




A Semiparametric Bayesian Approach to Dropout in Longitudinal Studies With Auxiliary Covariates

Tianjian Zhou, Michael J. Daniels & Peter Müller


To cite this article: Tianjian Zhou, Michael J. Daniels & Peter Müller (2020) A Semiparametric Bayesian Approach to Dropout in Longitudinal Studies With Auxiliary Covariates, Journal of Computational and Graphical Statistics, 29:1, 1-12, DOI: [10.1080/10618600.2019.1617159](https://doi.org/10.1080/10618600.2019.1617159)



To link to this article: <https://doi.org/10.1080/10618600.2019.1617159>

 [View supplementary material](#) 

 Published online: 02 Jul 2019.

 [Submit your article to this journal](#) 

 Article views: 544

 [View related articles](#) 

 [View Crossmark data](#) 

 Citing articles: 1 [View citing articles](#) 



A Semiparametric Bayesian Approach to Dropout in Longitudinal Studies With Auxiliary Covariates

Tianjian Zhou^{a,b}, Michael J. Daniels^c, and Peter Müller^d

^aDepartment of Public Health Sciences, The University of Chicago, Chicago, IL; ^bDepartment of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX; ^cDepartment of Statistics, University of Florida, Gainesville, FL; ^dDepartment of Mathematics, The University of Texas at Austin, Austin, TX

ABSTRACT

We develop a semiparametric Bayesian approach to missing outcome data in longitudinal studies in the presence of auxiliary covariates. We consider a joint model for the full data response, missingness, and auxiliary covariates. We include auxiliary covariates to “move” the missingness “closer” to missing at random. In particular, we specify a semiparametric Bayesian model for the observed data via Gaussian process priors and Bayesian additive regression trees. These model specifications allow us to capture nonlinear and nonadditive effects, in contrast to existing parametric methods. We then separately specify the conditional distribution of the missing data response given the observed data response, missingness, and auxiliary covariates (i.e., the extrapolation distribution) using identifying restrictions. We introduce meaningful sensitivity parameters that allow for a simple sensitivity analysis. Informative priors on those sensitivity parameters can be elicited from subject-matter experts. We use Monte Carlo integration to compute the full data estimands. Performance of our approach is assessed using simulated datasets. Our methodology is motivated by, and applied to, data from a clinical trial on treatments for schizophrenia. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received September 2017
Revised December 2018

KEYWORDS

Bayesian inference; Gaussian process; Longitudinal data; Missing data; Semiparametric model; Sensitivity analysis

1. Introduction

In longitudinal clinical studies, the research objective is often to make inference on a subject's full data response conditional on covariates that are of primary interest; for example, to calculate the treatment effect of a test drug at the end of a study. However, the vector of responses for a research subject is often incomplete due to dropout. Dropout is typically nonignorable (Rubin 1976; Daniels and Hogan 2008) and in such cases the joint distribution of the full data response and missingness needs to be modeled. In addition to the covariates that are of primary interest, we would often have access to some *auxiliary covariates* (often collected at baseline) that are not desired in the model for the primary research question. Such variables can often provide information about the missing responses and missing data mechanism. For example, missing at random (MAR) (Rubin 1976) might only hold conditionally on auxiliary covariates (Daniels and Hogan 2008). In this setting, auxiliary covariates should be incorporated in the joint model as well, but we should proceed with inference unconditional on these auxiliary covariates.

The full data distribution can be factored into the observed data distribution and the extrapolation distribution (Daniels and Hogan 2008). The observed data distribution can be identified by the observed data, while the extrapolation distribution cannot. Identifying the extrapolation distribution relies on untestable assumptions such as parametric models for the full

data distribution or identifying restrictions (Linero and Daniels 2018). Such assumptions can be indexed by unidentified parameters called *sensitivity parameters* (Daniels and Hogan 2008). The observed data do not provide any information to estimate the sensitivity parameters. Under the Bayesian paradigm, informative priors can be elicited from subject-matter experts and be placed on those sensitivity parameters. Finally, it is desirable to conduct a *sensitivity analysis* (Daniels and Hogan 2008; National Research Council 2011) to assess the sensitivity of inferences to such assumptions. The inclusion of auxiliary covariates can ideally reduce the extent of sensitivity analysis that is needed for drawing accurate inferences.

In this article, we propose a Bayesian semiparametric model for the joint distribution of the full data response, missingness, and auxiliary covariates. We use identifying restrictions to identify the extrapolation distribution and introduce sensitivity parameters that are meaningful to subject-matter experts and allow for a simple sensitivity analysis.

1.1. Missing Data in Longitudinal Studies

Literature about longitudinal missing data with nonignorable dropout can be mainly divided into two categories: likelihood-based and moment-based (semiparametric). Likelihood-based approaches include selection models (e.g., Heckman 1979; Diggle and Kenward 1994; Molenberghs, Kenward, and Lesaffre

1997), pattern mixture models (e.g., Little 1993; Little 1994; Hogan and Laird 1997), and shared-parameter models (e.g., Wu and Carroll 1988; Follmann and Wu 1995; Pulkstenis, Ten Have, and Landis 1998; Henderson, Diggle, and Dobson 2000). These three types of models differ from how the joint distribution of the response and missingness is factorized. Likelihood-based approaches often make strong parametric model assumptions to identify the full data distribution. For a comprehensive review see, for example, Daniels and Hogan (2008) or Little and Rubin (2014). Moment-based approaches, on the other hand, typically specify a semiparametric model for the marginal distribution of the response, and a semiparametric or parametric model for the missingness conditional on the response. Moment-based approaches are in general more robust to model misspecification since they make minimal distributional assumptions (see, e.g., Robins, Rotnitzky, and Zhao 1995; Rotnitzky, Robins, and Scharfstein 1998; Scharfstein, Rotnitzky, and Robins 1999; Tsiatis 2007; Tsiatis, Davidian, and Cao 2011).

There are several recent papers under the likelihood-based paradigm that are relevant to our approach, such as Wang et al. (2010), Linero and Daniels (2015), Linero (2017), and Linero and Daniels (2018). These papers specify Bayesian semiparametric or nonparametric models for the observed data distribution, and thus have similar robustness to moment-based approaches. However, existing approaches do not utilize information from auxiliary covariates. We will highlight more of our contribution and distinction compared to existing methods, in particular Linero and Daniels (2015) and Linero (2017), after we have introduced the required notation. In the presence of auxiliary covariates, Daniels, Wang, and Marcus (2014) model longitudinal binary responses using a parametric model under ignorable missingness. Our goal is to develop a flexible Bayesian approach to longitudinal missing data with nonignorable dropout that also allows for incorporating auxiliary covariates. As mentioned earlier, the reason to include auxiliary covariates is that we anticipate it will make the missingness “closer” to MAR.

1.2. Notation and Terminology

We introduce some notation and terminology as follows. Consider the responses for a subject i at J time points. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ})$ be the vector of longitudinal outcomes that was planned to be collected, $\tilde{\mathbf{Y}}_{ij} = (Y_{i1}, \dots, Y_{ij})$ be the history of outcomes through the first j times, and $\tilde{\mathbf{Y}}_{ij} = (Y_{i,j+1}, \dots, Y_{iJ})$ be the future outcomes after time j . Let S_i denote the dropout time or dropout pattern, which is defined as the last time a subject’s response is recorded, that is, $S_i = \max\{j : Y_{ij} \text{ is observed}\}$. Missingness is called *monotone* if Y_{ij} is observed for all $j \leq S_i$, and missingness is called *intermittent* if Y_{ij} is missing for some $j < S_i$. For monotone missingness, S_i captures all the information about missingness. In the following discussion, we will concern ourselves with monotone missingness. Dropout is called *random* (Diggle and Kenward 1994) if the dropout process only depends on the observed responses, that is, the missing data are MAR; dropout is called *informative* if the dropout process also depends on the unobserved responses, that is, the missing data are missing not at random (MNAR).

We denote by \mathbf{X}_i the covariates that are of primary interest, and $\mathbf{V}_i = (V_{i1}, \dots, V_{iQ})$ the Q auxiliary covariates that are not of primary interest. Those auxiliary covariates should be related to the outcome and missingness. The observed data for subject i is $(\tilde{\mathbf{Y}}_{iS_i}, S_i, \mathbf{V}_i, \mathbf{X}_i)$, and the full data is $(\mathbf{Y}_i, S_i, \mathbf{V}_i, \mathbf{X}_i)$. In general, we are interested in expectation of the form $E[t(\mathbf{Y}_i) \mid \mathbf{X}_i]$, where t denotes some functional of \mathbf{Y}_i . Finally, denote by $p(\mathbf{y}, s, \mathbf{v} \mid \mathbf{x}, \boldsymbol{\omega})$ the joint model for the full data response, missingness, and auxiliary covariates conditional on the covariates that are of primary interest, where $\boldsymbol{\omega}$ represents the parameter vector.

1.3. The Schizophrenia Clinical Trial

Our work is motivated by a multicenter, randomized, double-blind clinical trial on treatments for schizophrenia. The trial data were previously analyzed in Linero and Daniels (2015), which took a Bayesian nonparametric approach, but did not utilize information from the auxiliary covariates. For this clinical trial, the longitudinal outcomes are the positive and negative syndrome scale (PANSS) scores, which measure the severity of symptoms for patients with schizophrenia (Kay, Fiszbein, and Opfer 1987). The outcomes are collected at $J = 6$ time points corresponding to baseline, day 4 after baseline, and weeks 1, 2, 3, and 4 after baseline. The possible dropout patterns are $S_i = 2, 3, 4, 5, 6$. The covariate of primary interest is treatment, with $X_i = T, A$, or P corresponding to test drug, active control or placebo, respectively. In addition, we have access to $Q = 7$ auxiliary covariates including age, onset (of schizophrenia) age, height, weight, country, sex, and education level.

The dataset consists of $N = 204$ subjects, with 45 subjects for the active control arm, 78 subjects for the placebo arm, and 81 subjects for the test drug arm. Detailed individual trajectories and mean responses over time for the three treatment arms can be found in Appendix Figure A.1. The dropout rates are 33.3%, 20.0%, and 25.6% for the test drug, active control and placebo arms, respectively. Subjects drop out for a variety of reasons. Some reasons including adverse events (e.g., occurrence of side effects), pregnancy and protocol violation are thought to be random dropouts, while the other reasons such as disease progression, lack of efficacy, physician decision and withdraw by patient are thought to be informative dropouts. It is ideal to treat those reasons differently while making inference. The informative dropout rates are 29.6%, 15.6%, and 25.6% for the test drug, active control, and placebo arms, respectively. Detailed dropout rates for each dropout pattern can be found in Appendix Table A.1. The dataset has a few intermittent missing outcomes (1 for the test drug arm, 1 for the active control arm, and 2 for the placebo arm). We focus our study on monotone missingness and assume partial ignorability (Harel and Schafer 2009) for the few intermittent missing outcomes.

