

THE POTENTIAL FOR BIAS IN PRINCIPAL CAUSAL EFFECT ESTIMATION WHEN TREATMENT RECEIVED DEPENDS ON A KEY COVARIATE

BY CORWIN M. ZIGLER^{*,†} AND THOMAS R. BELIN[‡]

Harvard University[†] and University of California, Los Angeles[‡]

Motivated by a potential-outcomes perspective, the idea of principal stratification has been widely recognized for its relevance in settings susceptible to posttreatment selection bias such as randomized clinical trials where treatment received can differ from treatment assigned. In one such setting, we address subtleties involved in inference for causal effects when using a key covariate to predict membership in latent principal strata. We show that when treatment received can differ from treatment assigned in both study arms, incorporating a stratum-predictive covariate can make estimates of the “complier average causal effect” (CACE) derive from observations in the two treatment arms with different covariate distributions. Adopting a Bayesian perspective and using Markov chain Monte Carlo for computation, we develop posterior checks that characterize the extent to which incorporating the pretreatment covariate endangers estimation of the CACE. We apply the method to analyze a clinical trial comparing two treatments for jaw fractures in which the study protocol allowed surgeons to overrule both possible randomized treatment assignments based on their clinical judgment and the data contained a key covariate (injury severity) predictive of treatment received.

1. Introduction. All-or-none treatment noncompliance in the context of a randomized two-arm clinical trial is perhaps the simplest and most common example in health-sciences research of potential confounding by a posttreatment variable. One strategy to address confounding of treatment receipt with individual characteristics is the use of an instrumental-variable method [27] which has been linked to a potential-outcomes perspective on causal inference [1, 6, 22]. More recently, strategies for addressing potential confounding by posttreatment variables have been formalized using the framework of principal stratification [7], a central challenge of which is the classification of patients into latent subclasses, called principal strata, that facilitate causal treatment comparisons.

In the context of treatment noncompliance the target for inference is often the “complier average causal effect (CACE)” [22] which compares treatment outcomes with control outcomes in the principal stratum of “compliers” who would potentially receive whichever treatment is randomly assigned (as distinct from other principal strata where patients may always receive a particular treatment). Such comparisons within principal strata are known as “principal effects” and permit causal interpretation. Knowledge of membership in the stratum of compliers or in any other principal stratum requires knowledge of patients’ potential treatment receipts under both possible randomized assignments, but this information will never be observed in total for any individual in the population since treatment received is only observed for the actually assigned treatment.

A battery of now-standard assumptions underly methods for identifying and estimating the CACE in settings framed as treatment noncompliance [1], but more recent attention (e.g., [17, 24]) has been paid to the use of pretreatment covariates to increase precision or relax exclusion restrictions. One line of research focuses on settings where patients randomized to the control arm do not have access to the active treatment, that is, settings where the entire population would receive control if so assigned. The key feature of these

*Corresponding author

Keywords and phrases: Complier average causal effect, Noncompliance, Principal effect, Principal stratification

settings is that they allow patients who are assigned and receive active treatment to be identified as compliers, which further allows pretreatment covariates associated with membership in this stratum to be used in identifying which patients randomized to control are exchangeable with compliers. Specifically, these settings motivate so-called “two-stage” approaches that first use pretreatment covariates to estimate propensity scores [28] of membership in the complier stratum then estimate outcomes conditional on these so-called “principal scores” [5, 16, 24–26]. Although some previous research has framed the one-sided access to treatment as a nonessential detail that merely simplifies exposition, we aim to illuminate that added complexity can arise in more general settings where noncompliance exists in both treatment arms.

When treatment received can deviate from treatment assigned in both study arms, the use of pretreatment covariates to aid estimation of the CACE is more complicated because no patient is known to belong to the stratum of compliers, precluding estimation a model such as a propensity score model for membership in the complier stratum. Joint-estimation methods that simultaneously model stratum membership and outcomes have been employed in these settings [3, 8, 9, 15, 17, 29], which typically consist of two underlying strata in addition to the compliers: “never-takers” who would never receive the active treatment, and “always-takers” who would always receive the active treatment. Through use of standard assumptions that will be elaborated later, stratum membership for patients who receive a treatment different from that assigned can be regarded as having been revealed, with such individuals being either never-takers or always-takers, and covariates associated with membership in these two “noncomplier” strata can be identified. However, membership in the complier stratum is never directly observed because patients who receive the assigned treatment (and thus *might* be compliers) generally represent mixtures of compliers and never-takers (in the control arm) or compliers and always-takers (in the treatment arm). Since pretreatment covariates can only provide direct information about characteristics of noncompliers, the role of such covariates in estimating the CACE is to model which patients in the complier/noncomplier mixtures are noncompliers, thus indirectly estimating the remaining portion of the mixture to belong to the stratum of compliers.

In this article, we employ a joint-estimation method using a Gibbs sampling computational approach [10, 13] in a setting where noncompliance exists in both randomization arms. We aim to improve the estimate of the CACE through incorporation of a compliance-predictive model that uses a key covariate to select compliers from the complier/noncomplier mixtures. Our novel contribution is a detailed exposition of scenarios in which observed data predict membership in the noncomplier strata in a way that can select compliers in each treatment group from different portions of the covariate distribution, potentially implying that the estimated CACE is biased for the causal effect of treatment. After introducing the motivating oral-surgery application in Section 2, Section 3 formally defines a potential-outcomes inference framework and the assumptions necessary for estimation of the CACE. Section 4 develops the compliance-predictive model and corresponding estimation procedure. Section 5 uses simulated examples to illustrate some posterior checks and illuminate the potential for bias resulting from the compliance-predictive model, and Section 6 illustrates the impact of using the key covariate to predict compliance status in the oral-surgery setting. We conclude with a discussion.

