Causal Modeling with Principal Stratification to Assess Effects of Treatment with Partial Compliance, Noncompliance, and Principal Surrogacy in Longitudinal and Time-to-Event Settings

by

Xin Gao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Michael R. Elliott, Chair
Professor Roderick J.A. Little
Professor Jeremy M. G. Taylor
Associate Professor Ben B. Hansen

To my husband, daughter and parents

ACKNOWLEDGEMENTS

This dissertation could have never been completed without the help of many individuals. I want express my deepest gratitude to my advisor, Dr. Michael R. Elliott, for supporting me to explore the research topics and advising my dissertation work throughout my Ph.D. study. His broad knowledge, integrity and scholarship will serve as an academic role model for me, and his continuous support and inspiring guidance have been crucial for the dissertation's completion. I am deeply grateful to Drs. Roderick J. A. Little, Jeremy M. G. Taylor and Ben B. Hansen for serving as my dissertation committee members. Their insightful and constructive suggestions greatly improve the quality of my dissertation.

I particularly would like to thank late Dr. Thomas R. Ten Have at the University of Pennsylvania. I feel fortunate enough to have instructive discussions on my dissertation chapters and many other topics on research and career with him. His selflessness and encouragements will have long-lasting impacts on my career. I am also thankful to all faculty members of the Department of Biostatistics, in particular Drs. Douglas E. Schaubel and Yun Li. I learned tremendously from them through their lectures in survival analysis and discussions in casual inference. In addition, Drs. Morton B. Brown and Wen Ye have been instrumental to support and guide me in collaborative research during my assistantship at Biometrics and Outcomes Research Core.

Last but not least, I would like to dedicate this dissertation to my beloved daughter

Sophia, my husband Chen and my parents. My daughter's birth this May, the constant support and love from my family are the ultimate sources of power for me to complete my study. Without them, I would not have been able to go through the entire journey.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Much scientific research in the social and health sciences aims to understand the causal relationship between an intervention or exposure and an outcome, and a variety of statistical methods have been developed to answer these questions. However, in many situations, such causal relationships is not readily obtained even in well-conducted randomized clinical trials. Often, post-randomization variables are adjusted for or conditioned on at analysis stage to control for the potential confounding effects or other reasons such as surrogate endpoint validation. The causal interpretation for the effect of intervention is easily destroyed when using conventional approaches to adjust for post-randomization variables [45]. Under such circumstances, statistical models taking account of post-randomization variables while retaining the causal interpretation of the effect of intervention are highly desirable and challenging to develop.

This dissertation is motivated in part by the following three examples: 1) when possibly outcome-dependent non-compliance occurs in a longitudinal study where compliance can vary over time, evaluation of treatment effects may be confounded even with adjustment of compliance behavior; 2) when self-selected maximum dosage tolerances that relate with adverse effects exist and are distributed in an unbalanced

fashion between randomization arms, efficacy evaluation based on the primary endpoint may be confounded by the presence of self-selection; 3) surrogacy evaluation and validation is traditionally considered through a series of models that evaluate primary endpoints and putative surrogate markers marginally, as well as evaluate primary endpoints adjusted for surrogate markers. In all three examples, conventional analyses conditional on post-randomization variables are either problematic and/or lose causal interpretations of the treatment effect. In this dissertation, we develop and apply state-of-art causal inference to the above three topics under a potential outcome framework with a principal stratification approach.

One fundamental concept in causal modeling is the potential outcome, which is first proposed by Neyman [49] and defined as the value an outcome would take after assignment to a different treatment arm than the one actual observed [35]. One goal of causal inference is estimation of the causal effect of treatment, which is defined as the comparison of the potential outcomes under different treatment arms within an individual, and the average causal effect of treatment, which averages such comparison over the entire population [17, 50]. Under the potential outcome framework, Frangakis and Rubin proposed principal stratification [17]. Principal strata are determined by the joint distribution of potential values of the post-randomization adjustment variable of interest under different treatment arms. Because the potential outcome itself is considered to exist before treatment assignment, principal strata is also considered to exist before treatment assignment, and can be conditioned on or adjusted for without destroying the causal interpretation of a statistical model.

The first two chapters of this dissertation focus on noncompliance and partial compliance behavior in randomized clinical trials, which is very common in clinical studies. Often patients themselves choose whether or not to follow the assigned in-

tervention protocol, and thus noncompliance of treatment assignment is unavoidable in practice. Traditional statistical methods such as comparing outcomes by randomized treatment assignment – "intent-to-treat" (ITT) analysis — provides inference about the causal effects of the assignment, but not the actual treatment. Another traditional method compares outcomes among those adhering to the protocol –"per protocol" (PP) analysis – provides only evidence of association, not a causal interpretation, since treatment taken is a post-randomization variable [45].

To account for noncompliance, two major approaches have been developed with causal inference. The first approach considers noncompliance as a special case of mediation analysis, and focuses on estimating direct effects of treatment assignment and indirect effects of treatment assignment through the actual treatment taken [44]. This approach assumes that the treatment assigned and treatment taken can be manipulated independently, and subjects could be forced into compliance or non-compliance. The other approach estimates the causal effect of treatment by estimating the effect of treatment within principal strata [17]. In a two-arm randomized clinical trial, the principal strata consist of compliers (subjects who comply with the assignment under both treatment arms), always-takers (subjects who always take treatment regardless of assignment), never-takers (subjects who always take control regardless of assignment), and defiers (subjects who take the treatment opposite of what they are assigned). They conduct causal inference focusing on the effect of treatment assignment among the compliers and defiers, whose treatment taken changes with the treatment assigned; in the settings that defiers are assumed not to exist, only the complier stratum provides causal inference for the effect of treatment taken. In contrast to Robins and Greenland [44] , Frangakis and Rubin do not assume the ability to manipulate compliance behavior independently from treatment assignment, in-

stead they assume that only a subset of the population would be compliant with treatment, and focus inference on this stratum [17].

In the recent years, estimation of treatment effect in randomized clinical studies with noncompliance has been extended to longitudinal studies [60, 32, 15]. An important feature of longitudinal studies is that the outcomes and compliance behaviors are measured and recorded repeatedly over time, and thus the longitudinal data could reveal more information about the reason and mechanism of noncompliance. In Chapter II, we propose a Markov compliance and outcome model under the potential outcome framework with principal stratification approach. A unique feature of the proposed model is that we use the nature of the longitudinal study to assess the impact of the causal effect of treatment on the future compliance, as well as the causal effect of treatment within each principal stratum. We illustrate the model with a randomized clinical study designed to assess the effect of cognitive behavioral therapy for depression among a sample of suicide attempters.

Our work in Chapter III is motivated by the presence of partial compliance in the Interstitial Cystitis Collaborative Research Network (ICCRN) study. This two-arm randomized trial is designed to evaluate the efficacy of Amitriptyline for the treatment of interstitial cystitis. Subjects on both arms are assigned to a dose escalation schedule and recommended taking up to 75 mg. Primarily due to adverse events, many subjects opted to stay at or return to $\leq$ 25 mg or 50 mg. In this study, the severity of adverse events were reported and recorded in the trial for all subjects. Since observed adverse event is also a post-randomization variable, like observed partial compliance, we treat both of them and the clinical outcome together as potential outcomes in the proposed model. We incorporate adverse effect information to model the principal strata membership determined by dosage taken,

using the fact that subjects on the control arm can also change their placebo dosage to estimate their dosing behavior had they been under treatment .

Chapter IV focuses on causal modeling for surrogacy assessment. Research in surrogacy assessment has aroused much interests in recent years because clinical trials often encounter difficulties such as long follow-up periods, rare disease outcomes, or expensive medical costs. Surrogate markers may be able to be measured earlier and/or at lower cost, and may be used in lieu of primary outcomes to evaluate the effect of treatment if the (causal) effect of treatment is highly correlated with the (causal) effect of treatment of the ultimate outcome of interest. As in the noncompliance setting, a common drawback of previous regression-based methods to evaluate surrogate markers is that the estimated treatment effects lack causal interpretation due to the fact that the surrogate marker is measured after the treatment assignment [45]. To overcome this, Frangakis and Rubin proposed a principal surrogacy evaluation method based on the concept of principal stratification [17]. They proposed that an intermediate variable is a valid principal surrogate if it satisfies causal necessity, i.e., the causal effect of treatment on the primary outcome exists only when the causal effect of treatment on the intermediate variable exists. They also proposed two types of causal effects, associative and dissociative effects, to evaluate principal surrogacy. The associative effect on the outcome is defined as the comparison between the potential primary outcomes under different treatment arms when the value of surrogate markers are different under different treatment arms, and the dissociative effect on the outcome is defined as the comparison between potential primary outcomes under different treatment arms when the value of surrogate markers are same under different treatment arms. Current literature on principal surrogacy assessment has considered either a normally distributed primary outcome

or a binary primary outcome. For a time-to-event outcome in medical studies, how to evaluate the principal surrogacy has not been addressed yet. In Chapter IV, we follow the principal stratification approach, and propose a causal model to evaluate the principal surrogacy for time-to-event primary outcome. To take account of the correlation between the potential primary outcomes under different treatment arms, we introduce a shared-frailty model in conjunction with proportionality assumption [22]. We illustrate the proposed causal model in a randomized clinical trial designed to evaluate the efficacy of intensive-course fluorouracil combined with low-dose leucovorin as adjuvant therapy in patients with high-risk primary colorectal cancer to assess the principal surrogacy of 3-year disease free survival for overall survival.

All of the methods discussed above utilize a Bayesian inferential approach, using Markov chain Monte Carlo algorithms to accommodate complex missing data structures. Although our approach is Bayesian, we conduct detailed simulation studies in addition to the specific applications considered to investigate the repeated sampling properties of the proposed methods. We do this in order to make sure that our methods are "well calibrated" to assess the impact of differing priors on weakly or non-identified parameters in the proposed models [34].

CHAPTER II

Joint Modeling Compliance and Outcome for Causal Analysis in Longitudinal Studies

2.1 Introduction

Because randomized treatment assignment removes both observed and unobserved confounding, randomized studies provide a means to estimate the causal effect of treatment. However, since subjects can choose whether or not to comply with their assigned treatment in many circumstances, noncompliance or partial compliance is very common in randomized clinical studies. Traditional analysis methods include intent-to-treat (ITT) analysis, as-treated (AT) analysis and per-protocol (PP) analysis. ITT analyses provide a causal estimate of the effect of randomization, which can differ from the causal effect of the treatment in the presence of non-compliance. AT analyses ignore the randomization assignment, and compare the outcomes by the actual treatment received. PP analyses compare the outcomes for subjects who comply with the assigned treatment. However, because the latter two analyses methods condition on the treatment taken, which is a post-randomization variable, selection bias may exist and affect the AT and PP estimates of the causal effect of treatment [45].

A large literature has been developed in recent years to estimate effects of treatment via causal modeling under the potential outcome framework, with the goal

to compare the potential outcomes that would have been observed under various assignments of treatments. In fact, the idea of describing causal effect in terms of potential outcomes dates back to Neyman [49], and is now becoming widely used in the fields of economics, social and behavioral sciences, epidemiology and statistics.

In this chapter, we use the principal stratum approach formulated by Frangakis and Rubin [17]. For noncompliance problems, the principal strata are determined by the joint distributions of the compliance behaviors under both treatment arms. In the context of a two-arm randomized trial, principal strata consist of compliers (subjects who take treatment if and only if assigned to it), always-takers (subjects who take treatment regardless of assignment), not-takers (subjects who take control regardless of assignment), and defiers (subjects who take treatment if and only if assigned to control). Because potential outcomes are considered to exist before treatment assignment, principal strata are also considered to exist before treatment assignment, and can be conditioned on or adjusted for in regression without destroying the causal interpretation of a statistic model. In particular, the effect of treatment within the complier stratum, termed the complier average causal effect (CACE), is often of interest to investigators, and can be interpreted as the causal effect of treatment among the subpopulation who comply with the treatment assignment no matter to which treatment group they are assigned [25].

Previous research about noncompliance behavior in the randomized clinical studies with principal stratification has focused on obtaining a valid estimate of the effect of treatment within principal strata [10, 17, 25], and has been extended to longitudinal studies in recent years. These longitudinal studies vary in several features. For example, subjects can be randomized once at baseline [33, 32, 60] or multiple times over time [15]. Treatment can be applied once [60] or repeatedly over time

[15, 33, 32]. In this chapter, we focus on the longitudinal study when subjects are randomized once at baseline, treatments are applied repeatedly over time, and subjects' compliance behavior may change over time.

An important feature of a longitudinal study with this design is that the outcomes and compliance behaviors are measured and recorded repeatedly over time, and thus the data may reveal more information about the reason and mechanism of noncompliance. Possible reasons of noncompliance include forgetting to take the treatment, side effect or small effect of treatment. With an effective treatment, one might want to improve the compliance in future studies by means such as education of compliance or reduction in side effects. Motivated by this, we propose a Markov compliance and outcome model under the potential outcome framework using a principal stratification approach. A unique feature of the proposed model is the ability to assess the impact of the causal effect of the treatment on the future compliance, as well as the causal effect of treatment within principal strata. We illustrate the proposed causal model with a randomized clinical study designed to assess the effect of cognitive behavioral therapy (CBT) on depression among a sample of suicide attempters.

The remainder of this manuscript is organized as follows. Section 2.2 introduces the motivating study – CBT study. Section 2.3 describes our proposed Markov compliance and outcome model and Bayesian estimation method. Section 2.4 applies our proposed causal model to the CBT study. Section 2.5 studies the repeated sampling properties of the proposed model with simulation studies, followed by Section 2.6 discussing the implications of our findings and future extensions.

2.2 The Cognitive Behavioral Therapy (CBT) Study

The CBT study was a randomized longitudinal study designed to test the efficacy of a specially designed suicide prevention therapy. This study was designed to determine whether a brief psychosocial intervention could reduce depression severity over an 18-month follow up interval in a sample of subjects who had attempted suicide. The psychosocial intervention was built on clinical investigations regarding the psychopathological characteristics of suicide behaviors, and the central feature of this psychotherapy was identification of proximal thoughts, images, and core beliefs that were activated prior to the suicide attempt. The study sample consisted of 120 individuals who were initially identified in the emergency department at the Hospital of the University of Pennsylvania following a suicide attempt. After the subjects were medically cleared or stabilized in the emergency department, they were transferred to the psychiatric emergency department. Then eligible individuals were randomized into two groups: the control group, where patients received usual care, and the treatment group, where patients received the specially designed cognitive therapy besides the usual care.

The clinical outcome was Beck Depression Inventory (BDI), which means severity of depression. Subsequent in-person assessments of BDI were conducted at 1, 3, 6, 12, and 18 months following the baseline interview. Among the 120 subjects, 60 subjects were randomly assigned to the cognitive therapy group, and 60 subjects to the usual care group. We restrict our analysis to the 58 subjects in the treatment group and the 56 subjects in the control group with at least one BDI follow-up measurement. We summarize subjects' baseline covariates in Table 2.1. Randomization achieved balance across the observed baseline covariates between two treatment

Table 2.1: Demographic characteristics for the CBT study

	Treatment Assignment		
Baseline covariates	Treatment	Control	p value
Age, Mean (S.D.)	35 (10.0)	35 (10.3)	0.80
Gender, Female (%)	35(60.3)	34 (60.7)	0.97
Beck Depression Index	33 (12.1)	31 (15.9)	0.53
Beck Hopeless Scale	11 (5.5)	12 (6.3)	0.79
Suicide Ideation	28 (5.7)	29 (4.4)	0.48
Number of Previous Suicide Attempts	4 (5.4)	6 (13.6)	0.47
Self Reported Health Status	2 (1.1)	2 (1.1)	0.60
Positive Problem Orientation	9 (4.3)	9 (5.0)	0.91
Rational Problem Style	9 (5.1)	8 (5.3)	0.73
Impulsive Careless Style	10 (4.8)	9 (5.5)	0.22

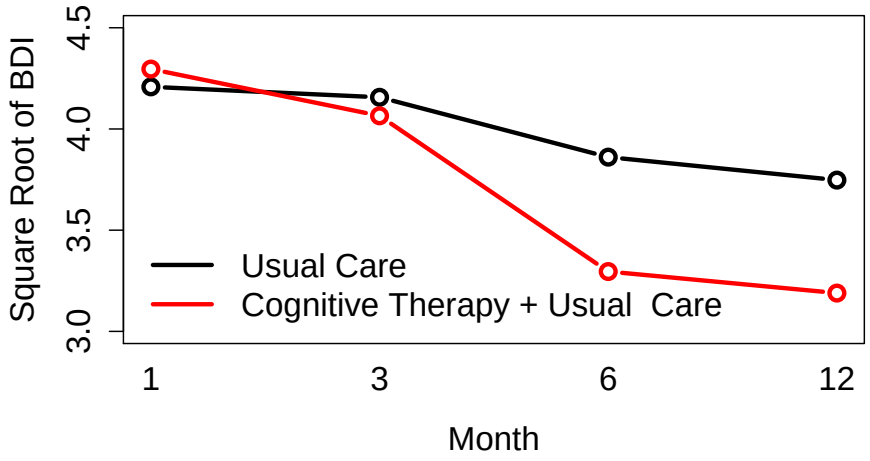


Figure 2.1: Mean of $\sqrt{BDI}$ for the CBT study

arms. Subjects' age ranged from 18 to 66 years, and 61% of them are female. We apply a square root transformation on BDI to improve the normality approximation for the clinical outcomes. Subjects randomized to the treatment group developed less severe depression on average than the subjects randomized to the control group from month 3 onward. We summarize the observed $\sqrt{BDI}$ of the study sample in Figure 2.1.

Subjects who were assigned to the treatment group and received CBT at least once

during a given follow up period are defined as having complied with the treatment. The noncompliance rates of subjects assigned to the treatment group are 7%, 9%, 36%, and 64% in follow up periods of one month, three months, six months and twelve months respectively. We do not include data for the 18-month follow-up in the analysis because noncompliance rate is 96%.

2.3 The Markov Compliance and Outcome Model

2.3.1 Notation

For subject $i = 1, ..., n$, we use Z_i to denote the treatment assignment (1 for treatment and 0 for control) and $\boldsymbol{X}_i$ to denote the baseline covariates. Compliance behaviors and outcomes are measured at fixed time points denoted as $1, ..., t, ..., T$. For subject i at follow-up period t, we use $Y_{i,t}(z_i)$ to denote the potential clinical outcome under treatment assignment z_i, and $D_{i,t}(z_i)$ to denote the potential treatment received under assignment z_i. With the joint distribution of $D_{i,t}(1)$ and $D_{i,t}(0)$, the principal strata $S_{i,t}$ at time t can be fully determined. As noted in Section 2.1, in the context of a two-arm randomized trial, the principal strata consist of compliers, always-takers, not-takers and defiers. The joint distribution of $D_{i,t}(1)$ and $D_{i,t}(0)$ fully determines the principal strata $S_{i,t}$ at time t. Since only one of $D_{i,t}(1)$ or $D_{i,t}(0)$ is observed, $S_{i,t}$ is latent for all subjects:

$$S_{i,t} = \begin{cases} c \text{ (complier)}, & D_{i,t}(z_i) = z_i \\ n \text{ (not-taker)}, & D_{i,t}(z_i) = 0 \\ a \text{ (always-taker)}, & D_{i,t}(z_i) = 1 \\ d \text{ (defier)}, & D_{i,t}(z_i) = 1 - z_i \end{cases}$$

We focus on two-arm randomized trials where subjects assigned to the control

group could not access treatment. Therefore principal strata consist of compliers and not-takers only, and subjects' principal strata become partially latent. The principal strata are observed for subjects randomized to the treatment group (compliers if take treatment and not-takers if take control), and remain latent and unobserved for subjects randomized to the control group.

2.3.2 Model Assumptions

We make two assumptions in the proposed Markov compliance and outcome model.

- *Ignorable Treatment Assignment Assumption* [47].

  This assumption means that the treatment assignment is independent of all (observed and unobserved) baseline variables and potential outcomes. Under ignorability, we do not need to model the assignment mechanism. It is reasonable in the CBT study because subjects were randomly assigned to either the treatment group or the control group.

- *Stable Unit Treatment Value Assumption (SUTVA)* [48].

  It comprises two subassumptions. The first subassumption implies there is no interference between subjects, i.e. the potential compliance and potential clinical outcome of individual i are independent of potential compliance and potential clinical outcome j ($i \neq j$). The first subassumption is reasonable for the CBT study because depression is not infectious, and subjects visited therapists by individuals, instead of in groups. The second subassumption assumes that there is no "hidden" version of the treatment, i.e. there are no systematic differences of treatments assigned within the treatment categories (CBT vs. usual care). This assumption is reasonable in the CBT study given that the usual care was

taken to standardize treatment provided in the CBT study.

- *Markov Dependence for Longitudinal Measurements*

  We assume a Markov relationship among longitudinal potential outcomes and principal strata memberships at different follow up time. In particular, we assume the potential outcomes at the end of the follow up period t depend not only on the principal stratification membership at the current time, but also on the principal strata in previous k_1 follow up periods as well as on the potential outcomes at the end of previous k_2 follow up periods. Similarly, we assume the principal strata in the follow up period t depend on the principal strata in previous l_1 follow up periods and the potential outcomes at the end of previous l_2 follow up periods.

Note that the ITT effect of the treatment within the not-taker stratum is often assumed zero in causal models for clinical studies, which is termed the exclusion restriction (ER) assumption [3]. This assumption is plausible in many studies, but not always. For example, in the CBT study, the clinical outcome is a subject's depression severity. The not-takers being assigned to receive cognitive therapy may experience stress as a result of failing to participate in the therapy, and thus become more depressed, but may not be stressed about not participating in the therapy if they were assigned to receive usual care. Therefore not-takers may develop different level of depression under different assignments, and ER assumption may not necessarily be met in this case.

2.3.3 Complete Data Likelihood

Under the potential outcome framework, the "complete" data include $\boldsymbol{Y} = (Y_{1,1}(1), Y_{1,1}(0), ..., Y_{n,T}(1), Y_{n,T}(0))$ and $\boldsymbol{S} = (S_{1,1}, ..., S_{n,T})$. Let $\boldsymbol{X}$ denote $n \times p$ matrix of base-

line covariates with i^{th} row $\boldsymbol{X}_i$. The complete data likelihood can be factored into a series of conditional likelihoods:

$$
\begin{aligned}
& L(\boldsymbol{Y}, \boldsymbol{S} | \boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \rho, \sigma) \\
= & \prod_{i=1}^{n} f(Y_{i,1}(1), Y_{i,1}(0), ..., Y_{i,T}(1), Y_{i,T}(0), S_{i,1}, ..., S_{i,T} | \boldsymbol{X}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \rho, \sigma) \\
= & \prod_{i=1}^{n} \prod_{t=2}^{T} f(S_{i,1} | \boldsymbol{X}_i, \boldsymbol{\alpha}) \times f(Y_{i,1}(1), Y_{i,1}(0) | S_{i,1}, \boldsymbol{\beta}, \rho, \sigma) \\
& \quad \times f(S_{i,t} | S_{i,t-1}, ..., S_{i,t-K_c^c}, Y_{i,t-1}(1), Y_{i,t-1}(0), ...Y_{i,t-K_c^y}(1), Y_{i,t-K_y^c}(0), \boldsymbol{\theta}) \\
& \quad \times f(Y_{i,t}(1), Y_{i,t}(0) | S_{i,t}, ..., S_{i,t-K_c^y}, Y_{i,t-1}(1), Y_{i,t-1}(0), ..., Y_{i,t-K_y^y}(1), Y_{i,t-K_y^y}(0), \boldsymbol{\gamma}, \rho, \sigma)
\end{aligned}
$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are used to parameterize the likelihood of principal strata in the first and the other follow up periods respectively, and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, ρ and σ are used to parameterize the likelihood of potential outcomes at the end of the first and the other follow up periods respectively, with ρ being the correlation between the potential outcomes under different treatment arms.

Typically, the correlation between the potential outcomes of a subject under different treatment arms is ignored, or a location shift between $Y_{i,t}(1)$ and $Y_{i,t}(0)$ ($\rho = 1$) is assumed [12]. In practice, because we could observe the potential outcome only under the arm the subject eventually takes, no information is available to estimate this within-subject correlation, at least in continuous settings. Viewing ρ as a sensitivity parameter, we assume a variety of correlations between 0 and 1, and investigate its sensitivity of the proposed Markov compliance and outcome. This correlation has little impact when making inference about ITT effects in large finite populations and no impact on such inference in superpopulations [49]. But in this setting, this correlation may have considerable impact on the estimates of parameters providing information about the impact of outcomes on compliance.

