Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2010)
59, *Part* 3, *pp.* 513–531

Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework

Qi Long,

Emory University, Atlanta, USA

Roderick J. A. Little

University of Michigan, Ann Arbor, USA

and Xihong Lin

Harvard University, Boston, USA

[Received April 2009. Revised November 2009]

Summary. Data analysis for randomized trials including multitreatment 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 multitreatment arms subject to non-compliance. One treatment effect of interest in the presence of non-compliance is the complier average causal effect, which is defined as the treatment effect for subjects who would comply regardless of the treatment assigned. Following the idea of principal stratification, we define principal compliance in trials with three treatment arms, extend the complier average causal effect and define causal estimands of interest in this setting. In addition, we discuss structural assumptions that are needed for estimation of causal effects and the identifiability problem that is 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-moments approach that was proposed by Cheng and Small in 2006 by using a hypothetical data set, and we further illustrate our approach with an application to a behavioural intervention study.

Keywords: Causal inference; Complier average causal effect; Multiarm trials; Non-compliance; Principal compliance; Principal stratification

1. Introduction

1.1. Non-compliance in trials involving multitreatment arms

Data analysis for randomized controlled trials 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 and Rubin, 1997a, b; Little and Yau, 1998; Peng *et al.*, 2004; Robins, 1994). 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

Address for correspondence: Qi Long, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Room 322, 1518 Clifton Road North East, Atlanta, GA 30322, USA.
E-mail: qlong@emory.edu

0035–9254/10/59513

variable methods of estimation. The basic idea is to classify participants as one of compliers, defiers, never takers and always takers according to their potential compliance status on 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 and 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 and 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 and Small (2006) proposed bounds on causal effects in three-arm trials subject to non-compliance, using method-of-moments estimates. To account for sampling uncertainty, they followed Horowitz and 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, e.g. 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 modelling missing data. Imbens and Rubin (1997a) applied this framework to the problem of non-compliance in randomized trials, specifically two-arm randomized trials. Their approach clarified the role that is 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 and Rubin (1997a) also discussed situations where relaxing assumptions such as the exclusion restriction 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 and Rubin (1997a) also discussed what could be learned in this case by 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 paper 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 behavioural intervention programmes that were aimed at enhancing the women’s ability to manage their disease. In addition to a usual care control treatment, two formats of a behavioural intervention were compared in this study: a group format, where 6–8 women meet for 2–2.5 h 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 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 detail 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 1 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 and Rubin (1997a) 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 non-compliance in trials with two active treatments, where a Bayes procedure was also used for inference. They proposed to model directly marginal distributions of the compliance status under each treatment on the basis of observed data. The marginal models are then used 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 to be independent of covariates. Our approach has several key differences. First, we model principal compliance directly and treat the principal compliance status as missing data in the analysis, avoiding the implicit assumption that the association parameter is independent of covariates. Although conceptually the approach in Roy *et al.* (2008) can be extended to trials with more treatment arms, modelling the distribution of principal compliance indirectly through association parameters becomes considerably more complicated and requires more implicit assumptions that the association parameters are independent of covariates. 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 that was proposed in Cheng and Small (2006) and attempt to clarify the differences between a Bayesian approach and a classical frequentist approach in the setting of interest. The rest of the paper is organized as follows. In Section 2, we introduce principal stratification of a population of interest on the basis of principal compliance status (Little *et al.*, 2009) and define causal estimands of interest; we further discuss structural assumptions and issues that are related to the identifiability of causal estimands of interest, and we 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 (DA) algorithm (Tanner and Wong, 1987) to simulate the posterior distributions of causal parameters. We compare our approach with the method-of-moments approach that was proposed in Cheng and Small (2006) by using a hypothetical data set. In Section 4, we illustrate our approach by using the WTP 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 illustrative purposes, we present our framework by using a randomized trial involving two active treatment arms (1 and 2) and one control arm (0). Later, we briefly discuss extensions to trials with more than two active treatments. 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 and 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 that is actually assigned, for each individual in the sample. We thus have a major identifiability problem. We first make some assumptions to reduce the scale of this problem.

