessens *et al.*, 1983).

Modeling Probability of Efficacy

Unlike the CRM for strictly increasing dose-toxicity curves, characterization of a non-monotonic dose-efficacy relationship depends heavily on the selection of an appropriate skeleton. If elicited efficacy probabilities in the skeleton do not fit the underlying true dose-efficacy relationship that plateaus at a certain dose level, the ability of the CRM in identifying the OBD using a one-parameter model may be severely compromised (Ivanova and Xiao, 2013). To address the potential problem of misspecification of the skeleton, we adopt the idea of BMA-CRM (Yin and Yuan, 2009b) by enumerating $K = J$ possible plateau skeletons (which we refer to as working models) and obtaining the efficacy probability estimates as weighted averages of the posterior means across all models considered.

Let M_k $(k = 1, ..., K)$ be the working model for the dose-efficacy relationship whose skeleton $(q_{k1}, ..., q_{kJ})$ reaches the plateau at dose d_k and remains constant at higher dose levels. Similar to toxicity, we use a power model representing the dose-efficacy relationship, $\pi^k_{Ej} = q_{kj}^{\exp(\theta_k)}$. According to Bayes' theorem, the posterior model probability for M_k given $\mathscr{D}$ is computed by

$$P(M_k|\mathscr{D}) = \frac{P(M_k) \cdot \int L(\theta_k|\mathscr{D}, M_k) f(\theta_k|M_k) d\theta_k}{\sum_{l=1}^{K} P(M_l) \cdot \int L(\theta_l|\mathscr{D}, M_l) f(\theta_l|M_l) d\theta_l}, \tag{3.5}$$

where $P(M_k)$ is the prior probability that working model M_k is the true efficacy model, $f(\theta_k|M_k)$ is the prior density of θ_k in model M_k, and the likelihood function under model M_k is given by

$$L(\theta_k|\mathscr{D}, M_k) \propto \prod_{j=1}^{J} \{\pi^k_{Ej}\}^{E_j} \{1 - \pi^k_{Ej}\}^{N_j - E_j}. \tag{3.6}$$

In this framework, the BMA estimate of the efficacy probability for dose d_j can be obtained by averaging posterior means of the model-specific efficacy probability $\hat{\pi}^k_{Ej}$, weighted by the posterior model probability,

$$\bar{\pi}_{Ej} = \sum_{k=1}^{K} \hat{\pi}_{Ej}^{k} P(M_k|\mathscr{D}). \tag{3.7}$$

An equal prior probability $P(M_k) = 1/K$ for different skeletons is often used if information regarding the preference of each skeleton is not prespecified. Compared to the CRM that depends upon a single model while ignoring other plausible models, adoption of the BMA-CRM is favored in that it takes into account all possible locations of the efficacy plateau, which will directly influence the recommendation of the OBD, and provides well-calibrated estimates by adaptively downweighing poorly fitted skeletons as more data are accumulated in the trial (Yin and Yuan, 2009b).

Dose Assignment Algorithm

In the practical use of our proposed design, we wish to utilize the toxicity and efficacy outcomes collected during the trial to guide the dose assignment sequentially, and to choose from a set of dose levels that are (1) safe and most efficacious if the efficacy probability increases monotonically with dose escalation, or (2) lowest on the efficacy plateau while maintaining safety if the efficacy probability stabilizes at a certain dose level. To that end, we construct a compound measure that estimates the posterior probability of the current dose being the OBD under various shapes of the efficacy plateau, and recommend the dose with the maximum value for the patients enrolled next.

Stage 1: adaptive randomization There is a concern with the CRM, and adaptive designs in general, that little data exist at the beginning of the trial and the prior distribution for the model parameter exhibits strong influence on parameter estimates (Mick and Ratain, 1993), leading to few dose levels being assigned to patients. In order to promote exploration and gain information for potentially promising dose levels, we introduce the idea of a first stage in which patients are assigned to doses through adaptive randomization

(Huang *et al.*, 2007; Pan *et al.*, 2014; Wages and Tait, 2015).

Specifically, we first define a set of acceptable doses by excluding overly toxic doses if their updated 95% lower credible limits of the toxicity probability exceed the maximum acceptable rate, $\mathscr{A} = \{d_j : LB(\hat{\pi}_{Tj}) \leq \Gamma\}$. We then assign the next patient to one of the acceptable doses based on a scaled randomization, with randomization probabilities equal to

$$R_j = \frac{\bar{\pi}_{Ej}}{\sum\limits_{d_j \in \mathscr{A}} \bar{\pi}_{Ej}}. \tag{3.8}$$

Unlike Wages and Tait (2015), who define acceptable doses using a clear cutoff for the estimated toxicity probability, we allow any dose with a estimated probability of toxicity marginally above Γ to be selected if its BMA estimate of the efficacy probability is sufficiently large, a modification that further varies dose assignments among patients. Given that adaptive randomization expedites information accumulation on different dose levels early in the trial, we allocate one-quarter the total sample size to the randomization, $N_1 = N/4$, as suggested by Wages and Tait (2015).

Stage 2: adaptive selection Once that sufficient data have been collected in Stage 1, we proceed to assign patients directly to the dose that the data suggest is most likely to be the OBD. To evaluate the probability that dose d_j is the OBD given observed data $\mathscr{D}$, we compute the following metric as an average of selection under all efficacy skeletons considered weighted by their posterior model probabilities $P(M_k|\mathscr{D})$, for $j = 2, 3, ..., J - 1$,

$$p_j^{OBD} = P(\pi_{Tj} \leq \Gamma < \pi_{Tj+1}|\mathscr{D}) \sum_{k=j}^{K} P(M_k|\mathscr{D}) + P(\pi_{Tj+1} \leq \Gamma|\mathscr{D}) P(M_j|\mathscr{D}) \tag{3.9}$$

where $P(\pi_{Tj} \leq \Gamma < \pi_{Tj+1}|\mathscr{D})$ is the posterior probability that the maximum acceptable toxicity Γ is bounded by the toxicity rates of the current dose d_j and the next higher dose d_{j+1}. For example, suppose we are investigating four dose levels ($J = 4$) of one agent in a hypothetical trial with $\Gamma = 0.3$. As illustrated in Figure 3.1, the posterior probability of selecting dose d_2 takes into account two possibilities: (a) when dose levels higher than d_2 are overly toxic ($\hat{\pi}_{Tj} > \Gamma, j = 3, 4$) and the dose level lower than d_2 is less effective ($\hat{\pi}^k_{E1} < \hat{\pi}^k_{E2}$), such as under skeletons M_2, M_3, and M_4 ($k = 2, 3, 4$); or (b) when the efficacy skeleton levels off at dose d_2 ($\hat{\pi}^k_{E2} = \hat{\pi}^k_{E3} = \hat{\pi}^k_{E4}, k = 2$) and at least two dose levels on the efficacy plateau are considered to be safe ($\hat{\pi}_{Tj} \leq \Gamma, j = 3$ or 4). Dose d_2 would not be selected as the OBD if the efficacy skeleton starts to plateau at d_1, regardless of the dose-toxicity relationship as shown in Figure 3.1 (c). Due to the boundary conditions, the posterior selection probabilities for the lowest and highest doses are

$$\begin{aligned} p_1^{OBD} &= P(\pi_{T2} > \Gamma|\mathscr{D}) + P(\pi_{T2} \leq \Gamma|\mathscr{D})P(M_1|\mathscr{D}) \\ p_J^{OBD} &= P(\pi_{TJ} \leq \Gamma|\mathscr{D})P(M_J|\mathscr{D}) \end{aligned}$$

Trial Conduct

The conduct of the trial is outlined as follows:

1. Start the trial by assigning the first patient to the lowest dose level d_1.

2. For patient $i = 2, ..., N_1$ in Stage 1, we obtain the set of acceptable doses $\mathscr{A}$ on the basis of available information, compute the BMA-CRM estimated efficacy probabilities $\bar{\pi}_{Ej}$, and randomize the next patient $i + 1$ to dose d_j with probability R_j as defined in Equation (3.8). Note that a patient cannot be assigned to an untried dose if that dose is more than one dose level above the dose of the previous patient, the so-called “no dose skipping” rule.

3. For patient $i = N_1+1, ..., N$ in Stage 2, we update the estimated toxicity and efficacy

probabilities $\hat{\pi}_{Tj}$ and $\bar{\pi}_{Ej}$ using Equations (3.2) and (3.7) given accumulated data, evaluate p_j^{OBD} for each dose level using Equation (3.9), and allocate the next patient $i+1$ to the dose with the largest value of p_j^{OBD}, subject to the "no dose skipping" rule.

4. At any point in the trial, terminate the study early for safety if the lower bound of the 95% credible interval in Equation (3.3) for the lowest dose exceeds the maximum acceptable toxicity rate, $LB(\hat{\pi}_{T1}) > \Gamma$.

5. Identify the OBD as the one with the largest posterior selection probability p_j^{OBD} after observing all data for the maximum sample size of N patients.

3.2.2 Maximum Likelihood Approach

Just as there exists a maximum likelihood version of the CRM (O'Quigley and Shen, 1996), we present a maximum likelihood version of our design that can be used as an alternative to the Bayesian approach described in Section 3.2.1. We apply the same models and skeletons for both toxicity and efficacy outcomes as in the Bayesian approach. Instead of numerical integration for the posterior means, estimates of toxicity and efficacy probabilities in Equations (3.2) and (3.7) can be replaced by their corresponding maximum likelihood estimates (MLEs),

$$\begin{aligned}\tilde{\pi}_{Tj} &= \text{arg}\max_{\beta}\ \text{log}L(\beta|\mathscr{D})\\ \tilde{\pi}^{k}_{Ej} &= \text{arg}\max_{\theta_k}\ \text{log}L(\theta_k|\mathscr{D}, M_k)\end{aligned}$$

While approximate variances and confidence intervals for model parameters β and $\theta_k, k = 1, ..., K$, using the observed information matrix are possible, we calculate them by the posterior variances of model parameters with a dispersed (variance=500) normal prior (Cheung, 2014).

The posterior model probability under the BMA-CRM for the efficacy outcomes can be linked with its likelihood function via the Bayes' factor, a Bayesian equivalent of the likelihood ratio test that compares model M_k with M_1 (Hoeting *et al.*, 1999),

$$BF_{k,1} = \frac{P(\mathscr{D}|M_k)}{P(\mathscr{D}|M_1)} \tag{3.10}$$

where $P(\mathscr{D}|M_k) = \int L(\theta_k|\mathscr{D}, M_k) f(\theta_k|M_k) d\theta_k$ is the marginal likelihood of model M_k integrated over parameter θ_k. If there is no preference over any efficacy working model, i.e. $P(M_1) = P(M_k), k = 2, ..., K$, the posterior model probability can be derived as

$$P(M_k|\mathscr{D}) = \frac{P(\mathscr{D}|M_k)}{\sum_{l=1}^{K} P(\mathscr{D}|M_l)} = \frac{BF_{k,1}}{\sum_{l=1}^{K} BF_{l,1}}. \tag{3.11}$$

Through an approximation of the Bayes' factor (Hoeting *et al.*, 1999; Jin *et al.*, 2015), given by

$$2 \text{ log } BF_{k,1} \approx 2\{\text{log } L(\tilde{\theta}_k|\mathscr{D}, M_k) - \text{log } L(\tilde{\theta}_1|\mathscr{D}, M_1)\} - (r_k - r_1) \cdot \text{log}N, \tag{3.12}$$

where r_k is the dimension of parameter θ_k, we can approximate the posterior model probability as

$$P(M_k|\mathscr{D}) \approx \frac{L(\tilde{\theta}_k|\mathscr{D}, M_k)}{\sum_{l=1}^{K} L(\tilde{\theta}_l|\mathscr{D}, M_l)}. \tag{3.13}$$

We note that in the Bayesian approach, prior information is essential to parameter estimation and sequential allocation for patients early in the trial. Without specifying priors, the maximum likelihood approach requires a start-up rule to begin because the MLE dose not exist when we have homogeneity among the outcomes. Therefore, we follow the same dose-finding algorithm described in Section 3.2.1, but apply an additional up-and-down scheme before Step (3) that dose escalation is enforced until the occurrence of the first toxicity and efficacy (possibly in different patients). Meanwhile, in order to control the number of patients treated at toxic doses, we de-escalate dose level if toxicity has been observed earlier than the first efficacy.

3.3 Simulations

3.3.1 Simulation Settings

To investigate the operating characteristics of our proposed design, we considered Phase I/II trials with $J = 6$ discrete dose levels in a variety of plausible settings. The actual dose-toxicity and dose-efficacy relationships are illustrated in Figure 3.2. Scenarios 1 to 11 were previously examined by (Wages and Tait, 2015), representing combinations of four toxicity curves (denoted T1 through T4) and three efficacy curves (denoted E1 through E3). The setting combining T4 and E3 was not considered because all dose levels had equal toxicity and efficacy rates and was not useful for assessing our design. T1 assumes that marginal toxicity increases for the first five doses and stabilizes at a high level, while all dose levels in T4 maintain a minimal toxicity rate and hence are deemed safe. T2 and T3 exhibit two strictly increasing dose-toxicity curves with a steeper increase for T2 than T3. In term of marginal efficacy, E1 reflects a monotonic increasing pattern, E2 assumes a efficacy plateau at higher dose levels, and E3 remains at a relative high efficacy rate regardless of the dose level. In addition, Scenario 12 is an example that efficacy probabilities level off within the range of safe doses. We specify the maximum acceptable toxicity rate $\Gamma = 1/3$ in Scenarios 1 to 11 and $\Gamma = 0.2$ in Scenario 12. Binary outcomes of toxicity and efficacy were generated independently, because their true association parameter is commonly unknown and there is numerical evidence that a dose-finding algorithm assuming independence between toxicity and efficacy is robust to the existence of correlation (Cunanan and Koopmeiners, 2014).

We compared our design to the approach of (Wages and Tait, 2015), denoted as WT15, which is a competing method developed to address nonmonotonic dose-efficacy relationship. According to the sample size distribution in WT15, we set the maximum sample size in our simulations to be $N = 64$, of which the first $N_1 = 16$ patients were subject to adaptive randomization in Stage 1 and the remaining 48 patients were a part of Stage 2. We chose the toxicity skeleton as $p_j = (0.01, 0.08, 0.15, 0.22, 0.29, 0.36)$, and $K = 6$ sets of efficacy skeletons

$$\begin{aligned}
q_{1j} &= (0.6, 0.6, 0.6, 0.6, 0.6, 0.6)\\
q_{2j} &= (0.5, 0.6, 0.6, 0.6, 0.6, 0.6)\\
q_{3j} &= (0.4, 0.5, 0.6, 0.6, 0.6, 0.6)\\
q_{4j} &= (0.3, 0.4, 0.5, 0.6, 0.6, 0.6)\\
q_{5j} &= (0.2, 0.3, 0.4, 0.5, 0.6, 0.6)\\
q_{6j} &= (0.1, 0.2, 0.3, 0.4, 0.5, 0.6)
\end{aligned} \tag{3.14}$$

which are expected to represent different prior opinions on where the true dose-efficacy curve starts to plateau. We placed a normal prior with mean 0 and variance 1.34 on the model parameters β and θ_k (O'Quigley and Shen, 1996). We assigned a prior model probability of $P(M_k) = 1/6$ to each efficacy skeleton. We used priors, skeleton rates and stopping rules as suggested in Wages and Tait (2015) when implementing their design.

Under each scenario and each design, we simulated 2,000 hypothetical trials and quantified the performance of each design by four metrics: (1) the average percentage of patients that were assigned to each dose level, (2) the average percentage of observed toxicity across simulations, (3) the average percentage of observed efficacy across simulations, and (4) the proportion of simulations in which each dose level was identified as the OBD at the end of the trial.

3.3.2 Simulation Results

For direct comparison among designs, we adopted an accuracy index (AI) that can summarize dose selection among all scenarios (Cheung, 2011):

$$AI = 1 - J \cdot \frac{\sum_{j=1}^{J} \rho_j \cdot P(\text{selection of } d_j)}{\sum_{j=1}^{J} \rho_j} \tag{3.15}$$

where ρ_j is a discrepancy measure between dose level d_j and the targeted OBD, defined as the sum of squared differences for the toxicity and efficacy rates, $\rho_j = (\pi_{Tj} - \pi_T^{OBD})^2 + (\pi_{Ej} - \pi_E^{OBD})^2$. Here, π_T^{OBD} and π_E^{OBD} are the actual toxicity and efficacy rates corresponding to the targeted OBD in each simulation scenario. As an index that reflects the accuracy of a design, AI penalizes for overly toxic and ineffective doses, so a larger AI indicates selected doses are concentrated in the neighborhood of the OBD, and the maximum attainable value of AI is 1. Likewise, AI can be used to evaluate the safety of a design by replacing $P(\text{selection of } d_j)$ with $P(\text{allocation of } d_j)$ in Equation (3.15) for the distribution of dose assignment.

The operating characteristics of our proposed design, using both Bayesian and maximum likelihood methods, relative to those of WT15 under various scenarios are presented in Tables 3.1 and 3.2. Overall, our proposed design is quite comparable to WT15 and can yield improved performance in terms of identification of the OBD and allocation of patients when the targeted OBD is on a plateau of the dose-efficacy curve. Our maximum likelihood approach behaves similarly to the Bayesian approach for the sample size we considered, although the Bayesian approach performs slightly better than the maximum likelihood approach.

