Harvard University
Harvard University Biostatistics Working Paper Series

Year 2010 *Paper* 117

Estimating Causal Effects in Trials Involving Multi-treatment Arms Subject to Non-compliance: A Bayesian Frame-work

Qi Long* Roderick J. Little†
Xihong Lin‡

*Emory University, qlong@sph.emory.edu
†University of Michigan, rlittle@umich.edu
‡Harvard University, xlin@hsph.harvard.edu

Estimating Causal Effects in Trials Involving Multi-Treatment Arms Subject to Non-compliance: A Bayesian Framework

Qi Long

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA.

Roderick J. A. Little

Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA.

Xihong Lin

Department of Biostatistics, Harvard University, Boston, MA, 02115, USA.

Summary. Data analysis for randomized trials including multi-treatment arms is often complicated by subjects who do not comply with their treatment assignment. We discuss here methods of estimating treatment efficacy for randomized trials involving multi-treatment arms subject to non-compliance. One treatment effect of interest in the presence of non-compliance is the complier average causal effect (CACE) (Angrist et al. 1996), which is defined as the treatment effect for subjects who would comply regardless of the assigned treatment. Following the idea of principal stratification (Frangakis & Rubin 2002), we define principal compliance (Little et al. 2009) in trials with three treatment arms, extend CACE and define causal estimands of interest in this setting. In addition, we discuss structural assumptions needed for estimation of causal effects and the identifiability problem inherent in this setting from both a Bayesian and a classical statistical perspective. We propose a likelihood-based framework that models potential outcomes in this setting and a Bayes procedure for statistical inference. We compare our method with a method of moment approach proposed by Cheng & Small (2006) using a hypothetical data set, and further illustrate our approach with an application to a behavioral intervention study (Janevic et al. 2003).

Keywords: Causal Inference, Complier Average Causal Effect, Multi-arm Trials, Non-compliance, Principal Compliance, Principal Stratification

1. Introduction

1.1. Non-compliance in Trials Involving Multi-Treatment Arms

Data analysis for randomized controlled trials (RCT) is often complicated by subjects who do not comply with their treatment assignment. Non-compliance in two-arm trials has been extensively studied (Angrist et al. 1996, Imbens & Rubin 1997*a*,*b*, Little & Yau 1998, Peng et al. 2004, Robins 2000). However there has been limited research on how to address non-compliance for trials involving two or more active treatments.

For two-arm randomized intervention trials, Angrist et al. (1996) proposed the complier average causal effect (CACE) as a valid estimand for treatment efficacy, and discussed instrumental variable (IV) methods of estimation. The basic idea is to classify participants as one of compliers(c), defiers(d), never-takers(n), and always-takers(a) according to their potential compliance status upon exposure to an active treatment and a control treatment. The CACE is defined as the average treatment effect for the subpopulation of compliers.

More recently, Frangakis & Rubin (2002) introduced the idea of principal stratification to adjust treatment comparisons for post-treatment variables, including treatment compliance. Any treatment effect defined within one principal stratum or combined principal strata is a valid causal estimand. However, methods for two-arm trials are not directly applicable to trials involving more than two treatments, since the usual identifying assumptions for two-arm trials are not sufficient to point identify the CACE and other causal estimands (Cheng & Small 2006).

Given this lack of identifiability, some have sought upper and lower bounds of the identification region of the parameters (Joffe 2001, Manski 2003, Shafer 1982, Walley 1991). Cheng & Small (2006) proposed bounds on causal effects in three-arm trials subject to non-compliance, using method of moment estimates. To account for sampling uncertainty, they followed Horowitz & Manski (2000) and Beran (1988) and constructed confidence intervals to cover the identification regions of the parameters of interest with fixed probability. This method seems to be restricted to outcomes with finite support, since useful bounds are not available for unbounded outcomes. In addition, it is not trivial to extend their method to more complicated designs, for example, a four-arm trial.

In a seminal paper, Rubin (1978) elucidated the role of randomization in the search for effective treatments and proposed a general Bayesian framework for estimating causal effects. It made clear the role of mechanisms for sampling trial subjects, assigning treatments, and modeling missing data. Imbens & Rubin (1997*a*) applied this framework to the problem of non-compliance in randomized trials, specifically two-arm randomized trials. Their approach clarified the role played by the treatment assignment mechanism and more importantly the complications that arise from the selective receipt of treatment due to possible non-compliance. For trials involving two treatment arms subject to non-compliance, Imbens & Rubin (1997*a*) also discussed situations where relaxing assumptions such as exclusion restriction (ER) and monotonicity (Angrist et al. 1996) leads to causal estimands that are not fully identified. They showed that the issues of identification are quite different from the Bayesian and the classical statistical perspectives, in that with proper prior distributions, posterior distributions are always proper even when the parameters of interest are only partially-identifiable in a classical statistical sense. Imbens & Rubin (1997*a*) also discussed what could be learned in this case using the proposed Bayesian framework. When trials involving multiple treatment arms are subject to non-compliance, we encounter similar yet more complex identifiability problems.

1.2. A Motivating Example

The article is motivated by the Women Take Pride (WTP) study (Janevic et al. 2003). The WTP study involved women aged 60 years and older with diagnosed cardiac disease, who were treated with daily heart medication. This study was conducted to evaluate behavioral intervention programs that were aimed at enhancing the women's ability to manage their disease. In addition to a usual care control treatment, two formats of an behavioral intervention were compared in this study: a Group format, where 6-8 women meet for 2-2.5 hours in a group setting; and a Self-directed format where the participant studies at home following an initial orientation session. Both formats consisted of six weekly units. The same material was presented in the two versions of interventions and only their formats differed. The WTP study utilized a Doubly Randomized Preference Trial (DRPT) design (Long et al. 2008), where some participants are randomized to a treatment in a random arm and some are allowed to choose their treatments in a choice arm. The design

is discussed in details by Long et al. (2008). The random arm is a typical three-arm randomized trial and is the primary motivation for our work. The women in the random arm were randomized to three groups: control, the Group treatment and the Self-Directed treatment. The WTP study was subject to non-compliance. In this paper, compliance was defined as whether a woman completed at least one unit of materials and it was shown that the compliance rates were 76% for both treatments in the Random arm and 100% for the control. Previous analysis has followed the intent-to-treat paradigm, and investigators have been interested in estimating the treatment efficacy after accounting for non-compliance.

In this paper, we propose a Bayesian approach in the spirit of Rubin (1978) and Imbens & Rubin (1997*a*) to estimate causal effects in trials with more than one active treatment that are subject to non-compliance such as the random arm in the WTP study. Roy et al. (2008) recently introduced another useful approach to adjust for noncompliance in trials with two active treatments, where a Bayes procedure was also used for inference. They proposed to directly model marginal distributions of the compliance status under each treatment based on observed data, and the marginal models are then used through a parametric form to construct a model for principal compliance (Little et al. 2009) after incorporating a parameter that captures the association between the marginal distributions and is implicitly assumed be independent of covariates. There are several key differences between our approach and theirs. First, our approach models the principal compliance directly and treat the principal compliance as missing data in the analysis, which avoids the implicit assumption that the association parameter is independent of covariates. While conceptually the approach in Roy et al. (2008) can be extended to trials with more treatment arms, it becomes considerably more complicated and requires more implicit assumptions when one needs to model a distribution of principal compliance indirectly through incorporating association parameters that are independent of covariates with multiple models that are postulated for marginal distributions of compliance within each treatment arm. Also, Roy et al. (2008) limited their discussion to binary outcomes, whereas our approach is developed for general outcomes, continuous or discrete.

In this paper, we focus on a comparison with the method proposed in Cheng & Small (2006) and an attempt to clarify the differences between a Bayesian approach and a classical statistical approach (or a frequentist's approach) in the setting of our interest. The rest of the paper is organized as follows. In Section 2, we introduce principal stratification of a population of interest based on principal compliance status (Little et al. 2009), and define causal estimands of interest; we further discuss structural assumptions and issues related to the identifiability of causal estimands of interest and contrast our Bayesian approach with classical statistical approaches. In Section 3, we propose a likelihood-based framework that models potential outcomes in a trial, and discuss a Bayes inference approach which uses a data augmentation algorithm (DA) (Tanner & Wong 1987) to simulate the posterior distributions of causal parameters, and we compare our approach with the method of moment approach proposed in Cheng & Small (2006) using a hypothetical data set. In Section 4, we illustrate our approach using a behavioral intervention study (Janevic et al. 2003). We make some concluding remarks in Section 5.