2. Motivating Oral-Surgery Clinical Trial. Our motivating example consists of 142 patients who were randomly assigned to receive treatment for jaw fractures in the form of Maxillomandibular Fixation (MMF, control) or Rigid Internal Fixation (RIF, active treatment). A degree of clinical flexibility was deemed essential to the protocol, allowing treatment decisions to depart from the randomized treatment assignment if deemed necessary by the treating surgeon. This clinical latitude gives rise to possible concerns that more severely injured patients were disproportionately selected into the more aggressive treatment arm, as it is well accepted in the surgical community that the MMF procedure, which is less expensive, is appropriate for less severe injuries while the RIF procedure, which is more resource-intensive, is appropriate for more severe injuries. Although the exact rationale for treatment decisions was not recorded, a continuously-scaled measure of injury severity (*SEV*) was calculated for each patient. This severity measure, originally

developed as the Mandible Injury Severity Score (MISS) [33], ranges from 0 (less severe) to 25 (extremely severe), and derives from anatomic and clinical characteristics of the constituent jaw fractures. The outcome of interest was a continuously-scaled General Oral Health Assessment Index (GOHAI) [2] measured at six months post-treatment, with higher values suggesting better oral-health quality of life. In the face of “non-compliance” (i.e., surgical judgment overriding the treatment assigned through the randomization protocol), one could conduct intention-to-treat and as-treated analyses [34], but the former addresses a question that is arguably not the only scientific question of interest, the latter can give rise to bias in estimates of the treatment effect, and neither accounts for the plausible effect that subjective treatment decisions had on the analysis.

3. Potential Outcomes, Principal Strata, and Causal Estimand. Definition of principal strata and causal estimands requires development of a potential-outcomes framework, often called the Rubin Causal Model [18, 30]. Following previous development in the setting of all-or-nothing treatment noncompliance in a two-arm clinical trial [1], we define potential outcomes and delineate the principal strata that arise in our motivating setting. We then outline the assumptions necessary for identifiability of the causal estimand of interest, the CACE.

3.1. Potential outcomes and principal strata. First, we define the relevant potential outcomes inherent in this clinical trial. Define \mathbf{Z} as the vector of random treatment assignments for all patients in the study, with i^{th} element Z_i equal to 0 for assignment to MMF and 1 for assignment to RIF. Let $\mathbf{D}(\mathbf{Z})$ be a vector with i^{th} element $D_i(\mathbf{Z})$ denoting the i^{th} patient’s received treatment under assignment \mathbf{Z} . Patients with $D_i(\mathbf{Z}) = 0$ would receive treatment with MMF under assignment \mathbf{Z} , while patients with $D_i(\mathbf{Z}) = 1$ would receive treatment with RIF under assignment \mathbf{Z} . Furthermore, we use $Y_i(\mathbf{Z}, \mathbf{D})$ to denote a patient’s potential GOHAI with respect to \mathbf{Z} and \mathbf{D} . We adopt the stable unit treatment value assumption (SUTVA) [30] here to indicate no interference between patients, allowing us to write $D_i(\mathbf{Z}) = D_i(Z_i)$ and $Y_i(\mathbf{Z}, \mathbf{D}) = Y_i(Z_i)$.

Principal strata in this setting are defined by all four possible values of the pair $(D_i(0), D_i(1))$. We call the principal stratum of patients with $(D_i(0) = 0, D_i(1) = 1)$ “compliers” who will receive the assigned treatment regardless of which treatment is assigned, and denote these patients as having $S_i = c$. Similarly, we can call the stratum with $(D_i(0) = 0, D_i(1) = 0)$ “never-takers” who will never receive treatment with RIF, denoting these patients with $S_i = n$, and the stratum of patients with $(D_i(0) = 1, D_i(1) = 1)$ “always-takers” who will always receive treatment with RIF, which we label with $S_i = a$. Finally, we define the principal stratum of “defiers” as those with $(D_i(0) = 1, D_i(1) = 0)$, or those who will always receive the treatment opposite of that assigned, with $S_i = d$.

Naturally, we observe only one component of $(D_i(0), D_i(1))$ and only one component of $(Y_i(0), Y_i(1))$. To draw out this distinction between observed and missing components, we write (D_i^{obs}, D_i^{mis}) , and (Y_i^{obs}, Y_i^{mis}) where the superscripts *obs* and *mis* denote the observed and missing potential outcomes, respectively.

3.2. Assumptions for identifiability of causal estimands. As no complete pair of potential outcomes is observable, we require additional assumptions for identifiability of causal estimands. In addition to SUTVA, we adopt a monotonicity assumption [21] disallowing the existence of the principal stratum of defiers, i.e., there are no patients who would receive MMF if assigned RIF but receive RIF if assigned MMF. This setting with noncompliance resulting from clinicians’ judgment is unlikely to produce a violation of the monotonicity assumption. The usefulness of monotonicity lies in its implication that patients with $D_i^{obs} = D_i(0) = 1$ must belong to the stratum of always-takers and those with $D_i^{obs} = D_i(1) = 0$ must belong to the stratum of never-takers. Stratum membership for those who received the assigned treatment remains unidentified, as patients with $D_i^{obs} = D_i(0) = 0$ represent a mixture of compliers and never-takers, while those with $D_i^{obs} = D_i(1) = 1$ represent a mixture of compliers and always-takers. The first three columns of Table 1 provide a summary of the possible principal strata for patients with each possible observed pattern of Z_i and D_i^{obs} .

TABLE 1

Possible principal strata for observed treatment assignment and receipt patterns and summary statistics for SEV and GOHAI in the motivating oral-surgery setting.