In Scenarios 1, 4, and 7 where the dose-efficacy relationship (E1) has a monotone increasing pattern, the OBD d_4 is the the highest dose whose toxicity probability stays below $\Gamma = 1/3$. We see our designs are more likely to identify the true OBD than WT15, with an approximate increase of 8-14 points in the selection probability of dose d_4 for the Bayesian approach and an approximate increase of 0-13 points for the maximum likelihood approach. Similarly, both of our designs assign patients more frequently to the OBD than WT15, with a range of 0-8 more patients. When the entire distributions of selection and dose assignments are considered through the AI, our Bayesian approach compares favorably to WT15 with regard to both the accuracy and safety of the design.

Scenarios 2, 5, and 8 represent situations in which the dose-efficacy relationship (E2) increases until dose d_4 and remains constant thereafter, so the lowest dose d_4 on the efficacy plateau that maintains safety is the OBD. Once again, our design correctly identifies the OBD more often than WT15, assigns more patients to dose d_4, and produces larger accuracy and safety summary scores than WT15.

Scenarios 3, 6, and 9 provide some insight into the operating characteristics of examined designs when all dose levels are equally effective (E3) and the lowest dose d_1 is the OBD. We observe that our proposed design tends to select and assign patients more often to dose levels above the OBD when compared to WT15. For example, 45% of simulations selected the true OBD in the design of WT15, as opposed to 25% and 23% in our Bayesian approach and maximum likelihood approach. And the numbers of patients treated at the OBD in our designs ($\sim$ 14-16) are approximately half the size treated in the WT15 ($\sim$ 29-30).

This differential in operating characteristics occurs because selection of the OBD now depends solely on the minimum toxicity probability. However, given variability of the binary outcomes and the limited sample size in the trial, observed data may not necessarily support the underlying flat efficacy skeleton over others, leading to a relatively small posterior model probability $P(M_1|\mathscr{D})$ and hence reduces the selection probability for the true OBD. To address this problem, we investigated the sensitivity of our design to the prior model probabilities in Scenarios 3, 6, and 9. Specifically, instead of giving equal probability to all models, we instead assigned a larger probability to the flat efficacy skeleton, i.e. $P(M_1) = 0.5$, with the remaining models receiving equal probability, i.e. $P(M_2) = \cdots = P(M_6) = 0.1$. Results when using these prior model probabilities are shown in Table 3.3; we found similar selection probability and improved allocation at the

OBD relative to WT15.

All doses in Scenarios 10 and 11 are safe because they all maintain a minimal toxicity probability (T4), so the OBD occurs at the dose with the maximum efficacy probability. In Scenario 10 with a monotone increasing toxicity, our design correctly identified d_6 as the OBD in at least 64% of simulations and d_6 was assigned to at least 57% of patients, which is considerably better than the corresponding values for WT15. Scenario 11 has three OBDs since the efficacy probabilities of doses d_4, d_5 and d_6 are the same. We observe that our design assigns 19% fewer patients at suboptimal doses and demonstrates larger AI scores than WT15. In Scenario 12 where toxicity increases across the range of doses and stays below Γ, while efficacy plateaus at higher doses, the benefits of our designs appear to be substantial, with at least an 11% increase in the selection probability of the OBD (d_4) and 21% increase in the selection of safe and maximum effective doses (d_4, d_5, and d_6) as compared to WT15. A similar gain in terms of dose assignment can be achieved using our proposed design.

Table 3.4 summarizes the average number of patients for whom toxicity or efficacy was observed under various settings. In general, our Bayesian approach and maximum likelihood approach provide toxicity and efficacy rates that are close to those with WT15, providing evidence that our design is as safe and accurate as WT15.

3.4 Discussion

In this work, we have proposed a dose-finding design that can use either a Bayesian approach based upon the idea of BMA-CRM (Yin and Yuan, 2009b) or a corresponding maximum likelihood alternative, for identification of the OBD in a Phase I/II trial when the dose-efficacy curve of the targeted agent reaches a plateau within the dose range of interest. By incorporating multiple efficacy skeletons as working models, our designs are robust to the misspecification of the dose-efficacy curve. By constructing the posterior selection probability, we are able to quantify the uncertainty for each dose being the OBD and specify dose assignment adaptively.

The simulation results demonstrate improved performances of our proposed design in most settings, as it assigns more patients to the OBD and identifies the OBD more often than the design of Wages and Tait (2015). However, in rare situations where the underlying true efficacy probability remains constant across all dose levels so that the lowest dose is the OBD, our design had a reduced likelihood of allocation and selection at the OBD if all dose-efficacy skeletons received equal probability. At this time, sensitivity analysis is needed to calibrate an appropriate weight for the flat efficacy skeleton. Moreover, we are exploring modifications to our proposed designs in the use of dual-agent Phase I/II dose-finding trials, which is an area of current methodological interest.

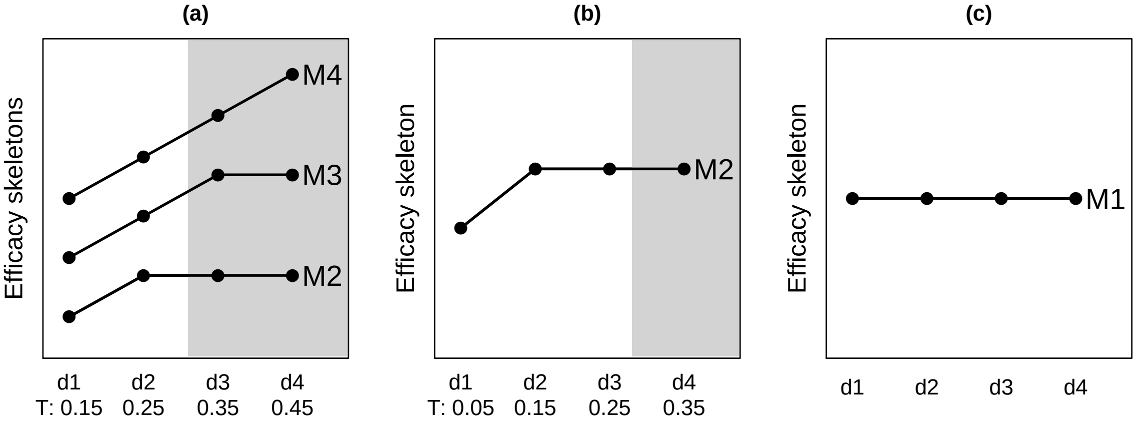

Figure 3.1: Possible situations in which dose d_2 is selected as the OBD in a hypothetical trial with four dose levels. Dose levels highlighted in gray are considered to be overly toxic.

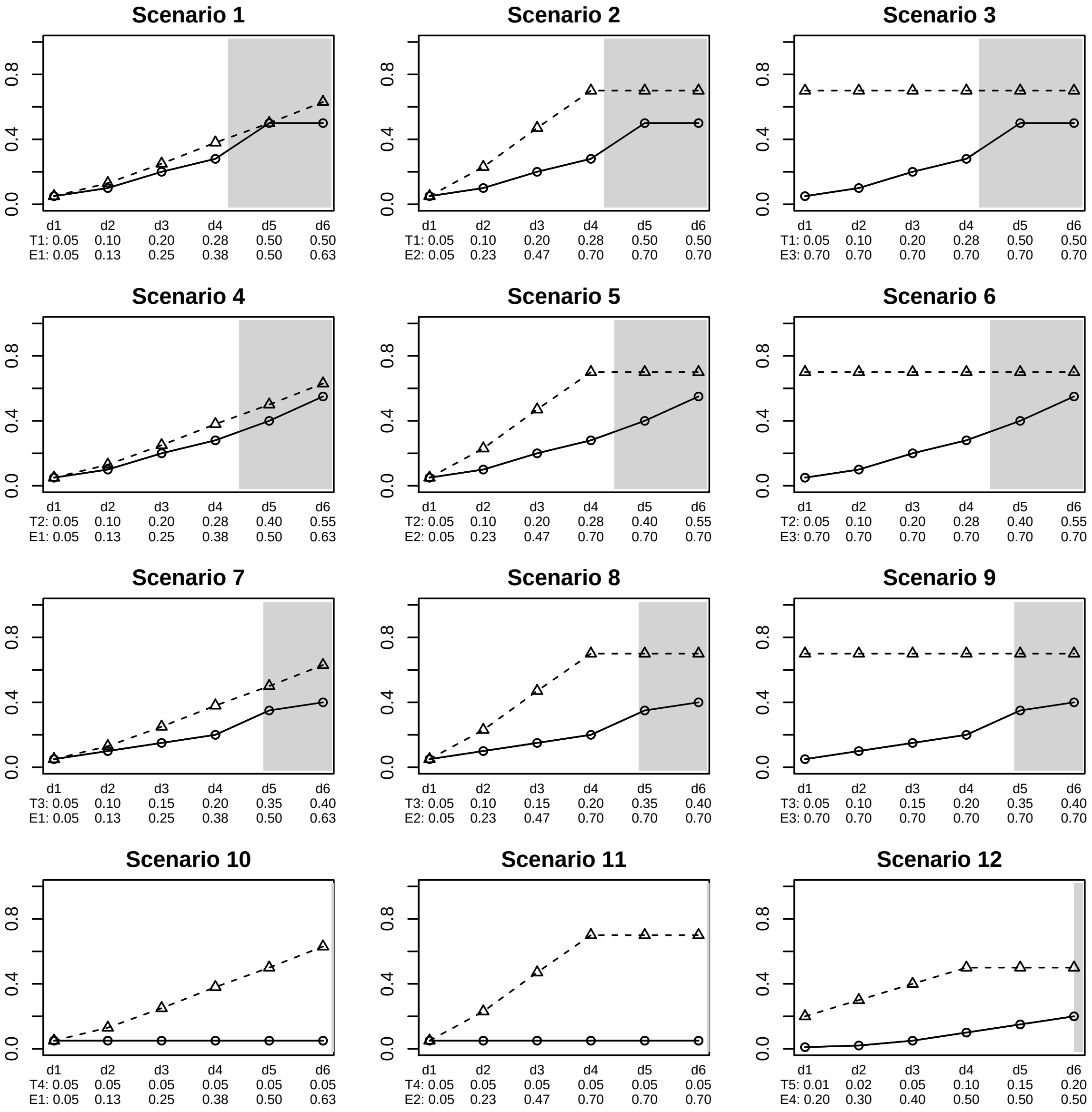

Figure 3.2: Illustration of twelve simulation scenarios. The solid lines with cycles represent does-toxicity relationships and the dashed lines with triangles represent dose-efficacy relationships. The true probabilities of toxicity and efficacy are listed under each scenario. Toxicity probabilities of dose levels highlighted in gray are greater than Γ.

Table 3.1: Summary of dose selection from Scenarios 1 to 12: average probability (%) of selection as the OBD at each dose level, and the accuracy index for the distribution of dose selection. The targeted OBD indicated in bold, and the design with the largest accuracy index underlined.

Scenario	Design	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	Dose 6	Accuracy Index
1	WT15	2.9	9.1	30.6	**51.1**	6.3	0.0	0.677
	Bayesian	0.2	3.0	32.8	**59.0**	5.0	0.0	0.815
	ML	0.8	6.5	37.4	**51.2**	4.2	0.0	0.751
2	WT15	1.0	4.3	32.2	**59.0**	3.3	0.2	0.753
	Bayesian	0.0	1.0	33.1	**64.2**	1.7	0.0	0.844
	ML	0.2	1.8	34.6	**61.6**	1.8	0.0	0.784
3	WT15	**45.6**	33.0	16.8	4.5	0.1	0.0	0.911
	Bayesian	**24.2**	25.4	35.0	14.9	0.4	0.0	0.786
	ML	**22.4**	23.4	34.6	18.8	0.8	0.0	0.752
4	WT15	2.7	8.4	32.1	**45.8**	10.7	0.3	0.685
	Bayesian	0.4	2.8	30.0	**57.6**	9.2	0.0	0.825
	ML	0.6	4.3	35.3	**51.8**	8.0	0.0	0.790
5	WT15	0.5	4.1	30.6	**54.1**	10.6	0.1	0.777
	Bayesian	0.0	0.5	29.9	**66.3**	3.2	0.0	0.867
	ML	0.2	1.9	31.2	**63.2**	3.6	0.0	0.830
6	WT15	**46.1**	32.9	15.7	4.8	0.5	0.0	0.900
	Bayesian	**24.8**	26.1	33.4	15.1	0.6	0.0	0.775
	ML	**22.6**	22.1	34.2	20.0	1.1	0.0	0.731
7	WT15	3.3	7.0	20.3	**39.2**	26.5	3.7	0.555
	Bayesian	0.3	1.5	11.6	**52.8**	31.7	2.2	0.707
	ML	0.5	3.2	14.4	**52.6**	27.6	1.8	0.706
8	WT15	0.3	3.8	14.7	**56.5**	22.7	2.0	0.817
	Bayesian	0.0	0.3	10.0	**72.8**	16.6	0.3	0.924
	ML	0.0	0.8	9.7	**71.3**	17.8	0.4	0.914
9	WT15	**45.2**	31.4	17.2	5.9	0.3	0.0	0.901
	Bayesian	**24.9**	25.4	26.3	19.9	3.5	0.0	0.736
	ML	**24.0**	21.2	24.2	25.8	4.8	0.0	0.683
10	WT15	2.6	5.7	13.8	18.6	24.1	**35.2**	0.566
	Bayesian	0.0	1.0	3.1	8.3	18.0	**69.6**	0.887
	ML	0.2	1.0	4.8	10.2	20.0	**63.8**	0.853
11	WT15	0.5	2.3	9.4	**37.7**	**30.0**	**20.1**	0.895
	Bayesian	0.0	0.2	3.0	**30.4**	**27.1**	**39.4**	0.983
	ML	0.0	0.3	3.6	**28.6**	**24.2**	**43.4**	0.978
12	WT15	8.7	18.6	25.6	**35.2**	10.0	1.9	0.264
	Bayesian	1.8	5.6	20.2	**49.9**	19.7	2.8	0.729
	ML	2.8	9.0	20.7	**46.3**	18.8	2.5	0.638

Table 3.2: Summary of dose assignment from Scenarios 1 to 12: the mean percentage of allocation at each dose level, and the accuracy index for the distribution of dose assignment. The targeted OBD indicated in bold, and the design with the largest accuracy index underlined.

Scenario	Design	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	Dose 6	Accuracy Index
1	WT15	9.9	17.0	29.4	**34.8**	7.6	1.3	0.405
	Bayesian	3.4	10.9	33.0	**39.6**	10.4	2.7	0.563
	ML	4.5	13.3	34.6	**35.3**	9.6	2.8	0.509
2	WT15	8.3	13.9	30.4	**39.9**	6.3	1.2	0.348
	Bayesian	2.5	8.8	34.3	**44.3**	7.9	2.2	0.597
	ML	3.3	9.6	33.8	**42.2**	8.4	2.7	0.556
3	WT15	**42.2**	32.5	18.5	6.0	0.7	0.1	0.878
	Bayesian	**21.3**	27.5	31.2	14.9	3.6	1.4	0.680
	ML	**19.6**	25.4	31.0	17.6	4.8	1.5	0.631
4	WT15	9.1	16.0	30.0	**31.3**	11.7	1.9	0.423
	Bayesian	3.1	9.4	31.1	**39.5**	13.9	3.1	0.603
	ML	3.8	12.0	32.9	**35.7**	12.6	3.0	0.556
5	WT15	7.9	14.0	28.6	**37.4**	10.6	1.5	0.363
	Bayesian	2.6	7.7	32.5	**45.1**	9.5	2.5	0.628
	ML	3.1	9.4	32.1	**43.0**	9.7	2.7	0.582
6	WT15	**42.1**	32.7	17.9	6.2	1.1	0.1	0.868
	Bayesian	**21.0**	28.0	31.2	15.0	3.6	1.2	0.693
	ML	**19.6**	23.7	31.4	18.4	5.3	1.6	0.629
7	WT15	9.5	14.8	22.7	**28.8**	18.7	5.5	0.335
	Bayesian	2.6	6.2	18.3	**40.7**	25.9	6.3	0.545
	ML	3.0	7.8	20.3	**40.2**	23.3	5.4	0.542
8	WT15	7.7	12.4	19.5	**39.6**	17.1	3.8	0.404
	Bayesian	2.3	5.0	18.3	**52.8**	17.5	4.0	0.717
	ML	2.4	5.4	18.2	**52.2**	17.5	4.3	0.706
9	WT15	**41.3**	32.4	18.3	6.7	1.2	0.1	0.870
	Bayesian	**21.3**	24.8	27.1	18.7	6.0	2.1	0.624
	ML	**19.4**	21.2	26.5	22.4	8.1	2.4	0.553
10	WT15	7.7	11.1	16.2	18.1	20.2	**26.7**	0.320
	Bayesian	2.0	3.0	5.6	10.7	17.6	**61.1**	0.763
	ML	2.1	3.4	7.1	12.0	18.7	**56.6**	0.730
11	WT15	6.6	9.1	14.3	**31.3**	**22.3**	**16.4**	0.519
	Bayesian	1.9	2.6	6.1	**24.8**	**22.7**	**42.0**	0.854
	ML	2.0	2.7	6.1	**23.2**	**20.8**	**45.3**	0.848
12	WT15	15.0	21.2	25.4	**25.9**	10.2	2.4	-0.001
	Bayesian	3.8	8.5	23.1	**37.3**	18.2	9.1	0.578
	ML	4.7	11.1	23.2	**35.0**	17.2	8.8	0.505

Table 3.3: Sensitivity analysis to the prior model probability from Scenarios 3, 6, and 9: the mean percentage of dose selection/assignment at each dose level, and the accuracy index for the distribution of dose selection/assignment. The targeted OBD indicated in bold.