2. The Problem

2.1. Principal Compliance and Stratification

For simplicity and illustration purposes, we present our framework using a randomized trial involving two active treatment arms (1 and 2) and one control arm (0), and we will briefly discuss extensions to trials with multi-treatment arms. Let R denote the random treatment assignment ($R = 0, 1, 2$), and $T(r)$ denote the treatment actually received when assigned treatment $R = r$. In full generality, there are 27 principal strata (Frangakis & Rubin 2002) defined by the set of 3^3 possible combinations $(T(0), T(1), T(2))$; all individuals in the population are assumed to belong to one of these strata. All that is observed about the principal strata is the value of $T(r)$ corresponding to the treatment r actually assigned, for each individual in the sample. We thus have a major identifiability problem. We make some assumptions to reduce the scale of this problem. We first assume

ASSUMPTION 1. *Subjects have no access to an active treatment if not assigned to that treatment.*

This is a type of monotonicity assumption in the sense of Angrist et al. (1996), and implies that (1) subjects assigned to control always take the control; and (2) subjects assigned one of the active treatments either take that treatment, or if they fail to comply, take the control treatment. Hence we know $T(0) = 0$, $T(1) = 1$ or 0, and $T(2) = 2$ or 0. This reduces the number of principal compliance strata from 27 to 4, based on subjects' potential compliance status under both active treatments. Following Little et al (2008), we define a principal compliance variable C for these strata, with values $C = 3$ for always-compliers who comply with both treatments ($T(0) = 0$, $T(1) = 1$, $T(2) = 2$), $C = 2$ for 2-only-compliers those who comply when assigned to treatment 2 but do not comply when assigned to treatment 1 ($T(0) = 0$, $T(1) = 0$, $T(2) = 2$), $C = 1$ for 1-only-compliers who comply when assigned to treatment 1 but do not comply when assigned to treatment 2 ($T(0) = 0$, $T(1) = 1$, $T(2) = 0$), and $C = 0$ for noncompliers who do not comply with either active treatment ($T_r = 0$ for $r = 0, 1, 2$). Principal compliance is unobserved in practice, and differs from observed compliance under the assigned treatment. For example, observed compliers in treatment 1 arm are a mixture of always-compliers ($C = 3$) and 1-only-compliers ($C = 1$). Let $\rho_c = \text{Prob}(C = c)$ denote the proportion of the population in principal compliance stratum c.

Consider a study with n subjects. For each subject i, let $Y_i(\mathbf{R}, \mathbf{T})$ denote the potential responses under randomization R and treatment receipt T, where $\mathbf{R}$ and $\mathbf{T}$ are the randomization assignment and treatment received for all subjects. We also let $\mu_{c,r,t}$ denote the expected value of Y in principal compliance stratum c when treatment $R = r$ is assigned and treatment $T = t$ is received (Table 1).

2.2. Structural Assumptions and Causal Estimands

In addition to Assumption 1, we consider several other structural assumptions.

ASSUMPTION 2. *Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978). The treatment receipt $T(r)$ and the outcome Y for subject i are not affected by the treatment assignments for other subjects.*

Given SUTVA, $Y_i(\mathbf{R}, \mathbf{T})$ can be written as $Y_i(R_i, T_i)$.

ASSUMPTION 3. *Exclusion Restriction (ER) (Angrist et al. 1996). The effect of treatment assignment R on the outcome Y is entirely through the effect of treatment receipt T on Y.*

Given SUTVA and ER assumption, we have $Y_i(r,t) = Y_i(r',t)$, and

$$\begin{cases} \mu_{0,0,0} = \mu_{0,1,0} = \mu_{0,2,0} \\ \mu_{2,0,0} = \mu_{2,1,0} \\ \mu_{1,0,0} = \mu_{1,2,0} \end{cases}$$

Hence, we can write $\mu_{c,t}$ instead of $\mu_{c,r,t}$ and $Y_i(T_i)$ instead of $Y_i(R_i, T_i)$, and Table 1 reduces to Table 2.

Following Angrist et al. (1996), we consider a comparison of mean outcomes to be *causal* if it compares means in the same population or subpopulation. Direct comparison of mean outcomes for observed compliers in treatment arm 1 and in treatment arm 2 is not causal without assumptions, because the observed compliers in each treatment arm are a mixture of two different principal compliance strata and hence are not from the same subpopulation. On the other hand, CACEs compare mean outcomes under two different treatments within a same principal compliance stratum, and hence are causal. In the setting of one treatment arm and 1 control arm, the CACE is uniquely defined as the difference in means between active treatment and control in the population of principal compliers. In our setting, a number of interesting CACEs can be defined. Three of particular interest are: 1) $\text{CACE}_{12} = \mu_{3,1} - \mu_{3,2}$, the CACE for comparing treatment 1 to treatment 2 for always-compliers(C=3); 2) $\text{CACE}_1 = (\rho_3\mu_{3,1} + \rho_1\mu_{1,1})/(\rho_3 + \rho_1) - (\rho_3\mu_{3,0} + \rho_1\mu_{1,0})/(\rho_3 + \rho_1)$, the CACE for comparing treatment 1 to control for always-compliers (C=3) and 1-only-compliers (C=1); 3) $\text{CACE}_2 = (\rho_3\mu_{3,2} + \rho_2\mu_{2,2})/(\rho_3 + \rho_2) - (\rho_3\mu_{3,0} + \rho_2\mu_{2,0})/(\rho_3 + \rho_2)$, the CACE for comparing treatment 2 to control for always-compliers(C=3) and 2-only-compliers(C=2). CACE_1 and CACE_2 are equivalent to those defined in (Angrist et al. 1996) for comparing treatment 1 vs control and treatment 2 vs control, respectively. CACE_{12}, however, is a new causal estimand. A simple approach to the three-arm problem is to estimate CACE_1 and CACE_2 using previously developed methods for comparing an active treatment to the control, and then compare CACE_1 and CACE_2. However, that comparison is not causal without assumptions, because CACE_1 and CACE_2 refer to different subpopulations. Methods for estimating CACE_{12} are more complex, but arguably CACE_{12} is the appropriate causal estimand, since a causal comparison of efficacy is only possible on the subpopulation of individuals who comply with both treatments. We note that Cheng & Small (2006) provided some discussion of the use of CACE_{12}. Other causal treatment effects can also be defined, for example, $\mu_{3,2} - \mu_{3,0}$, $\mu_{3,1} - \mu_{3,0}$ $\mu_{1,1} - \mu_{1,0}$ and $\mu_{2,2} - \mu_{2,0}$, but we view these as of secondary interest.

The relevance of a causal treatment effect in principal stratum $C = c$ increases with the proportion of the whole population that belongs to this principal stratum, that is, ρ_c. In particular when ρ_c is close to 0, the causal effect relates to a small part of the population and may not be considered of much interest. In some circumstances, we may be able to conjecture that a particular ρ_c is close to zero and therefore negligible, for example, if treatment 1 has less significant side effects than treatment 2 and the side effects are the sole reason for non-compliance, then the following assumption may be valid,

ASSUMPTION 4. $\rho_2 = 0$*, that is, subjects who would comply with treatment 2 would always comply with treatment 1.*

This assumption is also a type of monotonicity assumption in the sense of Angrist et al. (1996). We will see that when one or more principal stratum proportions are close to 0, estimation of valid causal effects is simplified and more informative results may be obtained. Hence, in practice it is important to identify situations where particular population proportions may be assumed negligible. In the WTP study, it is unclear whether Assumption 4 holds, therefore we will conduct a sensitivity analysis for the WTP study with or without Assumption 4.