The goal of this study is to estimate the change from baseline treatment effect,

$$r_x = E[Y_{i6} - Y_{i1} \mid X_i = x].$$

In particular, the treatment effect improvements over placebo, that is, $r_T - r_P$ and $r_A - r_P$, are of interest.

1.4. Overview and Contribution

We stratify the model by treatment, and suppress the treatment variable x to simplify notation hereafter. The extrapolation factorization (Daniels and Hogan 2008) is

$$p(\mathbf{y}, s, \mathbf{v} \mid \omega) = p(\tilde{\mathbf{y}}_s \mid \bar{\mathbf{y}}_s, s, \mathbf{v}, \omega_E) p(\bar{\mathbf{y}}_s, s, \mathbf{v} \mid \omega_O),$$

where the extrapolation distribution, $p(\tilde{\mathbf{y}}_s \mid \bar{\mathbf{y}}_s, s, \mathbf{v}, \omega_E)$, is not identified by the data in the absence of uncheckable assumptions or constraints on the parameter space. The observed data distribution $p(\bar{\mathbf{y}}_s, s, \mathbf{v} \mid \omega_O)$ is identified and can be estimated semiparametrically or nonparametrically. We factorize the observed data distribution based on pattern-mixture modeling (Little 1993),

$$p(\bar{\mathbf{y}}_s, s, \mathbf{v} \mid \omega_O) = p(\bar{\mathbf{y}}_s \mid s, \mathbf{v}, \boldsymbol{\pi}) p(s \mid \mathbf{v}, \boldsymbol{\varphi}) p(\mathbf{v} \mid \boldsymbol{\eta}), \quad (1)$$

where we assume distinct parameters $\omega_O = (\boldsymbol{\pi}, \boldsymbol{\varphi}, \boldsymbol{\eta})$ parameterizing the response model, the missingness and the distribution of the auxiliary covariates, respectively.

The model specification (1) brings two challenges:

1. For the models $p(\bar{\mathbf{y}}_s \mid s, \mathbf{v}, \boldsymbol{\pi})$ and $p(s \mid \mathbf{v}, \boldsymbol{\varphi})$, it is unclear how the auxiliary covariates are related to the responses and dropout patterns. For example, the auxiliary covariates contain height and weight, which might not have a linear and additive effect on the responses. For example, the responses might have a linear relationship with the body mass index, which is calculated by weight/height².
2. For the model $p(\bar{\mathbf{y}}_s \mid s, \mathbf{v}, \boldsymbol{\pi})$, the observed patterns are sparse. For example, the dropout pattern $S_i = 2$ for the active control arm has only 1 observation.

To mitigate challenge 1, we specify semiparametric models for $p(\bar{\mathbf{y}}_s \mid s, \mathbf{v}, \boldsymbol{\pi})$ and $p(s \mid \mathbf{v}, \boldsymbol{\varphi})$ via Gaussian process (GP) priors and Bayesian additive regression trees (BART), respectively. Such models, although still making some parametric assumptions, are highly flexible and robust to model misspecification. To address challenge 2, we use informative priors such as autoregressive (AR) and conditional autoregressive (CAR) priors to share information across neighboring patterns. Detailed model specifications will be described in Section 2.

We would like to emphasize the distinction between our approach and the approaches proposed in Linero and Daniels (2015) and Linero (2017). The earlier works by Linero and Daniels and Linero specified the observed data distribution (1) based on a working model for the full data constructed as a Dirichlet process mixture of selection models, that is, $p^*(\mathbf{y}, s \mid \omega) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_1) p(s \mid \mathbf{y}, \boldsymbol{\theta}_2) F(d\boldsymbol{\theta})$ with F following a Dirichlet process; the approaches did not consider covariates but they could be added in a very simple way by introducing them independently (from \mathbf{y} and s and from each other in the Dirichlet process mixture) as mentioned in the discussion of Linero (2017). These approaches can thus accommodate (auxiliary) covariates but they were not constructed to include them in a careful, efficient way. In contrast, the proposed approach was designed specifically to allow (auxiliary) covariates and to estimate an average (mean) treatment effect. To address the former, we use a pattern-mixture model parameterization, and exploit the expected structure including sparse patterns and similar covariate effects across patterns and over time via GP, AR/CAR

priors, and shrinkage priors. Later we will show through simulation studies that our approach indeed performs better than the approaches proposed in Linero and Daniels (2015) and Linero (2017) with a simple extension that accommodates auxiliary covariates.

The remainder of this article is structured as follows. In Section 2, we specify Bayesian (semiparametric) models for (1). In Section 3, we use identifying restrictions to identify the extrapolation distribution. In Section 4, we describe our posterior inference and computation approaches. In Section 5, we present simulation studies to validate our model and compare with results using other methods. In Section 6, we apply our method to a clinical trial on treatments for schizophrenia. We conclude with a discussion in Section 7.

2. Probability Model for the Observed Data

2.1. Model for the Observed Data Responses Conditional on Pattern and Auxiliary Covariates

We define the model for observed data responses conditional on drop out time and auxiliary covariates, that is, $p(\bar{\mathbf{y}}_s \mid s, \mathbf{v}, \boldsymbol{\pi})$, as follows. The distribution $p(\bar{\mathbf{y}}_s \mid s, \mathbf{v}, \boldsymbol{\pi})$ can be factorized as

$$p_s(\bar{\mathbf{y}}_s \mid \mathbf{v}, \boldsymbol{\pi}) = p_s(y_s \mid \bar{\mathbf{y}}_{s-1}, \mathbf{v}, \boldsymbol{\pi}) \cdots p_s(y_2 \mid y_1, \mathbf{v}, \boldsymbol{\pi}) p_s(y_1 \mid \mathbf{v}, \boldsymbol{\pi}), \quad (2)$$

where the subscript s corresponds to conditioning on dropping out pattern $S = s$.

We assume

$$\begin{aligned} & \left(Y_j \mid \bar{\mathbf{y}}_{j-1} = \bar{\mathbf{y}}_{j-1}, s, \mathbf{v}, a_0, a, \boldsymbol{\pi} \right) \\ & = \begin{cases} a_0(\mathbf{v}, s) + \varepsilon_{1s}, & j = 1; \\ a(y_{j-1}, \mathbf{v}, j, s) + \bar{\mathbf{y}}_{j-2}^T \boldsymbol{\phi}_{js} + \varepsilon_{js}, & j \geq 2, \end{cases} \end{aligned} \quad (3)$$

where $j = 1, \dots, s$; $s = 2, \dots, J$. Here a_0 and a are stochastic processes indexed by $\mathcal{U}_0 = \mathcal{V} \times \mathcal{J}_0$ and $\mathcal{U} = \mathcal{Y} \times \mathcal{V} \times \mathcal{J}$, respectively, where \mathcal{V} is the state space of \mathbf{v} , $\mathcal{J}_0 = \{2, \dots, J\}$ is the state space of s , $\mathcal{J} \subset \{1, \dots, J\}^2$ is the state space of (j, s) , and \mathcal{Y} is the state space of y_{j-1} . Furthermore, $\boldsymbol{\phi}_{js}$ is the vector of lag coefficients (of order 2 and above) for each time/pattern, and ε_{js} 's are independent Gaussian errors,

$$\varepsilon_{js} \sim N(0, \sigma_{js}^2).$$

To have a flexible mean model for Y_j as a function of previous response and covariates, we place GP priors (Rasmussen and Williams 2006) on a_0 and a ,

$$\begin{aligned} a_0(\mathbf{v}, s) & \sim \mathcal{GP} [\mu_0(\mathbf{v}, s), C_0(\mathbf{v}, s; \mathbf{v}', s')]; \\ a(y_{j-1}, \mathbf{v}, j, s) & \sim \mathcal{GP} [\mu(y_{j-1}, \mathbf{v}, j, s), C(y_{j-1}, \mathbf{v}, j, s; \\ & y'_{j-1}, \mathbf{v}', j', s')], \end{aligned}$$

with mean functions $\mu_0: \mathcal{U}_0 \rightarrow \mathbb{R}$ and $\mu: \mathcal{U} \rightarrow \mathbb{R}$ and covariance functions $C_0: \mathcal{U}_0 \times \mathcal{U}_0 \rightarrow \mathbb{R}^+$ and $C: \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}^+$, respectively. Specifically,

$$\begin{aligned} \mu_0(\mathbf{v}, s) & = \mathbf{v}^T \boldsymbol{\beta}_{0s} + b_{1s}; \\ \mu(y_{j-1}, \mathbf{v}, j, s) & = \psi_{js} y_{j-1} + \mathbf{v}^T \boldsymbol{\beta}_s + b_{js}, \end{aligned} \quad (4)$$

and

$$\begin{aligned} C_0(\mathbf{v}, s; \mathbf{v}', s') &= \kappa_0^2 D_0(\mathbf{v}, s; \mathbf{v}', s') + \tilde{\kappa}_0^2 I(\mathbf{v}, s; \mathbf{v}', s'); \\ C(y_{j-1}, \mathbf{v}, j, s; y'_{j-1}, \mathbf{v}', j', s') \\ &= \kappa^2 D(y_{j-1}, \mathbf{v}, j, s; y'_{j-1}, \mathbf{v}', j', s') \\ &\quad + \tilde{\kappa}^2 I(y_{j-1}, \mathbf{v}, j, s; y'_{j-1}, \mathbf{v}', j', s'). \end{aligned} \quad (5)$$

We use two different stochastic processes a_0 and a for $j = 1$ and $j \geq 2$. The reason is that for $j = 1$, a_0 represent the mean initial response with no past; for $j \geq 2$, a represents the mean at subsequent responses, with a past. In the mean functions (4), β_{0s} and β_s are the vectors of regression coefficients of the auxiliary covariates, ψ_{js} is the lag-1 coefficient, and b_{js} is the time/pattern specific intercepts. In the covariance functions (5), $D_0(a; b)$ and $D(a; b)$ are the exponential distances between a and b , defined by

$$\begin{aligned} D_0(\mathbf{v}, s; \mathbf{v}', s') &= \exp \left[-\frac{\|\underline{\mathbf{v}} - \underline{\mathbf{v}'}\|_2^2}{2\gamma_{v0}^2} - \frac{|\underline{s} - \underline{s}'|}{\gamma_{s0}} \right], \\ D(y_{j-1}, \mathbf{v}, j, s; y'_{j-1}, \mathbf{v}', j', s') \\ &= \exp \left[-\frac{\|y_{j-1} - y'_{j-1}\|_2^2}{2\gamma_y^2} \right. \\ &\quad \left. - \frac{\|\underline{\mathbf{v}} - \underline{\mathbf{v}'}\|_2^2}{2\gamma_v^2} - \frac{|j - j'|}{\gamma_j} - \frac{|\underline{s} - \underline{s}'|}{\gamma_s} \right]. \end{aligned}$$

Here κ_0^2 , γ_{v0} , γ_{s0} , $\tilde{\kappa}_0^2$, κ^2 , γ_y , γ_v , γ_j , γ_s , $\tilde{\kappa}^2$ are the hyperparameters. Details about the hyper-priors or choices of these hyperparameters are described in Appendix A.2. The values $\underline{\mathbf{v}}$, y_{j-1} , j , and \underline{s} are standardized values for \mathbf{v} , y_{j-1} , j , and s (details in Appendix A.2). For categorical covariates, the distance between \mathbf{v} and \mathbf{v}' is calculated by counting the number of different values. In addition, in (5), $I(a; b)$ is the Kronecker delta function that takes the value 1 if $a = b$ and 0 otherwise. The function $I(a; b)$ is used to introduce a small nugget for the diagonal covariances, which overcomes near-singularity of the covariance matrices and improves numerical stability. The GPs flexibly model the relationship between auxiliary covariates and the previous response with the current response (Y_j) and accounts for possibly nonlinear and nonadditive effects in the auxiliary covariates and previous response.

