BAYESIAN NONPARAMETRIC METHODS FOR CAUSAL INFERENCE AND PREDICTION

Bret Michael Zeldow

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Jason A. Roy, Associate Professor of Biostatistics

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Nandita Mitra, Professor of Biostatistics

Alisa Stephens-Shields, Assistant Professor of Biostatistics

Vincent Lo Re III, Associate Professor of Medicine and Epidemiology

Charles Leonard, Research Assistant Professor of Epidemiology

BAYESIAN NONPARAMETRIC METHODS FOR CAUSAL INFERENCE AND PREDICTION

ACKNOWLEDGEMENT

I would like to express my deepest gratitude for my advisor Jason Roy for his endless encouragement and brilliant ideas. His enthusiasm for our work consistently left me with renewed purpose and energy after our meetings. He is a great teacher and mentor, and I hope we can keep working together beyond my time at Penn.

I am also indebted to the rest of my committee–Alisa Stephens-Shields, Nandita Mitra, Vin Lo Re III, and Charlie Leonard–who have kindly given their time to make me a better statistician. Each one of you has been a great advocate for me and has provided valuable insight that has greatly improved this work. I feel extremely grateful to be able to know and work with all of you.

I would like to thank all the other faculty, students, and staff at Penn who helped contribute to my success. My funding has come from a variety of sources and I would like to acknowledge Phyllis Gimotty, Susan Ellenberg, Peter Snyder, Alisa Stephens-Shields, Jason Roy, and Justine Shults for their generosity.

Finally, I want to thank my family (Mom, Dad, and Whitney) and friends who have supported me financially and emotionally for many years. They laid the groundwork that made this work possible.

Lastly, I want to thank Jess Grody for her endless support, love, and encouragement. This journey wouldn't have been possible without her.

ABSTRACT

BAYESIAN NONPARAMETRIC METHODS FOR CAUSAL INFERENCE AND PREDICTION

Bret Michael Zeldow

Jason A. Roy

In this thesis we present novel approaches to regression and causal inference using popular Bayesian nonparametric methods. Bayesian Additive Regression Trees (BART) is a Bayesian machine learning algorithm in which the conditional distribution is modeled as a sum of regression trees. We extend BART into a semiparametric generalized linear model framework so that a portion of the covariates are modeled nonparametrically using BART and a subset of the covariates have parametric form. This presents an attractive option for research in which only a few covariates are of scientific interest but there are other covariates must be controlled for. Under certain causal assumptions, this model can be used as a structural mean model. We demonstrate this method by examining the effect of initiating certain antiretroviral medications has on mortality among HIV/HCV coinfected subjects. In later chapters, we propose a joint model for a continuous longitudinal outcome and baseline covariates using penalized splines and an enriched Dirichlet process (EDP) prior. This joint model decomposes into local linear mixed models for the outcome given the covariates and marginals for the covariates. The EDP prior that is placed on the regression parameters and the parameters on the covariates induces clustering among subjects determined by similarity in their regression parameters and nested within those clusters, sub-clusters based on similarity in the covariate space. When there are a large number of covariates, we find improved prediction over the same model with Dirichlet process (DP) priors. Since the model clusters based on regression parameters, this model also serves as a functional clustering algorithm where one does not have to choose the number of clusters beforehand. We use the method to estimate incidence rates of diabetes when longitudinal laboratory values from electronic health records are used to augment diagnostic codes for outcome identification. We later extend this work by using our EDP model in a causal inference setting using the parametric g-formula. We demonstrate this using electronic health record data consisting of subjects initiating second generation antipsychotics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Bayesian inference combines the full data likelihood of all observed and unobserved quantities with prior distributions for the unknown parameters. These priors reflect some degree of prior knowledge (or lack thereof) of the parameter values. By conditioning on observed data, we can calculate or approximate posterior distributions for the unknown parameters, combining information from the model assumptions, the prior distributions, and the observed data.

A potential drawback of Bayesian methods is that the full data likelihood must be specified. Data typically arise from complex scenarios that require many parameters to adequately describe it, but in usual statistical applications, often few parameters are of immediate scientific interest. For example, a researcher interested in the causal effect of a drug on a disease may not be concerned in reporting details as to why and in what situations doctors prescribe the drug, but such information may be essential for estimation of the parameter of interest. As such, it is necessary to include such information in the full data likelihood. Parameters describing parts of the likelihood that are not of scientific interest are often called nuisance parameters or the nuisance model. When the nuisance model is mispecified, such mispecification can affect the estimates for the parameters of interest in the form of bias, loss of efficiency, etc. Thus, the downside of the Bayesian setup is clear: correctly specifying a full data likelihood can be a daunting and even impossible task.

In classical (or frequentist) statistics, researchers have developed nonparametric or semiparametric methods which allow for all or part of the full data likelihood to remain unspecified. If the nuisance model is left unspecified, the researcher can proceed without fear of inducing bias due to the mispecification of the nuisance model. Fortunately, there is a Bayesian analog to these nonparametric and semiparametric methods, for which we use the umbrella term Bayesian nonparametrics. Bayesian nonparametrics are often more computationally intensive than their parametric counterparts but have experienced a boom in recent decades due to improvements in computing power. The idea behind Bayesian nonparametrics is simple. We cannot avoid full specification of the likelihood, but we can be as flexible as possible by introducing infinite dimensional parameters with appropriate priors. Commonly, these are priors on function spaces or probability measure spaces.

In this introduction, we will briefly demonstrate Bayesian nonparametrics with Dirichlet process priors, which are the most common nonparametric prior on the space of probability distributions. We also explore priors on function spaces that can be used in nonparametric Bayesian settings.

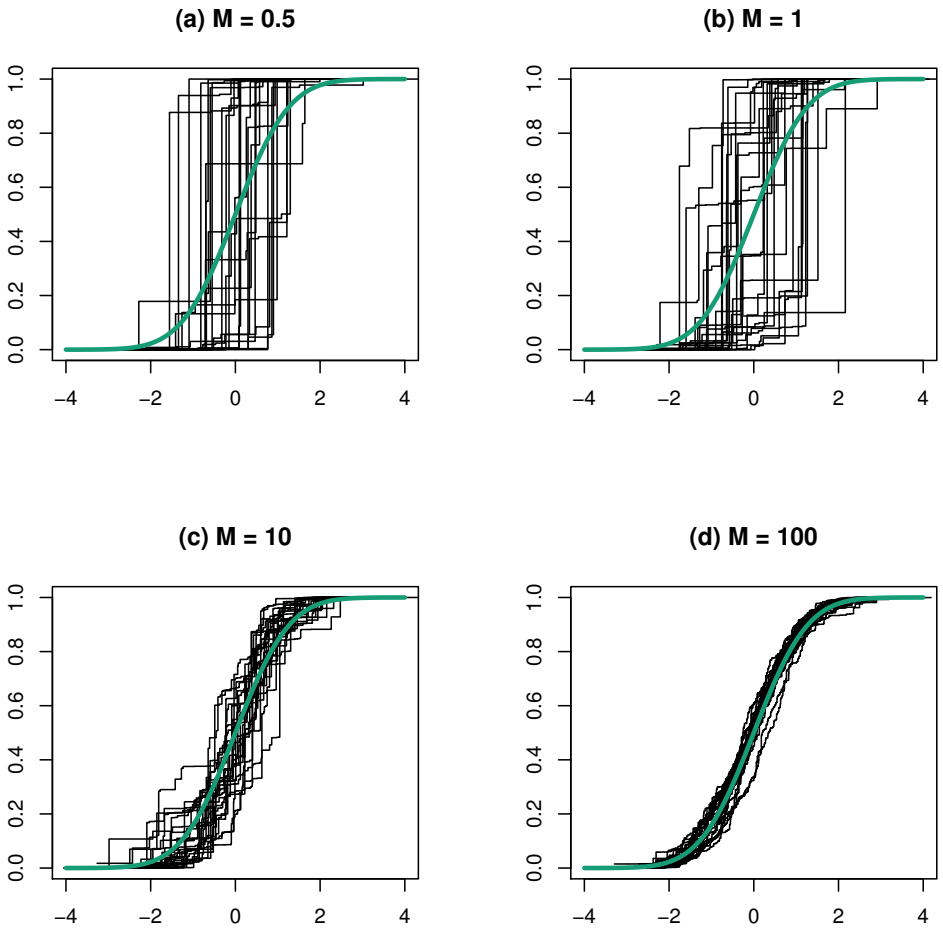
Dirichlet Process

The Dirichlet process (DP) is a popular Bayesian nonparametric prior (Ferguson, 1973) that can be used as a prior on the space of probability measures. The DP $G \sim \text{DP}(G_0, \alpha)$ is parameterized by a probability measure G_0 around which the DP is centered, and $\alpha > 0$, the mass parameter, which governs how close a draw from G is to G_0. Each draw from G is itself a probability measure. Two important properties of DPs are its discreteness–any draw from G can be written as an infinite sum of weighted point masses–and that G can weakly approximate any measure that has the same support as G_0 (Müller et al., 2015). These two properties are fundamental to the ubiquity of DPs in Bayesian nonparametrics, and we will demonstrate how this makes using a DP prior a departure from the usual parametric assumptions.

The base measure G_0 is the mean of G, written $E(G) = G_0$. As α grows larger, each draw from G is closer to G_0. The following is an illustration of DPs. Let G_0 be a normal distribution with mean 0 and variance 1. Figure 1.1 shows 25 draws from the DP G as α varies from 0.5, 1, 10, and 100. For each α, the distribution function for G_0 over the interval $[-4, 4]$ is shown in bold. When $\alpha = 0.5$, few atoms contain the majority of the mass and the draws are clearly distinct (but centered around) G_0. This is true as well for $\alpha = 1$, but the mass is more spread out across the atoms. As $\alpha = 10$, draws from G are noticeably nearer to G_0 and when $\alpha = 100$, draws from G are essentially G_0. The fact that the draws of G are step functions demonstrates its discreteness.

The discreteness of G has its drawbacks, however. When dealing with continuous density estimation, using a DP as the target distribution can be problematic. Instead, the DP is often used as a mixing distribution on a parametric distribution (Ferguson, 1983). That is,

Figure 1.1: Draws from a Dirichlet process.



Draws from a Dirichlet process $G \sim DP(G_0, \alpha)$ with base measure G_0, which is normal with mean 0 and variance 1. The mass parameter α varies between 0.5 and 100. (a) 25 draws when $\alpha = 0.5$. The draws are distinct from G_0 and consist mostly of a point that contains the majority of the mass. (b) 25 draws when $\alpha = 1$. Draws are still distinct from G_0 but the mass is spread around to several points. (c) Draws with $\alpha = 10$ are starting to resemble G_0. (d) With $\alpha = 100$, a draw from G is nearly G_0 itself.

$$
\begin{aligned}
y_i|\theta_i &\sim f(\cdot;\theta_i); \\
\theta_i|G &\sim G; \\
G &\sim \mathsf{DP}(G_0;\alpha),
\end{aligned}
\tag{1.1}
$$

where y_i is a continuous random variable with density $f_y(\cdot)$ parameterized by θ_i. Here the density of y_i is given a known parametric form. Each observation has its own θ_i but having been drawn from the discrete measure G, there is a positive probability of ties for θ_i among observations. Thus, some observations share the same θ_i. Integrating out the random probability measure yields an infinite mixture of parametric distribution.

$$f_G(y) = \int f(y;\theta) dG(\theta) \tag{1.2}$$

$$= \sum_{j=1}^{\infty} w_j f(y;\tilde{\theta}), \tag{1.3}$$

for some weights w_j depending on G.

Contrast this to the parametric Bayesian model below:

$$\begin{aligned} y_i|\theta &\sim f_y(\cdot;\theta); \\ \theta &\sim G_0. \end{aligned} \tag{1.4}$$

In the parametric version, the density is assumed to be of the form $f_y(\cdot;\theta)$. In the nonparametric version with a DP mixture, the density is an infinite mixture of $f_y(\cdot;\theta)$, which may assume arbitrary shape. In this density estimation example, we achieve greater flexibility merely by placing a DP prior in lieu of the parametric setup in which accuracy depends on correctly specifying the model in equation (1.4). In the following section, we continue our examination of nonparametric Bayesian priors focusing on function spaces.

Bayesian Additive Regression Trees

Consider an outcome Y and a vector of covariates X. To estimate Y given $X = x$, we may assume that $E(Y|X = x) = f(x)$ for some function $f(\cdot)$. The function $f(\cdot)$ can be parameterized by β so that $f(x;\beta) = x\beta$, as in linear regression (McCullagh, 1984). However, if we don't want to make that strong an assumption, we may consider the function $f(\cdot)$ as unknown and random and, under

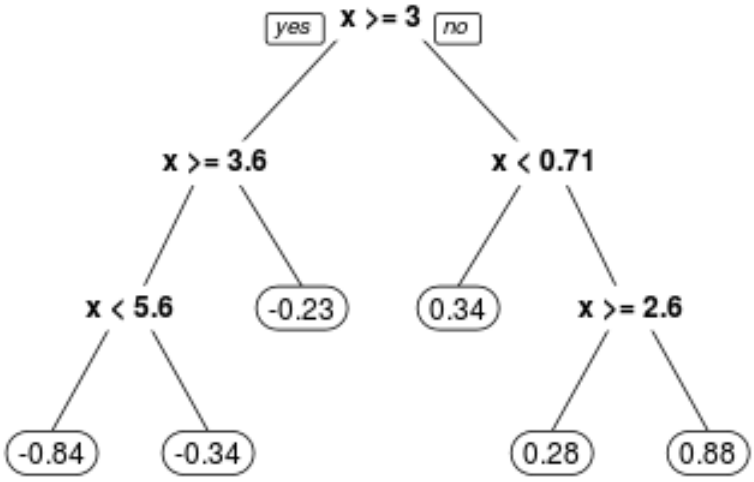
the Bayesian paradigm, place a prior on $f(\cdot)$ itself. One such option is the Gaussian process prior (Rasmussen, 2006). Note that we don't have to place the probability model directly on $f(\cdot)$. Instead, we can expand $f(\cdot)$ to be a sum of basis expansions ($f(x) = \sum \beta_i \phi_i(x)$) and put priors on the basis coefficients (Müller et al., 2015). In the next chapter of this dissertation, we adhere to this latter method using Bayesian Additive Regression Trees (BART) and write $f(\cdot)$ as a sum of Bayesian regression trees (Chipman, George, and McCulloch, 2010).

BART is a machine learning algorithm used to estimate an unknown function and make predictions of the outcome given covariates. To understand how BART works, it is necessary to understand terminology and methodology for a single regression tree (Chipman, George, and McCulloch, 1998). In the regression tree framework, the study population is split into subgroups based on a sequence of rules. Within each subgroup subjects have a similar mean response. Trees consists of interior nodes, splitting rules, and terminal nodes. Terminal nodes are the last node in a given sequence of interior nodes and splitting rules at which point the outcome Y is summarized. An example of a regression tree is shown in Figure 1.2. In this example, there is a single covariate predictive of a countinuous outcome. A subject with $x = 4$ would follow the leftmost path in the example figure, and the mean outcome of all subjects following this path (i.e., with $3.6 \leq x < 5.6$) is -0.84. Regression trees are widely available in off-the-shelf statistical software. However, they yield non-smooth estimates of the conditional distribution of Y given X, which may not be desirable in some applications. For an example of this, see Figure 1.3. We randomly choose points uniformly within the univariate predictor space $x \in [0, 2\pi]$. The outcome y is related to x through the relation $y = \sin(x) + \epsilon$ where ϵ is a normal error term. We assume the relationship between y and x is unknown and that the function relating y to x is the target of inference. Since the function $\sin(x)$ is non-linear, linear regression is the incorrect approach (solid line). Using a regression tree (dotted-dashed line) is a better fit than linear regression, but it still fails to capture the smoothness of the true function.

On the other hand, BART's sum-of-trees structure is adept at capturing the smoothness (dashed line in Figure 1.3). To implement BART, we set $f(x) = \sum \omega_i(x)$ where $\omega_i(x)$ is a regression tree and each $\omega_i(x)$ is restricted to be small (few terminal nodes). The sum-of-trees is more flexible and can better handle complex interactions and nonlinearities than a single tree. BART also has relatively few tuning parameters, which makes it an attractive option when one doesn't want to assume a parametric form for the unknown function. In the following chapter, we insert BART into a

generalized linear model setting where only a subset of the covariates are of scientific interest. The nuisance component is modeled with BART and the covariates of interest are modeled parametrically.

Figure 1.2: Example of a regression tree in a univariate covariate space.