2.3. Identifiability of Causal Estimands

We first define the point-identifiability or lack thereof in a classical statistical sense, that is, parameter(s) are not point identifiable if $F_{\theta_1} = F_{\theta_2}$ where F_θ is the probability distribution of the observables indexed by θ and θ_1 and θ_2 are two different values of θ. It has been long recognized in many settings classical statistical methods may have difficulties dealing with non-identifiable or partially-identifiable parameters (Balke & Pearl 1997, Cheng & Small 2006, Manski 2003, Neath & Samaniego 1997). In particular, Cheng & Small (2006) studied a similar design as ours, and they showed that the treatment effects within basic principal strata are only partially identified under certain assumptions, which means that given an unlimited number of observations, one could only place the parameter of interest in a set-valued identification region, where the values within this set (region) can not be distinguished based on the observables and the set is a strict subset of the parameter space. Specifically in the setting of our interest, under Assumption 1-3, there are 8 marginal means that are of interest (Table 2), and none of which is point identifiable; hence, all causal treatment effects discussed in Section 2.2 are only partially identifiable. For example, multiple values of CACE_{12} may lead to the same maximized observed data likelihood (Long 2005) or solve the same set of estimating functions, equation (1)-(4) in Cheng and Small (2006), and usually these values form a set-valued interval. Under Assumption 1-4, μ_{22} and μ_{20} are no longer applicable, hence the number of marginal means in Table 2 is reduced to six. It can also be shown that in this case $\mu_{32} - \mu_{30}$ becomes point-identifiable and the rest of causal estimands of interest remain not point identifiable.

Following Shafer (1982), Walley (1991), and Horowitz & Manski (2000), Cheng & Small (2006) argued that when a causal parameter of interest is partially identifiable, the identification region can be used as a way to conduct inference; and they also provided confidence intervals that cover the entire identification region with fixed probability. Alternatively, Imbens & Manski (2004) developed methods to construct confidence intervals that asymptotically cover the true value of the parameter with fixed probability, and showed that the confidence intervals for the identification region are wider than the confidence intervals for the true value of the parameter. Hence, the confidence intervals for the identification region, when used as the confidence intervals for the true value of the parameter, are likely to be conservative compared to the nominal level of coverage. However, Imbens & Manski (2004) did so in a considerably simpler setting and it is not trivial to extend their methods to the setting of our interest (Cheng & Small 2006).

In the above settings with non-identifiable or partially identifiable parameters, often times identifiability is a less serious issue with a Bayesian framework, one can still make interpretable inference using a Bayesian approach (Gustafson 2005, Imbens & Rubin 1997*a*, Lindley 1971, Neath & Samaniego 1997). In general, if the posterior distributions are proper, the usual Bayesian framework is valid and its credible intervals still bear their usual interpretation. Trials with multiple treatment arms are one of these settings. Hence, in

this setting a Bayesian approach has the potential to provide narrower confidence intervals and achieve more power, which makes a Bayesian approach more attractive. Even in the presence of potential improper posteriors, it is still possible to obtain meaningful results using a Bayesian approach (Gelfand & Sahu 1999).

We note another important difference between classical statistical methods such as a maximum likelihood (ML) approach and a Bayesian approach in multiple-parameter settings. When there are multiple parameters, the ML estimate(s) of one parameter are the value(s) that maximize the observed data likelihood jointly with ML estimates of the other parameters; whereas the marginal posterior distribution of one parameter is obtained by integrating out the other parameters. In other words, if we assume $L(\theta_1, \theta_2|\text{data})$ is the observed data likelihood with θ_1 denoting the parameter of interest and θ_2 denoting the other parameters, then the ML estimate of θ_1 maximizes the profile likelihood $L(\theta_1, \hat{\theta}_2(\theta_1)|\text{data})$ and the marginal posterior distribution of θ_1 with a prior $p(\theta_1, \theta_2)$ is proportional to $\int L(\theta_1, \theta_2|\text{data})p(\theta_1, \theta_2)d\theta_2$. Hence, the mode (or a region of modes) of the posterior distribution of one parameter (say, θ_1) does not necessarily correspond to its ML estimate (or a region of ML estimates), even if flat priors are used. Its 95% Bayesian credible interval can be quite different from its 95% ML confidence interval. While this is unlikely to happen when all parameters are point identifiable, this can happen when some parameters are only partially identifiable, When parameters are partially identifiable, there is usually a ridge or a plateau in the observed data likelihood surface (joint likelihood), which, however, may disappear after marginalizing the likelihood with respect to a subset of the parameters. In other words, in the presence of non-identifiability a simple step of marginalizing in a Bayesian analysis may have more profound impact on the statistical inference than it initially appears. We suspect that this marginalization step also helps produce narrower confidence intervals compared to those from a classical statistical approach. On the other hand, it is not obvious how to marginalize in a sensible way within the classical statistical framework.

3. A Bayesian Framework

In this section, we present a Bayesian framework for estimating causal parameters of interest such as CACE_{12}, for randomized trials involving two active treatment arms and one control arm. Throughout this section, we make Assumption 1-3. We first introduce some additional notation.

3.1. Notation

Following previous notation, for subject i, let R_i denote the random treatment assignment (2/1/0), C_i denote the true principal compliance stratum, where it takes a value of 0 for non-compliers, 1 for 1-only-compliers, 2 for 2-only-compliers, 3 for always-compliers, T_i denote the treatment actually received which is uniquely determined by C_i and R_i, that is, $T_i(C_i, R_i)$. Let Y_i denote the observed outcome for subject i, and $(Y_i(2), Y_i(1), Y_i(0))$ denote the potential outcome when the actual treatment received is 2, 1 and 0, respectively. We also let X_i denote a set of covariates that may be associated with the potential outcomes or the principal compliance status.

In a real trial, for each subject i $(i = 1, \ldots, n)$, we only get to observe the treatment assignment (R_i), and the treatment receipt given that particular treatment assignment T_i, one potential outcome $(Y_i = Y(T_i))$ and X_i. When subject i is not assigned to specific

active treatment, then its compliance status to that treatment is not observed. Hence we do not observe the principal compliance status C_i, in other words, C is a latent classification variable and is always missing in our setting. We note that C_i is observable in some other settings, for example, in a two arm trial (Little et al. 2009). However, since T_i is uniquely determined by R_i and C_i, the observed values of R_i and T_i may limit the feasible values of C_i, and we denote this set of feasible values by $C_{obs,i}$. For example, subjects with $R = 1$ and $T = 1$ can only belong to either $C = 1$ or $C = 3$ principal strata but not to $C = 2$ or $C = 0$ principal strata, and then $C_{obs,i} = \{1, 3\}$. Also, if subject i does not actually receive a treatment, then its potential outcome given that treatment is not observed. We note that i may be suppressed in our notation wherever it does not lead to confusion.

We define the complete data as $(Y_i, C_i, R_i, T_i, X_i)$ with $i = 1, \ldots, n$, which under Assumption 2 (SUTVA) constitute an independent and identically distributed sample. Then the observed data can be represented as $(Y_i, C_{obs,i}, R_i, T_i, X_i)$. Our objective is to relate the distribution of first the complete data $(Y_i, C_i, R_i, T_i, X_i)$ and then the observed data $(Y_i, C_{obs,i}, R_i, T_i, X_i)$ to the distribution of the potential outcomes $(Y_i(2), Y_i(1), Y_i(0))$. Thus, using the observed data, we can estimate the parameters associated with the distribution of the potential outcomes $(Y_i(2), Y_i(1), Y_i(0))$, which should bear causal interpretations.