We propose appropriate parametric models for each conditional likelihood:

1. *Principal strata in the first follow up period* $(S_{i,1})$ – *Probit regression.*

   We focus on the situation where principal strata are binary and consist only compliers and not takers, and we choose probit regression to conditional on the baseline covariates:

$$P(S_{i,1} = c|\boldsymbol{X}_i, \boldsymbol{\alpha}) = \Phi(\boldsymbol{X}_i, \boldsymbol{\alpha}) = \Phi(\alpha_0 + \boldsymbol{X}_i'\boldsymbol{\alpha}_1)$$

2. *Potential outcomes at the end of first follow up period* $(Y_{i,1}(1), Y_{i,1}(0))$ – *Bivariate normal distribution.*

   Conditional on the principal strata in the first follow up period, we assume the potential outcomes at the end of first follow up period follow a bivariate normal distribution with correlation ρ:

$$Y_{i,1}(1), Y_{i,1}(0)|S_{i,1}, \boldsymbol{\beta}, \rho, \sigma \sim MVN(\boldsymbol{\mu}_{i,1}, \Sigma);$$
$$\boldsymbol{\mu}_{i,1} = \begin{pmatrix} \beta_1 + \beta_{c1} I(S_{i,1} = c) \\ \beta_0 + \beta_{c0} I(S_{i,1} = c) \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

3. *Principal strata in follow up period* t $(S_{i,t})$ – *Probit regression.*

   For principal strata in follow up period t $(t > 1)$, we assume a single-order Markov dependence at this stage $(l_1 = l_2 = 1)$ at this stage. This Markov dependence allows for interactions between the effects of previous potential outcomes and principal strata, but is assumed to be independent of time:

$$\begin{aligned} & P(S_{i,t} = c|Y_{i,t-1}(1), Y_{i,t-1}(0), S_{i,t-1}, \boldsymbol{\theta}) \\ =& \Phi(\theta_t + \theta_0 Y_{i,t-1}(1) + \theta_y(Y_{i,t-1}(0) - Y_{i,t-1}(1)) + \theta_c I(S_{i,t-1} = c) + \\ & \theta_{yc}(Y_{i,t-1}(0) - Y_{i,t-1}(1)) I(S_{i,t-1} = c)) \end{aligned}$$

4. *Potential outcomes at the end of follow up period t* $(Y_{i,t}(1), Y_{i,t}(0))$ – *Bivariate normal distribution.*

   For potential outcomes at the end of follow up period t $(t > 1)$, we assume a bivariate normal distribution with a single-order Markov dependence on the previous potential outcomes and a zero-order Markov dependence on the principal strata at this stage, i.e. depending only on the principal strata membership at time t $(k_1 = 0, k_2 = 1)$. This Markov relationship is also assumed to be independent of time:

$$(Y_{i,t}(1), Y_{i,t}(0))|Y_{i,t-1}(1), Y_{i,t-1}(0), S_{i,t}, \boldsymbol{\gamma}, \rho, \sigma \sim MVN(\boldsymbol{\mu}_{i,t}, \Sigma);$$

$$\boldsymbol{\mu}_{i,t} = \begin{pmatrix} \gamma_{1t} + \gamma_0 Y_{i,t-1}(1) + \gamma_y (Y_{i,t-1}(0) - Y_{i,t-1}(1)) + \gamma_{c1t} I(S_{i,t} = c) \\ \gamma_{0t} + \gamma_0 Y_{i,t-1}(1) + \gamma_y (Y_{i,t-1}(0) - Y_{i,t-1}(1)) + \gamma_{c0t} I(S_{i,t} = c) \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

2.3.4 Estimates of Interest – Principal Effects

Imbens and Rubin (1997) defined the principal effect as the ITT effect of treatment within principal strata [25]. With the proposed Markov compliance and outcome model, we define the estimates of interest – principal effects – conditional on the current principal stratum membership and the previous potential outcomes:

$$ITT_{c,t} = E\left(Y_{i,t}(1) - Y_{i,t}(0)|S_{i,t} = c, Y_{i,t-1}(1), Y_{i,t-1}(0)\right) = (\gamma_{1t} - \gamma_{0t}) + (\gamma_{c1t} - \gamma_{c0t})$$

$$ITT_{n,t} = E\left(Y_{i,t}(1) - Y_{i,t}(0)|S_{i,t} = n, Y_{i,t-1}(1), Y_{i,t-1}(0)\right) = (\gamma_{1t} - \gamma_{0t})$$

Note that the proposed Markov compliance and outcome model allows us to obtain relatively simple estimates of within-principal-stratum ITT effects. If the dependence between the t and t-1 potential outcomes were to differ by *treatment assignment*, then

we replace γ_0 and γ_1 with γ_{0z} and γ_{1z}, yielding

$$ITT_{c,t} = E\left(Y_{i,t}(1) - Y_{i,t}(0)|S_{i,t} = c, Y_{i,t-1}(1), Y_{i,t-1}(0)\right) =$$
$$(\gamma_{1t} - \gamma_{0t}) + (\gamma_{c1t} - \gamma_{c0t}) + \ (\gamma_{01} - \gamma_{00})Y_{i,t-1}(0) + (\gamma_{11} - \gamma_{10})(Y_{i,t-1}(1) - Y_{i,t-1}(0))$$
$$ITT_{n,t} = E\left(Y_{i,t}(1) - Y_{i,t}(0)|S_{it} = n, Y_{i,t-1}(1), Y_{i,t-1}(0)\right) =$$
$$(\gamma_{1t} - \gamma_{0t}) + (\gamma_{01} - \gamma_{00})Y_{i,t-1}(0) + (\gamma_{11} - \gamma_{10})(Y_{i,t-1}(1) - Y_{i,t-1}(0))$$

One could view this model as assuming a sort of second-order failure of the ER, namely that treatment assignment by itself affects dependence between the t and t-1 potential outcomes. On the face it seems rather implausible, although in this situation $\gamma_{c1t} - \gamma_{c0t}$ still encapsulates the difference between the $\text{ITT}_{c,t}$ effect and the $\text{ITT}_{n,t}$ effect. A more serious situation would be that the previous principal stratum membership still carries information about the ITT effect even after conditioning on the current principal stratum membership and the previous potential outcomes. In this setting the principal strata at time t would be defined in terms of vectors of principal strata: for example, $((S_{i,t-1} = c, S_{i,t} = c), (S_{i,t-1} = c, S_{i,t} = n), (S_{i,t-1} = n, S_{i,t} = c), (S_{i,t-1} = n, S_{i,t} = n))$ if a one-degree Markov relationship with principal compliance held. Simple causal interpretation of these strata is no longer possible, although use of a clustering algorithm could allow for relatively interpretable strata such as high or low compliers [32, 33]. We do not pursue this extension in this dissertation.

2.3.5 Model Estimation

Because of the complex missing data structure, a Bayesian approach with Markov chain Monte Carlo (MCMC) algorithms becomes a natural estimation method [19, 55]. We use a Gibbs sampling algorithm to obtain random draws of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ^2 from their posterior distributions, and use Metropolis-Hastings within Gibbs algorithm to

obtain random draws of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ from their posterior distributions. The Metropolis-Hastings within Gibbs algorithm is also used to obtain random draws from the full conditional distributions of the missing potential outcomes and principal strata. The detailed descriptions of the posterior distributions of the parameters and the full conditional distributions of missing potential outcomes and principal strata are given in the Appendix. To check convergence, we calculate a measure of between and within-chain variance $\hat{R}$ [19]. $\hat{R} < 1.1$ is acceptable, and a $\hat{R}$ close to 1 indicates the good convergence of the parameters.

2.3.6 Model Fit Assessment

To assess goodness of fit, we consider the posterior predictive p-values [19]. Since the principal compliance of subjects randomized to the treatment group are equal to the observed compliance, we compare the posterior predictive distribution (PPD) of the principal compliance and assess their fit to the observed principal compliance. We make posterior predictive checks of the fitted models using the percentage of compliers in the treatment as the test statistics, and compare the observed percentage of compliers with their posterior predictive distribution to obtain PPD p values. A PPD p value close to 0.50 indicates good fit of the model to the data. Similarly, we make posterior predictive checks of the potential outcomes using the observed potential outcomes. We use a Chi-square type test of the form $T = \sum_{i=1}^{n} \sum_{j=1}^{n_i}((y_{i,t} - \mu_{i,t})^2 = \sigma^2)$, where $y_{i,t} = z_i * Y_{i,t}(1) + (1 - z_i)Y_{i,t}(0)$ and $\mu_{i,t} = z_i * \mu_{i,t}(1) + (1 - z_i)\mu_{i,t}(0)$. The distribution of this test statistic should be close to the Chi-square distribution with $\sum n_i$ degree of freedom. A Q-Q plot of the test statistics and the Chi-square statistics around the 45 degree line indicates good fit of the model to the data.

2.4 Application to the CBT Study

2.4.1 High Correlation of Potential Outcomes

We apply the proposed Markov compliance and outcome model to the CBT study. We assume relatively flat priors because we do not have strong prior knowledge for the CBT study. Specifically, we let $\boldsymbol{\beta} \sim MVN(\mathbf{0}, 10000I)$, $\boldsymbol{\gamma} \sim MVN(\mathbf{0}, 10000I)$, $\boldsymbol{\theta} \sim MVN(\mathbf{0}, 10I)$, $\sigma^2 \sim$ Inv-$\chi^2(1,1)$. Ten is chosen as the variance of prior of $\boldsymbol{\theta}$ because it provides a relatively flat prior on the probability scale when the range of outcomes is take into account. We run two chains of 100,000 after an initial burn-in of 50,000. The maximum value of $\hat{R}$ is $1.03 < 1.1$, indicating all parameters converge in distribution. In addition to the missing potential outcome under the arm to which the subject is not assigned, the CBT study have 19% of subjects having one or more of their BDI outcomes not measured or recorded. We assume a missing at random (MAR) mechanism for these missing data, i.e. conditional on the observed outcomes, compliance behavior and baseline covariates, the missingness mechanism is assumed to be random. For these missing data, we impute the potential outcomes under both treatment arms in each iteration of MCMC.

To be consistent with previous research that usually assumes (at least implicitly) $\rho = 1$ between the two potential outcomes under treatment and control, we focus first on the situation where this correlation is high ($\rho = 0.9$). Table 2.2 summarizes the estimated $\text{ITT}_{c,t}$ and $\text{ITT}_{n,t}$ for the cognitive therapy. For compliers, the cognitive therapy consistently lowers depression severity more effectively than the usual care. On average among compliers, the cognitive therapy lowers $\sqrt{BDI}$ 0.63 (95% CI: 0.05, 1.27) more than that under usual care at 1 month, and this effect increases to 2.22 (95% CI: 1.02, 3.35) at 12 months. For not-takers, because they would not

Table 2.2: $\text{ITT}_{c,t}$, $\text{ITT}_{n,t}$, ITT and AT effects for the CBT study (Median with 95% credible interval in parenthesis)

		Month			
	ρ	1	3	6	12
$\text{ITT}_{c,t}$	0.9	-0.63 (-1.27, -0.05)	-0.68 (-1.48, 0)	-1.09 (-1.87, -0.34)	-2.22 (-3.35, -1.02)
	0.5	-0.66 (-1.28, -0.03)	-0.74 (-1.52, 0.02)	-1.42 (-2.36, -0.50)	-2.40 (-3.67, -0.90)
	0.1	-0.72 (-1.32, -0.11)	-0.80 (-1.54, -0.07)	-1.49 (-2.43, -0.59)	-2.29 (-3.72, -0.70)
$\text{ITT}_{n,t}$	0.9	1.58 (0.38, 2.90)	1.87 (0.44, 3.29)	0.45 (-0.62, 1.56)	0.16 (-0.82, 1.20)
	0.5	1.87 (0.16, 3.62)	1.63 (-0.77, 3.63)	0.56 (-0.85, 1.89)	0.06 (-1.10, 1.25)
	0.1	1.85 (0.17, 3.59)	0.77 (-1.44, 2.97)	0.12 (-1.30, 1.58)	-0.26 (-1.44, 0.93)
ITT		-0.01 (-0.63, 0.60)	-0.22 (-0.85, 0.40)	-0.66 (-1.29, 0.03)	-0.81 (-1.45, -0.17)
AT		0.12 (-0.43, 0.66)	0.09 (-0.64, 0.47)	-0.75 (-1.33, -0.17)	-1.01 (-1.70, -0.31)

participate in the cognitive therapy regardless the group to which they are assigned, we can view $\text{ITT}_{n,t}$ as the effect of being assigned to receive cognitive therapy. The results show that subjects who are randomized to the cognitive therapy group but unwilling or unable to complete treatment during the first month have average $\sqrt{BDI}$ scores 1.58 (95% CI: 0.38, 2.90) higher than that if they had not been assigned to receive cognitive therapy. The results imply that being assigned to cognitive therapy – though not cognitive therapy itself – is harmful to not-takers at the beginning of the study. This effect decreases over time, with the 95% credible interval including 0 from 6 months on.

To compare our proposed model with the traditional analysis methods, we conduct ITT analyses and AT analyses on the CBT study, and summarize the results in Table 2.2. For ITT analyses, we fit a linear mixed model with a random intercept for each subject and fixed-effect indicators for follow-up periods, treatment assignment and their interaction. The analyses results show that being assigned to cognitive therapy lowers depression severity on average in all follow up periods, although these effects are smaller than the $\text{ITT}_{c,t}$ of cognitive therapy, and the CIs do not exclude 0 until 6 months. Note that the effect of treatment assignment is different from the effect of cognitive therapy in the CBT study because of the large proportion of non-

Table 2.3: Parameters governing principal strata for follow up period t ($t > 1$) for the CBT study (Median with 95% credible interval in parenthesis)

ρ	Causal Effect at $t-1$ in non-takers (θ_y)	Causal Effect at $t-1$ in compliers ($\theta_y + \theta_{yc}$)	Effect of Compliance at $t-1$ (θ_c)
0.9	3.19 (1.15, 6.26)	4.59 (1.73, 9.05)	-2.03 (-5.22, 0.53)
0.5	1.26 (-0.52, 4.05)	1.43 (-0.65, 3.60)	0.37 (-2.68, 2.96)
0.1	0.41 (-0.74, 2.01)	0.49 (-0.31, 1.75)	1.29 (-0.74, 2.87)

compliance. In AT analysis, we use a model similar to that in the ITT analysis but replace treatment assigned with treatment taken. The AT analysis shows that cognitive therapy lowers depression severity from 6 months after randomization. However, because of self-selection, subjects participating in the cognitive therapy and subjects participating in the usual care might no longer be comparable, and estimates of the effect of cognitive therapy from AT analysis might be biased.

In Table 2.3, we summarize the estimated parameters modeling principal strata in the follow up period t ($t > 1$). The coefficients θ_y and $\theta_y + \theta_{yc}$ estimate the impact of the causal effect of cognitive therapy at the end of follow up period $t-1$ on the principal strata at follow up period t among not-takers and compliers respectively. The positive 95% credible intervals indicates that the probability of being a complier at follow up period t increases as the causal effect of cognitive therapy at the end of follow up period $t-1$ increases. The quantity $\theta_y + \theta_{yc}$ is larger, indicating the effect is stronger for compliers than not-takers at follow up period $t-1$. The parameter θ_c shows the effect of principal compliance at $t-1$ on that at t among those for whom the treatment is neither harmful nor beneficial. The results show that the principal compliance at previous follow up period is not predictive among this group.

The parameters α_{gender} and α_{bdi} measure the impact of baseline covariates on the principal strata membership in the first follow up period. The 95% credible interval of α_{bdi} is positive ($\alpha_{\text{BDI}} = 0.7$ (0.47, 0.87)), indicating that subjects have higher

probability of being compliers in the first follow up period if subjects have more severe depression at the time of randomization. The parameter α_{gender} is estimated to be positive with a 95% credible interval covering zero ($\alpha_{gender} = 0.62$ (-0.30, 0.84)), indicating that women tend to be marginally more likely to comply in the first follow up period.

To assess goodness of fit, we check the PPD of subjects' principal strata and potential clinical outcomes. The p values of the PPD of principal strata are 0.83, 0.69, 0.53, and 0.47 for follow up periods 1, 3, 6, and 12 months respectively, indicating a good fit for principal strata. The Q-Q plot in Figure 2.2 (in Appendix) shows the distribution of the potential clinical outcome test statistics is close to the model-predicted chi-square distribution with 416 degrees of freedom, indicating that the normality assumption for the transformed BDI measure is reasonable.

2.4.2 Moderate or Small Correlation of Potential Outcomes

Because there is no data available to assess the correlation of the potential outcomes, we treat the correlation as a sensitivity parameter and consider its impact of different values on our estimates of interest. The previous section assumed high correlation ($\rho = 0.9$); here we consider moderate ($\rho = 0.5$) or small correlation ($\rho = 0.1$).

The results in Table 2.2 imply that both $\text{ITT}_{c,t}$ and $\text{ITT}_{n,t}$ are not very sensitive to the correlation of the potential outcomes. Under both moderate and small correlations, the cognitive therapy consistently has better effects on lowering depression severity than the usual care for compliers. The magnitudes of the difference are very similar to those under high correlation of potential outcomes. For not-takers, under both moderate and small correlations, being assigned to receive cognitive therapy

makes subjects' depression more severe than that if being assigned to receive usual care at the beginning of the study. This effect decreases as time increases, and disappears at 1 year after randomization.

The parameters modeling the principal strata are summarized in Table 2.3. The results show that, as the correlation of potential outcomes becomes smaller, the impact of the causal effect of cognitive therapy among not-takers at the previous time period (θ_y) and similarly the impact of the causal effect of cognitive therapy among compliers at the previous time period ($\theta_y + \theta_{yc}$) attenuates toward null, while the impact of being a complier at the previous time period (θ_c) deviates away from null simultaneously. These results are to be expected, because as the within-subject correlation increases, there will be less uncertainty in predicting the potential outcomes under the arm to which the subject is not assigned (the difference is a location shift when correlation is equal to one). On the contrary, when the within-subject correlation decreases, more random noise is present in the potential outcome difference, possibly due to the fluctuations in mood and anxiety of patients independent of the treatment. Thus the impact of the difference of the potential outcomes on future compliance is attenuated.

2.5 Simulation Studies

Although we utilize a Bayesian framework for our analysis, we are still interested in the repeated sampling properties of our proposed model. Because of the relatively long time necessary to fit the model to a single simulated data set, we simulate 50 data sets of 100 subjects each with 4 follow up periods. To make subjects' compliance behaviors and clinical outcomes similar to the CBT study data, each data set is simulated under $\boldsymbol{\alpha} = (\alpha_0,\ \alpha_{\text{gender}},\ \alpha_{\text{bdi}})$ = (-3.0, 0.6, 0.8), $\boldsymbol{\beta} = (\beta_1,\ \beta_0,\ \beta_{c1},\ \beta_{c0}) =$

Table 2.4: $\text{ITT}_{t,t}$ and $\text{ITT}_{n,t}$ with 50 simulations

	Coverage of 95% C.I.		Median (True Value)	
Month	$\text{ITT}_{c,t}$	$\text{ITT}_{n,t}$	$\text{ITT}_{c,t}$	$\text{ITT}_{n,t}$
1	94%	92%	-0.6 (-0.7)	1.3 (1.4)
3	92%	90%	-0.9 (-0.9)	1.6 (1.5)
6	92%	94%	-1.2 (-1.2)	0 (0.1)
12	90%	90%	-1.7 (-2.2)	-0.2 (-0.1)

Table 2.5: Parameters governing principal strata with 50 simulations

	α_{gender}	α_{bdi}	θ_y	θ_c	θ_{yc}
Coverage of 95% C.I.	94%	92%	90%	88%	96%
Median (True Value)	0.8 (0.6)	0.8 (0.8)	2.1 (2.1)	-1.1 (-1.2)	1.6 (1.1)

(3.1, 1.7, 1.2, 3.3), $\boldsymbol{\gamma} = (\gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{02}, \gamma_{03}, \gamma_{04}, \gamma_0, \gamma_y, \gamma_{c12}, \gamma_{c13}, \gamma_{c14}, \gamma_{c02}, \gamma_{c03}, \gamma_{c04}) = (2.5, 1.4, 1.0, 1.0, 1.3, 1.1, 0.6, 1.0, -2.0, -1.9, -1.9, 0.4, -0.6, 0.2)$, $\boldsymbol{\theta} = (\theta_2, \theta_3, \theta_4, \theta_0, \theta_y, \theta_c, \theta_{yc}) = (0.7, -2.6, -4.7, 0.6, 2.1, -1.2, 1.1)$, $\rho = 0.9$, and $\sigma^2 = 2.1$. We analyze the simulated data using proposed Markov compliance and outcome model.

We summarize the analysis results of the estimated $\text{ITT}_{c,t}$ and $\text{ITT}_{n,t}$ in each follow up period along with other parameters of interest in Table 2.4 and Table 2.5. Given the modest number of simulations due to the relatively long time necessary to fit the model to a single simulated data set, very accurate assessments of the repeated sampling properties of our Bayesian model are not possible; however all parameters appear to have relatively low amounts of bias and either approximately correct coverage or modest undercoverage.

2.6 Conclusions and Discussion

Our proposed model is different from previous research in that, in addition to estimating the causal effect of cognitive therapy, our proposed model allows us to estimate the impact of causal effect of cognitive therapy and principal strata in the follow up period t on principal strata in the current follow up period $t + 1$. In the

CBT study, the results show that the stronger the cognitive therapy is at the end of follow up period t, the more likely the subjects will be compliers in the follow up period $t+1$. This association is stronger for compliers than not-takers at time t. Our finding imply that subjects are sensing whether the treatment is effective for them, and adapting their compliance behavior accordingly. It is important to note that this result refers to the unobservable potential effect of the treatment on a given subject.

Similar to previous work, our proposed model accommodates time-varying latent principal strata [15, 32, 33]. However, our proposed model is different from their work in that our proposed model does not have a "super" principal strata, which summarizes the longitudinal pattern of compliance behaviors. Thus, in addition to "unpacking" the effect of potential outcomes at time $t-1$ on compliance at time t, our proposed model keeps the advantage of a valid interpretation of causal effect within each principal stratum at each follow up time t.

The proposed causal model is different from previous work in model assumptions. First, we relax the ER assumption, which is commonly assumed in causal modeling [3, 10, 25, 60]. In the CBT study, the main clinical outcome is depression severity. It is special because even being assigned to receiving treatment may affect it. The analysis results show that being assigned to receive the cognitive therapy is harmful for not-takers at the beginning of the study, though this effect eventually disappeared at the end of one year. Because avoiding enrollment of non-compliant individuals can be very difficult, especially in a mental health study, these results emphasize the importance of carefully monitoring non-takers, especially those who are non-compliant with treatment during the early period of the study. On the other hand, there might be possibility that failing to comply with one's assigned treatment status

is a marker of stress and other conditions that themselves impede the observed positive outcomes.

The robustness of outcome model assumption deserves further discussion. The proposed causal model assumes a bivariate normal distribution for the depression measure outcomes, such that the subjects who assigned to control group consist of compilers and non-takers that correspond to a mixture of two normal distributions. In addition, by the definition of randomization, the proportions of principal strata memberships among subjects assigned to the control group are same as those among subjects assigned to the treatment arm. As long as the proportions of principal strata membership are different, the causal effects are identifiable. In practice, the robustness of the estimation of the proposed causal model depends on the parametric assumptions; in particular, the results may be sensitive to the presence of skewness, failing to distinguish between a skewed unimodal distribution (under which the exclusion restriction may hold) and a bivariate normal distribution under which the exclusion restriction fails. Additional work should be done to explore the robustness of the proposed causal model under failures of the normality assumption.

Our research also differs in the way of modeling potential outcomes. Previous research typically assumed a location shift between the potential outcomes under treatment and control [12]. Such models assume a perfect correlation between the potential outcomes in all cases. In reality, it is reasonable to consider a correlation between 0 and 1. Similar to Jin and Rubin [26], we consider the correlation as a sensitivity parameter, and assess the sensitivity of analysis results to this correlation. Our results show that the principal effects are relatively insensitive to the choice of this correlation. There is more sensitivity to the correlation for the prediction of future compliance behavior. If we assume the within-subject correlation of potential

outcomes to be high, the effectiveness of the treatment for a subject is a much more important predictor for future compliance than previous compliance. If the within-subject correlation of potential outcomes is assumed to be low, then within-subject treatment effectiveness has little association, and previous compliance better predicts future compliance.

We build our Markov compliance class and outcome model assuming a single-order Markov relationship. This model has the advantage of clear interpretation of model parameters. The posterior predictive checks for compliance classes and potential outcomes show that the single order Markov relationship model provides sufficient fit for the CBT study. However, our proposed model could be extended to a higher-order Markov relationship when necessary. For example, the compliance class in the follow up period t may depend on compliance class in the follow up period $t-1$ and $t-2$, as well as the treatment effect at the end of follow up period $t-1$ and $t-2$, and their interactions. With a higher-order Markov relationship, there could be multiple interaction terms in the model, and interpretation of both the causal effect of treatment and predictors of compliance becomes more complex.

Other extensions are possible as well. The mixture model for the potential outcomes is not non-parametrically identified in the absence of the ER assumption [38]. Here we rely on the normality assumption to identify the mixture components associated with compliers and not-takers in the control group. An alternative to this approach would be to weaken or eliminate this parametric assumption and rely on either observed predictors of compliance [46] or prior distributional assumptions to induce posterior modes.

Appendix

Gibbs Sampler for Markov Compliance and Outcome Model Estimation

Let $\boldsymbol{Y}_{i,t}$ denote $(Y_{i,t}(Z=1), Y_{i,t}(Z=0))$ of subject i at the end of follow up period t. Let $M_{i,t}$ denote the design matrix of subject i for the potential outcomes at the end of the follow up period t.