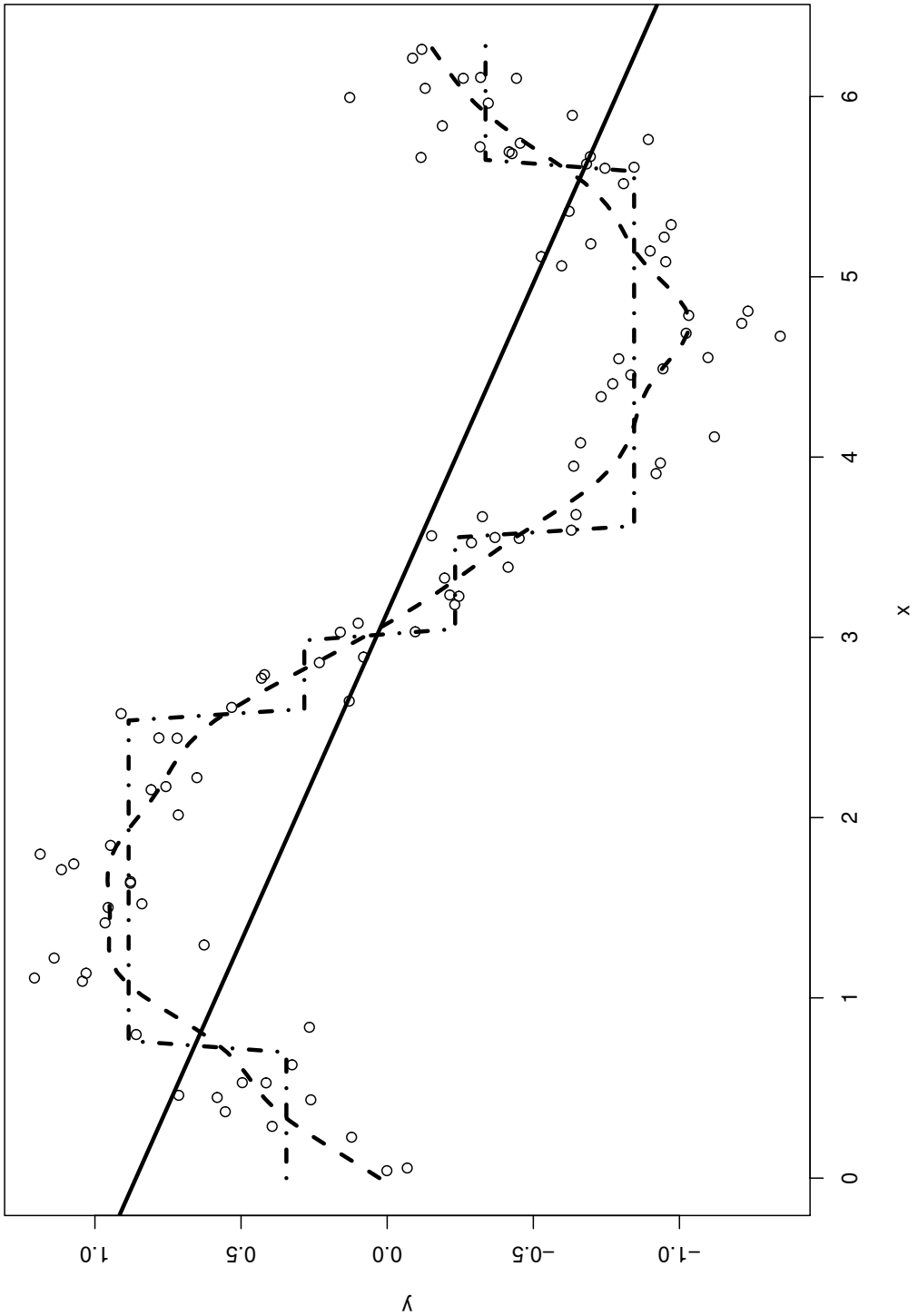


$x \in [0, 2\pi]$ and is related to the outcome y through the relation $y = \sin(x) + \epsilon$ where ϵ is a normal error term. Each interior node contains a splitting rule. If an observation satisfies the rule, the observation follows the leftmost path from that rule, until reaching the next splitting rule or terminal node. The summary at the terminal nodes refers to the mean outcome among all observations which follow the same sequence.

Bayesian methods in causal inference

Literature involving Bayesian methods in causal inference has grown in recent years. These include implementations of marginal structural models (Roy, Lum, and Daniels, 2016; Saarela et al., 2015) and the g-formula (Roy et al., 2017). In this dissertation, we aim to add to the literature by developing nonparametric Bayesian methods with emphasis on causal methods. In Chapter 2, we present a semiparametric regression model where only a small subset of covariates are of scientific interest using BART to control for confounding from other covariates. We show how this model can be used as a structural mean model, which has the advantage of avoiding g-estimation which is not possible for the probit and logit link functions, two popular link functions with binary outcomes. In Chapter 3, we present joint model for a continuous longitudinal outcome and the covariates using an enriched Dirichlet process, an improvement of a standard DP when the dimension of covariates is high. This offers improved prediction over competitor models and also serves as a functional clustering algorithm. In Chapter 4, we use this joint model for causal inference using the parametric g-formula.

Figure 1.3: Illustration of a BART fit with a univariate predictor space.



Here, $x \in [0, 2\pi]$ and mean response $y = \sin(x) + \epsilon$. The solid line is the fit using linear regression, the dashed line is the fit of BART, and the dashed-dotted line is the fit of a single tree.

CHAPTER 2

BAYESIAN SEMIPARAMETRIC REGRESSION AND STRUCTURAL MEAN MODELS WITH BART

Introduction

Semiparametric models, which include generalized estimating equations (GEE) and proportional hazards models, are some of the most commonly used models in statistics (Tsiatis, 2006). While the scope of semiparametrics is wide, the basic tenet is that we have a specific research question that is of interest (e.g., the effect of a treatment on an outcome) but in order to answer that question we must handle another part of the data that may not be of immediate scientific interest (e.g., adjusting for confounders). This latter part is referred to as the nuisance. A fully parametric model would need to model the nuisance parameters as well as the parameter of interest with a parametric form, but a model can be semiparametric by leaving the part for the nuisance parameters unspecified. Ideally, the semiparametric model can answer the scientific question of interest without inducing bias by the misspecification of the nuisance model.

The semiparametric framework is important in causal inference (Kennedy, 2016), which largely avoids fully parametric models for the aforementioned reasons. One of the most popular causal models, the marginal structural model, is semiparametric by leaving part of the conditional distribution of the outcome unspecified (Robins, Hernan, and Brumback, 2000). Marginal structural models were developed for longitudinal settings to adjust for time-varying confounding. A related but less used causal model, the structural mean model (SMM), is also semiparametric and was also developed for scenarios with time-varying confounding (Robins, 1986; Robins, 1994). Both marginal structural models and structural mean models can be used in the setting of an exposure at a single time point and still parameterize a meaningful causal contrast (Robins, 2000). Solving for the causal parameters in each of these models has been well documented in the literature (Hernán and Robins, 2018; Hernán, Brumback, and Robins, 2000). In particular, solving for the parameters of a SMM requires g-estimation, which amounts to solving estimating equations when the identity or log link function is used, as is typical for continuous or count outcomes (Hernán and Robins, 2018). When the outcome is binary and the common logit or probit link functions are preferred, no

easy solution exists for solving for the parameters of SMMs (Vansteelandt and Goetghebeur, 2003). Methods for this case have been proposed but require specifying a second model, which may introduce bias if specified incorrectly (Robins and Rotnitzky, 2004; Vansteelandt and Goetghebeur, 2003).

Recently, there have been Bayesian implementations of marginal structural models (Roy, Lum, and Daniels, 2016; Saarela et al., 2015). However, no Bayesian implementation of SMMs exists, though a fully parametric likelihood based model has been developed (Matsouaka and Tchetgen Tchetgen, 2014). Our aim is to develop the first fully Bayesian SMM, yielding posterior distributions for the causal parameters of interest while sidestepping the need for g-estimation and thus making estimation more robust when the outcome is binary. In doing so, we also find that our method is suitable for more general regression models, when causal assumptions might not be realistic or of interest, and can be used as a robust and intuitive semiparametric regression model in place of parametric regression. Our method can be easily implemented using our R package **semibart**, which is available on the author's website (https://www.github.com/zeldow/semibart).

The rest of the paper is organized as follows. Section 2 describes relevant background and a literature review. In Section 3, we describe the types of semiparametric models we are fitting, including SMMs and Bayesian semiparametric regression. Section 4 gives simulation results. In Section 5 we complete a data analysis on the effect of initiating certain antiretroviral drugs on death among adults with HIV infection who are newly initiating an antiretroviral regimen. In Section 6, we discuss strengths and limitations of our method.

Background

Let y be an outcome and let $\mathbf{X}$ be predictors of y. When y is continuous, consider the regression scenario $y_i = \omega(\mathbf{x}_i) + \epsilon_i$, with error terms ϵ assumed to be from from a distribution with mean zero. For non-continuous outcomes, we consider the model $E(y|\mathbf{X}) = g(\omega(\mathbf{X}))$ for a given link function g. The parametric linear regression which asserts that $\omega(\mathbf{x}_i;\beta) = \sum_{j=1}^{p} \beta_j x_{ij}$ with error terms $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ is easily solved using Bayesian methods when prior distributions are placed on β and the regression error variance σ^2 (Gelman et al., 2014). For the remainder of this paper, we focus on relaxing the assumption that $\omega(\mathbf{x}_i;\beta) = \sum_{j=1}^{p} \beta_j x_{ij}$, which can yield biased estimates if this assumption is far from the truth.

Our approach in modeling the relationship between Y and X targets the conditional mean of Y given X, which we denote as $\omega(\cdot)$. There is large statistical literature focused on modeling $\omega(\cdot)$ flexibly; we review some of these with added emphasis on Bayesian methods. We can think of $\omega(\cdot)$ as a random function by placing a prior distribution on the function space. One possible prior is the Gaussian process prior whose covariance structure can be specified such that the posterior captures nonlinear structures (Rasmussen, 2006). Other options for modeling $\omega(\cdot)$ include the use of basis functions (Müller et al., 2015) like splines (Eilers and Marx, 1996) or wavelets and placing prior distributions on the coefficients. Splines have been used extensively in Bayesian nonparametric and semiparametric regression. Biller (2000) presented a semiparametric generalized linear model where one variable is modeled using splines and the remaining variables were part of a parametric linear model(Biller, 2000). Holmes and Mallick (2001) developed a flexible Bayesian piecewise regression using linear splines (Holmes and Mallick, 2001). The approach in Denison et al (1998) involved piecewise polynomials and was able to approximate nonlinearities (Denison, Mallick, and Smith, 1998c). Biller and Fahrmeir (2001) introduced a varying-coefficient model with B-splines with adaptive knot locations (Biller and Fahrmeir, 2001).

Two of the most commonly used semiparametric methods that predict an outcome Y given covariates X are generalized additive models (GAM) (Hastie and Tibshirani, 1990) and multivariate adaptive regression splines (MARS) (Friedman, 1991), both of which were developed as frequentist procedures and are available in commonly used statistical software. GAM allows each predictor to have its own functional form using splines. The downside of GAM is that any interactions between covariates must be specified by the analyst, which can pose problems in high-dimensional problems in which there may be many multi-way interactions. Bayesian versions of GAM based on P-splines exist (Brezger and Lang, 2006) but do not have the widespread availability in statistical software that the frequentist version has. MARS is a fully nonparametric procedure which can automatically detect nonlinearities and interactions through basis functions also based on splines. A Bayesian MARS algorithm has also been developed (Denison, Mallick, and Smith, 1998b) but also lacks off-the-shelf software. A third option for nonparametric estimation of Y given X is Bayesian additive regression trees (BART), which like MARS, is adept at capturing nonlinearities and interactions between covariates, while being a fully Bayesian procedure (Chipman, George, and McCulloch, 2010). The details of BART are presented in more detail below.

Bayesian Additive Regression Trees

Bayesian additive regression trees (BART) is a machine learning algorithm designed to model an outcome as a function of covariates and a normal, additive error term. Let $Y = \omega(X) + \epsilon$ where Y is a continuous outcome, $\epsilon \sim N(0, \sigma^2)$, and $\omega(\cdot)$ is the unknown functional relating the predictors X to the outcome Y. For binary Y the probit link function is used, that is $\Pr(Y = 1|X) = \Phi(\omega(X))$, where $\Phi(\cdot)$ is the distribution function of a standard normal random variable. While other link functions for binary data (e.g., logit) are possible, using a probit link function simplifies Bayesian computations and is used in software implementing BART. BART estimates the function $\omega(\cdot)$ through a sum of regression trees, where a regression tree is a sequence of binary choices based on predictors X which yield predictions of Y within clusters of observations with similar covariate patterns. Classification and regression trees are typically frequentist procedures, but Bayesian versions of regression trees have also been developed (Chipman, George, and McCulloch, 1998; Denison, Mallick, and Smith, 1998a). The BART sum-of-trees model can be written as $\omega(x) = \sum_{i=1}^{m} \omega_i(x)$, where each $\omega_i(x)$ is itself a tree. Typically, the number of trees m is chosen to be large and each tree is restricted to have a small number of end nodes. This setup restricts the influence of any single tree while allowing detection of nonlinearities and interactions that would be not possible with one tree. An example of a BART fit to a nonlinear mean function $y = \sin(x) + \epsilon$ is shown in Figure 1.3 over a univariate predictor space x restricted to $[0, 2\pi]$, along with comparision to the fit of a single regression tree and linear regression.

The algorithm for BART utilizes Bayesian backfitting (Hastie, Tibshirani, et al., 2000). We review the algorithm for the case of continuous outcomes; the case for binary outcomes is a simple extension which utilizes the underlying normal latent variable formulation (Albert and Chib, 1993). Recall that $y_i = \sum_{j=1}^{m} \omega_j(x_i) + \epsilon_i$ where ϵ_i is assumed zero-mean normal with unknown variance σ^2. The algorithm iterates between updating the error variance σ^2 and updating the fit of the trees ω_j. The error variance σ^2 is updated by obtaining the residuals from the current fit and drawing the posterior from an inverse chi-square distribution when the conjugate inverse-chi square prior is used. Second, each tree ω_j is updated. For this step, we compute the residuals of the outcome by subtracting off the fit of the other $m - 1$ trees. When updating tree ω_1, the residuals $y_i^* = y_i - \sum_{i=2}^{m} \omega_j(x_i)$ are calculated. The fit for $\omega_1(\cdot)$ is updated through a proposed change to the tree (grow, prune, swap, or change) which is accepted or rejected through a Metropolis-Hastings step. The trees

$\omega_2(\cdot), \omega_3(\cdot), \ldots, \omega_m(\cdot)$ are all updated in the same fashion. More details are available in elsewhere (Chipman, George, and McCulloch, 2010). In the next section, we propose a semiparametric extension of BART, which we call semi-BART, where a small subset of covariates are allowed to have linear functional form and the rest are modeled with BART.

Semi-BART Model

Notation

Suppose we have n independent observations. Let Y denote the outcome, which we assume to be either binary or continuous. Denote by $\mathbf{L}$ the set of predictor variables. The outcome for individual $1 \leq i \leq n$ will be denoted as Y_i, with similar notation for covariates $\mathbf{L}_i$.

Semiparametric Generalized Linear Model

Our model imposes linearity on just a small subset of covariates of interest, while remaining flexible in modeling the rest of the covariates, whose exact functional form in relation to the outcome may be considered a nuisance. The predictors are partitioned into two distinct subsets so that $\mathbf{L} = \mathbf{L}_1 \cup \mathbf{L}_2$ and $\mathbf{L}_1 \cap \mathbf{L}_2 = \emptyset$. Here, $\mathbf{L}_1$ represents nuisance covariates that we must control for but is not of primary interest and $\mathbf{L}_2$ represents covariates that do have scientific interest. For continuous Y, we write $Y_i = \omega(\mathbf{L}_1) + h(\mathbf{L}_2; \psi) + \epsilon_i$, where $h(\cdot)$ is a linear function of its covariates in ψ (as in linear regression) but $\omega(\cdot)$ is a function with unspecified form. The errors ϵ_i are iid mean zero and normally distributed with unknown variance σ^2. More generally, we write $g\left[E(Y|\mathbf{L}_1, \mathbf{L}_2)\right] = \omega(\mathbf{L}_1) + h(\mathbf{L}_2)$, for a given link function g. We estimate $\omega(\cdot)$ using BART. Note that this implies that if $\mathbf{L_1} = \mathbf{L}$ and $\mathbf{L}_2 = \emptyset$, we have a nonparametric BART model. On the other hand if $\mathbf{L_1} = \emptyset$ and $\mathbf{L}_2 = \mathbf{L}$, we have a fully parametric regression model. While there is no restriction on the dimensionality of $\mathbf{L_1}$ and $\mathbf{L_2}$, in the typical case $\mathbf{L_1}$ is large enough that BART is a reasonable choice of an algorithm and $\mathbf{L_2}$ contains only a few covariates that are of particular interest.

Special Case: Structural Mean Models

We now consider the special case of a causal inference setting with observational data. We introduce further notation specific to this section. The exposure of interest is denoted A and can be either binary or continuous. The counterfactual Y^a denotes the outcome that would have been

observed under exposure $A = a$. For the special case of binary A, each individual has two counterfactual outcomes – Y^1 and Y^0 – but we observe at most one of the two, corresponding to the actual level of exposure received. That is, $Y = AY^1 + (1 - A)Y^0$. Let $\mathbf{X}$ be the set of confounders. We can further subset $\mathbf{X}$ into $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_2$ is the subset of variables that modify the causal effect of A and $\mathbf{X}_1$ are the other covariates. From the notation in the previous subsection, we can then think of $\mathbf{L}_2 = (A, \mathbf{X}_2)$ as the variables of primary interest and $\mathbf{L}_1 = \mathbf{X}_1$ as the variables that are not of interest but need to be controlled for.

Structural nested mean models are causal models developed by Robins to deal with time-varying confounding for longitudinal exposures (Robins, 1994, 2000). In the case of point treatment, structural nested mean models are referred to as structural mean models (SMMs) and parameterize a useful causal contrast even though time-varying confounding is not a concern (Vansteelandt and Joffe, 2014; Vansteelandt and Goetghebeur, 2003). This contrast encodes the mean effect of treatment among the treated given the covariates. The model can generally be written as:

$$g\left\{E\left(Y^a|\mathbf{X} = \mathbf{x}, A = a\right)\right\} - g\left\{E\left(Y^0|\mathbf{X} = \mathbf{x}, A = a\right)\right\} = h^*(x, a; \psi^*), \tag{2.1}$$

where g is a known link function. Here we switch from $h(\cdot; \psi)$ to $h^*(\cdot; \psi^*)$ to indicate that ψ^* represents a causal effect, but the two functions and parameters are otherwise identical. The goal of this paper is to provide a Bayesian solution to (2.1). First, we impose some restrictions on $h(\cdot; \psi^*)$. We require that under no treatment or when there is no treatment effect the function $h^*(\cdot; \psi^*)$ must equal 0. That is, $h^*(x, a; \psi^*)$ satisfies $h^*(x, a; 0) = h^*(x, 0; \psi^*) = 0$. Some examples of $h^*(\cdot; \psi^*)$ are $h^*(x, a; \psi^*) = \psi a$ or $h^*(x, a; \psi^*) = (\psi_1 + \psi_2 x)a$, when x is thought to be an effect modifier.