Treatment Assigned, Z_i	Treatment Received, $D_i^{obs}(Z_i)$	Possible principal Strata ($D_i(0), D_i(1)$)	n	Mean (SD) SEV	Mean (SD) GOHAI
0	0	compliers or never-takers ($D_i(0) = 0, D_i(1) = 0$ or 1)	53	12.8 (2.7)	42.8 (12.1)
0	1	always-takers ($D_i(0) = 1, D_i(1) = 1$)	9	14.0 (2.0)	42.8 (11.9)
1	1	compliers or always-takers ($D_i(0) = 1$ or 0, $D_i(1) = 1$)	40	13.2 (2.3)	44.5 (12.1)
1	0	never-takers ($D_i(0) = 0, D_i(1) = 0$)	40	12.2 (3.0)	41.7 (9.3)

Inference for causal effects in clinical trials with treatment noncompliance typically relies on another assumption, known as the exclusion restriction [1], stating that any effect of treatment assignment, Z , on the outcome, Y , must be via an effect of treatment received, D . After accounting for received treatment, random assignment no longer affects GOHAI, or $(Y(z)|Z = z, D(z) = d) = (Y(z)|D(z) = d)$ for $z = 0, 1$ and $d = 0, 1$.

With the above development, we define the CACE as the expected difference in potential GOHAI outcomes within the stratum of compliers:

$$CACE = E[Y_i(1) - Y_i(0)|S_i = c].$$

4. Bayesian Models for the CACE with the Compliance-Predictive Feature. We formulate our inference strategy with a phenomenological Bayesian model following [22]. The model is phenomenological in the sense described by [30, 31], where the inference builds on potentially observable quantities even though not all of the quantities will be observed. The relevant random variables for each patient are $Z_i, D_i(0), D_i(1), Y_i(0), Y_i(1)$, and X_i , where X_i denotes the i^{th} patient's SEV. We consider these random variables realizations from a joint distribution, with X_i, Z_i, D_i^{obs} , and Y_i^{obs} observed for each patient. Our goal is to model the conditional distributions of $Y_i(Z)$ conditional on principal stratum, which requires integration over missing values as a result of the unidentifiable mixtures over the latent S_i . This motivates a Gibbs-sampling strategy that first samples the missing S_i , thereby allowing assessment of the distributions of $Y_i(Z)$ conditional on the “complete compliance data” consisting of subpopulations without mixture components.

4.1. *Structure of Bayesian inference.* The joint distribution of the data can be factored as follows:

$$(4.1) \quad f(\mathbf{Z}, (\mathbf{Y}(0), \mathbf{Y}(1)), (\mathbf{D}(0), \mathbf{D}(1)), \mathbf{X}) = f(\mathbf{Z}, \mathbf{Y}, \mathbf{S}, \mathbf{X}) = f(\mathbf{Y}, \mathbf{S}, \mathbf{X}|\mathbf{Z})f(\mathbf{Z}) = f(\mathbf{Y}, \mathbf{S}, \mathbf{X})f(\mathbf{Z})$$

where the last equality holds due to randomization in the study design. We facilitate Bayesian inference by writing the joint distribution of \mathbf{Y}, \mathbf{S} and \mathbf{X} as the product of independently identically distributed random variables conditional on a generic parameter θ [4], where we denote the prior distribution of θ as $p(\theta)$ and the posterior distribution of θ as:

$$(4.2) \quad p(\theta|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{X}, \mathbf{Z}) \propto p(\theta) \int \prod f(Y_i^{obs}, Y_i^{mis}, D_i^{obs}, D_i^{mis}, X_i|\theta) dY_i^{mis} dD_i^{mis}.$$

As pointed out in [8] and [23], required integration over \mathbf{D}^{mis} proves computationally difficult in general, but as a result of randomization \mathbf{Y}^{mis} can be handled with standard randomization-based tools. Furthermore, the difficult integration over \mathbf{D}^{mis} leads us to consider the joint posterior of $(\theta, \mathbf{D}^{mis})$,

$$(4.3) \quad p(\theta, \mathbf{D}^{mis}|\mathbf{D}^{obs}, \mathbf{Y}^{obs}, \mathbf{X}, \mathbf{Z}) \propto p(\theta) \prod f(D_i^{obs}, D_i^{mis}, Y_i^{obs}, X_i|\theta),$$

which is proportional to a standard posterior distribution of θ had \mathbf{D}^{mis} been observed [23], further motivating the strategy of first drawing \mathbf{D}^{mis} and then sampling from the posterior distribution of θ conditional on complete compliance data. Posterior distributions of the relevant quantities follow from specification of both $p(\theta)$ and the models defined in Section 4.2. We describe our prior distributions for θ in the Section 4.4.

4.2. *Models for principal strata and outcomes.* To estimate the CACE, we further factor the joint distribution in (4.1) as $f(\mathbf{Y}|\mathbf{S}, \mathbf{X})f(\mathbf{S}|\mathbf{X})f(\mathbf{X})f(\mathbf{Z})$, and specify models for $f(\mathbf{Y}|\mathbf{S}, \mathbf{X})$ and $f(\mathbf{S}|\mathbf{X})$. As the population consists of three underlying strata, we follow the approach used in [8] and [3] whereby we model $f(\mathbf{S}|\mathbf{X})$ with two linked probit models, the first modeling membership in the never-taker stratum and the second modeling membership in the complier stratum conditional on exclusion from the never-taker stratum. We parameterize these models as

$$(4.4) \quad \begin{aligned} \Psi_n(X_i, \beta) &= P(S_i = n|X_i, \beta) = 1 - \Phi(\beta_{00} + \beta_{01}X_i), \\ \Psi_c(X_i, \beta) &= P(S_i = c|X_i, \beta) = \{1 - \Psi_n(X_i, \beta)\}\{1 - \Phi(\beta_{10} + \beta_{11}X_i)\}, \text{ and} \\ \Psi_a(X_i, \beta) &= P(S_i = a|X_i, \beta) = 1 - \Psi_n(X_i, \beta) - \Psi_c(X_i, \beta) \end{aligned}$$

where $\beta = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})$ and Φ is the standard normal cumulative distribution function. To facilitate computation, we represent these models as arising from underlying continuous random variables S_i^n and S_i^c ,

$$(4.5) \quad \begin{aligned} S_i &= n & \text{if } S_i^n &= \beta_{00} + \beta_{01}X_i + V_i \leq 0, \\ S_i &= c & \text{if } S_i^n > 0 \text{ and } S_i^c &= \beta_{10} + \beta_{11}X_i + U_i \leq 0, \text{ and} \\ S_i &= a & \text{if } S_i^n > 0 \text{ and } S_i^c > 0 \end{aligned}$$

where the V_i and U_i are independently distributed as $N(0, 1)$.

