Spring 2011

Causal and Design Issues in Clinical Trials

Rongmei Zhang
University of Pennsylvania, rongmei@mail.med.upenn.edu

Causal and Design Issues in Clinical Trials

Abstract

The first part of my dissertation focuses on post-randomization modification of intent-to-treat effects. For example, in the field of behavioral science, investigations involve the estimation of the effects of behavioral interventions on final outcomes for individuals stratified by post-randomization moderators measured during the early stages of the intervention (e.g., landmark analyses in cancer research). Motivated by this, we address several questions on the use of standard and causal approaches to assessing the modification of intent-to-treat effects of a randomized intervention by a post-randomization factor. First, we show analytically the bias of the estimators of the corresponding interaction and meaningful main effects for the standard regression model under different combinations of assumptions. Such results show that the assumption of independence between two factors involved in an interaction, which has been assumed in the literature, is not necessary for unbiased estimation. Then, we present a structural nested distribution model estimated with G-estimation equations, which does not assume that the post-randomization variable is effectively randomized to individuals. We show how to obtain efficient estimators of the parameters of the structural distribution model. Finally, we confirm with simulations the performance of these optimal estimators and further assess our approach with data from a randomized cognitive therapy trial.

The second part of my dissertation is on optimal and adaptive designs for dose-finding experiments in clinical trials with multiple correlated responses. For instance, in phase I/II studies, efficacy and toxicity are often the primary endpoints which are observed simultaneously and need to be evaluated together. Accordingly, we focus on bivariate responses with one continuous and one categorical. We adopt the bivariate probit dose-response model and study locally optimal, two-stage optimal, and fully adaptive designs under different cost constraints. We assess the performance of the different designs through simulations and suggest that the two-stage designs are as efficient as and may be more efficient than the fully adaptive deigns under a moderate sample size in the initial stage. In addition, two-stage designs are easier to construct and implement, and thus can be a useful approach in practice.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Marshall Joffe

Keywords

post-randomization, unmeasured confounding, ordinary least squares, G-estimation, dose-finding designs, bivariate endpoints

Subject Categories

Biostatistics | Clinical Trials | Design of Experiments and Sample Surveys | Statistical Methodology

CAUSAL AND DESIGN ISSUES IN CLINICAL TRIALS

RONGMEI ZHANG

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2011

Supervisor of Dissertation

*Signature*_____

Marshall Joffe

Associate Professor of Biostatistics

Graduate Group Chairperson

*Signature*_____

Daniel Heitjan, Professor of Biostatistics

Dissertation Committee

Thomas Ten Have, Professor of Biostatistics

Valerii Fedorov, Director of Research Statistics Unit, GlaxoSmithKline

Dylan Small, Associate Professor of Statistics

Gregory Brown, Research Associate Professor of Clinical Psychology in Psychiatry

Acknowledgments

First and foremost, I would like to thank my advisor, Thomas Ten Have, for his generous support and constant encouragement, for stimulating discussions and great ideas. He has been always willing to help me in every aspect of my study, research, and professional pursuit. I could not have imagined having a better advisor. I am deeply thankful to Valerii Fedorov, for his guidance and support throughout my PhD study. The rigorous scholarship and broad knowledge Dr. Fedorov has shown make our collaboration a fortune for my future scientific career. I would also like to deeply thank Marshall Joffe, who has inspired me many times in my research, and has always been nice and patient. Without their help and encouragement, I cannot finish my PhD study and this dissertation.

I would like to express my gratitude to other members in my committee, Dylan Small and Gregory Brown, for their nice advices and suggestions to improve my dissertation. I also appreciate Gregory Brown for using his data.

I would like to thank Yuehui Wu for inspired collaborations, and teaching me a lot about optimal designs theory and programming techniques. I would also like to thank Yimei Li and Jichun Xie for their helpful discussions, and Victoria Gamerman

and Matthew Guerra for carefully reading my manuscripts.

It has been a great pleasure studying at the Department of Biostatistics at the University of Pennsylvania. I have learned much from the faculties. I am also thankful to all of my student colleagues for the friendly and supportive environment they provided.

Finally, I would like to thank my husband, Yi Pan. Without his love, support, and encouragement, my accomplishments thus far would have been impossible. I would also like to thank my parents and parents-in-law for their love and understanding all these years.

ABSTRACT

CAUSAL AND DESIGN ISSUES IN CLINICAL TRIALS

Rongmei Zhang

Marshall Joffe

The first part of my dissertation focuses on post-randomization modification of intent-to-treat effects. For example, in the field of behavioral science, investigations involve the estimation of the effects of behavioral interventions on final outcomes for individuals stratified by post-randomization moderators measured during the early stages of the intervention (e.g., landmark analyses in cancer research). Motivated by this, we address several questions on the use of standard and causal approaches to assessing the modification of intent-to-treat effects of a randomized intervention by a post-randomization factor. First, we show analytically the bias of the estimators of the corresponding interaction and meaningful main effects for the standard regression model under different combinations of assumptions. Such results show that the assumption of independence between two factors involved in an interaction, which has been assumed in the literature, is not necessary for unbiased estimation. Then, we present a structural nested distribution model estimated with G-estimation equations, which does not assume that the post-randomization variable is effectively randomized to individuals. We show how to obtain efficient estimators of the parameters of the structural distribution model. Finally, we confirm with simulations the performance of these optimal estimators and further assess our approach with data

from a randomized cognitive therapy trial.

The second part of my dissertation is on optimal and adaptive designs for dose-finding experiments in clinical trials with multiple correlated responses. For instance, in phase I/II studies, efficacy and toxicity are often the primary endpoints which are observed simultaneously and need to be evaluated together. Accordingly, we focus on bivariate responses with one continuous and one categorical. We adopt the bivariate probit dose-response model and study locally optimal, two-stage optimal, and fully adaptive designs under different cost constraints. We assess the performance of the different designs through simulations and suggest that the two-stage designs are as efficient as and may be more efficient than the fully adaptive deigns under a moderate sample size in the initial stage. In addition, two-stage designs are easier to construct and implement, and thus can be a useful approach in practice.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Post-randomization analyses and causal inference

In clinical trials, randomization is the gold standard for evaluating the efficacy and effectiveness of intervention. Because of the random assignment to the treatment and control groups, the covariates across the groups are balanced. Therefore, the differences between the outcomes for the treatment and control groups can be attributed to the treatment, i.e., treatment is the cause. Post-randomization variables are measured after the treatment assignment and before the assessment of final outcomes of interest. These variables can potentially modify the effect of treatment assignment of the outcome.

Recently, there have been interest for assessing how post-randomization variables modify the effects of the randomized intervention on the outcome. For example, in the

field of behavioral science, investigations involve the estimation of the effects of behavioral interventions on final outcomes for individuals stratified by post-randomization moderators measured during the early stages of the intervention. In the area of cancer, there are landmark analyses where the effect of baseline randomized cancer treatments on an endpoint outcome such as survival or a reduction of continuous measures of cancer severity is stratified by an early response to the treatment in terms of a non-mortality measure of cancer severity (e.g.,Normand, 2007). More recently, in randomized HIV therapeutic vaccination trials of patients on anti-retroviral therapy, the focus is on the effect of the vaccine therapy on reducing viral load among those who stop anti-retroviral therapy after randomization to the therapy or control condition (Mogg et al. 2010).

In some biomedical sciences, intermediate post-randomization variables are often called mediators or surrogate markers. Information on post-randomization factors are often used to evaluate direct versus indirect effects. In a randomized trial comparing the effect of high-dose vs. low-dose 3-azido-3-deoxythymidine (AZT) for patients with HIV disease, subjects in the high-dose AZT group are less likely to receive prophylaxis therapy for Pneumocystis Carinii Pneumonia(PCP), a post-randomization treatment. A direct effect is the effect of AZT holding the same level of post-randomization treatment to a given level, e.g., receiving PCP prophylaxis therapy, and an indirect effect is the part of the effect of AZT mediated by PCP (Robins and Greenland, 1994).

Motivated by the wide interest of the post-randomization analyses in clinical trials,

the goal of the first part of my dissertation is to investigate the use of standard and causal approaches to assessing the modification of intent-to-treat effects of a randomized intervention by a post-randomization factor.

Standard approaches such as standard regression models with Ordinary Least Squares (OLS) are commonly used in post-randomization mediation or moderation analyses (Baron and Kenny, 1986). A crucial assumption is that there are no unmeasured confounders for the randomized intervention on post-randomization variables, i.e., a sequential ignorability assumption in causal inference. The no unmeasured confounding assumption holds for the randomized intervention due to the random assignment. However, post-randomization variables are usually not randomized and affected by treatment assignment. Therefore, there are often unmeasured confounders for post-randomization variable, and thus it introduces bias when we estimate the joint and main effects of the randomized intervention and the post-randomization variables.

Because of the vulnerability of standard approaches, a number of causal approaches have been developed based on the potential outcomes framework (Rubin, 1974). One popular causal approach is Principal Stratification (PS)(Frangakis and Rubin, 2002), where we stratify individuals into latent classes: compliers, defiers, never-takers and, always-takers. Robins (1992, 1994, 1998, 1999) developed a number of innovative approaches, which include Structural Nested Models (SNM) with G-estimation.

Chapter 2 of this dissertation focuses on the use of standard approaches in post-

randomization modification analyses. Based on the standard linear regression model with main effects of randomized intervention, post-randomization moderator, and their interaction, we show analytically the bias of the estimators of the corresponding interaction and meaningful main effects under different combinations of assumptions. Such assumptions involve the equality of the unmeasured confounding between the randomized intervention groups and the absence of an effect of the randomized intervention on the post-randomization moderator. We show that even in the presence of the unmeasured confounding for post-randomization moderator, less stringent assumptions are sufficient for unbiased OLS estimation of the randomized intervention effect conditional on the post-randomization moderator under the linear interaction model. In addition, our results show that the assumption of independence between two factors involved in an interaction, which has been assumed in the literature, is not necessary for unbiased estimation.

In Chapter 3, we present a Structural Nested Distribution Model(SNDM) estimated with G-estimation. Our causal approach does not assume that the post-randomization variable is effectively randomized to individuals (i.e., sequential ignorability). Under the working assumption of sequential ignorability and when sequential ignorability does not hold, we show how to obtain efficient estimators of the parameters of the SNDM. We use simulations to examine and verify the performance of these optimal estimators. The working assumption of sequential ignorability leads to simpler estimators, which are efficient under that assumption, but the more complex estimator not assuming sequential ignorability can substantially improve efficiency

when sequential ignorability does not hold. We use data from a randomized cognitive therapy trial to illustrate and further assess our approach.

1.2 Dose-finding experiments and optimal designs

The second part of my dissertation is on optimal and adaptive designs for dose-finding experiments. The primary goal in dose-finding studies is to establish the dose-response relationship or to find the target dose, e.g. Maximum Tolerated Dose(MTD). A number of statistical designs have been proposed and studied in dose-finding experiments. From the statistical view, we can divide them into two classes: non-parametric and parametric approaches. The traditional 3+3 design and Group-Up-and-Down design (Ivanova 2004, Gezmu and Flournoy, 2006) are often referred to as non-parametric designs, where the patient assignments are based on specific decision rules. They are usually intuitive and do not involve complicated calculations. Therefore, non-parametric designs are attractive because they are easy to understand and implement by a practitioner. However, non-parametric designs may require too many escalations to reach the target dose and get non-robust estimate for the target dose. For parametric designs, under a parametric framework, we specify a model for the dose-response relationship and estimate the unknown parameters. Therefore, parametric designs are also called model-based approaches, which include designs such as Best-intention designs (Wu, 1985), Continual Reassessment Method (CRM) (O'Quigley et al., 1990) and Optimal experimental design (Kiefer 1959, Fedorov,

1972). In Best intention design, researchers estimate the unknown parameters and the target dose in each step. The next patient will be assigned at the estimated target dose. In CRM, the parameters in the response model are continually updated and some dose escalation rules are applied. Bayes theorem is usually used in the CRM. Optimal designs are a class of designs where the optimality of the design depends on the statistical model and is assessed with respect to some statistical criteria. These statistical criteria are related to the variance-covariance matrix of the unknown parameters for the model.

In this dissertation, we focus on the study of Optimal designs because of three major advantages. First, optimal designs are mathematically rigorous and theoretically efficient. Second, they can be optimized under constraints, such as ethical and cost concerns, which are very important in dose-finding experiments. Third, optimal designs can accommodate multiple type of factors involving in the experiments (Atkinson, Donev, and Tobias, 2007).

The history of optimal designs can be traced back to the early twentieth century in a paper by Smith (1918). The core of the theory of optimal experimental design was developed during the fifties to the seventies. Kiefer (1959, 1960) contributed significantly to the development of optimal design theory. The first comprehensive book on the theory of optimal designs was written by Fedorov (1972). Silvey (1980) wrote a compact book to introduce the theory of optimal design, especially for linear models. Another introductory and popular book of the theory of optimal design was written by Atkinson and Donev (1992). Although optimal designs have a long

history, they are not very commonly used in the dose-finding experiments due to their complicated nature. The goal of the second part of my dissertation is to use optimal design framework to build an efficient design in dose-finding experiments, particularly, for those with correlated multiple responses.

In clinical trials, it is common that multiple endpoints are of interest. For instance, in phase I/II studies, efficacy and toxicity are often the primary endpoints which are observed simultaneously and need to be evaluated together. Motivated by this, we confine ourselves to bivariate responses and focus on the most analytically difficult case: a mixture of continuous and categorical responses. In Chapter 4, We show how to adopt the bivariate probit dose-response model to the use of mixutre responses and how to quantify the study goal by a utility function. Locally optimal designs, two-stage optimal designs, and fully adaptive designs are studied under different ethical and cost constraints in the experiments. We assess the performance of two-stage designs and fully adaptive designs via simulations. Our simulations suggest that the two-stage designs are as efficient as and may be more efficient than the fully adaptive designs if there is a moderate sample size in the initial stage. In addition, two-stage designs are easier to construct and implement, and thus can be a useful approach in practice.

Chapter 5 is the appendices, which include the technical proofs in previous chapters.

Chapter 2

Post-randomization Interaction Analyses in Clinical Trials with Standard Regression

2.1 Introduction

In the context of understanding how post-randomization moderators impact randomized interventions in randomized trials, this paper shows that even with unmeasured confounding of post-randomization moderators and outcome, Ordinary Least Squares Regression (OLSR) under a linear interaction model still leads to unbiased estimators of important effects involving the randomized intervention and post-randomization moderators if other less restrictive assumptions are hold. While these results pertain to any investigation of the interaction between the randomized factor

and post-randomization moderators on subsequent outcomes, the results are particularly important for assessing how post-randomization moderators measured early in a randomized intervention modify the effects of the randomized intervention later in or after the intervention. That is, we are interested in estimating the main effect of the randomized intervention on outcome that is modified by factors measured after randomization but early in the intervention. Such investigations are performed in such areas as cancer, HIV, and psychiatry. In the area of cancer, landmark analyses are done, where the effect of baseline randomized cancer treatments on an endpoint outcome such as survival or a reduction of continuous measures of cancer severity is stratified by an early response to the treatment in terms of a non-mortality measure of cancer severity (e.g., Normand, 2007). In randomized HIV prevention vaccination trials of participants at high risk for HIV, the focus is on the effect of the vaccine therapy on reducing viral load among those who become infected (The rgp120 HIV Vaccine Study Group, 2005; Gilbert, Bosch and Hudgens, 2003; Shepherd et al., 2006; Jemiai et al., 2007). More recently, in randomized HIV therapeutic vaccination trials of patients on anti-retroviral therapy, the focus is on the effect of the vaccine therapy on reducing viral load among those who stop anti-retroviral therapy after randomization to the therapy or control condition (Mogg et al. 2010). Finally, in the field of behavioral science, investigations of one aspect of "personalized medicine" involve the estimation of the effects of complex behavioral interventions on final endpoint outcomes for individuals stratified by post-randomization moderators measured during the early stages of the intervention (Faerber et al., 2010). Such early intervention

factors involve "common treatment factor" underlying personality (Lambert et al., 2003).

In addition to the scientific reasons for assessing post-randomization moderation, testing such an interaction has been proposed for the mediation context (see e.g., Vansteelandt 2009). More specifically, the strategy for modeling whether the intervention operates through the mediator or through other unmeasured factors depends on the presence or absence of the the interaction between the intervention and mediator. Consequently, the results of this paper pertain to both the scientific questions above and the methodological strategy for mediation analysis.

A number of causal approaches have been proposed for estimating randomized treatment effects on outcome stratified on a post-randomization moderator. Frangakis and Rubin (2002) proposed a general causal strategy called Principal Stratification (PS) where the post-randomization stratification factor is expressed as a combination of potential outcomes. Mogg et al. (2010) implemented such an approach for the HIV therapeutic context. Others have relied on less parametric approaches for stratifying on potential outcome variables for the post-randomization moderator (Gilbert, Bosch and Hudgens, 2003; Shepherd et al. 2006; Jemiai et al. 2007). Finally, Joffe, Small, and Hsu (2007) specify causal interaction models without stratifying on the potential outcomes of the post-randomization moderators, but still controlling for unmeasured confounding using weighted G-estimation techniques.

These causal approaches have been proposed because of the vulnerability of Ordinary Least Square (OLS) estimation of the linear interaction model to unmeasured

confounding of the post-randomization moderator vs. outcome relationship. However, we show that even in the presence of such confounding, less stringent assumptions are sufficient for unbiased OLS estimation of the randomized intervention effect conditional on the post-randomization moderator under the linear interaction model. Some of these less stringent assumptions are assessable with the observed data and thus preferable to work with than the no unmeasured confounding assumption.

To address these assumptions, we parameterize the linear interaction model such that the main effect for the randomized intervention represents the effect of this intervention on outcome for a given level of the post-randomization moderator, which can be binary or continuous. The interaction then represents the change in the effect of the randomized intervention on outcome given a change in level of the post-randomization moderator. The main effect for the post-randomization moderator and corresponding change represented by the interaction term are interpreted similarly with the roles of the randomized intervention and post-randomization moderator reversed. Such a parameterization is implemented in Joffe, Small, and Hsu (2007), but estimated causally using G-estimation.

Under this parameterization of the linear interaction model, less rigorous assumptions than no unmeasured confounding are sufficient for unbiased OLS estimation of the main effect for the randomized intervention and its interaction with the post-randomization moderator. Such assumptions involve the equality of the unmeasured confounding between the randomized intervention groups and the absence of an effect of the randomized intervention on the post-randomization moderator. However,

if the post-randomization moderator is related to baseline covariates, these assumptions are not sufficient, in which case the assumptions of no unmeasured confounding of both the post-randomization moderator vs. outcome and the baseline covariates vs. outcome relationships are required for unbiased OLS estimation of all parameters of the linear interaction model.

The remainder of the paper is organized as follows. In Section 2, we present the linear interaction models and analytic bias results for OLS estimation of the model parameters under different sets of assumptions. To confirm the analytic results, we conduct simulations under different assumptions in Section 3. In Section 4, we use the randomized Cognitive Therapy (CT) trial as an illustrative example. The paper concludes in Section 5 with a summary and discussion.

2.2 Statistical Models and Bias Results

To present bias results for OLS estimation of the linear interaction model, we present two versions of this model: 1) a "true" model that explicitly adjusts for an unmeasured confounder; and 2) an "analysis" model that does not adjust for the unmeasured confounder. We evaluate the bias of the OLS estimators of the parameters of the analysis model in terms of the parameters of the true model. By adjusting for the unmeasured confounder, the true model yields causal effects of the randomization intervention for a given level of the post-randomization moderator by adjusting for an unmeasured confounder. These results are presented with and

without effects for observed baseline covariates.

2.2.1 Notation

First, we define notation that is summarized graphically in Figure 2.1. The dichotomous randomized intervention is defined as R where $R = 1$ for assignment to the intervention group and 0 for assignment to the comparison group. We assume equal probability for each subject being assigned to the treatment and control groups. If this were not the case, the probability of randomization would explicitly appear in the formula, although the conclusions would remain the same. The continuous outcome is defined as Y, and the continuous or dichotomous post-randomization moderator is M. The vector of baseline covariates as $\tilde{X} = (X_1, \cdots, X_p)^T$. Finally, U represents the unmeasured confounder for post-randomization moderator M on the outcome Y. Note that all variables are defined for subject i, while we suppress the index i to simplify the notation.

2.2.2 Linear interaction model with unmeasured confounder

The targets of estimation are the parameters of the true linear interaction model that adjusts for the unmeasured confounder (U) of the outcome vs. post-randomization relationship. Figure 2.1 provides relationships among the variables underlying the true linear interaction model. These relationships characterized by the arrows are composed of those that explicitly correspond to parameters of the linear interaction model and those arrows that do not. The latter arrows represent relationships among

the randomized intervention (R), the post-randomization moderator (M), and the unmeasured covariate (U). In analyzing the bias of OLS estimation of the analysis model, we consider different sets of assumptions involving the arrows in Figure 2.1 for these three variables.

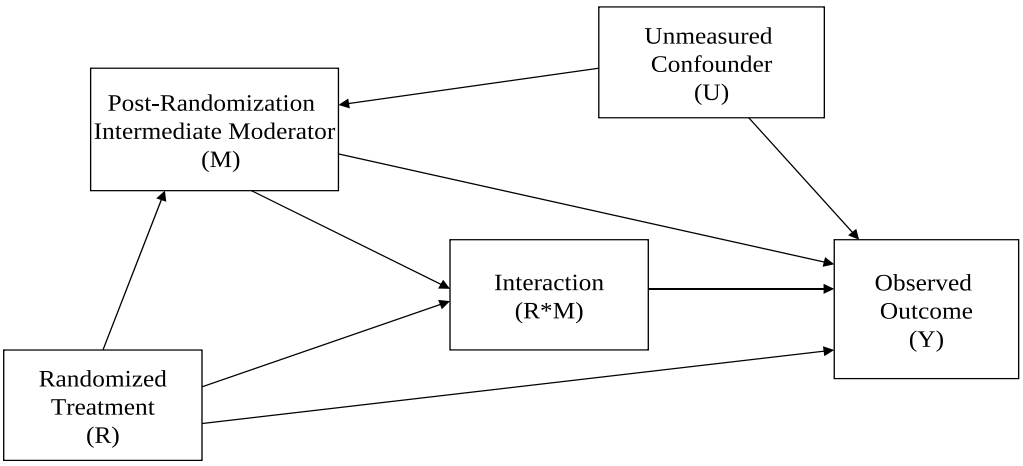

Figure 2.1: A graph example of the moderation analysis.

The true linear interaction model is now defined as:

$$Y = \theta_0 + \theta_R R + \theta_M M + \theta_{RM} RM + \theta_u U + \omega, \tag{2.2.1}$$

where ω is the random error. As the true model in which we account for all unmeasured confounding with U, it follows that ω is independent of all covariates in the model, i.e., $E(\omega|R, M, U) = 0$. Accordingly, the θ parameters can be interpreted in terms of as causal effects. Specifically, θ_R is the effect on the outcome due to assigning the randomized intervention to a patient relative to assigning the same patient to

the comparison group given the patient's post-randomization moderator is observed to be zero ($M = 0$). Similarly, θ_M represents the effect on Y due to changing the post-randomization moderator by a unit for a given patient assigned to the randomized comparison group ($R = 0$). Finally, θ_{RM} represents the interaction effect, and can be interpreted as the change in the effect of R on Y due to a unit change in M for an individual patient; or the change in the effect of M on Y when switching a patient's randomized intervention assignment from the comparison group ($R = 0$) to the intervention ($R = 1$). We note that this parameterization is equivalent to the one under a causal structural mean model provided by Joffe, Small, and Hsu (2007). Other parameterizations of the linear interaction model provide main effect parameters for R and M with different interpretations than the one we present. However, these parameterizations are all linearly dependent such that the model fit is the same.

Causal interpretation of the θ parameters in (2.2.1) requires the stable unit treatment value assumption (SUTVA) that applies to all estimation methods (Angrist, Imbens, and Rubin, 1996), regardless of the additional unique assumptions these methods make to yield such causal intepretations. It requires that a patient's treatment assignment and post-randomization moderator determination are not impacted by the assignments and determinations for other patients. Additionally, there is a single outcome for a given pair of levels for the intervention assignment and determination of the post-randomization moderator, regardless of the method of administration of the randomized intervention or the determination of the post-randomization moderator.

2.2.3 Analysis model without unmeasured confounder

To estimate the θ parameters in the true model in (2.2.1), we propose OLS estimation of the following analysis model:

$$Y = \beta_0 + \beta_R R + \beta_M M + \beta_{RM} RM + \epsilon, \tag{2.2.2}$$

where $\epsilon = \theta_U U + \omega$ is the random error composed of the error of the true model, ω, and the unmeasured confounder, U. Consequently, unbiased estimation of the causal θ parameters in the true model in (2.2.1) with OLS estimation of the β parameters in the analysis model in (2.2.2) is not ensured due to not adjusting explicitly for U in the analysis model. We now examine the assumptions that are necessary for ensuring such unbiased estimation of each of the θ parameters in (2.2.1).

2.2.4 Assumptions for unbiased estimation with analysis model

For unbiased estimation of the θ parameters in the true model in (2.2.1) with the estimators of the β parameters in the analysis model in (2.2.2), we start with the following assumptions:

(A1) Random error ϵ's are independent;

(A2) No unmeasured confounding for both randomized intervention and moderator on the outcome, i.e., $E(\epsilon|R, M) = 0$, no arrow from U to M or R in Figure 2.1;

(A3) Finite variance of random error, i.e., $\text{Var}(\epsilon|R, M) = \sigma$ is finite.

Note that Assumption (A2) is a sequential ignorability assumption, which is im-

plied by the following stronger sequential ignorability assumption:

$$\text{(A2.1)} \quad R \perp Y^{rm} \quad \text{and} \quad \text{(A2.2)} \quad M \perp Y^{rm}|R.$$

Here Y^{rm} is the potential outcome (Rubin 1974; Neyman 1990), representing the outcome that would be observed if a subject was assigned to treatment level r with post-randomization moderator level m. We note that the causal θ parameters in the true model (2.2.1) can be expressed in terms of these potential outcomes.

Under Assumptions A1, A2, and A3, the OLS estimators of all β parameters in the analysis model in (2.2.2) are unbiased with respect to the corresponding θ parameters in the true model in (2.2.1).

We now consider the bias of the OLS estimators of the β parameters with respect to the true model θ parameters when we relax the sequential ignorability assumption (Assumption A2) by not assuming that M is independent of U (arrow from U to M in Figure 2.1). Randomization guarantees the independence between R and U, so that $E(\epsilon|R) = 0$, which we refer to the baseline randomization assumption (Assumption A2.1).