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

This is a monotonicity assumption in the sense of Angrist *et al.* (1996) and implies that

(a) subjects assigned to the control always take the control and
(b) subjects assigned one of the active treatments either take that treatment or, if they fail to comply, take the control treatment.

Hence we know that $T(0)=0$, $T(1)=1$ or $T(1)=0$, and $T(2)=2$ or $T(2)=0$. This reduces the number of principal compliance strata from 27 to 4, on the basis of subjects' potential compliance status under both active treatments. Following Little *et al.* (2009), we define a principal compliance variable C for these strata, with values $C=3$ for always compliers who comply with both treatments ($T(0)=0$, $T(1)=1$ and $T(2)=2$), $C=2$ for 2-only compliers (those who comply when assigned to treatment 2 but do not comply when assigned to treatment 1; $T(0)=0$, $T(1)=0$ and $T(2)=2$), $C=1$ for 1-only compliers, who comply when assigned to treatment 1 but do not comply when assigned to treatment 2 ($T(0)=0$, $T(1)=1$ and $T(2)=0$), and $C=0$ for non-compliers, 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 the 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).

Table 1. 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 assumptions 1 and 2

Subpopulation defined by C	*Population proportion*	*Outcomes for the following values of 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	$\mu_{0,0,0}$	$\mu_{0,1,0}$	$\mu_{0,2,0}$

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; 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 the stable unit treatment value assumption, $Y_i(\mathbf{R},\mathbf{T})$ can be written as $Y_i(R_i,T_i)$.

Assumption 3 (exclusion restriction; 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 the stable unit treatment value and exclusion restriction assumptions, 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. A 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. In contrast, 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 one control arm, the CACE is uniquely defined as the difference in means between the active treatment and control in the population of principal compliers.

In our setting, some interesting CACEs can be defined. Three of particular interest are

(a) $\text{CACE}_{12}=\mu_{3,1}-\mu_{3,2}$, the CACE for comparing treatment 1 with treatment 2 for always compliers $(C=3)$;

(b) $\text{CACE}_1=(\rho_3\mu_{3,1}+\rho_1\mu_{1,1})/(\rho_3+\rho_1)-(\rho_3\mu_{3,0}+\rho_1\mu_{1,0})/(\rho_3+\rho_1)$, the CACE for comparing treatment 1 with control for always compliers $(C=3)$ and 1-only compliers $(C=1)$;

Table 2. Expected outcome $\mu_{c,t}$ for principal compliance stratum $C=c$ when actually receiving treatment $T=t$ under assumption 3 (the exclusion restriction) in addition to assumptions 1 and 2

Subpopulation defined by C	*Population proportion*	*Outcomes for the following values of 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	$\mu_{0,0}$	$\mu_{0,0}$	$\mu_{0,0}$

(c) $\text{CACE}_2 = (\rho_3\mu_{3,2} + \rho_2\mu_{2,2})/(\rho_3 + \rho_2) - (\rho_3\mu_{3,0} + \rho_2\mu_{2,0})/(\rho_3 + \rho_2)$, the CACE for comparing treatment 2 with 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 *versus* control and treatment 2 *versus* 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 by using previously developed methods for comparing an active treatment with the control, and then to 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 and Small (2006) provided some discussion of the use of CACE_{12}. Other causal treatment effects can also be defined, e.g. $\mu_{3,2} - \mu_{3,0}$, $\mu_{3,1} - \mu_{3,0}$, $\mu_{1,1} - \mu_{1,0}$ 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, i.e. 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 0 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. $\rho_2 = 0$, i.e. 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 shall 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 shall 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, i.e. 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 that classical statistical methods may have difficulties dealing with non-identifiable or partially identifiable parameters (Balke and Pearl, 1997; Cheng and Small, 2006; Manski, 2003; Neath and Samaniego, 1997). In particular, Cheng and Small (2006) studied a similar design to 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) cannot be distinguished on the basis of the observables and the set is a strict subset of the parameter space. Specifically, in our setting, under assumptions 1–3, there are eight marginal means that are of interest (Table 2), and none of which is point identifiable; hence, all causal treatment effects that were 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 (equations (1)–(4) in Cheng and Small (2006)) and usually these values form a set-valued interval. Under assumptions 1–4, μ_{22} and μ_{20} are no longer applicable; hence the number of marginal means in Table 2 is reduced to 6. 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 and Manski (2000), Cheng and 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 and 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 with the nominal level of coverage. However, Imbens and Manski (2004) did so in a considerably simpler setting and it is not trivial to extend their methods to our setting (Cheng and Small, 2006).