1. Draw of $\boldsymbol{\beta} \mid$ rest .

   The posterior distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned}
&\boldsymbol{\beta}| \text{ rest } \sim MVN(\hat{\boldsymbol{\mu}}_{n\beta}, \hat{\Sigma}_{n\beta}),\\
&\hat{\boldsymbol{\mu}}_{n\beta} = (\Sigma_\beta^{-1} + \sum_{i=1}^{n} M_{i,1}\Sigma^{-1}M_{i,1}')^{-1}(\Sigma_\beta^{-1}\boldsymbol{\mu}_\beta + \sum_{i=1}^{n} M_{i,1}\Sigma^{-1}\boldsymbol{Y}_{i,1}),\\
&\hat{\Sigma}_{n\beta} = (\Sigma_\beta^{-1} + \sum_{i=1}^{n} M_{i,1}\Sigma^{-1}M_{i,1}')^{-1}
\end{aligned}$$

2. Draw of $\boldsymbol{\gamma} \mid$ rest .

   The posterior distribution of $\boldsymbol{\gamma}$ is

$$\begin{aligned}
&\hat{\boldsymbol{\gamma}}| \text{ rest } \sim MVN(\hat{\boldsymbol{\mu}}_{n\gamma}, \hat{\Sigma}_{n\gamma}),\\
&\hat{\boldsymbol{\mu}}_{n\gamma} = (\Sigma_\gamma^{-1} + \sum_{i=1}^{n}\sum_{t=2}^{T} M_{i,t}\Sigma^{-1}M_{i,t}')^{-1}(\Sigma_\gamma^{-1}\boldsymbol{\mu}_\gamma + \sum_{i=1}^{n}\sum_{t=2}^{T} M_{i,t}\Sigma^{-1}\boldsymbol{Y}_{i,t}),\\
&\hat{\Sigma}_{n\gamma} = (\Sigma_\gamma^{-1} + \sum_{i=1}^{N}\sum_{t=2}^{T} M_{i,t}\Sigma^{-1}M_{i,t}')^{-1}
\end{aligned}$$

3. Draw of $\sigma^2 \mid$ rest .

   The posterior distribution of σ^2 is

$$\begin{aligned}
&\sigma^2| \text{ rest } \sim \text{Inv} - \chi^2(\nu_n, \psi_n),\\
&\nu_n = 2nT + \nu\\
&\psi_n = \left\{ \nu\psi + \frac{1}{1-\rho^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})'(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t}) - 2\rho(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})'(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})\right]\right\}\frac{1}{2nT+\nu}
\end{aligned}$$

4. Draw of counterfactual compliance.

The distribution of the unobserved compliance conditional on the observed data and parameters is

$$P(S_{i1} = c| \text{ rest }) = \begin{cases} \frac{f(Y_{i1}(1),Y_{i1}(0),S_{i1}=c)\Phi(\alpha_0^{'}+\boldsymbol{\alpha_1 x_i})}{f(Y_{i1}(1),Y_{i1}(0)|S_{i1}=c,\boldsymbol{\beta})\Phi(\alpha_0^{'}+\boldsymbol{\alpha_1 x_i})+f(Y_{i1}(1),Y_{i1}(0)|S_{i1}=n,\boldsymbol{\beta})(1-\Phi(\alpha_0+\boldsymbol{\alpha}_1^{'}\boldsymbol{x_i}))}; Z_i = 0 \\ 1; Z_i = 1, D_{i1}(1) = 1 \\ 0; Z_i = 1, D_{i1}(1) = 0 \end{cases}$$

$$P(S_{it} = c| \text{ rest }) = \begin{cases} \frac{p_c}{p_c+p_n} & Z_i = 0 \\ 1, & Z_i = 1, D_{it} = 1 \\ 0, & Z_i = 1, D_{it} = 0, \end{cases}$$

$$p_c = (\Phi(\mu_\theta^{t+1})I(S_{i,t+1} = c) + (1 - \Phi(\mu_\theta^{t+1}))I(S_{i,t+1} = n))f(Y_{it}(1), Y_{it}(0)|S_{it} = c, \boldsymbol{\gamma})\Phi(\mu_\theta^t),$$

$$p_n = (\Phi(\mu_\theta^{t+1})I(S_{i,t+1} = c) + (1 - \Phi(\mu_\theta^{t+1}))I(S_{i,t+1} = n))f(Y_{it}(1), Y_{it}(0)|S_{it} = n, \boldsymbol{\gamma})(1 - \Phi(\mu_\theta^t)),$$

$$\mu_\theta^t = \theta_t + \theta_0 Y_{i,t-1}(1) + \theta_y(Y_{i,t-1}(0) - Y_{i,t-1}(1)) + \theta_c I(S_{i,t-1} = c) +$$

$$\theta_{yc}(Y_{i,t-1}(0) - Y_{i,t-1}(1))I(S_{i,t-1} = c)$$

Q-Q plot of the test statistic with Chi-square distribution

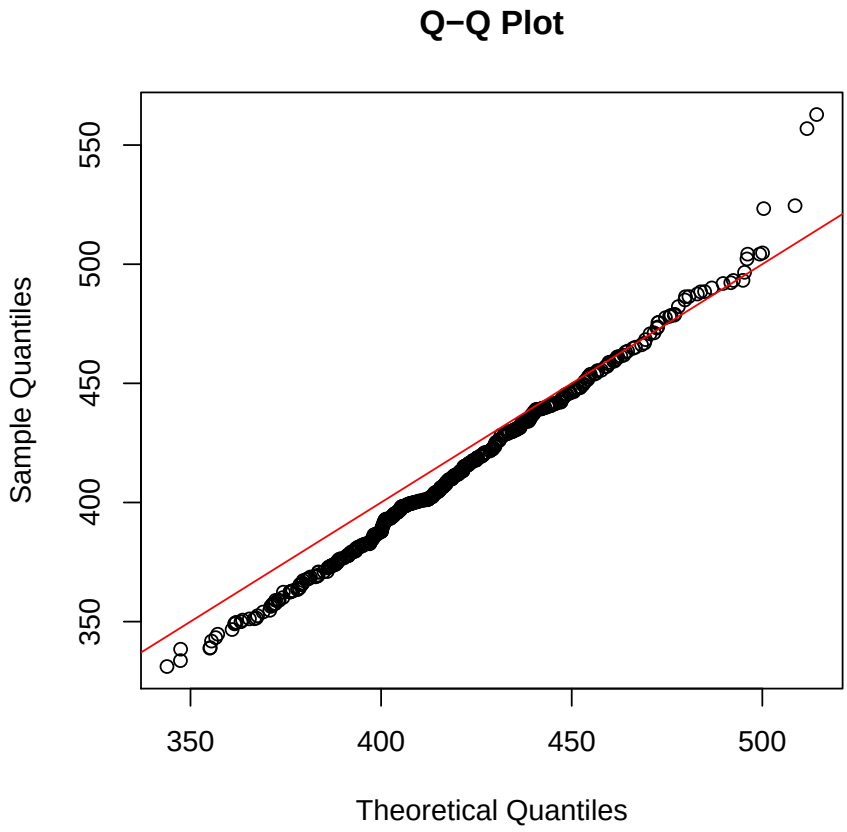


Figure 2.2: Q-Q plot of the test statistic with Chi-square distribution for the CBT study

CHAPTER III

Assessing the Causal Effect of Treatment in the Presence of Self-Selection of Dosage

3.1 Introduction

Much scientific research in the social and health sciences aims to understand the causal relationship between an intervention and outcome. Where ethically and logistically feasible, randomized assignments of intervention levels provide an effective means to assess these causal relationships. When subjects fail to follow their randomized intervention protocol, however, comparing results by randomized treatment arm – "intent to treat" analysis – provides inference about the causal effects of the assignment, but not the actual intervention. Comparisons among those adhering to the protocol –"per protocol" analysis – provides only evidence of association, not a causal interpretation, since treatment taken is a post-randomization variable [45].

Recent research in the field of causal inference is intended to help this gap. One fundamental concept in causal modeling is the potential outcome, which is defined as the value an outcome would take after assignment to a different treatment arm than the one actual observed [49]. A potential outcome is considered to exist before the action of randomization, and therefore can be considered and modeled as a pre-

randomization variable. The causal effect of the intervention is then defined as the comparison of the potential outcomes under different intervention arms [17, 25, 35, 50]. For example, the causal effect of taking a pill would be the comparison of one's headache now with what would have been if no pill had been taken [35]. One goal of causal inference is the estimation of the average causal effect of intervention, which is the comparison of the potential outcomes under different intervention arms within the individual, averaged over the whole population.

Over the years, two major approaches have been developed for causal inference in randomized trials with noncompliance. The first approach treats noncompliance as a special case of mediation analysis, and focuses on estimating direct effects of treatment assignment and indirect effects of treatment assignment through actual treatment taken [44]. This approach assumes that intervention assignments and intervention taken can be manipulated independently, so that, in principle at least, all subjects could be forced into compliance or non-compliance. Thus the (prescriptive) direct effect is defined as the expected value of the difference in the potential outcomes under different treatment assignments when the value of the intermediate variable (treatment taken) is held constant under different treatment arms, and the associated (prescriptive) indirect effect as the expected difference in the total effect (effect of assignment) and the prescriptive direct effect. Focusing on splitting the total effect into direct effect and indirect effect to understand the causal mechanisms of treatment assignment and treatment taken, this approach has been applied to many studies in the recent years [57, 58, 59].

The second major approach assesses the causal effect of intervention by estimating the causal effect within principal strata, defined by the joint distribution of the treatments taken under all possible assignments. Because non-compliance behavior

is a post-randomization variable, it cannot be simply conditioned on without destroying the causal interpretability of the treatment effect [45]. Consequently, as-treated (AT) analysis and per-protocol (PP) analysis violates the randomization rule, and thus bias may exist in the analysis results. Because the principal strata form a "pre-randomization" variable, it can be conditioned on or adjusted for while retaining a causal interpretation for the effect of intervention. In a two-arm randomized clinical trial with non-compliance, the principal strata consist of compliers (subjects who comply with the assignment under both treatment arm), always-takers (subjects who always take treatment regardless of treatment assignment), never-takers (subjects who always take control regardless of treatment assignment), and defiers (subjects who take the treatment opposite of what they are assigned). Causal inference focuses on the effect of treatment assignment among the compliers and defiers, whose treatment taken changes with the treatment assigned; in the typical settings that defiers are assumed not to exist, only the complier strata provides causal inference for the effect of treatment taken, since it is the only stratum in which treatment changes in parallel with randomized assignment. Thus, in contrast to the direct/indirect effect setting, principal stratification assumes that only a subset of the population will be compliant with treatment, and focuses inference on this stratum. This approach has been widely used in clinical and social studies when noncompliance behavior is present [26, 37, 42], including randomized studies with partial compliance [12], in longitudinal studies [32] and survival time analysis [9, 32] with partial compliance [26], and in surrogacy evaluations [56].

The study motivates our work is Interstitial Cystitis Collaborative Research Network (ICCRN) study, which is a randomized clinical trial designed to evaluate the efficacy of Amitriptyline for the treatment of interstitial cystitis. In this blinded two-

arm (treatment and placebo) randomized trial, subjects on both arms are put on a dose escalation schedule and encouraged to take up to 75 mg. However, primarily because of adverse events, many subjects opted to stay at or return to ≤ 25 mg or 50 mg. (Note that subjects on a sham drug can experience adverse events, the so-call "nocebo" effect [28], which can impact compliance behavior [4].) Thus this trial suffers from partial compliance [26], and traditional analyses that compare difference between the treated and control groups stratified by dose level, either unadjusted or adjusted for the observed adverse events, cannot provide a valid estimate of causal effect of treatment. Our goal is to stratify on the *treatment* dose tolerance, which we treat as a pre-randomization variable, observed for subjects assigned to treatment but not for those assigned to control. Thus treatment dose tolerance forms our principal strata, with our inferential target the effect of treatment assignment among subjects who are able to tolerate a specific maximum treatment dose.

Motivated by the ICCRN study, we develop a causal model following the principal stratification approach to estimate the causal effect of treatment. One advantage of the ICCRN study is that adverse events were reported and recorded in the trial for all subjects. We take advantage of this information by conditioning on them when modeling the principal strata membership under a "monotonicity" assumption that subjects experiencing more severe adverse events will be less tolerant of high drug dosages. However, because observed adverse event is a postrandomization variable, like observed compliance, it cannot be simply conditioned on as a covariate without destroying the causal interpretation for the model. Therefore, we include it in the manner of potential outcomes in our model, using the information about adverse events on one arm to inform about adverse events on the other arm under the assumption that adverse events will not be diminished by receiving treatment. Note

that, in contrast to the assumed relationship between adverse events and drug tolerance within a treatment arm, the assumed relationship between adverse events on different treatment arms within a subject cannot be directly assessed by the data, since subjects are observed to receive only one of the treatment arms, although there is information in the data to reject this "monotonicity" assumption if it is violated.

There is some precedent for our approach in the literature. The Lipid Research Clinics Coronary Primary Prevention Trial measured the effect of the drug cholestyramine and this study was analyzed by Efron [12]. The subjects in the study were measured for compliance (the proportion of the intended dose actually taken) and for cholesterol decrease. The compliance-response regression for the treatment arm showed a smooth increasing effect of the drug in cholesterol level with increasing compliance. Later this study was analyzed by Jin and Rubin [26], who utilized the principal stratification framework and found that given the compliance under control, the principal causal effect of treatment increases as compliance under treatment increases. Our approach differs from previous work in that a) we are presented with a dichotomous outcome, which makes identification more challenging, and b) we use adverse events to estimate principal stratification membership.

We introduce the details of ICCRN study in Section 3.2. Section 3.3 describes our proposed causal model. Section 3.4 applies our proposed model to the ICCRN study. Section 3.5 explores the repeated sampling properties of our model in a simulation study. Section 3.6 reviews our findings along with future work and conclusions. In addition to considering a problem of partial compliance, for which limited research is available, our work is innovative in that it includes predictors of compliance status (adverse events) observed post-randomization in a fashion that retains causal interpretability.

3.2 The Interstitial Cystitis Collaborative Research Network (ICCRN) Study

The ICCRN study is a randomized clinical trial designed to evaluate the efficacy of Amitriptyline, a tricyclic antidepressant, for the treatment of interstitial cystitis in newly diagnosed patients. Interstitial cystitis is a long-term (chronic) and painful inflammation of the bladder wall, and its cause remains unknown. Symptoms include pelvic pain, urinary discomfort, urinary frequency, urinary urgency, and pain during intercourse [40].

One of the main goals of this study is to evaluate the effects of Amitriptyline on interstitial cystitis using a two-arm randomized clinical trial design. There are 271 subjects enrolled and randomized to receive either oral Amitriptyline or a similar placebo "dose" regime. Dosing begins with 10 mg after randomization, increasing to 25 mg at 2 weeks, 50 mg at 4 weeks, and 75 mg at 6 weeks. Subjects may fail to increase or drop doses because of adverse events to a"maximum tolerated" level below 75 mg.

The outcome is a Global Response Assessment, measured at 12 weeks or study withdrawal. It measures overall improvement with therapy, and is used frequently as the primary end point in clinical trials of therapies for interstitial cystitis. The assessment asks "As compared to when you started the current study, how would you rate your overall symptoms now?" Subjects who indicate moderately improved or markedly improved to the assessment are considered responders to the treatment, and all others are considered non-responders.

Subjects' baseline demographic data by treatment arm for all patients ($n = 271$) are shown in Table 3.1. Approximately 83% of the patients are female and 74%

are white. The median age is 38 years (range 18 to 80 years). Baseline symptoms are summarized in terms of pain, urgency, and frequency, and averaged two baseline scores provides an overall baseline score. There are 40% of subjects with severe urination pain and 63% with severe urination frequency. There are 61% of subjects who report at least 11 voids/day at both baseline visits. There were no statistically significant differences in the distribution of demographic characteristics or any of the symptom measures at baseline between the two treatment arms.

Overall, 80% of participants report at least one adverse event, which are classified as minor, moderate or severe (Small number (16) of subjects who described "very severe" adverse events are classified as severe). A summary of the highest grade of adverse events by treatment arm is shown in Table 3.2. Frequently reported symptoms of adverse events on both treatment arms include constitutional symptoms (primarily fatigue, malaise), dermatology/skin, pain (primary headache), gastrointestinal disturbances (primary try mouth, constipation), musculoskeletal symptoms, neurological symptoms (primarily dizziness, somnolence), renal/genitourinary symptoms, pulmonary, ocular, visual, and infection. Because of adverse events, although subjects are encouraged to take 75 mg from 6 to 12 weeks, dose is ultimately the patient's decision and they may opt to return to or to stay at 25 mg or less (18 in control group and 46 in treatment group) or 50 mg (16 in control group and 25 in treatment group). We summarize the dose tolerance in Table 3.3.

3.3 Method

We face several challenges in estimating the causal effect of Amitriptyline dose in the ICCRN study. First, subjects self-selected dose in both treatment group and control group. Although subjects in the ICCRN study were encouraged to take a dose

Table 3.1: Demographic characteristics for the ICCRN study (p-value gives test of null hypothesis of no difference in means or proportions between treatment and control arms)

	Assignment		
	Treatment	Control	Total
Number of Subjects	135	136	271
Gender ($p = 0.43$)			
Male	20 (15%)	25 (18%)	45 (17%)
Female	115 (85%)	111 (82%)	226 (83%)
Age ($p = 0.30$)			
Mean ± s.d.	38.0±13.8	39.9±14.0	39.0±13.9
Median (Range)	35.7 (19-67)	39.6 (18-80)	37.7 (18-80)
Race ($p = 0.60$)			
White/Caucasion	96 (72%)	104 (76%)	200 (74%)
Black/African-American	18 (13%)	15 (11%)	33 (12%)
Other	21 (15%)	17 (13%)	38 (24%)
Baseline Pain Severity Score ($p = 0.11$)			
Minor (1-3)	6 (4%)	11 (8%)	17 (6%)
Moderate(4-6)	80 (59%)	64 (47%)	144 (53%)
Severe(7-10)	49 (36%)	60 (44%)	109 (40%)
Baseline Urgency Severity Score ($p = 0.64$)			
None/Minor(0-3)	5 (4%)	7 (5%)	12 (4%)
Moderate(4-6)	51 (38%)	56 (41%)	107 (40%)
Severe(7-10)	79 (59%)	72 (53%)	151 (56%)
Baseline Frequency Severity Score ($p = 0.99$)			
Minor (1-3)	2 (1%)	4 (3%)	6 (2%)
Moderate (4-6)	48 (36%)	47 (35%)	95 (35%)
Severe (7-10)	85 (63%)	84 (62%)	169 (63%)
Patient-Reported 24-hour Frequency ($p = 0.79$)			
<6 times at first visit	7 (5%)	4 (3%)	11 (4%)
7-10 times at both visits	22 (17%)	24 (18%)	46 (17%)
7-10 times, 11+ times	25 (19%)	23 (17%)	48 (18%)
11-14 times at both visits	34 (26%)	39 (29%)	73 (27%)
11-14 times, 15+ times	17 (13%)	21 (16%)	38 (14%)
15+ times at both visits	30 (23%)	24 (18%)	54 (20%)

Table 3.2: Highest grade of adverse events for the ICCRN study

	Assignment		
Observed Adverse Event Grade	Treatment	Control	Total
None	14 (10%)	36 (26%)	50 (18%)
Minor	51 (38%)	37 (27%)	88 (32%)
Moderate	62 (46%)	55 (40%)	117 (43%)
Severe	8 (6%)	8 (6%)	16 (6%)

Table 3.3: Observed dose tolerance for the ICCRN study

	Assignment		
Observed Dose Tolerance	Treatment	Control	Total
$\leq$ 25 mg	46 (34%)	18 (13%)	64 (24%)
50 mg	25 (19%)	16 (12%)	41 (15%)
75 mg	64 (47%)	102 (75%)	166 (61%)

of 75 mg, dose taken was patients' choice in the end. In this kind of clinical setting, compliance is not measured as all or none; instead patients take part of the assigned dose, whether active or placebo, and so even if randomization is binary, treatment eventually received is not. In addition, subjects assigned to the control group took a masked placebo. Therefore compliance has a different meaning in the treatment group and control group. Compliance determines the amount of active drug taken for subjects assigned to treatment group and also indicates some level of subjects' psychosomatic status. In the control group, only the psychosomatic component of compliance applies. The second challenge is the large number of adverse events in the study. Adjusting for adverse events observed post-randomization via regression will destroy the causal interpretation of the treatment effect. Therefore, we construct a potential adverse event variable and model it in the same manner as a potential outcome. Third, the often assumed exclusion restriction (ER) assumption (no causal effect of treatment within the principal stratum which takes same value of potential mediator under different treatment arms) does not apply in the ICCRN study. This is because nearly all subjects on the treatment arm took treatment, but at varying doses; hence there is not a potential stratum in which subjects have the same treatment taken regardless of treatment assignment. Without the ER assumption, the lack of identification issue faced by most causal models may get worse. Finally, we have the fundamental problem of causal modeling that we cannot observe all the potential outcomes for any subject, and therefore large amounts of the "complete"

data are missing.

3.3.1 Notation

For subject $i, i = 1, ..., n$, let Z_i denote randomization assignment: $Z_i = 1$ for treatment group, and $Z_i = 0$ for placebo group. We consider three pairs of potential outcomes including clinical outcome, dose taken and adverse event experience.

- $Y_i(Z_i)$ is used to denote the clinical outcome under assignment Z_i: $Y_i(Z_i) = 1$ for responder, $Y_i(Z_i) = 0$ for non-responder.
- $A_i(Z_i)$ is used to denote the adverse event experience under assignment Z_i: $A_i \in (0, 1, 2)$ for none, minor, moderate/severe respectively. Severe adverse events are combined with moderate ones because there are only 16 severe adverse events.
- $D_i(Z_i)$ is used to denote the dose taken under assignment Z_i: $D_i \in (1, 2, 3)$ for ≤ 25 mg, 50 mg, and 75 mg respectively.

3.3.2 Model Assumptions

Key assumptions to assist in the identifiability and estimation of the model include the ignorable treatment assignment assumption, the stable unit treatment value assumption (SUTVA), the monotonicity assumption, and the ignorable missingness mechanism assumption.

1. Ignorable treatment assignment assumes that, conditional on the observed baseline variables, the treatment assignment is independent of all baseline variables and potential outcomes, $Z_i \perp Y_i(Z_i), D_i(Z_i), A_i(Z_i), \boldsymbol{X}_i, \forall i \in 1, ..., n$ [47]. For the ICCRN study, the ignorable treatment assignment assumption is satisfied because of the randomized treatment assignment.

2. SUTVA consists of two subassumptions. The first subassumption assumes there is no interference between the potential outcomes of different subjects, $Y_i(Z_i), D_i(Z_i), A_i(Z_i) \perp Y_j(Z_j), D_j(Z_j), A_j(Z_j), i \neq j$ [48]. Because the painful bladder syndrome is not an infectious disease and we do not anticipate that treatment assignment to subject i will impact the action of subject j in any other fashion, this subassumption is reasonable for the ICCRN study. The second subassumption assumes that there is only one version of the treatment [48] so that the observed $y_i = z_i * Y_i(Z_i = z_i) + (1 - z_i) * Y_i(Z_i = z_i)$. This subassumption is satisfied in the ICCRN study because the subjects in the study population were treated with the same form of therapy.

3. The monotonicity assumption assumes monotonicity of the potential outcomes. Specifically, for ICCRN study, we assume the monotonicity of the adverse events and monotonicity of dose taken, but not for the actual outcome of interest.

    - The ICCRN study used a masked placebo for subjects randomized to the control group. In the data analysis step, adverse events which are present at baseline are excluded and only new or worsened events are included. Due to this, it is reasonable to assume monotonicity for the adverse events, i.e. the highest grade of the adverse events under control arm is no higher than that under treatment arm: $A_i(0) \leq A_i(1)$.

    - In the ICCRN study, subjects returned to or stayed at lower doses because they underwent adverse events. Since we assume $A_i(0) \leq A_i(1)$, it is reasonable to assume that the dose tolerance under treatment arm is not higher than that under control arm: $D_i(1) \leq D_i(0)$.

4. In the ICCRN study, whether the potential outcome is observed or missing

depends on thetreatment arm to which the subject is assigned. With randomized assignment, the missingness of the outcomes becomes random, thus the missingness mechanism of these missing data is missing complete at random (MCAR) [36]. Besides the missing counterfactual outcomes, there are subjects who dropped out from the study during the 12 weeks follow up (13% on control arm and 18% on treatment arm). Most of the withdrawals have the primary reason as being "lost to follow up", and the withdrawal rates and reasons provided are similar for each treatment arm. In the analysis, we assume missing at random (MAR) [36] mechanism for these missing data.

3.3.3 Complete Data Likelihood

The "complete" data under the potential outcome framework include ($A_i(0)$, $A_i(1)$, $D_i(0)$, $D_i(1)$, $Y_i(0)$, $Y_i(1)$, $i = 1, ..., n$). We summarize the underlying probabilities associated the "complete" data in Table 3.19 in the Appendix. The complete

data likelihood can be factored it into a series of conditional likelihoods:

$$
\begin{aligned}
L =& \prod_{i=1}^{n} f(Y_i(0), Y_i(1), D_i(0), D_i(1), A_i(0), A_i(1) | \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \\
=& \prod_{i=1}^{n} [f(Y_i(0), Y_i(1) | D_i(0), D_i(1), A_i(0), A_i(1), \boldsymbol{\gamma}) \\
& \times f(D_i(1) | D_i(0), A_i(0), A_i(1), \boldsymbol{\theta}) \\
& \times f(D_i(0) | A_i(0), A_i(1), \boldsymbol{\beta}) \\
& \times f(A_i(0), A_i(1) | \boldsymbol{\alpha})] \\
=& \prod_{i=1}^{n} [f(Y_i(0), Y_i(1) | D_i(0), D_i(1), \boldsymbol{\gamma}) \\
& \times f(D_i(1) | D_i(0), A_i(1), \boldsymbol{\theta}) \\
& \times f(D_i(0) | A_i(0), \boldsymbol{\beta}) \\
& \times f(A_i(0), A_i(1) | \boldsymbol{\alpha})]
\end{aligned}
$$

where $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ parameterize the conditional likelihoods of clinical outcome $Y_i(0), Y_i(1)$, dose taken under treatment $D_i(1)$, dose taken under control $D_i(0)$, and the highest grade of adverse events $A_i(0), A_i(1)$. The first equality follows from the SUTVA. The second equality factors the "complete" data into a series of conditional likelihoods, and the third equality follows from several model specification assumptions:

- Given the dose taken $(D_i(1), D_i(0))$, the clinical outcome $(Y_i(1), Y_i(0))$ is independent of the adverse event $(A_i(1), A_i(0))$.
- The dose tolerance under treatment $(D_i(1))$ depends only on the dose tolerance under control $(D_i(0))$ and the adverse event experience under treatment $(A_i(1))$.
- The dose tolerance under control $(D_i(0))$ depends only on the adverse event under control $(A_i(0))$.