3.2. Likelihood of the Data

For subject i, the distribution function of the complete data is

$$f(Y_i, C_i, R_i, T_i|X_i) = f(Y_i, C_i, T_i|R_i, X_i)f(R_i|X_i)$$

Since $f(R_i|X_i)$ is the treatment assignment model and is known due to the design, we can ignore the treatment assignment model and just focus on the $f(Y_i, C_i, T_i|R_i, X_i)$ in the statistical inference. Furthermore, we have

$$f(Y_i, C_i, T_i|R_i, X_i) = f(Y_i|C_i, T_i, R_i, X_i)f(T_i|C_i, R_i, X_i)f(C_i|R_i, X_i)$$

Since T_i is uniquely determined by C_i and R_i, $f(T_i(C_i, R_i)|C_i, R_i, X_i) = 1$. Due to the ER assumption and the random treatment assignment, $f(Y_i|C_i, T_i, R_i, X_i) = f(Y_i|C_i, T_i(C_i, R_i), X_i) = f(Y(T_i) = Y_i|C_i, X_i)$, which indicates that $f(Y_i|C_i, R_i, X_i)$ is determined by a model for the potential outcome $Y(T_i)$. Let α denote the set of parameters associated with the potential outcome model, that is, $f(Y(T_i) = Y_i|C_i, X_i, \alpha)$. Also due to the random treatment assignment, we have $f(C_i|R_i, X_i) = f(C_i|X_i, \beta)$, where β denotes the set of parameters associated with the model for the principal compliance C. Assume that α and β are distinct, and let $\theta = (\alpha, \beta)$. Given the exchangeability and independence among subjects, the complete data likelihood can be written as

$$\prod_i^n f(Y(T_i) = Y_i|C_i, X_i, \alpha)f(C_i|X_i, \beta) \tag{1}$$

where the first part models the potential outcomes $Y(t)$ and the second part models the principal compliance C. The observed data likelihood can be written as

$$L(\alpha, \beta) = \prod_i^n L_i(\alpha, \beta|Y_i, T_i, C_{obs,i}, X_i) \tag{2}$$

We now examine the observed data likelihood due to subject i, L_i. Based on our previous discussion for $C_{obs,i}$, it is straightforward to show that for subject i, the observed data $(Y_i, T_i, C_{obs,i}, X_i)$ follows a mixture distribution and the observed data likelihood is

$$L_i(\alpha, \beta | Y_i, T_i, C_{obs,i}, X_i) = \sum_{c \in C_{obs,i}} f(Y(T_i) = Y_i | C_i, X_i, \alpha) f(c | C_{obs,i}, X_i, \beta)$$

Hence, the observed data likelihood (2) is a product of different mixture distributions and only depends on the conditional distributions of potential outcomes $Y(t)$, $f(Y(t)|C, X, \alpha)$ $(t = 1, 2, 3)$, rather than the joint conditional distribution of $(Y(2), Y(1), Y(0))$. We shall see that the causal estimands of interest are only related to the parameters associated with these marginal distributions.

Let $f_{ct}(Y|X, \alpha_{ct})$ denote $f(Y(t) = Y|C = c, X, \alpha_{ct})$, that is, the conditional distribution of the potential outcome $Y(t)$ for subjects in principal compliance stratum $C = c$, where α_{ct} denotes a set of parameters associated with this distribution and α is then the collection of all α_{ct} that can be estimated from the data. Under Assumption 1-3, we know from Table 2 that $\alpha = (\alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{22}, \alpha_{20}, \alpha_{11}, \alpha_{10}, \alpha_{00})$ and the rest of α_{ct}'s are not applicable. The conditional distributions in (2) can be replaced by $f_{ct}(Y_i|X_i, \alpha_{ct})$. In addition, for each subject i, let $\rho_{i,c} = f(C_i = c|X_i, \beta)$ and $f_{i,ct} = f_{ct}(Y_i|\underline{x}_i, \alpha_{ct})$, and denote by $S(r, t)$ the set of subjects with $R = r$ and $T = t$. Under Assumption 1-3, Table 3 summarizes the structure of the observed data likelihood. The row totals are proportional to the contribution of subject i to the observed data likelihood, which accounts for all feasible values in $C_{obs,i}$ given the observed T and R and hence are from different mixture distributions. Given the observed data for subject i, each cell value represents the probability of the observed data (Y_i, X_i, T_i, R_i) when C_i is known. A value of 0 in a cell indicates that the corresponding value of C_i is not feasible based on the combination of observed R_i and T_i values. For example, when $R = 1$ and $T = 1$, the probability of $C = 2$ or $C = 0$ is 0. Then, the observed data likelihood (2) can be rewritten as follows

$$\begin{aligned} L(\beta, \alpha) \quad = \quad & \prod_{i \in S(1,1)} \frac{\rho_{i,3} f_{i,31} + \rho_{i,1} f_{i,11}}{\rho_{i,3} + \rho_{i,1}} \times \prod_{i \in S(1,0)} \frac{\rho_{i,2} f_{i,21} + \rho_{i,0} f_{i,01}}{\rho_{i,2} + \rho_{i,0}} \\ & \times \prod_{i \in S(2,2)} \frac{\rho_{i,3} f_{i,32} + \rho_{i,2} f_{i,22}}{\rho_{i,3} + \rho_{i,2}} \times \prod_{i \in S(2,0)} \frac{\rho_{i,1} f_{i,12} + \rho_{i,0} f_{i,02}}{\rho_{i,1} + \rho_{i,0}} \\ & \times \prod_{i \in S(0,0)} \frac{\rho_{i,3} f_{i,30} + \rho_{i,2} f_{i,20} + \rho_{i,1} f_{i,10} + \rho_{i,0} f_{i,00}}{\rho_{i,3} + \rho_{i,2} + \rho_{i,1} + \rho_{i,0}} \end{aligned} \tag{3}$$

Let $p(\theta)$ denote the prior distribution of $\theta = (\alpha, \beta)$, and then the posterior distribution of θ given the observed data is

$$\begin{aligned} f(\theta | R, T, Y, C_{obs,i}, X) \quad \propto \quad & p(\theta) \times \prod_{i \in S(1,1)} \frac{\rho_{i,3} f_{i,31} + \rho_{i,1} f_{i,11}}{\rho_{i,3} + \rho_{i,1}} \times \prod_{i \in S(1,0)} \frac{\rho_{i,2} f_{i,21} + \rho_{i,0} f_{i,01}}{\rho_{i,2} + \rho_{i,0}} \\ & \times \prod_{i \in S(2,2)} \frac{\rho_{i,3} f_{i,32} + \rho_{i,2} f_{i,22}}{\rho_{i,3} + \rho_{i,2}} \times \prod_{i \in S(2,0)} \frac{\rho_{i,1} f_{i,12} + \rho_{i,0} f_{i,02}}{\rho_{i,1} + \rho_{i,0}} \\ & \times \prod_{i \in S(0,0)} \frac{\rho_{i,3} f_{i,30} + \rho_{i,2} f_{i,20} + \rho_{i,1} f_{i,10} + \rho_{i,0} f_{i,00}}{\rho_{i,3} + \rho_{i,2} + \rho_{i,1} + \rho_{i,0}} \end{aligned} \tag{4}$$

It is obvious that the posterior distributions in (4) are proper. If we make Assumption 4 in addition to Assumption 1-3, then the observed data likelihood can be further simplified. Specifically, we can remove the column for $C_i = 2$ in Table 3 and the distribution of the observed data for a subject i is no longer a mixture distribution when $R = 1$ and $T = 0$, or when $R = 2$ and $T = 2$. In other words, we can change Table 3 and hence the observed data likelihood (2) and (3) accordingly when more or less assumptions are made.

It is straightforward to show that the marginal means defined in Table 2 and hence causal estimands of interest discussed in Section 2.2 can be expressed in terms of the parameters α_{ct}'s. For example, $\mu_{32} = \int\int Y f_{32}(Y|X, \alpha_{32}) d\nu(X) dY$ and $\mu_{31} = \int\int Y f_{31}(Y|X, \alpha_{31}) d\nu(X) dY$, and hence

$$\text{CACE}_{12} = \int\int Y f_{32}(Y|X, \alpha_{32}) d\nu(X) dY - \int\int Y f_{31}(Y|X, \alpha_{31}) d\nu(X) dY$$

where $\nu(X)$ is a probability measure on X. Hence we need to make inference about α_{ct}'s.