While expression (2.1) cannot be evaluated directly because of the unobserved counterfactuals, two assumptions are needed to identify it with observed data (Vansteelandt and Joffe, 2014).

1. Consistency: If $A = a$, then $Y^a = Y$;

2. Ignorability: $A \perp Y^0|X$.

The consistency assumption says that we actually get to see an individual's counterfactual corresponding to the exposure received. Ignorability ensures that there is no unmeasured confounding between the exposure A and the counterfactual under no treatment Y^0. Under these two assump-

tions together with the parametric assumption of $h^*(\cdot)$, the contrast on the left hand side of (2.1) is identified, and the SMM from (2.1) can be rewritten using observed variables as

$$g\left\{E\left(Y|X,A\right)\right\} = \omega(\mathbf{L}_2) + h^*(\mathbf{L}_1;\psi^*), \tag{2.2}$$

where $\omega(\mathbf{L}_1)$ is unspecified and $h^*(\mathbf{L}_2;\psi^*)$ is a linear function of $\mathbf{X}_2$ and A (Vansteelandt and Joffe, 2014). While we use the above assumptions for the remainder of this paper, the left hand side of (2.1) can be nonparametrically identified with a third assumption, dropping the parametric assumption of $h^*(\cdot)$. That is,

3. Positivity: $\Pr\left(A = a|\mathbf{X} = \mathbf{x}\right) > 0 \; \forall \; x$ such that $\Pr(\mathbf{X} = \mathbf{x}) > 0$.

The positivity assumption states that within all covariate levels $\mathbf{X} = \mathbf{x}$ that have positive probability of occurring, there is positive probability that an individual is treated. This assumption is violated in situations where treatment is deterministic at specific levels of $\mathbf{X} = \mathbf{x}$.

It should be noted that we have chosen a parametric form for all of $\mathbf{X}_2$, including the main effects of effect modifiers. In principal, one could include $\mathbf{X}_2$ in the nonparametric part and only model the interaction $\mathbf{X}_2 \times A$ parametrically. For example, if a researcher posits the relationship $h^*(x, a; \psi^*) = (\psi_1 + \psi_2 x)a$, the variable x in principle could be modeled nonparametrically. In simulations, we have found that including the covariates $\mathbf{X}_2$ into the BART model as well generally leads to poorer performance (bias, under coverage) of the causal effect posterior distributions. As a result, we would opt for the linear model $h^*(x, a; \psi^*) = (\psi_1 + \psi_2 x)a + \psi_3 x$ in our example. Further, in practice if researchers are interested in effect modification by $\mathbf{X}_2$ they might also be interested in interpreting the main effect.

Hill, 2011 has previously modeled causal effects on the treated using BART. The methods described in that paper correspond to our setting in equation (2.2) where g is the identity link function and ψ^* is a scalar describing only an effect of treatment with no effect modification. Our method extends this setup to settings with binary outcomes, continuous-valued treatment, or where low-dimensional summaries of effect modification are of interest. In settings with continuous outcomes, binary treatment, and no effect modification, the methods presented in Hill, 2011 may be preferred. We explore these differences using simulations.

Computations

The algorithm for semi-BART follows the BART algorithm with an additional step. We briefly reviewed the algorithm in Section 2.2.1. Below, we describe the basics of our algorithm for semi-BART. We are solving equation (2.2), where $\omega(\mathbf{L}_2)$ can be written as the sum-of-trees $\sum_{j=1}^{m} \omega_j(\mathbf{L}_2)$. Each tree $\omega_j(\mathbf{L}_2)$ has a vector of parameters $\boldsymbol{\theta}_j$ associated with it. The mean of the k^{th} endnode of the j^{th} tree is assumed to be normally distributed with mean μ_{jk} and variance σ^2_{jk}.

Recall that when the outcome is continuous, we assume independent errors distributed $N(0, \sigma^2)$. The algorithm for semi-BART for continuous outcomes is as follows. First, we initialize all values including the error variance σ^2, the parameters ψ^*, and the tree structure $\omega(\mathbf{L}_1)$ for all m trees. Next we begin our MCMC algorithm and iterate through the following steps. First update the m trees one at a time. When updating the j^{th} tree, subtract the fit of the remaining $m-1$ trees at their current parameter values as well as the fit of the linear part $h^*(\mathbf{L}_2; \psi^*)$ at the current value of ψ^* from the value of y for each individual. That is, we calculate $y_i^* = y_i - \omega_{-j}(\mathbf{L}_{1i}) - h^*(\mathbf{L}_{2i}; \psi^*)$, where $\omega_{-j}(\mathbf{L}_{1i})$ indicates the fit of the $m-1$ without the j^{th} tree. A modification of the j^{th} tree is now proposed. We either grow the tree (add a split point to what was previously an endnode), prune the tree (collapse two endnodes into one), change a splitting rule (for nonterminal nodes), or swap the rules between two nodes. Once a modification is proposed, we accept or reject this modification with a Metropolis-Hastings step (Chipman, George, and McCulloch, 1998). The parameters $\boldsymbol{\theta}_j$ are then updated from draws based on the conjugate priors (normal priors for μ_{jk} and inverse chi-squared for σ^2_{jk}).

Next we update ψ^*. To do this, we calculate the residuals after subtracting off the fit of all m trees. That is, calculate $y_i^* = y_i - \omega(\mathbf{L}_{1i})$. With a conjugate multivariate normal prior with mean $\boldsymbol{\psi}_0$ and variance $\sigma^2_{\psi}\mathbf{I}$ on ψ^* where $\mathbf{I}$ is the identity matrix of appropriate dimension, updating ψ^* is simply a draw from a multivariate normal distribution. The posterior for ψ is multivariate normal with covariance $\Sigma_\psi = \left[\frac{\mathbf{L}_2^T\mathbf{L}_2}{\sigma^2} + \frac{\mathbf{I}}{\sigma^2_\psi}\right]^{-1}$ and mean $\Sigma_\psi \left[\frac{\mathbf{L}_2 y^*}{\sigma^2} + \frac{\boldsymbol{\psi}_0}{\sigma^2_\psi}\right]$ (Gelman et al., 2014).

Finally, we update σ^2. We calculate the residuals by conditioning on the fitted trees $\theta_{\mathbf{i}}$ and the parametric part ψ^* and subtracting off the fit of all m trees and the linear part $h^*(\cdot)$. That is, calculate $y_i^* = y_i - \omega(\mathbf{L}_{1i}) - h(\mathbf{L}_{2i}; \psi^*)$. We use a conjugate inverse chi squared distribution for σ^2 and draw from an updated inverse chi squared distribution. We then return to updating each of the m trees and continue until the posterior distributions are well approximated.

The algorithm for binary outcomes with a probit link uses the underlying latent continuous variable formulation of (Albert and Chib, 1993) and is inserted into the algorithm in lieu of updating the error variance σ^2. Further details of the latent variable step as well as other steps pertaining to BART such as choosing a variable for a split point or choosing the splitting rule can be found in (Chipman, George, and McCulloch, 2010). The full implementation of our algorithm is available at https://www.github.com/zeldow/semibart.

Simulations

We used simulation to assess performance of our model under both continuous and binary outcomes. We generated five binary covariates ($x_1, \ldots, x_5$) from independent Bernoulli random variables and twenty continuous covariates ($x_6, \ldots, x_{25}$) from a multivariate normal distribution for a total of 25 predictors. The binary covariate x_1 is considered to be the treatment variable. The covariates $x_6, \ldots, x_{10}$ were generated with non-zero correlation with each other but independent of the rest, as were covariates $x_{11}, \ldots, x_{15}$, covariates $x_{16}, \ldots, x_{20}$, and covariates $x_{21}, \ldots, x_{25}$. Exact distributions for covariate generation are given in the Appendix (Section A). Outcomes were generated with both linear and non-linear mean functions and were related to only the first 10 covariates. In the continuous case, outcomes were generated from a normal distribution with standard deviations 0.1, 1, 2, and 3 (results in the main text are presented with a standard deviation of 1). In the linear cases, outcomes were given mean $\mu_{\ell,1} = 1+2x_1+2x_5+2x_6-0.5x_7-0.5x_8-1.5x_{10}$ when only the effect of x_1 was of interest, or $\mu_{\ell,2} = 1 + 2x_1 + 2x_5 + 2x_6 - 0.5x_7 - 0.5x_8 - 1.5x_{10} - x_1x_6$ when effect modification of x_6 on x_1 was also of interest. For nonlinear models, outcomes were given mean $\mu_{\text{nl},1} = 1 + 2x_1 + 2x_6 + \sin(\pi x_2 x_7) - 2\exp(x_3x_5) + \log\left(|\cos(\frac{\pi}{2}x_8)|\right) - 1.8\cos x_9 + 3x_3|x_7|^{1.5}$ or $\mu_{\text{nl},2} = 1+2x_1+2x_6+\sin(\pi x_2 x_7)-2\exp(x_3x_5)+\log\left(|\cos(\frac{\pi}{2}x_8)|\right)-1.8\cos x_9+3x_3|x_7|^{1.5}-x_1x_6$, for no effect modification and effect modification, respectively. Our goal was to predict the treatment effect of x_1 and when appropriate, the effect modification of x_6 on x_1. For the linear case, linear regression is the correctly specified model. In the non-linear cases, the treatment effect and the effect modification are linear but the rest of the covariates have a nonlinear relationship with the outcome. In addition to comparing semi-BART and linear regression, we also estimated effects with doubly robust g-estimation with logistic regression for the treatment model and linear regression for the outcome model. Since the treatment is generated independently, the treatment model is correctly specified with coefficients of 0. In cases with no effect modification, we also estimated

the treatment effect using Hill's estimate of the treatment effect on the treated using BART (Hill, 2011).

For binary outcomes, we generated 25 covariates in the same way, but outcomes were generated from a Bernoulli distribution. For the linear case with no effect modification, outcomes were generated with probability $p_{\ell,1} = \Phi(0.1+0.3x_1+0.1x_2+0.04x_6-0.02x_7+0.04x_9-0.03x_{10})$, where $\Phi(\cdot)$ is the distribution function for a standard normal variable. For the linear case with effect modification (of x_6 on x_1), outcomes were generated with probability $p_{\ell,2} = \Phi(0.1+0.3x_1+0.1x_2+0.04x_6-0.02x_7+0.04x_9-0.03x_{10}-0.1x_1x_6)$. In the nonlinear case, outcomes were generated with probability $p_{\text{nl},1} = \Phi(0.1+0.3x_1+0.04x_6-\sin(\pi x_2x_7)+\frac{1}{10}\exp(x_7/3)+\mathbb{1}_{[x_9>2]}\cos(x_8)+\mathbb{1}_{[x_9<1]}\cos(x_{10})-0.01x_7x_8x_{10})$ when there was no effect modification and $p_{\text{nl},2} = \Phi(0.1+0.3x_1+0.04x_6-\sin(\pi x_2x_7)+\frac{1}{10}\exp(x_7/3)+\mathbb{1}_{x_9>2}\cos(x_8)+\mathbb{1}_{x_9<1}\cos(x_{10})-0.01x_7x_8x_{10}-0.1x_1x_6)$ with effect modification between x_6 and x_1. Again, we are interested in extracting the treatment effect of x_1 on the outcome, and if applicable, the effect modification of x_6 on x_1. We estimate these effects from semi-BART and probit regression and compare results from the two models. For the linear case, probit regression is the correctly specified model.

For all scenarios, we generated 500 datasets each at sample sizes of $n = 250, 1000,$ and 5000. All models are compared on mean bias, 95% coverage probability of the confidence or credible interval, and mean squared error (MSE). For semi-BART, we used 10,000 total iterations the first 2,500 of which were burn-in. For the BART part of the model we used 200 trees. The prior distribution on the parameters of interest was independent mean zero normal with a standard deviation of 4, which is a diffuse prior given that the outcome was scaled and centered to be between $-\frac{1}{2}$ and $\frac{1}{2}$. For Hill's treatment on the treated with BART, we used all default values from the BayesTree package in R (Chipman, George, and McCulloch, 2010).

Continuous Outcome - No Effect Modification

The results of our simulations for continuous outcomes with no effect modification is shown in Table 2.1. Outcomes were generated with a standard deviation of 1. The true parameter for the treatment effect is $\psi = 2.0$. In the linear case (shown in the top half of the table and generated by mean function $\mu_{\ell,1}$), linear regression is the correctly specified model. As expected, it has lower MSE than the methods using BART, particularly at the smaller sample sizes. However, both semi-

BART and the pure BART results are unbiased and are nearly as efficient as linear regression at $n = 5000$ (MSE = 0.001 for all). G-estimation, being comprised of linear models as well, is nearly equivalent to linear regression in terms of bias, coverage, and MSE. The lower half of Table 2.1 shows the results when the mean function is largely non-linear, generated through mean function $\mu_{\text{nl},1}$. In this scenario, semi-BART and BART have much lower MSE than linear regression at all sample sizes. All methods are unbiased with good coverage. Note that at $n = 250$, the MSE for BART is 0.066 while the MSE for semi-BART is 0.104. Results for these simulations using outcomes drawn with standard deviations 0.1, 2, and 3 are displayed in Appendix B, Tables B.1, B.2, and B.3, respectively. The results are similar to those in Table 2.1.

Continuous Outcome - Effect Modification

The results of our simulations with a continuous outcome and a continuous effect modifier for the treatment effect are shown in Table 2.2. The true value for the treatment effect is $\psi_1 = 2.0$ and the true value for the effect modification of x_6 on the treatment is $\psi_2 = -1.0$. In the top half of the table denoting the linear case generated by mean function $\mu_{\ell,2}$, linear regression is the correctly specified model. At $n = 250$, linear regression has lower MSE for ψ_1 than semi-BART (0.126 vs. 0.153) and for ψ_2 (0.025, 0.031). At $n = 5000$, the two methods have the same rounded MSE (0.005 for ψ_1 and 0.001 for ψ_2). All methods are unbiased with coverage around the nominal level. In the non-linear case (generated by mean function $\mu_{\text{nl},2}$, the bias for ψ_1 using linear regression is slightly larger in absolute value than for semi-BART or g-estimation (-0.07 for linear regression and -0.03 for the rest). In terms of MSE, semi-BART is much more efficient than linear regression for both parameters and all sample sizes compared to linear regression and g-estimation with linear models. Results for these simulations using outcomes drawn with standard deviations 0.1, 2, and 3 are displayed in Appendix B, Tables B.4, B.5, and B.6, respectively. The results are similar to those in Table 2.2.

Binary Outcome - No Effect Modification

The simulation results with a binary outcome and no effect modification are shown in Table 2.3. The true value for the treatment effect is $\psi = 0.3$. In the linear case, outcomes are generated with probability $p_{\ell,1}$ and probit regression is the correctly specified model. In the non-linear case with probabilities $p_{\text{nl},2}$, there is some bias at $n = 250$ for both models, slightly larger for semi-BART (0.07

Table 2.1: Efficiency of Semi-Bart for a continuous outcome without effect modification.

		Semi-BART			Linear Regression			g-estimation			BART		
Mean Function	n	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	-0.02	0.96	0.039	-0.01	0.97	0.024	-0.01	0.94	0.024	-0.06	0.93	0.034
	1000	0.00	0.93	0.007	0.00	0.94	0.006	0.00	0.94	0.006	-0.00	0.93	0.007
	5000	-0.00	0.93	0.001	-0.00	0.92	0.001	-0.00	0.93	0.001	-0.00	0.92	0.001
Non-linear	250	-0.02	0.97	0.104	-0.01	0.95	0.319	-0.02	0.96	0.324	-0.03	0.94	0.066
	1000	-0.00	0.95	0.008	-0.01	0.95	0.076	-0.01	0.96	0.076	-0.01	0.94	0.009
	5000	-0.00	0.94	0.001	0.01	0.94	0.014	0.01	0.96	0.014	-0.00	0.95	0.001

The true value for the treatment effect is $\psi = 2.0$. The column titled BART refers to Hill's treatment effect on the treated from Hill (2011). The column g-estimation is also used for continuous outcomes only refers to doubly robust g-estimation using logistic regression for the treatment model and linear regression for the outcome model. We display the mean bias, the empirical coverage probability of the confidence/credible interval, and the empirical mean squared error over 500 simulated datasets.

Table 2.2: Efficiency of Semi-BART for a continuous outcome with effect modification.

			Semi-BART			Linear Regression			g-estimation		
Mean Function	n		Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	ψ_1	-0.01	0.96	0.153	0.02	0.96	0.126	0.01	0.93	0.141
		ψ_2	0.00	0.96	0.031	-0.01	0.95	0.025	-0.00	0.93	0.029
	1000	ψ_1	-0.00	0.96	0.029	-0.00	0.96	0.027	0.00	0.95	0.027
		ψ_2	0.00	0.97	0.006	0.00	0.96	0.005	0.00	0.95	0.005
	5000	ψ_1	0.01	0.95	0.005	0.00	0.96	0.005	0.00	0.96	0.005
		ψ_2	-0.00	0.96	0.001	-0.00	0.95	0.001	-0.00	0.96	0.001
Non-linear	250	ψ_1	-0.03	0.98	0.450	-0.07	0.94	1.570	-0.03	0.94	1.991
		ψ_2	0.01	0.96	0.102	0.03	0.94	0.332	0.01	0.92	0.432
	1000	ψ_1	-0.00	0.95	0.039	-0.01	0.94	0.332	-0.01	0.96	0.362
		ψ_2	0.01	0.94	0.008	0.01	0.94	0.073	0.00	0.96	0.082
	5000	ψ_1	-0.00	0.95	0.006	0.01	0.96	0.068	0.02	0.96	0.075
		ψ_2	0.00	0.95	0.001	-0.00	0.94	0.015	-0.01	0.96	0.017

The true value for the treatment effect is $\psi_1 = 2.0$ and the true value pertaining to effect modification between x_1 and x_6 is $\psi_2 = -1.0$. The column g-estimation is also used for continuous outcomes only refers to doubly robust g-estimation using logistic regression for the treatment model and linear regression for the outcome model

Table 2.3: Efficiency of Semi-BART for a binary outcome without effect modification.