In the above settings with non-identifiable or partially identifiable parameters, often identifiability is a less serious issue with a Bayesian framework; we can still make interpretable inference by using a Bayesian approach (Gustafson, 2005; Imbens and Rubin, 1997a; Lindley, 1971; Neath and 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 intervals and to achieve more power, which makes a Bayesian approach attractive. Even in the presence of potential improper posterior distributions, it is still possible to obtain meaningful results by using a Bayesian approach (Gelfand and Sahu, 1999).

We note another important difference between classical statistical methods such as maximum likelihood (ML) 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 that $L(\theta_1,\theta_2|\text{data})$ 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

$$\int L(\theta_1,\theta_2|\text{data})\,p(\theta_1,\theta_2)\,\mathrm{d}\theta_2.$$

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. Although 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 a more profound influence on the statistical inference than it initially appears. We suspect that this marginalization step also helps to produce narrower credible intervals compared with confidence intervals for the identification regions from a classical frequentist approach. It is not obvious how to marginalize in a sensible way within the classical statistical framework.

3. 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 assumptions 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, with value 0 for non-compliers, 1 for 1-only compliers, 2 for 2-only compliers and 3 for always compliers. Let T_i denote the treatment that is actually received, which is uniquely determined by C_i and R_i, i.e. $T_i(C_i, R_i)$. 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 observe only 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 a specific active treatment, then compliance 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, e.g. 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 $C_{\text{obs},i}$. 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 $C_{\text{obs},i}=\{1,3\}$. Also, if subject i does not actually receive a treatment, then its potential outcome given that treatment is not observed. We note that i is sometimes suppressed in our notation when this 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 (the stable unit treatment value assumption) constitute an independent and identically distributed sample. Then the observed data can be represented as $(Y_i, C_{\text{obs},i}, R_i, T_i, X_i)$. 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 $(Y_i, C_{\text{obs},i}, R_i, T_i, X_i)$ 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 that are associated with the distribution of the potential outcomes $(Y_i(2), Y_i(1), Y_i(0))$, which have 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 from the design, we can ignore the treatment assignment model and just focus on $f(Y_i, C_i, T_i | R_i, X_i)$ 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, $f\{T_i(C_i, R_i) | C_i, R_i, X_i\} = 1$. Owing to the exclusion restriction 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 that are associated with the potential outcome model, i.e. $f\{Y(T_i) = Y_i | C_i, X_i, \alpha\}$. Also the random treatment assignment implies that $f(C_i | R_i, X_i) = f(C_i | X_i, \beta)$, where β denotes the set of parameters that are associated with the model for C. Assume that α and β are distinct, and let $\theta = (\alpha, \beta)$. Given 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

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

We now examine L_i, the contribution to the observed data likelihood from subject i. On the basis of our previous discussion for $C_{\text{obs},i}$, it is straightforward to show that, for subject i, the observed data $(Y_i, T_i, C_{\text{obs},i}, X_i)$ follow a mixture distribution and the observed data likelihood is

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

Hence, the observed data likelihood (2) is a product of different mixture distributions and depends only 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 related only to the parameters that are associated with these marginal distributions.