Scenario	Design	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	Dose 6	Accuracy Index
				Dose selection				
3	Bayesian	**50.8**	27.3	17.7	4.0	0.1	0.0	0.913
	ML	**41.9**	26.9	23.2	7.8	0.3	0.0	0.868
6	Bayesian	**53.1**	25.8	16.6	4.4	0.1	0.0	0.909
	ML	**42.4**	24.4	24.4	8.4	0.6	0.0	0.850
9	Bayesian	**55.0**	23.2	15.2	5.7	1.0	0.0	0.896
	ML	**44.0**	23.2	22.6	8.7	1.4	0.0	0.853
				Dose assignment				
3	Bayesian	**69.5**	5.0	10.1	9.2	4.7	1.6	0.751
	ML	**68.4**	4.8	10.0	10.1	5.1	1.7	0.733
6	Bayesian	**69.8**	5.0	9.7	9.1	4.7	1.7	0.772
	ML	**68.2**	4.9	9.9	10.1	5.1	1.8	0.754
9	Bayesian	**69.6**	4.3	8.2	9.9	5.5	2.4	0.732
	ML	**68.2**	4.4	8.8	10.1	5.8	2.6	0.717

Table 3.4: The average percentage of toxicity and efficacy from Scenarios 1 to 12.

	Toxicity (%)			Efficacy (%)		
Scenario	WT15	Bayesian	ML	WT15	Bayesian	ML
1	22.3	25.3	24.4	28.0	31.9	30.8
2	22.8	25.2	25.2	50.8	56.2	55.6
3	11.1	16.6	17.8	69.8	69.8	70.0
4	22.5	25.6	24.7	28.9	33.0	31.9
5	23.0	25.0	24.8	51.2	56.9	56.2
6	10.9	16.6	17.7	69.8	70.0	69.8
7	19.7	23.1	22.3	32.0	38.0	36.6
8	19.8	21.6	21.7	54.5	61.9	61.6
9	9.7	14.2	15.3	70.0	70.0	70.2
10	5.0	5.0	5.0	40.0	53.3	52.0
11	5.0	5.0	5.0	57.8	66.2	66.2
12	6.6	9.7	9.2	38.6	44.8	43.9

CHAPTER IV

Exposure Enriched Outcome Dependent Designs for Longitudinal Studies of Gene-Environment Interaction

4.1 Introduction

Joint effects of genetic and environmental factors have been increasingly recognized in the development of many complex human diseases (Hunter, 2005). Investigation of gene-environment (GxE) interaction may not only provide biological insights into the etiology of these diseases, but also assist in the discovery of novel genetic or environmental risk factors (Dai *et al.*, 2012). However, GxE interaction studies are statistically challenging because of the prohibitive sample size requirement in each GxE configuration. The frequency of the risk allele, the distribution of the environmental exposure, and the effect size of the GxE interaction all contribute to the need for larger samples to detect such an interaction with adequate power (Thomas, 2010).

Despite the popularity of case-control and case-only designs, longitudinal cohort studies have long been recommended for GxE interactions because of better characterization of lifetime exposure history (Clayton and McKeigue, 2001), the ability to account for within-subject variability of the outcome, and the potential to delineate the dynamic temporal pattern of the genetic or GxE interaction effect, which is often missed in case-control studies by design. Typically, in such studies, extensive information has been collected over the course of the longitudinal follow-up, including prospectively assessed environmental exposures, repeatedly measured outcomes, and a detailed set of potential confounders. We consider a situation where genetic data are to be collected retrospectively with exposures and outcomes already measured; however, our proposed methods can be easily adapted to the collection of new expensive biomarkers of exposure when genetic data is already available. The ability to obtain both genetic and environmental data for all subjects in a large cohort under the budget constraint is often challenging due to cost. To focus the limited resources on informative subjects, there is a need to apply a principled strategy that prioritizes subjects for genotyping or exposure assay. This idea is akin to the two-phase sampling design that is commonly used for case-control studies, and we aim to extend it to longitudinal cohorts. Such a sampling design is also relevant when constructing an informative subsample using existing electronic health record (EHR) data to check a hypothesis on the interplay between genes and environmental exposures/biomarkers.

As highlighted in a recent discussion by Kraft and Aschard (2015), the small number of replicated GxE interactions in observational studies could be attributed to the lack of exposure variability in standard designs. There has been recommendations for exposure enriched sampling in case-control studies with binary exposure. For example, Ahn *et al.* (2013) developed a disease-exposure stratified sampling accompanied by a Bayesian analysis framework, Chen *et al.* (2012) explored several two-phase designs conditional on the exposure and case-control status, and Stenzel *et al.* (2015) evaluated the impacts of exposure enriched sampling designs and exposure measurement error on the power for tests of GxE interaction. All of them concluded with a consensus that an enriched selection of exposed subjects leads to improved power for GxE interactions, as long as exposure measurement error is not severe. Similarly, in cross-sectional studies with continuous exposure and outcome, a substantial reduction in the required number of subjects is achieved by selecting subjects with extreme exposure levels (Boks *et al.*, 2007).

In addition to exposure variability, it is also important to consider temporal variation in outcomes when constructing an informative subsample in a longitudinal study. For instance, Schildcrout and Heagerty (Schildcrout and Heagerty, 2008) introduced stratified sampling conditional on a binary response series. Schildcrout *et al.* (2012) proposed auxiliary variable dependent sampling when an inexpensive auxiliary variable related to the longitudinal binary response is available for repeated measures. For longitudinal continuous outcomes, Schildcrout *et al.* (2013) developed outcome trajectory dependent sampling that stratifies subjects by the summary measures of the individual outcome vector. In their work, a genetic main effect and a gene-by-time interaction effect were assessed without specific consideration of an environmental exposure. Improved efficiency of estimated coefficients were observed when sampling on a summary measure that is related to the targeted parameters.

To date, literature on sampling designs for longitudinal studies of GxE interaction is quite limited. In this work, we consider designs to select a subsample for genotyping or exposure assay on the basis of available data in an existing cohort/database. Specifically, we propose variants of two-phase design for longitudinal outcomes. We are interested in the GxE interaction, their joint effects, and potentially the time-varying GxE (GxExT) interaction.

Under the two-phase design, standard maximum likelihood analysis ignoring the sampling mechanism leads to biased estimates (Holt *et al.*, 1980). To correct for the biased design, some approaches consider only a subsample of individuals with a complete set of information on the exposure, genotype, outcome, and other relevant covariates, and make an adjustment using a weighted likelihood or a conditional likelihood (Robins *et al.*, 1994; Schildcrout and Heagerty, 2008; Schildcrout *et al.*, 2013), while others treat it as a missing data problem. For example, partial information on subjects whose genetic or environmental data are missing by design can be incorporated into the analysis, through an underlying distribution of the missing covariate, either estimated empirically or modeled parametrically (Lawless *et al.*, 1999; Weaver and Zhou, 2005). Multiple imputation strategies have been suggested for handling the selection bias (Schildcrout *et al.*, 2015). Furthermore, a full Bayesian analysis based on the joint likelihood of the entire cohort has been proposed (Ahn *et al.*, 2013). In order to improve estimation efficiency, we develop a conditional likelihood-based approach using data available from both phases in conjunction with our proposed designs, and investigate their statistical properties relative to the existing approaches.

We illustrate our methods using data from the Normative Aging Study (NAS), an ongoing longitudinal study of aging initiated by the Veterans Administration in 1963. In this study, subjects who underwent bone lead measurement between 1991 and 2002 were followed up for their blood pressure levels every three years. It has been documented in the existing NAS cohort that lead exposure was associated with increased pulse pressure (Perlstein *et al.*, 2007), a marker of arterial stiffness, and this association becomes stronger in subjects who are carriers of the risk alleles of the hemochromatosis (*HFE*) gene (Zhang *et al.*, 2010). We use this example to demonstrate the benefits of exposure enriched outcome trajectory dependent sampling for a study of GxE interaction when a quantitative trait in a longitudinal study is of interest.

The rest of the chapter is organized as follows. In Section 4.2, we describe five sampling designs for longitudinal studies of GxE interaction. Section 4.3 provides four likelihood approaches that can be utilized for parameter estimation and statistical inference. In Sections 4.4 and 4.5, we perform simulation studies and use the NAS example to evaluate the operating characteristics of different sampling designs and likelihood approaches.

Section 4.6 concludes with a summary of our findings and discussions.

4.2 Sampling Designs

4.2.1 Notation

Our study objective is to detect and quantify the joint (both main and interaction) effects of genetic and environmental factors on a continuous trait with repeated measurements in a longitudinal study. Let Y_{ij} denote the outcome for subject i measured at the jth follow-up for $i = 1, \ldots, N$ and $j = 1, \ldots, r_i$. Subject-specific design matrices of covariates for fixed and random effects are denoted by X_i and Z_i respectively. We characterize the response trajectory $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ir_i})'$ via a linear mixed effects model

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{4.1}$$

where β is a vector of fixed effects, b_i are subject-specific random effects, ϵ_i are measurement errors assumed to be normally distributed with mean zero and covariance matrix $R_i = \sigma_e^2 I_{r_i}$, and I_{r_i} is the r_i-dimensional identity matrix.

In this longitudinal setting, the subject-specific design matrix for fixed effects X_i is typically composed of a collection of confounding factors V_i, a variable representing time T_i, a baseline environmental exposure E_i (binary or continuous), a retrospectively collected genotype G_i that indicates the presence of the minor allele for a single nucleotide polymorphism (0 = no copy, 1 = at least one copy), and a GxE interaction term G_iE_i, e.g., $X_i = (V_i, T_i, E_i, G_i, G_iE_i)$. Additional interactions between time and exposure- or genotype-related covariates, such as E_iT_i, G_iT_i and $G_iE_iT_i$, could be included given evidence of significance or scientific justification. While many environmental exposures can change over time, we focus on baseline exposure in this present study. Extension to time-varying exposure is mentioned in the discussion.

Under this linear mixed effects model, we consider the subject-specific design matrix $Z_i = (1, T_i)$ for the random effects $b_i = (b_{0i}, b_{1i})'$, which are assumed to follow a bivariate normal distribution with mean zero and covariance matrix D that contains variance components σ_0^2, σ_1^2, and a correlation coefficient $\rho = corr(b_{0i}, b_{1i})$. Integrating over the random effects, the marginal distribution of the outcome follows a multivariate normal distribution with mean vector $\mu_i = X_i\beta$ and covariance matrix $\Sigma_i = Z_i D Z_i' + R_i$, i.e. $Y_i|X_i \sim (\mu_i, \Sigma_i)$. Statistical inference for fixed effects can be made by maximizing the marginal likelihood function:

(4.2)

$$L(\beta, \sigma) = \prod_{i=1}^{N} f(Y_i|X_i;\beta,\sigma) = \prod_{i=1}^{N} (2\pi)^{-\frac{r_i}{2}} \cdot |\Sigma_i|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}(Y_i - X_i\beta)'\Sigma_i{}^{-1}(Y_i - X_i\beta)\right\}$$

where $f(Y_i|X_i;\beta,\sigma)$ is the multivariate density of Y_i given X_i, and covariance matrix Σ_i with parameters $\sigma = (\sigma_0, \sigma_1, \rho, \sigma_e)$.

Suppose (Y_i, X_i^*), $X_i^* = (V_i, T_i, E_i)$, $i = 1, \ldots, N$, are collected for the entire cohort in an initial phase (Phase I), followed by a selection of the cohort with an expected sample size n $(n < N)$ for retrospective genotyping G_i in the second phase (Phase II). Let $S_i = 1$ $(S_i = 0)$ denote the inclusion (exclusion) of subject i in Phase II. For example, one might assign a constant selection probability $P(S_i = 1) = n/N$ to all subjects and draw samples via independent Bernoulli trials. This sampling scheme renders a missing completely at random pattern for G_i, so the standard maximum likelihood estimation should suffice and be regarded as a baseline for comparison. To investigate how a two-phase design can improve the efficiency of a longitudinal study of GxE interaction, we now describe five sampling schemes that take advantage of observed information in Phase I to guide the sample selection in Phase II.

4.2.2 Design 1: Exposure Stratified Sampling

Subjects observed in Phase I are partitioned by their environmental exposures into K mutually exclusive strata R^k, where $k = 1, ..., K$. That is, exposure is the only sampling variable, namely, $Q_i = E_i$, regardless of the outcome. Within each stratum, individuals are selected with a pre-specified stratum-specific probability $\pi(R^k) = P(S_i = 1|Q_i \in R^k) = n_k/N_k$, where N_k is the number of subjects falling into stratum R^k, and n_k is the expected number of subjects sampled in Phase II from R^k. Larger selection probabilities are allocated to the strata that are most informative. For binary exposure (stratum $k = E$ for exposed subjects, and $k = \bar{E}$ for unexposed), subjects with rare exposure are enriched to achieve a certain proportion $\lambda = n_E/n$. Note that although a balanced design ($\lambda = 0.5$) with a equal number of exposed and unexposed subjects is desired, the maximum level of enrichment is limited by the prevalence of exposure in the cohort and the overall sampling probability, $\lambda \leq P(E = 1) \cdot (n/N)^{-1}$. For continuous exposure, subjects are stratified into R^k ($k = 1, 2, 3$), where $R^1 = \{E_i \leq C_e^1\}$, $R^2 = \{C_e^1 < E_i \leq C_e^2\}$, and $R^3 = \{E_i > C_e^2\}$. For instance, to draw a sample of $n = 250$ subjects from the original cohort of $N = 1000$, one can choose cutpoints C_e^1 and C_e^2 as the 10^{th} and 90^{th} percentiles of the exposure distribution. Subjects from the two tails (strata R^1 and R^3) are preferentially sampled with a probability of 1.0, and a random sample from stratum R^2 is drawn to reach the genotyping capacity.

4.2.3 Design 2: Outcome Trajectory Dependent Sampling

To capture variation in individual outcome trajectories over time, Schildcrout *et al.* (2013) proposed a sampling design that is based upon summary statistics of the individual outcome vector, which we refer to as outcome trajectory dependent sampling in this paper. They specify the sampling variable Q_i as estimated effects from simple linear regression of the outcome for subject i on a single predictor time: $E[Y_{ij}] = \eta_{0i} + \eta_{1i}T_{ij} = Z_i\eta_i$, $i = 1, ..., N$, $j = 1, ..., r_i$, and $\eta_i = (\eta_{0i}, \eta_{1i})'$. For example, a univariate Q_i considers either estimated intercept $Q_i = \widehat{\eta}_{0i}$ or estimated slope of time $Q_i = \widehat{\eta}_{1i}$, while a bivariate $Q_i = \widehat{\eta}_i$ considers both jointly.

In the case that $Q_i = \widehat{\eta}_{0i}$, subjects in Phase I are partitioned into three strata: $R^1 = \{\widehat{\eta}_{0i} \in (-\infty, C^1]\}$, $R^2 = \{\widehat{\eta}_{0i} \in (C^1, C^2]\}$, and $R^3 = \{\widehat{\eta}_{0i} \in (C^2, +\infty)\}$, where strata R^1 and R^3 represent two tails of the sampling distribution and cutpoints (C^1, C^2) are determined by the percentiles of the empirical distribution of Q. When sampling from the bivariate Q_i, subjects lying in the center of the sampling distribution are stratified into $R^2 = \{(\widehat{\eta}_{0i}, \widehat{\eta}_{1i}) : C_0^1 < \widehat{\eta}_{0i} \leq C_0^2, C_1^1 < \widehat{\eta}_{1i} \leq C_1^2\}$, while others into $R^1 = \{(\widehat{\eta}_{0i}, \widehat{\eta}_{1i}) \notin R^2\}$. Cutpoints $(C_0^1, C_0^2, C_1^1, C_1^2)$ are determined by grid search of the empirical bivariate distribution of Q, ensuring the fraction of subjects fall into the central stratum R^2 using bivariate Q_i is the same as the fraction of R^2 using univariate Q_i. To capture a larger variability in the outcome trajectory, subjects from both strata R^1 and R^3 are sampled with probability 1.0.

An important property of this design is that Q_i is a linear combination of the individual outcome Y_i, i.e., $Q_i = \widehat{\eta}_i = W_iY_i$, where the weight $W_i = (Z_i'Z_i)^{-1}Z_i'$. Note that subjects with only one observation ($r_i = 1$) do not have OLS estimate for the intercept or slope of time due to the singular square matrix $Z_i'Z_i$. In this case, we assign $\widehat{\eta}_{0i} = y_{i1}$ and $\widehat{\eta}_{1i} = 0$. Given the marginal distribution assumption for $Y_i|X_i$ in the linear mixed model in (4.1), the sampling variable should also follow a normal distribution $Q_i|X_i \sim (\mu_{q_i} = W_i\mu_i, \Sigma_{q_i} = W_i\Sigma_iW_i')$. As such, the probability of subject i being sampled in Phase II given X_i, for a univariate Q_i, can be obtained:

(4.3)

$$P(S_i = 1|X_i) = \sum_{k=1}^{K} P(S_i = 1, Q_i \in R^k|X_i) = \sum_{k=1}^{K} \pi(R^k)\{F_{Q_i|X_i}(C^k) - F_{Q_i|X_i}(C^{k-1})\}$$

where sample selection given $Q_i \in R^k$ is assumed to be independent of $X_i | Q_i \in R^k$ so that $P(S_i = 1 | Q_i \in R^k, X_i) = P(S_i = 1 | Q_i \in R^k) = \pi(R^k)$, and $F_{Q_i|X_i}(\cdot)$ is the cumulative distribution function of $Q_i | X_i$. Under an outcome dependent design where missingness in G_i renders a missing at random mechanism, a closed form expression of this sampling probability makes it possible to calculate the conditional likelihood that corrects for the biased sampling.