		Semi-BART			Probit Regression		
Mean Function	n	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	0.07	0.92	0.059	0.04	0.92	0.052
	1000	0.03	0.93	0.011	0.01	0.95	0.009
	5000	0.01	0.95	0.002	-0.00	0.95	0.002
Non-linear	250	-0.02	0.91	0.058	-0.04	0.91	0.056
	1000	-0.01	0.95	0.012	-0.04	0.92	0.012
	5000	-0.00	0.94	0.002	-0.05	0.79	0.004

The true value for the treatment effect is $\psi = 0.3$.

for semi-BART and 0.04 for probit regression). The bias gets smaller for both models as the sample size increases. The MSE for probit regression is smaller than semi-BART at all sample sizes, but the difference is not as pronounced as with continuous outcomes in Table 2.1. In the non-linear case with outcomes generated with probability, $p_{\text{nl},1}$, there is slight initial bias for semi-BART (-0.02) and probit regression (-0.04). However, the bias for probit regression is persistent at all sample sizes whereas the bias vanishes in the semi-BART model. In terms of MSE, semi-BART and probit regression are similar, except perhaps at $n = 5000$ where the MSE with semi-BART is 0.002 versus 0.004 for probit regression.

Table 2.4: Efficiency of Semi-BART for a binary outcome with effect modification.

			Semi-BART			Probit Regression		
Mean Function	n		Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	ψ_1	0.07	0.92	0.322	0.04	0.93	0.291
		ψ_2	-0.01	0.93	0.063	-0.01	0.92	0.057
	1000	ψ_1	0.04	0.95	0.057	0.02	0.95	0.049
		ψ_2	-0.01	0.93	0.011	-0.01	0.93	0.010
	5000	ψ_1	0.00	0.94	0.010	-0.01	0.94	0.009
		ψ_2	0.00	0.94	0.002	0.00	0.94	0.002
Non-linear	250	ψ_1	0.00	0.94	0.294	-0.02	0.94	0.276
		ψ_2	-0.01	0.95	0.055	0.00	0.94	0.052
	1000	ψ_1	-0.01	0.94	0.054	-0.04	0.94	0.049
		ψ_2	0.00	0.95	0.011	0.01	0.94	0.010
	5000	ψ_1	0.00	0.95	0.011	-0.05	0.92	0.011
		ψ_2	-0.00	0.97	0.002	0.02	0.93	0.002

The true value for the treatment effect is $\psi_1 = 0.3$ and the true value for the effect modification parameter is $\psi_2 = -0.1$.

Binary Outcome - Effect Modification

Results for a binary outcome with a continuous effect modifier for the treatment effect are shown in Table 2.4. The true value for the treatment effect is $\psi_1 = 0.3$ and the true value for the effect modification parameter is $\psi_2 = -0.1$. In the linear case, outcomes were generated with probability $p_{\ell,2}$. Here, probit regression is more efficient than semi-BART at low sample sizes (the MSE for both parameters is about 1.1 times higher at $n = 250$). There is some bias at $n = 250$, that of semi-BART is higher than that of probit regression (0.07 verus 0.04 for ψ_1). However, at $n = 5000$, the results from the two models are nearly identical, as the bias in semi-BART went to 0. For the non-linear case, outcomes were generated with probability $p_{\text{nl},2}$. Here, probit regression shows some persistent bias for ψ_1, which is not the case for semi-BART. Despite this, the MSEs for probit regression and semi-BART are nearly identical, driven by the lower empirical variance of estimates from probit regression.

Data Application

To illustrate our method we analyzed data from the Veterans Aging Cohort Study (VACS) from 2002 to 2009, which is a cohort of HIV-infected patients being treated at Veterans Affairs facilities in the United States. Our study sample consisted of patients with HIV/Hepatitis C coinfection who were newly initiating antiretrovirals (including at least one nucleoside reverse transcriptase inhibitor [NRTI]) and had at least six months of observations recorded in VACS prior to initiation. Certain

NRTIs are known to cause mitochrondial toxicity. These mitochrondial toxic NRTIs (mtNRTIs) include didanosine, stavudine, zidovudine, and zalcitabine (Soriano et al., 2008). While these drugs are no longer part of first line HIV treatment regimens, they are still used in resource-limited settings or in salvage regimens (Günthard et al., 2016).

Exposure to mtNRTIs may increase the risk of hepatic injury which in turn may increase the risk of hepatic decompensation and death (Scourfield et al., 2011). The goal of this analysis was to determine if initiating an antiretroviral regimen containing a mtNRTI increased the risk of death versus antiretroviral containing a NRTI that is not a mtNRI. VACS data contains a number of variables confounding the relationship between mtNRTI use and death including subject demographics, year of antiretroviral initiation, HIV characteristics such as CD4 count and HIV viral load, concomitant medications, and laboratory measures relating to liver function.

One of the covariates included in our analysis is Fibrosis-4 (FIB-4), an index that measures hepatitic fibrosis with higher values indicating larger injury. Specifically FIB-4 > 3.25 (no units) indicates advanced hepatic fibrosis. FIB-4 can be calculated as:

$$[\text{age (years)} \times \text{AST (U/L)}] \,/\, \left[\text{platelet count}(10^9/\text{L}) \times \sqrt{\text{ALT (U/L)}}\right]$$

(Sterling et al., 2006). Here, AST stands for aspartate aminotransferase and ALT for alanine aminotransferase. There is some concern in that mtNRTI use in subjects with high FIB-4 will result in higher risk of liver decompensation and death than in subjects who have low FIB-4. Thus, we consider FIB-4 as a possible effect modifier of the effect of mtNRTIs on death.

The outcome is a binary indicator of death within a two-year period after the subject initiated antiretroviral therapy. While covariates were updated in the study, we only considered baseline values for this analysis. There were some missing values among the predictors that were handled through a single imputation. A previous analysis of this data used multiple imputation to handle missing covariates but found that results were very similar across imputations. All continuous covariates were centered at meaningful values. For example, age was centered around 50 years and year of study entry was centered at 2005.

In the first analysis we sought to determine the effect of mtNRTI use on death without considering

effect modification, and to this extent we fit a Bayesian SMM with a probit link. The estimand can be written as

$$\Phi^{-1}\left\{E\left(Y^a|\mathbf{X}=\mathbf{x},A=a\right)\right\}-\Phi^{-1}\left\{E\left(Y^0|\mathbf{X}=\mathbf{x},A=a\right)\right\}=\psi a, \tag{2.3}$$

where Y is the indicator of death, A represents whether mtNRTIs were part of the antiretroviral regimen at baseline ($A=1$ if mtNRTI were included in the regimen), and $\mathbf{X}$ all other covariates, including FIB-4. In the second and third analysis, we considered FIB-4 to be an effect modifier, once as a continuous covariate and once as a binary indicator which equaled 1 whenever FIB-4 >3.25. This estimand can be written as

$$\Phi^{-1}\left\{E\left(Y^a|\mathbf{X}=\mathbf{x},A=a\right)\right\}-\Phi^{-1}\left\{E\left(Y^0|\mathbf{X}=\mathbf{x},A=a\right)\right\}=\psi_1 a+\psi_2 a x_1, \tag{2.4}$$

where x_1 corresponds to the appropriate FIB-4 variable.

The analysis was conducted using $m=200$ trees with 20,000 total iterations (5,000 burn-in). The prior distribution on the ψ parameters were independent Normal$(0,4^2)$. In the first analysis the mean estimate of the posterior distribution for ψ was 0.15 (95% credible interval (CI): -0.02, 0.33). Notably the interval includes 0, but the direction of the point estimate indicates that subjects initiating antiretroviral therapy with an mtNRTI had greater risk of death within 2 years than subjects initiating therapy without an mtNRTI. We can interpret this coefficient in terms of $E\left(Y^0|\mathbf{X}=\mathbf{x},A=a\right)$ and $E\left(Y^a|\mathbf{X}=\mathbf{x},A=a\right)$ through the causal contrast in equation (2.3). Figure 2.1a shows the value of $E(Y^1|\mathbf{X}=\mathbf{x},A=1)$ as a function of $E(Y^0|\mathbf{X}=\mathbf{x},A=1)$ for $\psi=0.15$. As an example, suppose the unknowable quantity $E(Y^0|\mathbf{X}=\mathbf{x},A=1)=0.20$. This means that subjects treated with a mtNRTI ($A=1$) with covariates $\mathbf{X}=\mathbf{x}$ would have had a probability of death of 20% within 2 years had they been untreated ($A=0$). However, given $\psi=0.15$ we see that if $E(Y^0|\mathbf{X}=\mathbf{x},A=1)=0.20$ then $E(Y^1|\mathbf{X}=\mathbf{x},A=1)=0.24$, an increase of 4%. One can examine the change in probability for other base probabilities $E(Y^0|\mathbf{X}=\mathbf{x},A=1)$ by examining the graph in Figure 2.1a. The trace plot for this analysis is given as Figure A.1 in Section A.3 of the Appendix.

We conducted a second analysis with FIB-4 as a continuous effect modifier (centered around 3.25) with the same settings as the previous one. This analysis corresponds to the contrast from equation (2.4). Here, the estimate for the main effect of mtNRTI was $\psi_1=0.18$ (0.00, 0.36) and the

interaction between mtNRTI use and FIB-4 was $\psi_2 = 0.07$ (0.02, 0.12). The results can be viewed in Figure 2.1b. Again, for illustration, consider the special case where $E(Y^0|\mathbf{X} = \mathbf{x}, A = 1) = 0.20$. When FIB-4 is 3.25, then $E(Y^1|\mathbf{X} = \mathbf{x}, A = 1) = 0.25$. However, at larger values such as a FIB-4 of 5.25, $E(Y^1|\mathbf{X} = \mathbf{x}, A = 1) = 0.30$. The trace plot for this analysis is given as Figure A.2 in Section A.3 of the Appendix.

Finally we did a third analysis with FIB-4 as a binary effect modifier (> 3.25 vs. ≤ 3.25). Here we found that $\psi_1 = 0.07$ (-0.12, 0.26) and $\psi_2 = 0.38$ (0.07, 0.69). These results can be viewed in Figure 2.1c. Here, we see that if $E(Y^0|\mathbf{X} = \mathbf{x}, A = 1) = 0.20$, then $E(Y^1|\mathbf{X} = \mathbf{x}, A = 1) = 0.22$ for subjects with FIB-4 ≤ 3.25 and $E(Y^1|\mathbf{X} = \mathbf{x}, A = 1) = 0.35$ for subjects with FIB-4 > 3.25. The trace plot for this analysis is given as Figure A.3 in Section A.3 of the Appendix.

Discussion

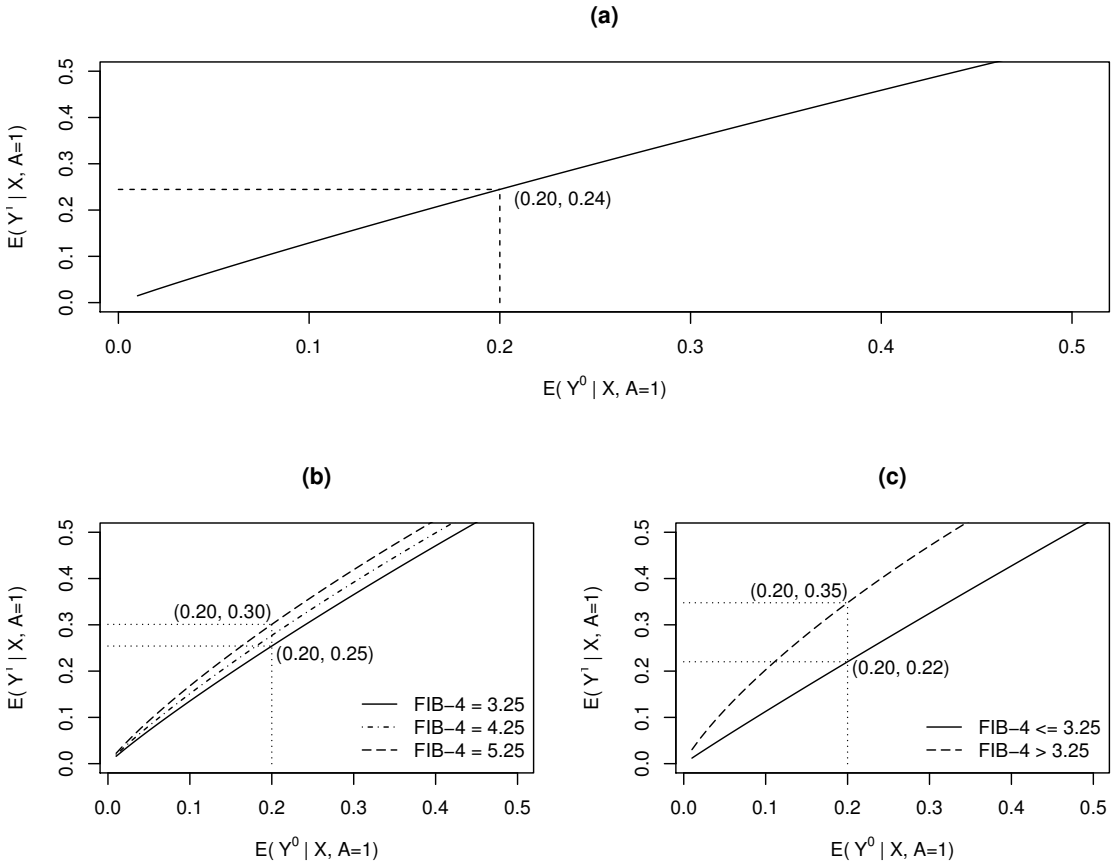
We presented a new Bayesian semiparametric model, which can be implemented with an R package **semibart** that is available from the author's GitHub page (https://github.com/zeldow/semibart). Our model allows for flexible estimation of the nuisance parameters while being fully parametric for covariates that are of immediate scientific interest, providing a viable and intuitive alternative to fully parametric regression. Under some causal assumptions, this model can as be interpreted as a SMM, which also provides the first fully Bayesian SMM. This is particularly useful in the case of binary outcomes where g-estimation is not possible. Vansteelandt (2003) provided approaches for estimating SMMs with binary outcomes in frequentist settings; our method is consistent with their suggestions but incorporates the added flexibility of BART (Vansteelandt and Goetghebeur, 2003) .

In simulations we showed that semi-BART performs nearly as well as probit and linear regression when the probit or linear model is correctly specified. On the other hand, when there is nonlinearity in the mean functions or many interactions between covariates, using semi-BART provided considerable benefits in our simulations in terms of efficiency (lower MSE for continuous outcomes) or bias (lower bias for binary outcomes). G-estimation is possible for SMMs with continuous outcomes. In our simulations with continuous outcomes, we used doubly robust g-estimation with linear models for both the treatment and the response, which we chose for simplicity and because the model for the treatment was correctly specified which would provide unbiased estimators. In practice, it may often be preferable to use machine learning algorithms to model one or both as one would not know

the true data generating distribution. Another method we examined was Hill's treatment effect on the treated which we used in simulations with continuous outcomes and no effect modifier for treatment (Hill, 2011). In this scenario, using Hill's method is preferable because all covariates including treatment can be modeled together using BART, whereas the semi-BART model utilizes the treatment variable in a separate step from the other covariates. However, the modeling advantage of semi-BART is that it provides a useful alternative to other Bayesian models when low-dimensional summaries of effect modification are of interest. Furthermore, when in the settings of SMMs with binary outcome, semi-BART is an alternative to other models, as g-estimation is not possible.

Some limitations of semi-BART are that it currently does not accommodate instrumental variables or longitudinal treatment measures. Furthermore, its Bayesian implementation makes handling issues such as censoring bias using inverse probability weights difficult (Robins, Hernán, and Wasserman, 2015). As seen in simulations, semi-BART performs best at higher sample sizes, though we found reasonable results at $n = 250$ with 25 covariates. Semi-BART is currently being extended to handle the logit link.

Figure 2.1: Effect of mtNRTIs on death using semi-BART on cohort of individuals with HIV-HCV coinfection newly initiating HAART.



Results of data application using semi-BART on cohort of individuals with HIV-HCV coinfection newly initiating HAART. $A = 1$ indicates receipt of HAART with a mtNRTI and $A = 0$ indicates receipt of HAART without a mtNRTI. The x-axis shows possible mean values for $E(Y^0|\mathbf{X}, A = 1)$ which indicates the mean probability of death if the treated $A = 1$ had in fact been untreated $A = 0$ given $\mathbf{X}$. This quantity is unknown so we consider a spectrum of reasonable values. The y-axis $E(Y^1|\mathbf{X}, A = 1)$ gives the effect of treatment A on the quantity given by the x-axis. No causal effect of A would be indicated by a line with slope 1 through the origin. (a) In this analysis, we only consider that effect of mtNRTI (A) on death (Y) with no effect modifiers. The figure shows that if $E(Y^0|\mathbf{X}, A = 1) = 0.20$ then $E(Y^1|\mathbf{X}, A = 1) = 0.24$, providing evidence that treatment A is harmful. The magnitude of the causal effect of A on Y is determine in part by the assumed value of $E(Y^0|\mathbf{X}, A = 1)$. (b) We consider the effect modification of mtNRTI on death by continuous FIB-4. The solid line indicates the causal effect curve when FIB-4 = 3.25 (the value we center FIB-4 around). At this value, assuming the base probability of death is 20%, that is $E(Y^0|\mathbf{X} = \mathbf{x}, A = 1) = 0.20$, we find that treatment increases this risk to 25%. However, the mean risk of death for individuals with a higher FIB-4 of 5.25 (indicated by the dashed line) is even higher at 30%. The dotted-dashed line shows the mean for FIB-4 = 4.25 and is inbetween the other two estimates. c) We consider the effect modification of mtNRTI on death by a dichotomized FIB-4.The solid line indicates the causal effect curve when FIB-4 ≤ 3.25. Assuming the base probability of death is 20%, that is $E(Y^0|\mathbf{X} = \mathbf{x}, A = 1) = 0.20$, we find that treatment increases the mean risk to 22%. However, the mean risk of death for individuals with high FIB-4 > 3.25 (indicated by the dashed line) is even higher at 35%.