After choosing suitable parametric models for each conditional likelihood, these model specification assumptions provide reasonable constraints to assist in identifiability. Overall we have 34 model parameters (5 in conditional likelihood of adverse event severity, 4 in conditional likelihood of dose tolerance under control, 7 in conditional likelihood of dose tolerance under treatment, and 18 in conditional likelihood of clinical outcome), while the observed data of ICCRN study provide 35 (36-1=35) sufficient statistics indicated by Table 3.20. The fact that the number of parameters in our model less than the sufficient statistics provided by the observed data makes the proposed model unsaturated. The model specifications for each conditional likelihood are as follows:

1. $f(Y_i(0), Y_i(1)|D_i(0), D_i(1), \boldsymbol{\gamma})$: Baseline multinomial logit model.

   Letting $p_{i,y_0,y_1} = P(Y_i(0) = y_0, Y_i(1) = y_1 |\ D_i(0), D_i(1), \boldsymbol{\gamma})$ and $\ln(\frac{p_{i,y_0,y_1}}{p_{i,0,0}}) = \gamma_{y_0,y_1} + \gamma_{d_0,d_1}^{y_0,y_1} I(D_i(0) = d_0, D_i(1) = d_1)$

2. $f(D_i(1)|D_i(0), A_i(1), \boldsymbol{\theta})$: Multinomial probit model.

   Let $p_{ij}^{d,a} = P(D_i(1) = j|D_i(0) = d, A_i(1) = a)$ and $\pi_{ij}^{d,a} = \sum_{k=1}^{j} p_{ij}^{d,a}$, with $\pi_{i3}^{d,a} \equiv 1$. Under monotonicity, we have

$$
\begin{aligned}
P(D_i(1)|D_i(0) = 1, A_i(1) = a) \quad & \pi_{i1}^{1,a} \equiv 1 \\
P(D_i(1)|D_i(0) = 2, A_i(1) = a) \quad & \pi_{i1}^{2,a} = \Phi\left(\theta_{12}^{D} + I(A_i(1) = a)\theta_{a2}^{A}\right) \\
& \pi_{i2}^{2,a} \equiv 1 \\
P(D_i(1)|D_i(0) = 3, A_i(1) = a) \quad & \pi_{i1}^{3,a} = \Phi\left(\theta_{13}^{D} + I(A_i(1) = a)\theta_{a3}^{A}\right) \\
& \pi_{i2}^{3,a} \equiv \Phi\left(\theta_{23}^{D} + I(A_i(1) = a)\theta_{a3}^{A}\right)
\end{aligned}
$$

3. $f(D_i(0)|A_i(0), \boldsymbol{\beta})$: Multinomial probit model.

   Let $p_{ij}^{a} = P(D_i(0) = j|A_i(0) = a)$ and $\pi_{ij}^{a} = \sum_{k=1}^{j} p_{ij}^{a}$. Then $\pi_{ij}^{a} = \Phi(\beta_j^D + I(A_i(0) = a)\beta_a^A)$, $j = 1, 2$, $\pi_{i3}^{a} \equiv 1$.

4. $f(A_i(0), A_i(1)|\boldsymbol{\alpha})$: Baseline multinomial logit model.

   Letting $p_{i,a_0,a_1} = P(A_i(0) = a_0, A_i(1) = a_1|\boldsymbol{\alpha})$, $\ln(\frac{p_{i,a_0,a_1}}{p_{i,0,0}}) = \alpha_{a_0,a_1}$

3.3.4 Estimates of Interest – Principal Effects

Our goal is to estimate the causal effect of treatment within the principal strata. For the ICCRN study, the principal strata consist of the subgroups of population with a given level of treatment dose tolerance ($\leq$ 25mg, 50 mg, or 75 mg). The outcome measure, global response assessment, is whether the subject is a responder or non-responder, therefore the responder rate difference becomes a natural estimate:

$$P(Y(1) = 1|D(1) = d_1) - P(Y(0) = 1|D(1) = d_1); d_1 \in (1, 2, 3)$$

This responder rate difference is estimated within the subgroup of the population with treatment dose tolerance of $d_1 = 1$ ($\leq$ 25 mg), $d_1 = 2$ (50 mg), or $d_1 = 3$ (75 mg). Note that the principal stratum $D_i(1)$ is not observed for subjects assigned to control group, thus we impute these subjects' principal strata membership, and use adverse events to help identify their principal strata membership.
Note $P(Y_i(1) = 1|D_i(1) = d_1) - P(Y_i(0) = 1|D_i(1) = d_1) = \frac{P(Y_i(1)=1, D_i(1)=d_1) - P(Y_i(0)=1, D_i(1)=d_1)}{P(D_i(1)=d_1)}$, where

$$
\begin{aligned}
& P(Y_i(z_i) = 1, D_i(1) = d_1) \\
&= \sum_{y_0=0}^{1} \sum_{d_0=1}^{3} \sum_{a_0=0}^{2} \sum_{a_1=0}^{2} P(Y_i(z_i) = 1, Y_i(1 - z_i) = y_0|D_i(1) = d_1, D_i(0) = d_0, \boldsymbol{\gamma}) \times \\
& P(D_i(1) = d_1|D_i(0) = d_0, A_i(1) = a_1, \boldsymbol{\theta}) \times P(D_i(0) = d_0|A_i(0) = a_0, \boldsymbol{\beta}) \times \\
& P(A_i(1) = a_1, A_i(0) = a_0|\boldsymbol{\alpha}); \\
& P(D_i(1) = d_1) \\
&= \sum_{d_0=1}^{3} \sum_{a_0=0}^{2} \sum_{a_1=0}^{2} P(D_i(1) = d_1|D_i(0) = d_0, A_i(1) = a_1, \boldsymbol{\theta}) \times \\
& P(D_i(0) = d_0|A_i(0) = a_0, \boldsymbol{\beta}) \times P(A_i(1) = a_1, A_i(0) = a_0|\boldsymbol{\alpha})
\end{aligned}
$$

3.3.5 Model Estimation

The observed data include Z_i, $Y_i(Z_i)$, $D_i(Z_i)$, $A_i(Z_i), i \in 1, ..., n$; and the probabilities associated with observed data are shown in Table 3.20 in the Appendix. To accommodate the complex missing data structure, we use a Bayesian scheme with a Markov chain Monte Carlo (MCMC) algorithm. Each iteration of the MCMC algorithm consists of two steps of subiterations. In the first step, we obtain a draw of each parameter from the posterior distribution conditional on the complete data. In the second step, we impute the missing potential outcomes conditional on the observed data and the updated parameters from the first step. We repeat the above two steps until all parameters converge in distribution. For the first step, the posterior distribution of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are not of closed form of a known parametric distribution, therefore we implement Metropolis-Hastings within Gibbs algorithm to obtain random draws from their posterior distributions. To obtain the draws of parameter $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from their posterior distribution, we utilize the Bayesian computation method proposed by Albert and Chib and add a data augmentation step [1]. With the added data augmentation step, the posterior distributions of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ become known distributions and parameters converge with fewer iterations. The priors of parameters are assumed as $\pi(\boldsymbol{\alpha}) \sim N(0, 10I)$, $\pi(\boldsymbol{\beta}) \sim N(0, 10I)$, $\pi(\boldsymbol{\theta}) \sim N(0, 10I)$, $\pi(\boldsymbol{\gamma}) \sim N(0, 10I)$. These priors are very weakly informative, but bound the model probabilities away from values extremely close to 0 or 1. In the second step, the missing outcomes conditional on the observed data and estimated parameters all follow binomial distribution or multinomial distribution. We include the details of the posterior distribution of parameters and conditional distributions of missing potential

Table 3.4: Per-protocol analysis for the ICCRN study

		Observed Dose Tolerance		
		$\leq$ 25 mg	50 mg	75 mg
Response in Treatment Group	Mean	0.577	0.720	0.683
	95% Confidence Interval	(0.387, 0.767)	(0.544, 0.896)	(0.566, 0.801)
	n	26	25	60
Response in Control Group	Mean	0.545	0.417	0.521
	95% Confidence Interval	(0.251, 0.840)	(0.138, 0.696)	(0.421, 0.621)
	n	11	12	96
Effect of Treatment	Mean	0.031	0.303	0.163
	95% Confidence Interval	(-0.319, 0.382)	(-0.026, 0.633)	(-0.008, 0.317)

outcomes in the Appendix.

3.4 Application to the ICCRN Study

3.4.1 Intent-to-Treat Analysis and Per-Protocol Analysis

We conduct a standard intent-to-treat analysis to the study sample. Subjects who withdrew from the study for any reason, and did not provide data on the primary endpoint, are excluded from the analysis. The global IC response rate is 0.667 on treatment and 0.513 on control ($p = 0.025$). We also conduct per-protocol analysis to the study sample and summarize the analysis results in Table 3.4. When classifying participants into subgroups by the maximum tolerance dose as low dose ($\leq$ 25 mg), medium dose (50 mg), and high dose (75 mg) based on their observed dose taken (treatment or control), the response rate is 0.577 on treatment and 0.545 on control ($p = 0.852$) for the low dose subgroup, 0.720 on treatment and 0.417 on control ($p = 0.156$) for the medium dose subgroup, and 0.683 on treatment and 0.521 on control ($p = 0.066$) for the high dose subgroup.

Assuming MAR for the missing outcomes of the subjects who dropped out from the study, we find that the treatment assignment has a marginally significant effect in reducing IC symptoms compared with placebo ($p = 0.06$) [36]. Note that the

Table 3.5: Subjects with missing outcomes for the ICCRN study

	Outcome Observed		Outcome Missing	
Outcome	Treatment	Control	Treatment	Control
Responder	74	61		
Non-Responder	37	58		
Total	111	119	24	17

Table 3.6: Maximum likelihood estimation for the ICCRN study

		Treatment	Control
Outcome	Responder (S.E)	0.332 (0.030)	0.257 (0.024)
	Non-Responder (S.E.)	0.166 (0.028)	0.245 (0.027)

MAR and MCAR assumptions correspond for the dose-stratified analyses.

3.4.2 Causal Model Analysis

We apply our proposed causal model to the ICCRN study, and use independent $N(0, 10)$ priors for the $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\gamma}$ of the multinomial logit and probit model parameters. To check the convergence of parameters, we run multiple chains with different starting points, and calculate $\hat{R}$ [19]. In general, $\hat{R} < 1.1$ indicates acceptable convergence. We run 4 chains, each with 400,000 iterations after 100,000 burn in but with different initial values of the parameters. The maximum value of $\hat{R}$ of all the parameters is 1.02, indicating the parameters converge in distribution.

Table 3.7 and Table 3.8 show the relationship of dose taken with adverse event experience. There was no evidence of a relationship between tolerated control dose and adverse event experience in the control arm. On the treatment arm, θ^A_{13} is estimated negative (-1.97, 95% CI (-6.88, 0.83)) indicating subjects who experienced (or would have experienced) minor adverse events under treatment were more likely to tolerate higher treatment doses, and θ^A_{23} is estimated positive (0.67, 95% CI (-0.47, 2.79)) indicating subjects who experienced (or would have experienced) moderate or severe adverse events under treatment were less likely to tolerate higher treatment

doses, although their 95% C.I. cover zero.

The key results of interest – the principal effects, along with the estimated response under treatment arm and control arm and principal strata membership – are summarized in Table 3.9. The analysis results show that the probability of having a global IC response is 0.564, 0.704 and 0.667 in the treatment arm for subjects with the treatment dose tolerances of $\leq$ 25 mg, 50 mg and 75 mg respectively, and 0.467, 0.486 and 0.559 for subjects in the control arm with the *treatment* dose tolerances of $\leq$ 25 mg, 50 mg and 75 mg respectively. Compared with response under treatment arm, the response under control arm has a wider credible interval, mainly due to the fact that the response is not observed and needs to be imputed in our proposed causal model. Comparing Table 3.9 with Table 3.4, we see that, while the results are broadly similar, the treatment effect in the higher treatment dose tolerances strata is more muted than in the per-protocol analysis, since the global response rate is higher when subjects on the control arm are assigned their estimated treatment dose tolerance, rather than their observed control dose tolerance. In neither the per-protocol nor the principal stratum analysis are the dose effects monotonic.

To better understand the differences between the AT and the causal principal stratification analyses, we show the joint distribution of the treatment and control dose tolerances in Table 3.10, and the global response rates by treatment and control dose tolerances in Table 3.11. Among treated subjects who can tolerate 75 mg of the treatment dose, the principal stratification model estimates the global response rate as 0.667. This corresponds well with the response rate of 0.683 among the treated subjects under 75 mg, since the principal strata of 75 mg of treatment dose is observed in the treatment arm (we might expect a slight discrepancy since the principal stratification estimate is based on a non-saturated model that predicts

Table 3.7: Parameters in the conditional likelihood of control dose tolerance for the ICCRN study

	β_d^D		β_a^A	
	β_2^D	β_3^D	β_1^A	β_2^A
Mean	-1.07	-0.63	-0.01	0.06
95% C.I.	(-1.55, -0.63)	(-1.07, -0.23)	(-0.62, 0.61)	(-0.45, 0.59)

Table 3.8: Parameters in the conditional likelihood of treatment dose tolerance for the ICCRN study

	$\theta_{d_1 d_0}^D$				
	θ_{12}^D	θ_{13}^D	θ_{23}^D		
Mean	0.84	0.37	-1.51		
95% C.I.	(0.27, 2.28)	(-4.89, 4.99)	(-4.13, -0.21)		
	θ_{a2}^A			θ_{a3}^A	
	θ_{12}^A	θ_{22}^A		θ_{13}^A	θ_{23}^A
Mean	0.11	-0.94		-1.97	0.67
95% C.I.	(-5.57, 6.11)	(-6.61, 5.33)		(-6.88, 0.83)	(-0.47, 2.79)

response as a function of dose tolerance and adverse events, rather than as a direct estimate from the observed cell). However, the principal stratification approach compares this response rate with the global response rate among *control* subjects who can tolerate 75 mg of the *treatment* dose, estimated at 0.559, for a principal effect of 0.107; whereas the AT approach compares this response rate with the global response rate among *control* subjects who can tolerate 75 mg of the *control* dose, which is a mixture of control subject global response rates of the cells in the last row of Table 3.11 corresponding to all subjects who could tolerate the highest level of the control dose, at the 0.105 : 0.127 : 0.491 mixing proportions from the last row of Table 3.10. The control global response rates for the subjects who could tolerate 75 mg of control but only 25 or 50 mg of treatment are lower than the global response rates for those who could tolerate 75 mg of treatment, thereby increasing the treatment effect in the AT analysis as compared with the principal stratification analysis.

Table 3.9: Response rate, principal effect and principal strata membership for the ICCRN study

		Treatment Dose Tolerance		
		$\leq$ 25 mg	50 mg	75 mg
Response under Treatment arm	Mean	0.564	0.704	0.667
	Median	0.564	0.708	0.668
	95% Credible Interval	(0.390, 0.736)	(0.525, 0.856)	(0.546, 0.780)
Response under Control arm	Mean	0.467	0.486	0.559
	Median	0.467	0.501	0.569
	95% Credible Interval	(0.212, 0.727)	(0.061, 0.888)	(0.344, 0.734)
Principal Effect	Mean	0.097	0.218	0.107
	Median	0.092	0.183	0.090
	95% Credible Interval	(-0.226, 0.413)	(-0.200, 0.690)	(-0.061, 0.354)
Principal Strata Membership	Mean	0.318	0.191	0.491
	Median	0.317	0.189	0.492
	95% Credible Interval	(0.245, 0.399)	(0.131, 0.258)	(0.408, 0.571)

Table 3.10: Posterior mean of $P(D(1) = d_1, D(0) = d_0)$ for the ICCRN study

	$D_i(1)$		
$D_i(0)$	1	2	3
1	0.153		
2	0.060	0.063	
3	0.105	0.127	0.491

Table 3.11: Posterior mean of $P(Y_i(1) = 1|D_i(1) = d_1, D_i(0) = d_0)$ and $P(Y_i(0) = 1|D_i(1) = d_1, D_i(0) = d_0)$ for the ICCRN study

	$D_i(1) = 1$		$D_i(1) = 2$		$D_i(1) = 3$	
$D_i(0)$	$P(Y_i(1) = 1)$	$P(Y_i(0) = 1)$	$P(Y_i(1) = 1)$	$P(Y_i(0) = 1)$	$P(Y_i(1) = 1)$	$P(Y_i(0) = 1)$
1	0.647	0.534				
2	0.734	0.342	0.529	0.504		
3	0.365	0.455	0.793	0.483	0.667	0.559

3.4.3 Model Fit Assessment

To assess the fit of the data, we calculate the posterior predictive distribution (PPD) p-values [19]. Because the adverse event severity $A_i(Z_i)$, dose taken $D_i(Z_i)$, and clinical outcome $Y_i(Z_i)$ in subjects randomized to the group Z_i are observed, we compare their PPD with the observed ones to assess the fit to the data. Let G_e denote the number of subjects having the eth grade of adverse event severity and let κ_e be the estimated probability of having the eth grade of adverse event severity. Let G_d denote the number of subjects having the dth level of dose taken, and let κ_d be the estimated probability of having the dth level of dose taken. Similarly, let G_y denote the number of subjects having the yth of level of clinical outcome, and let κ_y be the estimated probability of having the yth level of clinical outcome. We consider the χ^2-type statistics

$$S_e^{\text{obs}} = \sum_e \frac{(G_e^{\text{obs}} - N\kappa_e)^2}{N\kappa_e(1-\kappa_e)}; \text{ and } S_e^{\text{rep}} = \sum_e \frac{(G_e^{\text{rep}} - N\kappa_e)^2}{N\kappa_e(1-\kappa_e)}$$
$$S_d^{\text{obs}} = \sum_d \frac{(G_d^{\text{obs}} - N\kappa_d)^2}{N\kappa_d(1-\kappa_d)}; \text{ and } S_d^{\text{rep}} = \sum_d \frac{(G_d^{\text{rep}} - N\kappa_d)^2}{N\kappa_d(1-\kappa_d)}$$
$$S_y^{\text{obs}} = \sum_y \frac{(G_y^{\text{obs}} - N\kappa_y)^2}{N\kappa_y(1-\kappa_y)}; \text{ and } S_y^{\text{rep}} = \sum_y \frac{(G_y^{\text{rep}} - N\kappa_y)^2}{N\kappa_y(1-\kappa_y)}$$

where G_e^{obs}, G_d^{obs}, and G_y^{obs} are the observed statistics and G_e^{rep}, G_d^{obs}, and G_y^{obs} are the repeated statistics obtained from draws of the parameters from MCMC. The PPD p value is given by

$$\frac{\sum_l I(S_l^{\text{obs}} < S_l^{\text{rep}})}{\sum_l 1}$$

A PPD p value close to .50 indicates good fit of the model the the data [19]. We summarized the PPD p values in Table 3.12. The results show that PPD p values range from 0.45 to 0.73, indicating a reasonably good fit of our causal model to the data.

Table 3.12: PPD p values of the proposed causal model for the ICCRN study

	Treatment Arm	
	Treatment	Control
Adverse Event	0.57	0.68
Dose Tolerance	0.45	0.73
Outcome Response	0.61	0.61

Table 3.13: Distribution of control dose tolerance given control adverse events for simulation studies

	$D(0)=1$	$D(0)=2$	$D(0)=3$
$A(0)=0$	.16	.12	.72
$A(0)=1$	.16	.12	.72
$A(0)=2$	.16	.12	.72

3.5 Simulation Study

3.5.1 Data Simulation

Although we utilize a Bayesian approach, we are still interested in the repeated sampling properties of the proposed model. Given the relatively long computation time of analysis for each simulated data set, we do moderate number of simulations. We simulate 200 data sets, with 250 subjects in each data set. Data are simulated using the model in Section 3.3, under the monotonicity assumptions for the adverse events and dose tolerances. Table 3.13 shows the distribution of the control dose tolerance given control adverse events, Table 3.14 shows the distribution of the treatment dose tolerance given control dose tolerance and treatment adverse events, and Table 3.15 shows the distribution of clinical outcome by treatment arm given treatment dose tolerance. Similar to the ICCRN study, there is no relationship between control dose tolerance and control adverse event experience, whereas high levels of treatment adverse event experience are associated with reduced treatment dose tolerance. The simulation study parameters are set to yield higher clinical response rates under treatment than control, and a modest monotonically increasing treatment effect as dose level increases.

Table 3.14: Distribution of treatment dose tolerance given control dose tolerance and treatment adverse events for simulation studies with the proposed causal model

	$D(1)=1$	$D(1)=2$	$D(1)=3$
$D(0)=1$	1	0	0
$D(0)=2, A(1)=0$	.71	.29	0
$D(0)=2, A(1)=1$	.65	.35	0
$D(0)=2, A(1)=2$	.86	.14	0
$D(0)=3, A(1)=0$	.05	.20	.74
$D(0)=3, A(1)=1$	.02	.13	.85
$D(0)=3, A(1)=2$	.09	.26	.66

Table 3.15: Distribution of clinical outcome by treatment arm given treatment dose tolerance for simulation studies with the proposed causal model

	$P(Y(1)=1)$	$P(Y(0)=1)$	Treatment Effect
$D(1)=1$	.22	.16	.06
$D(1)=2$	.31	.23	.08
$D(1)=3$	.41	.31	.10

3.5.2 Analysis Results

A summary of the estimated response under each treatment arm and the associated principal effect within each principal stratum from 200 simulated data sets is shown in Table 3.16, including point estimates from the posterior mean and median, as well as the estimated repeated sampling coverage of the 95% credible interval. The true coverage probability for the 95% credible intervals associated with the response under each treatment arm and the treatment effect within each principal stratum is somewhat conservative, ranging from 93.5% to 99.5%. The posterior means of the response under both treatment and control arms within principal strata and consequently the principlal effects within the principal strata are very close to the true values, with the modest exception of the control arm response under treatment dose tolerances of less than 25 mg. Since principal stratum membership for subjects under control is latent, there is more uncertainty in the control arm response stratified by principal strata, leading to larger mean square errors and posteriors that are flatter and more difficult to summarize with a single point estimate such as a posterior

Table 3.16: Response rate, principal effect and principal strata membership with 200 simulations

		Treatment Dose Tolerance		
		$\leq$ 25 mg	50 mg	75 mg
Response under Treatment Arm	95% Coverage Probability	93.5%	95%	94.5%
	Mean (True Value)	0.239 (0.217)	0.317 (0.312)	0.406 (0.410)
	Median	0.234	0.312	0.405
	MSE	9.02×10^{-3}	1.72×10^{-2}	6.74×10^{-3}
Response under control Arm	95% Coverage Probability	98.5%	99.5%	98.5%
	Mean (True Value)	0.196 (0.157)	0.228 (0.233)	0.296 (0.307)
	Median	0.180	0.196	0.300
	MSE	1.40×10^{-2}	3.10×10^{-2}	1.13×10^{-2}
Principal Effect	95% Coverage Probability	99%	99.5%	97%
	Mean (True Value)	0.044 (0.060)	0.089 (0.078)	0.111 (0.103)
	Median	0.049	0.010	0.099
	MSE	1.86×10^{-2}	4.54×10^{-2}	1.74×10^{-2}
Principal Strata Membership	95% Coverage Probability	92%	96.5%	94%
	Mean (True Value)	0.279 (0.297)	0.172 (0.172)	0.543 (0.531)
	Median	0.278	0.171	0.544
	MSE	3.13×10^{-3}	1.85×10^{-3}	3.98×10^{-3}

mean. This can be seen by comparing, for a single simulation, the posterior distributions of the response under treatment in Figure 3.1 with the response under control in Figure 3.2 in the Appendix.

We tested the prior variance sensitivity by enlarging the variance of priors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ from 10 to 100. The results are very close to those with priors variance of 10, as is shown in Table 3.17.

We apply the traditional per-protocol analysis on the simulated data and summarize the results in Table 3.18. The estimated responses in the treatment group are close to the true values. However, the results show that the estimated response in the control group could be far off the true values, which could lead to the biased estimate of the effect of the treatment. Even though we simulate the data with monotonic true effects of treatment, the estimated effects are not monotonic. It results from the fact subjects with the observed control dose tolerance of 50 mg are a mix of subjects who can tolerate 50 mg and 25 mg on treatment, and this mix leads to the underestimation of the control response rate among those who can tolerate 50

Table 3.17: Response rate, principal effect and principal strata membership with 200 simulations with enlarged variance

		Treatment Dose Tolerance		
		$\leq$ 25 mg	50 mg	75 mg
Response under Treatment Arm	95% Coverage Probability	93.5%	95%	91.5%
	Mean (True Value)	0.224 (0.217)	0.308 (0.312)	0.391 (0.410)
	Median	0.219	0.301	0.389
	MSE	8.43×10^{-3}	1.79×10^{-2}	7.18×10^{-3}
Response under control Arm	95% Coverage Probability	97.5%	99.5%	96%
	Mean (True Value)	0.183 (0.157)	0.214 (0.233)	0.313 (0.307)
	Median	0.178	0.199	0.319
	MSE	1.21×10^{-2}	3.12×10^{-2}	9.95×10^{-3}
Principal Effect	95% Coverage Probability	96.5%	99%	94.5%
	Mean (True Value)	0.046 (0.060)	0.104 (0.078)	0.085 (0.103)
	Median	0.044	0.097	0.086
	MSE	1.56×10^{-2}	4.23×10^{-2}	1.62×10^{-2}
Principal Strata Membership	95% Coverage Probability	92%	95.5%	93.5%
	Mean (True Value)	0.277 (0.297)	0.170 (0.172)	0.542 (0.531)
	Median	0.279	0.165	0.554
	MSE	3.83×10^{-3}	1.76×10^{-3}	4.75×10^{-3}

Table 3.18: Per-protocol analysis with 200 simulations

		Observed Dose Tolerance		
		$\leq$ 25 mg	50 mg	75 mg
Response in Treatment Group	Mean (True Value)	0.227 (0.217)	0.316 (0.312)	0.417 (0.410)
	S.D.	0.068	0.100	0.061
Response in Control Group	Mean (True Value)	0.163 (0.157)	0.143 (0.233)	0.284 (0.307)
	S.D.	0.079	0.079	0.047
Effect of Treatment	Mean (True Value)	0.064 (0.060)	0.173 (0.078)	0.133 (0.103)
	S.D.	0.106	0.132	0.077

mg on treatment, due to the fact that subjects with dose tolerance of 25 mg under treatment have a lower control response rate.