4.2.4 Design 3: Exposure Enriched Outcome Trajectory Dependent Sampling

In Design 3, we combine the strategies of Design 1 and Design 2, and specify the sampling variable as a bivariate $Q_i = (\widehat{\eta}_{0i}, E_i)'$, $Q_i = (\widehat{\eta}_{1i}, E_i)'$, or a multivariate $Q_i = (\widehat{\eta}_{0i}, \widehat{\eta}_{1i}, E_i)'$. For binary exposure, we partition subjects into six strata: $R^{k,E} = \{\widehat{\eta}_{0i} \in (C^{k-1,E}, C^{k,E}]\}$ for exposed subjects and $R^{k,\bar{E}} = \{\widehat{\eta}_{0i} \in (C^{k-1,\bar{E}}, C^{k,\bar{E}}]\}$ for unexposed, $k = 1, 2, 3$, if the rate of change of the outcome is not considered in Q_i. Cutpoints $C^{k,E}$ and $C^{k,\bar{E}}$ are determined by the percentiles of the empirical distribution of Q. We set stratum-specific selection probabilities to ensure that subjects with extreme $\widehat{\eta}_{0i}$ are over-represented, and the proportion of exposed subjects in Phase II is enriched, $\lambda = \sum_k n_{k,E}/n$, where $n_{k,E}$ is the expected number of subjects sampled from stratum $R^{k,E}$. For continuous exposure, subjects are partitioned into two strata, of whom $R^2 = \{(\widehat{\eta}_{0i}, \widehat{\eta}_{1i}, E_i) : C_0^1 < \widehat{\eta}_{0i} \leq C_0^2, C_1^1 < \widehat{\eta}_{1i} \leq C_1^2, C_e^1 < E_i \leq C_e^2\}$ contains those located in the center of the sampling distribution and $R^1 = \{(\widehat{\eta}_{0i}, \widehat{\eta}_{1i}, E_i) \notin R^2\}$, if a multivariate Q_i is considered. Cutpoints for subject stratification and stratum-specific selection probabilities are chosen to match their counterparts in Design 2 with a bivariate Q_i.

As an extension to the outcome trajectory dependent sampling, this design tends to capture a larger exposure-outcome variation by incorporating an exposure enrichment strategy, while inheriting the favorable property of Design 2 that the conditional likelihood adjusting for the sampling bias can be derived analytically. For instance, if $Q_i = (\widehat{\eta}_{0i}, E_i)'$ and the binary exposure $E_i = 1$, the subject-specific correction for the sampling bias $P(S_i = 1|X_i)$ can be computed by (4.3) as an average of stratum-specific selection probabilities $\pi(R^{k,E})$ across $k = 1, 2, 3$, weighted by $P(Q_i \in R^{k,E}|X_i)$. Because personal exposure E_i is observed as a part of X_i, the sampling variable $Q_i|X_i$ indeed follows the same distribution as in Design 2 when sampling from $Q_i = \widehat{\eta}_{0i}$.

4.2.5 Design 4: Outcome Trajectory Dependent Sampling Using Best Linear Unbiased Predictors of Random Effects

In unbalanced longitudinal studies, the number of measurements available for each individual can be different. Subjects who are lost to follow-up early may have unstable OLS estimates, considering the small ratio of sample size to the number of parameters, $r_i/2$, in each simple linear regression. Furthermore, for subjects with only one measurement of the outcome, it is infeasible to fit a regression model. To handle the unbalanced nature of longitudinal data, and to describe the temporal pattern of the outcome while controlling for confounding factors, in Design 4, we propose to use the best linear unbiased predictors (BLUPs) from a linear mixed model with random intercept and random slope as the sampling variable.

Specifically, we first construct a mixed model using the Phase I data: $Y_i = \alpha X_i^* + a_i Z_i + e_i$, $i = 1, ..., N$, where α is the vector of population regression coefficients, $a_i = (a_{0i}, a_{1i})' \sim (0, D^*)$ is the vector of subject-specific random effects, and e_i is the measurement error assumed to be normally distributed with mean zero and covariance matrix $R_i^* = (\sigma_e^*)^2 \cdot I_{r_i}$. Under this sampling model, the empirical BLUPs of random effects for subject i can be obtained by $\hat{a}_i = \hat{D}^* Z_i' \hat{\Sigma}_i^{*-1}(Y_i - X_i^* \hat{\alpha})$, where $\hat{\alpha}$, $\hat{D}^*$, and $\hat{\Sigma}_i^* = Z_i \hat{D}^* Z_i' + \hat{R}_i^*$ are the restricted maximum likelihood (REML) estimates for fixed effects α, covariance parameter D^*, and the marginal covariance of $Y_i|X_i^*$, respectively. If we treat these REML estimates as fixed and define $\hat{\theta}^* = (\hat{\alpha}, \hat{D}^*, \hat{\sigma}_e^*)$ and $W_i = \hat{D}^* Z_i' \hat{\Sigma}_i^{*-1}$, the sampling variable based upon empirical BLUPs ($Q_i = \hat{a}_i$) should follow a normal distribution $Q_i|X_i, \hat{\theta}^* \sim (\mu_{q_i} = W_i\mu_i - W_i X_i^* \hat{\alpha}, \Sigma_{q_i} = W_i \Sigma_i W_i')$. Unlike Design 2, the distribution of the sampling variable Q_i under Design 4 depends on both the subject-specific design matrix X_i and the parameter estimates $\hat{\theta}^*$ obtained in Phase I. We note that plugging the REML estimates $\hat{\theta}^*$ into the distribution function of $Q_i|X_i$ may under-estimate parameter uncertainty, however, these fixed estimates allow us to derive the explicit formula for the mean and covariance matrix of the sampling variable, and thus markedly improves the computation of conditional likelihood.

Stratification of subjects, allocation of stratum-specific selection probabilities, and random sampling within each stratum are implemented in a similar fashion as in Design 2. Compared to Design 2, the major advantage of choosing BLUPs from a mixed model as Q_i over OLS estimates from a simple linear regression is that, BLUPs not only uses information on subject-specific (Y_i, T_i) but also can borrow strength from other subjects (Y_l, X_l^*), $l = 1, ..., N$, $l \neq i$, therefore, it is expected to better characterize the individual outcome trajectory with unbalanced data.

4.2.6 Design 5: Exposure Enriched Outcome Trajectory Dependent Sampling Using BLUPs of Random Effects

In Design 5, we propose sampling schemes that combine the exposure enrichment strategy in Design 1 with the outcome trajectory dependent sampling using BLUPs of random effects in Design 4. In the same way as how to convert Design 2 to Design 3, we first specify the joint sampling variable as $Q_i = (\widehat{a}_{0i}, E_i)'$, $Q_i = (\widehat{a}_{1i}, E_i)'$, or $Q_i = (\widehat{a}_{0i}, \widehat{a}_{0i}, E_i)'$, and then calculate the correction probability $P(S_i = 1|X_i)$ for each subject based upon identified distribution of $Q_i|X_i$. We emphasize that although exposure is adjusted in the mixed effect sampling model in Design 4, we believe enrich sample with exposed subjects can further help to increase the exposure-outcome variation.

Figure 4.1 provides a visualization of sample selection under each study design, given an overall selection probability of 0.25 from the original cohort of $N = 1000$ subjects, with a low exposure prevalence $P(E = 1) = 0.2$. For the brevity purpose, we do not present sample selection for Design 4 and Design 5, because they are implemented similarly to Design 2 and Design 3 but have OLS estimates replaced by BLUPs of random effects.

4.3 Likelihood Functions and Estimation Approaches

When data have been collected via one of the sampling designs described in Section 4.2, they consist of N subjects: $\{i : S_i = 1\}$ of whom have complete information (Y_i, X_i), which we refer to as complete-cases; while $\{i : S_i = 0\}$ of whom have partial information (Y_i, X_i^*), which we refer to as incomplete-cases. We now describe four likelihood functions one can use to estimate the regression parameters.

4.3.1 Unweighted Uncorrected Likelihood

Regardless of the sampling mechanism, one can naively perform the analysis for a standard prospective cohort, and make inference on parameters of interest in the linear mixed model (4.1) by equating the derivative of the unweighted uncorrected log-likelihood (UUL) to zero:

$$\sum_{i:S_i=1} \frac{\partial \log f(Y_i|X_i; \beta, \sigma)}{\partial \beta} = 0 \tag{4.4}$$

Solutions to this equation yield the maximum likelihood estimates of β (or σ). One problem of this naive analysis is that there is no guarantee for consistent estimates when sample selection in Phase II is in relation to the outcome. Moreover, there is a consequence of reduced precision when the likelihood function considers only complete-cases, while partial information on incomplete-cases are ignored.

4.3.2 Inverse Probability Weighted Likelihood

In a two-phase design, when the sample selection in Phase II is driven by exposure or outcome trajectory, selection bias may be introduced. To draw a valid inference that accounts for the sampling mechanism, one may consider the inverse probability weighted likelihood (IPWL), a modification of complete-case analysis that differentially weights subjects to adjust for the selection bias.

In particular, under our sampling designs, subject-specific selection probability $P(S_i = 1|Y_i, X_i)$ for the entire cohort $i = 1, ..., N$ can be obtained by matching the stratum-specific selection probability with identified personal stratum membership. Based upon information on complete-cases, consistent estimators of β (or σ) can be derived by solving the estimation equation

$$\sum_{i:S_i=1} \frac{\partial \log f(Y_i|X_i;\beta,\sigma)}{\partial \beta} \cdot [P(S_i = 1|Y_i, X_i)]^{-1} = 0 \tag{4.5}$$

Here, the contribution to the score function from a single subject in Phase II is weighted by the inverse of its sampling probability (Robins *et al.*, 1994). When a constant sampling probability is assigned, such as n/N in random sampling, it is easy to show that the IPWL becomes equivalent to the UUL. In IPWL, the idea of introducing weights into the standard likelihood is simple and the computation of equation (4.5) is straightforward, however, we lose estimation efficiency since the analysis is restricted to the complete-cases. Furthermore, the IPWL estimates can be quite variable when the stratum-specific selection probabilities get close to zero (Little and Rubin, 2014).

4.3.3 Complete-Case Conditional Likelihood

To adjust for biased sampling from outcome trajectory dependent sampling (Design 2), Schildcrout *et al.* (2013) developed an ascertainment corrected maximum likelihood for inference. In their analysis, subjects in the set of complete-cases contribute to the likelihood by a conditional probability of the outcome vector given being sampled in Phase II, $P(Y_i|X_i, S_i = 1; \beta, \sigma)$. By Bayes' theorem, its likelihood function, which we refer to as the complete-case conditional likelihood (CCL), can be derived as:

(4.6)

$$\begin{aligned} L^C(\beta, \sigma) &= \prod_{i:S_i=1} f(Y_i|X_i, S_i = 1; \beta, \sigma) \\ &= \prod_{i:S_i=1} \frac{P(S_i = 1|Y_i, X_i) f(Y_i|X_i; \beta, \sigma) f(X_i)}{P(S_i = 1|X_i; \beta, \sigma) f(X_i)} = \prod_{i:S_i=1} \frac{\pi(q_i) \cdot f(Y_i|X_i; \beta, \sigma)}{P(S_i = 1|X_i; \beta, \sigma)} \end{aligned}$$

Basically, subject-specific contribution to the CCL is composed of three terms: the multivariate density $f(Y_i|X_i; \beta, \sigma)$ in a standard analysis, the subject-specific correction term $P(S_i = 1|X_i; \beta, \sigma)$ that adjusts for the biased sampling, and the subject-specific selection probability $P(S_i = 1|Y_i, X_i) = \pi(q_i)$ determined by observed q_i, which is functionally independent of parameters β and σ.

We highlight the fact that all the sampling variables specified in Design 2 - Design 5 are linear functions of the outcome vector. Under the normality assumption of $Y_i|X_i$ in the response model, these sampling variables should also follow a normal distribution with its mean and covariance indexed by parameters β and σ, i.e., $Q_i|X_i \sim \big(\mu_{q_i}(\beta), \Sigma_{q_i}(\sigma)\big)$. Accordingly, the correction term for subject i can be computed as a weighted average of stratum-specific selection probabilities across all strata, $P(S_i = 1|X_i; \beta, \sigma) = \sum_{k=1}^{K} \pi(R^k) P(Q_i \in R^k|X_i; \beta, \sigma)$. This ensures the CCL be expressed in a closed form, thereby considerably conveniences the computation of the likelihood. Score functions of the CCL with respect to parameters β and σ are calculated analytically as in Schildcrout et al. (2013). One can obtain the CCL estimates by solving the corresponding score equations using the Newton-Raphson algorithm, and the covariance matrix by the inverse of the numerical derivative of the score function.

4.3.4 Full Conditional Likelihood

Since the likelihood functions defined in the UUL, IPWL and CCL consider only complete-cases while ignoring information available on the set of incomplete-cases, we propose a full conditional likelihood (FCL) that accounts for all subjects into the likelihood in hopes to gain efficiency in the parameter estimation. To describe the FCL, we first consider a binary genotype G_i and let

$$p(G_i|X_i^*;\gamma) = \frac{\exp(G_i \cdot X_i^*\gamma)}{1+\exp(X_i^*\gamma)} \tag{4.7}$$

denote the probability mass function of G_i given X_i^* through a logistic regression model with a nuisance parameter γ. Multinomial regression can be used in the case of a polychotomous G_i. By Bayes' theorem, we can write down the FCL that involves all subjects enrolled in the full cohort as

$$\begin{aligned} L^F(\beta,\sigma,\gamma) &= \prod_{i:S_i=1} f(Y_i,G_i|X_i^*,S_i=1;\beta,\sigma,\gamma) \prod_{i:S_i=0} f(Y_i|X_i^*,S_i=0;\beta,\sigma,\gamma) \\ &= \prod_{i:S_i=1} \frac{P(S_i=1|Y_i,G_i,X_i^*)f(Y_i|G_i,X_i^*;\beta,\sigma)p(G_i|X_i^*;\gamma)f(X_i^*)}{P(S_i=1|X_i^*;\beta,\sigma,\gamma)f(X_i^*)} \\ &\quad \cdot \prod_{i:S_i=0} \frac{P(S_i=0|Y_i,X_i^*)f(Y_i|X_i^*;\beta,\sigma,\gamma)f(X_i^*)}{P(S_i=0|X_i^*;\beta,\sigma,\gamma)f(X_i^*)} \\ &= \prod_{i:S_i=1} \frac{\pi(q_i)\cdot f(Y_i|X_i;\beta,\sigma)\cdot p(G_i|X_i^*;\gamma)}{P(S_i=1|X_i^*;\beta,\sigma,\gamma)} \prod_{i:S_i=0} \frac{[1-\pi(q_i)]\cdot f(Y_i|X_i^*;\beta,\sigma,\gamma)}{1-P(S_i=1|X_i^*;\beta,\sigma,\gamma)} \end{aligned} \tag{4.8}$$

Different from (4.6), complete-case contributes to the likelihood in (4.8) through a joint probability of the outcome and genotype $P(Y_i,G_i|X_i^*,S_i = 1;\beta,\sigma,\gamma)$, a common strategy in the missing data literature (Lawless *et al.*, 1999). For subjects with unknown G_i, their contribution to the likelihood in (4.8) is given by $P(Y_i|X_i^*;\beta,\sigma,\gamma) = \sum_{G_i\in\{0,1\}} f(Y_i|G_i,X_i^*;\beta,\sigma)p(G_i|X_i^*;\gamma)$. The FCL corrects for the sampling bias via $P(S_i = 1|X_i^*;\beta,\sigma,\gamma) = \sum_{G_i\in\{0,1\}} P(S_i = 1|G_i,X_i^*;\beta,\sigma)p(G_i|X_i^*;\gamma)$. Because the genotype variable renders a missing at random mechanism, the subject-specific selection probability determined by observed q_i, is independent of parameters (β,σ,γ), $P(S_i = 1|Y_i,X_i) = P(S_i = 1|Y_i,X_i^*) = \pi(q_i)$.

We estimate parameters of the FCL by direct maximization using the Newton-Raphson algorithm, with the initial values for (β,σ,γ) set equal to the standard maximum likelihood estimates. Estimated covariance can be calculated numerically after the final Newton-Raphson iteration.