For the noise variance σ_{js}^2 , we assume an inverse Gamma shrinkage prior,

$$\sigma_{js}^2 \mid \nu_\sigma \stackrel{\text{iid}}{\sim} \text{IG}(\lambda_\sigma, \lambda_\sigma \nu_\sigma), \quad j = 1, \dots, s, \quad s = 2, \dots, J,$$

with $E(1/\sigma_{js}^2) = 1/\nu_\sigma$ and $\text{var}(1/\sigma_{js}^2) = 1/\lambda_\sigma \nu_\sigma^2$. This prior shrinks the time/pattern specific variances to a common value, ν_σ . We put hyper-priors on λ_σ and ν_σ ,

$$\lambda_\sigma - 2 \sim \text{IG}(\lambda_1^{\lambda_\sigma}, \lambda_2^{\lambda_\sigma}), \quad \nu_\sigma \sim \text{Gamma}(\lambda_1^{\nu_\sigma}, \lambda_2^{\nu_\sigma}),$$

where we assume $\lambda_\sigma > 2$ to impose more shrinkage and borrowing of information.

Next, we consider the parameters in the mean functions (4). We allow the regression coefficients of the auxiliary covariates to vary by pattern. However, it is typical to have sparse patterns. As a result, we consider an informative prior that assumes

regression coefficients for neighboring patterns to be similar. In particular, we specify AR(1) type priors on β_{0s} and β_s . Let $\beta_0 = (\beta_{02}, \beta_{03}, \dots, \beta_{0J})$ and $\beta = (\beta_2, \beta_3, \dots, \beta_J)$ denote the coefficient vectors for the auxiliary covariates in Equation (4). We assume

$$\beta_0 \sim N \left[X_\beta \tilde{\beta}_0, \sigma_{\beta_0}^2 \Sigma_\beta(\rho_0) \right], \quad \beta \sim N \left[X_\beta \tilde{\beta}, \sigma_\beta^2 \Sigma_\beta(\rho) \right],$$

where $X_\beta^T = (I, I, \dots, I)$, $\tilde{\beta}_0$ and $\tilde{\beta}$ are the prior means for β_{0s} and β_s , respectively, and $\sigma_{\beta_0}^2 \Sigma_\beta(\rho_0)$ and $\sigma_\beta^2 \Sigma_\beta(\rho)$ are the AR(1) type covariance matrices. Details and hyper-priors on the hyper-parameters are described in Appendix A.2. The time/pattern specific intercepts are given CAR type priors (De Oliveira 2012; Banerjee, Carlin, and Gelfand 2014) as we expect them to be similar for neighboring patterns/times. Let $\mathbf{b}_0 = (b_{12}, b_{13}, \dots, b_{1J})$ and $\mathbf{b} = (b_{22}, b_{23}, b_{33}; \dots; b_{2J}, \dots, b_{JJ})$ denote the time/pattern-specific intercepts in Equation (4). We assume

$$\begin{aligned} \mathbf{b}_0 &\sim N \left(\mathbf{1} \tilde{b}_0, \sigma_{b_0}^2 (I - \gamma_{b_0} W_{b_0})^{-1} \mathcal{N}_{b_0} \right), \\ \mathbf{b} &\sim N \left(\mathbf{1} \tilde{b}, \sigma_b^2 (I - \gamma_b W_b)^{-1} \mathcal{N}_b \right), \end{aligned}$$

where \tilde{b}_0 and \tilde{b} are the prior means for \mathbf{b}_0 and \mathbf{b} , respectively, and $\sigma_{b_0}^2 (I - \gamma_{b_0} W_{b_0})^{-1} \mathcal{N}_{b_0}$ and $\sigma_b^2 (I - \gamma_b W_b)^{-1} \mathcal{N}_b$ are the CAR type covariance matrices. Details in Appendix A.2. The time/pattern specific lag-1 coefficients are given CAR type priors similar to the priors on b_{js} for the same reason. Let $\psi = (\psi_{22}; \psi_{23}, \psi_{33}; \dots; \psi_{2J}, \dots, \psi_{JJ})$ denote the time/pattern-specific coefficient vector for the lag-1 responses in Equation (4). We assume

$$\psi \sim N \left(\mathbf{1} \tilde{\psi}, \sigma_\psi^2 (I - \gamma_\psi W_\psi)^{-1} \mathcal{N}_\psi \right),$$

where $\tilde{\psi}$ is the prior mean for ψ , and $\sigma_\psi^2 (I - \gamma_\psi W_\psi)^{-1} \mathcal{N}_\psi$ is the CAR type covariance matrix. Again, more details in Appendix A.2. We complete the model with a prior for the higher-order (≥ 2) lag coefficients ϕ_{js} . Note that we assume the effect of higher-order lag responses on current response is linear. We do not include \bar{Y}_{j-2} in $a(\cdot)$ as the dimension of \bar{Y}_{j-2} varies for different time j . We expect to capture most of the nonlinear and nonadditive effects from lagged responses by including Y_{j-1} in $a(\cdot)$ since we expect most of the temporal effects come from the lag-1 response. We simply put normal priors with more prior mass around 0 to indicate the prior belief that higher-order lags have less impact on current response. Specifically,

$$\phi_{js} \sim N(\mathbf{0}, \sigma_\phi^2 I), \quad \sigma_\phi^2 \sim \text{IG}(\lambda_1^\phi, \lambda_2^\phi),$$

with $\lambda_1^\phi > \lambda_2^\phi$.

2.2. Model for the Pattern Conditional on Auxiliary Covariates

We model the hazard of dropout at time j with BART (Chipman, George, and McCulloch 2010),

$$p(S = j \mid S \geq j, \mathbf{v}, \boldsymbol{\varphi}) = F_N(f_j(\mathbf{v})),$$

where F_N denotes the standard normal cdf (probit link), and $f_j(\mathbf{v})$ is the sum of tree models from BART. The BART model captures complex relationships between auxiliary covariates and dropout including interactions and nonlinearities. We use the default priors for $f_j(\cdot)$ given in Chipman, George, and McCulloch (2010).

2.3. Model for the Auxiliary Covariates

We use a Bayesian bootstrap (Rubin 1981) prior for the distribution for \mathbf{v} . Suppose \mathbf{v} can only take the N discrete values that we observed, $\mathbf{V} \in \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$. The probability for each is

$$p(\mathbf{V} = \mathbf{v}_i | \boldsymbol{\eta}) = \eta_i, \quad (6)$$

where $\sum_{i=1}^N \eta_i = 1$. We place a Dirichlet distribution prior on $\boldsymbol{\eta}$,

$$(\eta_1, \dots, \eta_N) \sim \text{Dir}(\delta_{\eta_1}, \dots, \delta_{\eta_N}).$$

3. The Extrapolation Distribution

The extrapolation distribution for our setting can be sequentially factorized as

$$p_s(\tilde{\mathbf{y}}_s | \bar{\mathbf{y}}_s, \mathbf{v}, \boldsymbol{\omega}_E) = p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v}, \boldsymbol{\omega}_E) \cdot p_s(y_{s+2} | \bar{\mathbf{y}}_{s+1}, \mathbf{v}, \boldsymbol{\omega}_E) \cdots p_s(y_j | \bar{\mathbf{y}}_{j-1}, \mathbf{v}, \boldsymbol{\omega}_E). \quad (7)$$

The extrapolation distribution is not identified by the observed data. To identify the extrapolation distribution, we use identifying restrictions that express the extrapolation distribution as a function of the observed data distribution; see Linero and Daniels (2018) for a comprehensive discussion. For example, MAR (Rubin 1976) is a joint identifying restriction that completely identifies the extrapolation distribution. It is shown in Molenberghs et al. (1998) that MAR is equivalent to the available case missing value (ACMV) restriction in the pattern mixture model framework. The same statement is true when conditional on \mathbf{V} , in which case MAR is referred to as auxiliary variable MAR (A-MAR) (Daniels and Hogan 2008). ACMV sets

$$p_k(y_j | \bar{\mathbf{y}}_{j-1}, \mathbf{v}, \boldsymbol{\omega}_E) = p_{\geq j}(y_j | \bar{\mathbf{y}}_{j-1}, \mathbf{v}, \boldsymbol{\pi}),$$

for $k < j$ and $2 \leq j < J$, where the subscript $\geq j$ indicates conditioning on $S \geq j$. The latter involves averaging $p_s(\cdot)$ with respect to the missingness prior on s .