Such bias is presented in term of asymptotic bias as defined by the bias that exists after taking the limit in probability of sample moments for M and ϵ in the function for finite bias. We take this approach rather than obtaining the expectation of finite bias and then taking the limit of terms with increasing sample size, because the probability limits under the former approach offer informative limits representing moments of M given ϵ. This is not feasible with the latter approach of obtaining the expectation with respect to Y and M, an then taking the limit. Accordingly,

we express bias as $\text{Bias}(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \theta)$ as a function of M and ϵ plus notation for convergence in probability, $o_p(1)$.

Theorem 1. Under $P(R = 1) = P(R = 0)$=1/2 as in a randomized trial with 1:1 randomization, and (A1) and (A3) with $E(\epsilon|R) = 0$ (A2.2), the asymptotic bias of the OLS estimators of the β parameters in the analysis model in (2.2.2) relative to the θ parameters in the true model is

$$\text{Bias}\begin{pmatrix}\hat{\beta}_R\\ \hat{\beta}_M\\ \hat{\beta}_{RM}\end{pmatrix} = \begin{pmatrix} -\frac{E(M|R=1)\ E(M\epsilon|R=1)}{\text{Var}(M|R=1)} + \frac{E(M|R=0)\ E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \\ \frac{E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \\ \frac{E(M\epsilon|R=1)}{\text{Var}(M|R=1)} - \frac{E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \end{pmatrix} + \boldsymbol{o_p(1)} \quad (2.2.3)$$

where $\boldsymbol{o_p(1)} = (o_p(1), o_p(1), o_p(1))^T$.

A proof of Theorem 1 is shown in the appendix. Here, we present results for $\hat{\beta}_R, \hat{\beta}_M$, and $\hat{\beta}_{RM})^T$ but not $\hat{\beta}_0$, because this estimator of the intercept is not of interest in the analysis.

2.2.5 Alternate assumptions for unbiased estimation with analysis model

We now consider the assumptions beyond Assumptions A1, A2.1, and A3, that are necessary under which the asymptotic bias in (2.2.3) of the estimators β with respect to the θ parameters converges to zero. Accordingly, the bias equations in (2.2.3) show that the following equalities impact the asymptotic bias for the OLS estimators of

the β parameters with respect to the θ parameters:

(A2.2) $E(M\epsilon|R=1) = E(M\epsilon|R=0) = 0$ (A4) $E(M|R=1) = E(M|R=0)$

(A5) $E(M\epsilon|R=1) = E(M\epsilon|R=0)$ (A6) $\text{Var}(M|R=1) = \text{Var}(M|R=0)$

Assumption A2.2 is no unmeasured confounding for M and Y. Assumption A4 is no correlation between R and M, i.e., no ITT effect of R on M. Assumption A5 specifies that the confounding for the M vs. Y relationship is the same between randomized groups. Assumption A6 is no difference in the variance of M between randomized groups.

Given that Assumptions A1, A2.l, and A3 hold, different combinations of Assumptions (A2.2), (A4)-(A6) lead to the following results:

1. Assumption A2.2 implies no asymptotic bias for all three estimators, $\hat{\beta}_R$, $\hat{\beta}_M$, and $\hat{\beta}_{RM}$.

2. Assumptions A4, A5, and A6 imply no asymptotic bias for $\hat{\beta}_R$ and $\hat{\beta}_{RM}$, but not for $\hat{\beta}_M$.

3. Assumptions A5 and A6 imply no asymptotic bias only for $\hat{\beta}_{RM}$.

In summary, given that Assumptions A1, A2.1, and A3 hold:

1. Unbiased $\hat{\beta}_{RM}$ requires the least restrictive additional assumptions involving the equal confounding and equal variance of in the randomized groups (A5 and A6).

2. Unbiased $\hat{\beta}_R$ requires an additional assumption involving equal mean of M in the randomized groups in addition to the assumptions for $\hat{\beta}_{RM}$ to be unbiased (A4, A5, and A6).

3. Unbiased $\hat{\beta}_M$ requires no unmeasured confounding of the post-randomization vs. outcome relationship (A2.2).

We note it is possible to assess with observed data Assumptions A4 and A6, which entail estimating and testing the intervention-specific mean and variance of the post-randomization moderator. While we cannot assess with observed data Assumption A5 (equal confounding in the intervention groups), we can substitute observed confounders (e.g. baseline covariates) at least assess how much the three-way relationships of the baseline covariates with the post-randomization moderator and outcome differ between the randomized intervention groups. With less restrictive assumptions for unobserved confounder and relaxing the no confounding assumption A2.2 does provide a benefit some of which can be assessed from the observed data for the estimators of the main effect for the randomized intervention and the interaction ($\hat{\beta}_R$ and $\hat{\beta}_{RM}$).

Relationship between distribution properties and assumptions for unbiased estimation

We now consider the distributional properties for the joint distribution of R, M, and ϵ that lead to the above assumptions for unbiased estimation of θ_R and θ_{RM}. These properties involve the symmetry of this joint distribution. We begin with

Lemma 1 of Theorem 1 and proceed to Theorem 2, which addresses the relationships between the bias of the OLS estimators and the factors in Assumptions A2.2, A4, A5, and A6.

Lemma 1. Let $f(r, m, \epsilon)$ be the probability density function for the joint distribution of R, M, and ϵ. If $f(r, m, \epsilon) = f(1 - r, -m, -\epsilon)$, then Assumptions (A5) and (A6) hold, i.e., $E(M\epsilon|R = 1) = E(M\epsilon|R = 0)$ and $\text{Var}(M|R = 1) = \text{Var}(M|R = 0)$.

Theorem 2. Under $P(R = 1) = P(R = 0)$=1/2, $E(\epsilon|R) = 0$, if $f(r, m, \epsilon) = f(1 - r, -m, -\epsilon)$, the asymptotic bias of the OLS estimator in (2.2.2) is: $\text{Bias}(\hat{\beta}_R) = -4\frac{E(RM)E(M\epsilon)}{\text{Var}(M)} + o_p(1)$, $\text{Bias}(\hat{\beta}_M) = \frac{E(M\epsilon)}{\text{Var}(M)} + o_p(1)$, and $\text{Bias}(\hat{\beta}_{RM}) = o_p(1)$.

Proofs for Lemma 1 and Theorem 2 are shown in the appendix.

Theorem 2 tells us that when the joint symmetric condition holds, there is no asymptotic bias of the estimator of the interaction term even without sequential ignorability or correlation between R and M. In addition, under the symmetric distribution, the magnitude of asymptotic bias of the estimator of the main effect of R is proportional to the strength of the unmeasured confounding for M and Y, proportional to the strength of correlation between R and M, and inversely proportional to the variability of M. The magnitude and direction of the relationship between the asymptotic bias of two main effects depend only on the correlation between R and M. If $E(M) = 0$ (or M is centered by its means), without loss of generality, the direction of the relationship between the asymptotic bias of the two main effect estimators is inversely related to the direction of the correlation between R and M. The

impact of Assumptions A2.2, A4 to A6 and the symmetry of the joint distribution is summarized into Table 2.1.

Table 2.1: Impact of conditions for asymptotic bias of OLS estimator in the non-covariate model.

Assumptions[a]						
(A2.2)	(A4)	(A5)	(A6)	$Bias(\hat{\beta}_R)$	$Bias(\hat{\beta}_M)$	$Bias(\hat{\beta}_{RM})$
$\surd$	-	-	-	$o_p(1)$	$o_p(1)$	$o_p(1)$
$\times$	$\surd$	$\surd$	$\surd$	$o_p(1)$	$\frac{E(M\epsilon)}{Var(M)}$	$o_p(1)$
$\times$	$\times$	$\surd$	$\surd$	$\frac{[E(M\|R=0)-E(M\|R=1]\ E(M\epsilon)}{Var(M)}$	$\frac{E(M\epsilon)}{Var(M)}$	$o_p(1)$
Symmetric distribution[b]						
$\times$	$\times$	$\surd$	$\surd$	-4$\frac{E(RM)E(M\epsilon)}{Var(M)}$	$\frac{E(M\epsilon)}{Var(M)}$	$o_p(1)$

NOTE: (a) Assumptions are (can also be found in Section 2.1.1) :

(A2.2) $E(M\epsilon|R=1) = E(M\epsilon|R=0) = 0$, (A4) $E(M|R=1) = E(M|R=0)$,

(A5) $E(M\epsilon|R=1) = E(M\epsilon|R=0)$, (A6)$\text{Var}(M|R=1) = \text{Var}(M|R=0)$.

(b) The assumption of "Symmetric distribution" is $f(r,m,\epsilon) = f(1-r,-m,-\epsilon)$.

Bias relationships between estimators of R and $R*M$ effects

Because the main effect for the randomized intervention and its interaction with the post-randomization moderator are of clinical interest, we compare the asymptotic bias of the estimators of these two parameters The assumptions A4 and A6, which impact the comparison of bias between these the estimator of the R and $R*M$ parameters, can be assessed with observed data. However, the difference between randomization groups with respect to the confounding of the M on Y relationship (Assumption A5) cannot be assessed with observed data. Consequently, we assess how the bias equations in (2.2.3) for R and $R*M$ change with respect to the ratio,

$E(M\epsilon|R=1)/E(M\epsilon|R=0)=\alpha$:

$$
\begin{aligned}
& [\text{Bias}(\hat{\beta}_{RM})]^2 - [\text{Bias}(\hat{\beta}_R)]^2 \\
= \; & \left[\alpha - \frac{1+E(M|R=0)}{1+E(M|R=1)} \frac{\text{Var}(M|R=1)}{\text{Var}(M|R=0)}\right] \left[\alpha - \frac{1-E(M|R=0)}{1-E(M|R=1)} \frac{\text{Var}(M|R=1)}{\text{Var}(M|R=0)}\right] \\
& \cdot E^2(M\epsilon|R=0)\text{Var}^2(M|R=1)\frac{1}{E^2(M|R=1)} + o_p(1). \qquad (2.2.4)
\end{aligned}
$$

Given that moments for M conditional on R are estimable from the data, the function in (2.2.4) is a quadratic function of α, for which the solution can be inverted to produce bounds for the ratio $\alpha = E(M\epsilon|R=1)/E(M\epsilon|R=0)$. One can then compare resulting bounds for the separate asymptotic bias function for the estimators of R and $R*M$. In deriving such bounds, we provide an example (see Section 4. Data Analysis) for why it is informative to compare the bias between the estimators of R and $R*M$ effects in practice.

2.2.6 Model with baseline covariates and bias of OLS Estimator

In this section, we extend the previous results to accommodate baseline covariates under different conditions depending on combinations of associations between these covariates and the post-randomization moderator, the confounding of the covariate vs. outcome, and post-randomization moderator vs. outcome relationships. In Figure 2.2, these relationships correspond to the arrows between X and M; between X and U, and M and U, respectively. These additional arrows involving X result in additional sources of bias, which are presented analytically in this section.

Accordingly, we start with the true model with the baseline covariates:

$$Y = \theta_0 + \theta_R R + \theta_M M + \theta_{RM} RM + \delta_1 X_1 + \cdots + \delta_p X_p + \theta_u U + \omega, \tag{2.2.5}$$

where ω is the random error with $E(\omega|R, M, X_1, \cdots, X_p, U) = 0$. The effects of θ_R, θ_M, and θ_{RM} have the same interpretations as in the case without covariates in (2.2.2) except that the interpretations are now conditional on the observed values of baseline covariates. In addition, we assume that the effects of θ_R, θ_M, and θ_{RM} do not vary with baseline covariates.

As with the θ effects in the true model without baseline covariates relative to arrows in Figure 2.1, the θ and δ effects correspond to analagous arrows in Figure 2.2 with the baseline covariates added.

Adding baseline covarites to the analysis model in (2.2.2) results in the following analysis model:

$$Y = \beta_0 + \beta_R R + \beta_M M + \beta_{RM} RM + \gamma_1 X_1 + \cdots + \gamma_p X_p + \epsilon, \tag{2.2.6}$$

where $\epsilon = \theta_U U + \omega$ is the random error composed of the error of the true model, ω, and the unmeasured confounder, U.

For simplicity of notation, we rewrite the true model (2.2.5) and the analysis model (2.2.6) with matrix and vector expression as follows respectively:

$$\mathrm{Y} = \mathrm{Z}\boldsymbol{\theta} + \mathrm{X}\boldsymbol{\delta} + \boldsymbol{U}\theta_u + \boldsymbol{\omega}, \tag{2.2.7}$$

and

$$\mathrm{Y} = \mathrm{Z}\boldsymbol{\beta} + \mathrm{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{2.2.8}$$

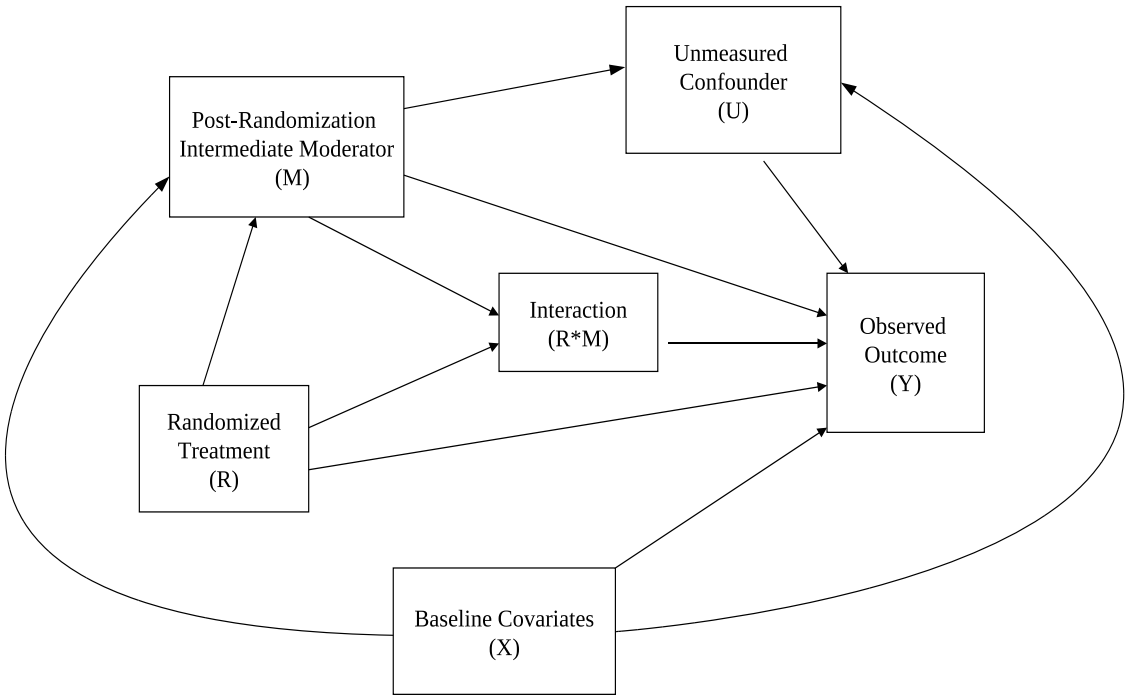

Figure 2.2: A graph example of the moderation analysis with baseline covariates.

where $\mathrm{Z}_{n\times 4}$ is the design matrix for R, M, and RM, $\mathrm{X}_{n\times p}$ is the design matrix for $\tilde{X} = (X_1, \cdots, X_p)^T$, $\boldsymbol{\theta} = (\theta_0, \theta_R, \theta_M, \theta_{RM})^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_R, \beta_M, \beta_{RM})^T$. Our parameters of interest in the analysis model in (2.2.8) are $\boldsymbol{\beta}$, but not $\boldsymbol{\gamma}$.

Similarly to (2.2.2), the OLS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for the analysis model in (2.2.6) are asymptotically unbiased with respect to the $\boldsymbol{\theta}$ parameters in the true model in (2.2.5) under Assumptions (A7) to (A9):

(A7) Random error ϵ's are independent;

(A8) No unmeasured confounding for the randomized intervention, moderator,

and baseline covariates on the outcome, i.e., $E(\epsilon|R, M, \tilde{X}) = 0$;

(A9) Finite variance of random error, i.e., $\text{Var}(\epsilon|R, M, \tilde{X}) = \sigma$ is finite.

Similar to Assumption (A2), Assumption (A8) is a sequential ignorability assumption, which is implied by the following stronger sequential ignorability assumption with the potential outcomes expression:

$$\text{(A8.1)} \;\; R \perp Y^{rm}|\tilde{X} \quad \text{and} \quad \text{(A8.2)} \;\; M \perp Y^{rm}|R, \tilde{X}.$$

We then consider the bias of the OLS estimators of the $\boldsymbol{\beta}$ with respect to $\boldsymbol{\theta}$ when we relax the sequential ignorability assumption (Assumption A8) by not assuming no unmeasured confounding for the moderator M on the outcome Y, i.e., an arrow from U to M allowed in Figure 2.2. The assumption of no unmeasured confounding for the randomized intervention R on the outcome Y, i.e. $E(\epsilon|R)$=0, is guaranteed by randomization and referred to the randomization assumption (Assumption A8.1). In addition, there may be unmeasured confounding for baseline covariates $\tilde{X}$ on outcome Y, i.e., an arrow from U to $\tilde{X}$ allowed in Figure 2.2. We present the bias by using the similar approach stated in the previous section.

Theorem 3. Under (A7) and (A9) and assuming $E(R\epsilon|\tilde{X}) = 0$, the asymptotic bias of the OLS estimator of $\boldsymbol{\beta}$ under the analysis model in (2.2.2) is

$$\text{Bias}(\hat{\boldsymbol{\beta}}) \;\; = \;\; \{\text{plim}(\text{Z}^T\text{Z})\}^{-1}\text{plim}(\text{Z}^T\boldsymbol{\epsilon}) + \text{plim}(\text{L})\{\text{plim}(\text{F})\}^{-1}\text{plim}((\text{ZL} - \text{X})^T\boldsymbol{\epsilon}) + o_p(1) \tag{2.2.9}$$

where $\text{L}_{4\times p} = (\text{Z}^T\text{Z})^{-1}\text{Z}^T\text{X}$, and $\text{F}_{p\times p} = \text{X}^T\text{X} - \text{X}^T\text{Z}(\text{Z}^T\text{Z})^{-1}\text{Z}^T\text{X}$.

The proof for Theorem 3 is shown in the appendix. Theorem 3 suggests that the asymptotic bias of OLS estimator $\hat{\boldsymbol{\beta}}$ under the analysis model in (2.2.6) consists of two terms. The first term, $\{\text{plim}(\text{Z}^T\text{Z})\}^{-1}\text{plim}(\text{Z}^T\boldsymbol{\epsilon})$, is the asymptotic bias as if we exclude the baseline covariates in the regression, which was discussed for the analysis model in (2.2.2) in the previous section. We can call this term "unmeasured confounding bias" (UM bias) since it is due to the unmeasured confounding of the moderator vs. outcome relationship. In the second term, $\text{L} = (\text{Z}^T\text{Z})^{-1}\text{Z}^T\text{X}$, can be considered as the OLS estimator in the pseduo multivariate regression of baseline covariates $\tilde{X}$ on R, M, and their interaction $R * M$. Note that this regression is "pseduo" because baseline covariates are measured before the post-randomization moderator. In addition, $(\text{X} - \text{ZL})$ is the residual, and F is the sum of residual square of this pseudo multivariate regression. Hence, the second term in the bias, $\text{plim}(\text{L})\{\text{plim}(\text{F})\}^{-1}\text{plim}((\text{ZL}-\text{X})^T\boldsymbol{\epsilon})$, can be called "the omitted variable bias" (OM bias) following Hosman, Hansen, and Holland (2009), which represents the part of bias of $\hat{\boldsymbol{\beta}}$ due to the omission of baseline covariates in (2.2.6).

We now consider the assumptions beyond Assumptions A7, A8.1, and A9 shown by the bias equations in (9) to impact the asymptotic bias if the OLS estimators of the β parameters in terms of the θ parameters of the true model:

(A8.2) $E(M\epsilon|R = 1, \tilde{X}) = E(M\epsilon|R = 0, \tilde{X}) = 0$

(A10) $E(M|\tilde{X}) = E(M)$

(A11) $E(\epsilon|\tilde{X}) = 0.$

Assumption A8.2 is no unmeasured confounding for M and Y given $\tilde{X}$. Assumption A10 is no correlation between M and $\tilde{X}$. Assumption A11 is no unmeasured confounding for $\tilde{X}$ and Y. We summarize the impact of these assumptions in Table 2.2.

Given that Assumptions A7, A8.1, and A9 hold, different combinations of Assumptions A8.2, A10 and A11 lead to the following results:

1. There is no overall asymptotic bias for all estimators $\hat{\boldsymbol{\beta}} = (\hat{\beta}_R, \hat{\beta}_M, \hat{\beta}_{RM})$ when Assumptions A8.2 and A10, or A11 hold.

2. The overall asymptotic bias for the estimator $\hat{\boldsymbol{\beta}}$ is equivalent to the asymptotic omitted variable bias when Assumption A8.2 holds.

3. The overall asymptotic bias for the estimator $\hat{\boldsymbol{\beta}}$ is equivalent to the asymptotic unmeasured confounding bias when Assumption A10 holds.

Table 2.2: Impact of conditions for asymptotic bias of OLS estimator in the covariate model.

Assumptions[a]					
(A8.2)	(A10)	(A11)	UM Bias	OM Bias	Overall Bias
√	√	-	$o_p(1)$	$o_p(1)$	$o_p(1)$
√	×	√	$o_p(1)$	$o_p(1)$	$o_p(1)$
√	×	×	$o_p(1)$	$\text{plim}(LF^{-1}(ZL - X)^T\epsilon)$	$\text{plim}(LF^{-1}(ZL - X)^T\epsilon)$
×	√	-	$\text{plim}((Z^TZ)^{-1}Z^T\epsilon)$	$o_p(1)$	$\text{plim}((Z^TZ)^{-1}Z^T\epsilon)$

NOTE: (a) Assumptions are (can also be found in Section 2.6) :
(A8.2) $E(M\epsilon|R = 1, \tilde{X}) = E(M\epsilon|R = 0, \tilde{X}) = 0$, (A10) $E(M|\tilde{X}) = E(M)$, (A11)$E(\epsilon|\tilde{X}) = 0$

Therefore, the asymptotic unmeasured confounding bias relies only on Assumption 8.2 - the relationship between M and Y. When there is no unmeasured confounding between M and Y, there will be no asymptotic unmeasured confounding bias. In

contrast, the asymptotic omitted variable bias may rely on all three assumptions above. When there is no association between M and $\tilde{X}$, there will be no asymptotic omitted variable bias. When there is an association between M and $\tilde{X}$, but no unmeasured confounding between M and Y, and no unmeasured confounding between $\tilde{X}$ and Y, there will be no asymptotic omitted variable bias.

If none of the above combination holds, the overall asymptotic bias will be the sum of the unmeasured confounding bias and omitted variable bias. The direction of the former depends on the sign of $\text{plim}\{\text{Z}^T\boldsymbol{\epsilon}\}$, which can be approximately considered as the sign of the effect of unmeasured confounding for M and Y. The direction of the latter depends on the sign of the correlation between M and $\tilde{X}$, and the sign of $\text{plim}\{(\text{ZL}-\text{X})^T\boldsymbol{\epsilon}\}$, i.e., the sign of effect of unmeasured confounding for the residual of projecting $\tilde{X}$ on Z and Y as $n \to \infty$. When these two biases have the opposite directions, adjusting for baseline covariates will reduce the magnitude of the asymptotic bias of $\hat{\boldsymbol{\beta}}$; while if they have the same direction, adjusting for baseline covariates will unfortunately increase the magnitude of the asymptotic bias of $\hat{\boldsymbol{\beta}}$.

2.3 Simulations

Simulation results are first presented for the non-covariate analysis model in (2.2.2), followed by simulation results for the covariate analysis model in (2.2.6). The true simulation models correspond to the relaxation of analysis Assumptions (A1) to (A3) for the non-covariate model and (A7), to (A9) for the covariate model.

All simulations were based on the following conditions. First, the true values for the model parameters that were estimable were obtained from the actual CT data analysis presented in Section 4. Second, the baseline CT data (randomization intervention R and baseline covariates $\tilde{X}$) were used to simulate M based on the regression models of M on $\tilde{X}$ and the random error within each of the randomization groups (τ if the subject is in the treatment group, and ξ if the subject is in the control group). The outcome Y was generated given R, simulated M, (the observed $\tilde{X}$ if under the covariate model), and the random error ϵ. Hence, the sample size of the data analysis for each simulation was that of the CT data (i.e., 94). The random errors (τ, ξ, ϵ) were assumed to be multivariate normal random variables with means zero, variances (σ_1^2, σ_0^2, σ^2), and covariances ($\rho\sigma_1\sigma_0$, $\rho_1\sigma_1\sigma$, $\rho_0\sigma_0\sigma$), where σ_1^2 and σ_0^2 were specified as the estimated variance of the observed M within randomized groups, σ^2 was specified as the estimated variance of observed Y, ρ was fixed to zero, while ρ_1 and ρ_0 were specified as different combinations of values (i.e., different confounding relationships for the moderator within each of the randomization groups).

The number of simulation iterations under each combination of model assumptions was 5000. We present the bias, coverage rate of 95% confidence intervals (the percentage of simulations for which the interval covered the true parameter), and Mean Squared Error (MSE).

Under the non-covariate model in (2.2.2), Table 2.3 presents the simulation results for four different values of ρ_1 and ρ_0 with the other model parameters specified as described above. Using the estimated quantities from the CT data, i.e.,

$E(M|R=1) = 0.468$, $E(M|R=0) = -0.430$, $Var(M|R=1) = 117.86$, and $Var(M|R=0) = 96.68$, (Note that M was centered by its mean before the data analysis was conducted.), we found the condition that if ρ_1/ρ_0 falls into the interval of $[0.4, 3]$, then $|Bias(\hat{\beta}_{RM})| < |Bias(\hat{\beta}_R)|$, otherwise $|Bias(\hat{\beta}_{RM})| > |Bias(\hat{\beta}_R)|$ (from expression (2.2.4) in Section 2.1.3). Therefore, four scenarios were considered to represent different unmeasured confounding relationships. The simulation results presented in Table 2.3 are consistent with our expectation. For the first two scenarios with $\rho_1/\rho_0 = 5$ or 0.2, $|Bias(\hat{\beta}_{RM})|$ are greater than $|Bias(\hat{\beta}_R)|$, and for the scenario with $\rho_1/\rho_0 = 1$, $|Bias(\hat{\beta}_{RM})|$ is less than $|Bias(\hat{\beta}_R)|$. For the scenario with $\rho_1 = \rho_0 = 0$ (i.e., with sequential ignorability), very small bias were detected for all three estimates, implying that they are asymptotically unbiased. The coverage rates for $\hat{\beta}_R$ and $\hat{\beta}_{RM}$ are higher than that for $\hat{\beta}_M$ among all four scenarios.