Let $f_{ct}(Y | X, \alpha_{ct})$ denote $f\{Y(t) = Y | C = c, X, \alpha_{ct}\}$, i.e. 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 that are associated with this distribution and α is then the collection of all α_{ct} that can be estimated from the data. Under assumptions 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 expression (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 $f_{i,ct} = f_{ct}(Y_i | X_i, \alpha_{ct})$, and denote by $S(r, t)$ the set of subjects with $R = r$ and $T = t$. Under assumptions 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 $C_{\text{obs},i}$ given the observed T and R_i 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 on the basis of 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

Table 3. Structure of the observed data likelihood for subject i for all possible combinations of R_i and T_i under assumptions 1–3†

R_i	T_i	*Results for the following principal compliances* C_i*:*				*Row total*
		3	*2*	*1*	*0*	
0	0	$\rho^i_3 f^i_{30}$	$\rho^i_2 f^i_{20}$	$\rho^i_1 f^i_{10}$	$\rho^i_0 f^i_{00}$	$\rho^i_3 f^i_{30}+\rho^i_2 f^i_{20}$ $+\rho^i_1 f^i_{10}+\rho^i_0 f^i_{00}$
1	1	$\rho^i_3 f^i_{31}$	0	$\rho^i_1 f^i_{11}$	0	$\rho^i_3 f^i_{31}+\rho^i_1 f^i_{11}$
	0	0	$\rho^i_2 f^i_{20}$	0	$\rho^i_0 f^i_{00}$	$\rho^i_2 f^i_{20}+\rho^i_0 f^i_{00}$
2	2	$\rho^i_3 f^i_{32}$	$\rho^i_2 f^i_{22}$	0	0	$\rho^i_3 f^i_{32}+\rho^i_2 f^i_{22}$
	0	0	0	$\rho^i_1 f^i_{10}$	$\rho^i_0 f^i_{00}$	$\rho^i_1 f^i_{10}+\rho^i_0 f^i_{00}$

†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.

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

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

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

It is obvious that the posterior distributions in expression (4) are proper. If we make assumption 4 in addition to assumptions 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 fewer assumptions are made.

It is straightforward to show that the marginal means that are defined in Table 2, and hence causal estimands of interest discussed in Section 2.2, can be expressed in terms of the parameters $\{\alpha_{ct}\}$. For example,

$$\mu_{32}=\int\int Y f_{32}(Y|X,\alpha_{32})\,\mathrm{d}\nu(X)\,\mathrm{d}Y,$$

$$\mu_{31}=\int\int Y f_{31}(Y|X,\alpha_{31})\,\mathrm{d}\nu(X)\,\mathrm{d}Y$$

and

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

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

3.3. Estimation and inference

The posterior distribution of θ in expression (4) is mathematically quite simple, but its computation is complicated because 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 DA algorithm (Tanner and Wong, 1987) to simulate the posterior distributions in expression (4) treating C as missing data. This DA 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 DA algorithm can be outlined as follows.

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

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

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 implement this algorithm, we need to specify a prior distribution $p(\theta)$ for θ. We propose to choose $p(\theta)$ to be non-informative but proper, and conjugate to the likelihood in the P-step when possible. The P-step can then be implemented by using a Gibbs sampler. The examples of these priors in some special cases can be found in the next two sections. To draw from the posterior distributions, we iterate between the I-step and P-step until the algorithm converges.

In the case of no covariates, $\rho_{i,c}=\rho_c$, 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 $\theta=(\alpha,\rho)$. Whereas the I-step in the DA algorithm does not change, the posterior distributions (5)–(7) in the P-step simplify to

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

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 approach proposed relies on the structural assumptions and the specification of the prior distribution. One can incorporate prior experience or preliminary results to determine the selection of the assumptions and the choice of priors whenever this information is available. When no prior information is available, we propose to conduct an additional sensitivity analysis to examine the influence of assumptions and the choice of prior distribution. First, we can conduct inference under different combinations of assumptions and examine how the causal estimands of interest vary. For example, we can consider the inclusion or exclusion of assumption 4 and its variants for other principal compliance strata 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, we can conduct inference under different specifications of the prior distribution. One could stay with conjugate priors in the P-step and compare the results by using different parameter values.