CHAPTER 3

OUTCOME IDENTIFICATION IN ELECTRONIC HEALTH RECORDS USING PREDICTIONS FROM AN ENRICHED DIRICHLET PROCESS MIXTURE

Introduction

Electronic health records (EHR), now a critical component of health care, make a large quantity of data available for researchers. Challenges in using EHR for statistical analyses, however, are well-documented (Sciences, Engineering, and Medicine, 2017). The focus of this paper is on the challenge of outcome identification. Many diseases can be identified in the data from diagnostic codes. However, this is unlikely to fully capture outcomes. EHR data often contain longitudinal measures from laboratory tests (labs) which can be used for the diagnosis of diseases and for disease monitoring. In practice, labs are sometime used to identify additional outcomes (beyond those identified from diagnostic codes). For instance, subjects at risk for diabetes can have fasting glucose labs monitored over time, which can be instrumental in diagnosing the disease (Association, 2014). From a statistical perspective, one challenge is that labs may be abundant for some subjects and sparse or missing for others. Unlike in planned observational studies with primary data collection, labs are not necessarily observed at ideal times. Correspondingly, it may be helpful to model these labs and to use this model to make predictions at time points of interest for EHR containing missing or sparse data. To this end, we propose a flexible joint model for the distribution of a continuous longitudinal outcome (lab values) and baseline covariates. The parameters from the joint model are all given a Dirichlet process (DP) prior with the enrichment proposed in Wade, Mongelluzzo, and Petrone, (2011). Our model provides a flexible framework for prediction as well as serving as a functional clustering algorithm in which one does not specify the number of clusters *a priori*.

The Dirichlet process (DP) mixture is a popular Bayesian nonparametric (BNP) model (Escobar and West, 1995; Ferguson, 1973, 1983) found in many applications, including topic modeling (Teh et al., 2004), survival analysis (Hanson and Johnson, 2004), regression (Hannah, Blei, and Powell, 2011), classification (Cruz-Mesía, Quintana, and Müller, 2007), and causal inference (Roy et al., 2017). Consider the regression setting of Shahbaba and Neal, (2009) and Hannah, Blei, and

Powell, (2011), where there is an outcome Y which we would like to regress on covariates X. In a Bayesian generalized linear model (GLM) setup (McCullagh, 1984), the predictors X are restricted to be a linear combination of the unknown regression parameters. Because of this, GLMs are not appropriate to model nonlinear response curves when the regression coefficients are given normally distributed priors (Gelman et al., 2014). In contrast, placing a DP prior on the regression coefficients (DP-GLM) instead of a parametric prior allows for nonlinearities despite the underlying GLM framework, and this flexibility can often be achieved with only modest additional computational burden. The power of the DP prior stems in part from its partitioning properties (Müller et al., 2015), where it clusters observations and fits local regressions among subjects with similar relationships between covariates and the outcome (Hannah, Blei, and Powell, 2011).

Wade, Mongelluzzo, and Petrone, 2011 showed that with a high number of covariates X, the likelihood contribution of X can dominate the posterior of the partition so that clusters form based more on similarity of covariates than on regression parameters. This leads to a high number of clusters with few observations per cluster and can result in poor predictive performance that can be improved by using an enriched DP (EDP) mixture instead of a DP mixture (Wade et al., 2014). The EDP mixture allows for nested clustering, where one can have clusters based solely on the regression coefficients governing Y on X and within those, nested clusters based on similarity in the covariate space. The benefits of the EDP mixture were demonstrated in simulation and in a real data analysis (Wade et al., 2014).

In this paper, we extend the EDP mixture model to longitudinal settings with a continuous outcome. Some alternatives to our EDP approach to longitudinal data have been proposed in the literature. Müller and Rosner, (1997) modeled blood concentrations in a pharmacokinetic study using DP mixtures with a DP prior on the covariate parameters and the regression coefficients. Li, Lin, and Müller, (2010) developed a flexible semiparametric mixed model with smoothing splines and a DP prior on the random effects with a uniform shrinkage prior for its hyperparameters. Das et al., (2013) fit a bivariate longitudinal model for sparse data with penalized splines for the effect of time and DP priors on the random effects. Quintana et al., (2016) developed a longitudinal model with random effects and a Gaussian process with DP mixtures on covariance parameters of the Gaussian process. This allows for flexible modeling of the correlation structure. Bigelow and Dunson, (2009) fit a joint model for a binary outcome and functional predictor where the functional

predictor was modeled with cubic B-splines whose basis coefficients were given a DP prior. Scarpa and Dunson, (2014) developed an enriched (unrelated to the enriched DP) stick-breaking process which incorporated curve features to better fit functional data.

Our model is unique in that the regression parameters and the parameters for the covariates are given an EDP prior rather than the usual DP prior. As a result, the partitions are not dominated by the covariates as may otherwise happen. Along with improved prediction over DP priors, our model serves as a functional clustering algorithm in which subjects with similar trajectories over time are likely to be part of the same cluster. This aspect also benefits from the EDP prior as functions cluster separately on the regression parameters and covariates. Functional clustering can illuminate distinct patterns among different groups of subjects. A review of functional clustering can be found in Jacques and Preda, (2014). Notably, frequentist and parametric Bayesian methods often require prior specification of the number of clusters, often chosen through model fit statistics. Our EDP model requires no such specification; new clusters may form and existing clusters may vanish throughout the Markov Chain Monte Carlo (MCMC) algorithm.

Our motivating example is a study of individuals who newly initiate a second-generation antipyschotic (SGA). SGAs are known to increase incidence of diabetes (De Hert et al., 2012; Newcomer, 2005). A previous analysis explored the value of incorporating elevated laboratory test results as part of the definition of the outcome of incident diabetes, defined by diagnosis codes and dispensing claims of antidiabetics (Flory et al., 2017). However, many subjects had no recorded lab values or had them measured outside the narrow study window. In this paper, we demonstrate our model by regressing each of three lab values indicative of diabetes (hemoglobin A1c, fasting glucose, and random glucose) on baseline covariates and time. Throughout the MCMC algorithm, values are predicted for each subject at the end of the individual's follow-up, either at one year post SGA initiation or earlier if censored prior to that time. We then combine each set of predictions with the observed data so that each draw can be thought of as an imputed lab values. We then calculate the incidence of diabetes using a multiple imputation procedure (Rubin, 2004). Lastly, we demonstrate how our model can be used for functional clustering by examining posterior clustering patterns resulting from the model.

The rest of the paper is organized as follows. In section 2, we write out the details of our model and describe key components. In section 3, we discuss computations and making predictions from our

model. In section 4, we test our method on simulated datasets. In section 5, we apply our method to the SGA dataset. We discuss the paper in section 6 including limitations and future directions.

Model

To motivate our model, first consider a hypothetical planned observational study, where the outcome of interest is diabetes status one year following initiation of an SGA. In that hypothetical study, we would collect laboratory data, such as hemoglobin A1c (HbA1c) at the end of the study. We might then classify people as having the outcome if their HbA1c value was $\geq$ 6.5%.

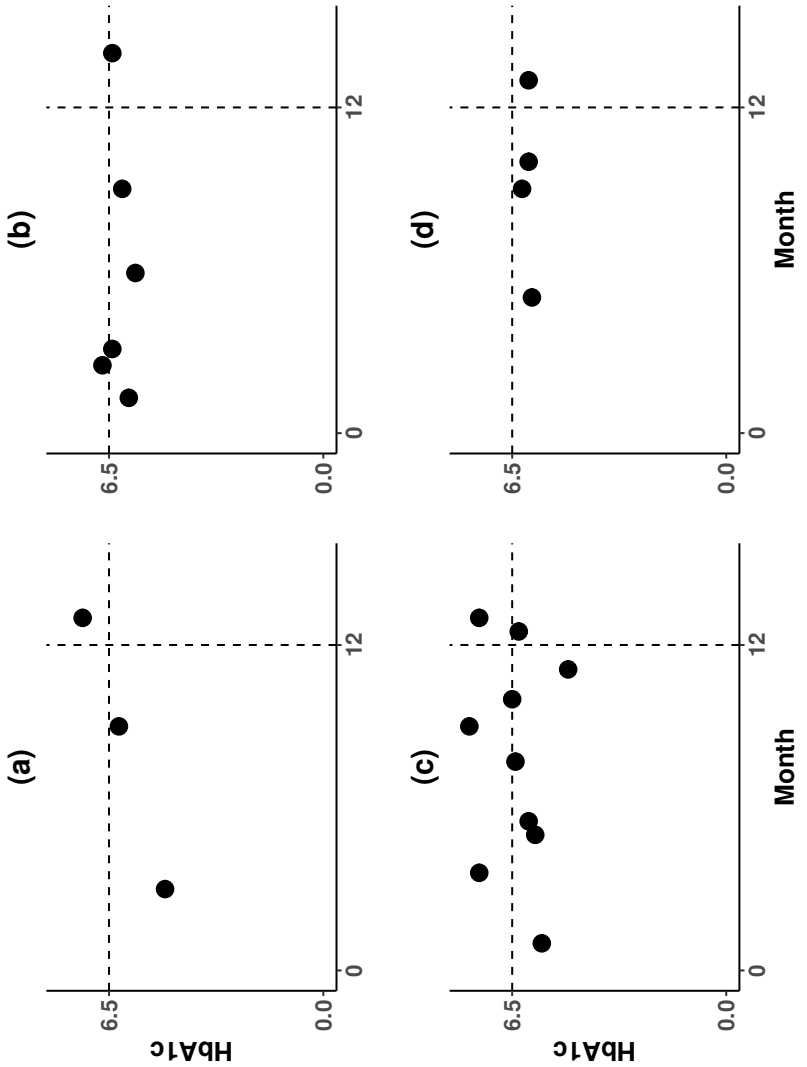
Now consider a study with the same goals, but using EHR data. Figure 3.1 shows four hypothetical subjects with longitudinal measurements of HbA1c over a period of about 15 months. We are interested in determining whether HbA1c levels are $\geq$ 6.5% at month 12. However, none of the four subjects have data collected precisely at month 12, so we need to interpolate from observed data to classify them as elevated or not at month 12. How we classify them is dependent on the algorithm used. Naive algorithms might include basing classification on the value closest to month 12, on the value closest to month 12 that is prior to month 12, or on the maximum value prior to month 12. For instance, it is clear that subject (a) can be classified as either elevated or not depending on the algorithm implemented. Subject (b) has many observations but only one is above the critical threshold and the overall trend suggests their value at month 12 would not elevated. The data for subject (c) has highly variable data and it is uncertain what their month 12 value would be. All subjects have varying degrees of uncertainty in their classifications. These naive classification methods do not use all of the data and do not account for uncertainty in the prediction/imputation.

Our BNP model, described below, was designed to impute outcomes at any time or times of interest, while fully utilizing all of the data (covariates and labs over time). It uses all available data and predicts the outcome at unobserved time points periodically throughout the MCMC algorithm. Thus, for each subject we estimate the distribution of the outcome at the time point of interest rather than just a single prediction.

Notation

Let y_{ij} denote the j^{th} occurrence ($1 \leq j \leq n_i$) of a continuous outcome for subject i, $i \in [1, \ldots, n]$, observed at time t_{ij}. Let $\mathbf{y}$ denote the vector of outcomes for all subjects and $\mathbf{y}_i$ denote the vector

Figure 3.1: Hypothetical example of data from electronic health records.



The four panels represent the hemoglobin A1c (HbA1c) values for four subjects over a 12+ month period. The vertical dashed line at month 12 indicates the time point of interest. The horizontal dashed line represents the critical value (6.5%) of HbA1c above which or equal to indicates diabetes. The cross marks indicate the observed values for each subject. None of the four subjects have values taken precisely at month 12 so interpolation is necessary. Panel (a) shows a subject who has a rising trajectory but doesn't cross the threshold until after month 12. Panel (b) shows a subject who crosses the threshold once prior to month 12 but is stable below the threshold for many other observations. Panel (c) shows a subject with highly variable data around the threshold before and after month 12. Panel (d) shows a subject below the threshold for all observed data.

of outcomes for the i^{th} subject. Let $\mathbf{t}_i$ denote the vector of time points at which $\mathbf{y}_i$ were recorded so that both $\mathbf{t}_i$ and $\mathbf{y}_i$ are length n_i. The covariates for subject i are measured at baseline and denoted by the p-dimensional vector $\mathbf{x}_i$. Without loss of generality, let the first p_1 values $\mathbf{x}_i$ be binary and

the remaining p_2 be continuous with $p = p_1 + p_2$. Let n denote the total number of subjects and N denote the total number of observations, accounting for multiple observations per subject.

We model the distribution the outcome y_{ij} as a function of covariates $\mathbf{x}_i$ and time t_{ij} jointly with the marginal distributions of $\mathbf{x}_i$. To allow for nonlinearities across time, we use splines with k pre-specified knots at $(q_1, \ldots, q_k)$ with $q_1 \leq \cdots \leq q_k$. Bigelow and Dunson, (2009) considered B-splines (Hastie and Tibshirani, 1990) and Li, Lin, and Müller, (2010) used P-splines. We opt for penalized, thin plate splines which have good mixing properties in Bayesian analysis (Crainiceanu, Ruppert, and Wand, 2005). The choice of penalized splines also allows us to choose a large number of knots, reducing the dependency of the model fit on the selection of knot locations. However, any number of basis expansions are possible, including wavelets (Ray and Mallick, 2006). For thin plate splines, let $\mathbf{Z}$ denote the N by k matrix with each row corresponding to the basis functions evaluated at each observed time point t. The matrix $\mathbf{Z}$ is calculated as $\mathbf{Z} = \mathbf{Z}_k \Omega_k^{-1/2}$, where the rows of $\mathbf{Z}_k$ are equal to $\left\{|t_{ij} - q_1|^3, \ldots, |t_{ij} - q_k|^3\right\}$ and the penalty matrix Ω_k is a $k \times k$ matrix where the (l, m)th entry is $|q_l - q_m|^3$ (Crainiceanu, Ruppert, and Wand, 2005). The penalty matrix prevents overfitting by penalizing the coefficients of Z_k. Each subject i in the sample contains a n_i by k submatrix $\mathbf{z}_i$ of $\mathbf{Z}$ which corresponds to the basis functions evaluated at each t_{ij}.

We fit the model

$$\begin{aligned}
\mathbf{y_i}|\mathbf{x_i}, \mathbf{t_i}, \boldsymbol{\beta_i}, \boldsymbol{\eta_i}, u_i, \sigma_i^2 &\sim \mathsf{N}(\mathbf{x_i^*}\boldsymbol{\beta_i} + \mathbf{z_i}\boldsymbol{\eta_i} + u_i, \sigma_i^2\mathbf{I}), && (3.1)\\
x_{ij}|\boldsymbol{\psi_i} &\sim \mathsf{N}(\mu_{ij}, \sigma^2_{\mu,ij}) \text{ (for continuous covariates)}; && (3.2)\\
x_{ij}|\boldsymbol{\psi_i} &\sim \text{Bernoulli}(p_{ij}) \text{ (for binary covariates)}; && (3.3)\\
u_i &\sim \mathsf{N}(0, \sigma_u^2);\\
(\boldsymbol{\theta_i}, \boldsymbol{\psi_i})|P &\sim P;\\
P &\sim \mathsf{EDP}(\alpha_\theta, \alpha_\psi, P_0);\\
\sigma_u^2, \alpha_\theta, \alpha_\psi &\sim \text{Inv-}\mathsf{Ga}(a_u, b_u) \times \mathsf{Ga}(a_\theta, b_\theta) \times \mathsf{Ga}(a_\psi, b_\psi);
\end{aligned}$$

where $\boldsymbol{\theta_i} = (\boldsymbol{\beta_i}, \sigma^2_{\beta,i}, \boldsymbol{\eta_i}, \sigma^2_{\eta,i}, \sigma_i^2)$ are the regression parameters. The notation $EDP(\alpha_\theta, \alpha_\psi, P_0)$ means that $P_\theta \sim DP(\alpha_\theta, P_{0\theta})$ and $P_{\psi|\theta} \sim DP(\alpha_\psi, P_{0\psi|\theta})$, where α_θ and α_ψ are positive valued parameters and $P_0 = P_{0\theta} \times P_{0\psi|\theta}$ is the base distribution with parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ independent.

Here,

$$P_{0\theta} \sim \underbrace{\text{Inv-Ga}(a_\beta, b_\beta)}_{\sigma^2_\beta} \times \underbrace{\text{N}(\boldsymbol{\beta_0}, \sigma^2_{\beta,i}\mathbf{I})}_{\beta} \times \underbrace{\text{Inv-Ga}(a_\eta, b_\eta)}_{\sigma^2_\eta} \times \underbrace{\text{N}(0, \sigma^2_{\eta,i}\mathbf{I})}_{\eta} \times \underbrace{\text{Inv-Ga}(a_y, b_y)}_{\sigma^2};$$

and

$$P_{0\psi|\theta} \sim \prod_{i=1}^{p_1} \text{Beta}(a_x, b_x) \times \prod_{i=p+1}^{p_1+p_2} \text{scaled Inv-}\chi^2(\nu_0, \tau_0^2) \times \text{N}(\mu_0, \tau^2/c),$$

where the first product is among binary covariates followed by a product over the continuous covariates. The notation $\mathbf{x}_i^*$ indicates the vector $\mathbf{x}_i$ with time t_{ij} possibly added, as would be the case if splines were omitted.