3.3. Estimation and Inference

Generally speaking, the posterior distribution of θ in (4) is mathematically not complicated, but its computation is complicated due to the fact that it involves mixture distributions. If C were observed, then the observed data likelihood would no longer involve mixture distributions and could be easily simulated. This leads us to employ a data augmentation (DA) algorithm (Tanner & Wong 1987) to simulate the posterior distributions in (4), which treats C as missing data when approximating the posterior distributions. This data augmentation algorithm is iterative and alternates between two steps, the I-step and the P-step, where I stands for imputation and P stands for drawing from the posterior distribution. The data augmentation algorithm can be outlined as follows:

(a) I-step: For each subject i, impute C_i for the "complete data" $(C_i, R_i, T_i, Y_i, X_i)$ using a draw. Specifically, given $(C_{obs,i}, R_i, T_i, Y_i, X_i)$ and θ drawn from current approximation to its posterior distribution, C_i is drawn from a multinomial distribution with sample size equal to 1 based on the conditional probabilities, $f(C|C_{obs,i}, R_i, T_i, Y_i, X_i)$. These conditional probabilities can be computed from Table 3 using the ratio of each cell probability to its row total.

(b) P-step: Given the imputed "complete data" $(C_i, R_i, T_i, Y_i, X_i)$, the posterior distribution becomes

$$f(\theta|C_i, R_i, T_i, Y_i, X_i) \propto p(\theta) \prod_{t=0,1,2} \prod_{c=(0,1,2,3)} \left\{ \prod_{C_i=c, T_{obs,i}=t} \rho_{i,c} f_{i,ct} \right\} \tag{5}$$

If we assume that α are independent of β, then we have

$$f(\beta|C_i, R_i, T_i, Y_i, X_i) \propto p(\beta) \prod_{c=(0,1,2,3)} \left\{ \prod_{C_i=c} \rho_{i,c} \right\}, \tag{6}$$

and

$$f(\alpha_{ct}|C_i, R_i, T_i, Y_i, X_i) \propto p(\alpha_{ct}) \prod_{C_i=c, T_i=t} f_{i,ct} \tag{7}$$

for all feasible values of c and t with $c = 0, 1, 2, 3$ and $t = 0, 1, 2, 3$.

To complete this algorithm, one needs to specify the priors of θ and we propose to choose $p(\theta)$ as general as possible such as flat priors, while they are still proper and conjugate to the likelihood in the P-step when possible. The P-step can then be implemented using a Gibbs sampler. The examples of these priors in some special cases can be found in the next two sections. To draw the posterior distributions, one needs to iterate between I step and P-step until the algorithm converges.

In case of no covariates adjustment, $\rho_{i,c} = \rho_c$, and β in the likelihood (3) and (4) can be replaced with $\rho = (\rho_1, \rho_2, \rho_3, \rho_4)$. Then $p(\theta)$ is the prior distribution of $\theta = (\alpha, \rho)$. While the I-step in the DA algorithm does not change, the posterior distributions (5)-(7) in the P-step simplify to the following:

$$f(\theta|C_i, R_i, T_i, Y_i, X_i) \propto p(\theta) \prod_{t=0,1,2} \prod_{c=(0,1,2,3)} \left\{ \prod_{C_i=c, T_{obs,i}=t} \rho_c f_{i,ct} \right\}$$

and

$$f(\rho|C_i, R_i, T_i, Y_i, X_i) \propto p(\rho) \prod_{c=0,1,2,3} \rho_c^{N_c},$$

where N_c is the number of the subjects in principal stratum $C = c$.

The proposed approach relies on the structural assumptions and the specification of priors. One can incorporate prior experience or preliminary results to determine the selection of the assumptions and the choice of priors whenever these information is available. When no prior information is available, we propose to conduct additional sensitivity analysis to examine the influence of assumptions and priors. First, one can conduct inference under different combinations of assumptions, and examine how the causal estimands of interest vary. For example, one can consider the inclusion or exclusion of Assumption 4 and its variants for other principal compliance stratum such as $\rho_1 = 0$ or $\rho_0 = 0$. In this case, the comparison should be focused on the causal estimands that remain applicable under these assumptions, such as CACE_{12}. Second, one can conduct inference under different specification of priors. One could stay with conjugate priors in the P-step and compare the results using different parameter values for these conjugate priors including the flat priors.

It is straightforward to extend the proposed framework to the estimation of causal estimands, in particular, CACE_{12}, for trials with multi-treatment arms. One can introduce more principal compliance strata and marginal means to Table 2 and 3, and define causal estimands of interest similar to those in Section 2.2. The observed data likelihood similar to (2) and (3) can then be constructed following the discussion in this section; hence a Bayesian inference approach can also be conducted.

3.4. A Hypothetical Example

We now compare our approach with a method of moment approach in Cheng & Small (2006) using a hypothetical example that was analyzed in Cheng & Small (2006). We shall see that the results suggest that our proposed method can obtain meaningful inference even when the treatment effects are only partially identifiable in a classical statistical sense. The data structure was detailed in Table 2 in Cheng & Small (2006). Suppose that we have a three-arm trial with a sample size of $n = 400$ in each arm, the control treatment denoted by 0 and two active treatments denoted by A and B, with a binary outcome, Y, and Y

equals to 1 for a successful outcome and 0 for a failure. All of those assigned to the control arm actually take control, among which 45% have successful outcomes ($Y = 1$). For those assigned to treatment A, 95% of the subjects actually take treatment 1 among which 95% have successful outcomes, and 5% actually take control, among which 20% have successful outcomes. For those assigned to treatment B, 80% of the subjects actually take treatment B among which 70% have successful outcomes, and 20% actually take the control among which 25% have successful outcomes. To make the notation consistent, we use 1 for A and 2 for B in our illustration.

We used the model described in Section 3 without covariates adjustment. Since the outcomes were binary, we assumed for subjects in principal stratum $C = c$ and treatment receipt $T = t$, the outcome success rate (Y) and principal compliance (C)

$$\begin{aligned} Y|C = c, T = t; \alpha_{ct} &\sim \text{Bernoulli}(\alpha_{ct}) \\ C|\rho &\sim \text{Multinomial}(\rho_0, \rho_1, \rho_2, \rho_3) \end{aligned}$$

where α_{ct} represents the probability of success for subjects in principal compliance stratum $C = c$ when taking treatment $T = t$. We used the following conjugate priors $\alpha_{ct} \sim \text{Beta}(a, b)$ and $\rho = (\rho_0, \rho_1, \rho_2, \rho_3) \sim \text{Dirichlet}(b_0, b_1, b_2, b_3)$ in our Bayesian inference, where values of (a, b) and (b_0, b_1, b_2, b_3) determine how informative these priors are. For this data analysis, we also conducted a sensitivity analysis using different parameter values for these conjugate priors. Specifically, let $a = b = b_0 = b_1 = b_2 = b_3 = \lambda$, where λ may take different values. When $\lambda = 1$, then uninformative flat priors are assumed for all parameters of interest. Given the model specifications, causal estimands of interest are then functions of μ_{ct}, for example, $\text{CACE}_{12} = \mu_{32} - \mu_{31}$. The data augmentation algorithm can be described as follows

(a) I-step: Given a draw of ρ, and α from their current approximate distribution and observed data, draw C_i for each i from a multinomial distribution with sample size equal to 1 with conditional probabilities computed using a simplified version of Table 3.
(b) P-step: Given observed data and current C_i drawn from the I-step,

$$\begin{aligned} \rho|\text{observed data}, C &\sim \text{Dirichlet}(n_0 + b_0, n_1 + b_1, n_2 + b_2, n_3 + b_3) \\ \alpha_{ct} &\sim \text{Beta}(m_{ct} + a, n_{ct} - m_{ct} + b) \end{aligned}$$

where n_c is the number of subjects in stratum $C = c$, n_{ct} is the number of subjects with $C = c$ and $T = t$, and m_{ct} is the number of successes with $C = c$ and $T = t$.

We used the DA algorithm to approximate the posterior distributions of the causal parameters for the hypothetical data. The approximate posterior distributions were obtained using 12,000 iterations from each of 20 independent runs of the DA algorithm after the first 10,000 iterations were discarded and each run started with different initial values drawn from uniform distributions over the range of the parameters. This scheme was used for all data analyses discussed in this paper. The 95% credible intervals were constructed from the marginal posterior distributions of parameters of interest.

We considered inference under two settings: one with Assumption 1-3 and the other with Assumption 1-4. As discussed previously, given Assumption 1-3, none of the causal treatment effects are point identifiable in the classical statistical sense; after adding Assumption 4 ($\rho_2 = 0$), only $\mu_{32} - \mu_{30}$ is point-identifiable in the classical statistical sense and

$\mu_{22} - \mu_{20}$ is no longer applicable. The results from our analysis are summarized in Table 4 for different prior specifications and two sets of assumptions.