Table 2.3: Simulation results for the non-covariate model in (2.2.2) based on the CT data. (N=94, $\theta_R = -5.39$, θ_M=0.74, $\theta_{RM} = -0.59$.)

Under Sequential Ignorability	Effect	Bias(%)	Coverage	MSE
	Randomization	-0.181(3%)	95.7%	6.63
No	Moderator	0.102(14%)	73.9%	0.04
$\rho_1 = 0.5$, $\rho_0 = 0.1$	Interaction	-0.315(53%)	90.4%	0.16
	Randomization	-0.199(4%)	95.5%	6.69
No	Moderator	0.503(68%)	18.3%	0.28
$\rho_1 = 0.1$, $\rho_0 = 0.5$	Interaction	0.420(-71%)	64.2%	0.23
	Randomization	-0.175(3%)	96.1%	6.75
No	Moderator	0.205(28%)	79.2%	0.07
$\rho_1 = 0.2$, $\rho_0 = 0.2$	Interaction	0.097(-16%)	94.4%	0.06
	Randomization	-0.035(0.6%)	95.8%	7.26
Yes	Moderator	0.002(0.3%)	95.0%	0.03
$\rho_1 = 0$, $\rho_0 = 0$	Interaction	0.004(0.7%)	95.7%	0.06

Table 2.4: Simulation results for the covariate model in (2.2.6) based on the CT data. UM and OM represents unmeasured confounding bias and omitted variable bias respectively. (N=94, $\theta_R = -5.85$, θ_M=0.48, $\theta_{RM} = -0.35$.)

Under Sequential Ignorability	Effect	UM Bias	OM Bias	Overall Bias(%)	Coverage	MSE
	Randomization	-0.250	-0.378	-0.628(11%)	94.3%	7.31
No	Moderator	0.099	0.031	0.130(27%)	89.9%	0.06
$\rho_1 = 0.5$, $\rho_0 = 0.1$	Interaction	0.317	0.063	0.380(-109%)	71.7%	0.22
	Randomization	-0.262	0.217	-0.045(0.8%)	95.1%	6.81
No	Moderator	0.503	0.125	0.628(131%)	12.2%	0.43
$\rho_1 = 0.1$, $\rho_0 = 0.5$	Interaction	-0.420	-0.100	-0.520(149%)	52.1%	0.34
	Randomization	-0.173	-0.051	-0.224(4%)	94.8%	7.55
No	Moderator	0.200	0.052	0.252(53%)	76.9%	0.11
$\rho_1 = 0.2$, $\rho_0 = 0.2$	Interaction	-0.034	-0.012	-0.046(13%)	94.8%	0.08
	Randomization	-0.006	0.007	0.001(0.02%)	94.8%	7.75
Yes	Moderator	-0.001	-0.001	-0.002(0.4%)	97.5%	0.05
$\rho_1 = 0$, $\rho_0 = 0$	Interaction	0.001	-0.001	$< 0.001 (< 0.01\%)$	95.0%	0.08

Under the similar setting of Table 2.3, Table 2.4 presents the simulation results for covariate model in (2.2.6). Under no sequential ignorability, the bias for the estimates of the three effects is the combined effects from unmeasured confounding bias and omitted variable bias, and the omitted variable bias cannot be ignored. Recall that the observed baseline covariates were used to generate the outcome, and they are not correlated with the random error ϵ, implying that no unmeasured confounding of $\tilde{X}$ and Y are assumed in this simulation setup. Therefore, from (2.2.9), when there is no unmeasured confounding of $\tilde{X}$ and Y, and under sequential ignorability, all three estimates should be asymptotically unbiased. The simulation results in Table 2.4 confirms this. Under sequential ignorability, the unmeasured confounding bias, omitted variable bias and overall bias for three estimates are all very small. Among all four scenarios, the coverage rates for the estimates for the randomization effect are always high (94.3% - 94.8%), the coverage rates for the estimates for the interaction

effect are fair (52% - 95%), and the coverage rate for the estimated for the moderator effect may be very low (12.2% - 97.5%).

2.4 Data Analysis

The motivation for assessing post-randomization moderators of treatment arose from the randomized trial of a cognitive therapy intervention for suicide attempters (Brown et al. 2005). The sample consisted of 120 suicide attempters who received medical or psychiatric evaluation at the Hospital of the University of Pennsylvania within 48 hours of the attempt and were recruited from the hospital emergency department from the original study. In this study, subjects were randomized to either receive or not receive 10 sessions of Cognitive Therapy (CT) for suicidal behavior and depression. Regardless of randomization assignment, everyone received usual care CT for suicidal behavior. In treating these outcomes, CT focuses on negative thinking (i.e. hopelessness and self-criticism), behavior problems (i.e. avoidance and passivity), strategies to instill hope (i.e. helping patient get a job), reducing suicide ideation, and needs related to the recent suicidal crisis. Under the common treatment factor hypothesis that intervention effects for behavioral outcomes are stronger in early responders to initial treatment (Haas et al. 2002; Lambert 2005), early suicide ideation targeted by CT will be examined as a post-randomization moderator of treatment on subsequent depression severity as an outcome given that CT impacts depression severity in addition to reducing the risk of suicide attempts.

For the purposes of illustrating the analytic results, the data analysis will focus on the moderation of the effect of CT on 6-month depression by 1-month suicide ideation. Because of drop-out the sample size is 94. Depression was measured by BDI-II, and suicide ideation was measured by Scale for Suicide Ideation-Worst (SSIW).

We used two analysis models described in previous sections on CT data for the outcome BDI-II at 6 months. They are: (1) the standard regression model with CT, SSIW and their interaction; (2) the standard regression model with CT, SSIW and their interaction, and baseline covariates such as gender, baseline BDI-II, Beck Hopelessness Scale (BHS), number of suicide attempts, physical health, etc., as the predictors. The analysis results are presented in Table 2.5.

Table 2.5: Data analysis results for the CT trial: non-covariate and covariate models of the standard regressions with OLS.

	Non-covariate model				Covariate Model			
Effect	Estimate	Std	t-value	P-val	Estimate	Std	t-value	P-val
Randomization	-5.39	2.73	-1.97	0.0515	-5.85	2.70	-2.17	0.0330
Moderator	0.74	0.19	3.81	0.0003	0.48	0.22	2.25	0.0269
Interaction	-0.59	0.27	-2.21	0.0297	-0.35	0.29	-1.18	0.2414

The covariate and non-covariate models yielded similar inference but the magnitudes of estimates differed. The two models agreed on the direction but not the significance of the interaction between CT and 1-month suicide ideation on 6-month depression. Under the common treatment factor model, one would expect some effect of CT on suicide ideation. The corresponding observed ITT effect size (mean group difference divided by the standard deviation of 1-month suicide ideation) for

this potential effect modifier suggests that 1-month suicide ideation may not correspond to the common treatment factor, as the effect size is very small (< 0.1). In turn, the estimator of the interaction may be unbiased in the presence of unmeasured confounding at least in the non-covariate case.

Among the eight baseline covariates included in the study, five of them are significantly correlated with the 1-month suicide ideation (p-value < 0.05), and four of them are significantly correlated with the 6-month depression (p-value < 0.05). Therefore, the omitted variable bias may not be ignored in (2.2.9).

The negative estimate for the interaction effect implies that the CT intervention alleviates the depression symptoms more for the patients with less suicide ideation than those with severe suicide ideation. Although one needs to be careful in interpreting main effects in the presence of interaction, the covariate and non-covariate models yielded estimates with same signs for both the main effects for CT and for suicide ideation. In the presence of the interaction, the significant main effect for CT is interpreted as the reduction of 6-month depression for those with no suicide ideation at 1-month. The significant increase in 6-month depression due to 1-month ideation pertains to the non-CT group. Although the parameter estimates in the two models have same signs, their values are somewhat different. This implies that adjusting for the measured baseline covariates in the regression model reflects confounding by them. One may conclude that there may be unmeasured factors that are also confounders of the effects in the regression models. Under such unmeasured confounding or departures from one of the other Assumptions in (A1)-(A3) or (A7)-

(A9) are violated, the two sets of estimates using covariate and non-covariate models are biased.

2.5 Discussion

In the context of assessing the modification of randomized intervention effects on outcome by early post-randomization moderators impacted by the intervention, we have investigated the bias of OLS estimators for the parameters of a standard regression interaction model. More specifically, we have presented bias results for the standard regression model with main effects and interaction for the randomized intervention and post-randomization moderator under departures from the assumptions for OLS estimation. Our results have several implications for assessing how post-randomization moderators modify the effect of randomized intervention on outcome in randomized trials.

First, the bias of the estimators of the different parameters with respect to the respective causal parameters depends in different ways on assumptions involving unmeasured confounding and differences between the randomization groups with respect to the moderator and unmeasured confounders. The OLS estimator of the main effect for the moderator is unbiased only when there is no unmeasured confounding of the moderator vs. outcome relationship, regardless of the other assumptions. In contrast, even in the presence of unmeasured confounding, the estimator of the randomized intervention main effect is unbiased when the magnitude of unmeasured confounding

and the mean and variance of the moderator are the same between the randomization groups. Further, in the presence of unmeasured confounding, the estimator of the interaction term is unbiased only when the magnitude of unmeasured confounding and variance of the moderator is the same between the randomization groups. We have shown that the symmetry of the joint distribution of the randomized intervention, moderator, and outcome can guarantee this condition.

In the presence of observed baseline covariates, the bias due to the above conditions ("unmeasured confounding bias") may be augmented by "omitted variable bias", due to the relationships among baseline covariates, moderator, outcome, and unmeasured confounder. Even without unmeasured confounding of the moderator and outcome, all OLS estimators under the regression interaction model may be biased due to these relationships. Figure 2.2 shows that either in the presence of unmeasured covariates or the baseline covariates, the moderator becomes a collider thus inducing confounding by its presence in the model.

The above results indicate that even in the presence of unmeasured confounding, unbiased inference may still be possible at least with respect to the randomized intervention main effect and its interaction with the moderator. The less of an impact of the randomized intervention on the distribution of the post-randomization moderator and its relationship with unmeasured confounders, the less bias there is for the estimators of the randomized intervention main effect and interaction terms. Such a result suggests more flexibility with OLS estimation in terms of the need of no unmeasured confounding assumptions. Apart from the scientific reasons for assess-

ing post-randomization moderation, testing such interaction has been proposed for the mediation context (see e.g., Vansteelandt 2009) in the absence of no unmeasured confounding assumptions, the reliance on the assumption of no relationship between the randomized intervention and the post-randomization moderator conflicts with the mediation context where the post-randomization factor is the mediator. Mediation requires that the randomized intervention-mediator interaction not be significant. One would then think that the reliance on the assumption of no relationship between the intervention and mediator in the absence of the no-unmeasured confounding assumption conflicts with the assumption for mediation that such a relationship does exist. However, the test for the interaction requires the no-unmeasured confounding assumption regardless of the relationship between the intervention and mediator. That is, there is no conflict between assessing treatment-post-randomization moderator interaction and assessing mediation.

Chapter 3

Optimal G-estimation Mediation Analyses under departure from Sequential Ignorability

3.1 Introduction

In clinical trials, especially for trials in behavioral intervention studies, the response for treatments may be very poor. Researchers are interested in not only the average treatment effect, but also how the effects vary among the subject groups. Conceptual theory in the behavioral science literature suggests that the treatment effect may depend on post-randomization factors. Therefore, there has been focus on treatment effect modification using post-randomization variable as moderators or mediators; one purpose is to identify early in the study patients who will respond

more efficiently or who will respond earlier to the treatment.

An example in a psychiatry study that motivates our research is a cognitive therapy trial (Brown et. al., 2005). The purpose of this trial is to evaluate the effect of cognitive therapy for recent suicide attempters. In this trial, at baseline, each patient is randomly assigned to either receive or not receive 10 sessions of Cognitive Therapy specifically developed for preventing suicide attempters. Regardless of randomization assignment, everyone received usual care from clinicians in the community as well as tracking and referral services from the study case managers. The outcome, the Beck Depression Inventory-II (BDI-II), is a score of depression severity, which was measured at 6 months after randomization. The post-randomization variable, the Scale for Suicide Ideation Worst (SSIW), was measured at 1-month after randomization but before the measurement of the final outcome. The scientific question is then whether the SSIW at 1-month modifies the treatment effect on the 6-month depression score BDI-II.

Post-randomization factors may mediate the effects of the treatment on the outcome in addition to moderating the effects; in the presence of possible mediation, information on post-randomization factors in often used to evaluate direct versus indirect effects in biomedical studies. In a randomized trial comparing the effect of high-dose vs. low-dose 3-azido-3-deoxythymidine (AZT) for patients with HIV disease, subjects in the high-dose AZT group are less likely to receive prophylaxis therapy for Pneumocystis Carinii Pneumonia(PCP), a post-randomization treatment. A direct effect is the effect of AZT holding the same level of post-randomization treatment

to a given level, e.g., receiving PCP prophylaxis therapy, and an indirect effect is the part of the effect of AZT mediated by PCP. To assess the benefits associated with AZT and the benefits of receiving PCP, Robins and Greenland (1994) used Structural Nested Failure Time models. Although the main question in this example was about mediation and not moderation as in the first example, a similar statistical approach may apply.

In this study, using the data from the first example, we investigate two types of statistical methods.

we consider and compare two methods that have been proposed to estimate the joint effects of the main treatment and the mediator/moderator. First, we consider standard regression, which has been widely used for investigating the joint effects of a randomized treatment and post-randomization variables (Baron and Kenny, 1986). It is a crucial assumption that the post-randomization variable is effectively randomized to individuals in addition to the randomized baseline intervention. In our data example, this assumption is dubious. The second approach is based on formal causal reasoning involving the potential outcomes framework (Rubin, 1974). Although the potential outcome framework has been set up for a long time, causal models that are appropriate for post-randomization modification analysis became available during the last two decades (Robins, 1992, Robins and Greenland, 1994). In this paper, we consider inference using Structural Nested Models and G-estimation without assuming the sequential ignorability.

The paper is organized as follows: Section 2 presents statistical methods and

analytical approaches, Section 3 presents a case-study, Section 4 provides the results of a simulation experiment, and Section 5 provides a discussion.

3.2 Statistical Models and Analytical Methods

We define all the observed and potential variables for subject i, while we generally suppress the index i to simplify the notation. For the observed variables, Y is the observed continuous outcome; R is the observed randomized intervention, 1 for subjects assigned treatment, 0 for control; M is the observed post-randomization factor, either continuous or binary; X is the vector of the observed baseline covariates other than randomization. We define Y^{rm} as the outcome variable that would be observed if subject i were randomized to level r of the intervention and then were to receive or exhibit level m of the post-randomization factor. Therefore, Y^{00} can be considered as the reference potential outcome, which is the outcome that would be observed for the subject to receive levels 0 for both R and M.

Causal effects are contrasts between different potential outcomes. $Y^{R0} - Y^{00}$ is the causal (manipulated) direct effect of the randomized intervention holding post-randomization factor M constant as 0 for the subject who receives the treatment R. $Y^{RM^R} - Y^{0M^R}$ is the causal (natural) direct effect of the randomized intervention holding post-randomization factor M to the level it would have attained had the subject received treatment R. $Y^{0M^R} - Y^{0M^0}$ is the causal (natural) indirect effect of the randomized intervention holding the intervention constant as 0 and manipu-

lating the post-randomization factor M to the level it would have attained had the subject received treatment R. In addition, $Y^{RM} - Y^{R0}$ is the causal effect of post-randomization factor M for the subject who receives the treatment R and at the level M of the post-randomization variable.

3.2.1 Models

Structural Nested Distribution Model SNDM

Robins(1992, 1999) developed a number of innovative methods to eliminate the bias of standard methods for estimating the causal effect of treatment. Following his idea, we define the baseline potential outcome Y^{00} in which the interaction between the randomized intervention and the post-randomization factor is allowed:

$$Y^{00} = Y - \psi_R R - \psi_M M - \psi_{RM} RM, \tag{3.2.1}$$

where ψ_R, ψ_M, and ψ_{RM} are the causal parameters. Loosely speaking, ψ_R is the main effect for the randomization factor R, ψ_M is the main effect for the post-randomization factor M, and ψ_{RM} is their interaction effect.

Standard Regression Model

To compare with the causal model in (3.2.3), we consider the corresponding standard linear regression model:

$$Y = \beta_X X + \beta_R R + \beta_M M + \beta_{RM} RM + \epsilon_S, \tag{3.2.2}$$

where $(\beta_R, \beta_M, \beta_{RM})$ are parameters for the association model, β_X is a vector of effects for baseline covariate values x, and ϵ_S is a mean zero error term with fixed variance. In this model, it is assumed that ϵ is uncorrelated with all the regressors. Note that in the absence of sequential ignorability, this assumption does not hold.

The parameters β_R, β_M, and β_{RM} are defined as comparisons of observed outcome expectations from different sample subgroups defined by R and M (Ten Have et al. 2007).Therefore, they are not casual contrasts of expectations under different conditions defined by R and M for the same individual. The comparisons of such subgroups will in general only equal the causal contrasts for an individual under certain conditions:

1. Sequential ignorability of both the baseline intervention and mediator given baseline covariates;

2. Independence among subjects;

3. Model assumptions including the correct association between baseline covariates X and outcome Y.

4. Finite variance of random error.

3.2.2 Assumptions for G-estimation of SNDMs

The necessary assumptions to obtain the unbiased estimators in the SNDMs are:

1. Stable Unit Treatment Value Assumption (SUTVA). This assumption consists of two sub-assumptions. First, there is a single value for each of the potential outcome variables (Y^{rm}) for a given subject i regardless of the randomization assignment or

mediation behavior of any other subject i'. Second, there is a single value for each of the potential outcome random variables (Y^{rm}) for a given subject i regardless of the method of administration of the randomized baseline intervention or the administration or occurrence of the mediator.(Angrist, Imbens, and Rubin, 1996; Ten Have et al. 2007; VanderWeele and Hernan, 2011)

2. Randomization Assignment or the ignorability of baseline assignment of intervention. Mathematically, this means that $Y^{rm} \perp R|X$. This assumption implies stochastic independence between the randomized baseline intervention, R, and potential outcomes, i.e., $P(Y^{rm}|R = r, X = \tilde{x}) = P(Y^{rm}|X = x)$. It also implies no imbalanced between randomization groups with respect to unmeasured confounders.

3. Model assumptions includes no interaction assumptions of $R*M*X$, $R*X$, and $M*X$. However, the SNDM model relaxes the assumption of the correct association of X and Y^{00}.

4. Independence of observations for standard error estimation.

3.2.3 Estimation for SNDM

We do not know what Y^{00} is. Under a causal theory, we can compute a putative value for Y^{00} from observed quantities as

$$Y^{00}(\Psi) \equiv Y - \psi_R R - \psi_M M - \psi_{RM} RM, \tag{3.2.3}$$

where $\Psi = (\psi_R, \psi_M, \psi_{RM})$.

If the putative value of the causal parameter Ψ is true, $Y^{00}(\Psi)$ can be viewed as

the potential outcome Y^{00} and will be independent of R given X. Estimation may be based on testing this independence for an assumed value of the causal parameter Ψ, which is the basic idea of G-estimation.

Under the assumptions for SNDM stated previously, we can obtain consistent estimators for ψ_R, ψ_M, and ψ_{RM} by solving appropriate unbiased estimating equations.

Let $q = Pr(R = 1)$ be the randomization probability (also the propensity score), and let $g(Y^{00}(\Psi), X)$ be a known function of potential outcomes $Y^{00}(\Psi)$ and X. The randomization assumption implies that $g(Y^{00}(\Psi), X)$ is independent of R conditional on X. Under the SNDM (3.2.3), we can obtain the correct Ψ through solving the following estimation equation.

$$E(U(\Psi)) = E\left\{\sum_i (R-q)g\{Y^{00}(\Psi), X\}\right\} = 0 \tag{3.2.4}$$

A consistent estimator Ψ can be obtained by solve the empirical version of (3.2.4):

$$U(\Psi) = \sum_i (R-q)g\{Y^{00}(\Psi), X\} = 0. \tag{3.2.5}$$

This method is known as G-estimation. The choice for $g\{Y^{00}(\Psi), X\}$ will not impact the consistency of the estimator, while some choices for $g\{Y^{00}, X\}$ will lead to attaining semi-parametric information bound, and other choices will lead to inefficient estimators. The optimal choice that produces the most efficient estimation for dichotomous randomized intervention R is (Robins(1992)):

$$\begin{aligned} g^{opt}(Y^{00}, X) &= E\{S_\Psi(\Psi, \theta; X, R, M, Y)|X, R = 1, Y^{00}\} \\ &\quad -E\{S_\Psi(\Psi, \theta; X, R, M, Y)|X, R = 0, Y^{00}\} \end{aligned} \tag{3.2.6}$$

where $S_\Psi(\Psi, \theta; X, R, M, Y)$ is the score function with respect to Ψ, computed from the full likelihood. θ represents the nuisance parameters.

In the framework of our post-randomization analyses, The full likelihood function is

$$\begin{aligned} & L(\Psi, \theta; X, R, M, Y) \\ = \; & L(\Psi, \theta; X, R, M, Y^{00}(\Psi)) \frac{\partial Y^{00}(\Psi)}{\partial Y} \\ = \; & f(X; \theta) f(R|X; \theta) f(Y^{00}(\Psi)|X; \theta) f(M|R, X, Y^{00}(\Psi); \theta) \end{aligned} \tag{3.2.7}$$

Note that $\frac{\partial Y^{00}(\Psi)}{\partial Y} = 1$ for model (3.2.3).

Under sequential ignorability and our model assumption (3.2.3), the score function for Ψ is

$$S_\Psi(\Psi, \theta; X, R, M, Y) = \frac{\partial \log f(Y^{00}(\Psi)|X; \theta)}{\partial \Psi} = -\frac{\partial \log f(Y^{00}(\Psi)|X; \theta)}{\partial Y^{00}(\Psi)} \begin{pmatrix} R \\ M \\ RM \end{pmatrix}. \tag{3.2.8}$$

Without sequential ignorability, if we assume randomization only, i.e., the distribution of M depends on the baseline potential outcome Y^{00}, then under our model

assumption (3.2.3), the score function for Ψ is

$$S_\Psi(\Psi,\theta;X,R,M,Y) = \frac{\partial \log f(Y^{00}(\Psi)|X;\theta)}{\partial\Psi} + \frac{\partial \log f(M|X,R,Y^{00}(\Psi);\theta)}{\partial\Psi}$$
$$= -\frac{\partial \log f(Y^{00}(\Psi)|X;\theta)}{\partial Y^{00}(\Psi)} \begin{pmatrix} R \\ M \\ RM \end{pmatrix} - \frac{\partial \log f(M|X,R,Y^{00};\theta)}{\partial Y^{00}(\Psi)} \begin{pmatrix} R \\ M \\ RM \end{pmatrix}. \tag{3.2.9}$$

Joffe and Brensigner (2003) proposed a weight scheme to gain efficiency of G-estimation based on the sequential ignorbility assumption. Although the models are different, we can apply their scheme on our case. In the followings, we assume that the potential outcome $Y^{00} \sim N(\mu(X),\sigma^2)$, and we discuss the optimal G-estimation under the score functions (3.2.8) and (3.2.9) respectively.

Optimal G-estimation with the sequential ignorability

Under (3.2.8) and the working assumption of $Y^{00} \sim N(\mu(X),\sigma^2)$, we have

$$g^{opt}(Y^{00},X) = \frac{Y^{00}-\mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ E\{M|X,R=1\} - E\{M|X,R=0\} \\ E\{M|X,R=1\} \end{pmatrix}. \tag{3.2.10}$$

Expression (3.2.10) is a vector of weight corresponding to ψ_R, ψ_M, and ψ_{RM}, respectively. It works for either continuous M or binary M. Following Joffe and Bresigner (2003), the part in parenthesis of (3.2.10) is the weight vector to identify the estimation equation in G-estimation. Specifically, the unit weight for main effect of R

implies that all subjects have the same effect of R, thus should be given the same weight. The weight for ψ_M, $E\{M|X, R = 1\} - E\{M|X, R = 0\}$, can be called the compliance score (Follman, 2000) since it is a measure of the effect of randomization R on the treatment received, which is the post-randomization factor M in our case. This can also be referred as the principle score (Hill, Waldfogel, and Brooks-gunn, 2002). Intuitively, this means that the subjects with higher effects of R on M given the baseline covariates X will provide more information on the effect of R on M and thus they should be given more weight. The weight for ψ_{RM} is the expectation of M in the treatment group. Since the subjects in the control group denoted as $R = 0$, the subjects in the treatment group ($R = 1$) with higher expected value of M will provide more information on the effect of R on $R * M$.

The difference between the expectations of two potential outcomes Y^{1M^1} and Y^{0M^0} can be written as the product of the optimal weight in (3.2.10) and causal parameters,

which is another intuitive way to explain the optimal weight.

$$
\begin{aligned}
& E\{Y^{1M^1} - Y^{0M^0}|X\} \\
= \; & E\{Y^{1M^1}|R=1,X\} - E\{Y^{0M^0}|R=0,X\} \\
= \; & E\{Y^{00} + \begin{pmatrix} 1 & M^1 & M^1 \end{pmatrix} \begin{pmatrix} \psi_R \\ \psi_M \\ \psi_{RM} \end{pmatrix} |R=1,X\} \\
& -E\{Y^{00} + \begin{pmatrix} 0 & M^0 & 0 \end{pmatrix} \begin{pmatrix} \psi_R \\ \psi_M \\ \psi_{RM} \end{pmatrix} |R=0,X\} \\
= \; & E\{Y^{00}|X\} - E\{Y^{00}|X\} \\
& + \begin{pmatrix} \psi_R & \psi_M & \psi_{RM} \end{pmatrix} \begin{pmatrix} 1 \\ E(M^1|R=1,X) - E(M^0|R=0,X) \\ E(M^1|R=1,X) \end{pmatrix} \\
= \; & \begin{pmatrix} \psi_R & \psi_M & \psi_{RM} \end{pmatrix} \begin{pmatrix} 1 \\ E(M|R=1,X) - E(M|R=0,X) \\ E(M|R=1,X) \end{pmatrix}
\end{aligned}
$$

Optimal G-estimation without sequential ignorability

Without sequential ignorability, the score function (3.2.9) depends on not only the distribution of Y^{00}, but also the distribution of M.