We assume that continuous variables are (locally) normally distributed and that binary predictors are Bernoulli. Other distributions can be used, but these distributions are convenient for their conjugacy properties. The parameter $\psi_{i,j}$ corresponds to the two dimensional parameter with mean μ_{ij} and variance $\sigma^2_{\mu,ij}$ if the j^{th} covariate is continuous or the one dimensional probability parameter p_{ij} if the j^{th} covariate is binary. Integrating out the subject specific parameters $\boldsymbol{\psi}_i$ and $\boldsymbol{\theta}_i$ as in Wade et al., (2014), our model can be thought as a countable mixture of linear mixed models where each subject is assigned to one of the mixture components.

We do not posit any *a priori* relationship between time and the outcome. In some applications where the overall trend may be known (for example, the amount of medication in blood may decrease over time after a drug is administered in a pharmacokinetic study), we may posit a model for equation (4.1) incorporating such knowledge, as in Müller and Rosner, (1997) which assumed a piecewise linear structure.

Clustering

A consequence of using the EDP prior on the regression coefficients is that subjects cluster based on their regression parameters $\boldsymbol{\theta}_i$ (that is, for some $i \neq j$, $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$), and within these clusters, will form sub-clusters based on their covariate parameters $\boldsymbol{\psi}_i$. Since $\boldsymbol{\theta}_i$ includes $\boldsymbol{\eta}_i$, the coefficients on the spline basis functions for time, subjects with similar trajectories of their outcomes over time will likely be assigned the same cluster. However, subjects are also clustered by the parameter σ_i^2, which governs variability of outcomes. Thus, it is possible to have clusters with small variability

that follow a precise trajectory over time, and it is possible to have clusters whose large variability defines the cluster, or some combination of the two. The total number of clusters depend on the data and the parameters α_θ and α_ψ, where values closer to 0 indicate fewer clusters.

For this paper, we use the term θ-cluster to indicate clusters based on the parameters $\boldsymbol{\theta}$. A ψ-cluster denotes a cluster nested within a θ-cluster and indicates closeness in the covariate space governed by covariate parameters $\boldsymbol{\psi}$. The ψ-clusters are only meaningful with respect to the θ-cluster in which it is nested.

While an advantage of the BNP appraoch is not having to select the number of clusters, this creates added difficulty in summarizing the clusters. We use the strategy employed in Medvedovic and Sivaganesan, (2002), which employed a distance metric based off of empirical pairwise probabilities of subjects being in the same cluster. To do this, we create a $n \times n$ matrix where each element indicates the number of times two corresponding subjects were in the same θ-cluster over all post burn-in MCMC iterations. From the rows of this matrix, we compute a distance matrix using the supremum norm. We then use Ward's hierarchical agglomerative clustering method implemented by Murtagh and Legendre, (2014). This last step requires choosing a number of clusters, which we choose from the median of the posterior distribution on the number of θ-clusters. R code for this calculation is provided in the Appendix B.4.

Computations

Draws from the posterior distribution of all parameters are obtained through Gibbs sampling. We use an extension of algorithm 8 by Neal (2000) (Neal, 2000) accommodating the nested partitioning of the EDP (Wade et al., 2014) and repeated measurements. Algorithm 8 involves generating m sets of auxiliary parameters corresponding to m clusters that currently have no members. Broadly, at each iteration we alternate between updating cluster membership for each subject, and then within each cluster we update the parameters $(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)$. Let $s_i = (s_{i,y}, s_{i,x})$ denote the cluster membership for the i^{th} subject, where $s_{i,y}$ denotes the θ-cluster corresponding to $\boldsymbol{\theta}_i$ and $s_{i,x}$ denotes the ψ-cluster nested within $s_{i,y}$ corresponding to $\boldsymbol{\psi}_i$. Let $\boldsymbol{\theta}^*_k$ denote the value of $\boldsymbol{\theta}$ corresponding to the k^{th} unique value of $s_{i,y}$. Similarly, let $\boldsymbol{\psi}^*_{j|k}$ denote the value of $\boldsymbol{\psi}$ corresponding to the j^{th} unique value of $s_{i,x}$ within the k^{th} unique value of $s_{i,y}$. Note that if $s_{i,y} = s_{j,y}$, then $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$. Furthermore, let n^θ_k denote the number of subjects in the k^{th} unique cluster of $s_{i,y}$ and $n^\psi_{j|k}$ denote the number

of subjects in the j^{th} unique cluster of $s_{i,x}$ nested within the k^{th} unique value of $s_{i,y}$. The notation $n_k^{-i,\theta}$ and $n_{j|k}^{-i,\psi}$ denote the size of the clusters with the i^{th} subject removed. Recall that the similar notation with no superscript, n_i, refers to the number of observations for the i^{th} individual.

The first step of our algorithm updates the value of s_i for every individual. First, remove individual i from their current cluster. The probability that an individual is in any given cluster depends on the current values of α_θ and α_ψ, the number of subjects within that cluster, the values of $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$ as well as the observed data. In choosing clusters, there are three possibilities: subjects can be assigned to an existing ψ-cluster within an existing θ-cluster, a new ψ-cluster within an existing θ-cluster, or a new θ-cluster and a new ψ-cluster. An individual is assigned to an existing cluster (k, j) with probability proportional to:

$$\frac{n_k^{-i,\theta} n_{j|k}^{-i,\psi}}{(n_k^{-i,\theta} + \alpha_\psi)(\alpha_\theta + n - 1)} \times \prod_{v=1}^{n_i} f_y(y_{i,v}; \boldsymbol{x}_i, \boldsymbol{\theta}_k^*) \times \prod_{l=1}^{p} f_{x,l}(x_{i,l}; \boldsymbol{\psi}_{j|k}^*).$$

An individual is assigned to a new ψ-cluster within the k^{th} existing θ-cluster with probability proportional to:

$$\frac{n_k^{-i,\theta} \alpha_\psi / m}{(n_k^{-i,\theta} + \alpha_\psi)(\alpha_\theta + n - 1)} \times \prod_{v=1}^{n_i} f_y(y_{i,v}; \boldsymbol{x}_i, \boldsymbol{\theta}_k^*) \times \prod_{l=1}^{p} f_{x,l}(x_{i,l}; \boldsymbol{\psi}_0^*).$$

An individual is assigned a new θ-cluster and a new ψ-cluster with probability proportional to:

$$\frac{\alpha_\theta / m}{\alpha_\theta + n - 1} \times \prod_{v=1}^{n_i} f_y(y_{i,v}; \boldsymbol{x}_i, \boldsymbol{\theta}_0^*) \times \prod_{l=1}^{p} f_{x,l}(x_{i,l}; \boldsymbol{\psi}_0^*).$$

These probabilities are then normalized to sum to 1.

The notation $\boldsymbol{\psi}_0^*$ and $\boldsymbol{\theta}_0^*$ refers to parameters from a cluster that currently has no members (also called auxiliary parameters, see (Neal, 2000)). They are generated randomly from the prior base distributions $P_{0\psi|\theta}$ and $P_{0\theta}$ for $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$. The notation $f_{x,l}(\cdot; \boldsymbol{\psi})$ corresponds to the normal density in equation (4.2) or the binomial density in equation (4.3) for continuous and binary, respectively, and $f_y(\cdot; \boldsymbol{x}_i, \boldsymbol{\theta})$ corresponds to the normal density from equation (4.1) evaluated with parameters $\boldsymbol{\theta}$. Once we calculate these probabilities, we draw cluster membership using a random multinomial distribution. This is done separately for each individual in the cohort.

Once cluster memberships for all individuals have been updated, the within cluster parameters $\boldsymbol{\theta}^*$

and $\boldsymbol{\psi}^*$ are updated. To update the regression parameters $\boldsymbol{\theta}_k^*$ for the k^{th} cluster, we consider only individuals with $s_{i,y} = k$. First, we update the regression variance σ_k^{2*} using a conjugate draw from an inverse gamma distribution and then update regression parameters $\boldsymbol{\beta}_k^*$ for covariates $\mathbf{x}$ from a draw with a multivariate normal distribution. Next, update the variance for the spline effects $\sigma_{b,k}^{2*}$ from a random draw from an inverse gamma distribution. Lastly, we update the coefficients $\boldsymbol{\eta}_k^*$ for the spline effects from a draw from a multivariate normal distribution. In essence, within each cluster we are fitting separate Bayesian mixed effects models and updating parameters accordingly (Zeger and Karim, 1991). Full posterior distributions for updating $\boldsymbol{\theta}^*$ are in Appendix B.2.

Next, we update covariate parameters $\boldsymbol{\psi}^*$. To update $\boldsymbol{\psi}_{j|k}^*$, we take subjects with $s_i = (k, j)$. If the l^{th} covariate is binary, then the distribution of x_l is assumed Bernoulli and the parameter ψ_l is updated from a Beta distribution with parameters $a_n = \sum_{s=(k,j)} x_{i,l} + a_x$ and $b_n = n_{j|k} - \sum_{s=(k,j)} x_{i,l} + b_x$. If the l^{th} covariate is continuous then the distribution of x_l is normal and the parameters $\psi_l = (\sigma_l^2, \mu_l)$ are updated from conjugate inverse-χ^2 and normal distributions, available in Appendix B.2.

It remains to update the random intercepts u_i, the variance σ_u^2, α_ψ, and α_θ. The new random intercepts u_i are calculated after taking the residuals from the current fit given covariates $\mathbf{x}_i$ and the rest of the current parameter values. The variance σ_u^2 is updated through a random draw from an inverse gamma distribution with shape $a_u + n$ and rate $b_u + \frac{\mathbf{u}^\intercal \mathbf{u}}{2}$. Finally, we update α_ψ and α_θ. α_θ is updated by generating a random value from a mixture of two gamma posteriors as in Escobar and West, (1995). α_ψ is updated through a Metropolis-Hastings step. The updates for these α parameters are equivalent to those in Roy et al., (2017) who also employ an EDP mixture model. Consult Appendix B.2 for expanded details of the MCMC algorithm.

Predictions

Predicting values for subjects who have observed data (that is, data at time points other than the time point of interest) is straightforward. At every iteration where we seek to make a prediction, each subject is assigned a cluster $s_{i,y}$ with corresponding $\boldsymbol{\theta}_i$. From this, we can predict from a single draw from a normal distribution given $\mathbf{x}_\mathbf{i}^*, \boldsymbol{\beta_i}, \mathbf{z_i}, \boldsymbol{\eta_i}, u_i, \sigma_i^2$ with mean $\mathbf{x}_\mathbf{i}^* \boldsymbol{\beta_i} + \mathbf{z_i} \boldsymbol{\eta_i} + u_i$ and variance σ^2.

For subjects missing outcome data, we must make predictions from their covariates $\mathbf{x_i}$. These sub-

jects may be part of the k^{th} existing cluster with parameters $\boldsymbol{\theta}_k^*$ or may be in an entirely new cluster. If they are part of the k^{th} existing θ-cluster, we use the current values from the corresponding parameters for that cluster (i.e., $\boldsymbol{\theta}_k^*$) and draw the prediction from a normally distribution with mean $\mathbf{x_i^*}\boldsymbol{\beta_k^*} + \mathbf{z_i}\boldsymbol{\eta_k^*}$ and variance σ_k^{2*}. If a subject is part of a new cluster that currently has no members, we generate $\boldsymbol{\theta}_i$ using the base distribution $P_{0\theta}$.

The probability that a subject is in the k^{th} existing θ-cluster is proportional to:

$$\frac{n_k^\theta}{\alpha_\theta + n} \times \left[\frac{\alpha_\psi}{\alpha_\psi + n_k^\theta} f_{x,0}(\mathbf{x}_i) + \sum_j \left(\frac{n_{j|k}^\psi}{\alpha_\psi + n_k^\theta} \prod_{l=1}^{p} f_{x,l}(x_{i,l}; \boldsymbol{\psi}_{j|k}^*) \right) \right],$$

where the summation iterates through all nested ψ-clusters for the k^{th} θ-cluster.

The probability that a subject is in a new θ-cluster is proportional to:

$$\frac{\alpha_\theta}{\alpha_\theta + n} \times f_{x,0}(\mathbf{x}_i),$$

where $f_{x,0}(\mathbf{x}_i) = \prod_{l=1}^{p} \int_\psi f_{x,l}(x_{i,l}) dP_{0\psi|\theta}$, the density integrated over the base measure evaluated at the observed data (Wade et al., 2014). This computation for binary and continuous covariates using our distributional and prior assumptions is shown in Appendix B.5. Since we used conjugate priors, this integration can be done analytically. When non-conjugate priors are used, Monte Carlo integration is an option.

Simulations

We used simulation to assess the predictive performance of our longitudinal model with splines and an EDP prior. For each simulated subject, we predicted the outcome at a specific time and compared it to the true value. Let $y_{i,t}$ be the i^{th} subject's true value at time t and let $\hat{y}_{i,t}$ be the prediction of $y_{i,t}$ from a given model. We computed the mean absolute prediction error L_1 and the mean squared prediction error L_2 over all simulated subjects.

$$\ell_1 = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_{i,t} - y_{i,t}|$$
$$\ell_2 = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_{i,t} - y_{i,t})^2$$

We simulated sample sizes of $n = 1000$ and $n = 5000$. Each individual was randomly assigned a minimum of 1 and a maximum of 5 repeated measurements corresponding to time points within the interval $t \in [0, 1]$ generated randomly from an independent uniform distribution. As before, let θ denote the regression parameters and ψ denote the covariate parameters. The true cluster structure had three θ-clusters. Within each θ-cluster, there were 3, 2, and 3 nested ψ-clusters. Thus, the total number of unique clusters was 8 while the total number of unique θ-clusters was 3. The structure of the clustering along with probabilities of being in each cluster are given in Figure 3.2. Each subject was assigned 20 simulated covariates from distributions whose parameters differed between ψ-clusters. Full data-generating details are available in Appendix B.1 and code is available upon request (code for the EDP and DP models are at https://www.github.com/zeldow/EDPlong and https://www.github.com/zeldow/DPlong).

Predictions were made for each subject at $t = 0.75$ and the true value $y_{i,t}$ was calculated based on the mean for the θ-cluster to which the individual belongs (mean function shown in Appendix B.1) and the random intercept. We generated 100 datasets and take the mean of ℓ_1 and ℓ_2 over all simulations, and then calculate

$$\bar{\ell}_1 = \frac{1}{100}\sum_{j=1}^{100} \ell_{1j}$$
$$\bar{\ell}_2 = \frac{1}{100}\sum_{j=1}^{100} \ell_{2j},$$

where ℓ_{1j} and ℓ_{2j} are ℓ_1 and ℓ_2 calculated on the j^{th} simulated dataset.

To assess the performance of our mixed model with an EDP prior, we compared it to two competitor models: a Bayesian mixed model with a DP prior and a linear mixed model (implemented by the

Table 3.1: Simulation results for $n = 1000$ showing mean L_1 and L_2 errors over 100 datasets for predictions at $t = 0.75$.

	EDP		DP		ME	
	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$
$\sigma^2 = 1; \sigma_u^2 = 0.15$	0.66	0.87	0.89	1.43	1.11	1.85
$\sigma^2 = 1; \sigma_u^2 = 0.5$	0.82	1.19	1.07	1.93	1.11	1.87
$\sigma^2 = 4; \sigma_u^2 = 0.15$	0.89	1.46	1.08	2.00	1.23	2.32
$\sigma^2 = 4; \sigma_u^2 = 0.5$	1.05	1.90	1.17	2.28	1.24	2.37

σ^2 indicates the simulated regression variance and σ_u^2 indicates the simulated random intercept variance. EDP indicates the longitudinal model with an enriched Dirichlet process prior. DP indicates the longitudinal model with a Dirichlet process prior. ME indicates a mixed effects model fit using the lmer package in R. Fit with penalized thin plate splines with 20 knots.

Table 3.2: Simulation results for $n = 5000$ showing mean L_1 and L_2 errors over 100 datasets for predictions at $t = 0.75$.

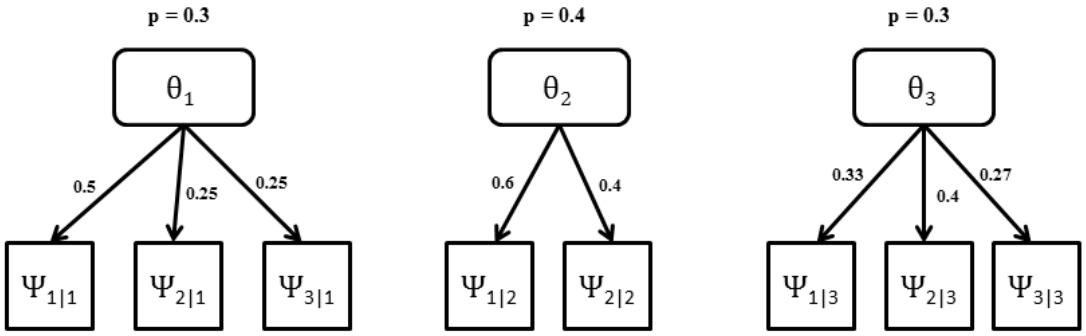
	EDP		DP		ME	
	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$
$\sigma^2 = 1; \sigma_u^2 = 0.15$	0.60	0.73	0.91	1.49	1.10	1.82
$\sigma^2 = 1; \sigma_u^2 = 0.5$	0.77	1.06	1.10	2.03	1.11	1.86
$\sigma^2 = 4; \sigma_u^2 = 0.15$	0.71	0.99	1.04	1.90	1.21	2.27
$\sigma^2 = 4; \sigma_u^2 = 0.5$	0.85	1.27	1.15	2.25	1.23	2.34

σ^2 indicates the simulated regression variance and σ_u^2 indicates the simulated random intercept variance. EDP indicates the longitudinal model with an enriched Dirichlet process prior. DP indicates the longitudinal model with a Dirichlet process prior. ME indicates a mixed effects model fit using the lmer package in R. Fit with penalized thin plate splines with 20 knots.

lme4 package (Bates et al., 2014) in R (R Core Team, 2017)). For each sample size, we varied the regression variance σ^2 and the random intercept variance σ_u^2 resulting in four simulation scenarios: (1) low variability in σ^2 and low variability in σ_u^2; (2) low variability in σ^2 and high variability in σ_u^2; (3) high variability in σ^2 and low variability in σ_u^2; and (4) high variability in σ^2 and high variability in σ_u^2. All models were fit using thin plate splines for the time effect. In Appendix B.3, we show results using cubic B-splines as well with 2 knots at $\frac{1}{3}$ and $\frac{2}{3}$.