3.6 Discussion and Conclusion

Subjects are sometimes put on escalating dosages in phase II clinical trials. However, subjects often choose to go back or stay at lower dosages, typically as a result of adverse events. When this occurs, estimating the effect of treatment conditional on the dosage tolerance is not straightforward. The proposed causal model constructs latent prerandomization principal strata based on the maximum tolerance dose under treatment to provide a valid estimate of the causal effect of treatment within each

of the principal strata. The adverse events are included and modeled as potential outcomes to help identify the subjects' principal strata membership. The ICCRN study analysis results show that the Amitriptyline does not significantly reduce interstitial cystitis symptoms when compared with placebo for subjects who can tolerate treatment of $\leq$ 25 mg, 50 mg, or 75 mg, although there is some weak evidence for more effective treatment for subjects with higher dose tolerance.

Our work of course has limitations and thus possible extensions. To deal with drop outs, we assume a simple MAR missingness mechanism. Potential extensions include allowing for non-ignorable missingness, such as latent ignorability which assumes ignorability only within principal strata [16, 42], although their practical effect in this application is likely to be modest, given the relatively small proportion of subjects who dropped out. A thorough discussion of a variety of missingness mechanisms in the principal stratification setting and their associated identification constraints are provided by Small and Cheng [53]. Identification in our setting, with our need to consider counterfactuals of adverse events, is also an issue, and is achieved here by making a variety of assumptions which, though reasonable and not countermanded by any of the observed data, may not be correct. We could extend this to causal inference settings when full identification is not possible, either by focusing on bounds of causal effects [10], or using Bayesian methods that do not require identified likelihoods if proper priors are utilized [14]. Adapting these approaches to the self-selection dosage setting remains future work.

Appendix

Relating Complete Data to Observed Data

Table 3.19: Cell probabilities associated with the complete data

		$Y_i(0), Y_i(1)$			
$A_i(0), A_i(1)$	$D_i(0), D_i(1)$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(1,1)	ω_{111}	ω_{112}	ω_{113}	ω_{114}
(0,0)	(2,1)	ω_{121}	ω_{122}	ω_{123}	ω_{124}
(0,0)	(2,2)	ω_{131}	ω_{132}	ω_{133}	ω_{134}
(0,0)	(3,1)	ω_{141}	ω_{142}	ω_{143}	ω_{144}
(0,0)	(3,2)	ω_{151}	ω_{152}	ω_{153}	ω_{154}
(0,0)	(3,3)	ω_{161}	ω_{162}	ω_{163}	ω_{164}
(0,1)	(1,1)	ω_{211}	ω_{212}	ω_{213}	ω_{214}
(0,1)	(2,1)	ω_{221}	ω_{222}	ω_{223}	ω_{224}
(0,1)	(2,2)	ω_{231}	ω_{232}	ω_{233}	ω_{234}
(0,1)	(3,1)	ω_{241}	ω_{242}	ω_{243}	ω_{244}
(0,1)	(3,2)	ω_{251}	ω_{252}	ω_{253}	ω_{254}
(0,1)	(3,3)	ω_{261}	ω_{262}	ω_{263}	ω_{264}
(0,2)	(1,1)	ω_{311}	ω_{312}	ω_{313}	ω_{314}
(0,2)	(2,1)	ω_{321}	ω_{322}	ω_{323}	ω_{324}
(0,2)	(2,2)	ω_{331}	ω_{332}	ω_{333}	ω_{334}
(0,2)	(3,1)	ω_{341}	ω_{342}	ω_{343}	ω_{344}
(0,2)	(3,2)	ω_{351}	ω_{352}	ω_{353}	ω_{354}
(0,2)	(3,3)	ω_{361}	ω_{362}	ω_{363}	ω_{364}
(1,1)	(1,1)	ω_{411}	ω_{412}	ω_{413}	ω_{414}
(1,1)	(2,1)	ω_{421}	ω_{422}	ω_{423}	ω_{424}
(1,1)	(2,2)	ω_{431}	ω_{432}	ω_{433}	ω_{434}
(1,1)	(3,1)	ω_{441}	ω_{442}	ω_{443}	ω_{444}
(1,1)	(2,2)	ω_{451}	ω_{452}	ω_{453}	ω_{454}
(1,1)	(3,3)	ω_{461}	ω_{462}	ω_{463}	ω_{464}
(1,2)	(1,1)	ω_{511}	ω_{512}	ω_{513}	ω_{514}
(1,2)	(2,1)	ω_{521}	ω_{522}	ω_{523}	ω_{524}
(1,2)	(2,1)	ω_{531}	ω_{532}	ω_{533}	ω_{534}
(1,2)	(3,1)	ω_{541}	ω_{542}	ω_{543}	ω_{544}
(1,2)	(3,2)	ω_{551}	ω_{552}	ω_{553}	ω_{554}
(1,2)	(3,3)	ω_{561}	ω_{562}	ω_{563}	ω_{564}
(2,2)	(1,1)	ω_{611}	ω_{612}	ω_{613}	ω_{614}
(2,2)	(2,1)	ω_{621}	ω_{622}	ω_{623}	ω_{624}
(2,2)	(2,1)	ω_{631}	ω_{632}	ω_{633}	ω_{634}
(2,2)	(3,1)	ω_{641}	ω_{642}	ω_{643}	ω_{644}
(2,2)	(3,2)	ω_{651}	ω_{652}	ω_{653}	ω_{654}
(2,2)	(3,3)	ω_{661}	ω_{662}	ω_{663}	ω_{664}

Table 3.20: Cell probabilities associated with the observed data

			$Y_i(Z_i)$	
Z_i	$A_i(Z_i)$	$D_i(Z_i)$	0	1
0	0	1	υ_{0010}	υ_{0011}
0	0	2	υ_{0020}	υ_{0021}
0	0	3	υ_{0030}	υ_{0031}
0	1	1	υ_{0110}	υ_{0111}
0	1	2	υ_{0120}	υ_{0121}
0	1	3	υ_{0130}	υ_{0131}
0	2	1	υ_{0210}	υ_{0211}
0	2	2	υ_{0220}	υ_{0221}
0	2	3	υ_{0230}	υ_{0231}
1	0	1	υ_{1010}	υ_{1011}
1	0	2	υ_{1020}	υ_{1021}
1	0	3	υ_{1030}	υ_{1031}
1	1	1	υ_{1110}	υ_{1111}
1	1	2	υ_{1120}	υ_{1121}
1	1	3	υ_{1130}	υ_{1131}
1	2	1	υ_{1210}	υ_{1211}
1	2	2	υ_{1220}	υ_{1221}
1	2	3	υ_{1230}	υ_{1231}

The probabilities associated with complete data (Table 3.19) and the probabilities associated with the observed data (Table 3.20) can be related as follows:

$$\upsilon_{0010} = \omega_{111} + \omega_{112} + \omega_{211} + \omega_{212} + \omega_{311} + \omega_{312}$$

$$\upsilon_{0011} = \omega_{113} + \omega_{114} + \omega_{213} + \omega_{214} + \omega_{313} + \omega_{314}$$

$$\upsilon_{0020} = \omega_{121} + \omega_{122} + \omega_{131} + \omega_{132} + \omega_{221} + \omega_{222} + \omega_{231} + \omega_{232} + \omega_{321} + \omega_{322} + \omega_{331} + \omega_{332}$$

$$\upsilon_{0021} = \omega_{123} + \omega_{124} + \omega_{133} + \omega_{134} + \omega_{223} + \omega_{224} + \omega_{233} + \omega_{234} + \omega_{323} + \omega_{324} + \omega_{333} + \omega_{334}$$

$$\upsilon_{0030} = \omega_{141} + \omega_{142} + \omega_{151} + \omega_{152} + \omega_{161} + \omega_{162} + \omega_{241} + \omega_{242} + \omega_{251} + \omega_{252} + \omega_{261} + \omega_{262} + \omega_{341} + \omega_{342} + \omega_{351} + \omega_{352} + \omega_{361} + \omega_{362}$$

$$\upsilon_{0031} = \omega_{143} + \omega_{144} + \omega_{153} + \omega_{154} + \omega_{163} + \omega_{164} + \omega_{243} + \omega_{244} + \omega_{253} + \omega_{254} + \omega_{263} + \omega_{264} + \omega_{343} + \omega_{344} + \omega_{353} + \omega_{354} + \omega_{363} + \omega_{364}$$

$$\upsilon_{0110} = \omega_{411} + \omega_{412} + \omega_{511} + \omega_{512}$$

$$\upsilon_{0111} = \omega_{413} + \omega_{414} + \omega_{513} + \omega_{514}$$

$$\upsilon_{0120} = \omega_{421} + \omega_{422} + \omega_{431} + \omega_{432} + \omega_{521} + \omega_{522} + \omega_{531} + \omega_{532}$$

$$\upsilon_{0121} = \omega_{423} + \omega_{424} + \omega_{433} + \omega_{434} + \omega_{523} + \omega_{524} + \omega_{533} + \omega_{534}$$

$$\upsilon_{0130} = \omega_{441} + \omega_{442} + \omega_{451} + \omega_{452} + \omega_{461} + \omega_{462} + \omega_{541} + \omega_{542} + \omega_{551} + \omega_{552} + \omega_{561} + \omega_{562} + \omega_{641} + \omega_{642} + \omega_{651} + \omega_{652} + \omega_{661} + \omega_{662}$$

$$\upsilon_{0131} = \omega_{443} + \omega_{444} + \omega_{453} + \omega_{454} + \omega_{463} + \omega_{464} + \omega_{543} + \omega_{544} + \omega_{553} + \omega_{554} + \omega_{563} + \omega_{564} + \omega_{643} + \omega_{644} + \omega_{653} + \omega_{654} + \omega_{663} + \omega_{664}$$

$$\upsilon_{0210} = \omega_{611} + \omega_{612}$$

$$\upsilon_{0211} = \omega_{613} + \omega_{614}$$

$$\upsilon_{0220} = \omega_{621} + \omega_{622} + \omega_{631} + \omega_{632}$$

$$\upsilon_{0221} = \omega_{623} + \omega_{624} + \omega_{633} + \omega_{634}$$

$$\upsilon_{0230} = \omega_{641} + \omega_{642} + \omega_{651} + \omega_{652} + \omega_{661} + \omega_{662}$$

$$\upsilon_{0231} = \omega_{643} + \omega_{644} + \omega_{653} + \omega_{654} + \omega_{663} + \omega_{664}$$

$$\upsilon_{1010} = \omega_{111} + \omega_{113} + \omega_{121} + \omega_{123} + \omega_{141} + \omega_{143}$$

$$\upsilon_{1011} = \omega_{112} + \omega_{114} + \omega_{122} + \omega_{124} + \omega_{142} + \omega_{144}$$

$$\upsilon_{1020} = \omega_{131} + \omega_{133} + \omega_{151} + \omega_{153}$$

$$\upsilon_{1021} = \omega_{132} + \omega_{134} + \omega_{152} + \omega_{154}$$

$$\upsilon_{1030} = \omega_{161} + \omega_{163}$$

$$\upsilon_{1031} = \omega_{162} + \omega_{164}$$

$$\upsilon_{1110} = \omega_{211} + \omega_{213} + \omega_{221} + \omega_{213} + \omega_{241} + \omega_{243} + \omega_{411} + \omega_{413} + \omega_{421} + \omega_{413} + \omega_{441} + \omega_{443}$$

$$\upsilon_{1111} = \omega_{212} + \omega_{214} + \omega_{222} + \omega_{214} + \omega_{242} + \omega_{244} + \omega_{412} + \omega_{414} + \omega_{422} + \omega_{414} + \omega_{442} + \omega_{444}$$

$$\upsilon_{1120} = \omega_{231} + \omega_{233} + \omega_{251} + \omega_{253} + \omega_{431} + \omega_{433} + \omega_{451} + \omega_{453}$$

$$\upsilon_{1121} = \omega_{232} + \omega_{234} + \omega_{252} + \omega_{254} + \omega_{432} + \omega_{434} + \omega_{452} + \omega_{454}$$

$$\upsilon_{1130} = \omega_{261} + \omega_{263} + \omega_{461} + \omega_{463}$$

$$\upsilon_{1131} = \omega_{262} + \omega_{264} + \omega_{462} + \omega_{464}$$

$$\upsilon_{1210} = \omega_{311} + \omega_{313} + \omega_{321} + \omega_{323} + \omega_{341} + \omega_{343} + \omega_{511} + \omega_{513} + \omega_{521} + \omega_{523} + \omega_{541} + \omega_{543} + \omega_{611} + \omega_{613} + \omega_{621} + \omega_{623} + \omega_{641} + \omega_{643}$$

$$\upsilon_{1211} = \omega_{312} + \omega_{314} + \omega_{322} + \omega_{324} + \omega_{342} + \omega_{344} + \omega_{512} + \omega_{514} + \omega_{522} + \omega_{524} + \omega_{542} + \omega_{544} + \omega_{612} + \omega_{614} + \omega_{622} + \omega_{624} + \omega_{642} + \omega_{644}$$

$$\upsilon_{1220} = \omega_{331} + \omega_{333} + \omega_{353} + \omega_{353} + \omega_{531} + \omega_{533} + \omega_{551} + \omega_{553} + \omega_{631} + \omega_{633} + \omega_{651} + \omega_{653}$$

$$\upsilon_{1221} = \omega_{332} + \omega_{334} + \omega_{352} + \omega_{354} + \omega_{532} + \omega_{534} + \omega_{552} + \omega_{554} + \omega_{632} + \omega_{634} + \omega_{652} + \omega_{654}$$

$$\upsilon_{1230} = \omega_{361} + \omega_{363} + \omega_{561} + \omega_{563} + \omega_{661} + \omega_{663}$$

$$\upsilon_{1231} = \omega_{362} + \omega_{364} + \omega_{562} + \omega_{564} + \omega_{662} + \omega_{664}$$

Gibbs Sampler for Principal Stratum Model Estimation

1. Draw of $\boldsymbol{\beta} \mid$ rest . The placebo dose tolerance $D_i(0)$ takes one of three ordered categories (1 for $\leq$ 25 mg, 2 for 50 mg, 3 for 75 mg). Letting $p_{ij} = P(D_i(0) = j), j = 1, 2, 3$, we define the cumulative probabilities $\pi_{ij} = \sum_{k=1}^{j} p_{ij}, j = 1, 2$. Then we can model the p_{ij} with $\pi_{ij} = \Phi(\beta_j^D - \boldsymbol{X}_i^T \boldsymbol{\beta}_a^A), i = 1, ..., N, j = 1, 2$, where $\boldsymbol{X}_i$ consists of indicator variables for $A_i(0) = a$ for $a = 1, 2$. We use the method of Albert and Chib ([1]) to obtain draws of $\boldsymbol{\beta} = (\beta_1^D, \beta_2^D, \beta_1^A, \beta_2^A)$.

Assume that there exists a latent continuous random variable C_i distributed $N(\boldsymbol{X}_i^T\boldsymbol{\beta}_a^A, 1)$, and we observe $D_i(0)$, where $D_i(0) = j$ if $\beta_{j-1}^D < C_i < \beta_j^D$ with $\beta_0^D = -\infty$ and $\beta_3^D = \infty$. Then this probit regression for $D_i(0)$ can be modeled using the latent variable C_i with normal regression.

With the conjugate $N(\mathbf{0}, V)$ prior, the posterior distribution of $\boldsymbol{\beta}^A$ conditional on $\boldsymbol{D}(0), \boldsymbol{C}, \beta_1^D$, and β_2^D is given by $N_2(\hat{\boldsymbol{\beta}}^A, B^A)$ where $\hat{\boldsymbol{\beta}}^A = ((V)^{-1}+X^TX)^{-1}(X^T\boldsymbol{C})$, and $\boldsymbol{B}^A = ((V)^{-1} + X^TX)^{-1}$.

The fully conditional posterior distribution of C_1,...,C_N are independent with

$C_i|\boldsymbol{\beta}^A, \beta_1^D, \beta_2^D, D_i(0) = j \sim N(\boldsymbol{X}_i^{A^T}\boldsymbol{\beta}^A, 1)$, truncated at the left(right) by $\beta_{j-1}^D(\beta_j^D)$

Finally the fully conditional density of $\beta_1^D, \beta_2^D|\boldsymbol{C}, \boldsymbol{D}(0), \boldsymbol{\beta}^A$ is uniform on the interval [max{max$\{C_i : D_i(0) = j\}$,β_{j-1}^D}, min{min$\{C_i, D_i(0) = j+1\}$,β_{j+1}^D}].

2. Draw of $\boldsymbol{\theta}$ | rest . The parameters for the conditional distributions of $D_i(1)$ | $D_i(0), A_i(1)$ (treatment dose tolerance given control dose tolerance and treatment adverse events) were obtained using the same data augmentation method described in 1).

3. Draw of $\boldsymbol{\alpha}$ | rest . Use A to denote $A_i(1), A_i(0), i = 1, ..., n$ and $\pi(\boldsymbol{\alpha})$ to denote the prior for $\boldsymbol{\alpha}$. The posterior distribution of $\boldsymbol{\alpha}$ is

$$f(\boldsymbol{\alpha}|\text{ rest }) \propto \exp\left(\prod_{i=1}^{n}\sum_{a_1=0}^{2}\sum_{a_0=0}^{a_1}\alpha_{a0,a1}I(A(1)_i = a_1, A(0)_i = a_0) - \log\left(\sum_{a_1=0}^{2}\sum_{a_0=0}^{a_1}\exp(\alpha_{a0,a1})\right)\right)$$
$$\times \pi(\boldsymbol{\alpha})$$

where $\alpha_{0,0} = 0$ for identifiability. This is not the closed form of a known distribution; thus we implement a Metroplis random walk algorithm to get random draws from their posterior distribution.

4. Draw of $\boldsymbol{\gamma} \mid$ rest . Use $D(0)$ to denote $D_i(0), i = 1, ..., n$, $D(1)$ to denote $D_i(1), i = 1, ..., n$, Y to denote $Y_i(1), Y_i(0), i = 1, ..., n$, T to denote $D(1), D(0), Y$, and to denote the prior for $\boldsymbol{\gamma}$. The posterior distribution of $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\gamma}| \text{ rest }) \propto \exp\left(\prod_{i=1}^{n}\sum_{y_0=0}^{1}\sum_{y_1=0}^{1}(\gamma_{y_0,y_1} + \gamma_{d_0,d_1}^{y_0,y_1} I(D_i(0) = d_0, D_i(1) = d_1))I(Y_i(0) = y_0, Y_i(1) = y_1\right.$$
$$\left.\log\left(\sum_{y_0=0}^{1}\sum_{y_1=0}^{1}\exp(\gamma_{y_0,y_1} + \gamma_{d_0,d_1}^{y_0,y_1} I(D_i(0) = d_0, D_i(1) = d_1))\right)\right) \times \pi(\boldsymbol{\gamma})$$

where $\gamma_{0,0} = 0$ and $\gamma_{d_0,d_1}^{0,0} = 0$ for all d_0, d_1 for identifiability. This is not the closed form of a known distribution; thus we implement a Metroplis random walk algorithm to get random draws from their posterior distribution.

5. Draw of counterfactual adverse event $A_i(1 - z_i)$. The distribution of the unobserved highest grade of the adverse events conditional on the observed data and parameters is

$$\begin{aligned} f(A_i(1 - z_i)|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, T) &= f(A_i(1 - z_i)|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, A_i(z_i), D_i, Y_i) \\ &\propto f(A_i, D_i, Y_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto f(A_i|\boldsymbol{\alpha}) f(D_i(0)|A_i(0), \boldsymbol{\beta}) f(D_i(1)|D_i(0), A_i(1), \boldsymbol{\theta}) f(Y_i|D_i, \boldsymbol{\gamma}) \\ &\propto p_{i,a_0,a_1} \times p_{i,d_0} \times p_{i,d_1} \times p_{i,y_0,y_1} \end{aligned}$$

$$\begin{cases} \sim \text{Multinomial}(p_{a_0}, p_{a_1}, p_{a_2}); & z_i = 0, a_{z_i} = 0 \\ \sim \text{Multinomial}(\frac{p_{a_1}}{p_{a_1}+p_{a_2}}, \frac{p_{a_2}}{p_{a_1}+p_{a_2}}); & z_i = 0, a_{z_i} = 1 \\ \sim \text{Multinomial}(0, 0, 1); & z_i = 0, a_{z_i} = 2 \\ \sim \text{Multinomial}(1, 0, 0); & z_i = 1, a_{z_i} = 0 \\ \sim \text{Multinomial}(\frac{p_{a_0}}{p_{a_0}+p_{a_1}}, \frac{p_{a_1}}{p_{a_0}+p_{a_1}}); & z_i = 1, a_{z_i} = 1 \\ \sim \text{Multinomial}(p_{a_0}, p_{a_1}, p_{a_2}); & z_i = 1, a_{z_i} = 2 \end{cases}$$

$$p_{a_0} = P(a_i(1-z_i) = 0, a_i(z_i)|\boldsymbol{\alpha}) \times P(d_i(0)|a_i(0), \boldsymbol{\beta}) \times P(d_i(1)|d_i(0), a_i(1), \boldsymbol{\theta})$$

$$p_{a_1} = P(a_i(1-z_i) = 1, a_i(z_i)|\boldsymbol{\alpha}) \times P(d_i(0)|a_i(0), \boldsymbol{\beta}) \times P(d_i(1)|d_i(0), a_i(1), \boldsymbol{\theta})$$

$$p_{a_2} = P(a_i(1-z_i) = 2, a_i(z_i)|\boldsymbol{\alpha}) \times P(d_i(0)|a_i(0), \boldsymbol{\beta}) \times P(d_i(1)|d_i(0), a_i(1), \boldsymbol{\theta})$$

6. Draw of placebo dose tolerance $D_i(0)$. The distribution of the unobserved placebo dose tolerance among those assigned to the treatment arm conditional on the observed data and parameters is

$$\begin{aligned} f(D_i(0)|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, T) &= f(D_i(0)|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, A_i, D_i(1), Y_i) \\ &\propto f(A_i, D_i, Y_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto f(D_i(0)|A_i(0), \boldsymbol{\beta}) f(D_i(1)|D_i(0), A_i(1), \boldsymbol{\theta}) f(Y_i|D_i, \boldsymbol{\gamma}) \\ &\propto p_{i,d_0} \times p_{i,d_1} \times p_{i,y_0,y_1} \\ &\begin{cases} \sim \text{Multinomial}(p_{d_{01}}, p_{d_{02}}, p_{d_{03}}); & d_i(1) = 1 \\ \sim \text{Multinomial}(\frac{p_{d_{02}} p_{d_{02}} + p_{d_{03}}}{,} \frac{p_{d_{03}}}{p_{d_{02}} + p_{d_{03}}}); & d_i(1) = 2 \\ = 3; & d_i(1) = 3 \end{cases} \end{aligned}$$

$$p_{d_{01}} = P(d_i(0) = 1|a_i(0), \boldsymbol{\beta}) \times P(d_i(1)|d_i(0) = 1, a_i(1), \boldsymbol{\theta}) \times P(\boldsymbol{y}_i|d_i(1), d_i(0) = 1, \boldsymbol{\gamma})$$

$$p_{d_{02}} = P(d_i(0) = 2|a_i(0), \boldsymbol{\beta}) \times P(d_i(1)|d_i(0) = 2, a_i(1), \boldsymbol{\theta}) \times P(\boldsymbol{y}_i|d_i(1), d_i(0) = 2, \boldsymbol{\gamma})$$

$$p_{d_{03}} = P(d_i(0) = 3|a_i(0), \boldsymbol{\beta}) \times P(d_i(1)|d_i(0) = 3, a_i(1), \boldsymbol{\theta}) \times P(\boldsymbol{y}_i|d_i(1), d_i(0) = 3, \boldsymbol{\gamma})$$