When the post-randomization factor M is continuous, we assume that $M \sim N(\mu_m, \sigma_m^2)$, where μ_m is a linear combination of X, R, and Y^{00}. When the post-

randomization factor M is binary, we assume that M is a logit regression of X, R, and Y^{00}, i.e., $logit(P(M=1)) = \log\left\{\frac{P(M=1)}{1-P(M=1)}\right\} = u_m$.

Under the above working assumptions, for continuous M,

$$\begin{aligned} g^{opt}(Y^{00}, X) &= \frac{Y^{00} - \mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ E\{M|X, R=1, Y^{00}\} - E\{M|X, R=0, Y^{00}\} \\ E\{M|X, R=1, Y^{00}\} \end{pmatrix} \\ &- \begin{pmatrix} 0 \\ \frac{\partial}{\partial Y^{00}}\left[E\{M|X, R=1, Y^{00}\} - E\{M|X, R=0, Y^{00}\}\right] \\ \frac{\partial}{\partial Y^{00}} E\{M|X, R=1, Y^{00}\} \end{pmatrix}. \end{aligned} \quad (3.2.11)$$

Compared with the results under sequential ignorability, first of all, the first part of $g^{opt}(Y^{00}, X)$ has the similar expression as (3.2.10), while now the expectation or probability of M are conditional on not only X and R, but also the baseline potential outcome Y^{00}. More importantly, the second part of the $g^{opt}(Y^{00}, X)$ are the extra terms under no sequential ignorability.

The details of the derivation of the optimal weight are attached in the appendix.

Variance estimation

The variance-covariance for $\widehat{\Psi}$ is estimated after convergence of the G-estimation algorithm with a sandwich estimator based on (3.2.5) as follows:

$$\mathrm{Var}(\widehat{\Psi}) = V_H^{-1}(\Psi) B(\Psi) (V_H^{-1}(\Psi))^T, \quad (3.2.12)$$

where V_H is the Hessian, a symmetric 3×3 matrix : $V_H(\Psi) = \sum \frac{\partial U(\Psi)}{\partial \Psi}$, $U(\Psi) = \sum_i (R-q) g\{Y^{00}(\Psi), X\}$, and $q = P(R=1|X)$; $B(\Psi) = [\sum U(\Psi) U(\Psi)^T]$. The

resulting estimate of $\mathrm{Var}(\widehat{\Psi})$ is used in Wald statistics for hypothesis testing and Wald confidence intervals for Ψ.

Following Robins(1992a), the fact that the probability of R is estimated can be taken into account. Robins (1992a) suggested to adjust the propensity score estimation into the estimation equations to obtain a less conservative estimation for $\mathrm{Var}(\widehat{\Psi})$. Assume that $q = P(R = 1|X) = \exp(X^T\beta)/(1 + \exp(X^T\beta))$, where β is a $p \times 1$ vector representing the parameters in the propensity score regression, and $X = (1, x_1, \cdots, x_{p-1})^T$ the baseline covariates. Denoting $U(\beta) = \sum(R - q)X$, $U(\beta, \Psi) = \begin{pmatrix} U(\beta) \\ U(\Psi) \end{pmatrix}$, then $V_H(\beta, \Psi) = \begin{pmatrix} \sum \frac{\partial U(\beta)}{\partial \beta} & 0 \\ \sum \frac{\partial U(\Psi)}{\partial \beta} & \sum \frac{\partial U(\Psi)}{\partial \Psi} \end{pmatrix}$, and $B(\beta, \Psi) = [\sum U(\beta, \Psi)U(\beta, \Psi)^T]$. Therefore, $\mathrm{Var}(\widehat{\Psi})$ is the 3×3 submatrix corresponding to the Ψ elements in the $(p+3)\times(p+3)$ matrix $\mathrm{Var}(\hat{\beta}, \widehat{\Psi}) = V_H^{-1}(\beta, \Psi)B(\beta, \Psi)(V_H^{-1}(\beta, \Psi))^T$.

3.3 Data Analysis

The data example used in our study is a psychiatry trial(Brown et. al., 2005). The purpose of this trial is to evaluate the effect of cognitive therapy for recent suicide attempters. At baseline, each patient is randomly assigned to either receive or not receive 10 sessions of Cognitive Therapy specifically developed for preventing suicide attempters. Regardless of randomization assignment, everyone received usual care from clinicians in the community as well as tracking and referral services from the study case managers. The outcome is a score of depression severity, Beck Depression Inventory-II (BDI-II), which is measured at 6-month after randomization.

The post-randomization variable, the Scale for Suicide Ideation Worst (SSIW) was measured at 1-month after randomization but before the measurement of the final outcome. We have several baseline covariates such as gender, age, baseline BDI-II, Beck Hopelessness Scale, etc.

We analyze the data using standard and causal methods. For standard approaches, we investigate two standard linear regression interaction models with Ordinary Lease Squares (OLS). The first OLS does not adjust baseline covariates and the second OLS does. For causal approaches, we use the two optimal G-estimation methods whose consistency depend only on initial randomization. In terms of efficiency, G-estimation I is optimal when sequential ignorability holds. G-estimation II is optimal even when sequential ignorability does not hold. The data analysis results are summarized in Table 3.1.

First, for the main effect of randomization, all the methods are similar in terms of estimates, standard deviation, and p-value. The negative estimates of the main effect of randomization from all the methods indicate that patients in the treatment group have lower BDI, which means less severe depression. For the main effect of moderator, all the methods lead to significant positive estimates($p < 0.05$), which implies that the patients with higher SSIW at 1-month have higher BDI or more severe depression, although the values of the estimates and standard deviation are very different among the methods. For the interaction effects, all the methods have negative estimates but the values are different. Within the two OLS and within the two G-estimations, the estimates, standard deviations are similar. The signification negative value of the

interaction effect from the two G-estimations implies that patients with higher SSIW at 1-month tend to have more treatment effect.

The differences between the two OLS indicate that the measured baseline covariates are confounders in the regression model for Y. They are likely not the only confounders because if they were, the OLS should be similar as the G-estimations. For the two OLS estimators, by adding baseline covariates, the standard deviation of the main effect of randomization decreases while the standard deviation of the main effect of moderator increases. The randomization is independent of baseline covariates and therefore, adjusting baseline covariates increases its efficiency. In contrast, the lower efficiency for the estimation of the moderator may be due to the higher association among the moderator and the baseline covariates. The differences between the OLS with baseline covariate adjustment and the two G-estimation approaches imply that these baseline covariates are not sufficient to control the confounding of the moderator.

3.4 Simulations

We now present simulation results for the effects of the randomized intervention, post-randomization factor, and their interaction given the example cognitive therapy trial. Each data set for each set of simulations is based on the characteristics of the example data set and the fitted SNDM by G-estimation II.

We use the observed baseline covariates based on the real cognitive therapy data,

Table 3.1: Data analysis results for the CT trial: the standard regression models with OLS and SNDMs with G-estimations.

Method	Effect	Estimate	Std	P-val
OLS (without baseline covariates adjustment)	Randomization	-5.39	2.73	0.05
	Moderator	0.74	0.19	< 0.01
	Interaction	-0.59	0.27	0.03
OLS	Randomization	-5.85	2.70	0.03
	Moderator	0.48	0.22	0.03
	Interaction	-0.35	0.29	0.24
G-estimation I	Randomization	-5.41	2.87	0.06
	Moderator	1.53	0.43	< 0.01
	Interaction	-1.63	0.54	< 0.01
G-estimation II	Randomization	-5.26	2.77	0.06
	Moderator	1.25	0.43	< 0.01
	Interaction	-1.53	0.45	< 0.01

and generate baseline randomized intervention from binomial with $p = 0.5$. We generate potential outcome and post-randomization factor based on our working assumptions. Specifically, we generate $Y^{00} = \mu(X) + \epsilon_Y$, where $\epsilon_Y \sim N(0, \sigma_Y^2)$. Next, we generate continuous post-randomization factor $M = \mu_M(X, Y^{00}, R) + \epsilon_M$, where $\epsilon_M \sim N(0, \sigma_M^2)$, and $\mu_M(X, R, Y^{00}) = X\theta_x + R\theta_r + XR\theta_{xr} + Y^{00}\theta_{Y^{00}} + XY^{00}\theta_{xY^{00}} + RY^{00}\theta_{rY^{00}} + XRY^{00}\theta_{xrY^{00}}$. The observed outcome is then generated by $Y = Y^{00} + \psi_R R + \psi_M M + \psi_{RM} RM$. Note that $\mu(X), \sigma_Y^2, \sigma_M^2$, θ's, and ψ's are all study-based estimates from the fitted SNDM by G-estimation II.

We analyze each data set with three methods. First, we use the standard linear regression model with OlS and with baseline covariates. We do not present the OLS without baseline covariates since there are somewhat similarity between the two OLS, and the estimates of the two OLS are expected to be biased due to the unmeasured

confounding. For causal approaches, as in the data analyses, we consider two G-estimation methods whose consistency depends only on the initial randomization: G-estimation optimal under sequential ignorability (G-estimation I) and G-estimation optimal even when sequential ignorability does not hold (G-estimation II). For each simulation, we simulate 1000 sets of data, and compute the absolute bias of the estimates, the mean squared error (MSE), and confidence interval coverage.

Table 3.2 to 3.4 have the same simulation set up except for their different sample sizes. Table 3.2 is under the original sample size in the cognitive therapy trial. We double the sample size in Table 3.3 and use five times the original sample size in Table 3.4 to study the asymptotic properties of the two G-estimation approaches. The results of Table 3.2 to 3.4 show that:

First, G-estimation I and II have smaller bias than the OLS estimators, especially for larger sample size and for the main effect of M and the interaction effect of $R \times M$.

Second, between G-estimation I and II, with the original sample size, G-estimation I has smaller bias and MSE. The worse performance of G-estimation II may be due to the bad estimation for the nuisance parameters in the model of post-randomization factor M. We have many nuisance parameters in the model of M. Asymptotically, it pays no price for it, while for a small sample size, it does not work well. With larger sample sizes, G-estimation II has smaller bias and MSE, and again especially true for the main effect of M and the interaction effect of $R \times M$. The simulation results match our analytical results in previous section. Comparing the g^{opt} under and not under sequential ignorability, i.e. (3.2.10) and (3.2.11), the second part of

(3.2.11) is the extra term. The first component in this term is zero, which implies that relaxing the sequential ignorability assumption may not improve the efficiency of main effect of randomization R. However, for the other two effects, the main effect of post-randomization factor M and its interaction effect with R, their efficiency will be improved.

Table 3.2: Simulation results (N=94; $\psi_R = -5.26$, $\psi_M = 1.25$, $\psi_{RM} = -1.53$)

Method	Effect	Coverage Rate	Bias(%)	MSE
OLS	Randomization	95%	0.369(-7.0%)	6.931
	Moderator	30%	-0.444(-35.4%)	0.226
	Interaction	11%	0.736(-48.1%)	0.590
G-estimation I	Randomization	94%	0.344(-6.5%)	9.442
	Moderator	91%	-0.059(-4.7%)	0.312
	Interaction	92%	0.086(-5.6%)	0.265
G-estimation II	Randomization	95%	0.301(-5.7%)	9.475
	Moderator	83%	-0.124(-9.9%)	0.542
	Interaction	81%	0.124(-8.1%)	0.652

Table 3.3: Simulation results (N=188; $\psi_R = -5.26$, $\psi_M = 1.25$, $\psi_{RM} = -1.53$)

Method	Effect	Coverage Rate	Bias(%)	MSE
OLS	Randomization	95%	-0.245(-4.7%)	3.448
	Moderator	1%	-0.620(-49.4%)	0.402
	Interaction	$< 1\%$	0.972(-63.6%)	0.973
G-estimation I	Randomization	94%	0.203(-3.9%)	4.425
	Moderator	93%	-0.030(-2.4%)	0.124
	Interaction	94%	0.039(-2.6%)	0.120
G-estimation II	Randomization	94%	0.222(-4.2%)	4.470
	Moderator	97%	-0.010(-0.8%)	0.084
	Interaction	93%	-0.001(-0.1%)	0.130

In the second set of simulations, we increase the variability over different levels of the covariate X of the effect of randomization R on post-randomization factor M,

Table 3.4: Simulation results (N=470; $\psi_R = -5.26$, $\psi_M = 1.25$, $\psi_{RM} = -1.53$)

Method	Effect	Coverage Rate	Bias(%)	MSE
OLS	Randomization	96%	0.157(-3.0%)	1.265
	Moderator	$< 1\%$	-0.629(-50.2%)	0.403
	Interaction	<1%	0.972(-63.5%)	0.956
G-estimation I	Randomization	95%	0.110(-2.1%)	1.569
	Moderator	94%	-0.009(-0.7%)	0.042
	Interaction	94%	0.013(-0.8%)	0.048
G-estimation II	Randomization	95%	0.105(-2.0%)	1.570
	Moderator	98%	0.006(0.4%)	0.020
	Interaction	95%	-0.006(0.4%)	0.042

using the double and five times of the original sample size in Table 3.5 and Table 3.6, respectively.

The OLS has lower coverage rate and larger bias and MSE under this situation for all three effects (except the MSEs for randomization effects, which are a little bit smaller). Between the two G-estimations, for the main effect of R, G-estimation I and II have similar bias and MSE. For the other two effects, G-estimation II has higher coverage rate, smaller bias, and smaller MSE. In addition, comparing with the first set of simulations under the same sample sizes, i.e., Table 3.3 vs. Table 3.5, and Table 3.4 vs. Table 3.6, we find that G-estimation II improves more on MSE in the second set of simulations. All these results imply that increasing the variability of the effect of randomization R on post-randomization factor M will lead to better performance of optimal G-estimations, especially G-estimation II in our case. The results are similar to Joffe and Brensinger (2003). The implications are that one should empirically look for baseline characteristics (X) which are associated with the

post-randomization factor M and with the effect of the randomization effect of R on M.

The performance on the estimators of the main effect of randomized intervention R is almost the same among all estimators and do not vary much among the different sample sizes. This is an artifact of the way the data were generated. Under an alternative specification, i.e. a different model of the post-randomization factor M, we do not observe this and two G-estimations have higher coverage rate and smaller bias and MSE for all three causal parameters.

In summary, through different sets of simulations, we show that G-estimations perform better than OLS. The baseline covariates are confounding, but they are not sufficient to control no unmeasured confounders. Between the two G-estimations, G-estimation I can be superior for a small sample sizes, even when sequential ignoribability does not hold, and G-estimation II is better with larger sample sizes, and the advantages improve greater when more variability over different levels of baseline covariates of the effect of the randomization on the post-randomization factor.

3.5 Discussion

In the context of assessing the modification of randomized intervention effects on outcome by early post-randomization moderators impacted by the intervention, we investigated the standard linear regression interaction model with OLS and causal structural nested distribution model with G-estimation. Although the interventions

Table 3.5: Simulation results (N=188; $\psi_R = -5.26$, $\psi_M = 1.25$, $\psi_{RM} = -1.53$) when increasing the variability over different levels of the covariate X of the effect of randomization R on post-randomization factor M.

Method	Effect	Coverage Rate	Bias(%)	MSE
OLS	Randomization	92%	0.064(-1.2%)	6.771
	Moderator	$< 1\%$	-0.650(-51.8%)	0.440
	Interaction	$< 1\%$	0.645(-42.2%)	0.433
G-estimation I	Randomization	91%	-0.034(0.6%)	7.753
	Moderator	92%	-0.036(-2.9%)	0.110
	Interaction	93%	0.035(-2.3%)	0.105
G-estimation II	Randomization	92%	-0.031(0.6%)	7.421
	Moderator	98%	-0.010(-0.8%)	0.049
	Interaction	98%	0.009(-0.6%)	0.048

Table 3.6: Simulation results (N=470; $\psi_R = -5.26$, $\psi_M = 1.25$, $\psi_{RM} = -1.53$) when increasing the variability over different levels of the covariate X of the effect of randomization R on post-randomization factor M.

Method	Effect	Coverage Rate	Bias(%)	MSE
OLS	Randomization	94%	0.161(-3.1%)	2.172
	Moderator	$< 1\%$	-0.661(-52.7%)	0.444
	Interaction	$< 1\%$	0.659(-43.1%)	0.441
G-estimation I	Randomization	93%	0.068(-1.3%)	2.345
	Moderator	94%	-0.012(-1.0%)	0.042
	Interaction	94%	0.012(-0.8%)	0.041
G-estimation II	Randomization	93%	0.082(-1.6%)	2.318
	Moderator	95%	0.005(-0.2%)	0.017
	Interaction	99%	0.005(0.3%)	0.017

are randomized, there are unmeasured confounding for the moderator since it is measured after randomization. Standard regression analyses are biased due to unmeasured confounding issue, while causal methods with G-estimation lead to consistent estimators even the unmeasured confounding issue exists.

We further show how to obtain efficient estimators of the parameters of the causal

model. G-estimation I is optimal under the sequential ignorability assumption and G-estimation II is optimal even when sequential ignorability does not hold. G-estimation I is easy to compute and G-estimation II is more complicated. Through the simulations, the efficiency of G-estimation II may be less if the nuisance parameter model is misspecified, e.g., when the sample size is small relative to the number of nuisance parameters, although this will not affect consistency.

Although our results are based on structural distribution models, they may apply for structural mean models, which are less restrictive. In terms of application, our study focus on the post-randomization analysis. However, the similar statistical approach can be applied on mediation analyses, as the example we discussed in the introduction section.

In our further research, we will extend our fixed baseline confounding variables to time-varying confounding. Another future work will be the repeated measures of the post-randomization factor. That is, when the post-randomization factor is a time-varying vector instead of a scalar, how to formulate the model.

Chapter 4

Optimal Dose-Finding Experiments with Correlated Continuous and Discrete Responses

4.1 Introduction

In clinical trials, when multiple endpoints are available, the conventional strategy is to model and analyze each endpoint separately. This approach ignores the information contained in the correlation among the endpoints, and lacks the ability to answer intrinsically multivariate questions (Teixeira-Pinto and Normand, 2009). The better approach is to model and evaluate multiple endpoints simultaneously in clinical trials, which necessitates the use of bivariate (multivariate) dose-response models (Li, Durham, and Flournoy, 1995; Thall and Russell, 1998; Ivanova, 2003; Thall and

Cook, 2004; Bekele and Shen, 2005; Dragalin and Fedorov, 2006; Whitehead et al., 2006; Zhou et al., 2006; Zohar and O'Quigley, 2006; Zohar and Chevret, 2007).

The multiple responses of interest can be continuous or categorical, with some categorical responses coming from categorization of the continuous responses. For example, in a Phase II lung cancer trial, the efficacy endpoint is often measured on a continuous scale, such as shrinkage in tumor size (Karrison et al., 2007). In contrast, the measure of toxicity is often categorical and described by multiple grades of Adverse Events (AEs), such as grade 0 for no AEs, grade 1 to 4 for the severity of the AEs from mild to life-threatening, and grade 5 for death (National Cancer Institute, 1999). If define grade 3 or higher as Dose Limiting Toxicity (DLT), one can convert the toxicity measure to a binary outcome, i.e. DLT/NoDLT (Dignam, Karrison, and Bryant, 2005; Ivanova, 2006). Catalano and Ryan (1992) discussed the bivariate endpoints in a toxicity experiment, in which one endpoint is a continuous variable for fetal weight, and the other endpoint is an unobserved (latent) variable corresponding to malformation. Sammel, Ryan and Legler (1997) discussed multiple mixed continuous and discrete outcomes in a prospective study of the effects of anticonvulsant medication.

One mathematical challenge to model the mixture responses is that there is no obvious multivariate distribution for the mixed variables. Two likelihood-based approaches have been discussed in literature. One is to factorize the joint distribution of the responses as the product of the marginal distribution of one response and the conditional distribution of the second response given the previous response (Pearson,

1900, 1909; Tate, 1955; Cox and Wermuth, 1992; Catalano and Ryan, 1992; Fitzmaurice and Laird, 1995). The other approach is to model the correlation among the multiple outcomes by introducing an unobserved (latent) variable (Arminger and Kuster, 1988; Sammel, Ryan and Legler, 1997). Alternatively, Liang and Zeger (1986), Prentice and Zhao (1991)used separate equations for each outcome and a working correlation matrix to model the correlation among outcomes.

In what follows, we confine ourselves to bivariate responses, i.e. efficacy and toxicity, which are the two primary responses often used in early phase clinical trials, while the proposed method can be easily generalized to multiple correlated endpoints scenarios. The likelihood approach is used to estimate bivariate model with the underlying bivariate normal distribution. This is not only convenient from the statistical point of view, but also appeals to toxicologists because it provides a natural and intuitive framework for the biological mechanism leading to adverse events (Catalano and Ryan, 1992). Essentially, there are three different classes of bivariate responses: both continuous, both categorical, and their mixture. We have investigated the dose-finding designs for the cases when both endpoints are binary or both are continuous (Fedorov and Wu, 2007a; Dragalin, Fedorov, and Wu, 2008) assuming the underlying bivariate normal distribution. In 2007 mODa paper, we briefly discussed the designs for all three cases and their relationship (Fedorov and Wu, 2007b). This paper is the extension of Fedorov, Wu, and Zhang (2010) with more scenarios and technical details for the most difficult case - the mixture of continuous and binary endpoints. Without loss of generality, continuous efficacy and binary toxicity are assumed. Our

primary goal is to implement optimal design techniques under the bivariate mixture of continuous and binary responses in dose-finding experiments.

In dose-finding studies, one of the main goals is to construct the dose-response relationship, i.e. the dose-response model. The estimates of model parameter, the prediction of responses, or both are of interest. As one of the approaches in experimental designs, optimal designs may answer a variety of questions, while the main issues are common: how to choose the dose levels, how to allocate patients to each dose level, and how to estimate the unknown parameters. Within the optimal design framework, general (multivariate) models for continuous responses were discussed in detail in Fedorov and Hackl (1997); for binary responses, several bivariate models have been studied such as Gumbel bivariate model (Heise and Myers, 1996; Dragalin and Fedorov, 2006), Cox bivariate model (Dragalin and Fedorov, 2006), and bivariate probit model (Fedorov and Wu, 2007a; Dragalin, Fedorov, and Wu, 2008). In contrast, little work has been done for mixed responses in optimal designs. Coffey and Glennings (2007) applied D-criterion to design experiments for multiple responses of different types. Their analysis are based on the first and second moment of responses. The choice of the working correlation matrix looks arbitrary, especially taking into account the different structure of responses (binary, count and continuous).

Our approach can be distinguished from the Coffey and Glennings's approach (2007) in several ways. First, we use the bivariate normal distribution to generate a model with mixed types of responses. The use of the underlying normal distribution allows to introduce correlation between responses of different types in a rather

natural way, see Tate (1955). Second, we incorporate ethical concerns and cost constraints. Furthermore, practical designs, such as two-stage and adaptive designs, are constructed and evaluated with extensive Monte Carlo simulations.

The major steps in our approach of optimal designs are:

1. Select a dose-response model and find the information matrix for a single observation;

2. Build a utility function that quantifies the target treatment effects; identify parameters of interest; and address ethical/cost aspects with a penalty function;

3. Identify/quantify the prior information in the design construction if it is available;

4. Select a criterion of optimality;

5. Construct locally optimal designs - they are the benchmarks;

6. If needed, build "practical" designs and compare them with the benchmarks.

This paper is organized as follows: Section 2 introduces the notation, our dose-response model, the information matrix for a single observation, the utility function and the penalty function. Different designs including locally optimal, two stage designs, and adaptive designs are discussed in Section 3. In Section 4, we illustrate our approach and compare various designs via examples and simulations. The paper concludes with a summary.

4.2 Model

4.2.1 Generalized bivariate probit model

Assume that the underlying efficacy and toxicity responses follow a bivariate normal distribution:

$$\mathbf{Z} \sim N(\boldsymbol{\eta}, \boldsymbol{\Sigma}) \tag{4.2.1}$$

where $\mathbf{Z}$ is a vector of responses, $\boldsymbol{\eta}$ is a vector of means, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix. In our discussion, let the first response Z_1 be efficacy, the second Z_2 be toxicity, $\boldsymbol{\eta} = (\eta_1, \eta_2)^T$, and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Responses can be observed either directly or indirectly, i.e. only some functions of them are available.

To link the results of this paper with previous publications (Dragalin and Fedorov, 2006; Dragalin, Fedorov, and Wu, 2008), we remind the reader that when responses are both binary, they can be described by a contingency table (see Table 4.1).

Table 4.1: Binary efficacy and binary toxicity.

		Toxicity 1	Toxicity 0	
Efficacy	1	p_{11}	p_{10}	$p_{1.}$
	0	p_{01}	p_{00}	$p_{0.}$
		$p_{.1}$	$p_{.0}$	1

Let Y_1 and Y_2 denote the binary responses for efficacy and toxicity, respectively.

They can be modelled by the bivariate probit model as dichotomizations of Z_1 and Z_2 from the bivariate normal distribution (Lesaffre and Molenberghs, 1991; Fedorov and Wu, 2007; Dragalin, Fedorov, and Wu, 2008); the probability of no efficacy and no toxicity is

$$p_{00} = P(Y_1 = 0, Y_2 = 0) = F(v_1, v_2; \mathbf{\Sigma}^*) = \int_{-\infty}^{v_2} \int_{-\infty}^{v_1} \frac{1}{2\pi|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{v}^T\mathbf{\Sigma}^{-1}\mathbf{v}\right\} d\mathbf{v}, \tag{4.2.2}$$

where $\mathbf{v} = (v_1, v_2)^T$, $v_k = (c_k - \eta_k)/\sigma_k$, $Y_k = I(Z_k > c_k)$, $k = 1$ or 2, $\mathbf{\Sigma}^* = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, c_k is the (known) cut-off point, ρ is the correlation between two responses, and σ_1, σ_2, η_1, and η_2 are parameters of marginal normal distributions.