We illustrate the analysis with two different models for $f(\mathbf{Y}|\mathbf{S}, \mathbf{X})$. The first model (Model A) entails a regression adjustment for the key covariate's association with the outcome:

$$(4.6) \quad \begin{aligned} f(Y_i(z)|X_i, S_i = n) &= g_n(Y_i|\alpha_0^n, \alpha_1^n, X_i, \sigma^2) \sim N(\alpha_0^n + \alpha_1^n X_i, \sigma^2), \\ f(Y_i(z)|X_i, S_i = a) &= g_a(Y_i|\alpha_0^a, \alpha_1^a, X_i, \sigma^2) \sim N(\alpha_0^a + \alpha_1^a X_i, \sigma^2) \quad \text{and} \\ f(Y_i(z)|X_i, S_i = c, Z_i = z) &= g_{cz}(Y_i|\alpha_0^{cz}, \alpha_1^{cz}, X_i, \sigma^2) \sim N(\alpha_0^{cz} + \alpha_1^{cz} X_i, \sigma^2) \quad \text{for } z = 0, 1, \end{aligned}$$

implying the exclusion restriction and the assumption that *GOHAI* outcomes are distributed with the same variance in each stratum and for each treatment receipt. For comparison purposes, we also conduct the analysis under another model (Model B) that does not explicitly incorporate X in the model for $Y(z)$, entailing the additional assumption that $Y(z) \perp\!\!\!\perp X|S$. That is, Model B incorporates the restriction that $\alpha_1^n = \alpha_1^a = \alpha_1^{c0} = \alpha_1^{c1} = 0$, representing a ‘‘standard’’ unadjusted CACE analysis.

The observed-data likelihood reflecting the mixtures over the latent S_i can be written as:

$$(4.7) \quad \begin{aligned} L_{\text{obs}}(\theta|\mathbf{Z}, \mathbf{D}^{\text{obs}}, \mathbf{Y}^{\text{obs}}, \mathbf{X}) &= \prod_{Z_i=1, D_i^{\text{obs}}=0} \{\Psi_n(X_i, \beta) \cdot g_n(Y_i|\alpha_0^n, \alpha_1^n, X_i, \sigma^2)\} \times \\ &\quad \prod_{Z_i=0, D_i^{\text{obs}}=1} \{\Psi_a(X_i, \beta) \cdot g_a(Y_i|\alpha_0^a, \alpha_1^a, X_i, \sigma^2)\} \times \\ &\quad \prod_{Z_i=0, D_i^{\text{obs}}=0} \{\Psi_n(X_i, \beta) \cdot g_n(Y_i|\alpha_0^n, \alpha_1^n, X_i, \sigma^2) + \Psi_c(X_i, \beta) \cdot g_{c0}(Y_i|\alpha_0^{c0}, \alpha_1^{c0}, X_i, \sigma^2)\} \times \\ &\quad \prod_{Z_i=1, D_i^{\text{obs}}=1} \{\Psi_a(X_i, \beta) \cdot g_a(Y_i|\alpha_0^a, \alpha_1^a, X_i, \sigma^2) + \Psi_c(X_i, \beta) \cdot g_{c1}(Y_i|\alpha_0^{c1}, \alpha_1^{c1}, X_i, \sigma^2)\} \end{aligned}$$

where $\theta = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}, \alpha_0^n, \alpha_1^n, \alpha_0^a, \alpha_1^a, \alpha_0^{c0}, \alpha_1^{c0}, \alpha_0^{c1}, \alpha_1^{c1}, \sigma^2)$ and the product over $Z_i = z, D_i^{obs} = d$ represents the product over all patients assigned treatment z who were observed to receive treatment d .

As a result of random assignment to treatment, the Y_i^{mis} in the stratum of compliers is sampled from the distribution $g_{c(1-Z_i)}$ and the CACE estimate is calculated as

$$CACE = E[Y_i(1) - Y_i(0) | X_i, S_i = c] = \frac{1}{n_c} \sum_{S_i=c} (Y_i(1) - Y_i(0))$$

where n_c is the number of patients with $S_i = c$ at the current iteration.

4.3. Sampling compliers within the compliance-predictive model. Despite the existence of three underlying strata, the step of the Gibbs sampler that determines patients' unknown compliance status does so via Bernoulli distributions reflecting the fact that patients who received the assigned treatment can belong to one of only two possible strata. Owing to these underlying two-component mixtures, the probability at a given iteration of the sampler that a patient with $Z_i = D_i^{obs} = z$ belongs to stratum of compliers is:

$$(4.8) \quad P(S_i = c | X_i, Y_i^{obs}, D_i^{obs}, Z_i, \theta) = \frac{\Psi_c(X_i, \beta) \cdot g_{cz}(Y_i | \alpha_0^{cz}, \alpha_1^{cz}, X_i, \sigma^2)}{(\Psi_c(X_i, \beta) \cdot g_{cz}(Y_i | \alpha_0^{cz}, \alpha_1^{cz}, X_i, \sigma^2) + \Psi_t(X_i, \beta) \cdot g_t(Y_i | \alpha_0^t, \alpha_1^t, X_i, \sigma^2))},$$

where $t = n$ if $z = 0$ and $t = a$ if $z = 1$. Examining these probabilities makes clear that the relative impacts of X_i and Y_i^{obs} on (4.8) depend on the extent to which X predicts stratum and on the amount of overlap between the distributions g_{cz} and g_t .