Cheng & Small (2006) analyzed this hypothetical data set using a method of moment approach and presented the results in their Table 4. This hypothetical data set was also analyzed using a maximum likelihood (ML) inference approach in Long (2005), where the identification regions were obtained and their confidence intervals were constructed based on 5000 bootstrap samples (Horowitz & Manski 2000). The identification regions using the ML approach were similar to those in Cheng & Small (2006) and their confident intervals were constructed to cover the identification regions with fixed probability in the spirit of Horowitz & Manski (2000) and Cheng & Small (2006) . The results found in Long (2005) were very close to those found in Cheng & Small (2006). Their results show that in general the addition of Assumption 4 shortens the identification regions as well as their confidence intervals, however the improvement is small. Under Assumption 1-3, the identification region for $\mu_{31} - \mu_{30}$ is (0.41,0.51) with a confidence interval of (0.34,0.58); under Assumption 1-4, the identification region changes to (0.44, 0.50) with a confidence interval of (0.37,0.57). Under Assumption 1-3, the identification region for $\mu_{11} - \mu_{10}$ is (0.39,0.79) with a confidence interval of (0.22,0.96); under Assumption 1-4, the identification region changes to (0.42, 0.73) with a confidence interval of (0.23,0.92) . More importantly, under Assumption 1-3, the identification region for $\mu_{32} - \mu_{10}$ is (0.16,0.23) with a confidence interval of (0.06,0.32); under Assumption 1-4, $\mu_{32} - \mu_{30}$ becomes point identifiable with an estimate of 0.20 and a confidence interval of (0.11,0.29). Our proposed Bayesian analysis shows similar trends, and our results also show that in terms of estimating $\mu_{32} - \mu_{30}$, our analysis without Assumption 4 is just as informative as with Assumption 4.

Compared to the results obtained using the flat priors ($\lambda = 1$), our sensitivity analysis shows that different prior specifications have minimal impact on the causal estimands defined in the principal compliance straum $C = 3$, that is, $\mu_{32} - \mu_{31}$, $\mu_{32} - \mu_{30}$, and $\mu_{31} - \mu_{30}$. However, the specification of priors has various degrees of impact on the causal estimands defined in the principal stratum $C = 2$ and $C = 1$. The prior close to the flat prior ($\lambda = 0.5$) has less impact than the priors that are strongly informative ($\lambda = 10$). The standard deviation of the posterior distribution and the width of credible intervals decreases considerably as a result of strong informative priors, hence the power of the analysis improves. Since the compliance rate is high in this study, the proportion of always-compliers ($C = 3$) is likely to be high whereas the proportion of other principal strata is likely to be low. In the stratum where the number of subjects is low, strong priors may dominate the observed data and have substantial impact on the causal estimands, in this case, those for $C = 1$ and $C = 2$, which is consistent with what we observed in this data analysis.

In general, when causal estimands are partially identifiable, our Bayesian 95% credible intervals under both set of assumptions and different prior specifications are considerably narrower than the corresponding 95% confidence intervals for identification regions found in Cheng & Small (2006). These findings are consistent with our discussion in Section 2.3. The one exception is for $\hat{\mu}_{32} - \hat{\mu}_{30}$ under Assumption 1-4. Using our approach, $\hat{\mu}_{32} - \hat{\mu}_{30}$ is 0.20 with a 95% credible interval is around (0.11, 0.29) for different prior specifications, which are similar to those found in Cheng & Small (2006). Since $\mu_{32} - \mu_{30}$ is point identifiable in this case, this result indicates that our approach and the approach proposed by Cheng & Small (2006) lead to comparable results when a parameter is point identifiable.

For the partially identifiable estimands in this hypothetical study, the improvement of efficiency using our approach does not lead to different conclusions regarding the causal estimands of interest for this hypothetical data. Under both sets of assumptions, our anal-

ysis shows that the 95% credible interval of CACE_{12} excludes 0, indicating a significant treatment effect between active treatment 1 and 2 among always-compliers. Our results also show significant treatment effect for always-compliers when comparing treatment 2 vs control and treatment 1 vs control, and for 1-only-compliers subpopulation when comparing treatment 1 vs control. However, the comparison between treatment 2 and control in 2-only-compliers subpopulation is inconclusive.

In summary, given this hypothetical data, we are able to obtain informative results. Specifically, under two sets of assumptions, treatment 1 is better than treatment 2 for always-compliers, and both are better than control for always-compliers and 1-only-compliers, whenever the comparisons are applicable. These findings are consistent with those in Cheng & Small (2006), but with improved precision.

4. Application to the WTP Data

In this section, we illustrate the proposed method with an application to the behavioral intervention study, the "Women Take Pride" (WTP) study (Janevic et al. 2003). We denote the three treatment groups by 0 for the usual care control treatment, 1 for the Group treatment, 2 for the Self-Directed treatment. The outcome of interest in this data analysis is the common cardiac bothersome score (Janevic et al. 2003) measured at Month 18. The common cardiac bothersome score ranges from 0 to 25 with higher scores indicating greater symptom effect. We created a binary outcome Y by comparing the measurement at Month 18 with that at baseline such that $Y = 1$ if the score does not increase, that is, symptoms do not worsen, and $Y = 0$ if otherwise. The compliance was defined as whether a woman completed at least one unit of materials. The primary objective of this data analysis was to estimate the effect of intervention programs after adjusting for non-compliance. For this study, Assumptions 1 and 3 hold, since patients did not have access to the alternative program if not assigned to that program. However, Assumption 2 may be questionable, since the interaction between patients in the group format may have an impact on the outcomes. For the purpose of exposition, we still make Assumption 2 in the data analysis. In addition, it is not clear whether Assumption 4 holds and we conduct a sensitivity analysis with or without Assumption 4.

We used the same model as described in Section 3.4 with the same conjugate priors, and analyzed the WTP data. We conducted our analysis under Assumption 1-3 with possible addition of Assumption 4 and its variations. Similar to the hypothetical data analysis, under Assumption 1-3, no causal treatment effect is point-identifiable when using the method of moment method proposed by Cheng & Small (2006) and a maximum likelihood (ML) analysis in Long (2005).

We first conducted sensitivity analysis using prior specifications as those in Section 3.4 and the conclusions were similar. Therefore only results using flat priors are reported and Table 5 summarizes these results under four different sets of assumptions: 1) Assumption 1-3; 2) Assumption 1-3 and Assumption 4, that is, 2-only-compliers ($C = 2$) do not exist; 3) Assumption 1-3 and $\rho_1 = 0$, that is, 1-only-compliers ($C = 1$) do not exist; 2) Assumption 1-3 and $\rho_0 = 0$, that is, always non-compliers ($C = 0$) do not exist. Under these assumptions, some causal estimands may not be applicable (Table 5). This study was also analyzed in Long (2005) using the ML analysis. As we discussed previously, Long (2005) showed that a ML analysis would lead to similar results as the method by Cheng & Small (2006) in this type of settings, therefore we only compare our results with those from the ML analysis.

We first focus on the results obtained under Assumption 1-3. The ML identification region of CACE_{12} $(= \mu_{32} - \mu_{31})$ is $(-0.26, 0.39)$ and its bootstrap 95% confidence interval is $(-0.37, 0.50)$ (Long, 2005). A Bayesian analysis using flat priors shows that the mean of its posterior distribution is 0.08 and its 95% credible interval is (-0.15, 0.28), which is considerably narrower than the 95% confidence interval for the identification region. However, since it still includes 0, there is no strong evidence indicating that either treatment is better than the other for the always-compliers ($C = 3$). For estimating $\mu_{32} - \mu_{30}$, the 95% confidence interval for its ML identification region is (-0.04,0.57) which includes 0, and its 95% Bayesian credible interval is (0.02,0.41) which excludes 0. Hence, based on this Bayesian analysis, there is some evidence indicating that the SD format(2) is better than the control for always-compliers in terms of improving the outcome. Similar to the arguments made in Section 4.4, due to the high compliance rates across treatment arms, there is little information based on which one could make inference about the treatment effects for the other two principal compliance strata (1-only-compliers and 2-only-compliers). This is reflected by the wide ranges of 95% credible intervals for $\mu_{22} - \mu_{20}$ and $\mu_{11} - \mu_{10}$, even though their widths are shorter than those of 95% "pseduo-confidence intervals".