7. Draw of treatment dose tolerance $D_i(1)$. The distribution of the unobserved treatment dose tolerance among those assigned to control conditional on the

observed data and parameters is

$$
\begin{aligned}
f(D_i(1)|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},T) &= f(D_i(1)|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},A_i,D_i(0),Y_i) \\
&\propto f(A_i,D_i,Y_i,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\
&\propto f(D_i(1)|D_i(0),A_i(1),\boldsymbol{\theta})f(Y_i|D_i,\boldsymbol{\gamma}) \\
&\propto p_{i,d_1} \times p_{i,y_0,y_1} \\
&\begin{cases} = 1; & d_i(0)=1 \\ \sim \text{Multinomial}(\frac{p_{d_{11}}}{p_{d_{11}}+p_{d_{12}}}, \frac{p_{d_{12}}}{p_{d_{11}}+p_{d_{12}}}); & d_i(0)=2 \\ \sim \text{Multinomial}(p_{d_{11}},p_{d_{12}},p_{d_{13}}); & d_i(0)=3 \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
p_{d_{11}} &= P(d_i(1)=1|d_i(0),a_i(1),\boldsymbol{\theta}) \times P(\boldsymbol{y}_i|d_i(1)=1,d_i(0),\boldsymbol{\gamma}) \\
p_{d_{12}} &= P(d_i(1)=2|d_i(0),a_i(1),\boldsymbol{\theta}) \times P(\boldsymbol{y}_i|d_i(1)=2,d_i(0),\boldsymbol{\gamma}) \\
p_{d_{13}} &= P(d_i(1)=3|d_i(0),a_i(1),\boldsymbol{\theta}) \times P(\boldsymbol{y}_i|d_i(1)=3,d_i(0),\boldsymbol{\gamma})
\end{aligned}
$$

8. Draw of counterfactual clinical outcome $Y_i(1-z_i)$. The distribution of the unobserved clinical outcome conditional on the observed data and parameters is

$$
\begin{aligned}
&f(Y_i(1-z_i)|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},T) \\
&= f(Y_i(1-z_i)|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},A_i,D_i,Y_i(z_i)) \\
&\propto f(A_i,D_i,Y_i,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\
&\propto f(Y_i|D_i,\boldsymbol{\gamma}) \\
&\propto p_{i,y_0,y_1} \\
&\sim \text{Bern}(p_{y_1}), p_{y_1} = \frac{P(y_i(1-z_i)=1,y_i(z_i)|\boldsymbol{d}_i,\boldsymbol{\gamma})}{P(y_i(1-z_i)=1,y_i(z_i)|\boldsymbol{d}_i,\boldsymbol{\gamma})+P(y_i(1-z_i)=0,y_i(z_i)|\boldsymbol{d}_i,\boldsymbol{\gamma})}
\end{aligned}
$$

Figure 3.1: Response under treatment for subjects with dose tolerance of (a). $\leq 25mg$; (b). $50mg$; (c). $75mg$ from one of 200 simulations

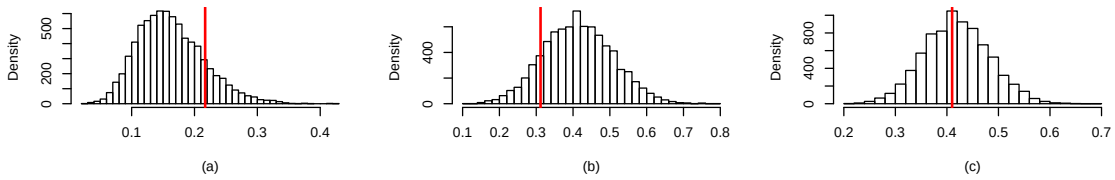


Figure 3.2: Response under control for subjects with dose tolerance of (a). $\leq 25mg$; (b). $50mg$; (c). $75mg$ from one of 200 simulations

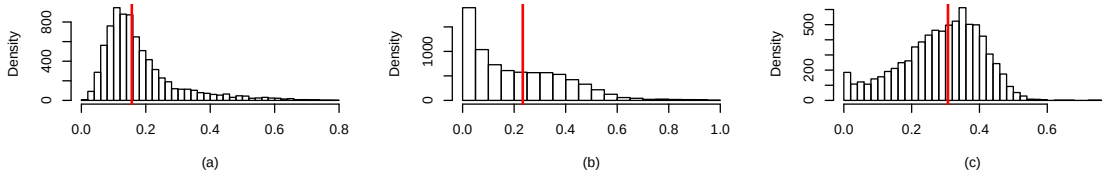


Figure 3.3: Principal effect for subjects with dose tolerance of (a). $\leq 25mg$; (b). $50mg$; (c). $75mg$ from one of 200 simulations

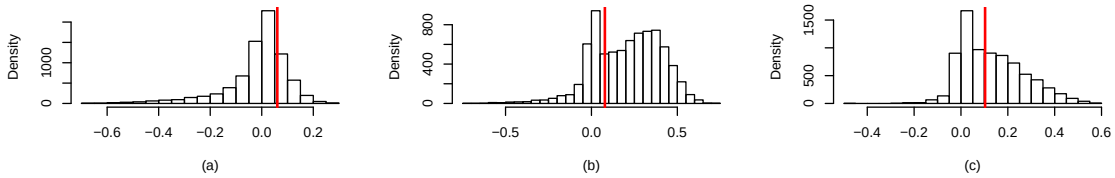
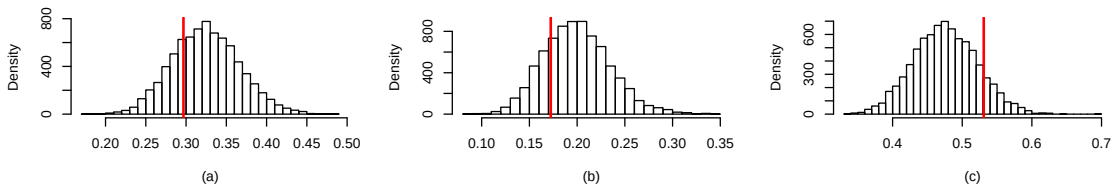


Figure 3.4: Principal strata membership percentage for subjects with dose tolerance of (a). $\leq 25mg$; (b). $50mg$; (c). $75mg$ from one of 200 simulations



CHAPTER IV

Principal Surrogacy in a Time-to-Event Setting

4.1 Introduction

In many circumstances, medical studies evaluating the effect of treatment encounter difficulties, such as long follow-up periods, rare disease outcomes, or high medical costs. In such settings, surrogate markers that can be measured earlier and/or at lower medical cost and used in lieu of primary outcomes to evaluate the effect of treatment are of great interest. Surrogate evaluation methods have been applied in many clinical areas, such as in preventive human immunodeficiency virus (HIV) vaccine clinical trials evaluating vaccine-induced immune responses as a surrogate marker for overall survival (OS) among HIV infected patients [21], and in colorectal cancer clinical trials evaluating disease-free survival (DFS) measured at a short time after randomization as a surrogate marker for OS [52].

The formal definition of surrogate marker was first proposed by Prentice [43]. He suggested modeling the treatment on the primary outcome adjusting for the surrogate marker, after confirming marginal associations between the treatment and outcome, between the surrogate and the outcome, and between the treatment and the surrogate. A valid surrogate marker requires no association between the treatment

and the primary outcome conditional on the surrogate marker. Freedman et al. argued Prentice's criteria is rather stringent and rarely satisfied in practice, since it requires a surrogate marker to fully capture the effect of treatment on the primary outcome [18]. When a less than perfect surrogate marker is available, Freedman et al. suggested measuring the proportion of the treatment effect explained by the surrogate marker, which compares the coefficients of treatment in the models with or without adjusting for the surrogate marker [18]. More recently, Buyse and Molenberghs proposed another set of surrogacy evaluation criteria that compare individual-level surrogacy (the individual-level association between the surrogate marker and the primary outcome after adjusting for treatment) and population-level surrogacy (the treatment effect on the primary outcome relative to that on the surrogate marker) [6]. These criteria was further studied in the setting of meta-analyses [5, 7].

The surrogacy evaluation concept was revisited by Frangakis and Rubin, and methods reviewed above are defined as "statistical surrogacy" [17]. A drawback of these method is that the surrogate marker is measured after the treatment assignment and cannot be simply adjusted for or conditioned on without destroying the causal interpretation of treatment effect in regression [45]. One approach that overcomes such a drawback is through the potential outcome framework. A potential outcome is defined as an outcome that would be observed under different treatment arms [35]. In causal modeling, we define outcomes under all possible treatments, not only the treatment actually assigned, and compare the values of the potential outcomes under different treatment for the same individual. Averaging such comparisons over the entire study population gives the population-level causal effect of treatment.

Several approaches have been proposed to assess surrogacy under the potential outcome framework. Taylor et al. studies the proportion of treatment effect ex-

plained by a binary surrogate marker for a binary primary outcome [56]. Another approach proposed by Robins and Greenland investigates surrogacy through treating the surrogate marker as a mediator and considering indirect effects of treatment (through the surrogate marker) and direct effects of treatment (around the surrogate marker) [44]. Assuming that the value of surrogate marker can be manipulated separately from the treatment assignment, the direct effect is the comparison of the primary outcome under different treatment arms when the value of the surrogate marker is held constant, while the indirect effect is the comparison of the primary outcomes under different treatment arms when the value of the surrogate marker is changed to that it would have been under treatment and control arms.

Alternatively, Frangakis and Rubin proposed a principal surrogacy evaluation method based on the concept of principal stratification [17]. A principal stratum is defined by the joint distribution of the intermediate variable under different treatment arms. They suggested that analysis should focus on the estimation of the treatment effect within principal strata. Because the potential outcome is considered to exist before the action of treatment assignment, the principal strata can also be considered as existing before the action of treatment assignment, and can be adjusted for or conditioned on in regression without destroying the causal interpretation of estimated treatment effect. They proposed that an intermediate variable is a valid principal surrogate if it satisfies causal necessity, i.e., the causal effect of treatment on the primary outcome exists only when the causal effect of treatment on the surrogate marker exists. They also proposed two types of causal effects, associative and dissociative effects, to evaluate principal surrogacy. The associative effect on the outcome is defined as the comparison between the potential primary outcomes under different treatment arms when the value of surrogate markers are different under

different treatment arms, and the dissociative effect on the outcome is defined as the comparison between potential primary outcomes under different treatment arms when the value of surrogate markers are same under different treatment arms. The distinction between this approach and the one proposed by Robins and Greenland [44] is that this approach does not assume the ability to manipulate the value of the surrogate marker independently from the treatment assignment.

More recently, the concept of principal surrogacy was revisited by Gilbert and Hudgens [20], in the context of a binary outcome and a continuous principal surrogate. They termed the causal necessity proposed by Frangakis and Rubin as average causal necessity, and defined the average causal sufficiency in terms of a risk difference when the difference between the surrogate under treatment and the surrogate under control is large than a given constant. They suggested a refined definition of principal surrogate as a biomarker satisfying both average causal necessity and average causal sufficiency.

The current literature on principal surrogacy assessment has thus far considered either a normally distributed primary outcome [11] or a binary primary outcome [31]. Here we extend this previous work to evaluate principal surrogacy when the outcome is a (possibly censored) time-to-event measure, which is a critical formulation for many clinical trials settings, particularly for cancer treatments. In this article, we propose a model to evaluate the principal surrogacy for a binary surrogate marker and a time-to-event primary outcome based on a proportional hazards modeling assumption. In addition, to take account of the correlation between the potential primary outcomes under different treatment arms, we introduce a shared-frailty (random-effect) model in conjunction with proportionality assumption [23, 29]. Like other random-effect models, the variance of frailty terms naturally describes the

population heterogeneity. However, unlike random effects in some statistical models, the variance of frailty terms in a proportional hazard (PH) model is identifiable even for the univariate case under certain conditions [30], as this merely represents a deviation from the proportionality assumption at the marginal model level. We illustrate the proposed causal model using randomized clinical trial to evaluate the principal surrogacy. This clinical trial was designed to evaluate the efficacy of intensive-course fluorouracil (5FU) combined with low-dose leucovorin as postoperative adjuvant therapy in patients with high-risk primary colorectal cancer, and the primary outcome was the OS. It has been of great interest to find valid surrogate markers that are available and may be assessed earlier than the primary endpoint in terms of efficacy in this disease setting. Here, the putative surrogate marker in consideration is censoring status of 3-year DFS, i.e., 1 (no event experienced, "favorable" result) or 0 (event experienced, "unfavorable" result).

The remainder of this manuscript is organized as follows. In Section 4.2, we propose a causal model under the potential outcome framework using the principal stratification approach to study principal surrogacy, extending the work of Frangakis and Rubin, and Gilbert and Hudgens into the time-to-event outcome setting. In Section 4.3, we illustrate the proposed causal model and estimation method on a randomized clinical trial for adjuvant colorectal cancer to evaluate 3-year DFS as a surrogate marker for OS. We carry out simulation studies for the proposed causal model to investigate its repeated sampling property in Section 4.4. In Section 4.5, we discuss the implications of our findings and future extensions.

4.2 Principal Stratification Model to Assess Surrogacy in a Time-to-Event Setting

4.2.1 Notation

Consider a binary randomized assignment $Z = 1$ or $Z = 0$ for treatment and control. The binary surrogate marker is denoted by S_z, where $S_z = 1$ for "favorable" results, and $S_z = 0$ for "unfavorable" results under treatment assignment $Z = z$. The primary time-to-event outcome is similarly denoted by T_z. In many time to event settings, there will also be a censoring indicator C_z. Subscripting with i to denote subject i, let δ_{iz} denote the event indicator variable $(T_{iz} < C_{iz})$, taking value 1 if subject i is observed to fail under arm z and 0 otherwise. Finally, let the observed outcome $Y_{iz} = T_{iz} \wedge C_{iz}$, hence Y_{iz} is a failure time T_{iz} if $\delta_{iz} = 1$ and a censoring time C_{iz} otherwise.

4.2.2 Principal Strata

The principal strata membership is determined by the joint distribution of the surrogate marker S_{i1} and S_{i0}. For a binary surrogate marker, subjects belong to one of four possible principal strata: $(S_{i1} = S_{i0} = 1)$, $(S_{i1} = S_{i0} = 0)$, $(S_{i1} = 1, S_{i0} = 0)$, and $(S_{i1} = 0, S_{i0} = 1)$, corresponding to situations in which the treatment has no impact on the surrogate marker $((S_{i1} = S_{i0} = 1), (S_{i1} = S_{i0} = 0))$, situations in which the surrogate marker is positively impacted by the treatment $(S_{i1} = 1, S_{i0} = 0)$, and situations in which the surrogate maker is negatively impacted by the treatment $(S_{i1} = 0, S_{i0} = 1)$.

4.2.3 Model Assumptions

We make several reasonable assumptions to assist with identifiability, including the ignorable treatment assignment assumption, the stable unit treatment value assumption (SUTVA), monotonicity assumption and the latent noninformative censoring assumption.

1. Treatment assignment is ignorable if conditional on the observed baseline covariates, the treatment assignment is independent of all baseline covariates and potential outcomes, $Z_i \perp Y_{i1}, Y_{i0}, S_{i1}, S_{i0}, \boldsymbol{X}_i, \forall i$ [47]. Under the ignorability, we do not need to model the treatment assignment mechanism.

2. SUTVA consists of two subassumptions. The first subassumption assumes there is no interference between the potential outcomes of different individuals, $Y_{i0}, Y_{i1}, S_{i0}, S_{i1} \perp Y_{j0}, Y_{j1}, S_{j0}, S_{j1}, i \neq j$ [48]. The second subassumption assumes that there exists only one form of the treatment, so that the observed $y_{iz} = z * Y_{iz} + (1 - z) * Y_{iz}$ [48].

3. The monotonicity assumption implies monotonicity of the potential outcomes. Specifically, for the application study, we assume the monotonicity of the putative surrogate marker, the censoring status of DFS at 3 years. It is satisfied in the application study due to the fact that it is well accepted that the treatment based on chemotherapy of 5FU has a better treatment effect than the control (surgery only).

4. Latent noninformative censoring assumes that conditional on observed baseline covariates and partially latent principal stratum membership, the event time and censoring time are independent of each other: $T_{i0}, T_{i1} \perp C_{i0}, C_{i1} \mid S_{i0}, S_{i1}, \boldsymbol{X}_i$. Note that this assumption is slightly different from the conventional noninforma-

tive censoring assumption, where the independence is conditional on observed baseline covariates only.

4.2.4 Complete Data Likelihood

To take account of the correlation between the potential primary outcomes under different treatment arms on the same individual, we construct a shared-frailty (random-effect) survival model in conjunction with proportionality assumption. A shared-frailty survival model includes a random effect term besides fixed effect terms to account for the unexplained variability of the model and the correlations among the survival times of subjects within clusters. Under the potential outcome framework, an individual is considered as a cluster, and the constructed shared-frailty survival model correlates potential primary outcomes under different treatment arms on the same individual. Let W_i denote the frailty random effect for subject i and $\boldsymbol{W} = (W_1, W_2, ..., W_n)$. Let $\boldsymbol{Y} = (Y_{11}, Y_{10}, ..., Y_{n1}, Y_{n0})$, $\boldsymbol{\delta} = (\delta_{11}, \delta_{10}, ..., \delta_{n1}, \delta_{n0})$, $\boldsymbol{S} = (S_{11}, S_{10}, ..., S_{n1}, S_{n0})$, and $\boldsymbol{Z} = (Z_1, ..., Z_n)$. Under the potential outcome framework, the "complete" data include $\boldsymbol{Y}, \boldsymbol{\delta}$, $\boldsymbol{S}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$. Let $\boldsymbol{X}$ denote $n \times p$ matrix of baseline covariates with i^{th} row $\boldsymbol{X}_i$. Denote the survival function by $S(.)$ and hazard function by $h(.)$. We factor the "complete" data likelihood:

$$
\begin{aligned}
&L(\boldsymbol{Y},\boldsymbol{\delta},\boldsymbol{S},\boldsymbol{W},\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\tau},\boldsymbol{\zeta}) \\
=&\prod_{i=1}^{n} f(Z_i) f(Y_{i1}, Y_{i0}, \delta_{i1}, \delta_{i0}, S_{i1}, S_{i0}|\boldsymbol{X}_i, W_i, \boldsymbol{\tau}, \boldsymbol{\zeta}) f(W_i) \\
=&\prod_{i=1}^{n} f(Z_i) f(Y_{i1}, Y_{i0}, \delta_{i1}, \delta_{i0}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \boldsymbol{\tau}) f(S_{i1}, S_{i0}|\boldsymbol{X}_i, \boldsymbol{\zeta}) f(W_i) \\
=&\prod_{i=1}^{n} f(Z_i) \prod_{z=0}^{1} f(Y_{iz}, \delta_{iz}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \tau) f(S_{i1}, S_{i0}|\boldsymbol{X}_i, \boldsymbol{\zeta}) f(W_i) \\
=&\prod_{i=1}^{n} f(Z_i) \prod_{z=0}^{1} f(Y_{iz}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \tau)^{\delta_{iz}} S(Y_{iz}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \tau)^{1-\delta_{iz}} f(S_{i1}, S_{i0}|\boldsymbol{X}_i, \boldsymbol{\zeta}) f(W_i) \\
=&\prod_{i=1}^{n} f(Z_i) \prod_{z=0}^{1} h(Y_{iz}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \tau)^{\delta_{iz}} S(Y_{iz}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \tau) f(S_{i1}, S_{i0}|\boldsymbol{X}_i, \boldsymbol{\zeta}) f(W_i)
\end{aligned}
$$

where $\boldsymbol{\tau}$ parameterizes the conditional distribution of primary outcomes given principal strata membership and frailty, and $\boldsymbol{\zeta}$ parameterizes the marginal distribution of principal strata membership. The first equality follows from the ignorable treatment value assumption and SUTVA. The second equality factors the complete data likelihood into the product of the conditional likelihood of primary outcomes given principal strata membership and frailty, and the marginal likelihood of principal strata membership. It follows from the marginal independence of frailty, leading to $W_i \perp S_{i1}, S_{i0}, \boldsymbol{X}_i, \boldsymbol{\zeta}$. The third equality follows because the primary outcomes of one individual under different treatment arms are independent of each other given the frailty. The fourth equality follows from the relationship between probability distribution functions, hazard functions, and survival functions.

We model principal strata membership using a baseline multinomial logit model. Let $p_{s_1 s_0} = P(S_{i1} = s_1, S_{i0} = s_0|\boldsymbol{X}_i, \boldsymbol{\zeta})$, then $\log(\frac{p_{s_1 s_0}}{p_{00}}) = \zeta_0^{s_1 s_0} + \boldsymbol{\zeta}_1^{s_1 s_0'} \boldsymbol{X}_i$, for $s_1, s_0 = (1,0)$, $(0,1)$, $(1,1)$.

For the primary outcome, we model the hazard function of Y_{iz} given principal

strata S_{i1}, S_{i0} under the PH assumption:

$$h(Y_{iz}|S_{i1}, S_{i0}, \boldsymbol{X}_i, W_i, \boldsymbol{\tau}) = h_0(Y_{iz}) W_i e^{\beta_{zs_1s_0} I(S_{i1}=s_1, S_{i0}=s_0, Z_i=z) + \boldsymbol{\beta}_X' \boldsymbol{X}_i}$$

where $h_0(.)$ is the baseline hazard function common to every subject, W_i is the random effect that inflates or deflates the baseline hazard function, and $\beta_{000} = 0$ for identifiability. Note that the proportionality assumption not only applies to the baseline covariates, but also to the principal strata membership.

In practice, the potential surrogate and outcome under the arm to which the subject is not assigned is unobservable, so there are large amount of missing data. Given this situation, we choose to model the baseline hazard parametrically, under a Weibull distribution assumption, so that $h_0(Y_{iz}) = \gamma Y_{iz}^{\alpha-1}$. This will ease estimation, while still being flexible since it has both shape and scale parameters.

4.2.5 Gamma Frailty and Identifiability

We anticipate some degree of within-subject correlation among the potential outcomes (as with the surrogate markers). We incorporate this into our model by assuming a common random effect or "frailty" W_i associated with the baseline hazard. We assume the frailty W_i follows a gamma distribution with a finite mean of 1 for both mathematical tractability and identifiability reasons [22, 29], such that $W_i \sim G(\eta, \eta^{-1})$, and η^{-1} is the variance of the W_i. Small values of η (small values of $\frac{1}{\eta}$) imply large amounts of heterogeneity between individuals, and thus high correlation between the potential outcomes of an individual under different treatment arms. In addition, the introduction of frailty incorporates the deviation from the proportionality assumption at the marginal model level.

The proposed shared gamma frailty PH causal model is constructed under the potential outcome framework, with "complete" data including the potential outcomes

under both treatment and control arms. However, in practice, we observe the potential outcome only under the arm to which the subject is assigned. Therefore, the correlation between the potential outcomes of the same subject cannot be identified in general. However, note that in the proposed shared gamma frailty PH causal model, while allowing for correlation between potential outcomes, the variance of the frailty is identifiable. Because an individual can be assigned to only one arm, we can consider the "observed" data as consisting of clusters with only one observation in each cluster. It has been shown that the variance of a frailty can be inferred from a frailty PH model even in the univariate case, as long as the hazard function models at least one covariate and the covariate is sufficiently variable [13, 30]. In our proposed causal model, we model the hazard function with both the principal strata membership and baseline covariates, which protects us from lack of identifiability of the frailty variance. We investigate the identifiability of frailty variance of the proposed shared gamma frailty PH causal model in the application by comparing prior and posterior distributions for the gamma frailty parameter.

4.2.6 Estimands to Assess Principal Surrogacy

Gilbert and Hudgens [20] define two measures of principal surrogacy in the setting of a binary outcome and continuous surrogate measure: average causal necessity given by $\text{risk}_1(s_1, s_0) = \text{risk}_0(s_1, s_0)$ if $s_1 = s_0$, and average causal sufficiency given by $\text{risk}_1(s_1, s_0) \neq \text{risk}_0(s_1, s_0)$ for all $|s_1 - s_0| > C$, where C is a constant >0 and $\text{risk}_z(s_1, s_0) = \Pr(Y_z = 1 | S_1 = s_1, S_0 = s_0)$. They also introduced the causal effect predictiveness (CEP) surface and the statistic $\text{CEP}^{\text{risk}}(s_1, s_0) = g(\text{risk}_1(s_1, s_0), \text{risk}_0(s_1, s_0))$, where $g(\cdot, \cdot)$ is a known contrast function satisfying $g(x, y) = 0$ if and only if $x = y$. This statistic can be used to assess principal surrogate based on

the refined definition, since $\text{CEP}^{\text{risk}}(s, s) = 0$ is equivalent to average causal necessity, and $\text{CEP}^{\text{risk}}(s_1, s_0) \neq 0$ for all $|s_1 - s_0| > C$ is equivalent to average causal sufficiency. Based on $\text{CEP}^{\text{risk}}(s_1, s_0)$, they proposed expected associative effect (EAE) and expected dissociative effect (EDE) to quantify how well the causal effect of treatment on surrogate marker can predict the causal effect of treatment on primary outcome. EAE is defined as the weighted average of $\text{CEP}^{\text{risk}}(s_1, s_0)$ when $s_1 \neq s_0$, and EDE is defined as the weighted average of $\text{CEP}^{\text{risk}}(s_1, s_0)$ when $s_1 = s_0$.

We extend these concepts into the setting of a binary surrogate and time-to-event outcome as follows. First, let $\pi_{s_1 s_0}(t)$ denote the principal hazard ratio, which is the hazard ratio within principal strata (s_1, s_0) at time t. Because we assume proportionality conditional on the frailty, we drop t from the notation of hazard ratios, and the notation of principal hazard ratio within stratum "$s_1 s_0$" is simplified to $\pi_{s_1 s_0}$. With the proposed shared frailty PH causal model, $\pi_{s_1 s_0} = exp(\beta_{1 s_1 s_0} - \beta_{0 s_1 s_0})$.