The results of the simulation study for $n = 1000$ are shown in Table 3.1. For all scenarios the EDP model outperformed the DP model and the mixed model in terms of mean L_1 and L_2 prediction error. The mean L_1 error for the EDP model ranged from 0.66 to 1.05. For the DP model, it ranged from 0.89 to 1.17, and for the standard mixed model it ranged from 1.11 to 1.24. The mean L_2 errors range from 0.87 to 1.90, 1.43 to 2.28, and 1.85 to 2.37 among the models for the four simulation scenarios, respectively. Given that the data were generated in clusters, it is unsurprising that the

Figure 3.2: Figure of structure of clusters for simulations.



Probabilities for being in a ψ-cluster are conditional on being in the appropriate θ-cluster. θ refers to the regression parameters and ψ refers to the covariate parameters.

two methods implementing clustering provide better predictions than the standard mixed effects model. However, we see that the EDP model yielded more precise prediction than the DP model based on L_1 and L_2 prediction error.

The results in Table 3.2 display the results for the simulations with $n = 5000$. Again, the EDP model outperforms the DP model which outperforms the mixed model. For the most part, there are no large differences in the relative performance of methods at the two sample sizes. Using cubic B-splines (see Appendix B.3) in lieu of thin plate splines also did not have considerable effect on prediction error. Overall the results for thin-plate splines were slightly improved over those of B-splines, but more research needs to be done and more scenarios examined.

Table 3.3: Simulation results for $n = 1000$ showing mean L_1 and L_2 errors over 100 datasets for predictions at $t = 0.75$ when the standard mixed effects model is correctly specified.

	EDP		DP		ME	
	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$
$\sigma^2 = 1; \sigma_u^2 = 0.15$	0.31	0.16	0.31	0.16	0.28	0.12
$\sigma^2 = 1; \sigma_u^2 = 0.5$	0.56	0.50	0.56	0.50	0.38	0.23
$\sigma^2 = 4; \sigma_u^2 = 0.15$	0.34	0.18	0.34	0.18	0.35	0.20
$\sigma^2 = 4; \sigma_u^2 = 0.5$	0.58	0.52	0.58	0.52	0.52	0.42

σ^2 indicates the simulated regression variance and σ_u^2 indicates the simulated random intercept variance. EDP indicates the longitudinal model with an enriched Dirichlet process prior. DP indicates the longitudinal model with a Dirichlet process prior. ME indicates a mixed effects model fit using the lmer package in R. Fit with cubic B-splines with 2 knots.

Lastly, we performed simulations where all subjects were part of the same cluster so that the linear mixed model was correctly specified and was expected to work best. The results for $n = 1000$ are included in Table 3.3. Over 100 simulated datasets, the correctly specified standard mixed model outperformed both the EDP and DP models in almost all scenarios. Results from EDP and DP models showed no difference up to two decimal places. This interesting finding was due to the fact that the EDP model did not split θ-clusters in subclusters, rendering the difference between the DP and EDP models irrelevant. Overall, we found that the EDP and DP models concentrated around one large θ-cluster with scattered observations in other θ-clusters. With $n = 1000$, the L_1 error for the mixed model ranged from 0.28 to 0.52, while for the DP and EDP models, it ranged from 0.31 to 0.58. The largest difference in favor of the standard mixed model occurred with low regression variance $\sigma^2 = 1$ and high random intercept variance $\sigma_u^2 = 0.5$ (L_1 error: 0.38 versus 0.56; L_2 error: 0.23 versus 0.50). Other scenarios showed either no difference or only a modest improvement for the standard mixed model. One possible explanation for this discrepancy is that there is an identifiability problem in the models with DP or EDP priors in which the algorithm has difficulty determining if σ_u^2 is smaller and there are many clusters or if σ_u^2 is large and there are few clusters. Thus, in this scenario the EDP and DP models split the sample into more clusters than was necessary and prediction suffered accordingly. On the other hand, the reverse scenario with high regression variance and low random intercept variance showed no difference between the three models, indicating that the nonparametric prior performed fine in more likely situations.

Data Analysis

Sentinel is an intiative of the US Food and Drug Administration with 19 data partners (*Sentinel*). Under Sentinel, a distributed database has been established that collects EHR and administrative health plan data to assess safety in approved medical products, particularly drugs and vaccines. As part of a workgroup effort to understand and use laboratory results data in the Sentinel Distributed Database (SDD) (*Analytic Methods for Using Laboratory Test Results In Active Database Surveillance*), Flory et al., (2017) used the MSDD to calculate incidence rates of diabetes among new initiators of second generation antipsychotics (SGAs), which are known to increase the risk of Type II diabetes mellitis (T2DM) (De Hert et al., 2012; Newcomer, 2005). T2DM is often diagnosed based on elevated levels of hemoglobin A1c (HbA1c), serum glucose, or capillary glucose (Association, 2014). In Flory et al., (2017) incidence rates for T2DM were computed from two outcomes: (O1)

diagnosis codes and dispensement of antidiabetic medication and (O2) diagnosis codes, dispensement of antidiabetic medication as well as an elevated diabetes labs. Lab values were considered elevated if fasting glucose ≥ 126 mg/dl, random glucose ≥ 200 mg/dl, or HbA1c ≥ 6.5v%. Including diabetes labs increased the number of T2DM cases, but missingness was differential among the sites analyzed, affecting some sites more than others. In this paper, we extend some of the results of Flory et al., (2017) using predictions from our longitudinal EDP model.

We restricted our analysis to site one of Flory et al., (2017), which corresponds to a small integrated delivery system. As in that publication, our cohort was restricted to participants at least 21 years of age who had at least 183 days of health plan enrollment prior to initiating a SGA (aripiprazole, olanzapine, quetiapine, and risperidone). We included those who had first dispensement of a SGA between 1 January 2008 and 31 October 2012. Any individuals with evidence of diabetes prior to initiation of the SGA, including diagnosis of diabetes, receipt of an antidiabetic medication, or an elevated diabetes lab, were excluded. Follow-up began at first dispensement of a SGA and continued until discontinuation of insurance, death, occurrence of the outcome, or end of 365 days, whichever came first. The outcome was incident diabetes within 365 days of study, equal to that of outcome O1 above. We also define a new outcome O3, which consists of O1 and predicted elevated lab values.

The motivation for using our EDP longitudinal model for this problem is as follows. Our interest lies in calculating the incidence of diabetes within one year of initiating a SGA, supplementing the outcome with information from recorded lab values. The previous analysis was limited by restricting to lab values within one year of follow-up. However, lab values after one year can be informative as well, particularly those drawn soon after study end. Over 30% of the subjects from site one did not have any lab values recorded between 1 and 365 days of SGA initiation. Subjects with lab values recorded had differential amounts of data recorded within that study window, ranging from 1 to 4 records for HbA1c, 1 to 5 of fasting glucose, and 1 to 115 of random glucose. Lastly, the approach in Flory et al., (2017) treats any instance of a lab value exceeding the threshold as part of the outcome even if only one measurement among many exceeded the threshold. Because of this, uncertainty stemming from measurement error was inadequately accounted for. Our model incorporates such uncertainty through the regression variance component σ^2 as well as the fact that cluster membership s_y changes throughout the algorithm.

We fit EDP longitudinal models for each of three lab values (HbA1c, fasting glucose, and random glucose) separately. Models were fit with the entire history of the subject's lab values until initiation of an anti-diabetic medication. Our dataset had a total of $n = 3,764$ study participants. Among these, 680 subjects contributed 1,003 observations for HbA1c. For fasting glucose, 2,032 subjects contributed 4,110 observations. For random glucose, 3,013 subjects contributed 21,614 observations. We used 200,000 iterations with 40,000 burn in period. Throughout the 160,000 post burn in iterations, predictions were drawn at 800 evenly spaced iterations. Each subject had predictions made at day 365, unless their study censoring time was prior to that, at which point we made predictions at that censoring time. All predicted values were appended to the original dataset resulting in 800 imputed datasets. Each imputed dataset consists of the original data, including diabetes diagnoses and dispensement of antidiabetics, along with three predicted values for HbA1c, random glucose, and fasting glucose. The outcome O3 was calculated for each imputed dataset. From this, we then calculate the incidence of diabetes and use multiple imputation methods to combine estimates across imputations (Rubin, 2004). Overall, the HbA1c model took 4.7 hours of runtime, the fasting glucose model 22.4 hours, and the random glucose model 63.2 hours.

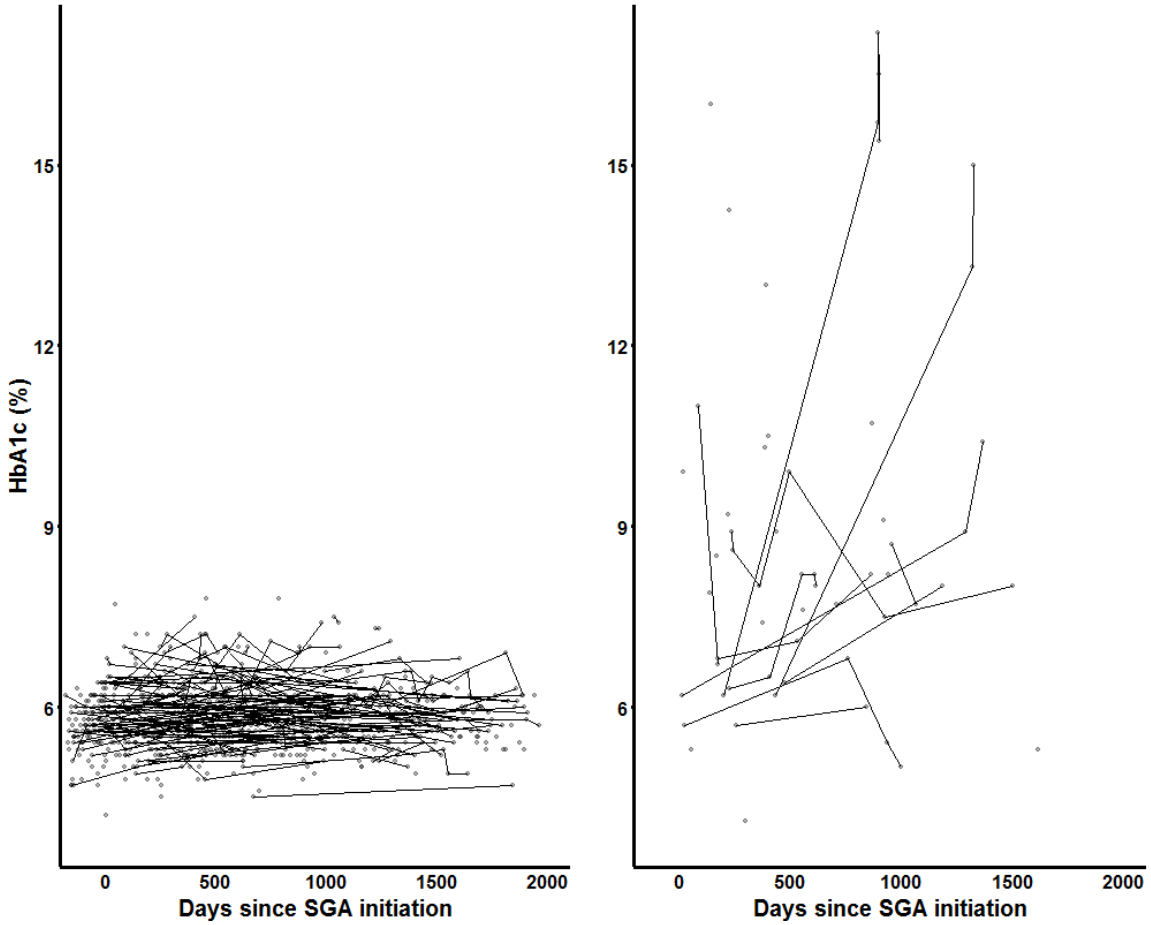
In total, 89 participants were diagnosed with diabetes through diagnosis codes or dispensement of anti-diabetic medication. The total number of outcomes O3 ranged from 146 to 394 outcomes with a median of 200 throughout the 800 imputations. This resulted in an incidence of 0.059 events per person-year (95% confidence interval: 0.043–0.080). This result is similar to the incidence found in Flory et al., (2017) for site one among those with recorded lab values, except the confidence interval is wider, reflecting greater uncertainty in classification using lab values.

Clustering

We also examined clustering resulting from our model. There is a multitude of reasons one may be interested in clustering in the present example. First, it can show heterogeniety (or lack thereof) of outcome features among groups of individuals. The cluster itself may be able to predict outcomes. For example, if we know that a certain individual is in a cluster with rising HbA1c values over time, we know that their likelihood of a diabetes diagnosis is increased compared to a group with flat trajectories over time. Further, once we have identified the clustering structure, we can examine the distributions of covariates within cluster and determine covariates that may be affecting the differences among groups.

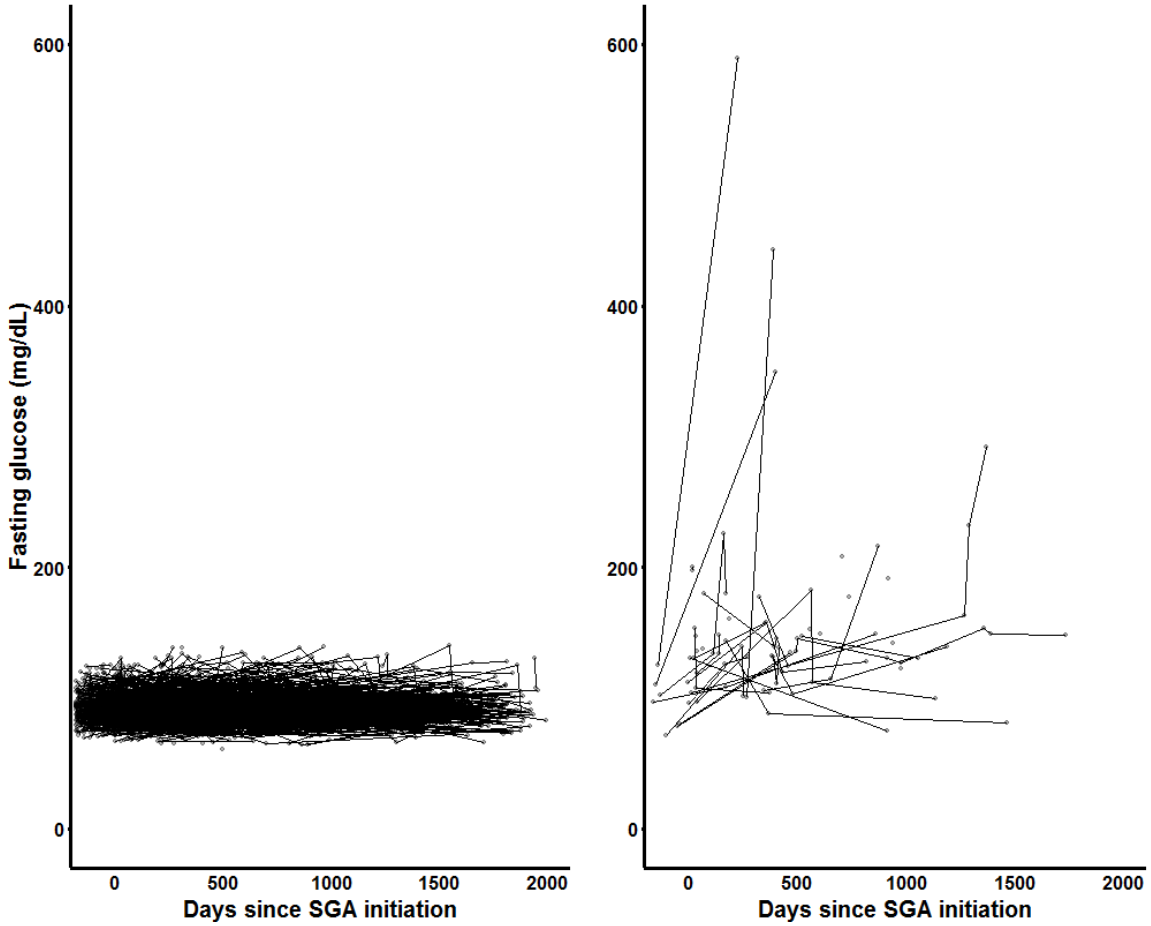
Recall that we refer to clusters based on regression parameters as θ-clusters and the nested clusters based on covariates as ψ-clusters. For illustrative purposes, we focus strictly on functional clustering using θ-clusters. Other applications may have interest in summarizing ψ-clusters as well. Given that within the MCMC algorithm, not only cluster membership but the number of clusters can change, we condense the results into a single point estimate for the posterior cluster structure. For the HbA1c and fasting glucose models, the posterior number of clusters concentrated around two. For random glucose, the posterior number of clusters concentrated around three. All models were initialized to have two θ-clusters. When we initialized the number of θ-clusters to 10, results eventually converged to similar answers for each of the outcomes. However, computation time was considerably longer when initialized with a large number of θ-clusters.

Figure 3.3: Clustering results for HbA1c model.



The model settled on two θ-clusters which are shown in the figure. The larger cluster has 650 subjects and the smaller cluster has 30 subjects.

Figure 3.4: Clustering results for fasting glucose model.