4.4 Simulation Study

4.4.1 Description of Simulation Settings

We investigated the performance of the five sampling designs proposed in Section 4.2 using the four likelihood approaches described in Section 4.3 under various simulation settings. Following the general form of linear mixed model in (4.1), we generated individual level data with $r_i = 5$ repeated measures of the continuous outcome at equally spaced observation times $Ti = \{T_{i1},...,T_{i5}\} = \{-1,-0.5,...,1\}$ for subject $i = 1,...,N$. This model involves a random intercept and a random slope of time $Z_i = (1,T_i)$, with its marginal mean for subject i given by

$$X_i\beta = \beta_0+\beta_T T_i+\beta_E E_i+\beta_{ET}E_iT_i+\beta_G G_i+\beta_{GE}G_iE_i+\beta_{GT}G_iT_i+\beta_{GET}G_iE_iT_i \tag{4.9}$$

We considered a binary genotype with a minor allele frequency of $P(G_i = 1) = 0.1$. We examined over a range of combinations for different exposure types, interaction models, and G-E associations, as presented in Table 4.1. For simulation settings with binary exposure, we used a prevalence rate of $P(E_i = 1) = 0.2$; and for simulation settings with continuous exposure, we used a standard normal distribution $E_i \sim (0, 1)$. When a two-way interaction model was considered, we assumed a genetic modification effect on the exposure-outcome association that was constant over time, in addition to the main effects of the genotype and exposure. In the three-way interaction model where all parameters in (4.9) were set to be non-zero, we assumed that both main and interaction effects of the genotype and exposure are time-dependent. Under simulations when genotype and personal exposure were correlated, we controlled the strength of G-E association by a logistic regression model defined by logit$\{P(G_i = 1|E_i; \gamma)\} = \gamma_0 + \gamma_E E_i$, where the association parameter $\gamma_E = 0.2$ represents an odds ratio of 1.22. To maintain comparability across simulation settings, parameters of fixed effects were selected to reflect the contribution of time (10-20%), exposure (5%), genotype (1%), and GxE interaction (0.5-1%) in explaining the variance of the outcome. Confounding factors were not considered in these models. For simulating the random effect for subject i, we set $b_i = (b_{0i}, b_{1i})' \sim (0, D)$, where its variance components $\sigma_0^2 = \sigma_1^2 = 1$ and $\rho = 0$. The error term $\epsilon_i \sim (0, \sigma_e^2 I_{r_i})$ with σ_e set to 1 or 2. Moreover, we examined above simulation settings with both balanced and unbalanced data. Under longitudinal design with unbalanced data, 10% of remaining subjects were randomly selected as dropouts at each follow-up observation, so by the end of the study about 65% of subjects in the original cohort underwent five repeated measurements of the outcome, implying a missing completely at random mechanism with a monotone pattern.

Table 4.2 shows how the sampling variable Q_i is related to the environmental exposure and/or outcome trajectory, and provides guidance on the allocation of stratum-specific selection probabilities under the five sampling designs. To reflect a moderate budgetary constraint, we assumed that $n = 250$ from the original cohort of $N = 1000$ subjects are sampled in Phase II under a two-way interaction model. Previous studies suggested that sampling subjects more towards the extremes of Q_i led to larger efficiency gains (Boks *et al.*, 2007; Schildcrout *et al.*, 2013), hence we used stratum sizes $(N_1, N_2, N_3) = (100, 800, 100)$ for unvariate continuous Q_i and $(N_1, N_2) = (200, 800)$ for bivariate or multivariate continuous Q_i. Given the rare binary exposure in our settings, exposed subjects were enriched to reach a proportion of $\lambda = 0.5$ in the selected sample. In Design 3 or 5, when the bivariate Q_i depends upon the mixture of a continuous intercept (or slope) and a binary exposure, we considered stratum sizes $(N_{1,E}, N_{2,E}, N_{3,E}; N_{1,\bar{E}}, N_{2,\bar{E}}, N_{3,\bar{E}}) = (50, 100, 50; 50, 700, 50)$. Similarly, with a multivariate Q_i depending upon both intercept and slope and a binary exposure, subjects were partitioned into strata with different sizes $(N_{1,E}, N_{2,E}; N_{1,\bar{E}}, N_{2,\bar{E}}) = (100, 100; 100, 700)$. To detect the three-way GxExT interaction, $n = 500$ subjects out of $N = 5000$ were sampled in Phase II. While the same enrichment proportion $\lambda = 0.5$ was targeted, adjustments to the stratum sizes were made, $(N_1, N_2, N_3) = (200, 4600, 200)$ and $(N_{1,E}, N_{2,E}, N_{3,E}; N_{1,\bar{E}}, N_{2,\bar{E}}, N_{3,\bar{E}}) = (100, 800, 100; 100, 3800, 100)$, in order to increase the variability in Q_i given the genotyping capacity.

The efficiency of our proposed sampling designs and likelihood approaches were quantified by three evaluation metrics: bias, relative efficiency, and detection power. Bias is estimated as the average difference between the estimator and parameter over 1000 Monte Carlo runs. Relative efficiency is defined as the ratio of the average mean squared error (MSE) from random sampling with UUL, to the average MSE from a two-phase design with one of the likelihood approaches (UUL, IPWL, CCML, and FCL). Power is estimated as the proportion of correctly rejecting a non-zero effect using a two-sided Wald test at a significance level of 0.05.

4.4.2 Summary of Simulation Results

A change in the exposure-outcome relationship for a given genotype subgroup is of interest in GxE interaction studies, so we focused on the detection and estimation of two particular effects: the GxE interaction effect β_{GE} (or GxExT interaction effect β_{GET} under a three-way interaction model), and the joint exposure effect $\beta_E+\beta_{GE}$ (or $\beta_E+\beta_{ET}+\beta_{GE}+\beta_{GET}$ under a three-way interaction model) among carriers of the risk allele ($G_i = 1$).

<u>Bias:</u> Table 4.3 presents estimated bias for GxE interaction and joint exposure effect when the data were generated from a two-way GxE interaction model with rare exposure and balanced data. Among the four likelihood approaches, FCL effect estimates are closest to the true parameters in all of the designs considered, with the largest bias relative to the parameter no greater than 6%. Estimated bias using the CCL is small ($< 10\%$), but not for the main effect of exposure under Design 3, leading to the over-estimation of the joint exposure effect $\beta_E + \beta_{GE}$. The UUL that ignores the design for analysis yields severely biased estimates (12% - 123%) when the sampling variable is related to the individual mean of the outcome vector, e.g., $\widehat{\eta}_{0i}$ and $\widehat{a}_{0i}$. IPWL estimate, considering the limited sample size, produces modestly biased estimates. For example, under Design 2 with $Q_i = \widehat{\eta}_{0i}$, estimated bias of β_{GE} and $\beta_E + \beta_{GE}$ in the UUL are 1.23 and -0.76, as compared to -0.27 and -0.32 in the IPWL, respectively. No significant bias for the UUL estimates was observed under Design 1.

We examine the impact of different sampling designs and likelihood approaches on estimated bias of time-varying GxE interaction and joint exposure effect under a three-way GxExT interaction model (Table 4.6). Likewise, the FCL yields nearly unbiased estimates with the smallest differences to the true parameters, followed by the CCL and IPWL. Note that substantial bias for β_{ET} and β_{GET} has been observed in the UUL estimate when the sampling mechanism is based upon individual slope of the outcome vector, such as $\widehat{\eta}_{1i}$ and $\widehat{a}_{1i}$. This is because subjects with greater temporal variation are more likely to be sampled, bringing bias into the estimation of time-varying effects. In alternative settings not reported in this paper, the benefits of using the FCL over other likelihood approaches are preserved regardless of the exposure type, G-E association, and longitudinal data structure (balanced or not).

Relative efficiency: Figure 4.2 illustrates the efficiency of GxE interaction and joint exposure effect under each design and likelihood combination relative to the UUL estimates using random sampling, given a rare exposure and balanced data. We note that estimation efficiency for the GxE interaction effect is improved by two ways, increasing the variation of the cross-product interaction term among sampled subjects by exposure enrichment, as in Design 1; or increasing the variation of the outcome by relating the sampling variable to the individual outcome trajectory, as in Designs 2 and 4. For example, using the FCL, estimated relative efficiency for β_{GE} under Design 1 and Designs 2 and 4 based upon an intercept estimate are 2.03, 2.29, and 2.25, respectively. An increase in efficiency is achieved under designs that consider both personal exposure and estimated intercept in the sample selection, such as Designs 3 and 5 with the estimated relative efficiency for β_{GE} exceeding 3.50.

With the same likelihood approach, we see little difference in the estimated efficiency between Designs 2 and 4, as well as Designs 3 and 5. This is because subject stratification under these two designs are highly concordant, approximately 90% of subjects in Phase I are partitioned into the same stratum, despite model differences in characterizing the individual outcome trajectory over time. To handle unbalanced data with a larger time effect (β_T corresponds to 20% variation in the outcome) and a smaller within-subject correlation

($\sigma_e = 2$), sampling design based upon BLUPs provides a larger efficiency improvement over OLS estimates (Figure 4.4). For example, when estimating β_{GE}, Design 4 is 39$\%$ (1.56/1.12=1.39) more efficient than Design 2, and Design 5 is 26$\%$ (3.94/3.12=1.26) more efficient than Design 3.

When the data are simulated from a three-way GxExT interaction model, estimated relative efficiency for the time-varying effects β_{GET} and $\beta_E + \beta_{ET} + \beta_{GE} + \beta_{GET}$ are high when sampling is based upon estimated slope of the outcome vector instead of the estimated intercept (Figure 4.5), which is consistent with prior findings (Schildcrout *et al.*, 2013, 2015). It has been reported that the assumption of G-E independence can improve the efficiency of odds ratio of the GxE interaction in case-control studies (Chatterjee and Chen, 2007). However, no appreciable difference has been observed in the presence of a moderate G-E association, while keeping other parameters unchanged (results not shown). This suggests that estimation of the GxE interaction effect on a longitudinal outcome is insensitive to the incorporation of a G-E association at a realistic strength ($OR_{GE} = 1.22$). Moreover, due to increased resolution and enriched sampling at two tails, continuous exposure under examined designs shows a similar pattern, but leads to a larger efficiency gain for β_E and β_{GE} than a binary exposure. For example, relative efficiency for β_{GE} under Design 1 with the FCL is 2.03 with a binary exposure, but increases to 2.82 with a continuous exposure.

Overall, the FCL provides most efficient estimates (15$\%$ - 29$\%$ gain in the relative efficiency over the CCL, and 20$\%$ - 663$\%$ gain over the UUL) across all designs under different simulation settings. Additionally, while the efficiency gain of the FCL over the UUL is modest for β_{GE} (for instance, $2.03/1.69 = 1.20$ under Design 1 with a two-way GxE interaction model), it becomes substantial for covariates not related to the genotype, such as β_E ($3.54/1.54 = 2.30$). This is because information available on subjects not sampled in Phase II are recovered using the FCL. Due to the use of a sandwich-type variance estimator, IPWL estimates are found to be less efficient than the CCL. In addition, the impact of sampling design outweighs the impact of likelihood approach as long as the selection bias is corrected by a conditional likelihood.

Detection power: Table 4.4 lists estimated power of detecting a causal GxE interaction and a joint exposure effect among $G_i = 1$ under a two-way GxE interaction model with rare exposure and balanced data. When testing β_{GE}, we observe that oversampling subjects by environmental exposure (Design 1) or estimated intercept (Design 2) are approximately 50% more powerful than random sampling ($\sim 0.36/0.24 = 1.50$), if the sampling bias is corrected by a conditional likelihood. Sampling based upon both environmental exposure and estimated intercept (Design 3) leads to a considerable improvement that is 2.21 times the power from random sampling ($\sim 0.53/0.24 = 2.21$). When testing $\beta_E + \beta_{GE}$, adequate power ($> 95\%$) can be achieved by using any of the sampling designs, as opposed to a 76% power from random sampling. Moreover, we find the power for detecting the GxE interaction and the joint exposure effect under Design 2 (or Design 3) is similar to Design 4 (or Design 5) due to the high correlation between OLS estimates and the BLUPs of the individual outcome trajectory. No significant power gain has been observed when replacing OLS estimates with the BLUPs in the sampling design for unbalanced longitudinal data.

We also estimated sample size required to achieve 80% power for studying GxE interaction under different sampling designs. Assuming a full cohort of 5000 subjects among which 20% are exposed, and an effect size of one unit reduction in the outcome for risk allele carriers when exposed, random selection of 1100 subjects in Phase II provides adequate power to detect the GxE interaction, while a subsample of 250 using exposure enriched outcome trajectory dependent sampling (Design 5) should suffice, leading to a substantial reduction in the sample size requirement.

Comparisons among different likelihood approaches suggest that the FCL is most powerful at detecting the GxE interaction and joint exposure effect. The CCL behaves similarly to the FCL when the sampling variable is related to the outcome vector, but is less likely to identify exposure-related effects when the sampling is exposure biased. For example, under Design 3 with $Q_i = (\widehat{\eta}_{0i}, E_i)$, the FCL yields the greatest power at 53% and 100% for β_{GE} and $\beta_E + \beta_{GE}$, whereas the CCL is able to detect these effects at an estimated 15% and 86%. This may relates to the fact the CCL restricts its analysis on complete-cases while ignoring the exposure information collected on incomplete-cases. One should note that the IPWL tends to have inflated detection power at the cost of reduced efficiency due to the use of a sandwich variance estimator. The naive UUL constantly provides severely biased and least powered estimate. We also examined sensitivity of our designs and likelihood approaches to various simulation settings and found no qualitative differences.

4.5 Data Example: the Normative Aging Study

Since year 1991, participants of the NAS were invited to a bone lead assessment using a K-x-ray fluorescence instrument, which provides an index of cumulative lead exposure. The outcome of interest is the difference between systolic blood pressure and diastolic blood pressure (pulse pressure, PP), which was measured at the time of bone lead assessment (baseline, 1991-2002) and followed up every three years until 2013, with a median follow-up time of 12.1 years. Indeed, lead exposure has been associated with increased PP (Perlstein *et al.*, 2007). Zhang *et al.* (2010) observed a significant GxE interaction between polymorphisms in the *HFE* gene and cumulative lead exposure on PP . In this example, we aim to illustrate the utility of exposure enriched outcome trajectory dependent sampling and FCL approach in the analysis of *HFE* by lead interaction.

We focused on 720 subjects from the NAS cohort who were successfully assessed for cumulative lead exposure at the patella bone and genotyped for the *HFE* gene. Subjects with compound heterozygotes were excluded because, between two major *HFE* variant alleles (*C282Y* and *H63D*) the association between lead exposure and PP was found to be exclusive among *H63D* variant carriers (having one or two *H63D* variant alleles but no *C282Y* variant allele) (Ko *et al.*, 2013). This results in a full cohort of 706 subjects (descriptive characteristics in Table 4.5), of whom more than 96% had at least two measurements, contributing to a total of 3265 observations. The majority (97%) of the subjects were Caucasian, with an average age of 66.3 ± 7.2 at the baseline measurement and a risk allele frequency of 21.8%. Patella bone lead concentration was measured continuously, but dichotomized to reflect a relatively rare binary exposure with a prevalence of 0.1 (High:$\geq$52 μg/g; Low: $<$52 μg/g).

For illustration purposes, we assume that personal genotype data were not available by the end of longitudinal follow-up, and the budget constraint allows retrospective genotyping for only 200 subjects. Full cohort analysis aligned with the findings in Zhang *et al.* (2010) that the mean PP was estimated to be 7.61 mm Hg (95% CI: [1.89, 13.33]) higher for the high patella lead group than the low patella lead group among the *H63D* variant carriers. For wild types, the difference in the mean PP between the high and low exposure groups was estimated to be -1.57 mm Hg (95% CI: [-4.24, 1.10]). Supported by the the Akaike information criterion (AIC), this analysis used a mixed effects model with random intercept and random slope of time, adjusted for baseline age, body mass index, education level, hypertension, and Type II diabetes as fixed effects.

In addition to random sampling, we examined five designs described in Section 4.2. We initially included a lead by time interaction, a *H63D* by time interaction, and a *H63D* by lead by time interaction in the mixed model, and found none of these interactions were significant in the full cohort analysis, thereby we considered sampling variables Q_i in Designs 2-5 that depend on the intercept estimate of the outcome trajectory $\widehat{\eta}_{0i}$ and $\widehat{a}_{0i}$, rather than the slope or bivariate estimate, provided the results in the simulation studies. In particular, we specified stratum sizes $(N1; N2; N3) = (71; 564; 71)$ for Designs 2 and 4, and $(N_{1,E}, N_{2,E}, N_{3,E}; N_{1,\bar{E}}, N_{2,\bar{E}}, N_{3,\bar{E}}) = (7, 57, 7; 58, 519, 58)$ for Designs 3 and 5. Due to the limited genotyping capacity and low exposure prevalence, the maximum stratum sizes N_E under Design 1 and $N_E = N_{1,E} + N_{2,E} + N_{3,E}$ under Designs 3 and 5 were no greater than 71 ($\approx 706 \times 0.1$), leading to the proportion of high patella lead subjects in Phase II at most $\lambda = 71/200$. Stratum-specific selection probabilities were computed accordingly by Table 4.2. Because of the superior performance in the simulation, we used the FCL for the estimation of regression coefficients.