4.4. Additional model specifications and statistical computing details. We treat the elements of θ to be *a priori* independent, using conditionally-conjugate normal distributions for the $\beta, \alpha_0^n, \alpha_1^n, \alpha_0^a, \alpha_1^a, \alpha_0^{cz}, \alpha_1^{cz}$, and a conditionally-conjugate gamma distribution for the precision parameter $\frac{1}{\sigma^2}$. The distributions for $(\alpha_0^n, \alpha_0^a, \alpha_0^{cz})$ are centered at the overall sample mean *GOHAI* with variances of 100, and the distributions for $(\alpha_1^n, \alpha_1^a, \alpha_1^{cz})$ are centered at 0 with variances of 100. The prior distribution for the precision parameter is gamma with shape and scale parameter set to 0.01. Prior distributions for the elements of β are centered at 0 with variance 5.

After a burn-in of 5,000 iterations, each chain is run for 5,000 additional iterations, saving every 10^4 sample. For each model, three chains are run from different starting values, and the potential scale-reduction statistics [11, 12] are calculated for each parameter to assess convergence. All parameters in all models had potential scale-reduction statistics less than or equal to 1.06, suggesting satisfactory convergence. For each model, the three chains are combined to calculate posterior estimates.

5. Illustration of the Potential for Bias in the CACE Using Simulated Data. To illustrate that the compliance-predictive model can imply complier treatment groups with different characteristics and to illustrate our graphical diagnostic, we examine in detail a simulated scenario where X is predictive of stratum membership and the true CACE=0. Details of this simulation and a broader simulation study appear in a supplementary web appendix.

To investigate the relationships between X and stratum membership under Model A, we examine posterior-predictive distributions of the probabilities in (4.8) for a hypothetical group of patients having an X distribution mirroring that in the observed data. Figure 1(a) displays, for $z = 0, 1$, histograms of the observed X distributions in patients with $Z_i = D_i^{obs} = z$, with histogram bars shaded according to the mean posterior-predictive probability of membership in the complier stratum for a value of X at that point of the histogram and for Y_i^{obs} equal to the mean value observed in patients with $Z_i = D_i^{obs} = z$. Note the different shading patterns in the two histograms. For $Z_i = D_i^{obs} = 0$, histogram bars are darker as X increases (more severely injured patients are more likely compliers), while for $Z_i = D_i^{obs} = 1$, histogram bars are lighter as X increases

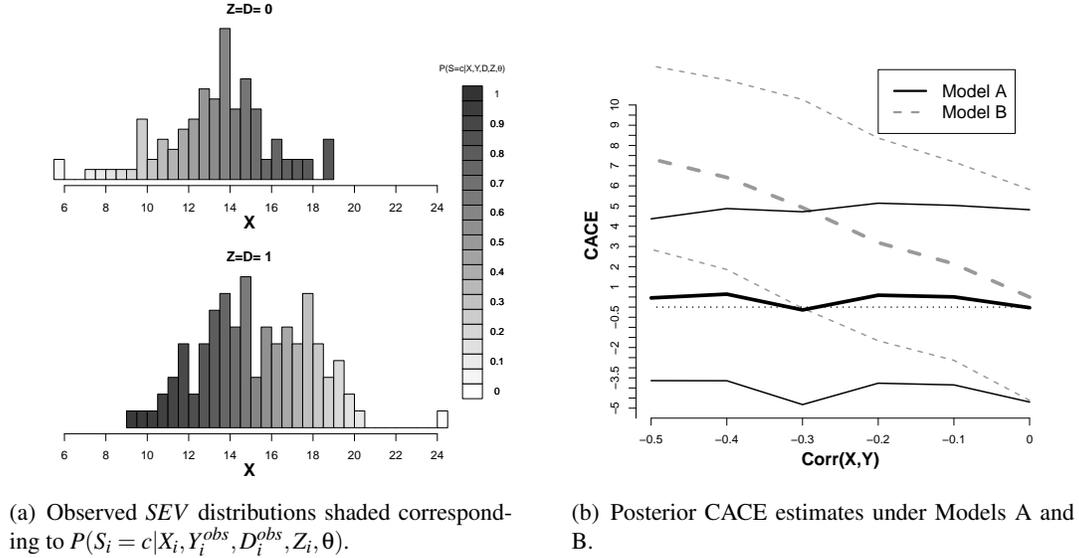


FIG 1. Results from simulated datasets where X predicts stratum membership. As the procedure selects compliers from opposite ends of the severity distribution (1(a)), estimates of the CACE can become particularly susceptible to model misspecification (1(b)). Thick lines in (1(b)) are posterior means and thin lines are 95% posterior intervals. For each value of $Corr(X, Y)$, posterior summaries are averaged over 50 Monte Carlo simulations. All simulations have $CACE=0$ (horizontal dotted line).

(more severely injured patients are less likely compliers). For example, note that patients with X in the range $[9, 12]$ in the $Z_i = D_i^{obs} = 1$ group have probability of membership in the complier stratum near 1.0, while patients with the same range of SEV in the $Z_i = D_i^{obs} = 0$ group have probability of membership in the complier stratum in the range $[0.1 - 0.4]$. The implication for estimates of the CACE is that over the course of the sampler, patients in the observed mixture of compliers and always-takers ($Z_i = D_i^{obs} = 1$) with lower X will more often contribute to the CACE than patients with comparable X in the observed mixture of compliers and never-takers ($Z_i = D_i^{obs} = 0$). The opposite sampling disparity holds for patients with higher X . If X is also related to the primary outcome, Y , a situation such as that depicted in Figure 1(a) leaves estimates of the CACE particularly vulnerable to misspecified models that incorrectly extrapolate to areas of the X distribution where there is limited data. For example, estimation of a typical unadjusted CACE (Model B) would represent one such misspecified model, and could lead to vastly different estimates of the CACE. To illustrate this point, Figure 1(b) displays posterior estimates of the CACE using both Model A and Model B in scenarios where X is related to stratum membership and with varying magnitudes of the relationship between X and Y . We see that under Model B, the imbalanced sampling of compliers evident from Figure 1(a) leads to bias in the estimated CACE that is increasing in $|Corr(X, Y)|$, providing misleading results even when the association between X and Y is modest and in some cases estimating a significant treatment effect when there in fact is none. The same bias is not depicted under Model A because even though there is limited data on comparable compliers in some areas of the X distribution, extrapolation of Model A to these areas of the distribution correctly reflects the underlying relationship; that is, there is no model misspecification. The supplementary web appendix considers simulations under a broader range of relationships between X and both stratum membership and Y and further indicates the potential for bias in the CACE when using a compliance-predictive covariate.