When the missingness is not at random, a partial identifying restriction (Linero and Daniels 2018) is the missing nonfuture dependence (NFD) assumption (Kenward, Molenberghs, and Thijs 2003). NFD states that the probability of dropout at time j depends only on $\bar{\mathbf{y}}_{j+1}$. Similarly, when conditional on \mathbf{V} , auxiliary variable NFD (A-NFD) assumes

$$p(S = j | \bar{\mathbf{y}}_j, \mathbf{v}, \boldsymbol{\omega}) = p(S = j | \bar{\mathbf{y}}_{j+1}, \mathbf{v}, \boldsymbol{\omega}).$$

Within the pattern-mixture framework, NFD is equivalent to the nonfuture missing value (NFMV) restriction (Kenward, Molenberghs, and Thijs 2003). Under A-NFD, we have

$$p_k(y_j | \bar{\mathbf{y}}_{j-1}, \mathbf{v}, \boldsymbol{\omega}_E) = p_{\geq j-1}(y_j | \bar{\mathbf{y}}_{j-1}, \mathbf{v}, \boldsymbol{\pi}), \quad (8)$$

for $k < j-1$ and $2 < j \leq J$. NFMV leaves one conditional distribution per incomplete pattern unidentified: $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v})$. To identify $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v})$, we assume a location shift τ_{s+1} (Daniels and Hogan 2000),

$$[Y_{s+1} | \bar{\mathbf{Y}}_s, S = s, \mathbf{V}, \boldsymbol{\omega}] \stackrel{d}{=} [Y_{s+1} + \tau_{s+1} | \bar{\mathbf{Y}}_s, S \geq s+1, \mathbf{V}, \boldsymbol{\omega}], \quad (9)$$

where $\stackrel{d}{=}$ denotes equality in distribution, and τ_{s+1} measures the deviation of the unidentified distribution $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v})$ from ACMV. In particular, ACMV holds when $\tau_{s+1} = 0$; τ_{s+1} is a *sensitivity parameter* (Daniels and Hogan 2008). To help calibrate the magnitude of τ_{s+1} , we set

$$[\tau_{s+1} | \bar{\mathbf{Y}}_s = \bar{\mathbf{y}}_s, \mathbf{V} = \mathbf{v}] = \tilde{\tau} \cdot \Delta_{s+1}(\bar{\mathbf{y}}_s, \mathbf{v}), \quad (10)$$

where $\Delta_{s+1}(\bar{\mathbf{y}}_s, \mathbf{v})$ is the standard deviation of $(Y_{s+1} | \bar{\mathbf{y}}_s, s, \mathbf{v})$ under ACMV, and $\tilde{\tau}$ represents the number of standard deviations that $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v})$ is deviated from ACMV. Similar strategies to calibrate sensitivity parameters based on the observed data can be found in Daniels and Hogan (2008) and Kim et al. (2017). Importantly, note that, based on this calibration, for a fixed $\tilde{\tau}$ we would have a smaller Δ using auxiliary covariates and thus a smaller deviation from ACMV, in comparison to unconditional on \mathbf{V} .

4. Posterior Inference and Computation

4.1. Posterior Sampling for Observed Data Model Parameters

We use a Markov chain Monte Carlo (MCMC) algorithm to draw samples from the posterior $\mathbf{w}_O^{(l)} \stackrel{\text{iid}}{\sim} p(\mathbf{w}_O | \{\bar{\mathbf{y}}_{is_l}, s_l, \mathbf{v}_l\}_{i=1}^N)$, $l = 1, \dots, L$. Note that we use distinct parameters $\boldsymbol{\pi}, \boldsymbol{\varphi}, \boldsymbol{\eta}$ for $p(\bar{\mathbf{y}}_s | s, \mathbf{v}, \boldsymbol{\pi})$, $p(s | \mathbf{v}, \boldsymbol{\varphi})$ and $p(\mathbf{v} | \boldsymbol{\eta})$, and the parameters are also a priori independent, $p(\boldsymbol{\pi}, \boldsymbol{\varphi}, \boldsymbol{\eta}) = p(\boldsymbol{\pi})p(\boldsymbol{\varphi})p(\boldsymbol{\eta})$. Therefore, the posterior distribution of \mathbf{w}_O can be factored as

$$p(\mathbf{w}_O | \{\bar{\mathbf{y}}_{is_l}, s_l, \mathbf{v}_l\}_{i=1}^N) = p(\boldsymbol{\pi} | \{\bar{\mathbf{y}}_{is_l}, s_l, \mathbf{v}_l\}_{i=1}^N) p(\boldsymbol{\varphi} | \{s_l, \mathbf{v}_l\}_{i=1}^N) p(\boldsymbol{\eta} | \{\mathbf{v}_l\}_{i=1}^N), \quad (11)$$

and posterior simulation can be conducted independently for $\boldsymbol{\pi}, \boldsymbol{\varphi}$ and $\boldsymbol{\eta}$. Gibbs transition probabilities are used to update $\boldsymbol{\pi}$ (details in Appendix A.2), the R packages `bartMachine` (Kapelner and Bleich 2016) and `BayesTree` (Chipman and McCulloch 2016) are used to update $\boldsymbol{\varphi}$, and $\boldsymbol{\eta}$ is updated by directly sampling from its posterior $\boldsymbol{\eta} | \{\mathbf{v}_l\}_{i=1}^N \sim \text{Dir}(1 + \delta_{\eta_1}, \dots, 1 + \delta_{\eta_N})$.

4.2. Computation of Expectation of Functionals of Full-Data Responses

Our interest lies in the expectation of functionals of \mathbf{y} , given by

$$\begin{aligned} E[t(\mathbf{y})] &= \int_{\mathbf{y}} t(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} \\ &= \int_{\mathbf{y}} t(\mathbf{y}) \left[\sum_s \int_{\mathbf{v}} p_s(\tilde{\mathbf{y}}_s | \bar{\mathbf{y}}_s, \mathbf{v}) p_s(\bar{\mathbf{y}}_s | \mathbf{v}) p(s | \mathbf{v}) p(\mathbf{v}) d\mathbf{v} \right] d\mathbf{y}. \end{aligned} \quad (12)$$

Once we have obtained posterior samples $\{\mathbf{w}_O^{(l)}, l = 1, \dots, L\}$, the expression (12) can be computed by Monte Carlo integration. Since the desired functionals are functionals of \mathbf{y} , computing (12) involves sampling pseudo-data based on the posterior samples. We note that this is an application of G-computation

(Robins 1986; Scharfstein et al. 2014; Linero and Daniels 2015) within the Bayesian paradigm (see Appendix Algorithm A.1).

In detail, at step 1, we draw $\mathbf{V}^* = \mathbf{v}_i$ with probability $p(\mathbf{V} = \mathbf{v}_i \mid \eta^{(l)}) = \eta_i^{(l)}$. At step 2, we draw S^* by sequentially sampling from $R \sim \text{Bernoulli}[p(S^* = j \mid S^* \geq j, \mathbf{v})]$. If $R = 1$, take $S^* = j$; otherwise proceed with $p(S^* = j+1 \mid S^* \geq j+1, \mathbf{v})$, $j = 2, \dots, J$. At step 3, we first draw $y_1^* \sim N(a_0(\mathbf{v}^*, s^*), \sigma_{1s^*}^2)$ and then sequentially draw $y_j^* \sim N(a(y_{j-1}^*, \mathbf{v}^*, j, s^*) + \bar{\mathbf{y}}_{j-2}^{*T} \boldsymbol{\phi}_{js^*}, \sigma_{js^*}^2)$, $j = 2, \dots, s^*$ as in (2), where $a_0(\mathbf{v}^*, s^*)$ and $a(y_{j-1}^*, \mathbf{v}^*, j, s^*)$ are generated by a GP prediction rule (Rasmussen and Williams 2006). At step 4, we sequentially draw y_j^* for $j = s^* + 1, \dots, J$ as in (7) from the unidentified distributions, now identified using identifying restrictions. When the ACMV restriction is specified, step 4 involves generating the random $p_{\geq j}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v})$, which is defined as

$$p_{\geq j}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v}) = \sum_{k=j}^J \alpha_{kj}(\bar{\mathbf{y}}_{j-1}, \mathbf{v}) p_k(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v}), \quad (13)$$

and

$$\begin{aligned} \alpha_{kj}(\bar{\mathbf{y}}_{j-1}, \mathbf{v}) &= p(S = k \mid \bar{\mathbf{y}}_{j-1}, S \geq j, \mathbf{v}) \\ &= \frac{p(\bar{\mathbf{y}}_{j-1} \mid S = k, \mathbf{v}) p(S = k \mid S \geq j, \mathbf{v})}{\sum_{k=j}^J p(\bar{\mathbf{y}}_{j-1} \mid S = k, \mathbf{v}) p(S = k \mid S \geq j, \mathbf{v})}. \end{aligned}$$

The distribution in (13) is a mixture distribution over patterns. We sample from (13) by first drawing $K = k$ with probability α_{kj} , $k = j, \dots, J$, then drawing a sample from $p_k(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v})$. When the NFMV restriction is specified, step 4 also involves generating the random $p_{\geq j-1}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v})$, where

$$\begin{aligned} p_{\geq j-1}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v}) &= \alpha_{j-1, j-1}(\bar{\mathbf{y}}_{j-1}, \mathbf{v}) p_{j-1}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v}) \\ &\quad + [1 - \alpha_{j-1, j-1}(\bar{\mathbf{y}}_{j-1}, \mathbf{v})] p_{\geq j}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v}). \end{aligned}$$

Sampling from $p_{\geq j-1}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v})$ is done by first sampling $Y_j^* \sim p_{\geq j}(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v})$ as in (13). Then draw $R \sim \text{Bernoulli}[\alpha_{j-1, j-1}]$. If $R = 1$, apply the location shift (9), otherwise, retain Y_j^* . See Appendix A.4 for more details of steps 3 and 4.

5. Simulation Studies

We conduct several simulation studies similar to the data example to assess the operating characteristic of our proposed model (denoted as GP hereafter). We simulate responses for $J = 6$ time points and fit our model to estimate the change from baseline treatment effect, that is, $E[Y_j - Y_1]$. We set the prior and hyper-prior parameters at standard noninformative choices. See Appendix A.5 for exact values. For comparison, we consider four alternatives:

(1, LM) a linear pattern-mixture model that consists of a linear regression model for $p_s(y_j \mid \bar{\mathbf{y}}_{j-1}, \mathbf{v})$, a sequential logit model for $p(s \mid \mathbf{v})$, and a Bayesian bootstrap model for $p(\mathbf{v})$, as in Equations (16), (15), and (6), respectively;

(2, LM-) a linear pattern-mixture model without \mathbf{V} that consists of a linear regression model for $p_s(y_j \mid \bar{\mathbf{y}}_{j-1})$ and a Bayesian bootstrap model for $p(s)$;

(3, DPM) a working model for the full data, constructed as a Dirichlet process mixture of selection models, $p^*(\mathbf{y}, s, \mathbf{v} \mid \boldsymbol{\omega}) = \int p(\mathbf{y} \mid \mathbf{v}, \boldsymbol{\theta}_1) p(s \mid \mathbf{y}, \mathbf{v}, \boldsymbol{\theta}_2) p(\mathbf{v} \mid \boldsymbol{\theta}_3) F(d\boldsymbol{\theta})$ with F following a Dirichlet process. As suggested in Linero (2017), we use a linear regression model for $p(\mathbf{y} \mid \mathbf{v}, \boldsymbol{\theta}_1)$ and a sequential logit model for $p(s \mid \mathbf{y}, \mathbf{v}, \boldsymbol{\theta}_2)$. For $p(\mathbf{v} \mid \boldsymbol{\theta}_3)$, similar to Shahbaba and Neal (2009), we assume independent normal distributions for continuous \mathbf{V} 's, Bernoulli distributions for binary \mathbf{V} 's and multinomial distributions for categorical \mathbf{V} 's; and

(4, DPM-) a working model for the full data without \mathbf{V} , $p^*(\mathbf{y}, s \mid \boldsymbol{\omega}) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_1) p(s \mid \mathbf{y}, \boldsymbol{\theta}_2) F(d\boldsymbol{\theta})$, which was proposed by Linero and Daniels (2015) and Linero (2017).