Under four different sets of structural assumptions, our results in Table 5 show that the causal effect $\mu_{32} - \mu_{30}$ remains significant. Additional assumptions reduce the number of parameters, and hence may improve efficiency. Furthermore, the addition of assumption $\rho_2 = 0$ or $\rho_1 = 0$ lead to the point identifiability of $\mu_{32} - \mu_{30}$ and $\mu_{31} - \mu_{30}$, respectively. Specifically, the addition of assumption $\rho_2 = 0$ or $\rho_1 = 0$ shortens the Bayesian credible intervals for all causal effects defined in the principal stratum $C = 3$, and makes the estimates of $\mu_{32} - \mu_{31}$ close to becoming significant. The impact of assumption $\rho_0 = 0$ is relatively small due to the high compliance rates. In practice, caution needs to be exercised when adding structural assumption, since these assumptions may lead to biased estimates when they do not hold.

In summary, the results from our data analysis show that the Self-Directed treatment was better than the control for always compliers and the other causal comparisons were not statistically significant. In the settings of our interest, our results also seem to indicate that Bayesian inference can potentially achieve greater power in detecting significant treatment effects compared to the method of moment (Cheng & Small 2006) or the ML approach (Long 2005), which use the confidence intervals for the identification regions.

5. Discussion

For multi-arm trials subject to non-compliance, we propose a likelihood-based framework and a Bayesian inference approach. A data augmentation algorithm is used to approximate the marginal posterior distribution of causal parameters of interest. We also propose sensitivity analysis to investigate the impact of structural assumptions and priors. The proposed method is compared to a method of moment approach in Cheng & Small (2006) using a hypothetical data set used in Cheng & Small (2006) and the WTP study (Janevic et al. 2003). Our results show that the 95% Bayesian credible intervals are in general narrower than the estimated 95% confidence intervals for the identification regions of causal parameters, and that additional structural assumptions have the potential to improve the power of an analysis, if they hold.

In settings concerned in this paper, our proposed method has some attractive features compared to existing methods that compute the identification regions and their confidence

intervals. The framework is conceptually straightforward, and is not different from cases where parameters are point-identifiable in classical statistical sense. It is very flexible and can be easily applied to model different types of outcomes and extended to accommodate covariates adjustment, additional structural assumptions, and more complex designs such as the DRPT design in Long et al. (2008), which is a subject for future research. The proposed method may achieve greater power in terms of detecting significant treatment effects, especially when existing substantive knowledge can be incorporated to the priors. Furthermore, the interpretation of the credible intervals remains the same and it is straightforward to evaluate the properties of posterior distributions of causal parameters of interest. However, when using the proposed Bayesian approach in the settings of our interest, it is possible that the marginal posterior distribution of a parameter still concentrates its mass and remains flat over a part of the parameter space, in which case caution needs to be exercised in constructing 95% credible intervals. Consequently, in the presence of partially identifiable parameters, it is of future interest to systematically study and compare the properties of Bayesian credible intervals for the true values of parameters and confidence intervals for identification regions.

In addition to the DRPT design, our Bayesian approach can be extended to accommodate other interesting features of the WTP study. First, the WTP study includes the intervention of a group format, which allows interaction between participants; consequently, the outcome variable may be correlated between subjects assigned to the same group and Assumption 2 is questionable. To address this issue, one can introduce multivariate distributions for modeling Y for subjects of a same group in the group treatment arm, and one needs to change the complete data likelihood (1) and observed data likelihood (2) accordingly. Second, all participants in the WTP study completed between 0 to 6 weekly units and hence partial compliance was present. To adjust for partial compliance, one can still use the principal compliance framework by extending the approach proposed in Jin & Rubin (2008) to the case of multi-arm trials.

The proposed Bayesian approach shares one limitation with existing methods, though to a lesser degree. For complex designs with more treatment arms, the proposed analysis, while valid, may not be very informative, for example, it is likely that all credible intervals include 0. To improve the power of the analysis, it may require incorporating existing substantive knowledge into the priors and making strong structural assumptions.

Acknowledgements

We thank Dr. Noreen Clark for providing the WTP data, and the Associate Editor and one referee for their valuable suggestions, which helped to improve the paper considerably. This research was supported by National Cancer Institute grant R01CA76404.

References

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

Balke, A. & Pearl, J. (1997), 'Bounds on treatment effects from studies with imperfect compliance', *Journal of the American Statistical Association* **92**, 1171–1176.

Beran, R. (1988), 'Balanced simultaneous confidence sets.', *Journal of the American Statistical Association* **83**, 679–697.

Cheng, J. & Small, D. (2006), 'Bounds on causal effects in three-arm trials with noncompliance.', *Journal of the Royal Statistical Society: Series B* **68**, 815–836.

Frangakis, C. E. & Rubin, D. B. (2002), 'Principal stratification in causal inference.', *Biometrics* **58**, 21–29.

Gelfand, A. E. & Sahu, S. K. (1999), 'Identifiability, improper priors, and gibbs sampling for generalized linear models.', *Journal of the American Statistical Association* **94**, 247–253.

Gustafson, P. (2005), 'On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables.', *Statistical Science* **20**, 111–140.

Horowitz, J. & Manski, C. F. (2000), 'Nonparametric analysis of randomized experiments with missing covariate and outcome data.', *Journal of the American Statistical Association* **95**, 77–84.

Imbens, G. W. & Manski, C. F. (2004), 'Confidence intervals for partially identified parameters.', *Econometrica* **72**, 1845–1857.

Imbens, G. W. & Rubin, D. B. (1997*a*), 'Bayesian inference for causal effects in randomized experiments with noncompliance.', *Annals of Statistics* **25**, 305–327.

Imbens, G. W. & Rubin, D. B. (1997*b*), 'Estimating outcome distributions for compliers in instrumental variables models.', *Review of Economic Studies* **64**, 555–574.

Janevic, M. R., Janz, N. K., Lin, X., Pan, W., Sinco, B. R. & Clark, N. M. (2003), 'The role of choice in health education interventional trials: a review and case study.', *Social Science and Medicine* **56**, 1581–1594.

Jin, H. & Rubin, R. (2008), 'Principal stratification for causal inference with extended partial compliance.', *Journal of the American Statistical Association* **103**, 101–111.

Joffe, M. M. (2001), 'Using information on realized effects to determine prospective causal effects.', *Journal of the Royal Statistical Society: Series B* **63**, 759–774.

Lindley, D. V. (1971), *Bayesian Statistics: a Review*, Philadelphia,PA: SIAM.

Little, R. J. A., Long, Q. & Lin, X. (2009), 'A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance.', *Biometrics* p. in press.

Little, R. J. A. & Yau, L. (1998), 'Statistical techniques for analyzing data from prevention trials: treatment of no-shows using rubin's causal model.', *Psychological Methods* **3**, 147–159.

Long, Q. (2005), *Emerging issues in causal inference for intervention trials*, PhD. Dissertation, University of Michigan.

Long, Q., Little, R. J. A. & Lin, X. (2008), 'Causal inference in hybrid intervention trials involving treatment choice.', *Journal of the American Statistical Association* **103**, 474–484.

Manski, C. F. (2003), *Partial identification of probability distributions*, New York: Springer.

Neath, A. A. & Samaniego, F. J. (1997), 'On the efficacy of bayesian inference for nonidentifiable models.', *American Statistician* **51**, 225–232.

Peng, Y., Little, R. J. A. & Raghunathan, T. (2004), 'An extended general location model for causal inferences from data subject to non-compliance and missing values.', *Journal of the American Statistical Association* **60**, 598–608.

Robins, J. M. (2000), 'Correcting for non-compliance in randomized trials using structural nested mean models', *Communications in Statistics* **23**, 2379–2412.