The marginal probabilities for efficacy and toxicity are defined as $p_{1.} = P(Y_1 = 1) = 1 - F(v_1)$ and $p_{.1} = P(Y_2 = 1) = 1 - F(v_2)$ respectively, where $F(\cdot)$ is the cumulative density function of standard normal distribution. Note that Table 1 has only three independent entries. For instance, given p_{11}, $p_{.1}$ and $p_{1.}$, other probabilities are $p_{10} = p_{1.} - p_{11}$, $p_{01} = p_{.1} - p_{11}$, and $p_{00} = 1 - p_{1.} - p_{.1} + p_{11}$.

For mixed responses such as continuous efficacy (Y_1) and binary toxicity (Y_2), define

$$Y_1 = Z_1, \quad Y_2 = \begin{cases} 1, & \text{if } Z_2 \geq c_2 \\ 0, & \text{otherwise.} \end{cases} \tag{4.2.3}$$

The bivariate mixed responses Y_1 and Y_2 are now described in Table 2. The marginal probability of toxicity is $p_{.1} = P(Y_2 = 1) = \int P(Y_2 = 1|Y_1 = y_1)\varphi(y_1)dy_1$, where $\varphi(y_1)$ is the marginal probability density of efficacy Y_1, i.e. the probability

density of normal distribution with mean η_1 and standard deviation σ_1. We denote the conditional probability of Y_2 given Y_1 as $p_{1|y_1} = P(Y_2 = 1|Y_1 = y_1) = 1 - F(u_2)$, where $u_2 = (v_2 - \rho(y_1 - \eta_1)/\sigma_1)/\sqrt{1-\rho^2}$ and $v_2 = (c_2 - \eta_2)/\sigma_2$. Note that in Table 4.2, as compared to Table 4.1, the marginal probability of efficacy in $(y_1, y_1 + dy_1)$ is $P(y_1 < Y_1 < y_1 + dy_1) = \varphi(y_1)dy_1$.

Table 4.2: Continuous efficacy and binary toxicity.

		Toxicity		
		1	0	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Efficacy	$(y_1, y_1 + dy_1)$	$p_{1\|y_1}\varphi(y_1)dy_1$	$p_{0\|y_1}\varphi(y_1)dy_1$	$\varphi(y_1)dy_1$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$p_{\cdot 1}$	$p_{\cdot 0}$	1

Given the unknown parameters $\boldsymbol{\vartheta} = (\eta_1, v_2, \rho, \sigma_1)^T$, the likelihood for a single observation $(Y_1 = y_1, Y_2 = y_2)$ can be expressed as

$$L(y_1, y_2; \boldsymbol{\vartheta}) = [p_{1|y_1}]^{y_2}[1 - p_{1|y_1}]^{1-y_2}\varphi(y_1), \tag{4.2.4}$$

and the log-likelihood for a single observation is:

$$l(y_1, y_2; \boldsymbol{\vartheta}) \quad \propto \quad y_2 \log\{1 - F(u_2)\} + (1 - y_2)\log\{F(u_2)\} - \log\sigma_1 - \frac{(y_1 - \eta_1)^2}{2\sigma_1^2}. \tag{4.2.5}$$

For the model given by (4.2.3), (4.2.4) and (4.2.5), we call $\boldsymbol{\vartheta} = (\eta_1, v_2, \rho, \sigma_1)^T$ the elemental parameters. In practice, these parameters may depend on some covariates

such as doses of various compounds (drugs), age, gender, etc. Note that although η_1 and σ_1 can be estimated separately, their counterparts η_2 and σ_2 cannot. Only $(c_2 - \eta_2)/\sigma_2$ is estimable (Dragalin, Fedorov, and Wu, 2008).

4.2.2 Information Matrix for a Single Observation

In experimental design, the information matrix plays a crucial role since it is the basis for the formulation of the optimality criterion (cf. Fedorov and Hackl, 1997). Because the information matrix of independent observations is the sum of the information matrices of the single observations, the derivation of the information matrix for a single observation is discussed.

Elemental Information Matrix

For a single observation, the information matrix for model (4.2.3) is (e.g., Tate, 1955):

$$\boldsymbol{\mu}[\boldsymbol{\vartheta}(\boldsymbol{\theta})] = \begin{pmatrix} \frac{1-\rho^2+\rho^2 a_0}{\sigma_1^2(1-\rho^2)} & \frac{\rho a_0}{\sigma_1(1-\rho^2)} & \frac{\rho(\rho v_2 a_0 - a_1)}{\sigma_1(1-\rho^2)^2} & \frac{\rho^2 a_1}{\sigma_1^2(1-\rho^2)} \\ - & \frac{a_0}{(1-\rho^2)} & \frac{\rho v_2 a_0 - a_1}{(1-\rho^2)^2} & \frac{\rho a_1}{\sigma_1(1-\rho^2)} \\ - & - & \frac{a_2 - 2\rho v_2 a_1 + \rho^2 v_2^2 a_0}{(1-\rho^2)^3} & \frac{\rho(\rho v_2 a_1 - a_2)}{\sigma_1(1-\rho^2)^2} \\ - & - & - & \frac{2(1-\rho^2)+\rho^2 a_2}{\sigma_1^2(1-\rho^2)} \end{pmatrix}, \qquad (4.2.6)$$

where

$$a_k(v_2, \rho) = \int_{-\infty}^{+\infty} \frac{t^k \varphi(t) \varphi^2(\frac{v_2 - \rho t}{\sqrt{1-\rho^2}})}{F(\frac{v_2-\rho t}{\sqrt{1-\rho^2}}) \left[1 - F(\frac{v_2-\rho t}{\sqrt{1-\rho^2}})\right]} \, dt \quad \text{and} \quad k = 0, 1, 2, \qquad (4.2.7)$$

and "$-$" in the lower left of (4.2.6) stands for the corresponding element of the symmetric matrix.

In (4.2.7), except when $\rho = 0$, we have to use numerical integration to find a_k. If Σ is known, then $\boldsymbol{\vartheta} = (\eta_1, v_2)^T$ and their information matrix is the upper left 2×2 submatrix of (4.2.6).

Information matrix for regression models

To move from the elemental parameters $\boldsymbol{\vartheta}$ to parameters $\boldsymbol{\theta}$ for regression models, recall that if $\boldsymbol{\vartheta} \in R_{m'}$ is a continuous function of $\boldsymbol{\theta} \in R_m$ then

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{J}\boldsymbol{\mu}[\boldsymbol{\vartheta}(\boldsymbol{\theta})]\mathbf{J}^T, \tag{4.2.8}$$

where matrix $\mathbf{J}$ is the $m \times m^{'}$ Jacobian matrix of transformation $\boldsymbol{\vartheta}_{=}\boldsymbol{\vartheta}(\boldsymbol{\theta})$ (cf. Lehmann 1983):

$$\mathbf{J} = \frac{\partial \boldsymbol{\vartheta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \left|\left| \frac{\partial \vartheta_\beta(\theta)}{\partial \theta_\alpha} \right|\right|_{\alpha=1,\beta=1}^{m,m^{'}}. \tag{4.2.9}$$

If we assume that $\eta_1 = \boldsymbol{\theta}_1^T \mathbf{f}_1(x)$ and $v_2 = \frac{c_2 - \eta_2}{\sigma_2} = \boldsymbol{\theta}_2^T \mathbf{f}_2(x)$, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \rho, \sigma_1)^T$, then

$$\mathbf{J} = \begin{pmatrix} \mathbf{f}_1(x) & 0 & 0 & 0 \\ 0 & \mathbf{f}_2(x) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1-\rho^2+\rho^2 a_0}{\sigma_1^2(1-\rho^2)}\mathbf{f}_1\mathbf{f}_1^T & \frac{\rho a_0}{\sigma_1(1-\rho^2)}\mathbf{f}_1\mathbf{f}_2^T & \frac{\rho(\rho v_2 a_0 - a_1)}{\sigma_1(1-\rho^2)^2}\mathbf{f}_1 & \frac{\rho^2 a_1}{\sigma_1^2(1-\rho^2)}\mathbf{f}_1 \\ - & \frac{a_0}{(1-\rho^2)}\mathbf{f}_2\mathbf{f}_2^T & \frac{\rho v_2 a_0 - a_1}{(1-\rho^2)^2}\mathbf{f}_2 & \frac{\rho a_1}{\sigma_1(1-\rho^2)}\mathbf{f}_2 \\ - & - & \frac{a_2 - 2\rho v_2 a_1 + \rho^2 v_2^2 a_0}{(1-\rho^2)^3} & \frac{\rho(\rho v_2 a_1 - a_2)}{\sigma_1(1-\rho^2)^2} \\ - & - & - & \frac{2(1-\rho^2)+\rho^2 a_2}{\sigma_1^2(1-\rho^2)} \end{pmatrix}. \tag{4.2.10}$$

Whenever it does not lead to a confusion, we omit x or $\boldsymbol{\theta}$ in our notation, but one should remember that in (4.2.10), v_2, a_1, a_2, $\mathbf{f}_1$, and $\mathbf{f}_2$ all depend on x. (4.2.10) is the information matrix for all unknown parameters in the regression model. In this paper, we assume that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are always unknown, σ_1 is always known, and ρ is either known or unknown. Accordingly, $\boldsymbol{\theta}$ is $(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ or $(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \rho)^T$, and the information matrix is the upper left 2×2 or 3×3 block matrix in (4.2.10). The true dimension of the matrix will depend on the dimensions of vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The parameterization for the binary toxicity response is taken to be $v_2 = \boldsymbol{\theta}_2^T \mathbf{f}_2(x)$, since the probability of toxicity is fully determined by v_2 in our model. The information matrix for different parameterizations, such as $\eta_2 = \boldsymbol{\theta}_2^T \mathbf{f}_2(x)$ (under known c_2 and σ_2), can be easily obtained through the Jacobian transformation.

4.2.3 Utility Function

A utility function needs to be specific to the target of a particular dose-finding study. It can coincide with the dose-response probability, or be other function of unknown parameters. Fedorov and Wu (2007) suggested a utility function based on the joint probability of efficacy without toxicity, i.e. $p_{10}(x, \boldsymbol{\theta}) = P(Y_1 = 1 | Y_2 = 0) P(Y_2 = 0)$, and defined the best dose as $x^* = \arg\max_{x \in \mathfrak{X}} p_{10}(x, \boldsymbol{\theta})$. Similar ideas can be found in Li, Durham, and Flournoy (1995); Kpamegan and Flournoy (2001); Zohar and O'Quigley (2006); Ivanova (2006). Another utility could be a "distance" function which measures the distance of the response probabilities from a "desirable" point (Thall and Cook, 2004; Dragalin, Fedorov, and Wu, 2008).

We emphasize that although the utility function $p_{10}(x, \boldsymbol{\theta})$ is based on dichotomized responses, this does not mean that either efficacy or toxicity should be dichotomized during, or before statistical analysis. Dichotomization during the analysis may lead to significant loss of information. For example, when dichotomizing a normal distribution by a cut-off point, the loss in terms of (Fisher's) information is at least 36% (Fedorov, Mannino, and Zhang, 2009). Whenever possible, one should generate the binary utility only when analysis is completed (Fedorov and Wu, 2007).

To obtain a utility function for mixed responses of continuous efficacy and binary toxicity that is consistent with $p_{10}(x, \boldsymbol{\theta})$, we define our utility function as the product of the mean of efficacy without toxicity multiplied by the probability of having no toxicity:

$$\begin{aligned} \zeta(x, \boldsymbol{\theta}) &= E(Y_1|Y_2 = 0)P(Y_2 = 0) \\ &= \eta_1 F(v_2) - \rho\sigma_1\varphi(v_2) = \boldsymbol{\theta}_1^T \mathbf{f}_1(x) F(\boldsymbol{\theta}_2^T \mathbf{f}_2(x)) - \rho\sigma_1\varphi(\boldsymbol{\theta}_2^T \mathbf{f}_2(x)). \end{aligned} \tag{4.2.11}$$

The best dose is defined as the dose whose utility reaches the maximum of (4.2.11) within the design region:

$$x^*(\boldsymbol{\theta}) = \arg\max_{x \in \mathfrak{X}} \zeta(x, \boldsymbol{\theta}). \tag{4.2.12}$$

Of course any of the utility functions described in Fedorov and Wu (2007) can be used for the model in (4.2.3), but (4.2.11) shows the flexibility of this approach.

4.2.4 Penalty function

In drug development studies, there are always ethical concerns and cost constraints associated with different doses. These constraints can be quantified by a penalty function $\phi(x, \boldsymbol{\theta})$. Dragalin and Fedorov (2006) made one of the first attempts to address the penalized optimal design problem in the dose-finding context. For two binary responses, they introduced a penalty function involving both efficacy and toxicity, i.e. a function of the probabilities of efficacy and toxicity. Lai and Robbins (1978, 1982) introduced a cost function $\phi(x, \theta) = (x - x^*)^2$, where x^* is the target dose. A target dose, for instance, could be MTD in phase I trials or MSD in phase II trials. In our study, we use the following penalty functions similar to the cost function of Lai and Robbins (1978):

$$\phi(x, \boldsymbol{\theta}) = r(x - x^*(\boldsymbol{\theta}))^2 + c, \tag{4.2.13}$$

where $x^*(\boldsymbol{\theta}) = \arg\max_{x \in \mathfrak{X}} \zeta(x, \boldsymbol{\theta})$ is the "best dose" and $\zeta(x, \boldsymbol{\theta})$ is the utility function defined in (4.2.11). The parameter r quantifies the risk of the dose. The larger the value of r, the more penalty is added to the doses which are far away from the best dose. The traditional optimal design can be viewed as a special case of penalized optimal design, in which one adds constant penalty across the entire design region, i.e., $r = 0$ in (4.2.13). The parameter c can be considered as the cost for each observation or unit cost. Since every observation has a cost, c should be positive. It can be shown that optimal design characteristics depend only on the ratio of c/r.

The left plot in Figure 4.1 shows efficacy, toxicity, utility and penalty functions

under our model. The true parameters of the model in this example were estimated from a clinical trial of prevention of venous thromboembolism (VTE) in total joint replacement to assess new anticoagulants (see details in Dragalin, Fedorov, and Wu, 2008). The efficacy is the lowered VTE incidence rate and the toxicity is having major bleeding event. For the parameters of the penalty function, $c = 1$ and $r = 10$ are used. Under our model, the mean efficacy η_1 and the probability of toxicity $p_{\cdot 1}$ increase as the doses increase. The utility function, or "success" curve $\zeta(x, \boldsymbol{\theta})$, increases first and then decreases. The penalty function $\phi(x, \boldsymbol{\theta})$ reaches the minimum when the utility function reaches its maximum. The "best dose" $X^*(\boldsymbol{\theta})$ is the maxima of the "success" curve.

The utility functions for the different values of ρ are shown in the right plot of Figure 4.1. As ρ increases from -0.9 to 0.9, the best dose $X^*(\boldsymbol{\theta})$ decreases from 1.2 to 1.0. In addition, the plot shows that when the dose level is less than 0.6, the utility function almost linearly depends on the dose and is not very sensitive to ρ. This is because when dose is low (< 0.6), the probability of toxicity $p_{\cdot 1}$ does not change much from zero (see left in Figure 4.1), and the effect of efficacy dominates the utility function. When the dose is high, the probability of toxicity significantly increases as the dose increases. Both efficacy and toxicity affect the utility and, therefore, the utility is more sensitive to the correlation between efficacy and toxicity.

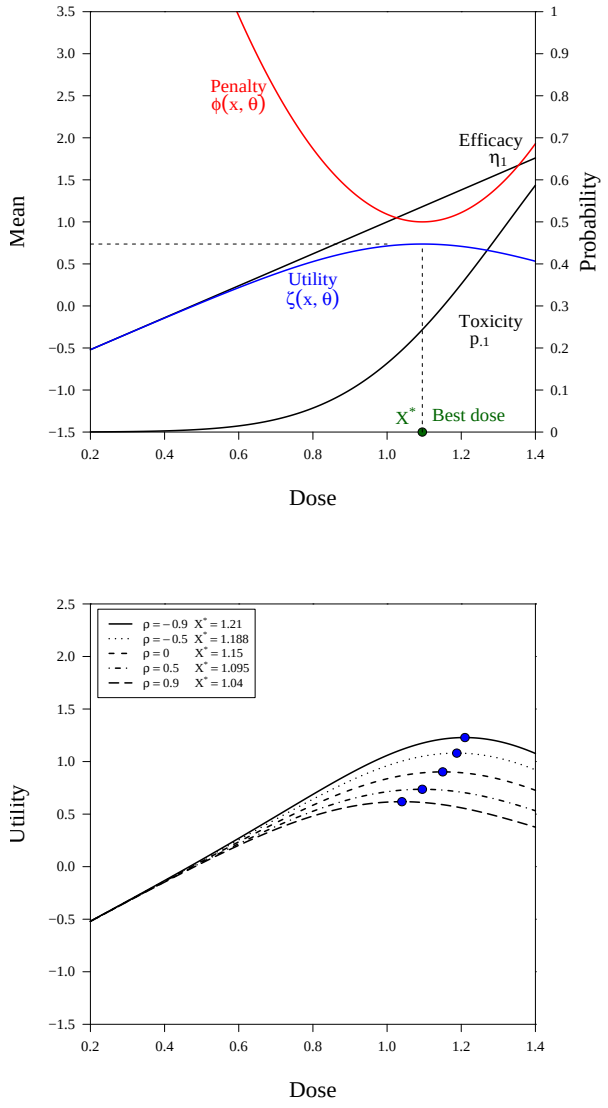

Figure 4.1: Under the mixed responses model with $\boldsymbol{\theta} = (-0.9, 1.9, 3.98, -3)$, $\sigma_1 = 1$: (1) Left plot: relationship between efficacy η, toxicity $p_{1.}$, utility $\zeta(x, \boldsymbol{\theta})$ and penalty $\phi(x, \boldsymbol{\theta})$ with $\rho = 0.5$. The left y-axis is for efficacy, utility, and penalty; the right y-axis is for toxicity; (2) Right plot: utility functions with different ρ's.

4.3 Optimal Designs

4.3.1 Locally optimal designs

Assume that the sample size at design point x_i in the design region $\mathfrak{X}$ is n_i, and the weight for x_i is $\lambda_i = n_i/N$, where $i = 1, \ldots, k$, $N = \sum_{i=1}^{k} n_i$. Let $\xi = \{x_i, \lambda_i\}_1^k$ denote the design. From (4.2.10) and additivity of information matrices for independent observations, we have

$$\mathbf{M}(\xi, \boldsymbol{\theta}) = \sum_{i=1}^{k} \lambda_i \boldsymbol{\mu}(x_i, \boldsymbol{\theta}) = N^{-1} \sum_{i=1}^{k} n_i \boldsymbol{\mu}(x_i, \theta) = \mathbf{M}_N(\xi, \boldsymbol{\theta})/N, \tag{4.3.1}$$

where $\mathbf{M}(\xi, \boldsymbol{\theta})$ is called normalized information matrix and $\mathbf{M}_N(\xi, \boldsymbol{\theta})$ is the information matrix.

Under mild regularity conditions (cf. Rao 1973, Ch. 4a), the Maximum Likelihood Estimator (MLE) is strongly consistent and $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})/\sqrt{N}$ is asymptotically ($N \to \infty$) normal with zero mean and variance-covariance matrix $\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})$.

We can define penalized locally optimal design as in Dragalin and Fedorov (2006), Dragalin, Fedorov, and Wu (2008):

$$\xi^*(\boldsymbol{\theta}) = \arg\min_{\xi} \Psi[\mathbf{M}(\xi, \boldsymbol{\theta})/\Phi(\xi, \boldsymbol{\theta})], \tag{4.3.2}$$

where Ψ is some convex function called the criterion of optimality, and $\Phi(\xi, \boldsymbol{\theta})$ is the total penalty normalized by the total sample size with the following definition,

$$\Phi(\xi, \boldsymbol{\theta}) = \int_{\mathfrak{X}} \phi(x, \boldsymbol{\theta}) \xi(dx). \tag{4.3.3}$$

Note that matrix $\mathbf{M}(\xi, \boldsymbol{\theta})/\Phi(\xi, \boldsymbol{\theta})$ may be viewed as information gained per unit

of penalty. Indeed, let

$$\Phi_N(\xi, \boldsymbol{\theta}) = \sum_{i=1}^{k} n_i \phi(x_i, \boldsymbol{\theta}) = N\Phi(\xi, \boldsymbol{\theta}), \tag{4.3.4}$$

then the information gained per unit of penalty is

$$\frac{\mathbf{M}_N(\xi, \boldsymbol{\theta})}{\Phi_N(\xi, \boldsymbol{\theta})} = \frac{N\mathbf{M}(\xi, \boldsymbol{\theta})}{N\Phi(x, \boldsymbol{\theta})} = \frac{\mathbf{M}(\xi, \boldsymbol{\theta})}{\Phi(\xi, \boldsymbol{\theta})}. \tag{4.3.5}$$

The choice of the criterion of optimality Ψ, is driven by the study goals. For instance, it could be the determinant of the variance-covariance matrix of all estimated parameters, the variance of the estimator of the best dose, or the variance of the estimator of the parameter of the utility function, etc. The former is D-optimality, which is well described in the statistical literature (e.g., Fedorov and Hackl, 1997). If estimation of the best dose is the only goal, for sufficiently large samples, it is necessary to minimize the variance of its estimator as a function of unknown parameters. This variance can be viewed as a generalized version of L-criterion (Dragalin, Fedorov, and Wu, 2008). In our study, both the best dose and the utility function involve all unknown parameters in the model; therefore, we choose D-optimality as our design strategy to obtain the overall accuracy.

The sensitivity function $\psi(x, \xi, \theta)$ is related to the directional derivative of the design criterion Ψ, and completely determines the location of the support points of the optimal design ξ^* (cf. Fedorov and Hackl, 1997). For penalized locally D-optimal design, the design criterion is

$$\xi^*(\boldsymbol{\theta}) = \arg\min_{\xi} \log |\mathbf{M}(\xi, \boldsymbol{\theta})/\Phi(\xi, \boldsymbol{\theta})|^{-1}, \tag{4.3.6}$$

and the sensitivity function is

$$\psi(x, \xi, \boldsymbol{\theta}) = \text{tr}[\boldsymbol{\mu}(x, \boldsymbol{\theta})\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})] - m\phi(x, \boldsymbol{\theta})/\Phi(\xi, \boldsymbol{\theta}), \tag{4.3.7}$$

where m is the number of unknown parameters in the model.

A design ξ^* is locally D-optimal if and only if the following inequality holds for $\forall x \in \mathfrak{X}$

$$\psi(x, \xi^*, \boldsymbol{\theta}) \leq 0, \tag{4.3.8}$$

where the equality holds for all $x's$ that are support points of ξ^*. Inequality (4.3.8) can be viewed as another generalized version of the Kiefer-Wolfowitz general equivalence theorem (Fedorov and Hackl, 1997). We would like to emphasize again that although locally optimal designs have restrictions in practice, they are important to calibrate other designs like adaptive designs.

Numerical algorithm

To construct the optimal design, we use the first order exchange algorithm (Fedorov and Wu, 2007). It is an iterative procedure that shuffles the design points within the design region in order to improve the design criterion. At the s^{th} iteration, a "good" (the most informative) design point from the candidate set with certain weight is added into the current design point set. This (forward) step can be expressed as

$$x_s^+ = \arg\max_{x \in \mathfrak{X}} \psi(x, \xi_{s-1}, \boldsymbol{\theta}). \tag{4.3.9}$$

In the second (backward) step, a "bad" (the least informative) design point with certain weight is deleted from the current design:

$$x_s^- = \arg\min_{x \in \mathfrak{X}} \psi(x, \xi_{s-1}, \boldsymbol{\theta}). \tag{4.3.10}$$

As $s \to \infty$, ξ_s converges to locally optimal designs. The algorithm stops when the design criterion cannot be improved (Fedorov and Hackl, 1997).

4.3.2 Two-stage Designs

For non-linear models, the information matrices $\mathbf{M}(\xi, \boldsymbol{\theta})$ in (4.2.6) and (4.2.10) depend on the unknown parameters $\boldsymbol{\theta}$. Therefore, the locally optimal design depends on the unknown parameters $\boldsymbol{\theta}$ and cannot be implemented in practice. Dragalin, Fedorov, and Wu (2008) proposed a penalized optimal two-stage design under the bivariate probit model for practical dose-finding experiments.

The idea is that at the initial design stage, the researcher collects N_0 observations based on design ξ_0 and obtains the initial estimates for unknown parameters, $\widehat{\boldsymbol{\theta}}_0$. Then in the second design stage, $\widehat{\boldsymbol{\theta}}_0$ are treated as the "true" parameters, on which locally optimal designs or penalized locally optimal designs are constructed for the remaining N_1 subjects. At the end of the second stage, the unknown parameters are re-estimated using all $N_0 + N_1$ observations.

Assuming that the information matrix for ξ_0 is $N_0\mathbf{M}(\xi_0, \boldsymbol{\theta}) \simeq \Sigma_0^{-1}$, the D-optimal design for the second stage (see derivations of (4.3.11) and (4.3.12) in the appendix)

is

$$\xi^*(\boldsymbol{\theta}) = \arg\min_{\xi} \log \left| \frac{\pi \mathbf{M}(\xi_0, \boldsymbol{\theta}) + (1-\pi)\mathbf{M}(\xi, \boldsymbol{\theta})}{\pi \Phi(\xi_0, \boldsymbol{\theta}) + (1-\pi)\Phi(\xi, \boldsymbol{\theta})} \right|^{-1}, \tag{4.3.11}$$

where $\pi = N_0/(N_0 + N_1)$.