6. Using SEV to Predict Principal Strata in the Motivating Oral-Surgery Study. As described in Table 1, patients in the oral-surgery example who had assignment to MMF overruled (known always-takers) had higher average X than the rest of the sample, patients who had assignment to RIF overruled (known

never-takers) had lower average X than the rest of the sample, and there was a relatively high estimated proportion of never-takers (50.0%) and a relatively low proportion of always-takers (14.5%). The oral-surgery example had missing Y for a substantial proportion of the patients. Based on observed data, the nonresponse rates were 48.4%, 46.2% in the $Z = 0, 1$ arms, respectively, and 55.6%, 45.0% in the observed always-takers, never-takers, respectively. To prevent complication of our illustrative goal, we assume in the models for the oral-surgery data that 1) the S_i are independent of the missing indicator and 2) the missing Y are latently ignorable conditional on S_i and Z_i [6]. The implication of these assumptions for the computation is that missing Y are drawn at each iteration from the distribution for patients' current stratum membership conditional on current values of the parameters. Furthermore, the small number of observed Y values precludes useful estimation of all of the α parameters in (4.6), leading us to alter Model A to Model A* that includes the constraint that $\alpha_1^0 = \alpha_1^1 = \alpha_1^{c0} = \alpha_1^{c1}$.

The observed relationship between X and membership in the never-taker and always-taker strata (Table 1) prompts examination of the probabilities of selection into the stratum of compliers within the compliance-predictive model. Figure 2(a) shows the posterior predictive distributions of the Bernoulli probabilities in (4.8) for hypothetical patients with Y equal to the observed sample mean, X across the range observed in the data, and $Z_i = D_i^{obs} = z$ for $z = 0, 1$. The unequal proportions of underlying strata are reflected in this figure by the fact that the probability of being sampled as a complier is consistently higher for the patients with $Z_i = D_i^{obs} = 1$ than for those in the other treatment arm; the low estimated proportion of always-takers ($9/62=0.14$) implies that most patients with $Z_i = D_i^{obs} = 1$ belong to the stratum of compliers.

The wide spread of the posterior predictive distributions in Figure 2(a) suggests that X has limited utility for identifying which patients are compliers, but there is some indication that the relationship between X and the probability of membership in the complier stratum is slightly different at the high end of the X distribution depending on the value of Z_i and D_i^{obs} . To assess the potential for these relationships to affect the sampling of compliers, we examine in Figure 2(b) the posterior-predictive probabilities of membership in the complier stratum for hypothetical patients with X distributions identical to those observed in the sample with $Z_i = D_i^{obs} = z$ and with Y equal to the mean value observed in patients with $Z_i = D_i^{obs} = z$, for $z = 0, 1$. This illustration provides limited evidence that patients with different values of Z_i and D_i^{obs} are sampled as compliers from different areas of their respective X distributions. There is a slight positive association between X and membership in the complier stratum in the $Z_i = D_i^{obs} = 0$ patients (evidenced by the darkening of the histogram bars as X increases) that differs from the negative association in the $Z_i = D_i^{obs} = 1$ patients (evidenced by the lightening of the histogram bars as X increases), but the amount of uncertainty in these posterior probabilities likely precludes any serious effect on the estimated CACE.

Overall, the information in Figure 2 does not provide any strong indication that the compliance-predictive model estimates a CACE calculated from compliers with different injury characteristics in the two treatment groups. To explore the sensitivity to alternative models for stratum membership, we adapt Model A* to replace the probit models in (4.4) with a multinomial logit model along the lines of that used in [17], and refer to this as Model C*. Using Model C*, figures analogous to Figure 2(a) and 2(b) appear largely indistinguishable from those under Model A* and are not pictured. Table 2 summarizes posterior CACE estimates from a compliance-predictive analysis under Model A*, Model B, and Model C*, as well as from an analysis following [22] that does not explicitly use the SEV covariate at all and places a noninformative conditionally-conjugate Dirichlet prior distribution on the population proportions of principal strata. None of these models provides evidence of a treatment effect, and all three compliance-predictive models offer slightly decreased precision, most likely due to the lack of information contained in the SEV covariate regarding stratum membership and the inclusion of extraneous model parameters.

7. Discussion. Using covariates to model membership in latent principal strata has many advantages in estimating the CACE. We provide a detailed illustration of the subtlety involved in using a key covariate when noncompliance exists in both treatment arms. In particular, we show that when a covariate is related