Roy, J., Hogan, J. W. & Marcus, B. H. (2008), 'Principal stratification with predictors of compliance for randomized trials with 2 active treatments', *Biostatistics* **9**, 277–289.

Rubin, D. B. (1978), 'Bayesian inference for causal effects: the role of randomization.', *Annals of Statistics* **6**, 34–58.

Shafer, G. (1982), 'Belief functions and parametric models.', *Journal of the Royal Statistical Society: Series B* **44**, 322–352.

Tanner, M. & Wong, W. (1987), 'The calculation of posterior distributions by data augmentation.', *Journal of the American Statistical Association* **82**, 528–550.

Walley, P. (1991), *Statistical reasoning with imprecise probabilities*, London: Chapman and Hall.

Table 1. The expected outcome $\mu_{c,r,t}$ for principal compliance stratum $C = c$ when assigned to treatment $R = r$ and actually receiving treatment $T = t$ under Assumption 1 and 2

Subpopulation Defined by C	Population Proportion	R		
		0	1	2
3	ρ_3	$\mu_{3,0,0}$	$\mu_{3,1,1}$	$\mu_{3,2,2}$
2	ρ_2	$\mu_{2,0,0}$	$\mu_{2,1,0}$	$\mu_{2,2,2}$
1	ρ_1	$\mu_{1,0,0}$	$\mu_{1,1,1}$	$\mu_{1,2,0}$
0	ρ_0	$\mu_{0,0,0}$	$\mu_{0,1,0}$	$\mu_{0,2,0}$

Table 2. The expected outcome $\mu_{c,t}$ for principal compliance stratum $C = c$ when actually receiving treatment $T = t$ under Assumption 3 (ER) in addition to Assumption 1 and 2.

Subpopulation Defined by C	Population Proportion	R		
		0	1	2
3	ρ_3	$\mu_{3,0}$	$\mu_{3,1}$	$\mu_{3,2}$
2	ρ_2	$\mu_{2,0}$	$\mu_{2,0}$	$\mu_{2,2}$
1	ρ_1	$\mu_{1,0}$	$\mu_{1,1}$	$\mu_{1,0}$
0	ρ_0	$\mu_{0,0}$	$\mu_{0,0}$	$\mu_{0,0}$

Table 3. The structure of the observed data likelihood for subject i for all possible combinations of R_i and T_i under Assumption 1-3. Each cell value represents the probability of the observed data (Y_i, X_i, R_i, T_i) if the value of C_i were known, and the conditional probability of C_i given the observed data is computed as the ratio of each cell entry to its row total.

R_i	T_i	Principal Compliance C_i				Row total
		3	2	1	0	
0	0	$\rho_3^i f_{30}^i$	$\rho_2^i f_{20}^i$	$\rho_1^i f_{10}^i$	$\rho_0^i f_{00}^i$	$\rho_3^i f_{30}^i + \rho_2^i f_{20}^i + \rho_1^i f_{10}^i + \rho_0^i f_{00}^i$
1	1	$\rho_3^i f_{31}^i$	0	$\rho_1^i f_{11}^i$	0	$\rho_3^i f_{31}^i + \rho_1^i f_{11}^i$
	0	0	$\rho_2^i f_{20}^i$	0	$\rho_0^i f_{00}^i$	$\rho_2^i f_{20}^i + \rho_0^i f_{00}^i$
2	2	$\rho_3^i f_{32}^i$	$\rho_2 f_{22}^i$	0	0	$\rho_3^i f_{32}^i + \rho_2^i f_{22}^i$
	0	0	0	$\rho_1^i f_{10}^i$	$\rho_0^i f_{00}^i$	$\rho_1^i f_{10}^i + \rho_0^i f_{00}^i$

Table 4. Bayesian analysis of the hypothetical data using different prior specifications under two sets of assumptions. Mean is the mean of the Bayesian posterior distribution; SD, the standard deviation of the posterior distribution; CI, the 95% Bayesian credible interval; NA, an estimand is not applicable. λ represents different conjugate prior specifications and $\lambda = 1$ corresponds to the uninformative flat priors.

	Assumption 1-3			Assumption 1-4		
	Mean	SD	CI	Mean	SD	CI
Causal Effects	Prior distributions with $\lambda = 1$					
$\mu_{32} - \mu_{31}$	-0.26	0.04	(-0.33,-0.19)	-0.27	0.03	(-0.33,-0.20)
$\mu_{32} - \mu_{30}$	0.20	0.05	(0.11,0.29)	0.20	0.04	(0.12,0.28)
$\mu_{31} - \mu_{30}$	0.46	0.04	(0.37,0.54)	0.47	0.04	(0.39,0.54)
$\mu_{22} - \mu_{20}$	0.17	0.38	(-0.62,0.83)	NA	NA	NA
$\mu_{11} - \mu_{10}$	0.58	0.12	(0.31,0.79)	0.56	0.14	(0.26,0.79)
	Prior distributions with $\lambda = 0.5$					
$\mu_{32} - \mu_{31}$	-0.26	0.04	(-0.33,-0.18)	-0.27	0.03	(-0.34,-0.20)
$\mu_{32} - \mu_{30}$	0.20	0.05	(0.11,0.29)	0.20	0.04	(0.12,0.28)
$\mu_{31} - \mu_{30}$	0.46	0.04	(0.37,0.54)	0.47	0.04	(0.39,0.55)
$\mu_{22} - \mu_{20}$	0.14	0.47	(-0.85,0.93)	NA	NA	NA
$\mu_{11} - \mu_{10}$	0.59	0.14	(0.28,0.82)	0.55	0.16	(0.21,0.82)
	Prior distributions with $\lambda = 10$					
$\mu_{32} - \mu_{31}$	-0.24	0.04	(-0.30,-0.18)	-0.25	0.03	(-0.31,-0.19)
$\mu_{32} - \mu_{30}$	0.20	0.05	(0.12,0.29)	0.20	0.04	(0.12,0.28)
$\mu_{31} - \mu_{30}$	0.45	0.04	(0.37,0.52)	0.45	0.04	(0.38,0.52)
$\mu_{22} - \mu_{20}$	0.11	0.15	(-0.19,0.39)	NA	NA	NA
$\mu_{11} - \mu_{10}$	0.48	0.08	(0.28,0.63)	0.43	0.10	(0.21,0.61)

Table 5. Bayesian analysis for the outcome of interest (common cardiac symptom bothersome score at month 18) in the WTP study under different sets of assumptions using flat priors ($\lambda = 1$). Mean is the mean of the Bayesian posterior distribution; SD, the standard deviation of the posterior distribution; CI, the 95% Bayesian credible interval; NA, an estimand is not applicable.

	Mean	SD	CI	Mean	SD	CI
Causal Effects	Assumption 1-3			Assumption 1-3 and $\rho_2 = 0$		
$\mu_{32} - \mu_{31}$	0.08	0.10	(-0.15,0.28)	0.07	0.05	(-0.03,0.17)
$\mu_{32} - \mu_{30}$	0.19	0.10	(0.02,0.41)	0.12	0.06	(0.01,0.23)
$\mu_{31} - \mu_{30}$	0.12	0.12	(-0.07,0.40)	0.05	0.06	(-0.07,0.18)
$\mu_{22} - \mu_{20}$	-0.16	0.34	(-0.77,0.60)	NA	NA	NA
$\mu_{11} - \mu_{10}$	-0.33	0.35	-0.90,0.42)	-0.18	0.39	(-0.86,0.65)
	Assumption 1-3, $\rho_1 = 0$			Assumption 1-3, $\rho_0 = 0$		
$\mu_{32} - \mu_{31}$	0.09	0.05	(-0.01,0.20)	0.05	0.15	(-0.23,0.35)
$\mu_{32} - \mu_{30}$	0.13	0.06	(0.02,0.25)	0.26	0.12	(0.04,0.49)
$\mu_{31} - \mu_{30}$	0.04	0.06	(-0.08,0.15)	0.21	0.14	(-0.06,0.47)
$\mu_{22} - \mu_{20}$	0.00	0.39	(-0.74,0.76)	-0.17	0.20	(-0.59,0.16)
$\mu_{11} - \mu_{10}$	NA	NA	NA	-0.36	0.26	(-0.84,0.06)