We use noninformative priors for the two parametric models (1) and (2), and use the default prior choices in Linero and Daniels (2015) and Linero (2017) for the Dirichlet process mixture models (3) and (4). For each simulation scenario below, we generate 500 datasets with $N = 200$ subjects per dataset. See Appendix Section A.7 for further details on computing times.

5.1. Performance Under MAR

We first evaluate the performance of our model under the ACMV restriction (MAR). Since this restriction completely identifies the extrapolation distribution, this simulation study validates the appropriateness of our observed data model specification. We consider the following three simulation scenarios.

Scenario 1. We test the performance of our approach when the data are generated from a simple linear pattern-mixture model to assess loss of efficiency from using an unnecessary complex modeling approach. For each subject, we first simulate $Q = 4$ auxiliary covariates from a multivariate normal distribution

$$\mathbf{V} \stackrel{\text{iid}}{\sim} N(\mathbf{0}, \Sigma_{\mathbf{V}}). \quad (14)$$

We then generate dropout time using a sequential logit model

$$\text{logit } P(S = s \mid S \geq s, \mathbf{V}) = \zeta_s + \mathbf{V}^T \boldsymbol{\xi}_s. \quad (15)$$

Next, we generate $\bar{\mathbf{Y}}_s$ from

$$\begin{aligned} (Y_j \mid \bar{\mathbf{Y}}_{j-1}, S = s, \mathbf{V}) &\sim N\left(\mu_{js}(\bar{\mathbf{Y}}_{j-1}, \mathbf{V}), \sigma_{js}^2\right), \quad \text{for } j = 1, \dots, s \\ \text{where } \mu_{js}(\bar{\mathbf{Y}}_{j-1}, \mathbf{V}) &= \begin{cases} \mathbf{V}^T \boldsymbol{\beta}_{0s} + b_{1s} & \text{if } j = 1 \\ Y_{j-1} \psi_{js} + \mathbf{V}^T \boldsymbol{\beta}_s + b_{js} + \bar{\mathbf{Y}}_{j-2}^T \boldsymbol{\phi}_{js} & \text{if } j \geq 2 \end{cases} \end{aligned} \quad (16)$$

Finally, the distribution of $\bar{\mathbf{Y}}_s$ is specified under the ACMV restriction (for calculating the simulation truth of the mean estimate).

The parameters in (14)–(16) are chosen by fitting the model to the test drug arm of the schizophrenia clinical trial (after standardizing the responses and the auxiliary covariates with mean 0 and standard deviation 1). See Appendix A.5 for details.

Table 1. Summary of simulation results under MAR.

Model	Bias	CI width	CI coverage	MSE
Scenario 1				
GP	-0.013(0.004)	0.294(0.002)	0.909(0.012)	0.014(0.000)
LM	-0.005(0.004)	0.379(0.001)	0.969(0.007)	0.017(0.000)
LM-	0.004(0.004)	0.385(0.002)	0.969(0.007)	0.018(0.001)
DPM	-0.013(0.004)	0.355(0.002)	0.954(0.009)	0.018(0.001)
DPM-	-0.009(0.004)	0.343(0.001)	0.947(0.009)	0.016(0.001)
Scenario 2				
GP	0.037(0.010)	0.967(0.005)	0.943(0.010)	0.122(0.004)
LM	0.247(0.010)	1.021(0.004)	0.819(0.017)	0.189(0.006)
LM-	0.330(0.010)	1.094(0.005)	0.783(0.018)	0.243(0.007)
DPM	0.183(0.011)	1.188(0.006)	0.924(0.012)	0.192(0.005)
DPM-	0.302(0.011)	1.054(0.006)	0.781(0.019)	0.228(0.008)
Scenario 3				
GP	-0.005(0.007)	0.666(0.002)	0.958(0.009)	0.057(0.002)
LM	0.008(0.007)	0.705(0.002)	0.968(0.008)	0.061(0.002)
LM-	0.026(0.007)	0.707(0.002)	0.964(0.008)	0.061(0.002)
DPM	-0.008(0.008)	0.778(0.002)	0.984(0.006)	0.070(0.002)
DPM-	-0.001(0.007)	0.669(0.002)	0.953(0.010)	0.058(0.002)

NOTES: Values shown are averages over repeat sampling, with numerical Monte Carlo standard errors in parentheses. GP, LM, LM-, DPM, DPM- represent the proposed semiparametric model, the linear regression model with covariates, the linear regression model without covariates, the Dirichlet process mixture model with covariates, and the Dirichlet process mixture model without covariates, respectively. CI width and coverage are based on 95% credible intervals.

Scenario 2. We consider a scenario where the covariates and the responses have more complicated structures to test the performance of our model when linearity does not hold. For simplicity, for each subject, we simulate $Q = 3$ auxiliary covariates from $\mathbf{V} \stackrel{iid}{\sim} N(\mathbf{0}, \Sigma_{vv})$. The responses and drop out times are generated in the same way as in scenario 1, but we include interactions and nonlinearities by replacing \mathbf{V} in Equations (15) and (16) (case $j = 1$) with $\dot{\mathbf{V}} = (V_1, V_2, V_3, V_1 \times V_2, V_1 \times V_3, V_2 \times V_3, V_1^2, V_2^2, V_3^2)$ and replacing \mathbf{V} in Equation (16) (case $j \geq 2$) with $\dot{\mathbf{V}} = (V_1, V_2, V_3, V_1 \times V_2, V_1 \times V_3, V_2 \times V_3, V_1^2, V_2^2, V_3^2, V_1 \times Y_{j-1}, V_2 \times Y_{j-1}, V_3 \times Y_{j-1}, \sqrt{|Y_{j-1}|})$. The regression coefficients ξ_s , β_{0s} , and β_s change accordingly. See Appendix A.5 for further details.

Scenario 3. We consider a scenario with a very different structure from our model formulation. In particular, we consider a lag-1 selection model with a mixture model for the joint distribution of \mathbf{Y} and \mathbf{V} . We generate

$$\begin{aligned}
 K &\sim \text{Categorical}(\boldsymbol{\pi}), \\
 \Omega^{(K)} &\sim \mathcal{W}^{-1} \left((v - J - Q - 1)\Omega_0^{(K)}, \nu \right), \\
 \begin{pmatrix} \mathbf{Y} \\ \mathbf{V} \end{pmatrix} | K &\sim N \left[\boldsymbol{\mu}^{(K)}, \Omega^{(K)} \right], \\
 \text{logit } P(S = s | S \geq s, \mathbf{Y}, \mathbf{V}) &= \zeta_s + \psi_s Y_s + \mathbf{V}^T \boldsymbol{\xi}_s,
 \end{aligned} \tag{17}$$

where $\mathcal{W}^{-1}((v - J - Q - 1)\Omega_0, \nu)$ is an inverse-Wishart distribution with precision parameter ν and mean Ω_0 . See Linero and Daniels (2015) for further details on this type of model. Formulating a joint distribution as in (17) allows us to impose complicated relationships between \mathbf{Y} and \mathbf{V} (Müller, Erkanli, and West 1996). We consider $Q = 3$ auxiliary covariates and 5 mixture components. We assume $\boldsymbol{\mu}^{(K)}$ and $\Omega_0^{(K)}$ correspond to a linear model of $(\mathbf{Y} | \mathbf{V})$ and have the form

$$\boldsymbol{\mu}^{(K)} = \begin{pmatrix} \boldsymbol{\mu}_y^{(K)} \\ \mathbf{0} \end{pmatrix}, \quad \Omega_0^{(K)} = \begin{pmatrix} \Sigma_{yy}^{(K)} & \Sigma_{yv}^{(K)} \\ \Sigma_{vy}^{(K)} & \Sigma_{vv} \end{pmatrix}.$$

In particular, we generate $\boldsymbol{\mu}^{(K)}$ and $\Omega_0^{(K)}$ according to Linero and Daniels (2015) by fitting the mixture model to the active control arm of the schizophrenia clinical trial. See Appendix A.5 for further details.

The simulation results are summarized in Table 1. For scenario 1, the true data generating model is the linear regression model with \mathbf{V} , that is, the LM model. The five models have similar performance in terms of MSE. The 95% credible interval of the GP model has a frequentist coverage rate less than 95% due to the prior information, that is, the GP priors and the AR/CAR priors, being quite strong and the sample size ($N = 200$) being relatively small. Therefore, the Bayesian credible interval is unlikely to have the expected frequentist coverage. The LM- and DPM- models (which ignore \mathbf{V}) do not perform worse than the LM and DPM models. The reason is probably that the (linear) effects of different \mathbf{V} 's on $t(\mathbf{Y})$ cancel out in the integration (12). For scenario 2, the true data generating model does not match any of the five models used for inference. The GP model significantly outperforms the other models in all aspects. The result suggests that when the model is misspecified, the GP model has much more robust performance. We note that the DPM and DPM- models, although being nonparametric, perform worse than the GP model. The reason is that the GP model is designed specifically to incorporate auxiliary covariates. It better exploits the structure of the data, which allows it to more readily capture nonlinear and nonadditive effects and handle sparse patterns, in particular with small sample sizes. We also note that when \mathbf{Y} and S do not have a linear relationship with \mathbf{V} , ignoring \mathbf{V} results in more significant bias than including \mathbf{V} (even mistakenly). For scenario 3, the true data generating model is a mixture of linear regression models, similar to the specification of the DPM model. The five models again have similar performance. For a pattern-mixture model, the marginal distribution of the responses \mathbf{Y} is a mixture distribution over patterns, which explains the good performance of the GP, LM, and LM- models. For all the three scenarios, the GP model always gives narrower credible intervals and has lower bias, in particular versus the models without auxiliary covariates.

Table 2. Summary of simulation results for Scenario 2 under MNAR.