Figure 4.3 shows average estimated exposure effects among subjects who are carriers of the *H63D* variant or wild types under different designs based upon 500 replicated Phase II samples. Consistent with simulation studies, we found that point estimates of β_E and $\beta_E + \beta_{GE}$ using the FCL were close to results from the full cohort analysis, and estimated efficiency of β_E and $\beta_E + \beta_{GE}$ were considerably improved by our examined designs. For example, we observe that outcome trajectory dependent designs had an estimated relative efficiency of 1.2-1.3 for $\beta_E + \beta_{GE}$ compared to random sampling with a standard analysis, whereas exposure enriched designs were approximately 1.6-2.6 times more efficient than random sampling, given the rare exposure in this example. More importantly, we highlight that incorporation of the exposure enrichment strategy enables detection of the deleterious exposure effect among *H63D* variant carriers under Designs 1, 3 and 5. Specifically, the expected PP was estimated to be 7.55 mm Hg (95% CI: [1.08, 14.02]) higher for the high patella lead group among the *H63D* variant carriers under Design 1, 6.77 mm Hg (95%

CI: [0.00, 13.55]) higher under Design 3, and 6.86 mm Hg (95% CI: [1.43, 12.29]) higher under Design 5. However, this exposure effect, also seen in the full cohort analysis, was considered to be statistically not significant under random sampling, Design 2, or Design 4. We realize that there could be unmeasured confounders in the NAS cohort, yet their potential influence on the GxE interaction was not addressed in our data analysis.

4.6 Discussion

While novel analysis and powerful tests have been proposed to enhance the detection of multiplicative GxE interaction with repeated measures data (Mukherjee *et al.*, 2012a; Ko *et al.*, 2013), it remains relatively less addressed as to how sampling design would affect the statistical inference about the GxE interaction in a longitudinal cohort study. In this paper, we described five study designs that prioritize subjects for retrospective genotyping by leveraging environmental exposure information and individual outcome trajectory during the sample selection. We derived a conditional likelihood using data from both phases and compared it with three alternative complete-case based strategies. Our results indicate that the FCL provides nearly unbiased estimation and enhanced precision (15% - 663% gain in relative efficiency) over existing alternatives. Among competing sampling schemes we considered, exposure enriched outcome trajectory dependent design outperforms others in terms of estimation efficiency and detection power of the GxE interaction. In addition, we found sampling based upon personal exposure and estimated intercept ($\widehat{\eta}_{0i}$ or $\widehat{a}_{0i}$) can improve efficiency of the time-stationary GxE interaction, while sampling based upon personal exposure and estimated slope ($\widehat{\eta}_{1i}$ or $\widehat{a}_{1i}$) can improve efficiency of the time-varying GxE (GxExT) interaction. Therefore, we recommend the use of exposure enriched outcome trajectory dependent design coupled with the FCL-based approach to evaluate the GxE interaction.

To characterize individual outcome trajectory, we compared two classes of regression estimates: OLS estimates for intercept and slope of time from simple linear regressions, as employed in Designs 2 and 3; and BLUPs for random intercept and random slope of time from a linear mixed model, as employed in Designs 4 and 5. Both classes applied dimension reduction in constructing summary features of the longitudinal outcome, and shared the property that analytical distribution of these features can be obtained in a closed form. However, we emphasize that BLUPs can be advantageous, with a 25% - 39% gain in the relative efficiency for the GxE interaction effect over the OLS estimates when accommodating unbalanced data. This is because instead of using the subject-specific information as in OLS estimates, BLUPs can borrow strength from other subjects in the mixed model, making its estimates more robust to the missing data.

We acknowledge this study has several limitations that could be addressed in the future. First, we focus on a time-stationary environmental exposure, but many such exposures change over time in practice (Aschard *et al.*, 2012). For cohort studies that collect longitudinal exposure data, it would be helpful to utilize the time-varying exposure to guide the sample selection in Phase II. Inspired by a recent discovery of gene-by-longitudinal environmental exposure interaction in a case-control study (Wei *et al.*, 2014), one may consider decomposing the time-varying exposure trajectory into a few unrelated components via the functional principal component analysis and explore sampling designs in terms of these components. Secondly, we consider only a linear time trend in the longitudinal outcome with a random intercept and random slope in the sampling model. To handle the possible non-linear time effect, one may regress the outcome on multiple functions of time such as polynomial terms or more general parametric spline basis, and then incorporate these complex smooth features of the outcome trajectory into the sampling mechanism.

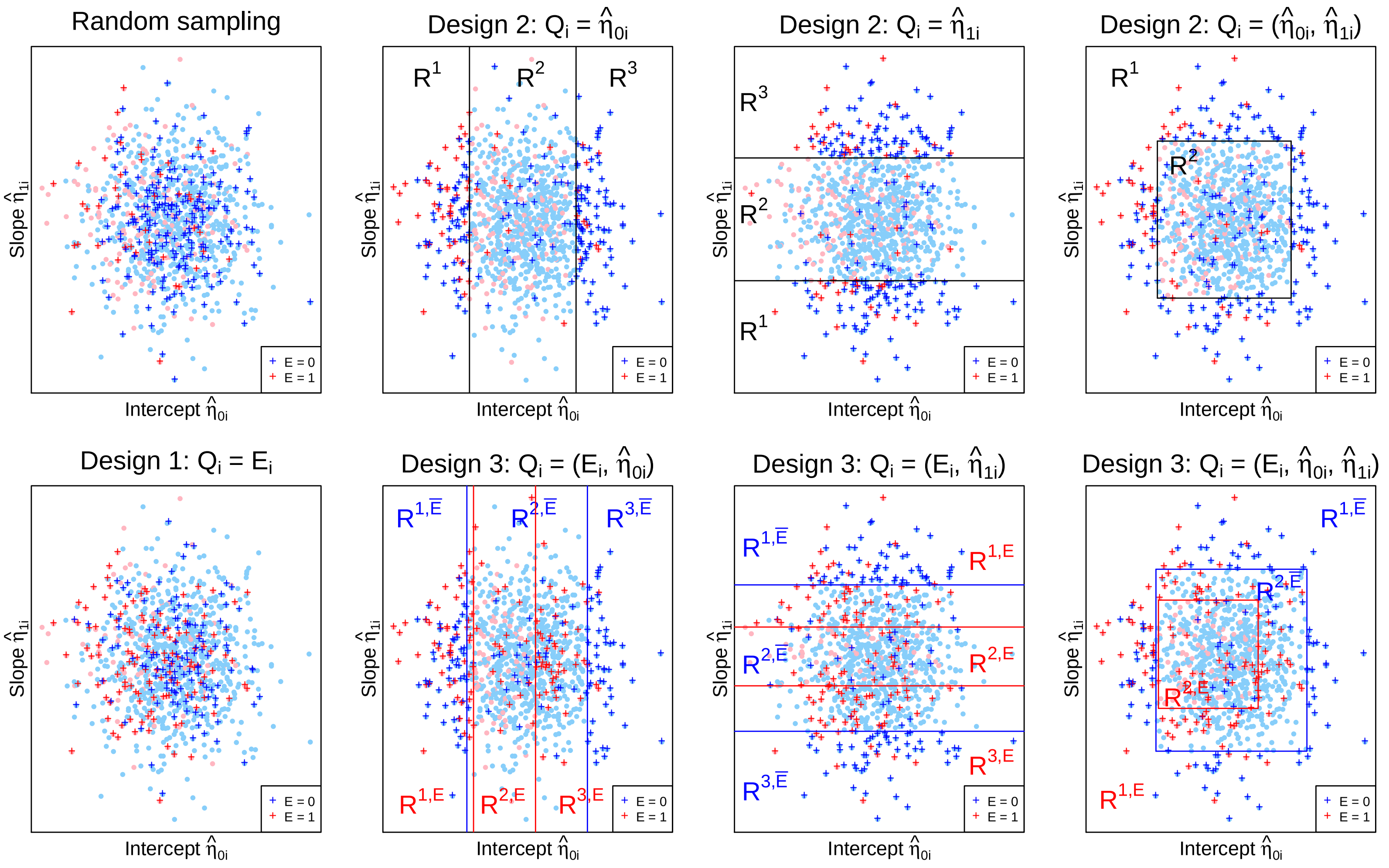

Figure 4.1: Sample selection under different study designs.

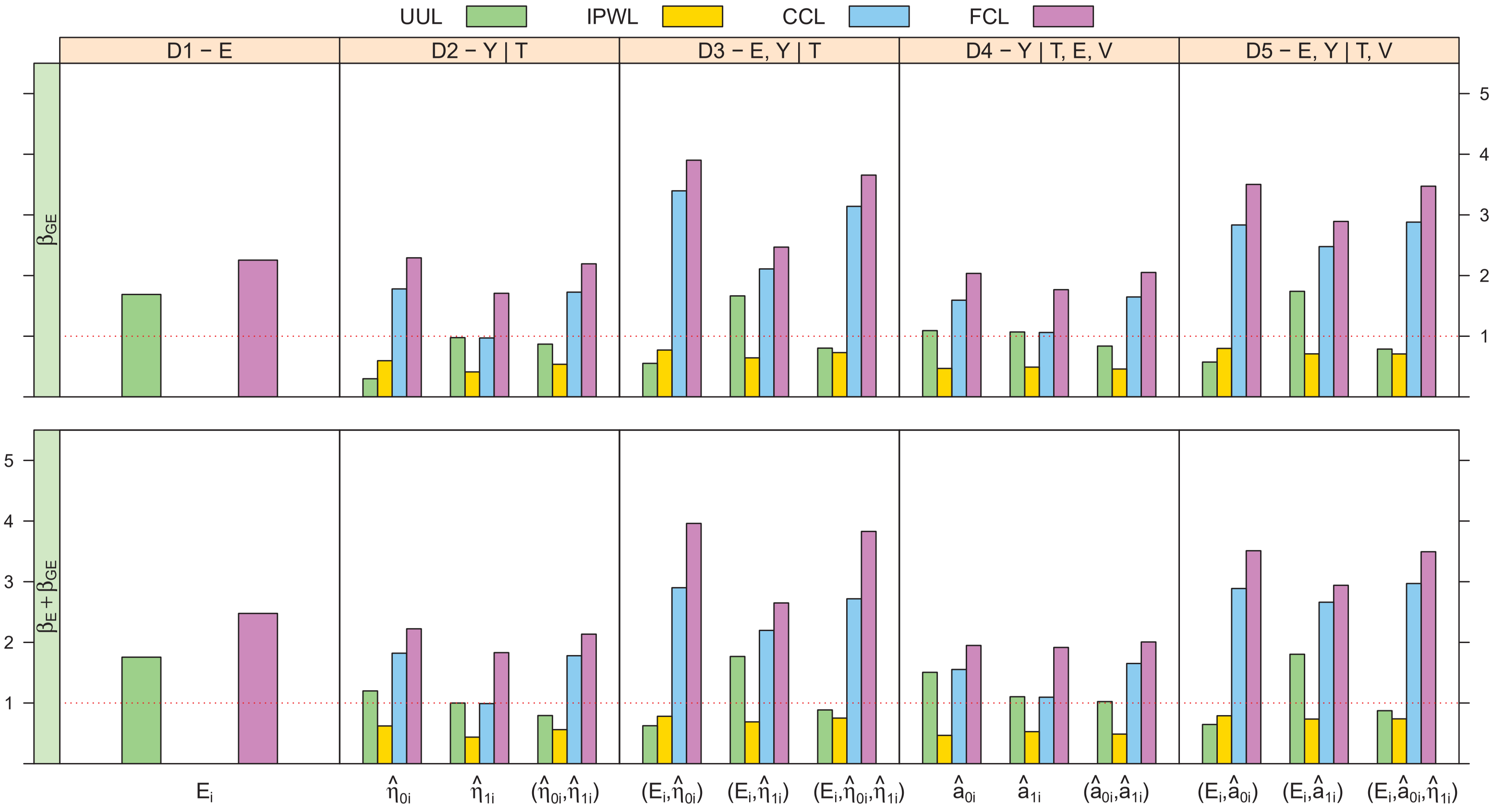

Figure 4.2: Relative efficiency of parameter estimates under the two-way GxE interaction model with a rare exposure and balanced data.

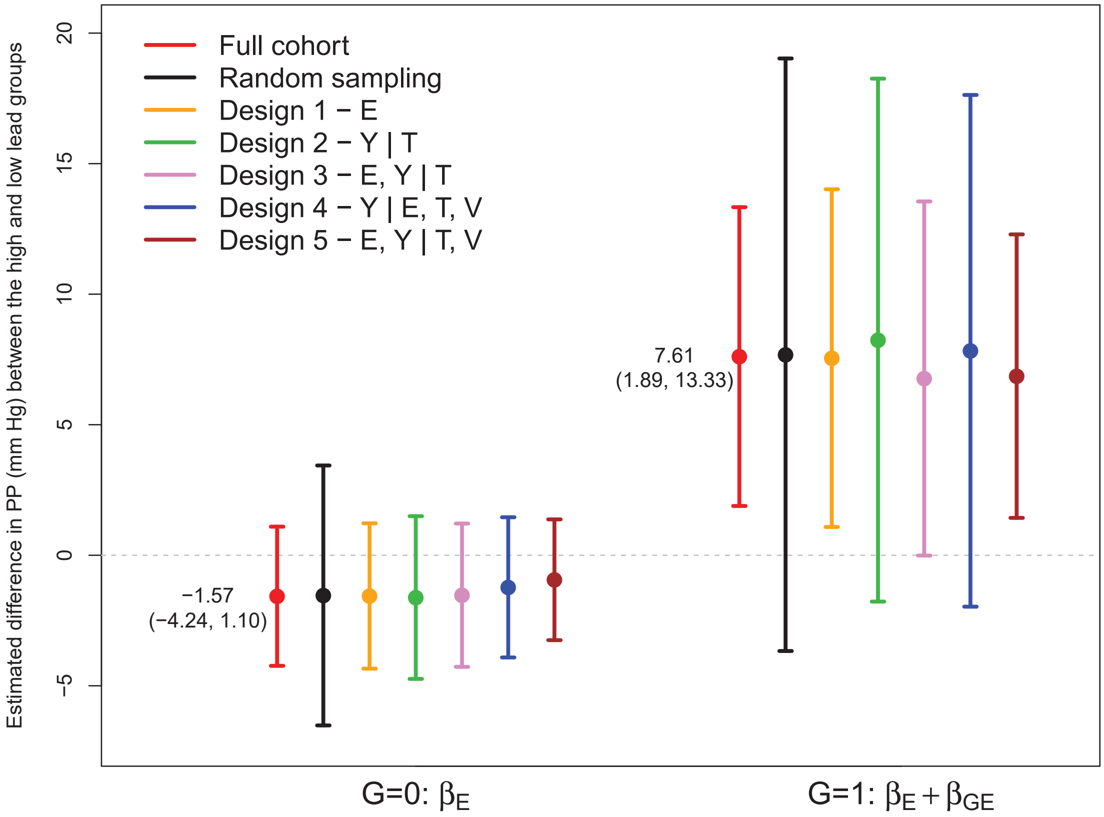

Figure 4.3: NAS results: average estimated effects under different study designs using the FCL (n=200).

Table 4.1: Description of different simulation settings.

Exposure	GxE interaction	G-E association	Parameters of fixed effects $(\beta_0, \beta_T, \beta_E, \beta_{ET}, \beta_G, \beta_{GE}, \beta_{GT}, \beta_{GET})$
Binary	Two-way	Independent Dependent	(10, -0.7, -1.0, 0, -0.6, -1.0, 0, 0)
	Three-way	Independent Dependent	(10, -0.7, -1.0, -1.0, -0.6, -1.0, -1.0, -1.5)
Continuous	Two-way	Independent Dependent	(10, -0.7, -0.4, 0, -0.6, -0.5, 0, 0)
	Three-way	Independent Dependent	(10, -0.7, -0.4, -0.4, -0.6, -0.5, -1.0, -1.0)

Table 4.2: Configurations of examined two-phase longitudinal designs.