The necessary and sufficient condition for D-optimality of ξ^* (see Appendix) is

$$\begin{aligned} &\operatorname{tr}\left\{\boldsymbol{\mu}(x, \boldsymbol{\theta}) \left[\pi \mathbf{M}(\xi_0, \boldsymbol{\theta}) + (1-\pi)\mathbf{M}(\xi^*, \boldsymbol{\theta})\right]^{-1}\right\} - \frac{m\phi(x, \boldsymbol{\theta})}{\pi \Phi(\xi_0, \boldsymbol{\theta}) + (1-\pi)\Phi(\xi^*, \boldsymbol{\theta})} \\ &\leq \operatorname{tr}\left\{\mathbf{M}(\xi^*, \boldsymbol{\theta}) \left[\pi \mathbf{M}(\xi_0, \boldsymbol{\theta}) + (1-\pi)\mathbf{M}(\xi^*, \boldsymbol{\theta})\right]^{-1}\right\} - \frac{m\Phi(\xi^*, \boldsymbol{\theta})}{\pi \Phi(\xi_0, \boldsymbol{\theta}) + (1-\pi)\Phi(\xi^*, \boldsymbol{\theta})}. \end{aligned} \tag{4.3.12}$$

The first order iterative algorithm ((4.3.9) and (4.3.10)) is used to construct the optimal design in the second stage, where the sensitivity function $\psi(x, \xi, \boldsymbol{\theta})$ is defined as

$$\psi(x, \xi, \boldsymbol{\theta}) = \operatorname{tr}\left\{\boldsymbol{\mu}(x, \boldsymbol{\theta}) \left[\pi \mathbf{M}(\xi_0, \boldsymbol{\theta}) + (1-\pi)\mathbf{M}(\xi, \boldsymbol{\theta})\right]^{-1}\right\} - \frac{m\phi(x, \boldsymbol{\theta})}{\pi \Phi(\xi_0, \boldsymbol{\theta}) + (1-\pi)\Phi(\xi, \boldsymbol{\theta})}.$$

In an actual two-stage design, the true parameters $\boldsymbol{\theta}$ are replaced by their estimates $\hat{\boldsymbol{\theta}}_0$ after the analysis of the initial stage data. In (4.3.11) and (4.3.12), $\hat{\boldsymbol{\theta}}_0$ and $M(\xi_0, \widehat{\boldsymbol{\theta}}_0)$ (or Σ_0) may be viewed as a prior information that, for instance, has been accumulated via other studies or existing publications related to the considered drug.

Note that for the actual two-stage design, the matrix $\pi \mathbf{M}(\xi_0, \hat{\boldsymbol{\theta}}_0) + (1-\pi) M(\xi^*, \hat{\boldsymbol{\theta}}_0)$ is not the exact normalized information matrix any more because ξ^* depends on $\hat{\boldsymbol{\theta}}_0$. However, on intuitive level we can make the following conjecture. If the initial design ξ_0 has a regular information matrix for any $\boldsymbol{\theta} \in R_m$, where the admissibility set R_m includes the true value $\boldsymbol{\theta}_{true}$ as an internal point, then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\theta}}(N_0)$ will be strongly consistent, i.e. converges almost surely

to $\boldsymbol{\theta}_{true}$ when $N_0 \to \infty$. Consequently, the sensitivity function $\psi(x, \xi, \hat{\boldsymbol{\theta}}_0)$ will converge almost surely to $\psi(x, \xi, \boldsymbol{\theta}_{true})$ uniformly with respect to x and ξ. Obviously some smoothness of $f_1(x)$ and $f_2(x)$ is needed together with the compactness of $\mathfrak{X}$ (Rao, 1973, Ch. 2c). Consequently, the solution $\xi^*(\hat{\boldsymbol{\theta}}_0)$ will converge to $\xi^*(\boldsymbol{\theta}_{true})$ and $\pi\mathbf{M}(\xi_0, \hat{\boldsymbol{\theta}}_0) + (1-\pi)M(\xi^*, \hat{\boldsymbol{\theta}}_0)$ will converge to the "true" information matrix $\pi\mathbf{M}(\xi_0, \boldsymbol{\theta}_{true}) + (1-\pi)M(\xi^*, \boldsymbol{\theta}_{true})$. Of course the above statement is only a conjecture without any rigorous mathematical proof and that is why we resort to Monte-Carlo simulations to confirm its validity for our specific case.

4.3.3 Fully Adaptive Designs

An alternative design is the fully adaptive design. Similar to the two-stage designs, the researcher obtains the initial estimates for unknown parameters. For fully adaptive designs, the researcher then constantly updates estimates of unknown parameters when new observations come in, and stops when the procedure converges or all the resources are consumed. Box and Hunter (1965) proposed sequential assignments by maximizing a sensitivity function; Dragalin and Fedorov (2006) proposed this strategy for two correlated binary responses in the dose-response studies. In each step of the iterative algorithm, given a penalty function, we assign the patient to the dose which maximizes the sensitivity function (e.g., Dragalin and Fedorov, 2006), i.e.,

$$x_N = \arg\max_{x \in \mathfrak{X}} \psi(x, \xi_{N-1}, \hat{\boldsymbol{\theta}}_N). \tag{4.3.13}$$

Note that (4.3.13) coincides with the forward step of the first order algorithm

(4.3.9) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}_N$.

Similar to two-stage designs, in fully adaptive designs, the observations are not independent. Since ξ_N is random, $\pi\mathbf{M}(\xi_0, \boldsymbol{\theta}_{true}) + (1-\pi)\mathbf{M}(\xi_N^*, \boldsymbol{\theta}_{true})$ is random and hence it is not the actual information matrix. However, the distribution of each element of this matrix after "large" number of observations heavily gravitates towards to the corresponding element in $\pi\mathbf{M}(\xi_0, \boldsymbol{\theta}_{true}) + (1-\pi)\mathbf{M}(\xi^*, \boldsymbol{\theta}_{true})$, see the results of simulation, Lai (2001), and Rosenberger and Hughes-Oliver (1999).

4.4 Examples

In this section, we construct locally optimal, two-stage designs and fully adaptive designs. The performances among different designs are assessed via simulation studies.

For illustration and comparison purposes, the true parameters of the model are the same as in the prevention of VTE trials (Dragalin, Fedorov, and Wu, 2008). The efficacy endpoint is the lowered VTE incidence rate and the toxicity endpoint is having major bleeding event. In the continuous-binary model given by (4.2.3)-(4.2.5), we assume that σ_1 is a known parameter with a value of 1, while ρ is either known (Section 4.1 to 4.2) or unknown (Section 4.3). For the other two elemental parameters η_1 and v_2, we assume that

$$\eta_1 = \theta_{11} + \theta_{12}x, \quad \text{and} \quad v_2 = (c_2 - \eta_2)/\sigma_2 = \theta_{21} + \theta_{22}x,$$

where $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = (-0.9,\ 1.9,\ 3.98, -3)$. The design region is restricted to [0.2,

1.4]. A moderate correlation between efficacy and toxicity, i.e. $\rho = 0.5$, is assumed unless stated otherwise (see Figure 1). Note that the sensitivity function ψ (4.3.7) depends on the ratio $\phi(x, \boldsymbol{\theta})/\Phi(\xi, \boldsymbol{\theta})$, which depends on c/r (not c and r separately). Thus without loss of generality, we select $c = 1$ and discuss different scenarios that $r = 0$, 1, 10 and 100.

4.4.1 Locally optimal designs

Locally optimal designs are benchmarks for other designs. Under our model, the locally D-optimal design may be a two-point or three-point design depending on the parameter r (see the left plot in Figure 4.2). When $r = 0$, locally D-optimal design is a three-point design having two points at the boundaries of the design region [0.2, 1.4] with weight 20% and 47%, respectively, and one point in the middle with weight 33%. When r is small, the optimal design is a three-point design that is close to the design with $r = 0$. As r gets larger, the optimal design puts higher weights on the points around the best dose ($X^*(\boldsymbol{\theta}) = 1.095$) and may reduce to a two-point design. We found that the threshold for r between two and three design points for the locally optimal design is around $r = 2.6$ under our model. For instance, when $r = 10$, the two optimal dose levels are 0.93 and 1.4, with corresponding weights 60% and 40%. When $r = 100$, the two optimal dose levels (1.06 and 1.35) are closer to the best dose.

Let us define the relative efficiency as

$$\left| \frac{M(\xi', \boldsymbol{\theta})/\Phi(\xi', \boldsymbol{\theta})}{M(\xi, \boldsymbol{\theta})/\Phi(\xi, \boldsymbol{\theta})} \right|^{1/m}. \tag{4.4.1}$$

A value of (4.4.1) less than one indicates that design ξ is more efficient than design ξ' with respect to the D criterion, i.e. a smaller cost is needed for ξ than ξ' to achieve the same precision. This metric is model and penalty dependent.

The left table in Table 4.3 lists the relative efficiency for the four locally optimal designs (see the left plot in Figure 4.2). The value of r' is the assumed penalty parameter used to build designs, and the value of r is the "actual" penalty parameter to evaluate the efficiency. For example, the relative efficiency for the designs with $r' = 1$ and $r = 10$ is 0.79. This number shows the efficiency of the design built under the assumption that $r = 1$ if the actual $r = 10$. The first row in Table 4.3 indicates that the efficiency of the traditional locally optimal design built under the assumption that $r = 0$, monotonically decreases as r increased. The table allows to evaluate the robustness of the design with respect to potential uncertainty in cost selection.

4.4.2 Two-stage Designs

Because we know little about the model parameters, at the first stage of the two-stage designs N_0 patients are allocated according to a uniform design ξ_0 with five equally spaced doses {0.2, 0.5, 0.8, 1.1, 1.4 } and equal weights. Note that the first stage design can be any design that the investigator feels comfortable with. The responses collected in the first stage are used to obtain the preliminary model parameters estimates $\widehat{\boldsymbol{\theta}}_0$. In the second stage, penalized locally D-optimal designs are built for the remaining N_1 patients by treating $\widehat{\boldsymbol{\theta}}_0$ as the "true" parameters.

Examples of the two-stage designs with different penalty parameters are shown in

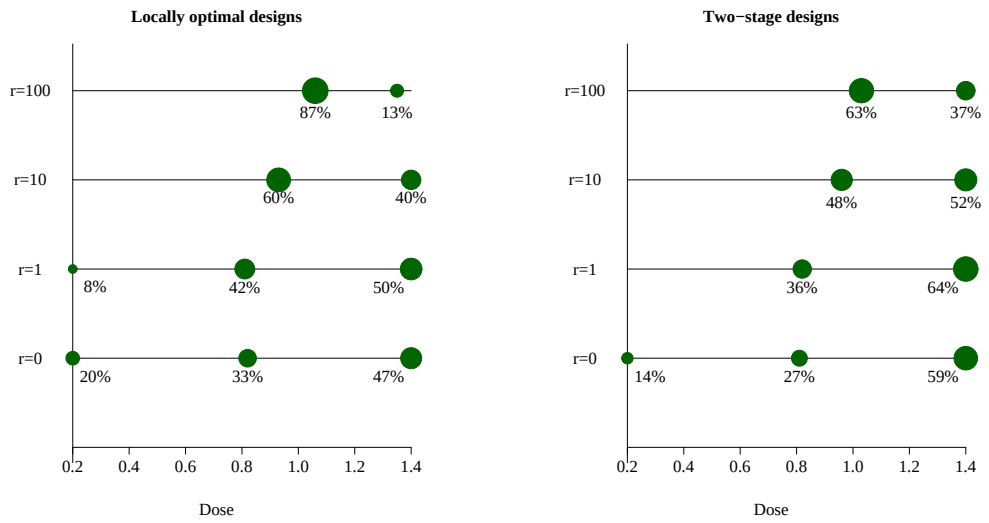

Figure 4.2: Allocation of the doses for optimal designs built with different values of r in the penalty function. The size of each point represents the corresponding weight which is labelled below each point.**Left:** locally optimal designs; **Right**: the second stage designs in the two-stage designs. True values of the unknown model parameters are used in the second stage and five-point uniform design is used in the initial stage.

the right plot in Figure 4.2, in which the second stage designs are constructed based on the true values of the unknown parameters. Similar to the left table, the right table in Table 4.3 shows the relative efficiency for the optimal two-stage designs with various values of r' and r. Again, the efficiency of the optimal two-stage design built under the assumption that $r = 0$ drops as the actual r increases. Compared to the left table, the relative efficiency is higher since the same uniform design is used for the initial stage for all the two-stage designs in the examples.

One thousand simulations are performed for the two-stage designs with $r = 0$,

Table 4.3: The relative efficiency for the bivariate mixture of continuous and binary responses.

Locally optimal designs

r' \ r	0	1	10	100
0	1	0.96	0.60	0.23
1	0.97	1	0.79	0.34
10	0.77	0.88	1	0.65
100	0.36	0.42	0.65	1

Two-stage designs

r' \ r	0	1	10	100
0	1	0.97	0.80	0.69
1	0.98	1	0.97	0.92
10	0.94	0.97	1	0.99
100	0.88	0.92	0.98	1

10, and 100 respectively. In each scenario, $N_0 = 80$ patients are assigned to the five doses via the uniform design in the initial stage, and then the rest of $N_1 = 120$ patients are assigned according to the estimated optimal designs. MLEs are calculated using the non-linear optimization subroutine "nlptr" in SAS. For this specific dose-response model, the probability of toxicity is very low within the dose range 0.2 to 0.8. There is a small probability that one may get no events for all five dose levels in the initial stage, which would lead to unreasonable MLEs. To fix this problem, we add a regularization term to the likelihood function for the initial stage. This regularization term has minimal impact on the parameter estimates if enough events are observed (Tarantola, 2004, Wahba, 1990).

Figure 4.3 shows the results of simulations. The left panel presents the allocations of the optimal doses in the second stage for the different designs. The circles in the

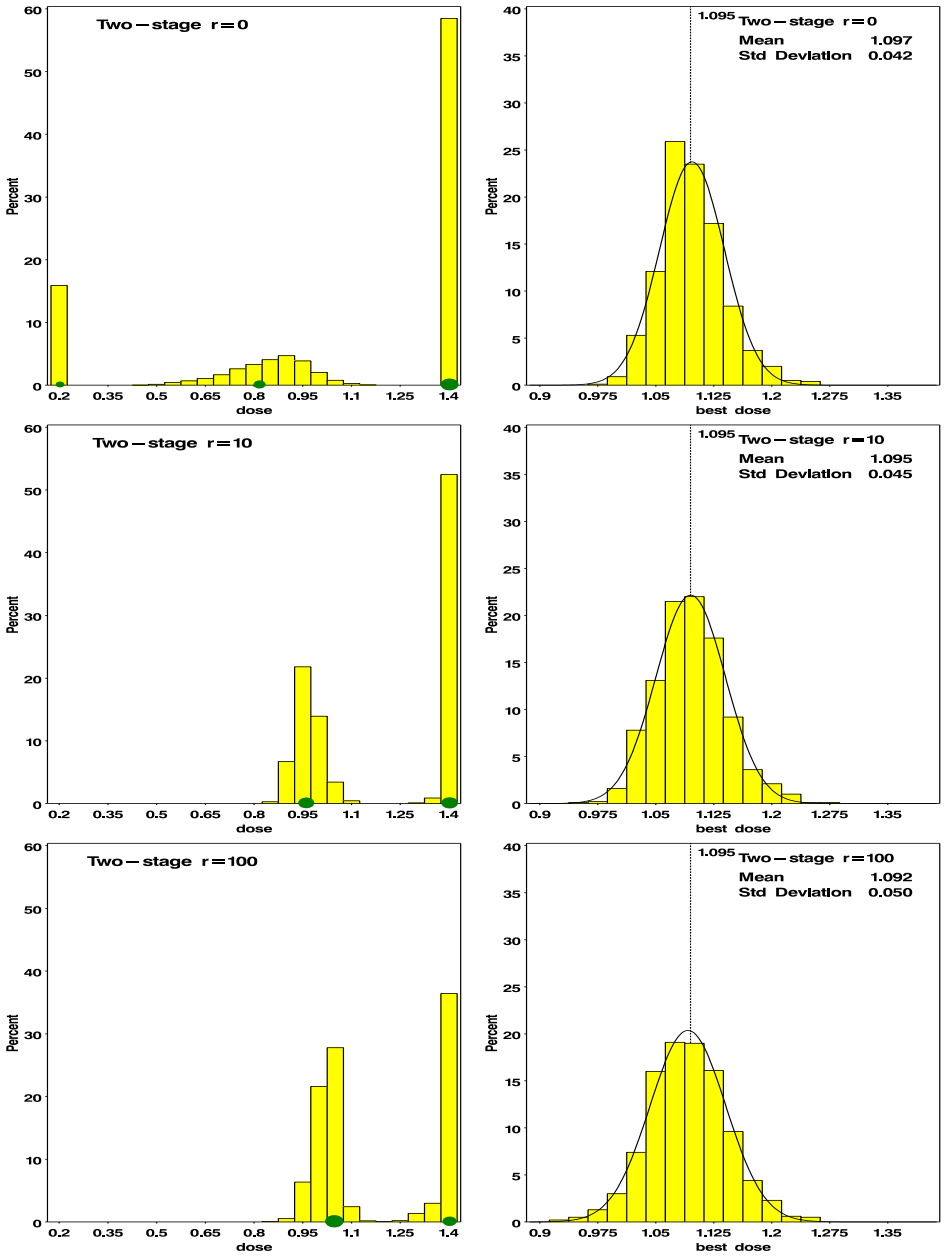

Figure 4.3: 1000 simulated two-stage designs with $r = 0$, 10 and 100. $N_0 = 80$; $N1 = 120$. **Left panel**: Locations of design points in the second stage; **Right panel**: Distributions of the predicted "best dose" $\hat{X}^*$.

histograms represent the theoretical second stage optimal doses in Figure 4.2, and the sizes of the circles are proportional to their weights. These plots demonstrate that the locations of the design points are around the theoretical second stage optimal designs, and close to the corresponding locally optimal designs as well (see Figure 4.2).

The right panel of Figure 4.3 are the distributions for the predicted best dose $\hat{X}^*$ for different two-stage designs using the final estimated parameters for all 200 subjects. The reference lines indicate the best dose under true parameters ($X^* = 1.095$). The curves denote the fitted normal density with the sample mean of $\hat{X}^*$, and the variance coinciding with the asymptotic variance of $\hat{X}^*$ for the corresponding two-stage optimal designs with true parameters (see Table 4.3). The means of the estimated best doses for all three designs are close to the true best dose. Although the two-stage design with $r = 0$ has smaller variances for the estimated best dose, the allocation for the doses is spread far away from the best dose and thus cause the large penalties.

4.4.3 Fully Adaptive Designs

Similar to the two-stage designs, 80 patients are assigned to the uniform design at the initial stage and 200 patients are assumed in total. We update the model parameters after each subject and assign the next subject following a given strategy. For D-adaptive designs, given currently available observations, we allocate the next observation to the dose which maximizes the sensitivity function in (4.3.13). D-

adaptive designs with the penalty parameters $r = 0$, 10, and 100 are studied. For the purpose of comparison, we also construct the best intention adaptive design, in which the next patient is assigned to the predicted best dose (the dose that maximizes the utility function, see (4.2.11) and (4.2.12)) based on the current accumulated data, i.e.,

$$x_N = \arg\max_{x \in \mathfrak{X}} \zeta(x, \hat{\boldsymbol{\theta}}_N). \tag{4.4.2}$$

Similar designs are rather popular in clinical trials such as the intuitive "best intention" design proposed by Lai and Robbins (1978), the adaptive design for maximization of the probability p_{10} by Li, Durham and Flournoy (1995), Continuous Reassessment Method (CRM) by O'Quigley, Pepe, and Fisher (1990), and Escalation with Overdose Control (EWOC) by Babb, Rogatko, and Zacks (1998). All these designs have some in common with so called "self-tuning" optimizer problem, see Pronzato (2000).

Figure 4.4 shows the allocations of the patients (the left panel) and the distributions of the predicted "best dose" (the right panel) for the three D-adaptive designs and the best intention adaptive design when the total number of patients reaches 200. One thousand simulations are performed for each design.

The allocations of the patients in the simulations are around the theoretical second stage optimal designs, and close to the corresponding locally optimal designs as well (see Figure 4.2). For the estimates of the best dose, the D-adaptive design with $r = 0$ has the smallest variability and the best intention design has the largest variability. In addition, the distributions of the predicted best dose in the adaptive designs have

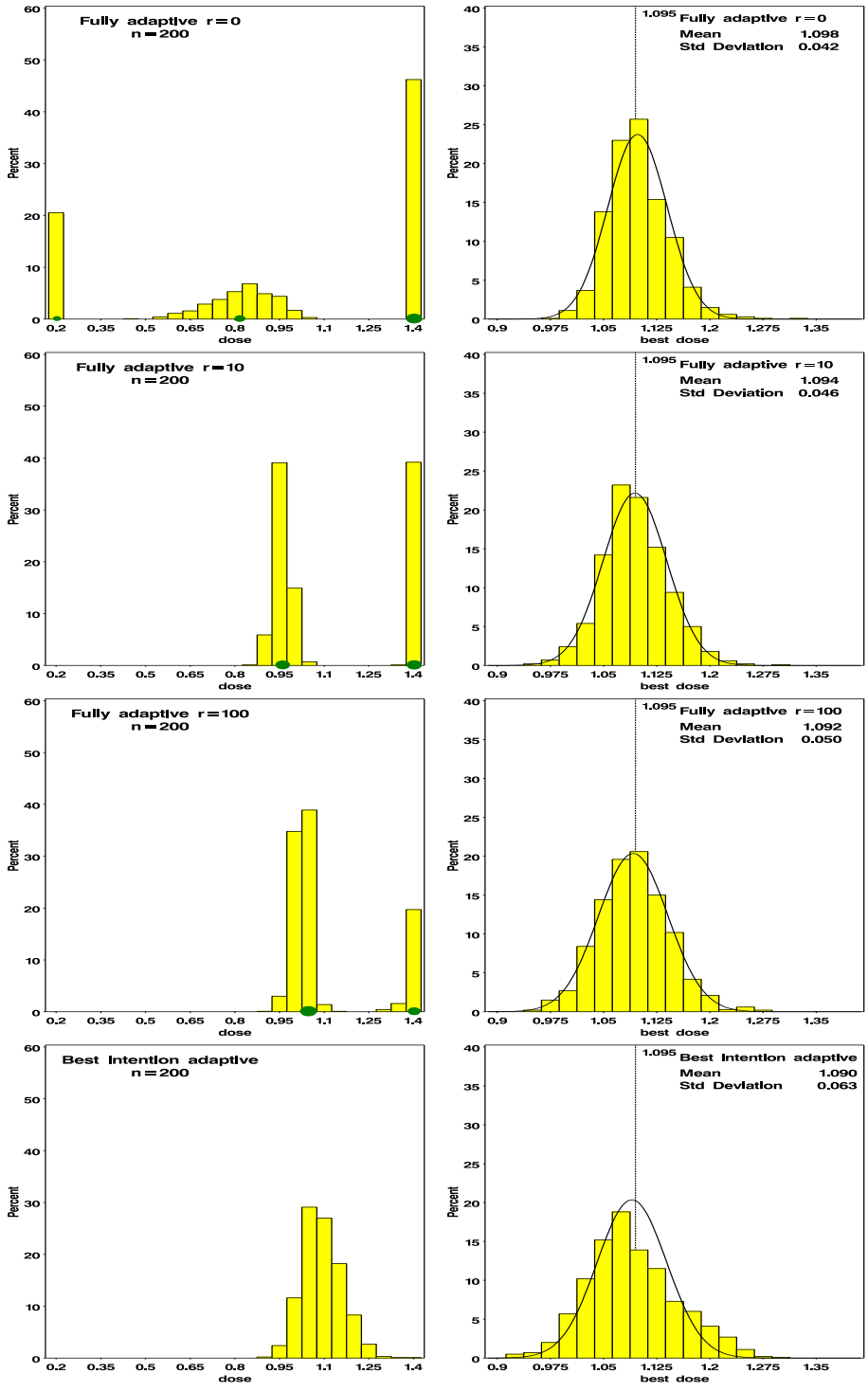

Figure 4.4: 1000 simulated fully adaptive designs with $r = 0$, 10 and 100, and best intention adaptive design; $N_0 = 80$; **Left panel**: Locations of design points at 200^{th} patient; **Right panel**: Distributions of the predicted "best dose" $\hat{X}^*$.

the means and standard deviations close to the corresponding two-stage designs (see Figure 4.3).

4.4.4 Unknown correlation ρ

In this section, we construct the penalized locally D-optimal design when ρ is assumed as an unknown parameter. To study the effect of different correlations between efficacy and toxicity on the optimal designs, we consider the following true values of ρ: $-0.9, -0.5, 0, 0.5$ and 0.9, while the other parameters remain the same. The results (Figure 4.5) show that the optimal designs are not very sensitive to the changes of the correlation ρ.

We expect that when ρ is assumed to be a unknown model parameter, the two-stage designs and D-adaptive designs have similar changes to the corresponding locally optimal designs.

4.4.5 Partition of sample sizes in two-stage designs

In practice, people may be interested in the choice of the ratio of the sample sizes of two stages in the two-stage design. This is a trade-off between the accuracy of estimators from the initial stage and the number of patients put on the more efficient design. For example, a very large sample size of the initial stage leads to more accurate estimators from the initial stage but leaves a small sample size for optimal design, and consequently the final estimators may be less accurate. In contrast, a very small sample size may result in inaccurate estimates from the initial stage, and consequently

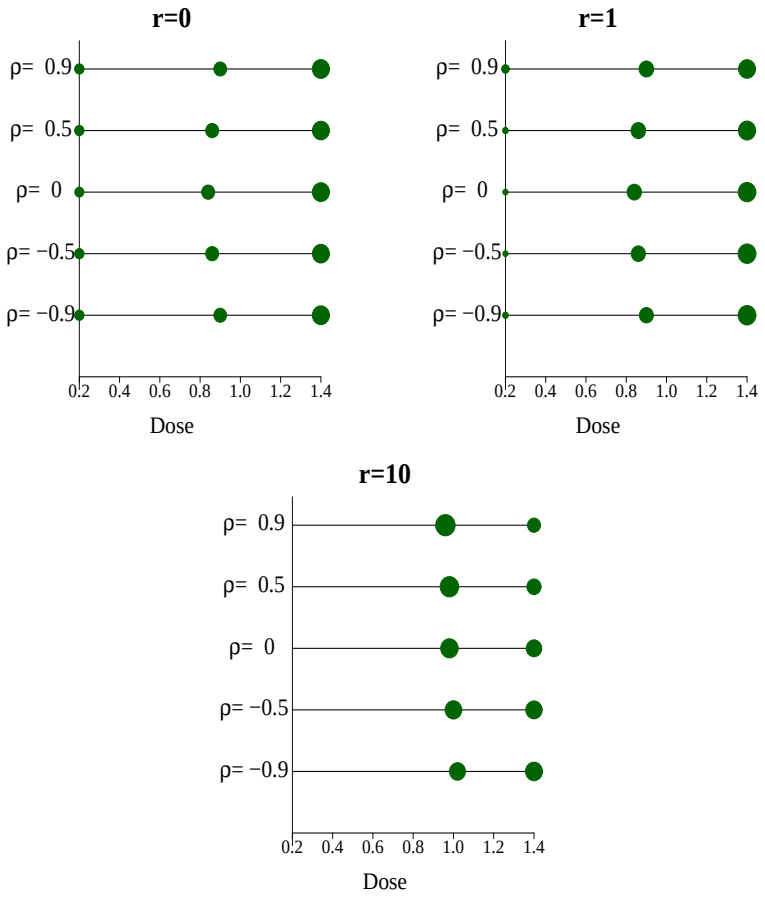

Figure 4.5: Allocation of the doses for locally optimal designs under different unknown correlation parameters of ρ and with different values of r in the penalty function. The size of each point represents the corresponding weight.

the design in the second stage may be far away from the local optimal design. To investigate this issue, given the total sample size 200, we construct simulations under different sample sizes of first stage (N_0) with values of 20, 40, 60, 80, 100 and 120 with $r = 10$ in the penalty function, and compare the final results.