Model	$E(\tilde{\tau})$	Bias	CI width	CI coverage	MSE
GP	-0.25	-0.024(0.011)	1.032(0.005)	0.957(0.009)	0.133(0.004)
	0	0.065(0.011)	1.050(0.005)	0.949(0.010)	0.140(0.004)
	0.25	0.157(0.011)	1.069(0.005)	0.909(0.012)	0.165(0.005)
	0.5	0.256(0.011)	1.091(0.005)	0.851(0.015)	0.210(0.007)
	-0.25	0.156(0.011)	1.122(0.005)	0.918(0.012)	0.170(0.005)
LM	0	0.271(0.011)	1.146(0.005)	0.866(0.015)	0.223(0.007)
	0.25	0.389(0.011)	1.170(0.005)	0.755(0.019)	0.307(0.009)
	0.5	0.513(0.011)	1.199(0.005)	0.626(0.021)	0.424(0.012)
	-0.25	0.222(0.010)	1.215(0.005)	0.909(0.012)	0.204(0.006)
	0	0.352(0.010)	1.237(0.005)	0.842(0.016)	0.284(0.008)
LM-	0.25	0.487(0.010)	1.266(0.006)	0.710(0.020)	0.403(0.011)
	0.5	0.626(0.011)	1.300(0.005)	0.528(0.022)	0.567(0.014)
	-0.25	0.077(0.011)	1.275(0.006)	0.979(0.006)	0.178(0.004)
	0	0.185(0.011)	1.289(0.006)	0.954(0.009)	0.210(0.005)
	0.25	0.298(0.011)	1.308(0.007)	0.884(0.014)	0.269(0.008)
DPM	0.5	0.415(0.012)	1.332(0.007)	0.795(0.018)	0.358(0.010)
	-0.25	0.179(0.011)	1.167(0.006)	0.932(0.011)	0.184(0.005)
	0	0.304(0.011)	1.197(0.006)	0.851(0.016)	0.252(0.008)
	0.25	0.435(0.011)	1.234(0.006)	0.712(0.020)	0.357(0.011)
	0.5	0.571(0.012)	1.278(0.006)	0.579(0.022)	0.504(0.014)

NOTES: Values shown are averages over repeat sampling, with numerical Monte Carlo standard errors in parentheses. CI width and coverage are based on 95% credible intervals. The values of $E(\tilde{\tau})$, -0.25, 0, 0.25, and 0.5, correspond to prior specifications $\text{Unif}(-0.75, 0.25)$, $\text{Unif}(-0.5, 0.5)$, $\text{Unif}(-0.25, 0.75)$ and $\text{Unif}(0, 1)$, respectively.

In summary, the semiparametric approach (GP) loses little when a simple parametric alternative holds, and it significantly outperforms the other approaches when the model used for inference is misspecified. The simulation results suggest that the semiparametric approach accommodates complex mean models and is more favorable compared with the parametric approaches and even simple nonparametric alternatives.

5.2. Performance Under MNAR

To assess the sensitivity of our model to untestable assumptions for the extrapolation distribution, we fit our model to simulated data under an NFD restriction (8). We consider simulation scenarios 2 and 3 as in Section 5.1, where the simulation truth is still generated under MAR. We complete our model with a location shift (Equations (9) and (10)). Recall that the sensitivity parameter $\tilde{\tau}$ measures the deviation of our model from MAR, and the simulation truth corresponds to $\tilde{\tau} = 0$. The sensitivity parameter $\tilde{\tau}$ is given four different priors: $\text{Unif}(-0.75, 0.25)$, $\text{Unif}(-0.5, 0.5)$, $\text{Unif}(-0.25, 0.75)$, $\text{Unif}(0, 1)$. All the four priors contain the simulation truth. Compared to fixing the value of $\tilde{\tau}$, using a uniform prior conveys uncertainty about the identifying restriction. For example, using a point mass prior $\tilde{\tau} = 0$ implies MAR with no uncertainty, while using a prior such that $E[\tilde{\tau}] = 0$ and $\text{var}[\tilde{\tau}] > 0$ implies MAR with uncertainty.

The simulation results for scenarios 2 and 3 are summarized in Table 2 and Appendix Table A.3, respectively. When the sensitivity parameter $\tilde{\tau}$ is centered at the correct value 0, the GP model significantly outperforms the alternatives under scenario 2 and performs as well as the alternatives under scenario 3. Comparing with the simulation results under MAR (Table 1), the use of a uniform prior for $\tilde{\tau}$ induces more uncertainty on inference resulting in the wider credible intervals. We also note that, when $\tilde{\tau}$ is not centered at 0, the models using V still perform better than the model not using V . This is due to the calibration of the location shift (Equations (9) and (10)). For the same $\tilde{\tau}$ we would have a smaller deviation from ACMV using V

compared to not using V . This property makes the missingness “closer” to MAR and reduces the extent of sensitivity analysis with the inclusion of V .

6. Application to the Schizophrenia Clinical Trial

We implement inference under the proposed model for data from the schizophrenia clinical trial described in Section 1.3. The dataset was first used in Linero and Daniels (2015). Recall the quantity of interest is the change from baseline treatment effect, $r_x = E[Y_{i6} - Y_{i1} \mid X_i = x]$, where $x = T, A$, or P correspond to treatments under test drug, active control, or placebo, respectively. We are particularly interested in the treatment effect improvements over placebo, that is, $r_T - r_P$ and $r_A - r_P$. Also, recall that we have $Q = 7$ auxiliary covariates including age, onset (of schizophrenia) age, height, weight, country, sex and education level. Details of computing specifications and times, as well as convergence diagnostics, are summarized in Appendix Section A.7.

6.1. Comparison to Alternatives and Assessment of Model Fit

We first compare the fit among the proposed model and alternatives. We consider the linear pattern-mixture models and Dirichlet process mixture of selection models with and without auxiliary covariates, as we have used in the simulation studies. We use the log-pseudo marginal likelihood (LPML) as the model selection criteria, where $\text{LPML} = \left[\sum_{i=1}^N \log(\text{CPO}_i) \right] / N$, CPO_i is the conditional predictive ordinate (Geisser and Eddy 1979) for observation i and $\text{CPO}_i = p\left(\tilde{Y}_{iS_i}, S_i, V_i \mid \{\tilde{Y}_{i' S_{i'}}, S_{i'}, V_{i'}\}_{i'=1, i' \neq i}^N\right)$. LPML can be straightforwardly estimated using posterior samples $\{\omega_O^{(l)}, l = 1, \dots, L\}$ (Gelfand and Dey 1994), without the need to refit the model N times. A model with higher LPML is more favorable compared to models with lower LPMLs. We fit the five models to

Table 3. Comparison of LPML (the second column) and inference results (the third and fourth columns) under MAR (the first five rows) and CCA (the last row).

Model	LPML	$r_T - r_P$	$r_A - r_P$
GP	-31.93	0.60(-5.07, 7.01)	-6.08(-13.90, 1.72)
LM	-32.61	-1.26(-8.59, 5.74)	-7.24(-15.00, 0.05)
LM-	-32.71	-1.94(-9.00, 5.07)	-8.13(-15.30, -1.00)
DPM	-39.25	0.44(-10.14, 10.42)	-7.66(-25.27, 9.67)
DPM-	-32.58	-1.69(-8.03, 4.78)	-5.44(-12.61, 2.27)
CCA	-	-3.23(-8.63, 2.18)	-3.82(-10.18, 2.55)

NOTES: For the inference results under MAR and CCA, values shown are posterior means, with 95% credible intervals in parentheses.

the data and calculate the LPML by taking the summation of the LPML under each treatment arm. The results are summarized in Table 3. The proposed semiparametric model (GP) has the largest LPML over the alternatives. In particular, the LPML improvement over the linear pattern-mixture model without covariates (LM-) for the GP model is much higher than the LM and DPM- models. This is not surprising in light of the earlier simulation results. We also compare inferences on treatment effect improvements over placebo under the MAR assumption using the five models, as well as a complete case analysis (CCA) based on the empirical distribution of the subjects who have complete outcomes. The results are summarized in Table 3. We point out the two plus points shifts between the GP model and the other models for the test drug versus placebo comparison and for the active drug versus placebo. The DPM model has the lowest LPML and the widest credible intervals. The poorer performance of the DPM model is probably due to the small sample size of each treatment arm (e.g., 45 subjects for the active control arm) and the relatively large number of covariates ($Q = 7$). Inference under the DPM model has large variability with small sample sizes, and the covariates can dominate the partition structure (Wade 2013). Further interpretation of the results of the GP model can be found in Section 6.2. The CCA (which implicitly assumes missing completely at random) is inefficient and is generally very unrealistic for longitudinal data.

Next, we assess the “absolute” goodness of fit of the proposed model. We estimate the cumulative dropout rates and observed-data means at each time point and under each treatment using the proposed model by

$$p(S \leq j | x) = \int p(S \leq j | \mathbf{v}, x)p(\mathbf{v} | x)d\mathbf{v}, \quad \text{and}$$

$$E(Y_j | S \geq j, x) = \int E(Y_j | S \geq j, \mathbf{v}, x)p(\mathbf{v} | S \geq j, x)d\mathbf{v}.$$

We then compare those estimates with results obtained from the empirical distribution of the observed data (that implicitly averages over the empirical distribution of the auxiliary covariates). Despite some small differences, there is no evidence for lack of fit. The comparison is shown in Figure 1.

6.2. Inference

A large portion of subjects dropout for reasons that suggest the missing data are MNAR (see Section 1.3). To identify the extrapolation distribution, we make the NFD assumption (8). Recall that the NFD assumption leaves one conditional distribution per incomplete pattern unidentified: $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v}, x)$. To

better identify $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v}, x)$, rather than simply assuming a location shift (9), we make use of information regarding the type of dropout. Let $Z_i = 1$ or 0 denote subject i drops out for informative or noninformative reasons, respectively. We model Z conditional on observed data responses, pattern, auxiliary covariates, and treatment with BART,

$$P(Z = 1 | \bar{\mathbf{Y}}_s, S = s, \mathbf{V}, X = x) = F_N(f_{sx}(\bar{\mathbf{Y}}_s, \mathbf{V})).$$

Recall F_N is the standard normal cdf, and $f_{sx}(\bar{\mathbf{Y}}_s, \mathbf{V})$ is the sum of tree models from BART.