Sampling scheme	Sampling variable	Exposure	# strata	Stratum-specific selection probabilities $\pi(R^1), ..., \pi(R^K) = (n_1/N_1, ..., n_K/N_K)$
Random*	-	-	1	n/N
Design 1 - E	$Q_i = E_i$	Binary	2	$(\lambda n/N_E, (1-\lambda)n/N_{\bar{E}})$
		Continuous	3	$(N_1/N_1, (n-N_1-N_3)/N_2, N_3/N_3)$
Design 2 - Y|T	$Q_i = \widehat{\eta}_{0i}$	-	3	$(N_1/N_1, (n-N_1-N_3)/N_2, N_3/N_3)$
	$Q_i = \widehat{\eta}_{1i}$	-	3	$(N_1/N_1, (n-N_1-N_3)/N_2, N_3/N_3)$
	$Q_i = (\widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	-	2	$(N_1/N_1, (n-N_1)/N_2)$
Design 3 - E, Y|T†	$Q_i = (E_i, \widehat{\eta}_{0i})$	Binary	6	$(N_{1,E}/N_{1,E}, (\lambda n - N_{1,E} - N_{3,E})/N_{2,E}, N_{3,E}/N_{3,E},$ $N_{1,\bar{E}}/N_{1,\bar{E}}, ((1-\lambda)n - N_{1,\bar{E}} - N_{3,\bar{E}})/N_{2,\bar{E}}, N_{3,\bar{E}}/N_{3,\bar{E}})$
		Continuous	2	$(N_1/N_1, (n-N_1)/N_2)$
	$Q_i = (E_i, \widehat{\eta}_{1i})$	Binary	6	$(N_{1,E}/N_{1,E}, (\lambda n - N_{1,E} - N_{3,E})/N_{2,E}, N_{3,E}/N_{3,E},$ $N_{1,\bar{E}}/N_{1,\bar{E}}, ((1-\lambda)n - N_{1,\bar{E}} - N_{3,\bar{E}})/N_{2,\bar{E}}, N_{3,\bar{E}}/N_{3,\bar{E}})$
		Continuous	2	$(N_1/N_1, (n-N_1)/N_2)$
	$Q_i = (E_i, \widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	Binary	4	$(N_{1,E}/N_{1,E}, (\lambda n - N_{1,E})/N_{2,E},$ $N_{1,\bar{E}}/N_{1,\bar{E}}, ((1-\lambda)n - N_{1,\bar{E}})/N_{2,\bar{E}})$
		Continuous	2	$(N_1/N_1, (n-N_1)/N_2)$
Design 4 - Y|T, E, V	$Q_i = \widehat{a}_{0i}$	-	3	$(N_1/N_1, (n-N_1-N_3)/N_2, N_3/N_3)$
	$Q_i = \widehat{a}_{1i}$	-	3	$(N_1/N_1, (n-N_1-N_3)/N_2, N_3/N_3)$
	$Q_i = (\widehat{a}_{0i}, \widehat{a}_{1i})$	-	2	$(N_1/N_1, (n-N_1)/N_2)$
Design 5 - E, Y|T, V†	$Q_i = (E_i, \widehat{a}_{0i})$	Binary	6	$(N_{1,E}/N_{1,E}, (\lambda n - N_{1,E} - N_{3,E})/N_{2,E}, N_{3,E}/N_{3,E},$ $N_{1,\bar{E}}/N_{1,\bar{E}}, ((1-\lambda)n - N_{1,\bar{E}} - N_{3,\bar{E}})/N_{2,\bar{E}}, N_{3,\bar{E}}/N_{3,\bar{E}})$
		Continuous	2	$(N_1/N_1, (n-N_1)/N_2)$
	$Q_i = (E_i, \widehat{a}_{1i})$	Binary	6	$(N_{1,E}/N_{1,E}, (\lambda n - N_{1,E} - N_{3,E})/N_{2,E}, N_{3,E}/N_{3,E},$ $N_{1,\bar{E}}/N_{1,\bar{E}}, ((1-\lambda)n - N_{1,\bar{E}} - N_{3,\bar{E}})/N_{2,\bar{E}}, N_{3,\bar{E}}/N_{3,\bar{E}})$
		Continuous	2	$(N_1/N_1, (n-N_1)/N_2)$
	$Q_i = (E_i, \widehat{a}_{0i}, \widehat{a}_{1i})$	Binary	4	$(N_{1,E}/N_{1,E}, (\lambda n - N_{1,E})/N_{2,E},$ $N_{1,\bar{E}}/N_{1,\bar{E}}, ((1-\lambda)n - N_{1,\bar{E}})/N_{2,\bar{E}})$
		Continuous	2	$(N_1/N_1, (n-N_1)/N_2)$

*Random sampling assigns a constant probability to all subjects.

-Exposure type does not affect the sample variable and selection probabilities.

†In Designs 3 and 5 with binary exposure, the stratum-specific selection probabilities are presented in the order of $\pi(R^{1,E}), \pi(R^{2,E}), \pi(R^{3,E}), \pi(R^{1,\bar{E}}), \pi(R^{2,\bar{E}}), \pi(R^{3,\bar{E}})$ for bivariate Q_i, and $\pi(R^{1,E}), \pi(R^{2,E}), \pi(R^{1,\bar{E}}), \pi(R^{2,\bar{E}})$ for multivariate Q_i.

Table 4.3: Estimated bias for GxE interaction and joint exposure effects among $G_i = 1$ under the two-way GxE interaction model with a rare exposure and balanced data. Estimates biased by at least 10% in bold.

Sampling	Sampling	UUL		IPWL		CCL		FCL	
scheme	variable	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$
Random sampling	-	0.03	0.03	-	-	-	-	-	-
Design 1 - E	E_i	0.01	0.01	-	-	-	-	0.02	0.01
Design 2 - Y$\|$T	$\widehat{\eta}_{0i}$	**1.23**	**-0.76**	**-0.27**	**-0.32**	0.04	0.02	0.03	0.02
	$\widehat{\eta}_{1i}$	0.06	0.06	0.05	0.07	0.06	0.06	0.06	0.05
	$(\widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	**0.46**	**-0.57**	**-0.29**	**-0.32**	0.01	-0.01	0.01	-0.01
Design 3 - E, Y$\|$T	$(E_i, \widehat{\eta}_{0i})$	**0.71**	**0.63**	0.10	0.09	0.02	**0.27**	0.01	0.02
	$(E_i, \widehat{\eta}_{1i})$	0.04	0.03	0.06	0.04	0.03	0.03	0.04	0.04
	$(E_i, \widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	**0.43**	**0.39**	0.02	0.02	0.03	**0.30**	0.02	0.03
Design 4 - Y$\|$T, E, V	$\widehat{a}_{0i}$	**-0.38**	**-0.24**	**-0.33**	**-0.33**	-0.07	-0.08	-0.04	-0.05
	$\widehat{a}_{1i}$	0.05	0.04	0.05	0.04	0.05	0.04	0.04	0.04
	$(\widehat{a}_{0i}, \widehat{a}_{1i})$	**-0.44**	**-0.33**	**-0.36**	**-0.36**	-0.05	-0.05	-0.03	-0.03
Design 5 - E, Y$\|$T, V	$(E_i, \widehat{a}_{0i})$	**0.69**	**0.62**	0.07	0.08	0.09	0.08	0.01	0.01
	$(E_i, \widehat{a}_{1i})$	0.03	0.03	0.05	0.05	0.03	0.03	0.03	0.03
	$(E_i, \widehat{a}_{0i}, \widehat{a}_{1i})$	**0.43**	**0.40**	0.04	0.05	0.01	0.03	0.02	0.02

Table 4.4: The power (%) of detecting the non-zero genetic and GxE interaction effects under the two-way GxE interaction model with a binary exposure and G-E independence assumption.

Sampling	Sampling	UUL		IPWL		CCL		FCL	
scheme	variable	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$
Random sampling	-	24	76	-	-	-	-	-	-
Design 1 - E	E_i	37	93	-	-	-	-	36	97
Design 2 - Y|T	$\widehat{\eta}_{0i}$	2	87	60	85	36	95	35	95
	$\widehat{\eta}_{1i}$	23	75	33	71	23	74	33	88
	$(\widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	6	94	52	84	36	96	38	97
Design 3 - E, Y|T	$(E_i, \widehat{\eta}_{0i})$	4	38	36	81	15	86	53	100
	$(E_i, \widehat{\eta}_{1i})$	35	92	34	76	33	94	37	97
	$(E_i, \widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	8	55	36	81	21	94	52	100
Design 4 - Y|T, E, V	$\widehat{a}_{0i}$	18	75	54	83	34	97	34	97
	$\widehat{a}_{1i}$	24	75	32	73	24	76	33	89
	$(\widehat{a}_{0i}, \widehat{a}_{1i})$	18	80	53	85	39	97	39	97
Design 5 - E, Y|T, V	$(E_i, \widehat{a}_{0i})$	4	38	36	82	47	97	54	99
	$(E_i, \widehat{a}_{1i})$	35	92	33	77	31	96	38	97
	$(E_i, \widehat{a}_{0i}, \widehat{a}_{1i})$	7	55	37	80	46	99	53	98

Table 4.5: Baseline characteristics of 706 participants in the Normative Aging Study (NAS)

Variable	Mean ± SD, N (percent)
Baseline age (years)	66.3 ± 7.2
Body Mass Index (kg/m^2)	27.9 ± 3.7
Pulse pressure (mmHg)	55.3 ± 15.1
Cumulative patella lead (μg/g)	26.5 [20.8]*
Race (white)	683 (97%)
Education (>12 years)	396 (56%)
Type II diabetes	72 (10%)
Hypertension	447 (63%)
Number of repeated measures on pulse pressure per subject	
1–2	137 (19%)
3–4	221 (31%)
5–6	202 (29%)
7–8	146 (20%)

*Median [interquartile range] for lead exposure whose distribution is right skewed.

4.7 Appendix

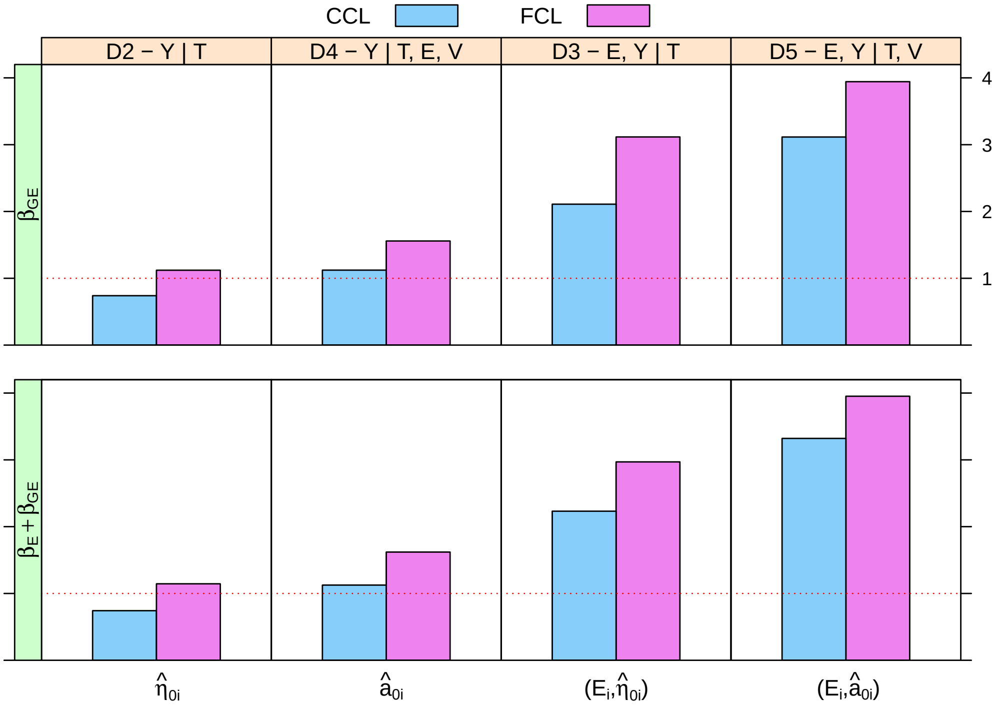

Figure 4.4: Relative efficiency of parameter estimates under the two-way GxE interaction model with a rare exposure and unbalanced data.

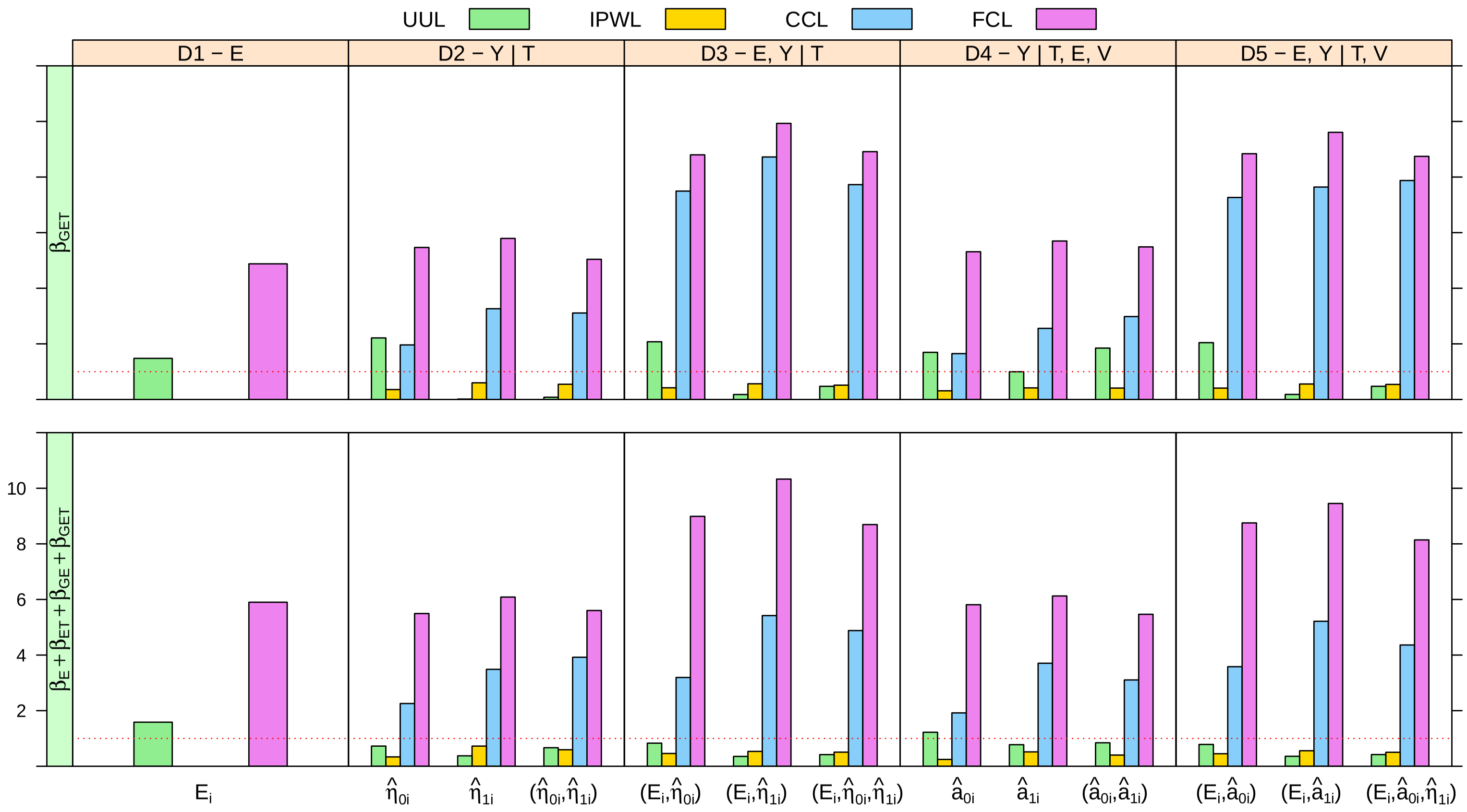

Figure 4.5: Relative efficiency of parameter estimates under the three-way GxExT interaction model with a rare exposure and balanced data.

Table 4.6: Estimated bias for the GxExT interaction and joint exposure effect among $G_i = 1$ under the three-way GxExT interaction model with a rare exposure and balanced data. Estimates biased by at least 10% in bold.

Sampling	Sampling	UUL		IPWL		CCL		FCL	
scheme	variable	GxExT	Joint E	GxExT	Joint E	GxExT	Joint E	GxExT	Joint E
Random sampling	-	-0.02	0.02	-	-	-	-	-	-
Design 1 - E	E_i	-0.01	-0.02	-	-	-	-	0.01	-0.01
Design 2 - Y|T	$\widehat{\eta}_{0i}$	-0.02	-0.41	0.02	-0.27	-0.02	-0.04	0.01	0.01
	$\widehat{\eta}_{1i}$	**3.32**	**0.96**	-0.09	-0.13	0.03	0.01	0.01	-0.01
	$(\widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	**1.36**	**0.52**	-0.07	-0.12	0.01	0.01	0.00	0.01
Design 3 - E, Y|T	$(E_i, \widehat{\eta}_{0i})$	-0.01	0.36	0.03	-0.01	-0.01	0.29	0.01	0.02
	$(E_i, \widehat{\eta}_{1i})$	**0.92**	**1.01**	0.01	0.03	-0.06	0.42	0.01	0.02
	$(E_i, \widehat{\eta}_{0i}, \widehat{\eta}_{1i})$	**0.52**	**0.90**	-0.02	-0.04	-0.07	0.44	0.01	0.02
Design 4 - Y|T, E, V	$\widehat{a}_{0i}$	-0.01	-0.37	0.04	-0.35	-0.02	-0.08	0.02	0.01
	$\widehat{a}_{1i}$	**0.30**	**0.66**	-0.13	-0.17	0.01	0.01	-0.03	0.01
	$(\widehat{a}_{0i}, \widehat{a}_{1i})$	**0.69**	**0.57**	-0.13	-0.19	-0.02	-0.01	-0.01	-0.01
Design 5 - E, Y|T, V	$(E_i, \widehat{a}_{0i})$	-0.02	0.38	0.01	-0.01	-0.01	-0.02	0.03	0.01
	$(E_i, \widehat{a}_{1i})$	**0.91**	**1.01**	-0.02	-0.02	-0.02	0.12	-0.01	0.01
	$(E_i, \widehat{a}_{0i}, \widehat{a}_{1i})$	**0.52**	**0.90**	-0.01	-0.04	-0.01	0.11	-0.02	-0.02

CHAPTER V

Summary

In the first two chapters of this dissertation, we developed study designs that extend Phase I clinical trials into two directions: from single-agent to dual-agent, and from a single endpoint of toxicity to bivariate endpoints of toxicity and efficacy. Specifically, in Chapter II, we developed a nonparametric two-stage adaptive BCD that can be easily implemented for dual-agent Phase I trials. The basic idea of our design was to divide the entire trial into two stages and apply the BCD, with modification, in each stage. We compared the operating characteristics of our design to four competing parametric approaches via simulation in several numerical examples. Under all simulation scenarios we examined, our method performed well in terms of identification of the MTC and allocation of patients relative to the performance of its competitors. Our design inherits the favorable statistical properties of the BCD, is competitive with existing designs, and promotes patient safety by limiting patient exposure to toxic combinations whenever possible. In our design, stopping rule criteria and the distribution of the total sample size among the two stages are context-dependent, and both need careful consideration before adopting our design in practice. Hence, one interesting direction worth exploration is the sample size calculation given availability of asymptotic distribution for dose allocation in the original BCD. As a nonparametric approach, our design has the risk of reduced efficiency in the estimation of DLT rates when compared to parametric approaches. However, considering the small sample size in Phase I trials, we do not need to get precise estimates, instead, we just need to identify the combination with its DLT rate closest to the MTC.