Three measures are investigated :

(1) The information per penalty ($|(\pi\mathbf{M}(\xi_0, \boldsymbol{\theta}) + (1-\pi)\mathbf{M}(\xi, \boldsymbol{\theta}))/(\pi\Phi(\xi_0, \boldsymbol{\theta}) + (1-$

$\pi)\Phi(\xi, \boldsymbol{\theta}))|^{1/m})$ which indicates the efficiency of the design;

(2) The estimate of the general MSE of $\boldsymbol{\theta}$, which quantifies the overall accuracy of all unknown parameters and is defined as $|E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T|^{1/m} = N|\sum_{i=1}^{s}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T/s|^{1/m}$, where s is the number of simulations and N is the total sample size;

(3) The Root Mean Square Error (RMSE) of the predicted best dose which measures the performance of the design in terms of the estimated best dose.

For comparison purposes, the corresponding values for the locally two-stage D-optimal design (i.e., the true parameters are used in the second stage), the penalized locally D-optimal designs, and the five-dose uniform designs are also calculated as the reference lines.

The information per penalty for the locally two-stage designs (the top left plot in Figure 4.6), decreases as N_0 increases, while in the simulations, the values may drop when N_0 is small and the range of the quantiles may be large due to the less accurate MLE in the first stage. Both the estimate of the general MSE of $\boldsymbol{\theta}$ (the top right plot in Figure 4.6) and the RMSE of the predicted best dose (the bottom plot in Figure 4.6) decrease first and then increase as N_0 increases. All these results suggest that a moderate sample size (say between 80 and 100) in the first stage has relatively higher efficiency for the design and better performance in terms of the prediction of the overall parameters and the best dose.

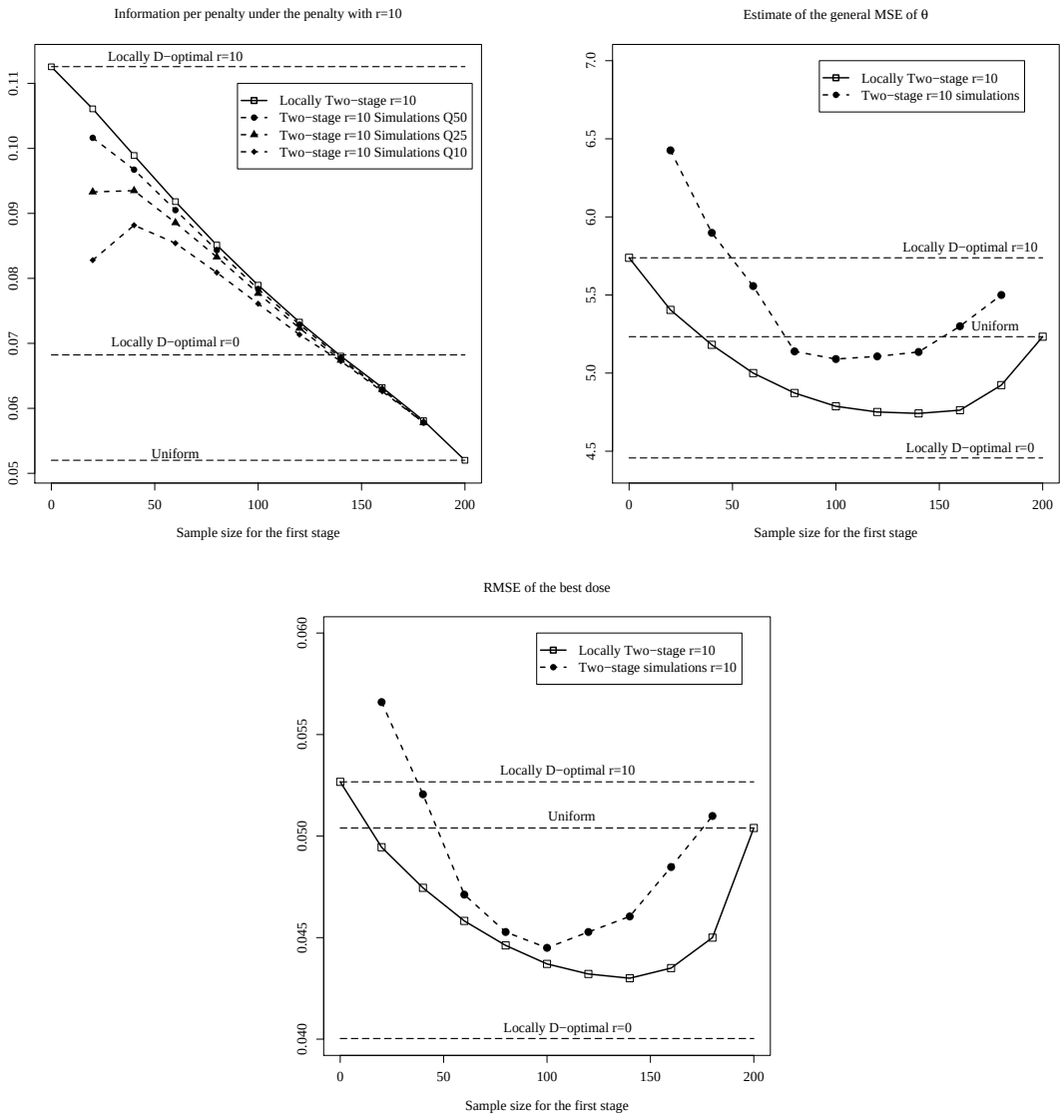

Figure 4.6: For different designs: locally D-optimal design built with $r = 0$, locally D-optimal design built with $r = 10$, five-point uniform deign, locally and simulated two-stage designs built with $r = 10$ and different partition of sample sizes, respectively, (1) **Top-left**: information per penalty, (2)**Top-right**: estimation of the general MSE of $\boldsymbol{\theta}$, and (3)**Bottom**: RMSE of the estimated best dose.

4.5 Conclusion

Under a bivariate model for correlated continuous efficacy and discrete toxicity responses, we developed dose-finding procedures based on optimal experimental design theory. A utility function is used to quantify the targeted treatment effects, and a penalty function is introduced to address ethical issues and cost constraints in a drug development setting.

Various designs including locally optimal, two-stage designs, and fully adaptive designs were constructed and compared. We used the locally optimal designs as benchmarks with which all other designs were compared. In examples, we varied the penalty function to illuminate its impact on the optimal designs and compared the relative efficiency among different designs. In practice, the selection of the penalty assigned to each dose should be extensively discussed with researchers.

In our simulations the two-stage designs work well to predict the best dose, and the assignment of the patients is very close to the locally optimal two-stage designs when a decent sample size is used in the initial stage (e.g. 40% of the total 200 observations). We provided a rather straightforward evaluation of the sample size of the first stage based on Monte Carlo simulations. It is computationally intensive but we are not aware of any analytical tool. Looking at the histograms provided for various estimators one can notice that two-stage designs perform equally well or better than fully adaptive designs. It might contradict to the intuition of many who think that fully adaptive designs use the available information more efficiently than two-

stage designs. However, recalling (Fedorov, 1972) that fully adaptive designs mimic the first order algorithm with "forward" steps only, one can understand why two-stage designs may be better. Indeed in two-stage designs given $\hat{\boldsymbol{\theta}}$, we build an exact optimal compliment to the first stage. Therefore, if unknown parameters are reasonably well estimated after the initial stage, two-stage designs can be superior than fully adaptive designs. Taking into account that the logistics for two-stage designs is much simpler than for fully adaptive designs, the two-stage designs are preferable in most cases.

This study is an example of how general probit model can be introduced and combined with optimal design theory in dose-finding experiments. The generalization of our work can be easily extend to multiple mixture responses, multilevel categorical responses, and multistage optimal designs.

Chapter 5

Appendices

5.1 The proofs related to Chapter 2

5.1.1 Proof of Theorem 1

The OLS estimator of $\boldsymbol{\beta}^T = (\beta_0, \beta_R, \beta_M, \beta_{RM})$ for the model in (2.2.2) is $\hat{\boldsymbol{\beta}} = (Z^T Z)^{-1} Z^T \mathrm{Y} = \boldsymbol{\theta} + (Z^T Z)^{-1} Z^T \epsilon$, where Z is the design matrix in regression analysis, and $\mathrm{Y} = (Y_1, \ldots, Y_n)^T$, the vector of outcome. Under Assumption (A1) and $E(\epsilon|R) = 0$, also assuming that $(Z^T Z)$ has the full rank, we have

$$
\begin{aligned}
&\hat{\boldsymbol{\beta}} - \boldsymbol{\theta} = (Z^T Z)^{-1} Z^T \epsilon \\
&= \begin{pmatrix} 1 & \sum_i R_i/n & \sum_i M_i/n & \sum_i R_i M_i/n \\ \sum_i R_i/n & \sum_i R_i^2/n & \sum_i R_i M_i/n & \sum_i R_i^2 M_i/n \\ \sum_i M_i/n & \sum_i R_i M_i/n & \sum_i M_i^2/n & \sum_i R_i M_i^2/n \\ \sum_i R_i M_i/n & \sum_i R_i^2 M_i/n & \sum_i R_i M_i^2/n & \sum_i R_i^2 M_i^2/n \end{pmatrix}^{-1} \begin{pmatrix} \sum_i \epsilon_i/n \\ \sum_i R_i \epsilon_i/n \\ \sum_i M_i \epsilon_i/n \\ \sum_i R_i M_i \epsilon_i/n \end{pmatrix} \\
&= \begin{pmatrix} 1 & E(R) & E(M) & E(RM) \\ E(R) & E(R^2) & E(RM) & E(R^2 M) \\ E(M) & E(RM) & E(M^2) & E(RM^2) \\ E(RM) & E(R^2 M) & E(RM^2) & E(R^2 M^2) \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ E(M\epsilon) \\ E(RM\epsilon) \end{pmatrix} + \boldsymbol{o_p(1)},
\end{aligned}
$$

Given $P(R=1) = P(R=0) = 1/2$, we have $E(R) = 1/2$, $E(RM) = E(R^2M) = 1/2E(M|R=1)$, $E(RM^2) = E(R^2M^2) = 1/2E(M^2|R=1)$, and $E(RM\epsilon) = 1/2E(M\epsilon|R=1)$. Substituting them into the above expression leads to

$$
\begin{aligned}
\text{Bias}(\hat{\boldsymbol{\beta}}) &= \text{plim}(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}) + \boldsymbol{o_p(1)} \\
&= \begin{pmatrix} \frac{E(M|R=0)E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \\ -\frac{E(M|R=1)\ E(M\epsilon|R=1)}{\text{Var}(M|R=1)} + \frac{E(M|R=0)\ E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \\ \frac{E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \\ \frac{E(M\epsilon|R=1)}{\text{Var}(M|R=1)} - \frac{E(M\epsilon|R=0)}{\text{Var}(M|R=0)} \end{pmatrix} + \boldsymbol{o_p(1)}.
\end{aligned}
$$

5.1.2 Proof of Lemma 1

Under $f(r,m,\epsilon) = f(1-r,-m,-\epsilon)$ and $P(R=1) = P(R=0)$,

$$\begin{aligned}
(1)\ E(M\epsilon|R=1) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{m\epsilon f(r=1,m,\epsilon)}{P(R=1)}dmd\epsilon \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{m\epsilon f(r=0,-m,-\epsilon)}{P(R=1)}dmd\epsilon \\
&= \int_{+\infty}^{-\infty}\int_{+\infty}^{-\infty}\frac{(-m)(-\epsilon) f(r=0,m,\epsilon)}{P(R=1)}d(-m)d(-\epsilon) \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{m\epsilon f(r=0,m,\epsilon)}{P(R=1)}dmd\epsilon \ = E(M\epsilon|R=0).
\end{aligned}$$

$$\begin{aligned}
(2)E(M|R=1) &= \int_{-\infty}^{+\infty}\frac{mf(r=1,m)}{P(R=1)}dm = \int_{-\infty}^{+\infty}\frac{mf(r=0,-m)}{P(R=1)}dm \\
&= \int_{+\infty}^{-\infty}\frac{(-m)f(r=0,m)d(-m)}{P(R=1)} = \int_{-\infty}^{+\infty}\frac{(-m)f(r=0,m)}{P(R=1)}dm \\
&= -E(M|R=0).
\end{aligned}$$

$$\begin{aligned}
&(3)\quad E(M^2|R=1) = \int_{-\infty}^{+\infty}\frac{m^2 f(r=1,m)}{P(R=1)}dm = \int_{-\infty}^{+\infty}\frac{m^2 f(r=0,-m)}{P(R=1)}dm \\
&= \int_{+\infty}^{-\infty}\frac{m^2 f(r=0,m)}{P(R=1)}d(-m) = \int_{-\infty}^{+\infty}\frac{m^2 f(r=0,m)}{P(R=1)}dm = E(M^2|R=0).
\end{aligned}$$

From (2) and (3), we have $\mathrm{Var}(M|R=1) = E(M^2|R=1) - E^2(M|R=1) = E(M^2|R=0) - E^2(M|R=0) = \mathrm{Var}(M|R=0)$.

5.1.3 Proof of Theorem 2

From Lemma 1, we know that if $f(r,m,\epsilon) = f(1-r,-m,-\epsilon)$, and under $P(R=1) = P(R=0) = 1/2$, then $E(M\epsilon|R=1) = E(M\epsilon|R=0)$ and $\mathrm{Var}(M|R=1) = \mathrm{Var}(M|R=0)$. We have $\mathrm{Cov}(R,M) = E(RM) - E(R)E(M) = E(RM) =$

$\frac{1}{4}[E(M|R=1) - E(M|R=0)]$. Substituting these into the bias expression of Theorem 1 leads to $\text{Bias}(\hat{\beta}_R) = -4\frac{E(RM)E(M\epsilon)}{\text{Var}(M)} + o_p(1)$, $\text{Bias}(\hat{\beta}_M) = \frac{E(M\epsilon)}{\text{Var}(M)} + o_p(1)$, and $\text{Bias}(\hat{\beta}_{RM}) = o_p(1)$.

And also $\text{Bias}(\hat{\beta}_R)/\text{Bias}(\hat{\beta}_M) = -[E(M|R=1) - E(M|R=0)] = -4E(R,M) + o_p(1)$.

5.1.4 Proof of Theorem 3

The OLS estimator for the model in (2.2.6) is

$$\hat{\boldsymbol{\beta}}_f = (\mathrm{X}_f^T\mathrm{X}_f)^{-1}\mathrm{X}_f^T\mathrm{Y} = \boldsymbol{\theta}_f + (\mathrm{X}_f^T\mathrm{X}_f)^{-1}\mathrm{X}_f^T\boldsymbol{\epsilon}$$

, where $\mathrm{X}_f = (\mathrm{Z}, \mathrm{X})$ and $\boldsymbol{\theta}_f = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T)^T$.

Therefore,

$$\hat{\boldsymbol{\beta}}_f - \boldsymbol{\theta}_f = (\mathrm{X}_f^T\mathrm{X}_f)^{-1}\mathrm{X}_f^T\boldsymbol{\epsilon} = \begin{pmatrix} \mathrm{Z}^T\mathrm{Z} & \mathrm{Z}^T\mathrm{X} \\ \mathrm{X}^T\mathrm{Z} & \mathrm{X}^T\mathrm{X} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{Z}^T\boldsymbol{\epsilon} \\ \mathrm{X}^T\boldsymbol{\epsilon} \end{pmatrix}$$

$$= \begin{pmatrix} (\mathrm{Z}^T\mathrm{Z})^{-1}(I + (\mathrm{Z}^T\mathrm{X})\mathrm{F}^{-1}(\mathrm{Z}^T\mathrm{X})^T(\mathrm{Z}^T\mathrm{Z})^{-1}) & -(\mathrm{Z}^T\mathrm{Z})^{-1}(\mathrm{Z}^T\mathrm{X})\mathrm{F}^{-1} \\ -\mathrm{F}^{-1}(\mathrm{Z}^T\mathrm{X})^T(\mathrm{Z}^T\mathrm{Z})^{-1} & F^{-1} \end{pmatrix} \begin{pmatrix} \mathrm{Z}^T\boldsymbol{\epsilon} \\ \mathrm{X}^T\boldsymbol{\epsilon} \end{pmatrix}$$

$$= \begin{pmatrix} (\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\boldsymbol{\epsilon} + \mathrm{L}\mathrm{F}^{-1}(\mathrm{Z}\mathrm{L} - \mathrm{X})^T\boldsymbol{\epsilon} \\ \mathrm{F}^{-1}\mathrm{X}^T\boldsymbol{\epsilon} - \mathrm{F}^{-1}\mathrm{X}^T\mathrm{Z}\text{Ł} \end{pmatrix},$$

where $\mathrm{L} = (\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathrm{X}$ and $\mathrm{F} = \mathrm{X}^T\mathrm{X} - \mathrm{X}^T\mathrm{Z}(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T X$.

Consequently,

$$
\begin{aligned}
& Bias(\hat{\boldsymbol{\beta}}) \\
= \;& \text{plim}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta} + \boldsymbol{o_p(1)} \\
= \;& \{\text{plim}(\mathrm{Z}^T\mathrm{Z})\}^{-1}\text{plim}(\mathrm{Z}^T\boldsymbol{\epsilon}) + \text{plim}\{\mathrm{LF}^{-1}(\mathrm{ZL}-\mathrm{X})^T\boldsymbol{\epsilon}\} + \boldsymbol{o_p(1)} \\
= \;& \{\text{plim}(\mathrm{Z}^T\mathrm{Z})\}^{-1}\text{plim}(\mathrm{Z}^T\boldsymbol{\epsilon}) + \text{plim}(\mathrm{L})\{\text{plim}(\mathrm{F})\}^{-1}\text{plim}((\mathrm{ZL}-\mathrm{X})^T\boldsymbol{\epsilon}) + \boldsymbol{o_p(1)}.
\end{aligned}
$$

5.2 The derivation of optimal weight of G-estimation

The optimal choice of $g^{opt}(Y^{00}(\Psi), X)$ for dichotomous R is (Robins(1992a)):

$$
\begin{aligned}
g^{opt}(Y^{00}, X) &= E\{S_\Psi(\Psi, \theta; X, R, M, Y)|X, R=1, Y^{00}\} \\
&- E\{S_\Psi(\Psi, \theta; X, R, M, Y)|X, R=0, Y^{00}\} \qquad (5.2.1)
\end{aligned}
$$

where $S_\Psi(\Psi, \theta; X, R, M, Y)$ is the score function w.r.t Ψ, and here θ represents the nuisance parameters.

To study $g^{opt}(Y^{00}, X)$, we start with the likelihood function based on the observable data:

$$
\begin{aligned}
L(\Psi, \theta; X, R, M, Y) &= L(\Psi, \theta; X, R, M, Y^{00}(\Psi))\frac{\partial Y^{00}(\Psi)}{\partial Y} \\
&= L(\Psi, \theta; X, R, M, Y^{00}(\Psi)) \\
&\quad \left(\because \frac{\partial Y^{00}(\Psi)}{\partial Y} = 1 \;\text{ for }\; \text{model(3.2.3)}\right) \\
&= f(X;\theta)f(R|X;\theta)f(Y^{00}(\Psi)|X, R;\theta)f(M|R, X, Y^{00}(\Psi);\theta) \\
&= f(X;\theta)f(R|X;\theta)f(Y^{00}(\Psi)|X;\theta)f(M|R, X, Y^{00}(\Psi);\theta) \\
&\quad \left(\because R \perp Y^{00}(\Psi)|X\right)
\end{aligned}
$$

The log likelihood function is :

$$\begin{aligned} l(\Psi,\theta;X,R,M,Y) &= \log f(X;\theta)+\log f(R|X;\theta) \\ &+ \log f(Y^{00}(\Psi)|X;\theta)+\log f(M|R,X,Y^{00}(\Psi);\theta) \quad (5.2.2) \end{aligned}$$

Thus the score function is :

$$\begin{aligned} S_\Psi(\Psi,\theta;X,R,M,Y) &= \frac{\partial l(\Psi,\theta;X,R,M,Y)}{\partial\Psi} = \frac{\partial l(\Psi,\theta;X,R,M,Y)}{\partial Y^{00}(\Psi)}\frac{\partial Y^{00}(\Psi)}{\partial\Psi} \\ &= -\frac{\partial \log f(Y^{00}(\Psi)|X;\theta)}{\partial Y^{00}(\Psi)}\begin{pmatrix} R \\ M \\ RM \end{pmatrix} - \frac{\partial \log f(M|X,R,Y^{00};\theta)}{\partial Y^{00}(\Psi)}\begin{pmatrix} R \\ M \\ RM \end{pmatrix} \\ &= S_{\{1\}\Psi}(\Psi,\theta;X,R,M,Y)+S_{\{2\}\Psi}(\Psi,\theta;X,R,M,Y), \quad (5.2.3) \end{aligned}$$

where $S_{\{1\}\Psi}(\Psi,\theta;X,R,M,Y) = -\frac{\partial \log f(Y^{00}(\Psi)|X;\theta)}{\partial Y^{00}(\Psi)}\begin{pmatrix} R \\ M \\ RM \end{pmatrix}$

and $S_{\{2\}\Psi}(\Psi,\theta;X,R,M,Y) = -\frac{\partial \log f(M|X,R,Y^{00};\theta)}{\partial Y^{00}(\Psi)}\begin{pmatrix} R \\ M \\ RM \end{pmatrix}$.

Expression (5.2.3) consists of two parts, which depend on the distributions of $f(Y^{00}|X)$ and $f(M|X,R,Y^{00})$. Accordingly, denoting

$$\begin{aligned} g^{opt}_{\{i\}}(Y^{00},X) &= E(S_{\{i\}\Psi}(\Psi,\theta;X,R,M,Y)|X,R=1,Y^{00}) \\ &- E(S_{\{i\}\Psi}(\Psi,\theta;X,R,M,Y)|X,R=0,Y^{00}), \end{aligned}$$

where $i=1$, or 2, we can write g^{opt} as

$$g^{opt}(Y^{00},X) = g^{opt}_{\{1\}}(Y^{00},X)+g^{opt}_{\{2\}}(Y^{00},X).$$

Under the sequential ignorability assumption, i.e. $M \perp Y^{00}|R, X$, the second term, $g^{opt}_{\{2\}}(Y^{00}, X)$ will be ignored (Joffe and Brensinger, 2003). Here we relax this assumption so both terms should be kept in g^{opt}. In what follows, for simplicity we assume that the potential outcome Y^{00} and post-randomization factor M follow the normal distributions.

If we assume $Y^{00} \sim N(\mu(X), \sigma^2)$, then

$$\begin{aligned} f(Y^{00}|X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{(Y^{00}-\mu(X))^2}{2\sigma^2}\Big\}, \\ S_{\{1\}\Psi}(\Psi, \theta; X, R, M, Y) &= \left\{-\frac{Y^{00}-\mu(X)}{\sigma^2}\right\} \begin{pmatrix} -R \\ -M \\ -RM \end{pmatrix}. \end{aligned}$$

Thus

$$g^{opt}_{\{1\}}(Y^{00}, X) = \frac{Y^{00}-\mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ E\{M|X, R=1, Y^{00}\} - E\{M|X, R=0, Y^{00}\} \\ E\{M|X, R=1, Y^{00}\} \end{pmatrix} \tag{5.2.4}$$

Expression (5.2.4) is a general form for $g^{opt}_{\{1\}}(Y^{00}, X)$, and can be used for any type of M (continuous, binary, categorical). In other words, under sequential ignorability, (5.2.4) is the optimal weight for G-estimation. The explicit form of conditional expectation of M depends on the model and distribution of M that we assume (the working assumption in G-estimation). For $g^{opt}_{\{1\}}(Y^{00}, X)$, it also depends on the distribution assumption of M. Therefore, in the following, we will discuss $g^{opt}_{(Y^{00},X)}$ when

M is continuous or binary respectively.

5.2.1 Continuous post-randomization factor M

Assume $M \sim N(\mu_m(X, R, Y^{00}), \sigma_m^2)$, where the mean $\mu_m(X, R, Y^{00})$ is a linear combination of X, R, and Y^{00}. For example. all two-way and three way interactions are included, then

$$\begin{aligned} \mu_m(X, R, Y^{00}) &= X\theta_x + R\theta_r + XR\theta_{xr} + Y^{00}\theta_{Y^{00}} + \\ & \quad XY^{00}\theta_{xY^{00}} + RY^{00}\theta_{rY^{00}} + XRY^{00}\theta_{xrY^{00}}, \end{aligned} \tag{5.2.5}$$

or equivalently,

$$\begin{aligned} M &= X\theta_x + R\theta_r + XR\theta_{xr} + Y^{00}\theta_{Y^{00}} + XY^{00}\theta_{xY^{00}} \\ & \quad + RY^{00}\theta_{rY^{00}} + XRY^{00}\theta_{xrY^{00}} + \epsilon_m, \end{aligned} \tag{5.2.6}$$

with $E(\epsilon_m|X, R, Y^{00}) = 0$ and $E(\epsilon_m^2|X, R, Y^{00}) = \sigma_m^2$.