It is straightforward to extend the proposed framework to the estimation of causal estimands, in particular, CACE_{12}, for trials with multitreatment 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 is similar to expressions (2) and (3), and Bayesian inference can then be constructed along the lines that have been discussed in this section.

3.4. Hypothetical example

We now compare our approach with a method-of-moments approach in Cheng and Small (2006) by using a hypothetical example that was analysed in Cheng and 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 and 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 $Y=1$ for a successful outcome and $Y=0$ for a failure. Participants who are assigned to the control arm actually take the control, among which 45% have successful outcomes ($Y=1$). For those who are assigned to treatment A, 95% of the subjects actually take treatment 1, of which 95% have successful outcomes, and 5% actually take the control, of which 20% have successful outcomes. For those assigned to treatment B, 80% of the subjects actually take treatment B, of which 70% have successful outcomes, and 20% actually take the control, of 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 that was described in Section 3 without covariate 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

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

where α_{ct} represents the probability of success for subjects in principal compliance stratum $C=c$ when taking treatment $T=t$. We assume the conjugate prior distributions $\alpha_{ct} \sim \text{beta}(a, b)$ 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 $\lambda=1$, uninformative flat prior distributions are assumed for all parameters. Given the model specification, causal estimands of interest are then functions of μ_{ct}, e.g. $\text{CACE}_{12}=\mu_{32}-\mu_{31}$. The DA 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,

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

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 by using 12 000 iterations from each of 20 independent runs of the DA algorithm with different initial values drawn from uniform distributions over the range of the parameters. The first 10 000 iterations from each run were discarded. This scheme was used for all data analyses that are 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 assumptions 1–3 and the other with assumptions 1–4. As discussed previously, given assumptions 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 various prior specifications and two sets of assumptions.

Cheng and Small (2006) analysed this hypothetical data set by using a method-of-moments approach and presented the results in their Table 4. This hypothetical data set was also analysed by using an ML inference approach in Long (2005), where the identification regions were obtained and their confidence intervals were constructed on the basis of 5000 bootstrap samples (Horowitz and Manski, 2000). The identification regions by using the ML approach were similar to those in Cheng and Small (2006) and their confidence intervals were constructed to cover the identification regions with fixed probability in the spirit of Horowitz and Manski (2000) and Cheng and Small (2006). The results that are found in Long (2005) were very close to those found in Cheng and 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 assumptions 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 assumptions 1–4, the identification region changes to (0.44, 0.50) with a confidence interval of (0.37,0.57). Under assumptions 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 assumptions 1–4, the identification region changes to (0.42, 0.73) with a confidence interval of (0.23,0.92). More importantly, under assumptions 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 assumptions 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.

Table 4. Bayesian analysis of the hypothetical data by using different prior specifications under two sets of assumptions†

Causal effects	*Results with assumptions 1–3*			*Results with assumptions 1–4*		
	Mean	*SD*	*CI*	*Mean*	*SD*	*CI*
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)

†Mean is the mean of the Bayesian posterior distribution, SD the standard deviation of the posterior distribution, CI the 95% Bayesian credible interval and NA denotes that an estimand is not applicable. λ represents different conjugate prior specifications and $\lambda = 1$ corresponds to the uninformative flat priors.

Compared with the results that were obtained by using the flat priors ($\lambda = 1$), our sensitivity analysis shows that different prior specifications have minimal effect on the causal estimands defined in the principal compliance stratum $C = 3$, i.e. $\mu_{32}-\mu_{31}$, $\mu_{32}-\mu_{30}$ 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 that is close to the flat prior ($\lambda = 0.5$) has less effect than the priors that are strongly informative ($\lambda = 10$). The standard deviation of the posterior distribution and the width of credible intervals decrease considerably as a result of strong informative priors; hence the power of the analysis improves. Since the rate of compliance 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 a substantial effect 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 sets of assumptions and different prior specifications are considerably narrower than the corresponding 95% confidence intervals for identification regions that were found in Cheng and 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 assumptions 1–4. Using our approach, $\hat{\mu}_{32}-\hat{\mu}_{30}$ is 0.20 with a 95% credible interval around (0.11, 0.29) for different prior specifications, which is similar to those found in Cheng and Small (2006). Since $\mu_{32}-\mu_{30}$ is point identifiable in this case, this result suggests that our approach and the approach that was proposed by Cheng and 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 by using our approach does not lead to different conclusions regarding the causal estimands of interest for these hypothetical data. Under both sets of assumptions, our analysis 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 a significant treatment effect for always compliers when comparing treatment 2 *versus* control and treatment 1 *versus* control, and for the 1-only-compliers subpopulation when comparing treatment 1 *versus* control. However, the comparison between treatment 2 and control in the 2-only-compliers subpopulation is inconclusive.