The indicator Z is used to help identify $p_s(y_{s+1} | \bar{\mathbf{y}}_s, \mathbf{v}, x)$. We assume

$$\begin{aligned} [Y_{s+1} | \bar{\mathbf{Y}}_s, S = s, \mathbf{V}, X, \boldsymbol{\omega}] &\stackrel{d}{=} P(Z = 1 | \bar{\mathbf{Y}}_s, S = s, \mathbf{V}, X) \\ &\cdot [Y_{s+1} + \tau_{s+1} | \bar{\mathbf{Y}}_s, S \geq s + 1, \mathbf{V}, X, \boldsymbol{\omega}] \\ &+ P(Z = 0 | \bar{\mathbf{Y}}_s, S = s, \mathbf{V}, X) \cdot [Y_{s+1} | \bar{\mathbf{Y}}_s, S \geq s + 1, \mathbf{V}, X, \boldsymbol{\omega}], \end{aligned} \quad (18)$$

which is a mixture of an ACMV assumption and a location shift. We refer to Equation (18) as a MAR/MNAR mixture assumption. The idea is that, if a subject drops out for a reason associated with MAR, we impute the next missing value under ACMV; otherwise, we impute the next missing value by applying a location shift. The sensitivity parameter τ_{s+1} is interpretable to subject-matter experts, thus prior on τ_{s+1} can be created. Suppose two hypothetical subjects A and B have the same auxiliary covariates and histories up to time s , and suppose subject B drops out for an informative reason at time s while subject A remains on study. Then, the response of subject B at time $(s + 1)$ is stochastically identical to the response of subject A at time $(s + 1)$ after applying the location shift τ_{s+1} . As the prior for τ_{s+1} , we assume $\tau_{s+1} \geq 0$ as we expect subject B would have a higher PANSS score at time $(s + 1)$ than subject A. The magnitude of τ_{s+1} is calibrated as in Equation (10),

$$[\tau_{s+1} | \bar{\mathbf{Y}}_s = \bar{\mathbf{y}}_s, \mathbf{V} = \mathbf{v}, X = x] = \tilde{\tau}_x \cdot \Delta_{s+1,x}(\bar{\mathbf{y}}_s, \mathbf{v}). \quad (19)$$

We assume a uniform prior on $\tilde{\tau}_x$, $\tilde{\tau}_x \sim \text{Unif}(0, 1)$, as it is thought unlikely that the deviation from ACMV would exceed a standard deviation (Linero and Daniels 2015).

Figure 2 summarizes change from baseline treatment effect improvements of the test drug and active drug over placebo. We implement inference under both the MAR and the mixture of MAR/MNAR (Equations (18) and (19)) assumptions. For the test drug arm, the treatment effect improvement $r_T - r_P$ has posterior mean 0.60 and 95% credible interval $(-5.07, 7.01)$ under MAR, and posterior mean 0.91 and 95% credible interval $(-5.29, 7.81)$ under MAR/MNAR mixture. There is no evidence that the test drug has better performance than placebo. The MAR/MNAR mixture assumption slightly increases the posterior mean of $r_T - r_P$ as the test drug arm has a slightly higher informative dropout rate than the placebo arm (Appendix Table A.1). For the active drug arm, the treatment effect improvement $r_A - r_P$ has posterior mean -6.08 and 95% credible interval $(-13.90, 1.72)$ under MAR, and posterior mean -6.45 and 95% credible interval $(-14.34, 1.75)$ under MAR/MNAR mixture. There appears to be some evidence that the active drug has better effect than placebo. The MAR/MNAR mixture assumption

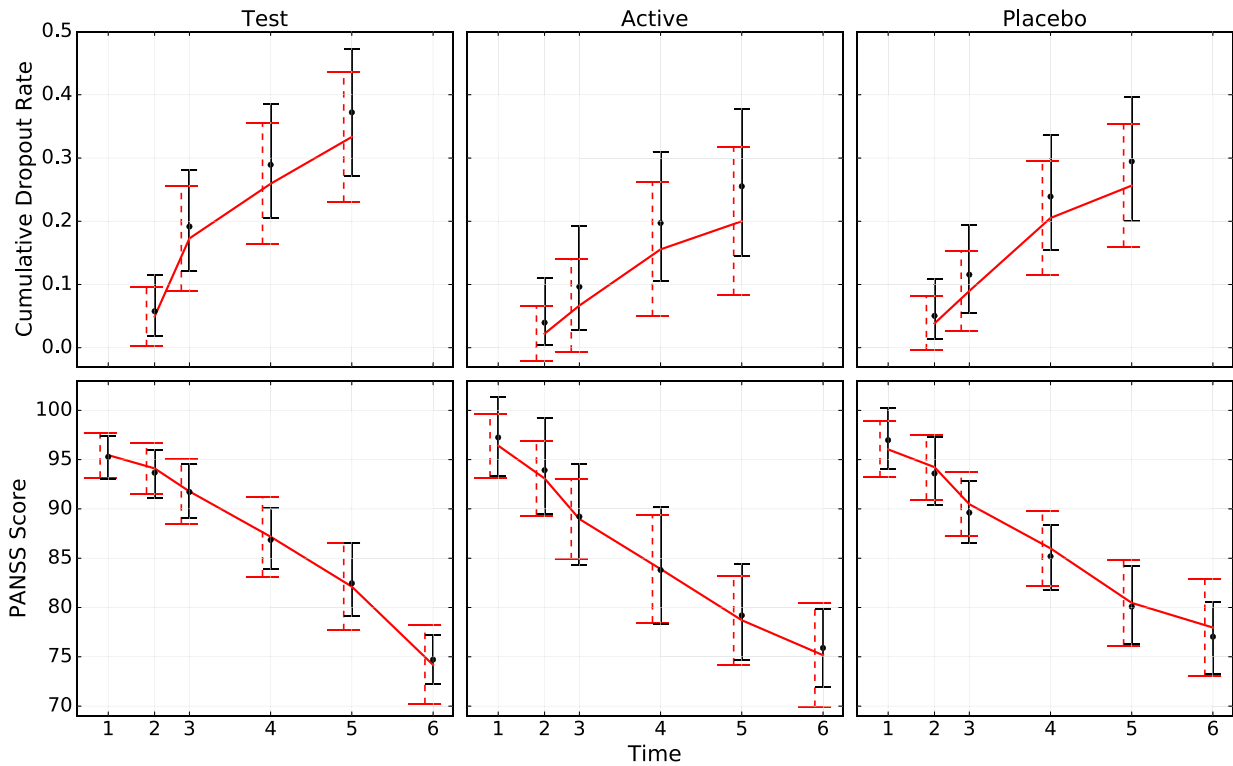


Figure 1. Cumulative dropout rates (top) and means of the observed data (bottom) over time obtained from the model versus the ones obtained from the empirical distribution of the observed data. The solid line represents the empirical values, dots represent the posterior means, dashed error bars represent frequentist 95% confidence intervals, and solid error bars represent the model's 95% credible intervals.

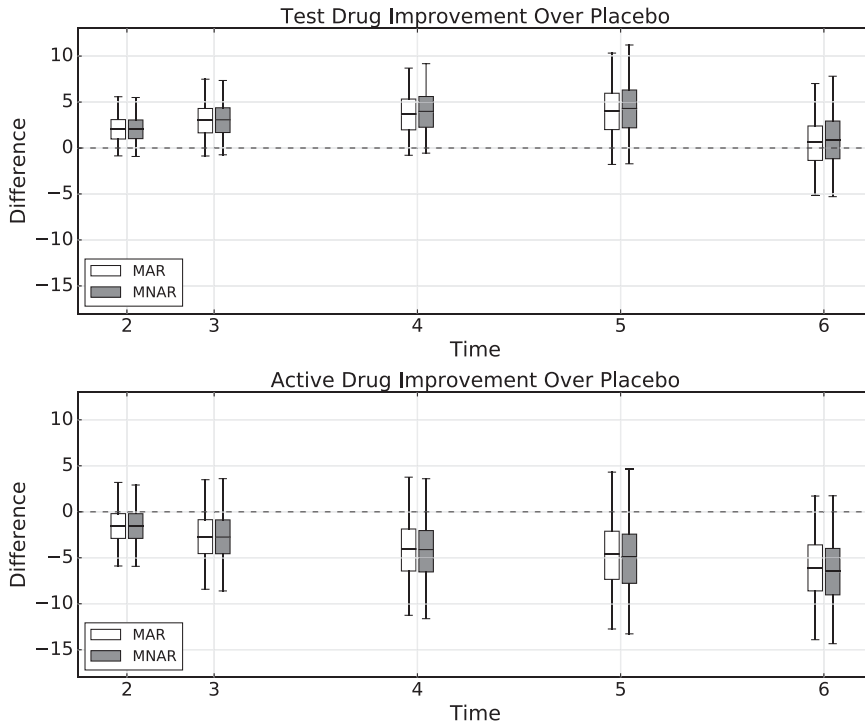


Figure 2. Change from baseline treatment effect improvements of the test drug (top) and active drug (bottom) over placebo over time. Smaller values indicate more improvement compared to placebo. The dividing line within the boxes represents the posterior mean, the bottom and top of the boxes are the first and third quartiles, and the ends of the whiskers show the 0.025 and 0.975 quantiles.

slightly decreases the posterior mean of $r_A - r_P$ as the active drug arm has a slightly lower informative dropout rate than the placebo arm (Appendix Table A.1). Also, in both scenarios, the MAR/MNAR mixture assumption induces more uncertainty on

the inferences (wider credible intervals), as we have discussed in Section 5.2.

The same dataset was previously analyzed in Linero and Daniels (2015), which concluded that there is little evidence that

the test drug is superior to the placebo and some evidence of an effect of the active control. Our analysis is consistent with the previous analyses. See Appendix Table A.4 for a detailed comparison.

6.3. Sensitivity Analysis

To assess the sensitivity of inferences on treatment effect improvements ($r_T - r_P$ and $r_A - r_P$) to the informative priors on the sensitivity parameters ($\tilde{\tau}_T$, $\tilde{\tau}_A$ and $\tilde{\tau}_P$), we consider a set of point-mass priors for each $\tilde{\tau}_x$ along the $[0, 1]$ grid. The detailed figure showing how inferences on $r_T - r_P$ and $r_A - r_P$ change for different choices of $\tilde{\tau}_T$, $\tilde{\tau}_A$, and $\tilde{\tau}_P$ is in Appendix Figure A.2. The sensitivity analysis corroborates our conclusion that there is no evidence that the test drug has better performance than placebo. For all the choices of $\tilde{\tau}_T$ and $\tilde{\tau}_P$, the posterior probability of $r_T - r_P < 0$ does not reach the 0.95 posterior probability cutoff. On the other hand, the sensitivity analysis shows that there is some evidence that the active drug is superior than placebo. For all the combinations of $\tilde{\tau}_A$ and $\tilde{\tau}_P$, the posterior probability of $r_A - r_P < 0$ is greater than 0.79. For most favorable values of $\tilde{\tau}_A$ and $\tilde{\tau}_P$, the posterior probability of $r_A - r_P < 0$ is greater than 0.95, although it only occurs when $\tilde{\tau}_A$ is substantially smaller than $\tilde{\tau}_P$. In summary, for all the choices of $\tilde{\tau}_x$, we do not reach substantially different results, which improves our confidence on the previous conclusions.