Let $\text{risk}_{iz}(s_1, s_0) = h(y_{iz}|s_1, s_0, x_i, w_i)$, and $\text{CEP}^{\text{risk}}(s_1, s_0) = log(\pi_{s_1 s_0})$,which is the difference of two intercept terms $\beta_{1 s_1 s_0} - \beta_{0 s_1 s_0}$, simple and recognizable from the hazard modeling perspective. Note that $\text{CEP}^{\text{risk}}(s, s) = 0 \ \forall \ s$ is equivalent to average causal necessity, and $\text{CEP}^{\text{risk}}(s_1, s_0) \neq 0$ for all $|s_1 - s_0| > C$ is equivalent to average causal sufficiency.

We define *expected associative effect* (EAE) and *expected dissociative effect* (EDE)

under the proposed shared gamma frailty PH causal model:

$$\begin{aligned} EAE =& \frac{E[w(S_{i1}=s_1, S_{i0}=s_0)\mathrm{CEP}^{\mathrm{risk}}(s_1,s_0)|s_1=s_0]}{E[w(S_{i1}=s_1, S_{i0}=s_0)|s_1=s_0]} \\ =& \frac{E[w(S_{i1}=s_1, S_{i0}=s_0)log(\pi_{s_1s_0})|s_1=s_0]}{E[w(S_{i1}=s_1, S_{i0}=s_0)|s_1=s_0]} \end{aligned}$$

$$\begin{aligned} EDE =& \frac{E[w(S_{i1}=s_1, S_{i0}=s_0)\mathrm{CEP}^{\mathrm{risk}}(s_1,s_0)|s_1\neq s_0]}{E[w(S_{i1}=s_1, S_{i0}=s_0)|s_1\neq s_0]} \\ =& \frac{E[w(S_{i1}=s_1, S_{i0}=s_0)log(\pi_{s_1s_0})|s_1\neq s_0]}{E[w(S_{i1}=s_1, S_{i0}=s_0)|s_1\neq s_0]} \end{aligned}$$

where $w(\cdot,\cdot)$ is a nonnegative weight function. In our binary example we consider $w(S_{i1}=s_1, S_{i0}=s_0) = I(S_{i1}=s_1, S_{i0}=s_0)$.

4.2.7 Model Estimation

We face the usual limitation of causal modeling that we cannot observe $Y_{i,1-z}, S_{i,1-z}$ $\forall i$. The frailty W_i is also unobserved quantity, thus our model faces a large amount of missing data. We propose to use Bayesian estimation with a Markov chain Monte Carlo (MCMC) algorithm for model estimation. To obtain the joint posterior distribution of parameters γ, α, $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$ and η, we use a Gibbs sampler. The posterior distributions of α, $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$ and η are not of closed form of any known parametric distributions, therefore we implement Metropolis-Hastings within Gibbs algorithms to obtain random draws from their posterior distributions. In each iteration of MCMC, we first obtain a random draw of each parameter conditional on the "complete" data and other parameters; we then impute the missing outcomes conditional on the observed data and the updated parameters. These two steps are repeated until all parameters converge in distribution. We run multiple chains with different initial values of parameters, and calculate a measure of between and within-chain variance ($\hat{R}$) to check the convergence of parameters [19]. $\hat{R} < 1.1$ is acceptable, and a $\hat{R}$

close to 1 indicates the good convergence of the parameters. Details of the conditional posterior distributions of parameters and conditional distributions of missing outcomes are given in the Appendix.

4.3 Application

Multiple adjuvant chemotherapies based on fluorouracil (5FU) for resected high-risk colon cancers were studied in 80's and 90's [24, 39, 52], and evidence of improved DFS and OS were robustly observed in a number of randomized phase III clinical trials with different chemotherapy regimens based on 5FU. A pooled analysis based on 18 randomized trials by Sargent et al. suggested that, in adjuvant colon cancer setting, 5-year OS is adequate to serve as a valid primary endpoint, and 3-year DFS is an appropriate endpoint to replace the primary endpoint for efficacy analysis [52].

We use one of these trials, North Central Cancer Treatment Group (NCCTG) 874651, to illustrate our proposed method [41]. This trial was originally conducted and led by NCCTG during 1988-1989, and was designed to compare intensive-course 5FU with low-dose leucovorin as postoperative adjuvant chemotherapy versus surgery only in stage II and III colon cancer patients. The objective of the study was to evaluate the efficacy of proposed chemotherapy regimen in terms of OS. The original study design was a 3-arm randomized trial with two different experimental chemotherapy regimens and one control arm (surgery only), and stratified by a number of important histological factors. The study was closed early due to the emerging positive results for 5FU-based adjuvant regimens from a national intergroup trial [39]. For illustration purposes, we combine the two treatment arms as they have very similar impacts on DFS and OS, and ignore the stratification variable in our analysis thereafter. More information about this study can be found by O'Connell et al [41].

Table 4.1: Demographic characteristics for the NCCTG trial 874651

	Treatment Assignment	
Variable	Treatment	Control
Number of Patients	255	153
Age, Mean (S.D.) ($p = 0.52$)	60.8 (10.7)	61.6 (10.8)
Gender ($p = 0.05$)		
Male	130 (51.0%)	94 (61.4%)
Female	125 (49.0%)	59 (38.6%)
Tumor Stage ($p = 0.92$)		
stage II	47 (18.4%)	28 (18.3%)
stage III	208 (81.6%)	125 (81.7%)
Censoring Status of 3-Year DFS ($p = 0.25$)		
Fail	70 (28.2%)	51 (34.2%)
Disease Free	178 (71.8%)	98 (65.8%)

We summarize subjects' baseline covariates and 3-year DFS in Table 4.1. Corresponding p values from Chi-square tests and t-tests indicate randomization achieved baseline balance with respect to demographic characteristics with the possible exception of gender. Overall there are 178 (71.8%) subjects in the treatment group, and 98 (65.8%) subjects in the control group who survived disease free at 3 years after randomization. We conduct log-rank test to compare the risk of death in treatment and control arms. The log-rank test yields Chi-square statistics $\chi^2 = 2.1$ with 1 degrees of freedom and p value of 0.15, indicate that the data cannot conclude that the subjects in the treatment group have significantly lower risk of death compared to subjects in the control group.

4.3.1 Conventional PH Models Analysis

We first fit the data with semiparametric Cox PH model with and without adjusting for the *observed* surrogate marker (censoring status of DFS at 3 years). Subjects' age, gender and tumor stage are included as baseline covariates, where age is normalized with sample mean and standard deviation to have mean 0 and variance 1. We summarize the multivariate analysis results in Table 4.2. After adjusting for age and gender, there is a strong effect of tumor stage (hazard ratio = 1.93, $p = .02$); treat-

ment reduces the risk of death relative to control (hazard ratio = 0.76, $p = 0.12$), although this difference is not statistically significant. To check the Weibull baseline hazard assumption in the NCCTG trial 874651, we fit the data assuming a Weibull baseline hazard. The estimated covariate coefficients from the PH model assuming Weibull baseline hazard are nearly identical to those estimated from Cox PH model, indicating that the Weibull baseline hazard assumption is reasonable in this NCCTG trial under proportionality assumption, since Cox PH model leaves the baseline hazard function unspecified and estimates it nonparametrically. To assess the statistical surrogacy, we include the censoring status of 3-year DFS in the hazard function. After adjustment, the hazard ratio is closer to 1 than not adjusting for (hazard ratio = 0.88, $p = 0.49$), and the observed surrogate marker has a significant effect on the primary outcome with $p < 0.001$. Although the lack of a significant treatment effect invalidates the use of Prentice's criteria, the substantial reduction in the adjusted effect of treatment, together with the significant association with the surrogate marker (hazard ratio = 33.52, $p < 0.001$), suggests that the putative surrogate marker is a valid (statistical) surrogate marker for overall survival.

To have a proper comparison with our proposed shared gamma frailty PH causal model, we further fit the data with univariate gamma frailty Cox PH model assuming Weibull baseline hazard. The univariate gamma frailty Cox PH model reveals another aspect of this data set, namely the heterogeneity-induced non-proportionality in terms of OS. The analysis results suggest a slightly more pronounced but insignificant treatment effect (hazard ratio = 0.64, $p = 0.13$) than the marginal model, and similar trends for other baseline covariates. This observation is expected as marginal models tend to provide attenuated results relative to frailty models [2]. In addition, adjusting for the putative surrogate marker further reduces the treatment effect (haz-

Table 4.2: Estimates of covariates coefficients and parameters with conventional PH models for the NCCTG trial 874651

Covariate/Parameter	Coef	exp(Coef)	S.E.(Coef)	p value
Cox PH Model				
Treatment vs. Control	-0.28	0.76	0.18	0.12
Tumor Stage III vs. II	0.66	1.93	0.28	0.02
Gender, Female vs. Male	-0.15	0.86	0.18	0.39
Age	-.003	1.00	0.01	0.76
Cox PH Model Adjusted For Observed Surrogate Marker				
Treatment vs. Control	-0.12	0.88	0.18	0.49
Observed Surrogate Marker, "Favorable" vs. "Unfavorable"	3.51	33.52	0.25	<.001
Tumor Stage III vs. II	0.16	1.18	0.28	0.55
Gender, Female vs. Male	-0.31	0.73	0.18	0.08
Age	0.02	1.02	0.01	0.06
Gamma Frailty PH Model Assuming Weibull Baseline Hazard				
Treatment vs. Control	-0.44	0.64	0.29	0.13
Tumor Stage III vs. II	0.87	2.39	0.38	0.02
Gender, Female vs. Male	-0.29	0.75	0.28	0.31
Age	<.001	1.00	0.01	0.99
Variance of Frailty	1.87			
Gamma Frailty PH Model Assuming Weibull Baseline Hazard Adjusted For Observed Surrogate Marker				
Treatment vs. Control	-0.25	0.78	0.25	0.32
Observed Surrogate Marker, "Favorable" vs. "Unfavorable"	4.11	60.95	0.43	<.001
Tumor Stage III vs. II	0.27	1.31	0.36	0.46
Gender, Female vs. Male	-0.48	0.62	0.25	0.06
Age	0.01	1.01	0.01	0.26
Variance of Frailty	0.61			

ard ratio = 0.78, $p = 0.32$), with a very strong association between surrogate marker and primary endpoint (hazard ratio = 60.95, $p < 0.001$), similar to marginal model findings.

4.3.2 Shared Gamma Frailty PH Causal Model Analysis

We apply the proposed shared gamma frailty PH causal model and Bayesian estimation method on the NCCTG trial to estimate the causal effect of treatment and assess principal surrogacy of the censoring status of 3-year DFS for OS. The first two model assumptions we have discussed are satisfied in this randomized clinical trial. First, the ignorable treatment assignment assumption is satisfied because of the

randomized treatment assignment. The first subassumption of SUTVA is satisfied because the colon cancer is not an infectious disease and we do not anticipate that primary outcomes or surrogate markers of subject i impact any of those of subject j in any fashion ($i \neq j$). The second subassumption of SUVTA is also satisfied because the treatments subjects in the NCCTG trial received were reasonably well controlled. Next, there is no reason to suspect non-informative censoring assumption is violated, since the censoring was administrative due to the end of the trial, rather than due to loss of follow-up. Our third assumption,"monotonicity"for the surrogate marker, i.e. subjects who would have remission or death within 3 years under treatment but disease free beyond 3 years under control do not exist, is not testable, but reasonable for an effective treatment.

To check convergence of distributions of the parameters, we run multiple chains with different starting points. The priors of parameters are assumed as $\pi(\eta) \sim Gamma(2,2)$, $\boldsymbol{\beta} \sim N(\mathbf{0}, 10I)$, $\boldsymbol{\zeta} \sim N(\mathbf{0}, 10I)$, $\alpha \sim Gamma(2, 0.75)$, and $\gamma \sim Gamma(1,1)$. These priors are very weakly informative, but bound the model probabilities away from values extremely close to 0 or 1. Parameters converge in their distributions after 300,000 iterations with 100,000 used as burn-in. We calculate the $\hat{R}$ based on multiple chains to check the convergence of parameters. The maximum $\hat{R}$ is 1.02, indicating that the parameters converge in distribution.

Under the monotonicity assumption, subjects who did not have 3-year DFS event on the control arm must belong to principal stratum ($S_{i1} = 1, S_{i0} = 1$); similarly those who had 3-year DFS event on the treatment arm must belong to principal stratum ($S_{i1} = 0, S_{i0} = 0$). Subjects who did not have a 3-year DFS event under the treatment arm may belong to ($S_{i1} = 1, S_{i0} = 1$) or ($S_{i1} = 1, S_{i0} = 0$); similarly subjects who had a 3-year DFS event under the control arm may belong to ($S_{i1} =$

$0, S_{i0} = 0)$ or $(S_{i1} = 1, S_{i0} = 0)$. Based on this and observed covariates used to predict principal stratum membership, subjects' principal strata membership are imputed accordingly in each iteration of MCMC.

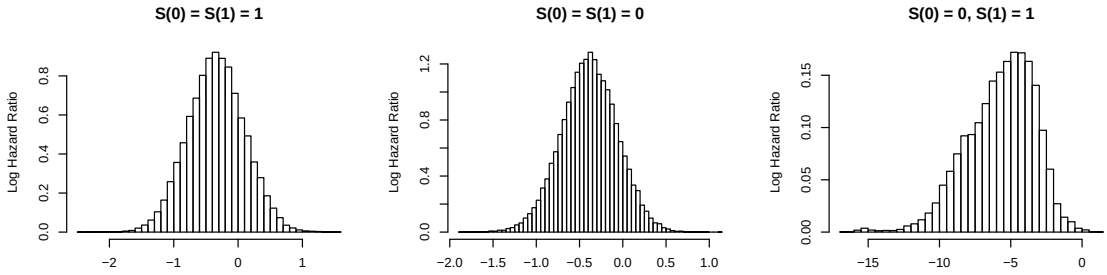
The analysis suggests the proportion of the subjects having "favorable" surrogate markers under both treatment arms ($S_{i1} = S_{i0} = 1$) is 64% (95% CI, 61%-68%), the proportion of subjects having "unfavorable" surrogate markers under both treatment arms ($S_{i1} = S_{i0} = 0$) is 28% (95% CI, 26%-29%), and the proportion of subjects having "favorable" surrogate marker under treatment arm and "unfavorable" surrogate marker under control arm ($S_{i1} = 1, S_{i0} = 0$), i.e., those for whom the treatment changes the surrogate marker, is 8% (95% CI, 4%-12%). The principal hazard ratios with point estimates of medians and their 95% credible intervals are summarized in Table 4.3. For subjects who would not benefit from the treatment in terms of surrogate markers ($S_{i1} = S_{i0} = 1$ or $S_{i1} = S_{i0} = 0$), the principal hazard ratios of treatment to control is 0.71 and 0.68 respectively. Both of the 95% credible intervals cover 1, indicating the treatment does not have significant effects compared with the control on primary endpoints. For subjects who would indeed benefit from the treatment in terms of surrogate markers ($S_{i1} = 1, S_{i0} = 0$), the hazard ratio is 4.42×10^{-3} with 95% credible interval of $(2.25 \times 10^{-5}, 0.15)$, indicating the treatment has significant effects compared with the control on primary endpoints. We further plot the distributions of log principal hazard ratios in Figure 4.1.

We summarize the expected associative effect (EAE) and expected dissociative effect (EDE) of treatment compared with control in Table 4.4. Because we make monotonicity assumption for censoring status of 3-year DFS, the EAE is equal to the log principal hazard ratio for subjects with $(S_{i1} = 1, S_{i0} = 0)$. The 95% C.I. does not cover 0, indicating the putative surrogate marker satisfies the one sided average

Table 4.3: Principal hazard ratios assuming various hyperpriors for gamma frailty parameter η for the NCCTG trial 87651 (variance of frailty = $\frac{1}{\eta}$) (median with 95% credible interval in parenthesis)

	Principal Strata		
$\pi(\eta)$	$S_{i1} = S_{i0} = 1$	$S_{i1} = S_{i0} = 0$	$S_{i1} = 1, S_{i0} = 0$
$Gamma(2, 2)$	0.71 (0.30, 1.68)	0.68 (0.36, 1.27)	4.42×10^{-3} (2.25×10^{-5}, 0.15)
$Gamma(0.1, 10)$	0.67 (0.29, 1.60)	0.67 (0.34, 1.28)	5.91×10^{-3} (8.59×10^{-5}, 0.13)
$Gamma(1, 1)$	0.67 (0.28, 1.60)	0.66 (0.33, 1.30)	4.52×10^{-3} (2.55×10^{-5}, 0.13)
$Gamma(1, 2)$	0.67 (0.29, 1.63)	0.67 (0.33, 1.30)	4.58×10^{-3} (2.34×10^{-5}, 0.18)
$Gamma(5, 1)$	0.68 (0.29, 1.62)	0.69 (0.37, 1.26)	4.84×10^{-3} (1.53×10^{-5}, 0.18)
$Gamma(8, 0.5)$	0.69 (0.30, 1.58)	0.69 (0.37, 1.27)	5.86×10^{-3} (3.98×10^{-5}, 0.14)

Figure 4.1: Distribution of Log principal hazard ratios for the NCCTG trial 874651

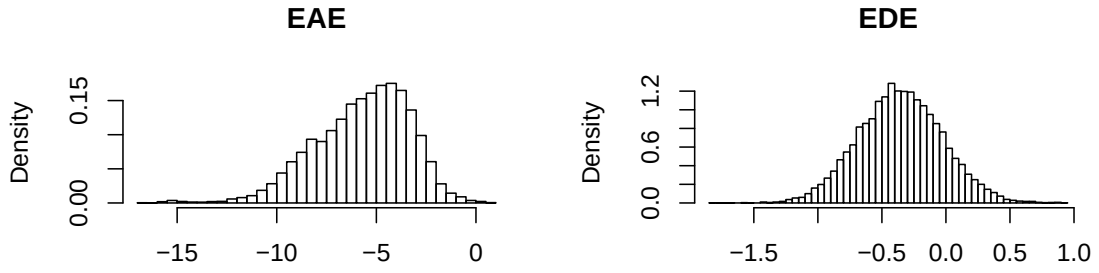


causal sufficiency that $\text{risk}_1(S_{i1} = s_1, S_{i0} = s_0) < \text{risk}_0(S_{i1} = s_1, S_{i0} = s_0)$ when $s_1 > s_0$, i.e. the clinical benefit of treatment on surrogate marker predicts clinical benefit of treatment on primary outcome. The EDE has median value -0.36 with 95% C.I. covering 0, indicating the putative surrogate marker satisfies the average causal necessity that $\text{risk}_1(S_{i1} = s_1, S_{i0} = s_0) = \text{risk}_0(S_{i1} = s_1, S_{i0} = s_0)$ when $s_1 = s_0$, i.e. subjects who would not benefit from the treatment on the surrogate marker will not benefit from the treatment on the primary outcome. Because the censoring status of 3-year DFS satisfies both the average causal necessity and average causal sufficiency for OS in NCCTG trial, we conclude that the censoring status of 3-year DFS is a valid principal surrogate marker for OS in this NCCTG trial. The posterior distributions of EAE and EDE are presented in Figure 4.2.

Table 4.4: Expected associative effect and expected dissociative effect for the NCCTG trial 874651

	Expected Associative Effect	Expected Dissociative Effect
Median	-5.42	-0.36
95% C.I.	(-10.70, -1.90)	(-0.99, 0.29)

Figure 4.2: Distribution of expected associative effect and expected dissociative effect for the NCCTG trial 874651

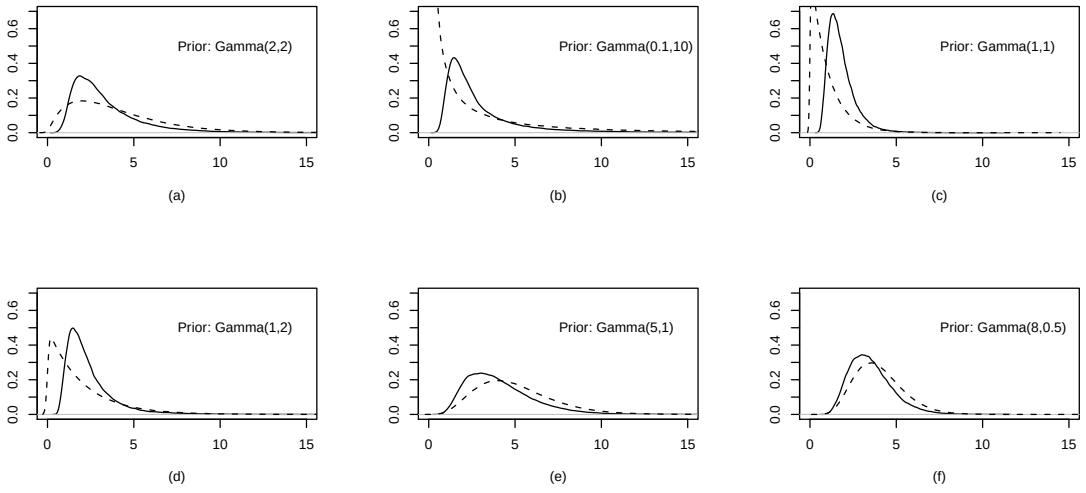


4.3.3 Identifiability and Sensitivity of Results to the Frailty Variance Prior

To investigate the sensitivity of the proposed model to the assumption about the prior for the gamma frailty parameter, we analyze NCCTG trial 874651 assuming prior of gamma frailty parameter η follows gamma distribution with mean and variance (variance of frailty $= \frac{1}{\eta}$) differing for $\mu = 1$, and compare the prior distributions and posterior distributions plotted in Figure 4.3.

When we assume relative uninformative priors of η ($Gamma(2, 2)$ and $Gamma(0.1, 10)$ in Figure 3(a)-(b)), the posterior distributions are both bell shaped with similar posterior means (Mean = 3.33 and Mean = 3.09 respectively). To further investigate identifiability of η, we assume the prior of η follows a gamma distribution, either with a mean value smaller than 3 ($Gamma(1, 1)$ and $Gamma(1, 2)$ in Figure 3(c)-(d)) or larger than 3 ($Gamma(5, 1)$ and $Gamma(8, 0.4)$ in Figure 3(e)-(f)). When the prior of η has a small mean, the mean of posterior distributions shift right (Mean = 1.82 and Mean = 2.39 for $Gamma(1, 1)$ and $Gamma(1, 2)$ respectively). When the prior

Figure 4.3: Prior distributions and posterior distributions of gamma frailty parameter for the NCCTG trial 874651 (solid line denotes posterior distribution, dashed line denotes prior distribution)



of η with a large mean, the mean of posterior distributions shift left (Mean = 4.10 and Mean = 3.53 for $gamma(5,1)$ and $Gamma(8,0.5)$ respectively). Although the posterior distribution of η is moderately sensitive to the choice of the prior, the main results of interest, the principal hazard ratios, are quite stable under a wide variety of prior assumptions, as is shown in Table 4.3.

4.4 Simulation Studies

4.4.1 Data Simulation

We simulate 200 data sets, with 400 subjects in each data set. One half (200) subjects were assigned at random to treatment and 200 to control. The simulation parameters are based on those estimated from our application. The binary surrogate marker is simulated with a baseline multinomial logit model adjusted for baseline covariates, including a continuous variable ($X_1 \sim N(0,1)$) and a categorical variable ($X_2 \sim$ Bernoulli(0.54)), with $\boldsymbol{\zeta} = (\zeta_0^{10}, \zeta_0^{11}, \zeta_{x1}^{10}, \zeta_{x2}^{10}, \zeta_{x1}^{11}, \zeta_{x2}^{11})$ = (-2.5, -1.6,

-2.5, -0.5, -0.1, 0.7). Assuming a monotonic surrogate marker, we have $P(S_{i1} = S_{i0} = 1) = 0.644$, $P(S_{i1} = S_{i0} = 0) = 0.194$, and $P(S_{i1} = 0, S_{i0} = 1) = 0.162$. Frailty for each subject is assumed to follow a gamma distribution with mean of 1 and variance of $\frac{1}{3}$. Given the surrogate marker and frailty, failure times T_{i0} and T_{i1} are independent of each other, and T_{iz} is simulated with the proposed shared gamma frailty PH model assuming Weibull baseline hazard (shape = 2.5, scale = 0.1, $\boldsymbol{\beta} = (\beta_{010}, \beta_{011}, \beta_{100}, \beta_{110}, \beta_{111}, \beta_{x1}, \beta_{x2}) = (0.8, 2.2, 0, -0.8, 2.2, 0.1, -0.5)$), so that the hazard ratio is 1.0 for subjects within strata where the treatment has no impact on the surrogate marker ($S_{i1} = S_{i0} = 1$) and ($S_{i1} = S_{i0} = 0$), and 0.2 for subjects within the stratum where the treatment has a positive effect ($S_{i1} = 1, S_{i0} = 0$).

We simulate censoring time to be 5-year administrative censoring. Note that when generating counterfactual potential primary outcomes, the corresponding censoring distribution is assumed to be noninformative to the event time, and the inference of event times should not be influenced by the corresponding censoring. Therefore, to simplify our computations, we may either assume the counterfactual event times are always observable, or the counterfactual censoring distribution is exactly same as the observed censoring distribution. The latter conceptually provides less information and may be less efficient in computation. Nevertheless, in this simulation study, we adopt the latter convention and observe satisfactory results. Overall, the outcome is censored 17.9% of the time on average on the treatment arm and 24.6% of the time on average on the control arm.

4.4.2 Analysis Results

We analyze each simulated data set with the proposed principal surrogacy model. The principal hazard ratios are summarized in in Table 4.5, including point estimates

Table 4.5: Principal hazard ratios with 100 simulations

	Principal Strata		
	$S_{i1} = S_{i0} = 1$	$S_{i1} = S_{i0} = 0$	$S_{i1} = 1, S_{i0} = 0$
95% Coverage Probability	90%	92%	88%
Median (True Value)	0.98 (1.00)	1.13 (1.00)	0.20 (0.20)
Mean Square Error	0.09	0.60	0.07

of median and repeated sampling coverage of 95% credible interval. The 95% credible intervals associated with the principal hazard ratios undercover slightly, with coverage probabilities ranging from 88% to 92%. The posterior medians of the principal hazard ratios are very close to the true values (we use posterior medians because of the skewness in the posterior distributions). The parameters modeling the marginal distribution of subjects' principal strata membership (ζ), and the parameters modeling the conditional distribution of subjects' primary outcome all have good coverage, ranging from 92% to 98%, and the posterior means are very close to the true values.