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

4. Application to the women take pride data

In this section, we illustrate the proposed method with an application to the behavioural intervention study, the 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 and 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 effects. 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, i.e. symptoms do not worsen, and $Y=0$ if otherwise. The compliance was defined as whether a woman completed at least 1 unit of materials. The primary objective of this data analysis was to estimate the effect of intervention programmes after adjusting for non-compliance. For this study, assumptions 1 and 3 hold, since patients did not have access to the alternative programme if they were not assigned to that programme. However, assumption 2 may be questionable, since the interaction between patients in the group format may have an influence 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 analysed the WTP data by using the model that was described in Section 3.4 with the same conjugate prior distributions. We conducted our analysis under assumptions 1–3 with possible addition of assumption 4 and its variations. As with the hypothetical data analysis, no causal treatment effect is point identifiable under assumptions 1–3 when using the method-of-moments method that was proposed by Cheng and Small (2006) and the ML analysis in Long (2005).

We conducted a sensitivity analysis using prior specifications as in Section 3.4, and the conclusions were similar. Therefore only results by using flat priors are reported, and Table 5 summarizes these results under four different sets of assumptions:

(a) assumptions 1–3;
(b) assumptions 1–3 and assumption 4, i.e. 2-only compliers ($C=2$) do not exist;
(c) assumptions 1–3 and $\rho_1=0$, i.e. 1-only compliers ($C=1$) do not exist;
(d) assumptions 1–3 and $\rho_0=0$, i.e. always non-compliers ($C=0$) do not exist.

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 by using flat priors ($\lambda = 1$)†

Causal effects	*Mean*	*SD*	*CI*	*Mean*	*SD*	*CI*
	Assumptions 1–3			*Assumptions 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)
	Assumptions 1–3 and $\rho_1 = 0$			*Assumptions 1–3 and $\rho_0 = 0$*		
$\mu_{32} - \mu_{31}$	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)

†Mean is the mean of the Bayesian posterior distribution, SD the standard deviation of the posterior distribution, CI the 95% Bayesian credible interval and NA denotes that an estimand is not applicable.

Under these assumptions, some causal estimands may not be applicable (Table 5). This study was also analysed in Long (2005) by using ML. As we discussed previously, Long (2005) showed that ML would lead to similar results to those of the method by Cheng and Small (2006) in this type of setting; therefore we compare our results only with those from the ML analysis.

We first focus on the results that were obtained under assumptions 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, on the basis of this Bayesian analysis, there is some evidence indicating that the standard deviation format (2) is better than the control for always compliers in terms of improving the outcome. Similarly to the arguments that were made in Section 4.4, owing to the high compliance rates across treatment arms, there is little information 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 95% confidence intervals for the identification regions.

Our results in Table 5 show that the causal effect $\mu_{32} - \mu_{30}$ remains significant under four different sets of structural assumptions. Additional assumptions reduce the number of parameters and hence may improve efficiency. Furthermore, the addition of assumption $\rho_2 = 0$ or $\rho_1 = 0$ leads 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 significant. The effect of assumption $\rho_0 = 0$ is relatively small owing to the high compliance rates. In practice, caution needs to be exercised when adding structural assumptions, 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 that we studied, our results also seem to indicate that Bayesian inference can potentially achieve greater power in detecting significant treatment effects compared with the method of moments (Cheng and Small, 2006) or the ML approach (Long, 2005), which use the confidence intervals for the identification regions.