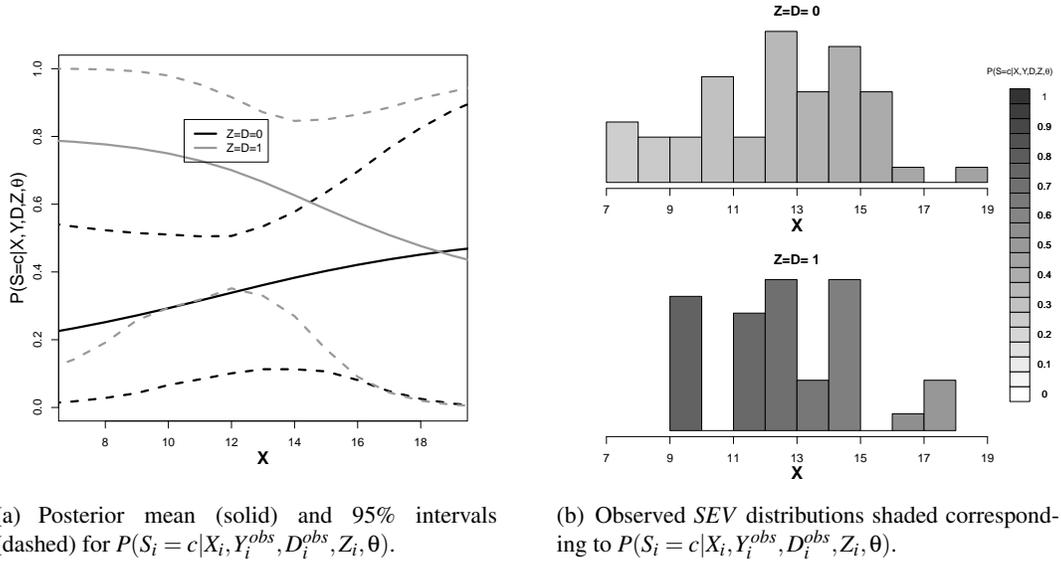


FIG 2. Posterior-predicted probabilities of membership in the complier stratum for hypothetical patients with $Z_i = D_i^{obs} = z$ in the oral-surgery study under Model A*, $z = 0, 1$.

TABLE 2

Posterior estimates of the CACE in the motivating oral-surgery study using a model without the compliance-predictive covariate and using three compliance-predictive strategies.

Modeling Strategy	Posterior Mean	SD	2.5%	97.5%
Compliance-predictive Model A*	2.65	7.3	-8.9	20.9
Compliance-predictive Model B	0.17	7.0	-13.0	16.0
Compliance-predictive Model C*	1.95	6.7	-9.6	18.9
Model without SEV	0.74	6.4	-11.0	13.6

to stratum membership, a joint-estimation method can imply treatment groups in the latent stratum of compliers with different covariate characteristics. The resulting danger of comparing compliers with different characteristics can be alleviated with modeling assumptions that correctly extrapolate the treatment effect to areas of the covariate distribution where compliers are not estimated to exist in both treatment arms. However, this differential sampling of patients into the complier stratum poses a serious threat to the CACE under model misspecification, including calculation of the standard unadjusted CACE when a covariate predicts stratum membership. We propose simple graphical posterior checks that indicate the extent to which the estimated CACE relies on compliers that have different covariate characteristics, potentially characterizing the danger for model misspecification to bias the estimated CACE.

Our aim is not to discourage the use of covariates that are predictive of latent stratum membership but rather to shed light on the subtleties involved and to provide guidance on how to detect whether a compliance-predictive model endangers estimates of the CACE. Our motivating oral-surgery example is somewhat unique in its availability of a key covariate that was thought to influence the treatment received, but the possibility of covariates relating to stratum membership can arise elsewhere, as with the randomized encouragement design considered in [17] where age and presence of chronic obstructive pulmonary disease (COPD) were thought to influence whether patients were in the underlying stratum of individuals who would always receive a flu vaccination regardless of random encouragement to do so. The authors of that work include compliance-predictive models to relax exclusion restrictions and provide posterior estimates of model parameters suggestive of a different relationship between age and COPD and the probability of membership in the complier stratum dependent on the values of Z_i and D_i^{obs} . Whether their model tended to consider a complier stratum consisting of younger patients without COPD in the $Z = 1$ arm and older patients with COPD in the $Z = 0$ arm could be assessed by examining posterior probabilities of stratum membership across the observed ranges of these covariates.

We present models that do and do not adjust the CACE for levels of the key covariate. We frame the choice not to model Y conditional on both S and X (as in Model B) as a form of model misspecification, but in real applications researchers are confronted with the decision to calculate the familiar unadjusted CACE or to specify a more detailed model for $f(\mathbf{Y}|\mathbf{S}, \mathbf{X})$ and estimate an adjusted CACE. We show that when a covariate is used to model stratum membership, estimation of the unadjusted CACE can produce biased results. Thus, we recommend that the CACE be adjusted for any covariates used to model stratum membership, which is contrary to previous recommendations that stratum-predictive covariates need not be included in models for outcomes within strata [9]. Furthermore, specification of a more detailed model for $f(\mathbf{Y}|\mathbf{S}, \mathbf{X})$ does not guarantee correctness, and we provide a framework to assess whether model misspecification poses a particular danger to estimation of a covariate-adjusted CACE that can depend on areas of the covariate distribution where there is limited data.

The core features of the scenario presented here, namely that a compliance-predictive model must respect the presence of three underlying strata while a patient of unknown stratum can belong to one of only two strata, can have conflicting impacts. One way to characterize these issues is to view modeling membership in the complier stratum not as selection of compliers but rather as a process for selection of “non-compliers” from both treatment arms since, no matter how predictive, the compliance-predictive feature is anchored to observed information on always-takers and never-takers and can only indirectly model membership in the stratum of primary interest. Some applications focus on treatment effects within principal strata analogous to always-takers [14, 19, 20, 29, 32] and are less susceptible to the type of bias depicted here because, as in settings where noncompliance exists in only one treatment arm, the data provide direct evidence on the relationship between covariates and the stratum of primary interest.

We have characterized scenarios that lend themselves to the use of a compliance-predictive covariate but leave an opening for bias in the estimation of the CACE. Such scenarios warrant careful model checking; in Sections 5 and 6 we propose steps to investigate the potential for bias. Future research on methods that use stratum-predictive covariates to estimate the CACE when the data consist of three underlying strata would

prove valuable in settings where it is appealing to use covariates to aid identifiability or improve precision of causal estimates.

Acknowledgements. This work received support from NIDA grants 1-R01-DA025680 and 1-R01-DA016850 as well as NIMH grants 5-P30-MH58017 and 1-P30-MH082760.

References.