Then under (5.2.5) and (5.2.6),

$$E\{M|X, R = 1, Y^{00}\} = \theta_r + X(\theta_x + \theta_{xr}) + Y^{00}(\theta_{Y^{00}} + \theta_{rY^{00}}) + XY^{00}(\theta_{xY^{00}} + \theta_{rxY^{00}}) \tag{5.2.7a}$$

$$E\{M|X, R = 0, Y^{00}\} = X\theta_x + Y^{00}\theta_{Y^{00}} + XY^{00}\theta_{xY^{00}} \tag{5.2.7b}$$

$$E\{M|X, R = 1, Y^{00}\} - E\{M|X, R = 0, Y^{00}\} = \theta_r + X\theta_{xr} + Y^{00}\theta_{rY^{00}} + XY^{00}\theta_{rxY^{00}} \tag{5.2.7c}$$

Substituting (5.2.7) into (5.2.4), we obtain

$$g^{opt}_{\{1\}}(Y^{00}, X) =$$
$$\frac{Y^{00} - \mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ \theta_r + X\theta_{xr} + Y^{00}\theta_{rY^{00}} + XY^{00}\theta_{rxY^{00}} \\ \theta_r + X(\theta_x + \theta_{xr}) + Y^{00}(\theta_{Y^{00}} + \theta_{rY^{00}}) + XY^{00}(\theta_{xY^{00}} + \theta_{rxY^{00}}) \end{pmatrix} \tag{5.2.8}$$

Again under (5.2.5) and (5.2.6), we have

$$\begin{aligned} & S_{\{2\}\Psi}(\Psi, \theta; X, R, M, Y) \\ = & \left\{ \frac{M - \mu_m}{\sigma_m^2} \right\} (\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}}) \begin{pmatrix} -R \\ -M \\ -RM \end{pmatrix} \\ = & -\frac{\epsilon_m}{\sigma_m^2} (\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}}) \begin{pmatrix} R \\ M \\ RM \end{pmatrix}. \end{aligned}$$

Therefore,

$$
\begin{aligned}
& g^{opt}_{\{2\}}(Y^{00}, X) \\
= \; & -\Big\{ E\big\{ \frac{\epsilon_m}{\sigma_m^2}(\theta_{Y^{00}} + X\theta_{xY^{00}} + \theta_{rY^{00}} + X\theta_{xrY^{00}}) \begin{pmatrix} 1 \\ M \\ M \end{pmatrix} | X, R = 1, Y^{00} \big\} \\
& \quad -E\big\{ \frac{\epsilon_m}{\sigma_m^2}(\theta_{Y^{00}} + X\theta_{xY^{00}}) \begin{pmatrix} 0 \\ M \\ 0 \end{pmatrix} | X, R = 0, Y^{00} \big\} \Big\} \\
= \; & -\frac{Y^{00} - \mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ \theta_r + X\theta_{xr} + Y^{00}\theta_{rY^{00}} + XY^{00}\theta_{rxY^{00}} \\ \theta_r + X(\theta_x + \theta_{xr}) + Y^{00}(\theta_{Y^{00}} + \theta_{rY^{00}}) + XY^{00}(\theta_{xY^{00}} + \theta_{rxY^{00}}) \end{pmatrix}.
\end{aligned} \tag{5.2.9}
$$

Thus

$$
\begin{aligned}
& g^{opt}(Y^{00}, X) = g^{opt}_{\{1\}}(Y^{00}, X) + g^{opt}_{\{2\}}(Y^{00}, X) \\
= \; & \frac{Y^{00} - \mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ \theta_r + X\theta_{xr} + Y^{00}\theta_{rY^{00}} + XY^{00}\theta_{rxY^{00}} \\ \theta_r + X(\theta_x + \theta_{xr}) + Y^{00}(\theta_{Y^{00}} + \theta_{rY^{00}}) + XY^{00}(\theta_{xY^{00}} + \theta_{rxY^{00}}) \end{pmatrix} \\
& - \begin{pmatrix} 0 \\ \theta_{rY^{00}} + X\theta_{xrY^{00}} \\ \theta_{Y^{00}} + \theta_{rY^{00}} + X(\theta_{xY^{00}} + \theta_{xrY^{00}}) \end{pmatrix}.
\end{aligned} \tag{5.2.10}
$$

5.2.2 Binary post-randomization factor M

Assume that M is binary with logit link,

$$\begin{aligned} & logit(P(M=1)) = \log\big\{\frac{P(M=1)}{1-P(M=1)}\big\} \\ = \;& X\theta_x + R\theta_r + XR\theta_{xr} + Y^{00}\theta_{Y^{00}} + XY^{00}\theta_{xY^{00}} + RY^{00}\theta_{rY^{00}} + XRY^{00}\theta_{xrY^{00}} \\ = \;& u_m, \end{aligned}$$

$$\begin{aligned} P(M=1) &= \frac{e^{u_m}}{1+e^{u_m}} = \frac{1}{1+e^{-u_m}}, \\ P(M=0) &= 1 - \frac{e^{u_m}}{1+e^{u_m}} = \frac{1}{1+e^{u_m}}. \end{aligned}$$

Denote

$$\begin{aligned} logit(P(M=1|X,R=1,Y^{00})) &= \theta_r + X(\theta_x + \theta_{xr}) + Y^{00}(\theta_{Y^{00}} + \theta_{rY^{00}}) \\ &+ XY^{00}(\theta_{xY^{00}} + \theta_{xrY^{00}}) = u_m^{r_1} \\ logit(P(M=1|X,R=0,Y^{00})) &= X\theta_x + Y^{00}\theta_{Y^{00}} + XY^{00}\theta_{xY^{00}} = u_m^{r_0}. \end{aligned}$$

Then

$$\begin{aligned} P(M=1|X,R=1,Y^{00}) &= \frac{e^{u_m^{r_1}}}{1+e^{u_m^{r_1}}} = \frac{1}{1+e^{-u_m^{r_1}}}, \\ P(M=0|X,R=1,Y^{00}) &= 1 - \frac{e^{u_m^{r_1}}}{1+e^{u_m^{r_1}}} = \frac{1}{1+e^{u_m^{r_1}}}. \end{aligned}$$

Thus we have

$$g^{opt}_{\{1\}}(Y^{00},X) = \frac{Y^{00}-\mu(X)}{\sigma^2} \begin{pmatrix} 1 \\ \frac{1}{1+e^{-\mu_m^{r_1}}} - \frac{1}{1+e^{-\mu_m^{r_0}}} \\ \frac{1}{1+e^{-\mu_m^{r_1}}} \end{pmatrix}. \tag{5.2.11}$$

The log likelihood of binary M and its partial derivative w.r.t Y^{00} are:

$$\begin{aligned}
\log f(M) &= \log\{P(M=1)^M P(M=0)^{(1-M)}\} \\
&= M\log P(M=1) + (1-M)\log P(M=0) \\
\frac{\partial \log f(M)}{\partial Y^{00}} &= M\frac{\partial \log P(M=1)}{\partial Y^{00}} + (1-M)\frac{\partial \log P(M=0)}{\partial Y^{00}} \\
&= M\frac{1}{1+e^{u_m}}(\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}}) \\
&\quad -(1-M)\frac{e^{u_m}}{1+e^{u_m}}(\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}}) \\
&= (\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}})(M - \frac{e^{u_m}}{1+e^{u_m}}) \\
&= (\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}})(M - P(M=1)).
\end{aligned} \tag{5.2.12}$$

Thus we have,

$$\begin{aligned}
& S_{\{2\}\Psi}(\Psi, \theta; X, R, M, Y) \\
= \; & -(\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}})(M - P(M=1)) \begin{pmatrix} R \\ M \\ RM \end{pmatrix},
\end{aligned} \tag{5.2.13}$$

$$
\begin{aligned}
& g^{opt}_{\{2\}}(Y^{00}, X) \\
= \; & -(\theta_{Y^{00}} + X\theta_{xY^{00}} + \theta_{rY^{00}} + X\theta_{xrY^{00}}) E\{(M - P(M=1)) \begin{pmatrix} 1 \\ M \\ M \end{pmatrix} | X, R=1, Y^{00}\} \\
& -(\theta_{Y^{00}} + X\theta_{xY^{00}}) E\{(M - P(M=1)) \begin{pmatrix} 0 \\ M \\ 0 \end{pmatrix} | X, R=0, Y^{00}\} \\
= \; & -(\theta_{Y^{00}} + X\theta_{xY^{00}} + R\theta_{rY^{00}} + XR\theta_{xrY^{00}})(M - P(M=1)) \begin{pmatrix} R \\ M \\ RM \end{pmatrix}. \qquad (5.2.14)
\end{aligned}
$$

Therefore, for binary M

$$
g^{opt}(Y^{00}, X) \;=\; g^{opt}_{\{1\}}(Y^{00}, X) + g^{opt}_{\{2\}}(Y^{00}, X) \qquad (5.2.15)
$$

5.3 The necessary and sufficient condition of optimality for the two-stage designs.

For two-stage designs, we assume that the design for the first stage ξ_0 is available and fixed, with sample size N_0. For simplicity of notation, we omit $\boldsymbol{\theta}$ in the information matrix, the penalty, and the sensitivity function, but readers should keep in mind that they all depend on $\boldsymbol{\theta}$.

When the total penalty for the two stages is limited by C, the optimal design can

be defined as

$$\xi^*(\boldsymbol{\theta}) = \arg\min_{\xi} \Psi[N_0\mathbf{M}(\xi_0) + N_1\mathbf{M}(\xi)] \text{ s.t. } N_0\Phi(\xi_0) + N_1\Phi(\xi) \leq \mathcal{C}, \quad (5.3.1)$$

where Ψ is convex and homogeneous.

Denoting $\pi = N_0/(N_0 + N_1)$, we can rewrite (5.3.1) as

$$\xi^*(\boldsymbol{\theta}) = \arg\min_{\xi} \Psi[N(\pi\mathbf{M}(\xi_0) + (1-\pi)\mathbf{M}(\xi))] \text{ s.t. } N(\pi\Phi(\xi_0) + (1-\pi)\Phi(\xi)) \leq \mathcal{C}. \quad (5.3.2)$$

Assume that π is fixed. The constraint in (5.3.2) implies $N(\xi) = \mathcal{C}/(\pi\Phi(\xi_0) + (1-\pi)\Phi(\xi))$, and thus

$$\begin{aligned} \xi^*(\boldsymbol{\theta}) &= \arg\min_{\xi} \Psi[(\pi\mathbf{M}(\xi_0) + (1-\pi)\mathbf{M}(\xi))\ \mathcal{C}/(\pi\Phi(\xi_0) + (1-\pi)\Phi(\xi))] \\ &= \arg\min_{\xi} \gamma(C)\Psi[(\pi\mathbf{M}(\xi_0) + (1-\pi)\mathbf{M}(\xi))/(\pi\Phi(\xi_0) + (1-\pi)\Phi(\xi))], \end{aligned} \quad (5.3.3)$$

where γ is a non-increasing function.

Due to the homogeneity assumption (Fedorov and Hackl, 1997, Ch.2.2), the optimization problem for the second stage in two-stage design is equivalent to

$$\xi^*(\boldsymbol{\theta}) = \arg\min_{\xi} \Psi\left[\frac{\pi\mathbf{M}(\xi_0) + (1-\pi)\mathbf{M}(\xi)}{\pi\Phi(\xi_0) + (1-\pi)\Phi(\xi)}\right], \quad (5.3.4)$$

or

$$\xi^* = \arg\min_{\xi} \Psi\left[\frac{\mathbf{A} + \mathbf{M}(\xi)}{a + \Phi(\xi)}\right], \quad (5.3.5)$$

where $\mathbf{A} = \frac{\pi}{1-\pi}\mathbf{M}(\xi_0)$ and $a = \frac{\pi}{1-\pi}\Phi(\xi_0)$. Both $\mathbf{A}$ and a are fixed given the fixed initial design ξ_0.

In general, the criterion in (5.3.5) is not a convex function of $\xi \in \Xi(\mathfrak{X})$. However, this criterion is quasiconvex (see Avriel, 2003 Ch.6.1). For this class of functions most results used in convex optimization stay valid.

A real-valued function f, defined on a convex set $X \subset R^n$, is said to be quasiconvex if

$$f(q_1 x^1 + q_2 x^2) \leq \max[f(x^1), f(x^2)]. \tag{5.3.6}$$

Consider the design $\bar{\xi} = (1-\alpha)\xi^* + \alpha\xi$, where ξ^* is the optimal design and ξ is some arbitrary design. For D-optimality, the directional derivative of Ψ in the direction of $\bar{\xi}$ is

$$\frac{\partial}{\partial\alpha} \log \left| \frac{\mathbf{A} + \mathbf{M}(\xi)}{a + \Phi(\xi)} \right|_{\alpha=0}^{-1} = m\frac{\Phi(\xi) - \Phi(\xi^*)}{a + \Phi(\xi^*)} - \operatorname{tr}\left\{ (\mathbf{A} + \mathbf{M}(\xi^*))^{-1}(\mathbf{M}(\xi) - \mathbf{M}(\xi^*)) \right\}, \tag{5.3.7}$$

where m is the number of unknown parameters.

Ψ is quasi-convex and the necessary and sufficient condition of the optimality of ξ^* is nonnegativeness of the directional derivative (5.3.7), i.e.,

$$\operatorname{tr}(\mathbf{A} + \mathbf{M}(\xi^*))^{-1}\mathbf{M}(\xi) - \frac{m\Phi(\xi)}{a + \Phi(\xi^*)} \leq \operatorname{tr}(\mathbf{A} + \mathbf{M}(\xi^*))^{-1}\mathbf{M}(\xi^*) - \frac{m\Phi(\xi^*)}{a + \Phi(\xi^*)}. \tag{5.3.8}$$

This statement is of little help because it should be verified for all possible ξ. However, instead of looking through all possible designs, one can actually verify inequality (5.3.8) only for designs atomized at a single point (Fedorov and Hackl, 1997). Indeed, with $M(\xi) = \int_{\mathfrak{X}} \mu(x)\xi(dx)$ and $\Phi(\xi) = \int_{\mathfrak{X}} \phi(x)\xi(dx)$, one may conclude that

for any design ξ, there exists $\tilde{x} \in \mathfrak{X}$ such that

$$\begin{aligned} & \operatorname{tr}(\mathbf{A}+\mathbf{M}(\xi^*))^{-1}\mathbf{M}(\xi) - \frac{m\Phi(\xi)}{a+\Phi(\xi^*)} \\ = & \int_{\mathfrak{X}} [\operatorname{tr}(\mathbf{A}+\mathbf{M}(\xi^*))^{-1}\mu(x) - \frac{m\phi(\xi)}{a+\Phi(\xi^*)}]\xi(dx) \\ = & \operatorname{tr}(\mathbf{A}+\mathbf{M}(\xi^*))^{-1}\mu(\tilde{x}) - \frac{m\phi(\tilde{x})}{a+\Phi(\xi^*)}. \end{aligned} \tag{5.3.9}$$

Therefore, one may conclude that the inequality

$$\begin{aligned} & \operatorname{tr}(\mathbf{A}+\mathbf{M}(\xi^*))^{-1}\mu(x) - \frac{m\phi(x)}{a+\Phi(\xi^*)} \\ & \leq \operatorname{tr}(\mathbf{A}+\mathbf{M}(\xi^*))^{-1}\mathbf{M}(\xi^*) - \frac{m\Phi(\xi^*)}{a+\Phi(\xi^*)},\ \forall x \in \mathfrak{X}, \end{aligned} \tag{5.3.10}$$

is the necessary and sufficient condition of optimality of ξ^*. The equality holds for all x which are support points of ξ^*.

Thus the sensitivity function for the second stage in the two-stage design is

$$\psi(x) = \operatorname{tr}(\mathbf{A}+\mathbf{M}(\xi))^{-1}\mu(x) - \frac{m\phi(x)}{a+\Phi(\xi)}. \tag{5.3.11}$$

Bibliography

Aiken, L.S., and West, S.G. (1991), *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identificatino of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**, 444-455.

Arminger G and Kusters U (1988). Latent trait models indicators of mixed measurement level. *Latent Trait and Latent Class Models*. pp 51-73. New York: Plenum.

Atkinson A and Donev A (1992). *Optimum Experimental Designs*. Oxford University Press, Oxford.

Atkinson A, Donev A and Tobias R (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford.

Avriel M (2003). *Nonlinear Programming: Analysis and Methods*. Dover Publications.

Babb J, Rogatko A, and Zacks S (1998). Cancel phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17, 1103-1120.

Baron, R.M., and Kenny, D.A. (1986). The Moderator-Mediator variable distinction in social psychological reserach: conceptual, strategic, and statistical consideration. *Journal of Personality and Social Psychology*, **51**, 1173-1182.

Bekele N and Shen Y (2005). A bayesian approach to jointly modelling toxicity and biomarker expression in a Phase I/II dose-finding trial. *Biometrics*, 61, 344-354.

Box G and Hunter W (1965). Sequential design of experiments for nonlinear models. In Proceedings of the IBM Scientific Computing Symposium on Statistics, 113-137, October 21-23, 1963.

Brown, G.K., Ten Have, T., Henriques, G.R., Xie, S.X., Hollander, J.E., and Beck, A.T. (2005). Cognitive therapy for the prevention of suicide attempts: a randomized controlled trial. *Journal of the American Medical Association*, **294**, 563-570.

Catalano P and Ryan L (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87, 651-658.

Coffey T and Glennings C (2007). D-Optimal designs for mixed discrete and continuous outcomes analyzed using nonlinear models. *Journal of Agricultural, Biological, and Environmental Statistics*, 12, 78-95.

Cox D and Wermuth N (1992). Response models for mixed Binary and quantitative variables. *Biometrika*, 79, 441-461.

Dragalin V and Fedorov V (2006). Adaptive designs for dose-finding based on efficacy-toxicity response. *Journal of Statistical Planning and Inference*, 136, 1800-1823.

Dragalin V, Fedorov V, and Wu Y (2008). Two-stage desingn for dose-finding that accounts for both efficacy and safety. *Statistics in Medicine*, Vol 27, 5156-5176.

Dignam J, Karrison T, and Bryant J (2005). Chapter 8, Design and Analysis of Oncology Clinical Trials. *Oncology: An Evidence-Based Approach*. New York: Springer-Verlag.

Dunn, G and Bentall, R (2007). Modeling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, **26**, 4719-4745.

Faerber, J.A., Joffe, M. M., Zhang, R., Brown, G. K., Beck, A. T., Ten Have, T. R (2010). Submitted to *Psychological Methods*.

Fedorov V (1972). *Theory of Optimal Exeriments.* New York: Academic Press.

Fedorov V and Hackl P (1997). *Model-Oriented Design of Experiments.* Lecture Notes in Statistics. 125. New York: Springer-Verlag.

Fedorov V and Wu Y (2007a). Dose finding designs for continuous responses and binary utility. *Journal of Biopharmaceutical Statistics*, 17, 1085-1096.

Fedorov V and Wu Y (2007b). Generalized probit model in design of dose finding experiments. *mODa8 - Advances in Model-Oriented Design and Analysis*, 67-73, Physics-Verlag.

Fedorov V, Wu Y, and Zhang R (2010). Dose Finding Experiments: Responses of Mixed Type. *mODa9 - Advances in Model-Oriented Design and Analysis*, 65-72, Physics-Verlag.

Fedorov V, Mannino F, and Zhang R (2009). Consequences of dichotomization. *Pharmaceutical Statistics*, 8: 50-61.

Fitzmaurice G and Laird N (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* 90, 845- 852.

Frangakis, C.E. and Rubin, D.B. (2002). Principlal stratification in causal inference. *Biometrics*, **58**, 21-29.

Frangakis, C., Brookmeyer, R., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a Needle Exchange Program. *Journal of the American Statistical Association*, **97**, 284-292.

Follan, D.A. (2000). On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, **95**, 1101-1109.

Gezmu M and Flournoy N (2006). Group up-and-down designs for dose-finding. *Journal of Statistical Planning and Inference*, 136, 1749-1764.

Gilbert, P. B., Bosch, R. J. and Hudgens, M. G. (2003) Sensitivity analysis for the

assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, **59**, 531-541.

Haas, E., Hill, R., Lambert, M., and Morrell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology*, **58**, 1157-1172.

Heise M and Myers R (1996). Optimal designs for bivariate logistic regression. *Biometrics*, 52, 613-624.

Hill, J., Waldfogel, J, and Brooks-gunn, J. (2002). Differential effects of high-quality child care, *Journal of Policy Analysis and Management*, **21**, 601-627.

Hosman, C.A., Hansen, B.B., and Holland, P.W. (2009). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. To appear *the Annals of Applied Statistics*.

Ivanova A (2003). A new dose-finding design for bivariate outcomes. *Biometrics*, 59, 1003-1009.

Ivanova A (2004). Zoom-in designs for dose-finding in oncology. *UNC Technical Report 04-03*.

Ivanova A (2006). Chapter 4, Dose-Finding in Oncology - Nonparameteric methods, *Dose Finding in Drug Development*, New York: Springer-Verlag.

Jemiai, Y., Rotnitzky, A., Shepherd, B., Gilbert P.B. (2007). Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a

postrandomization event occurs. *Journal of the Royal Statistical Society Series B*, **69**, 879-901.

Joffe, M.M. and Brensinger C.(2003). Weighting in instrumental variables and G-estimation. *Statistics in Medicine*, **22**, 1285-1303.

Joffe, M.M, Small, D., and Hsu, C. (2007). *Statistical Science*, **22**, 74-97.

Karrison T, Maitland M, Stadler W and Ratain M (2007). Design of Phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *J Natl Cancer Inst*, 99, 1455-1461.

Kpamegan E and Flournoy N (2001). An optimizing up and down design. *Optimum Design 2000*, Atkinson A and Bogacka B eds., pp. 211-224, the Netherlands: Kluwer Academic Publishers.

Kiefer E (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society Series B*, 21, 273-319.

Lai T and Robbins H (1978). Adaptive design in regression and control. *Proc. Natl. Acad. Sci. USA*, Vol. 75, No. 2, 586-587, February 1978.

Lai T and Robbins H (1982). Iterated least squares in multiperiod control. *Advances in Applied Mathematics*, Vol. 3, 50-73.

Lai T (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinca*, Vol. 11, 303-351.

Lambert, M.J. (2005). Early response in psychotherapy: Further evidence for the importance of common factors rather than placebo effects. *Journal of Clinical Psychology*, **61**, 855-869.

Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., and Khan, S. R. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology*, **84**, 277-295.

Lesaffre E and Molenberghs M (1991). Multivariate probit analysis: a neglected procedure in medical statisticis. *Statisticis in Medicine*, Vol. 10, 1391-1403.

Lehmann E (1983). Theory of point estimation. New York: Springer-Verlag, 1983.

Li Z, Durham, S and Flournoy, N (1995). An adaptive design for maximization of a contingent binary response. *Adaptive Designs*, IMS Lecture-Notes - Monograph Series 25, Flournoy N and Rosenberger W, eds., pp. 179-196, Hayward: Institute of Mathematical Statistics.

Liang K and Zeger S (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

National Cancer Institute (1999). Common Toxicity Criteria Manual (CTC) v2.0. Available at *http* : *//ctep.cancer.gov/protocolDevelopment/ electronic_applications/ctc.htm*.

Normand, S.L. (2007). Evaluating the optimal time of angiography: landmark or off the mark? *Circulation*, **23**, 2656-2657.

Mogg, R., Joffe, M., Mehta, M., and Ten Have, T (2010). A Causal Selection Model to Compare Treatment Groups in a Subset Selected Post- Randomization with Application to an HIV Antiretroviral Immunotherapy Trial. Submitted for publication.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, **5**, 465-472.

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

Pearson K (1900). Mathematical contributions to the theory of evolution VII: On the Correlation of Characters not Quantitatively Measurable. *Philosophical Transactions of the Royal Society of London, Series A. Containing Papers of a Mathematical or Physical Character*. Vol 195, 1-47.

Pearson K (1909). On a new method for determining the correlation between a measured character A and a character B. *Biometrika*, Vol 7, 96-105.

Prentice R and Zhao L (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47, 825-839.

Pronzato L(2000). Adaptive optimisation and D-optimum experimental design. *Annals of Statistics*, Vol 28, 1743-1761.

Rao C (1973). Linear Statistical Inference and Its Applications, 2^{nd} Editon. New York: J. Wiley.

The rgp120 HIV Vaccine Study Group (2005) Placebo-controlled Phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Disease*, **191**, 654-665.

Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**, 321-334.

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, **23**, 2379-2412.

Robins, J.M. and Greenland, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, **89**, 737-749.

Robins, J.M. (1998). Marginal structural models. *1997 Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, 1-10. Reproduced courtesy of the American Statistical Association.

Robins, J.M. (1999). Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology: The Environment and Clinical Trials: Halloran M.E. and Berry D., Editors.* Springer-Verlag, 95-134.

Rochon, J. (1999). Issues in adjusting for covariates arising postrandomization in clinical trials. *Drug Information Journal*, **33**, 1219-1228.

Rosenberger W and Hughes-Oliver J (1999). Inference from a sequential design: Proof of a conjecture by Ford and Silvey. *Statistics and Probability Letters*, Vol. 44, 177-180.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Jouranl of Educational Psychology*, **66**, 688-701.

Sammel M, Ryan L and Legler J (1997). Latent variable models for mixed discrete and continous outcomes. *Journal of Royal Statistical Society B*, 59, 667-678, 1997.

Shepherd, B. E., Gilbert, P. B., Jemiai, Y. and Rotnitzky, A. (2006) Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*, **62**, 332-342.

Smith K (1918). On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they Give towards a Proper Choice of the Distribution of Observations. *Biometrika*, 12, 1-85.

Silvey S(1980). Optimal Design. London: Chapman and Hall.

Tarantola A (2004). Inversse Problem Theory. *Society for Industrial and Applied Mathematics.* ISBN 0-89871-572-5.

Tate R (1955). The Theory of Correlation Between Two Continuous Variables when One is Dichotomized. *Biometrika*, 42: 205-216.

Teixeira-Pinto A and Normand S (2009). Correlated bivariate continuous and binary outcomes: issues and applications. *Statistics in Medicine*, 28, 1753-1773.

Thall P and Cook J (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, 60, 84-693.

Thall P and Russel K (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, 54, 251-264.

VanderWeele, T.J. and Hernan, M.A. (2011). Compound treatments and transportability of causal inference.*Epidemiology, in press*

Vansteelandt, S. (2009). Estimating direct effects in cohort and case-control studies. *Epidemiology*, **20**, 851-860.

Wahba G (1990). Spline Models for Observational Data. *Society for Industrial and Applied Mathematics.*

Whitehead J, Zhou Y, Stevens J, Blakey G, Price J, Leadbetter J (2006). Bayesian decision proceduses for dose-escalation based on evidence of undesirable events and therapeutic benefit. *Statistics in Medicine*, 25, 37-53.

Wu C (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association*, 80, 974-984.

Zhou Y, Whitehead J, Bonvini E and Stevens J(2006). Bayesian decision procedures for binary and continuous bivariate dose-escalation studies. *Pharmaceutical Statistics*, 5, 125-133.

Zohar S and Chevret S (2007). Recent developments in adaptive designs for phase I/II dose-finding studies. *Journal of Biopharmaceutical Statistics*, 17, 1071-1083.

Zohar S and O'Quigley J (2006). Optimal designs for estimating the most successful dose. *Statistics in Medicine*, 25, 4311-4320.