5. Discussion

For multiarm 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 analyses to investigate the effect of structural assumptions and priors. The method proposed is compared with a method-of-moments approach in Cheng and Small (2006) by using a hypothetical data set that was used in Cheng and 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 that were considered in this paper, our proposed method has some attractive features compared with 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 the classical statistical sense. It is very flexible and can be easily applied to model different types of outcomes and extended to accommodate covariate adjustment, additional structural assumptions and more complex designs such as the doubly randomized preference trial design in Long *et al.* (2008), which is a subject for future research. The method proposed may achieve greater power in terms of detecting significant treatment effects, especially when existing substantive knowledge can be incorporated in 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 these settings, 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. It is of interest to study the properties of Bayesian credible intervals and frequentist confidence intervals for partially identified parameters in these settings.

In addition to the doubly randomized preferential trial 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 who are assigned to the same group and assumption 2 is questionable. To address this issue, one can introduce multivariate distributions for modelling 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 and 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 that was proposed in Jin and Rubin (2008) to the case of multiarm trials.

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

Acknowledgements

We thank Dr Noreen Clark for providing the WTP data, and the Associate Editor and a referee for their valuable suggestions, which helped to improve the paper considerably. This research was supported by National Cancer Institute grant R37CA76404.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Balke, A. and Pearl, J. (1997) Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Ass.*, **92**, 1171–1176.

Beran, R. (1988) Balanced simultaneous confidence sets. *J. Am. Statist. Ass.*, **83**, 679–697.

Cheng, J. and Small, D. S. (2006) Bounds on causal effects in three-arm trials with non-compliance. *J. R. Statist. Soc.* B, **68**, 815–836.

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

Gelfand, A. E. and Sahu, S. K. (1999) Identifiability, improper priors, and gibbs sampling for generalized linear models. *J. Am. Statist. Ass.*, **94**, 247–253.

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

Horowitz, J. and Manski, C. F. (2000) Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Statist. Ass.*, **95**, 77–84.

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

Imbens, G. W. and Rubin, D. B. (1997a) Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.*, **25**, 305–327.

Imbens, G. W. and Rubin, D. B. (1997b) Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.*, **64**, 555–574.

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

Jin, H. and Rubin, R. (2008) Principal stratification for causal inference with extended partial compliance. *J. Am. Statist. Ass.*, **103**, 101–111.

Joffe, M. M. (2001) Using information on realized effects to determine prospective causal effects. *J. R. Statist. Soc.* B, **63**, 759–774.

Lindley, D. V. (1971) *Bayesian Statistics: a Review*. Philadelphia: Society for Industrial and Applied Mathematics.

Little, R. J. A., Long, Q. and Lin, X. (2009) A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*, **65**, 640–649.

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

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

Long, Q., Little, R. J. A. and Lin, X. (2008) Causal inference in hybrid intervention trials involving treatment choice. *J. Am. Statist. Ass.*, **103**, 474–484.

Manski, C. F. (2003) *Partial Identification of Probability Distributions*. New York: Springer.

Neath, A. A. and Samaniego, F. J. (1997) On the efficacy of bayesian inference for nonidentifiable models. *Am. Statistn*, **51**, 225–232.

Peng, Y., Little, R. J. A. and Raghunathan, T. (2004) An extended general location model for causal inferences from data subject to non-compliance and missing values. *Biometrics*, **60**, 598–607.

Robins, J. M. (1994) Correcting for non-compliance in randomized trials using structural nested mean models. *Communs Statist. Theory Meth.*, **23**, 2379–2412.

Roy, J., Hogan, J. W. and 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. *Ann. Statist.*, **6**, 34–58.
Shafer, G. (1982) Belief functions and parametric models (with discussion). *J. R. Statist. Soc.* B, **44**, 322–352.
Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Statist. Ass.*, **82**, 528–550.
Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.