7. Discussion

In this work, we have developed a semiparametric Bayesian approach to inference for monotone missing data with non-ignorable missingness in the presence of auxiliary covariates. Under the extrapolation factorization, we flexibly model the observed data distribution and specify the extrapolation distribution using identifying restrictions. We have shown that the inclusion of auxiliary covariates in the model could in general improve the accuracy of inferences and reduce the extent of a sensitivity analysis. We have also shown more accurate inferences can be obtained by using the proposed semiparametric Bayesian approach compared to using more restrictive parametric approaches and simple Bayesian nonparametric approaches.

The computational complexity in our application is manageable since the schizophrenia clinical trial dataset contains only 204 subjects. With much larger datasets computation becomes challenging. However, posterior simulation can be conducted in parallel for fitting the models of the observed responses, patterns, and auxiliary covariates (see Equation (11)). For each individual component, see Banerjee, Dunson, and Tokdar (2013), Hensman, Fusi, and Lawrence (2013), and Datta et al. (2016) for a scalable GP implementation and Pratola et al. (2014) for a scalable BART implementation. G-computation is easily scalable because it requires drawing independent hypothetical datapoint using each posterior sample.

When the number of auxiliary variables grows, it might be desirable to perform variable selection. Variable selection can be done through exploratory analysis, for example, fitting linear regression or spline regression models. Alternatively, it can be done more formally for each component of Equation (1). See

Savitsky, Vannucci, and Sha (2011) for variable selection for Gaussian process priors and Linero (2018) for variable selection for BART.

A possible extension of our work is to consider continuous time dropout. The GP is naturally suitable for the continuous case. Another extension would be more flexible incorporation of auxiliary covariates beyond the mean. Extending our method to nonmonotone missing data without imposing the partial ignorability assumption could be done with alternative identifying restrictions described in Linero and Daniels (2018) and possibly, a slightly modified semiparametric model. In the setting of binary outcomes, our method can be naturally extended by using a probit link. To identify the extrapolation distribution under NFD, we assume a location shift. Alternatively, we can consider exponential tilting (Rotnitzky, Robins, and Scharfstein 1998; Birmingham, Rotnitzky, and Fitzmaurice 2003).

Supplementary Materials

Appendix: Appendix showing more details of the schizophrenia clinical trial dataset, prior specification, MCMC implementation, G-computation, simulation studies, real data analysis, convergence diagnostics, and computing times.

Python package: Python package `bspmis` (with an R interface) containing code to perform the simulation studies and real data analysis described in the article (GNU zipped tar file).

Funding

Zhou, Daniels, and Müller were partially supported by NIH CA 183854.

References

- Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013), "Efficient Gaussian Process Regression for Large Datasets," *Biometrika*, 100, 75–89. [11]
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: CRC Press. [4]
- Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. M. (2003), "Pattern-Mixture and Selection Models for Analysing Longitudinal Data With Monotone Missing Patterns," *Journal of the Royal Statistical Society, Series B*, 65, 275–297. [11]
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [4]
- Chipman, H., and McCulloch, R. (2016), "BayesTree: Bayesian Additive Regression Trees," R Package Version 0.3-1.4. [5]
- Daniels, M. J., and Hogan, J. W. (2000), "Reparameterizing the Pattern Mixture Model for Sensitivity Analyses Under Informative Dropout," *Biometrics*, 56, 1241–1248. [5]
- (2008), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Boca Raton, FL: CRC Press. [1,2,3,5]
- Daniels, M., Wang, C., and Marcus, B. (2014), "Fully Bayesian Inference Under Ignorable Missingness in the Presence of Auxiliary Covariates," *Biometrics*, 70, 62–72. [2]
- Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016), "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets," *Journal of the American Statistical Association*, 111, 800–812. [11]
- De Oliveira, V. (2012), "Bayesian Analysis of Conditional Autoregressive Models," *Annals of the Institute of Statistical Mathematics*, 64, 107–133. [4]
- Diggle, P., and Kenward, M. G. (1994), "Informative Drop-Out in Longitudinal Data Analysis," *Journal of the Royal Statistical Society, Series C*, 43, 49–93. [1,2]

- Follmann, D., and Wu, M. (1995), “An Approximate Generalized Linear Model With Random Effects for Informative Missing Data,” *Biometrics*, 51, 151–168. [2]
- Geisser, S., and Eddy, W. F. (1979), “A Predictive Approach to Model Selection,” *Journal of the American Statistical Association*, 74, 153–160. [8]
- Gelfand, A. E., and Dey, D. K. (1994), “Bayesian Model Choice: Asymptotics and Exact Calculations,” *Journal of the Royal Statistical Society, Series B*, 56, 501–514. [8]
- Harel, O., and Schafer, J. L. (2009), “Partial and Latent Ignorability in Missing-Data Problems,” *Biometrika*, 96, 37–50. [2]
- Heckman, J. (1979), “Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153–162. [1]
- Henderson, R., Diggle, P., and Dobson, A. (2000), “Joint Modelling of Longitudinal Measurements and Event Time Data,” *Biostatistics*, 1, 465–480. [2]
- Hensman, J., Fusi, N., and Lawrence, N. D. (2013), “Gaussian Processes for Big Data,” in *Uncertainty in Artificial Intelligence*, eds. A. Nicholson and P. Smyth, Palo Alto, CA: Association for Uncertainty in Artificial Intelligence Press, pp. 282–290. [11]
- Hogan, J. W., and Laird, N. M. (1997), “Mixture Models for the Joint Distribution of Repeated Measures and Event Times,” *Statistics in Medicine*, 16, 239–257. [2]
- Kapelner, A., and Bleich, J. (2016), “bartMachine: Machine Learning With Bayesian Additive Regression Trees,” *Journal of Statistical Software*, 70, 1–40. [5]
- Kay, S. R., Flszbein, A., and Opfer, L. A. (1987), “The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia,” *Schizophrenia Bulletin*, 13, 261–276. [2]
- Kenward, M. G., Molenberghs, G., and Thijs, H. (2003), “Pattern-Mixture Models With Proper Time Dependence,” *Biometrika*, 90, 53–71. [5]
- Kim, C., Daniels, M. J., Marcus, B. H., and Roy, J. A. (2017), “A Framework for Bayesian Nonparametric Inference for Causal Effects of Mediation,” *Biometrics*, 73, 401–409. [5]
- Linero, A. R. (2017), “Bayesian Nonparametric Analysis of Longitudinal Studies in the Presence of Informative Missingness,” *Biometrika*, 104, 327–341. [2,3,6]
- (2018), “Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection,” *Journal of the American Statistical Association*, 113, 626–636. [11]
- Linero, A. R., and Daniels, M. J. (2015), “A Flexible Bayesian Approach to Monotone Missing Data in Longitudinal Studies With Nonignorable Missingness With Application To an Acute Schizophrenia Clinical Trial,” *Journal of the American Statistical Association*, 110, 45–55. [2,3,6,7,8,9,10]
- (2018), “Bayesian Approaches for Missing Not at Random Outcome Data: The Role of Identifying Restrictions,” *Statistical Science*, 33, 198–213. [1,2,5,11]
- Little, R. J. (1993), “Pattern-Mixture Models for Multivariate Incomplete Data,” *Journal of the American Statistical Association*, 88, 125–134. [2,3]
- (1994), “A Class of Pattern-Mixture Models for Normal Incomplete Data,” *Biometrika*, 81, 471–483. [2]
- Little, R. J., and Rubin, D. B. (2014), *Statistical Analysis With Missing Data*, New York: Wiley. [2]
- Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997), “The Analysis of Longitudinal Ordinal Data With Nonrandom Drop-Out,” *Biometrika*, 84, 33–44. [2]
- Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998), “Monotone Missing Data and Pattern-Mixture Models,” *Statistica Neerlandica*, 52, 153–161. [5]
- Müller, P., Erkanli, A., and West, M. (1996), “Bayesian Curve Fitting Using Multivariate Normal Mixtures,” *Biometrika*, 83, 67–79. [7]
- National Research Council (2011), *The Prevention and Treatment of Missing Data in Clinical Trials*, Washington, DC: National Academies Press. [1]
- Pratola, M. T., Chipman, H. A., Gattiker, J. R., Higdon, D. M., McCulloch, R., and Rust, W. N. (2014), “Parallel Bayesian Additive Regression Trees,” *Journal of Computational and Graphical Statistics*, 23, 830–852. [11]
- Pulkstenis, E. P., Ten Have, T. R., and Landis, J. R. (1998), “Model for the Analysis of Binary Longitudinal Pain Data Subject to Informative Dropout Through Remedication,” *Journal of the American Statistical Association*, 93, 438–450. [2]
- Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press. [3,6]
- Robins, J. (1986), “A New Approach to Causal Inference in Mortality Studies With a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 7, 1393–1512. [6]
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data,” *Journal of the American Statistical Association*, 90, 106–121. [2]
- Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), “Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse,” *Journal of the American Statistical Association*, 93, 1321–1339. [2,11]
- Rubin, D. B. (1976), “Inference and Missing Data,” *Biometrika*, 63, 581–592. [1,5]
- (1981), “The Bayesian Bootstrap,” *The Annals of Statistics*, 9, 130–134. [5]
- Savitsky, T., Vannucci, M., and Sha, N. (2011), “Variable Selection for Nonparametric Gaussian Process Priors: Models and Computational Strategies,” *Statistical Science*, 26, 130–149. [11]
- Scharfstein, D., McDermott, A., Olson, W., and Wiegand, F. (2014), “Global Sensitivity Analysis for Repeated Measures Studies With Informative Dropout: A Fully Parametric Approach,” *Statistics in Biopharmaceutical Research*, 6, 338–348. [6]
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), “Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models,” *Journal of the American Statistical Association*, 94, 1096–1120. [2]
- Shahbaba, B., and Neal, R. (2009), “Nonlinear Models Using Dirichlet Process Mixtures,” *Journal of Machine Learning Research*, 10, 1829–1850. [6]
- Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, New York: Springer Science & Business Media. [2]
- Tsiatis, A. A., Davidian, M., and Cao, W. (2011), “Improved Doubly Robust Estimation When Data Are Monotonely Coarsened, With Application To Longitudinal Studies With Dropout,” *Biometrics*, 67, 536–545. [2]
- Wade, S. (2013), “Bayesian Nonparametric Regression Through Mixture Models,” Ph.D. thesis, Bocconi University. [9]
- Wang, C., Daniels, M., Scharfstein, D. O., and Land, S. (2010), “A Bayesian Shrinkage Model for Incomplete Longitudinal Binary Data With Application to the Breast Cancer Prevention Trial,” *Journal of the American Statistical Association*, 105, 1333–1346. [2]
- Wu, M. C., and Carroll, R. J. (1988), “Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process,” *Biometrics*, 44, 175–188. [2]