In Chapter III, we extended the BMA-CRM that considers DLT alone and proposed a design for identification of the OBD in Phase I/II trials when the dose-efficacy curve plateaus within the dose range of interest. We incorporated multiple sets of prespecified efficacy probabilities and used BMA to enhance the robustness of our designs to various non-monotonic dose-efficacy curves. During the trial, dose assignment is determined adaptively in two stages, with a first stage that uses adaptive randomization based upon the efficacy probability estimates, and a second stage that uses estimates of the posterior probabilities that each dose is the OBD. We presented both Bayesian and maximum likelihood approaches to estimation. The simulation results demonstrated that our design is able to identify the OBD effectively and allocates patients to doses at and around the OBD frequently when compared to a competing approach designed for non-monotonic dose-efficacy curves. Despite that varying functional forms of the parameterized working model for dose-toxicity and dose-efficacy exist, in our design, we have limited our use to the simple one-parameter power model. If this model assumption may not hold suggested by preclinical data, we can certainly adopt more complex models. It should also be noted that when the underlying true efficacy remains minimal and constant across all dose levels, appropriate design-based tuning parameters need to be carefully calibrated before implementation. A relevant future direction of our design is to consider modifications that can be applied to dual-agent Phase I/II trials in the presence of dose-efficacy plateau. Furthermore, we are pursuing work that can extend our design to accommodate censored outcome due to lagged time in the determination of treatment efficacy.

To investigate GxE interaction in longitudinal cohort studies, in Chapter IV, we proposed exposure enriched outcome trajectory dependent designs that can inform sample selection by leveraging individual exposure and outcome trajectory, and developed a FCL-based analysis that corrects for the biased sampling. We compared the performance of our proposed designs and analysis to combinations of different sampling designs and estimation approaches via simulation. We observed that the FCL provides improved estimates for the GxE interaction and joint exposure effects over uncorrected complete-case analysis, and the exposure enriched outcome trajectory dependent design outperforms other designs in terms of estimation efficiency and detection power for the GxE interaction compared to random selection of subjects. We also illustrated the utility of our designs and analysis in an example from the Normative Aging Study, a longitudinal study of Boston area veterans. In the future, this work can be extended in two promising directions, to explore sampling designs that make use of the time-varying environmental exposure if long-term exposure history is available, and to accommodate non-linear time trend in the longitudinal outcome in the selection of subjects.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahn, J., Mukherjee, B., Gruber, S. B., and Ghosh, M. (2013). Bayesian semiparametric analysis for two-phase studies of gene-environment interaction. *The Annals of Applied Statistics*, **7**(1), 543–569.

Asakawa, T., Hirakawa, A., and Hamada, C. (2014). Bayesian model averaging continual reassessment method for bivariate binary efficacy and toxicity outcomes in phase I oncology trials. *Journal of Biopharmaceutical Statistics*, **24**(2), 310–325.

Aschard, H., Lutz, S., Maus, B., Duell, E. J., Fingerlin, T. E., Chatterjee, N., Kraft, P., and Van Steen, K. (2012). Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Human Genetics*, **131**(10), 1591–1613.

Berenson, J. R., Yellin, O., Patel, R., Duvivier, H., Nassir, Y., Mapes, R., Abaya, C. D., and Swift, R. A. (2009). A phase I study of samarium lexidronam/bortezomib combination therapy for the treatment of relapsed or refractory multiple myeloma. *Clinical Cancer Research*, **15**(3), 1069–1075.

Boks, M., Schipper, M., Schubart, C., Sommer, I., Kahn, R., and Ophoff, R. (2007). Investigating gene-environment interaction in complex diseases: increasing power by selective sampling for environmental exposure. *International Journal of Epidemiology*, **36**(6), 1363–1369.

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

Braun, T. M. (2014). The current design of oncology phase I clinical trials: progressing from algorithms to statistical models. *Chinese Clinical Oncology*, **3**(1), 2–12.

Braun, T. M. and Alonzo, T. A. (2011). Beyond the 3+ 3 method: expanded algorithms for dose-escalation in phase I oncology trials of two agents. *Clinical Trials*, **8**(3), 247–259.

Braun, T. M. and Jia, N. (2013). A generalized continual reassessment method for two-agent phase I trials. *Statistics in Biopharmaceutical Research*, **5**(2), 105–115.

Braun, T. M. and Wang, S. (2010). A hierarchical bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics*, **66**(3), 805–812.

Bril, G., Dykstra, R., and Pillers, C. (1984). Isotonic regression in two independent variables. *Applied Statistics*, **33**(3), 352–357.

Chatterjee, N. and Chen, Y.-H. (2007). Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 123–142.

Chatterjee, N. and Mukherjee, B. (2008). Statistical approaches to studies of gene-gene and gene-environment interaction. *Molecular Epidemiology in Cancer*, pages 145–169.

Chen, J., Kang, G., VanderWeele, T., Zhang, C., and Mukherjee, B. (2012). Efficient designs of gene-environment interaction studies: implications of Hardy-Weinberg equilibrium and gene-environment independence. *Statistics in Medicine*, **31**(22), 2516–2530.

Cheung, K. (2014). dfcrm: Dose-finding by the continual reassessment method. package version 0.2–2.

Cheung, Y. K. (2011). *Dose finding by the continual reassessment method*. CRC Press.

Cheung, Y. K. (2013). Sample size formulae for the bayesian continual reassessment method. *Clinical Trials*, **10**(6), 852–861.

Chevret, S. (2006). *Statistical methods for dose-finding experiments*. Statistics in practice. John Wiley and Sons Ltd.

Clayton, D. and McKeigue, P. M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *The Lancet*, **358**(9290), 1356–1360.

Conaway, M. R., Dunbar, S., and Peddada, S. D. (2004). Designs for single-or multiple-agent phase I trials. *Biometrics*, **60**(3), 661–669.

Cunanan, K. and Koopmeiners, J. S. (2014). Evaluating the performance of copula models in phase I/II clinical trials under model misspecification. *BMC Medical Research Methodology*, **14**(1), 51–61.

Dai, J. Y., Logsdon, B. A., Huang, Y., Hsu, L., Reiner, A. P., Prentice, R. L., and Kooperberg, C. (2012). Simultaneously testing for marginal genetic association and gene-environment interaction. *American Journal of Epidemiology*, **176**(2), 164–173.

Durham, S. D. and Flournoy, N. (1995). Up-and-down designs I: stationary treatment distributions. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, **25**, 139–157.

Durham, S. D., Flournoy, N., and Montazer-Haghighi, A. A. (1995). Up-and-down designs II: exact treatment moments. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, **25**, 158–178.

Durham, S. D., Flournoy, N., and Rosenberger, W. F. (1997). A random walk rule for phase I clinical trials. *Biometrics*, **53**(2), 745–760.

Fan, S. K., Venook, A. P., and Lu, Y. (2009). Design issues in dose-finding phase I trials for combinations of two agents. *Journal of Biopharmaceutical Statistics*, **19**(3), 509–523.

Gandhi, L., Bahleda, R., Tolaney, S. M., Kwak, E. L., Cleary, J. M., Pandya, S. S., Hollebecque, A., Abbas, R., Ananthakrishnan, R., Berkenblit, A., *et al.* (2014). Phase I study of neratinib in combination with temsirolimus in patients with human epidermal growth factor receptor 2–dependent and other solid tumors. *Journal of Clinical Oncology*, **32**(2), 68–75.

Gasparini, M. (2013). General classes of multiple binary regression models in dose finding problems for combination therapies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**(1), 115–133.

Giovagnoli, A. and Pintacuda, N. (1998). Properties of frequency distributions induced by general "up-and-down" methods for estimating quantiles. *Journal of Statistical Planning and Inference*, **74**(1), 51–63.

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

Hamberg, P., Ratain, M. J., Lesaffre, E., and Verweij, J. (2010). Dose-escalation models for combination phase I trials in oncology. *European Journal of Cancer*, **46**(16), 2870–2878.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**(4), 382–401.

Holt, D., Smith, T., and Winter, P. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society: Series A (General)*, **143**(4), 474–487.

Huang, X., Biswas, S., Oki, Y., Issa, J.-P., and Berry, D. A. (2007). A parallel phase I/II clinical trial design for combination therapies. *Biometrics*, **63**(2), 429–436.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Reviews Genetics*, **6**(4), 287–298.

Ivanova, A. (2006). Escalation, group and a+ b designs for dose-finding trials. *Statistics in Medicine*, **25**(21), 3668–3678.

Ivanova, A. and Wang, K. (2004). A non-parametric approach to the design and analysis of two-dimensional dose-finding trials. *Statistics in Medicine*, **23**(12), 1861–1870.

Ivanova, A. and Xiao, C. (2013). Dose finding when the target dose is on a plateau of a dose–response curve: comparison of fully sequential designs. *Pharmaceutical Statistics*, **12**(5), 309–314.

Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G., and D Durham, S. (2003). Improved up-and-down designs for phase I trials. *Statistics in Medicine*, **22**(1), 69–82.

Jain, R. K., Lee, J. J., Hong, D., Markman, M., Gong, J., Naing, A., Wheler, J., and Kurzrock, R. (2010). Phase I oncology studies: evidence that in the era of targeted therapies patients on lower doses do not fare worse. *Clinical Cancer Research*, **16**, 1289–1297.

Jin, I. H., Huo, L., Yin, G., and Yuan, Y. (2015). Phase I trial design for drug combinations with Bayesian model averaging. *Pharmaceutical Statistics*, **14**(2), 108–119.

Ko, Y.-A., Saha-Chaudhuri, P., Park, S. K., Vokonas, P. S., and Mukherjee, B. (2013). Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. *Genetic Epidemiology*, **37**(6), 581–591.

Korn, E. L., Arbuck, S. G., Pluda, J. M., Simon, R., Kaplan, R. S., and Christian, M. C. (2001). Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology*, **19**(1), 265–272.

Kraft, P. and Aschard, H. (2015). Finding the missing gene-environment interactions. *European Journal of Epidemiology*, **30**(5), 353–355.

Kramar, A., Lebecq, A., and Candalh, E. (1999). Continual reassessment methods in phase I trials of the combination of two drugs in oncology. *Statistics in Medicine*, **18**(14), 1849–1864.

Kuzuya, K., Ishikawa, H., Nakanishi, T., Kikkawa, F., Nawa, A., Fujimura, H., Iwase, A., Arii, Y., Kawai, M., Hattori, S.-e., Sakakibara, K., Sasayama, E., Furuhashi, Y., Suzuki, T., and Mizutani, S. (2001). Optimal doses of paclitaxel and carboplatin combination chemotherapy for ovarian cancer: a phase I modified continual reassessment method study. *International Journal of Clinical Oncology*, **6**(6), 271–278.

Lawless, J., Kalbfleisch, J., and Wild, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(2), 413–438.

Lee, S. M. and Cheung, Y. K. (2009). Model calibration in the continual reassessment method. *Clinical Trials*, **6**(3), 227–238.

LeTourneau, C., Dieras, V., Tresca, P., Cacheux, W., and Paoletti, X. (2010). Current challenges for the early clinical development of anticancer drugs in the era of molecularly targeted agents. *Targeted Oncology*, **5**, 65–72.

Leung, D. H.-Y. and Wang, Y.-G. (2002). An extension of the continual reassessment method using decision theory. *Statistics in Medicine*, **21**(1), 51–63.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

LoRusso, P. M., Boerner, S. A., and Seymour, L. (2010). An overview of the optimal planning, design, and conduct of phase I studies of new therapeutics. *Clinical Cancer Research*, **16**, 1710–1718.

Mick, R. and Ratain, M. J. (1993). Model-guided determination of maximum tolerated dose in phase I clinical trials: evidence for increased precision. *Journal of the National Cancer Institute*, **85**(3), 217–223.

Møller, S. (1995). An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Statistics in Medicine*, **14**(9), 911–922.

Moore, S. C., Gunter, M. J., Daniel, C. R., Reddy, K. S., George, P. S., Yurgalevitch, S., Devasenapathy, N., Ramakrishnan, L., Chatterjee, N., Chanock, S. J., *et al.* (2012). Common genetic variants and central adiposity among Asian–Indians. *Obesity*, **20**(9), 1902–1908.

Mukherjee, B., Ko, Y.-A., VanderWeele, T., Roy, A., Park, S. K., and Chen, J. (2012a). Principal interactions analysis for repeated measures data: application to gene–gene and gene–environment interactions. *Statistics in Medicine*, **31**(22), 2531–2551.

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (2012b). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology*, **175**(3), 177–190.

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

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

Paller, C. J., Bradbury, P. A., Ivy, S. P., Seymour, L., LoRusso, P. M., Baker, L., Rubinstein, L., Huang, E., Collyar, D., Groshen, S., *et al.* (2014). Design of phase I combination trials: recommendations of the clinical trial design task force of the NCI investigational drug steering committee. *Clinical Cancer Research*, **20**(16), 4210–4217.

Pan, H., Xie, F., Liu, P., Xia, J., and Ji, Y. (2014). A phaseI/II seamless dose escalation/expansion with adaptive randomization scheme (SEARS). *Clinical Trials*, **11**(1), 49–59.

Paoletti, X. and Kramar, A. (2009). A comparison of model choices for the continual reassessment method in phase I cancer trials. *Statistics in Medicine*, **28**(24), 3012–3028.

Perlstein, T., Weuve, J., Schwartz, J., Sparrow, D., Wright, R., Litonjua, A., Nie, H., and Hu, H. (2007). Cumulative community-level lead exposure and pulse pressure: the normative aging study. *Environmental Health Perspectives*, **115**(12), 1696–1700.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W., and Kahaner, D. K. (1983). *Quadpack: a subroutine package for automatic integration*. Springer Verlag.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**(427), 846–866.

Rosenberger, W. F. and Haines, L. M. (2002). Competing designs for phase I clinical trials: a review. *Statistics in Medicine*, **21**(18), 2757–2770.

Schildcrout, J. S. and Heagerty, P. J. (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, **9**(4), 735–749.

Schildcrout, J. S., Mumford, S. L., Chen, Z., Heagerty, P. J., and Rathouz, P. J. (2012). Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. *Statistics in Medicine*, **31**(22), 2441–2456.

Schildcrout, J. S., Garbett, S. P., and Heagerty, P. J. (2013). Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, **69**(2), 405–416.

Schildcrout, J. S., Rathouz, P. J., Zelnick, L. R., Garbett, S. P., and Heagerty, P. J. (2015). Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *The Annals of Applied Statistics*, **9**(2), 731–753.

Stenzel, S. L., Ahn, J., Boonstra, P. S., Gruber, S. B., and Mukherjee, B. (2015). The impact of exposure-biased sampling designs on detection of gene-environment interactions in case-control studies with potential exposure misclassification. *European Journal of Epidemiology*, **30**(5), 413–423.

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

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

Thall, P. F. and Russell, K. E. (1998). 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.

Thall, P. F., Millikan, R. E., Mueller, P., and Lee, S.-J. (2003). Dose-finding with two agents in phase I oncology trials. *Biometrics*, **59**(3), 487–496.

Thomas, D. (2010). Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, **11**(4), 259–272.

Ting, N. (2006). *Dose finding in drug development*. Springer Science & Business Media.

Wages, N. A. and Tait, C. (2015). Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *Journal of Biopharmaceutical Statistics*, **25**(5), 903–920.

Wages, N. A., Conaway, M. R., and O'Quigley, J. (2011). Continual reassessment method for partial ordering. *Biometrics*, **67**(4), 1555–1563.

Wang, K. and Ivanova, A. (2005). Two-dimensional dose finding in discrete dose space. *Biometrics*, **61**(1), 217–222.

Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, **100**(470), 459–469.

Wei, P., Tang, H., and Li, D. (2014). Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genetic Epidemiology*, **38**(7), 638–651.

Yin, G. and Yuan, Y. (2009a). Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58**(2), 211–224.

Yin, G. and Yuan, Y. (2009b). Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association*, **104**(487), 954–968.

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

Yuan, Y. and Yin, G. (2008). Sequential continual reassessment method for two-dimensional dose finding. *Statistics in Medicine*, **27**(27), 5664–5678.

Zang, Y., Lee, J. J., and Yuan, Y. (2014). Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical Trials*, **11**(3), 319–327.

Zhang, A., Mukherjee, B., Nie, H., Hu, H., Park, S. K., Wright, R. O., Weisskopf, M. G., and Sparrow, D. (2010). HFE H63D polymorphism as a modifier of the effect of cumulative lead exposure on pulse pressure: the normative aging study. *Environmental Health Perspectives*, **118**(9), 1261–1266.

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