- [1] ANGRIST, J. D., IMBENS, G. W., AND RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 434, 444–455.
- [2] ATCHISON, K. (1997). The general oral health assessment index. In *Measuring Oral Health and Quality of Life*. 71–80.
- [3] BARNARD, J., FRANGAKIS, C. E., HILL, J. L., AND RUBIN, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association* **98**, 462, 299–324.
- [4] DE FINETTI, B. (1974). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons.
- [5] FOLLMANN, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association* **95**, 452, 1101–1109.
- [6] FRANGAKIS, C. E. AND RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 2 (June), 365–379, <http://biomet.oxfordjournals.org/cgi/content/abstract/86/2/365>.
- [7] FRANGAKIS, C. E. AND RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 1 (Mar.), 21–29, <http://www.blackwell-synergy.com/doi/abs/10.1111/j.0006-341X.2002.00021.x>.
- [8] FRANGAKIS, C. E., RUBIN, D. B., AND ZHOU, X. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* **3**, 2 (June), 147–64, <http://www.ncbi.nlm.nih.gov/pubmed/12933609>.
- [9] GALLOP, R., SMALL, D. S., LIN, J. Y., ELLIOTT, M. R., JOFFE, M., AND TEN HAVE, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine* **28**, 7, 1108–1130, <http://dx.doi.org/10.1002/sim.3533>.
- [10] GELFAND, A. E. AND SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 410 (June), 398–409, <http://www.jstor.org/stable/2289776>.
- [11] GELMAN, A., CARLIN, J. B., STERN, H. S., AND RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, New York.
- [12] GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 4, 457–472.
- [13] GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 6, 721–741.
- [14] GILBERT, P. B., BOSCH, R. J., AND HUDGENS, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59**, 3, 531–541.
- [15] GRIFFIN, B. A., MCCAFFREY, D. F., AND MORRAL, A. R. (2008). An application of principal stratification to control for institutionalization at follow-up in studies of substance abuse treatment programs. *Annals of Applied Statistics* **2**, 3 (Nov.), 1034–1055, <http://arxiv.org/abs/0811.1831>.
- [16] HILL, J. L., BROOKS-GUNN, J., AND WALDFOGEL, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology* **39**, 4, 730–744.
- [17] HIRANO, K., IMBENS, G. W., RUBIN, D. B., AND ZHOU, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 1 (Mar.), 69–88, <http://biostatistics.oxfordjournals.org/cgi/content/abstract/1/1/69>.
- [18] HOLLAND, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 396 (Dec.), 945–960, <http://www.jstor.org/stable/2289064>.
- [19] HUDGENS, M. G. AND HALLORAN, M. E. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association* **101**, 51–64, <http://ideas.repec.org/a/bes/jnlasa/v101y2006p51-64.html>.
- [20] HUDGENS, M. G., HOERING, A., AND SELF, S. G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* **22**, 14, 2281–2298, <http://dx.doi.org/10.1002/sim.1394>.
- [21] IMBENS, G. W. AND ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–467.
- [22] IMBENS, G. W. AND RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* **25**, 1 (Feb.), 305–327, <http://www.jstor.org/stable/2242722>.
- [23] JIN, H. AND RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman data. *Journal of the American Statistical Association* **103**, 481, 101–111.
- [24] JO, B. AND STUART, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine* **28**, 23 (Oct.), 2857–2875, <http://www.ncbi.nlm.nih.gov/pubmed/19610131>.

- [25] JOFFE, M. M., SMALL, D., AND HSU, C. Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science* **22**, 1, 74–97.
- [26] JOFFE, M. M., TEN HAVE, T. R., AND BRENSINGER, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics* **4**, 3, 327–340.
- [27] MCCLELLAN, M., MCNEIL, B. J., AND NEWHOUSE, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association* **272**, 11 (Sept.), 859–866, <http://jama.ama-assn.org/cgi/content/abstract/272/11/859>.
- [28] ROSENBAUM, P. R. AND RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 1 (Apr.), 41–55, <http://www.jstor.org/stable/2335942>.
- [29] 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**, 2 (Apr.), 277–289, <http://www.ncbi.nlm.nih.gov/pubmed/17681993>.
- [30] RUBIN, D. B. (1978a). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6**, 1 (Jan.), 34–58, <http://www.jstor.org/stable/2958688>.
- [31] RUBIN, D. B. (1978b). Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. 20–34.
- [32] SHEPHERD, B. E., GILBERT, P. B., JEMIAI, Y., AND ROTNITZKY, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62**, 2 (June), 332–342, <http://www.blackwell-synergy.com/doi/abs/10.1111/j.1541-0420.2005.00495.x>.
- [33] SHETTY, V., ATCHISON, K., DER-MATIROSIAN, C., WANG, J., AND BELIN, T. R. (2007). The mandible injury severity score: Development and validity. *Journal of Oral and Maxillofacial Surgery* **65**, 4 (Apr.), 663–670, <http://www.sciencedirect.com/science/article/B6WKF-4N8JDW3-K/2/bb21580f3c127f83826fcdabd74a4e65>.
- [34] SHETTY, V., ATCHISON, K., LEATHERS, R., BLACK, E., ZIGLER, C., AND BELIN, T. R. (2008). Do the benefits of rigid internal fixation of mandible fractures justify the added costs? Results from a randomized controlled trial. *Journal of Oral and Maxillofacial Surgery* **66**, 11 (Nov.), 2203–12, <http://www.ncbi.nlm.nih.gov/pubmed/18940481>.

CORWIN M. ZIGLER
HARVARD SCHOOL OF PUBLIC HEALTH
BUILDING 2, 4TH FLOOR
655 HUNTINGTON AVE
BOSTON, MA 02115
E-MAIL: czigler@hsph.harvard.edu

THOMAS R. BELIN
DEPARTMENT OF BIostatISTICS
UCLA SCHOOL OF PUBLIC HEALTH
51-267 CENTER FOR HEALTH SCIENCES
LOS ANGELES, CA 90095-1772
E-MAIL: tbelin@mednet.ucla.edu