The frailty parameter η has coverage probability of 94% with median 2.8 (S.E. = 3.1, true value = 3), showing that the shared gamma frailty parameter is reasonably well identified.

4.5 Conclusions and Future Extensions

In the recent years, surrogacy evaluation in medical studies has aroused much interest, because it can possibly shorten the duration of medical studies and lower medical costs. One drawback in early research of surrogate evaluation method is that the surrogate is a post-randomization variable, and when conditioning on a post-randomization variable, the estimated effect of treatment on the outcome loses a causal interpretation [45]. In this article, we construct a shared gamma frailty PH causal model under the potential outcome framework, using principal stratification approach to evaluate principal surrogacy without destroying the causal interpretation

of the treatment effect on the outcome.

Previous research on principal surrogacy considers binary primary outcomes or normally distributed continuous outcomes. We propose a shared frailty Weibull PH model to evaluate principal surrogacy for time-to-event primary outcomes with a binary surrogate marker. We model the correlation between the potential primary outcomes of the same subject under different treatment arms using a shared gamma frailty, and estimate the model in a Bayesian framework to more easily account for the complex missing data patterns typical of causal inference settings.

When applying the proposed shared gamma frailty PH causal model to the NCCTG trial 874651, we find a strong treatment effect (principal hazard ratio near zero) for the 8% of subjects estimated to benefit from the treatment on the surrogate markers ($S_{i1} = 1, S_{i0} = 0$), and principal hazard ratios not significantly different from 0 for the 92% of subjects who would not benefit from the treatment on the surrogate markers ($S_{i1} = S_{i0} = 1$ or $S_{i1} = S_{i0} = 0$). The estimated EAE with 95% C.I. not covering 0 and EDE with 95% C.I. covering 0 imply the censoring status of 3-year DFS is a valid principal surrogate marker for OS. This finding is similar, but not the same as previous findings by Sargent et al., the 3-year DFS can be used as a (statistical) surrogate marker for overall survival for efficacy evaluation [52]. Note that this result provides clear evidence for the surrogate marker values of 3-year DFS, in contrast to the traditional regression settings, where a somewhat weak treatment effect signal muddles the assessment of surrogate marker value of 3-year DFS.

We model frailty parameter η with gamma priors. Comparison of posterior distributions of η assuming different priors indicates the principal hazard ratio is not sensitive to the choice of the prior for η.

A number of extensions of principal surrogacy for time-to-event outcomes are

possible. For a binary surrogate marker, the CEP^{risk} consists of only 4 points. For continuous or time-to-event surrogate markers, the CEP^{risk} becomes a real surface , with x-axis and y-axis being surrogate marker values under treatment and control arm respectively (for a two arm randomized clinical trial design), and the z-axis being CEP^{risk}.

Another important extension is to extend the current model for a single trial setting to the multiple trial setting. In this setting, hierarchical models with a second level of clusters can be used to accommodate heterogeneity between subjects in different trials.

Our current proposed method may also be extended to accommodate a cured fraction of patients, i.e., cure rate models [54]. Such model considers survival probabilities consisting of a cured fraction and an uncured fraction, which occur often in cancer clinical trials and likely in our application study too (5-year survival greater than 50%). Such a mixture model could thus be developed in conjunction with our proposed method to simultaneously estimate the proportion of cured patients, and to evaluate principal surrogacy. Surrogacy evaluation has the potential to be more accurately estimated by disentangling cured from not-yet-failed patients.

Finally, although it appears to be sufficient for our application of interest, we model frailty with a gamma distribution primarily because of mathematical convenience. Its simple density leads to relatively straightforward estimation using a Gibbs sampler. Identifiability is maintained for a Weibull hazard function even without regressors, as long as the frailty distribution has a finite mean. In the future, we may extend the frailty model to the natural exponential family, such as the inverse Gaussian distribution or positive stable distribution. We would like to extend of the current PH model assuming a parametric Weibull baseline hazard to a semiparamet-

ric Cox PH model where the baseline hazard is estimated nonparametrically. This will require inference based on the partial likelihood, instead of full likelihood in the current model. This also opens up the more general topic of estimating partial likelihood model with missing data in a Bayesian setting, an area where there is little existing research [8, 27, 51].

Appendix

Gibbs Sampler for Shared Gamma Frailty PH Causal Model Estimation

Let $\boldsymbol{X}_i$ denote baseline covariates for subject i, and

$$\theta_{iz} = \exp(\beta_{zs_1s_0} I(S_{i1} = s_1, S_{i0} = s_0, Z_i = z) + \boldsymbol{\beta}_x^{'} \boldsymbol{X}_i).$$

1. Draw of $\eta|$ rest .

   Let $\pi(\eta)$ denote the distribution of η. The posterior distribution of η is

   $$f(\eta|\text{rest}) \propto \left(\prod_{i=1}^{n} W_i\right)^{\eta-1} \eta^{n\eta} \frac{exp\left(-\eta \sum_{i=1}^{n} W_i\right)}{[\Gamma(\eta)]^n} \times \pi(\eta)$$

2. Draw of $\gamma|$ rest .

   We choose the conjugate prior $\pi(\gamma) \sim Gamma(\rho_1, \rho_2)$. The posterior distribution of γ is:

   $$\gamma|\text{rest} \sim Gamma\left(\rho_1 + \sum_{i=1}^{n}\sum_{z=0}^{1} \delta_{iz}, \left(\rho_2^{-1} + \frac{1}{\alpha}\sum_{i=1}^{n}\sum_{z=0}^{1} Y_{iz}^{\alpha}\theta_{iz}W_i\right)^{-1}\right)$$

3. Draw of $\alpha|$ rest .

   Let $\pi(\alpha)$ denote the prior for α. The posterior distribution of α is:

   $$f(\alpha|\text{rest}) \propto \left(\prod_{i=1}^{n}\prod_{z=0}^{1} Y_{iz}^{\delta_{iz}}\right)^{\alpha-1} \exp\left(-\frac{\gamma}{\alpha}\sum_{i=1}^{n}\sum_{z=0}^{1} Y_{iz}^{\alpha}\theta_{iz}W_i\right) \times \pi(\alpha)$$

4. Draw of $\beta|$ rest .

   Let $\pi(\boldsymbol{\beta})$ denote the prior for $\boldsymbol{\beta}$. The posterior distribution of $\boldsymbol{\beta}$ is:

$$f(\boldsymbol{\beta}|\text{ rest }) \propto \exp\left(\boldsymbol{\beta}' \sum_{i=1}^{n}\sum_{z=0}^{1} \delta_{iz}\boldsymbol{X}_{is_1s_0} - \frac{\gamma}{\alpha}\sum_{i=1}^{n}\sum_{z=0}^{1} Y_{iz}^{\alpha}\theta_{iz}W_i\right) \times \pi(\boldsymbol{\beta})$$

   where $\boldsymbol{X}_{is_1s_0}$ denote the hazard function covariates for subject i within principal strata (s_1, s_0).

5. Draw of $\zeta|$rest.

   Let $\pi(\boldsymbol{\zeta})$ denote the prior for $\boldsymbol{\zeta}$. The posterior distribution of $\boldsymbol{\zeta}$ is

$$f(\boldsymbol{\zeta}|\text{ rest }) \propto \exp\left(\prod_{i=1}^{n}\sum_{s_1=0}^{1}\sum_{s_0=0}^{s_1}(\zeta_{s_1,s_0} + \zeta_x^{s_1,s_0}\boldsymbol{X}_i) - \log\left(\sum_{s_1=0}^{1}\sum_{s_0=0}^{s_1}\exp(\zeta_{s_1,s_0} + \zeta_x^{s_1,s_0}\boldsymbol{X}_i)\right)\right)$$
$$\times \pi(\boldsymbol{\zeta})$$

   where $\zeta_{0,0} = 0$ and $\zeta_x^{0,0} = 0$ for identifiability.

   The posterior distributions of η, α, $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are not the closed form of a known distribution; thus we implement a Metroplis random walk algorithm within Gibbs sampler to get random draws from their posterior distributions.

6. Draw of counterfactual surrogate marker $S_{i(1-z_i)}$.

   Let $S_i = (S_{i1}, S_{i0})$, and $Y_i = (Y_{i1}, Y_{i0})$. The distribution of the unobserved

surrogate marker conditional on the observed data and parameters is

$$
\begin{aligned}
f(S_{i(1-z_i)}|\boldsymbol{\zeta},\boldsymbol{\beta},\alpha,\gamma,\eta,T) &= f(S_{i(1-z_i)}|\boldsymbol{\zeta},\boldsymbol{\beta},\alpha,\gamma,\eta,S_{i(z_i)},W_i,Y_i) \\
&\propto f(S_i,W_i,Y_i,\boldsymbol{\zeta},\boldsymbol{\beta},\alpha,\gamma,\eta) \\
&\propto f(S_i|\boldsymbol{\zeta},\boldsymbol{X}_i)f(Y_i|W_i,S_i,\boldsymbol{\beta}) \\
&\propto p_{i,s_1,s_0} \times p_{i,s_1,s_0} \\
&\begin{cases}
\sim \text{BERN}(p_{s_1},p_{s_0}); & z_i=0, s_{z_i}=0 \\
= I(S_{i(1-z_i)}=1)\text{or BERN}(1); & z_i=0, s_{z_i}=1 \\
= I(S_{i(1-z_i)}=0)\text{or BERN}(0); & z_i=1, s_{z_i}=0 \\
\sim \text{BERN}(p_{s_1},p_{s_0}); & z_i=1, s_{z_i}=1
\end{cases}
\end{aligned}
$$

$$
\begin{aligned}
p_{s_1} &= P(s_{i(1-z_i)}=1, s_i(z_i)|\boldsymbol{X}_i,\boldsymbol{\zeta}) \times P(y_i|w_i,s_i,\boldsymbol{\beta}) \\
p_{s_0} &= P(s_{i(1-z_i)}=0, s_i(z_i)|\boldsymbol{X}_i,\boldsymbol{\zeta}) \times P(y_i|w_i,s_i,\boldsymbol{\beta})
\end{aligned}
$$

7. Draw of counterfactual event time $T_{i(1-z_i)}$

   The distribution of the unobserved event time conditional on the observed data and parameter is

$$
\begin{aligned}
f(T_{i(1-z_i)}|\boldsymbol{\zeta},\boldsymbol{\beta},\alpha,\gamma,\eta,T) =& f(T_{i(1-z_i)}|\boldsymbol{\zeta},\boldsymbol{\beta},\alpha,\gamma,\eta,W_i,S_i) \\
=& f(T_{i(1-z_i)}|W_i,S_i,\boldsymbol{\beta}) \\
=& h(T_{i(1-z_i)}|W_i,S_i,\boldsymbol{\beta})S(T_{i(1-z_i)}|W_i,S_i,\boldsymbol{\beta}) \\
=& \gamma T_{i(1-z_i)}^{\alpha-1} W_i e^{\beta_{zs_1s_0} I(S_{i1}=s_1,S_{i0}=s_0,Z_i=1-z_i)+\boldsymbol{\beta}_X' \boldsymbol{X}_i} \times \\
& e^{-\frac{\gamma}{\alpha}T_{i(1-z_i)}^{\alpha}W_i e^{\beta_{zs_1s_0} I(S_{i1}=s_1,S_{i0}=s_0,Z_i=1-z_i)+\boldsymbol{\beta}_X' \boldsymbol{X}_i}}
\end{aligned}
$$

CHAPTER V

Summary and Future Extensions

5.1 Summary of Results

This dissertation focuses on causal modeling under the potential outcome framework using a principal stratification approach to estimate effects of treatment when there is noncompliance or partial compliance in longitudinal randomized clinical trials, and assess principal surrogacy in the time-to-event setting. Under traditional analytic approaches, post-randomization variables such as compliance behavior in randomized clinical trials and dosage tolerance in developmental toxicology studies may confound the estimation of treatment effects. Similarly, assessment of surrogacy is challenging because almost by definition the surrogate marker must be observed after treatment assignment. By framing these problems using the "potential outcome" paradigm we can finesse the post-randomization variable issue by developing "principal strata" that can be viewed as pre-randomization variables, and the causal effects of treatment can be calculated accordingly. Surrogacy can be analogously assessed in the framework of potential outcomes when problems are formulated in causal pathways. A major challenge we face in general is the complex missing data structure, where at least half of the "complete data" are missing. As a result, Bayesian

approaches are employed for estimation and statistical inference.

Chapter II focuses on causal modeling of compliance behavior in longitudinal studies, where compliance behavior and clinical outcome are both measured longitudinally. The uniqueness of the proposed model lies in the joint modeling of longitudinal compliance and outcomes, and their Markov relationship at consecutive time points, such that we can estimate not only the causal effect of the treatment, but also the impact of such causal effects on "future" compliance. We demonstrate the proposed method with a cognitive behavior treatment study for suicide attempters, which shows a substantial increased complier average causal treatment effects relative to an intent-to-treat analysis, and provides evidence that the study subjects could sense the causal effect of treatment and comply if they sense such efficacy.

Chapter III focuses on partial compliance in medical research with escalating dosage schedules, when efficacy of intervention and maximum tolerable dose are both of interest. Such study designs make the actual received dosage a post-randomization variable, and thus a simple comparison of clinical outcomes between randomization groups stratified by treatment-versus-control dosage level received may no longer have a causal interpretation. Within the potential outcome framework, we define pre-randomization "principal strata" by the potential dosage tolerance under treatment, and estimate the treatment effect within the population of subgroups with given dosage tolerance under active treatment. In our proposed causal model, we consider a pair of post-randomization variables, including potential dosage tolerance under all possible treatment assignments and potential adverse event severity under all possible treatment assignment. We then subsequently estimate all model parameters using a Bayesian approach to accommodate the complex missing data structure. The proposed method is applied to an efficacy and safety study for painful bladder

syndrome. We find that the principal stratification analysis suggests that the standard per-protocol analysis may overestimate the causal effect of treatment at higher doses, since the per-protocol analysis includes control subjects who could not tolerate the higher treatment doses, and whose response under control was lower on average, making the treatment effect larger in the as-treated analysis.

In many circumstances, medical studies evaluating the efficacy of an intervention need either long follow-up periods or expensive or invasive procedures to obtain the primary outcome. This motivates the considerable attention to surrogate evaluation in recent years, which aims to use alternative measures ("surrogate marker(s)") in lieu of the primary outcome to evaluate the efficacy of an intervention. However, conventional surrogate evaluation methods fail to provide a causal interpretation, as surrogate markers that are conditioned on in regression are post-randomization variables. Principal surrogacy, defined based on the concept of principal stratification, overcomes such shortcomings. The current literature of principal surrogacy focuses on normally distributed continuous primary outcomes or binary outcomes. In Chapter IV, we propose a shared gamma frailty proportional hazard causal model to study principal surrogacy for time-to-event primary outcomes. The proposed model is constructed under the potential outcome framework using a principal stratification approach, and a gamma frailty model is used to correlate the potential outcomes of an individual under different treatment arms. With the proposed model, we define the principal hazard ratio, expected associative effect and expected dissociative effect to evaluate principal surrogacy. We again use a Bayesian approach to accommodate the complicated missing data structure. We use simulations to study the repeated sampling properties of the proposed model. We illustrate the proposed model and estimation method with a randomized clinical trial of colorectal cancer to study the

censoring status of 3-year disease free survival as the principal surrogate for 5-year overall survival as the primary outcome. We find that the censoring status of 3-year disease free survival can serve as a principal surrogate for 5-year overall survival, as there is little evidence of reduction of risk among subjects whose censoring status of 3-year disease free survival status is estimated not to change as a result of treatment, but there is strong evidence of risk reduction among subjects with censoring status of 3-year disease free survival as a result of treatment.

5.2 Proposed Future Extensions

Many assumptions in causal modeling under the potential outcome framework are untestable in practice, and while are often reasonable based on substantive considerations, may be doubtful in some settings.

- The first subassumption of the SUTVA implies that there is no interference between subjects. This assumption is standard in randomized clinical trials. It may not be satisfied in infectious disease setting when the potential outcomes of different individuals may be dependent on each other.

- The monotonicity assumption adds restriction on the potential outcomes under different treatment arms, and often eliminates the existing possibility of "defier" principal strata, and sometimes is a strong assumption about the nature of the population. When reasonable, this assumption may be replaced with a weaker version of itself, "stochastic monotonicity assumption", which may be implemented through Bayesian estimation method with MCMC algorithms [14].

- The exclusion restriction assumption requires no causal effect of treatment within the principal strata with same value of potential strata variable under

different treatment arms, which may not be satisfied in some cases, such as the CBT study with patients having depression and treated with cognitive therapy. We do not assume exclusion restriction in Chapter II, and found in the end that there is indeed a negative effect to be randomized to the treatment group for not-takers. In surrogate evaluation method, we do not assume exclusion restriction because one goal of surrogacy assessment is to estimate the degree to which the effect of treatment on the surrogate impacts the effect of treatment on the primary outcome.

- The missingness mechanism is assumed to be missing at random through the three projects in this dissertation. Possible extensions include allowing for non-ignorable missingness, such as latent ignorability, which assumes ignorability only within principal strata [42].

The non-identifiability issue is often faced in causal inference due to large amount of missing data relative to a "complete data" framework and its often complex missingness structure. In Chapter II, we rely on the normality assumption to identify the mixture components associated with the complier and not-taker groups in the control arm. An alternative to this approach would be to weaken or eliminate this parametric assumption and rely on either observed predictors of compliance or prior distributional assumptions to induce posterior modes [46]. In Chapter III, identification is also an issue, and is achieved by making a variety of assumptions which, though reasonable and not contradicted by any of the observed data, may not be correct. We could extend this to causal inference settings when full identification is not possibly considered, either by focusing on bounds of causal effects [10], or use of Bayesian methods that do not require identified likelihoods if proper priors are utilized [14]. In Chapter IV, we investigate the identifiability of variance of frailty of

the proposed shared gamma frailty PH causal model in the application by comparing priors for the gamma parameter with posteriors for the gamma parameter. It has been shown that the variance of a frailty can be inferred from a frailty PH model even in the univariate case, as long as the hazard function model has at least one covariate and the covariate is sufficiently variable. In our proposed causal model, we include both the principal strata membership and baseline covariates in the hazard function to protect us from the unidentifiability of the frailty variance.

For surrogacy assessment with causal inference under the potential outcome framework using principal stratification approach, a number of extensions for time-to-event outcomes are desirable, such as extension to continuous or time-to-event surrogate markers with CEP^{risk} becoming a real surface, to a multiple trial setting from the current single trial setting, and to cure models to accommodate a cured fraction of patients, and to semiparametric Cox PH model with missing data and Bayesian estimation method.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] P.K. Andersen, J.P. Klein, K.M. Knudsen, and Palacios R.T. Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, 53(4):1475–1484, 1997.

[3] J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[4] A.J. Barsky, R. Saintfort, M.P. Rogers, and J.F. Borus. Nonspecific medication side effects and the nocebo phenomenon. *Journal of the American Medical Association*, 287(5):622–627, 2002.

[5] T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, and D. Renard. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(4):405–422, 2001.

[6] M. Buyse and G. Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54(3):1014–1029, 1998.

[7] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67, 2000.

[8] M.H. Chen, J.G. Ibrahim, and Q.M. Shao. Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika*, 93(4):791–807, 2006.

[9] P.Y. Chen and A.A. Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

[10] J. Cheng and D.S. Small. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836, 2006.

[11] M.J. Daniels and M.D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16(17):1965–1982, 1998.

[12] B. Efron and D. Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

[13] C. Elbers and G. Ridder. True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies*, 49(3):403–409, 1982.

[14] M.R. Elliott, T.E. Raghunathan, and Y. Li. Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, 11(2):353–372, 2010.

[15] C.E. Frangakis, R.S. Brookmeyer, R. Varadhan, M. Safaeian, D. Vlahov, and S.A. Strathdee. Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association*, 99(465):239–249, 2004.

[16] C.E. Frangakis and D.B. Rubin. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379, 1999.

[17] C.E. Frangakis and D.B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[18] L.S. Freedman, B.I. Graubard, and A. Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11(2):167–178, 1992.

[19] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. CRC press, 2004.

[20] P.B. Gilbert and M.G. Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4):1146–1154, 2008.

[21] P.B. Gilbert, M.L. Peterson, D. Follmann, M.G. Hudgens, D.P. Francis, M. Gurwith, W.L. Heyward, D.V. Jobes, V. Popovic, S.G. Self, F. Sinangil, D. Burke, and P.W. Berman. Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of hiv-1 infection in a phase 3 hiv-1 preventive vaccine trial. *Journal of Infectious Diseases*, 191(5):666–677, 2005.

[22] P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273, 1995.

[23] P. Hougaard. Analysis of multivariate survival data. 2000.

[24] J.G. Ibrahim, M.H. Chen, and D. Sinha. *Bayesian survival analysis*. Springer, 2005.

[25] G.W. Imbens and D.B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25(1):305–327, 1997.

[26] H. Jin and D.B. Rubin. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111, 2008.

[27] J.D. Kalbfleisch. Nonparametric Bayesian analysis of survival time data. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 40(2):214–221, 1978.

[28] W.P. Kennedy. The nocebo reaction. *Medical World*, 95:203–205, 1961.

[29] J.P. Klein. Semiparametric estimation of random effects using the Cox model based on the em algorithm. *Biometrics*, 48(3):795–806, 1992.

[30] M.R. Kosorok, B.L. Lee, and J.P. Fine. Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics*, 32(4):1448–1491, 2004.

[31] Y. Li, J.M.G. Taylor, and M.R. Elliott. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*, 66(2):523–531, 2009.

[32] J.Y. Lin, T.R. TenHave, and M.R. Elliott. Longitudinal nested compliance class model in the presence of time-varying noncompliance. *Journal of the American Statistical Association*, 103(482):462–473, 2008.

[33] J.Y. Lin, T.R. TenHave, and M.R. Elliott. Nested markov compliance class model in the presence of time-varying noncompliance. *Biometrics*, 65(2):505–513, 2009.

[34] R.J.A. Little. Calibrated bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26(2):162–174, 2011.

[35] R.J.A. Little and D.B. Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21(1):121–145, 2000.

[36] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley-Interscience, 2002.

[37] Q. Long, R.J.A. Little, and X. Lin. Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association*, 103:474–484, 2008.

[38] Q. Long, R.J.A. Little, and X. Lin. Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(3):513–531, 2010.

[39] C.G. Moertel, T.R. Fleming, J.S. Macdonald, D.G. Haller, J.A. Laurie, P.J. Goodman, J.S. Ungerleider, W.A. Emerson, D.C. Tormey, J.H. Glick, M.H. Veeder, and J.A. Mailliard. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, 322(6):352–358, 1990.

[40] NCBI/NIH. *Interstitial cystitis*. National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, 2012. Access on Oct. 2012, available at `http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001508`.

[41] M.J. O'Connell, J.A. Mailliard, M.J. Kahn, J.S. Macdonald, D.G. Haller, R.J. Mayer, and H.S. Wieand. Controlled trial of fluorouracil and low-dose leucovorin given for 6 months as postoperative adjuvant therapy for colon cancer. *Journal of Clinical Oncology*, 15(1):246–250, 1997.

[42] Y. Peng, R.J.A. Little, and T.E. Raghunathan. An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics*, 60(3):598–607, 2004.

[43] R.L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4):431–40, 1989.

[44] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[45] P.R. Rosenbaum. The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):656–666, 1984.

[46] J. Roy, J.W. Hogan, and B.H. Marcus. Principal stratification with predictors for compliance for randomized trials with 2 active treatments. *Biostatistics*, 9(2):277–289, 2008.

[47] D.B. Rubin. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58, 1978.

[48] D.B. Rubin. Comment: Randomization analysis of experimental data: the fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[49] D.B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

[50] D.B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–31, 2005.

[51] D.J. Sargent. A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics*, 54(4):1486–1497, 1998.

[52] D.J. Sargent, H.S. Wieand, D.G. Haller, R. Gray, J.K. Benedetti, M. Buyse, R. Labianca, J.F. Seitz, C.J. O'Callaghan, G. Francini, A. Grothey, M. O'Connell, P.J. Catalano, C.D. Blanke, D. Kerr, E. Green, N. Wolmark, T. Andre, R.M. Goldberg, and A.D. Gramont. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology*, 23(34):8664–8670, 2005.

[53] D.S. Small and J. Cheng. Discussion of "identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data". *Biometrics*, 65(3):682–686, 2008.

[54] J.P. Sy and J.M.G. Taylor. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[55] M.A. Tanner and W.H. Wong. An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, 29(1):23–32, 1987.

[56] J.M.G. Taylor, Y. Wang, and R. Thiébaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4):1102–1111, 2005.

[57] T.R. TenHave, M.M. Joffe, K.G. Lynch, G.K. Brown, S.A. Maisto, and A.T. Beck. Causal mediation analyses with rank preserving models. *Biometrics*, 63(3):926–934, 2007.

[58] T.J. VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, 2009.

[59] S. Vansteelandt. Estimating direct effects in cohort and case – control studies. *Epidemiology*, 20(6):851–860, 2009.

[60] L.H.Y. Yau and R.J. Little. Inference for the complier-average causal effect from longitudinal

data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, 96(456):1232–1244, 2001.