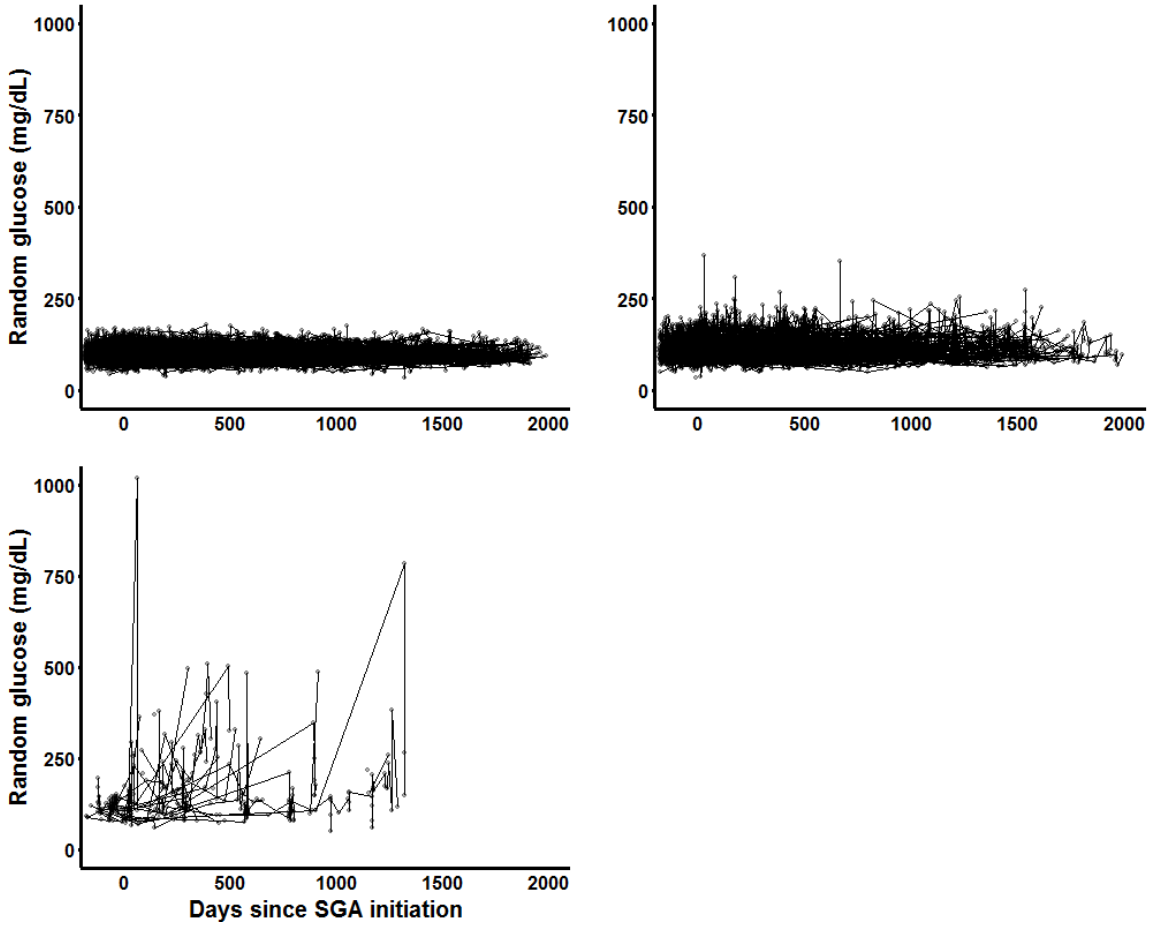


The model settled on two θ-clusters. The larger cluster has 1997 subjects and the smaller cluster has 35 subjects.

For both the model with HbA1c as the outcome and the model with fasting glucose as the outcome, our algorithm settled on two distinct clusters as seen in Figure 3.3 and Figure 3.4, respectively. For HbA1c, the first cluster contained 650 observations consisting of trajectories that mostly stay within the values 5% and 8%. The remaining 30 observations in the second cluster consisted of highly variable trajectories that had spikes in their values. In the fasting glucose model, the first cluster had 1997 members and consisted of tight trajectories below the threshold of 126 mg/dL, while the second cluster housed the remaining 35 subjects mostly of subjects whose trajectory at some point contains a spike or is somehow indicative of higher variability.

The model with random glucose as the outcome settled on three clusters which can be seen in

Figure 3.5: Clustering results for random glucose model.



The model settled on three θ-clusters with 2563, 419, and 31 subjects.

Figure 3.5. The largest cluster had 2563 subjects who had relatively flat trajectories with small within-subject variability. The second largest cluster contained 419 subjects who had trajectories with slightly more variability than the first cluster. The third cluster contains 31 subjects with large spikes and characterized by larger variability than the other clusters.

Discussion

In this paper, we presented a joint model for a continuous longitudinal outcome and the baseline covariates. The model is partitioned into the product of a linear mixed model for the outcome given the covariates and the marginal distributions for the covariates. The use of the EDP prior in a longitudinal model is an extension of the model developed by Wade, Mongelluzzo, and Petrone,

(2011), which itself is an extension of the DP prior. Through the nested clustering of the EDP prior, where subjects are clustered separately for their regression trajectories and similarity in the covariate space, our model allows for improved prediction over the same model with the usual DP prior. This improvement was demonstrated in simulation scenarios in which the EDP longitudinal model outperformed both a standard mixed model and a longitudinal model with a DP prior when the data generating distribution contained a nested clustering structure. When the simulation scenario was simplified so that there was no underlying cluster structure and the linear mixed model was correctly specified, using the nonparametric EDP prior did not excessively diminish predictive performance. Our model also serves as a functional clustering algorithm, the first to use an EDP prior. In our model setup, the EDP prior is particularly useful because it allows the functional to cluster solely on functional features rather than non-functional components (i.e., closeness in the covariate space).

One limitation of the present model is that it can only incorporate baseline covariates. In many longitudinal settings, covariates may be updated throughout the study. One possibility to incorporate this into our model would be to use the dynamic DP, which allows for distributions to evolve in discrete time (Rodriguez and Ter Horst, 2008). From the current state of the literature, DPs which evolve throughout time are less thoroughly developed and more difficult to implement. The extension of our model to handle time-varying covariates is a topic for future research.

Throughout the paper we made several modeling choices that could be changed or generalized. For example, the value for α_ψ could depend on θ so that the mass parameter is written as $\alpha_\psi(\theta)$. This would allow the number of subclusters to differ depending on the value of θ. Further, we made the assumption that the values of ψ and θ were independent through the fact that $P_0 = P_{0\theta} \times P_{0\psi|\theta}$. This assumption simplifies calculations but can be relaxed if needed. These two changes were discussed in Wade, Mongelluzzo, and Petrone, (2011). Lastly, we focused on continuous outcomes, but our methods can be extended to more general settings such as binary or count outcomes or different link functions with various additional computational challenges (non-conjugacy for one). Bayesian computations for generalized linear mixed models are provided in Zhao et al., (2006) and references therein.

We demonstrated our model with data from the Sentinel Distributed Database, where we used predicted lab measurements to augment incidence rates of diabetes among subjects initiating certain anti-psychotics. Our incidence rates were similar to those found in a previous paper on the same

study population (Flory et al., 2017), but our estimates gave wider confidence intervals, reflecting greater uncertainty about the incorporation of labs as part of the diabetes diagnosis. Our model is well-suited for other applications as well, such as with data arising from studies using wearable devices or studies of symptoms of chronic conditions with interest in detecting patterns among patients.

CHAPTER 4

PARAMETRIC G-FORMULA FOR A LONGITUDINALLY RECORDED OUTCOME USING AN ENRICHED DIRICHLET PROCESS PRIOR

Introduction

Randomized trials are the ideal experiment to address questions of efficacy for a treatment or intervention (Hernán and Robins, 2018). Due to a variety of reasons ranging from ethical considerations to time or financial constraints, we often must resort to analyzing data from a non-randomized source instead (Black, 1996). In such cases, we can try to emulate a randomized experiment using observational data (Hernán and Robins, 2016). One source of non-randomized data is electronic health records (EHR), an increasingly used system designed to store medical data with the idea of centralizing individual medical records (Schoen et al., 2012). EHR, often made available for data analysis, pose unique challenges beyond the lack of a randomized intervention. In contrast to randomized trials and planned observational studies, data are not always recorded at ideal times, and certain individuals may have ample data while others have little to none.

In using EHR to emulate a randomized trial, one must define the point of entry in the trial and the time frame across which the outcome is assessed (Hernán and Robins, 2016). We can often find appropriate markers representing study entry such as the first use of a medication, initial entry into the EHR system, or the diagnosis of a condition. Specifying the follow-up period can be a more difficult task since there is no guarantee that outcomes are recorded at the desired times. In the previous chapter, we addressed this problem from a prediction standpoint and proposed a joint model for a continuous outcome and baseline covariates using an enriched Dirichlet process (EDP) prior (Wade, Mongelluzzo, and Petrone, 2011). The joint model can be decomposed into the product of a linear mixed model for the outcome and marginals for the covariates. In this paper, we extend our model to use with the parametric g-formula (Robins, 1986). We demonstrate this in simulation and using EHR in which the outcome may not be measured at the time point of interest.

There have been several recent papers proposing Bayesian methods for causal inference, which has typically been in the domain of classical semiparametrics (Hill, 2011; Roy, Lum, and Daniels,

2016; Roy et al., 2017). Bayesian methodology requires full specification of the likelihood, which can be restrictive in the sense that a large portion of the likelihood, called the nuisance component, is not of scientific interest but is still subject to model mispecification (Gelman et al., 2014). On the other hand, classical semiparametrics allow for the nuisance component to be left unspecified. While a Bayesian approach can never avoid specification of the full likelihood, we can use Bayesian nonparametrics to relax modeling assumptions by introducing an infinite dimensional parameter with an appropriate prior. Typically, this is done with a Dirichlet process mixture (DPM), the most popular Bayesian nonparametric model (Ferguson, 1983). In this paper, the EDP prior that we use is an extension of the common DP prior, allowing for improved prediction when there are many covariates (Wade et al., 2014).

The chapter is organized as follows. In section 2, we describe our model and define causal effects along with identifying assumptions. In section 3, we detail the algorithm for the parametric g-formula within our joint model. In section 4, we use simulation to assess small sample properties. In section 5, we complete a data analysis using new initiators of second generation antipsychotics to assess their causal effect on risk of diabetes. In the last section, we discuss the results and suggest future research directions.

Model

Dirichlet Process

The Dirichlet process (DP) is the most popular Bayesian nonparametric prior (Ferguson, 1973; Müller et al., 2015). A DP $P \sim DP(\alpha P_0)$ is parameterized by a positive-valued mass parameter α and a centering distribution P_0 defined on a sample space S. Each draw from P is itself a probability measure, centered around P_0, meaning the DP is suited to be a prior on the space of distributions. When $\{B, B^c\}$ is a partition of S, we have that $E(P(B)) = P_0(B)$ and that var$(P(B)) = P_0(B)(1 - P_0(B))/(1 + \alpha)$ (Müller et al., 2015). From the variance relation, it is clear that the parameter α controls the variability of P around P_0, where high values of α imply smaller variability and less deviation from P_0. Two additional properties of DPs are that P has the same support S as P_0, and that draws from P are almost surely discrete, even if P_0 is continuous. An example of draws from a DP is displayed in Figure 1.1.

In density estimation problems, the probability model can be placed directly on the unknown density using a DP, but the DP is more frequently used as a mixing distribution for the parameters of a parametric kernel. These mixture models are called DP mixtures (DPMs) (Ferguson, 1983). An important example of a DPM can be found in Hannah, Blei, and Powell, (2011) in which they fit a generalized linear model (GLM) with a DP as the mixing distribution for subject-specific regression coefficients β_i. The discreteness of the DP implies that for two subjects $i \neq j$, there is a positive probability that $\beta_i = \beta_j$. Thus, the DP induces clustering among subjects through shared values of β. If, in addition, we consider the covariates random with a DP on their parameters, the clustering occurs jointly on β and these covariate parameters.

Wade, Mongelluzzo, and Petrone, (2011) showed that when the dimension of the covariates is large, clustering can form based more on similarity in the covariate space rather than β. This impacts predictive performance as more clusters form than needed to adequately describe the heterogeniety of the regression parameters β. To address this problem, the enriched DP (EDP) was proposed which allows for clusters based on the covariate parameters to be nested within clusters for β. It was shown that predictive performance improved when an EDP prior was used in lieu of a DP prior in applications with many covariates (Wade et al., 2014). In the previous chapter, the idea of using an EDP prior was extended within a mixed model framework to handle longitudinal data.

Regression

For subject $i = 1, \ldots, n$, let n_i be the total number of observations that the i^{th} subject contributes with $N = \sum_{i=1}^{n} n_i$. Let the outcome y_{ij} correspond to the j^{th} measurement for the i^{th} subject recorded at time t_{ij}. The vector $\mathbf{y_i}$ corresponds to all outcomes for the i^{th} subject and the vector $\mathbf{y}$ designates the $N \times 1$ vector of all subjects combined. Let $\mathbf{t_i}$ and $\mathbf{t}$ denote the corresponding vectors of time. The treatment, denoted a_i, is administered at baseline, and we let $\mathbf{x_i}$ be the $p \times 1$ vector of covariates measured at baseline. To facilitate nonlinear effects across time, we define the matrix $\mathbf{z}$ to be a spline basis matrix for time where each individual has submatrix $\mathbf{z_i}$. Capital letters Y, A, and X refer to the random variables.

In the previous chapter, we showed how to estimate the joint distribution of (Y, A, X) by decomposing the model into the product of the outcome Y given time, A, and X alongside the marginal

distributions of X and A. For the outcome process, we used a linear mixed model where the effects of $\mathbf{x}$ and a on $\mathbf{y}$ are parameterized by $\boldsymbol{\beta}$ and the time (spline) effect of $\mathbf{z}$ on y is parameterized by $\boldsymbol{\eta}$. Together let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2)$ be the regression parameters, where σ^2 is the regression variance parameter. To complete this model, the marginal distributions for the covariates X and treatment A are assumed to have a Bernoulli distribution if binary and a normal distribution if continuous. These distributions are parameterized by the vector $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p)$. As such, if the k^{th} covariate is binary, then $\boldsymbol{\psi}_k$ is the univariate probability parameter. If it is continuous, then $\boldsymbol{\psi}_k$ is the two-dimensional parameter with a mean and variance. If we place a DP prior on $\boldsymbol{\theta}$, and $\boldsymbol{\psi}$ so that $(\boldsymbol{\theta}, \boldsymbol{\psi}) \sim P$, where $P \sim DP(\alpha, P_0)$ for some P_0, our model implies that subjects will cluster on similar values of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. As mentioned in the previous section, if the dimension of $\boldsymbol{\psi}$ is large, clusters can form based largely on $\boldsymbol{\psi}$ and predictive performance can suffer. To address this issue, we use the enriched DP (EDP) prior.

When using an EDP prior, we instead write $P \sim EDP(\alpha_\theta, \alpha_\psi, P_0)$ indicating $P_\theta \sim DP(\alpha_\theta, P_{0\theta})$ and $P_{\psi|\theta} \sim DP(\alpha_\psi, P_{0\psi|\theta})$ where $P_0 = P_{0\theta} \times P_{0\psi|\theta}$. As with the DP prior before, the EDP prior is also discrete and induces clustering on $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. However, this formulation yields nested clustering where clusters for $\boldsymbol{\psi}$ are nested within clusters for $\boldsymbol{\theta}$. The additional mass parameters, α_θ and α_ψ, control the number of clusters for θ and ψ, respectively.

The full model for the observed data with an EDP prior is

$$\begin{aligned}
\mathbf{y_i}|\mathbf{x_i}, \mathbf{t_i}, \boldsymbol{\beta_i}, \boldsymbol{\eta_i}, u_i, \sigma_i^2 &\sim \mathsf{N}(\mathbf{x_i^*}\boldsymbol{\beta_i} + \mathbf{z_i}\boldsymbol{\eta_i} + u_i, \sigma_i^2\mathbf{I}), && (4.1)\\
u_i &\sim \mathsf{N}(0, \sigma_u^2);\\
x_{ij}|\boldsymbol{\psi_i} &\sim \mathsf{N}(\mu_{ij}, \sigma_{\mu,ij}^2) \text{ (for continuous covariates)}; && (4.2)\\
x_{ij}|\boldsymbol{\psi_i} &\sim \text{Bernoulli}(p_{ij}) \text{ (for binary covariates)}; && (4.3)\\
(\boldsymbol{\theta_i}, \boldsymbol{\psi_i})|P &\sim P;\\
P &\sim \mathsf{EDP}(\alpha_\theta, \alpha_\psi, P_0);\\
\sigma_u^2, \alpha_\theta, \alpha_\psi &\sim \text{Inv-Ga}(a_u, b_u) \times \mathsf{Ga}(a_\theta, b_\theta) \times \mathsf{Ga}(a_\psi, b_\psi);
\end{aligned}$$

where $\boldsymbol{\theta_i} = (\boldsymbol{\beta_i}, \sigma^2_{\beta,i}, \boldsymbol{\eta_i}, \sigma^2_{\eta,i}, \sigma^2_i)$ are the regression parameters. The centering distributions are

$$P_{0\theta} \sim \underbrace{\text{Inv-Ga}(a_\beta, b_\beta)}_{\sigma^2_\beta} \times \underbrace{\text{N}(\boldsymbol{\beta_0}, \sigma^2_{\beta,i}\mathbf{I}))}_{\beta} \times \underbrace{\text{Inv-Ga}(a_\eta, b_\eta)}_{\sigma^2_\eta} \times \underbrace{\text{N}(0, \sigma^2_{\eta,i}\mathbf{I})}_{\eta} \times \underbrace{\text{Inv-Ga}(a_y, b_y)}_{\sigma^2};$$

and

$$P_{0\psi|\theta} \sim \text{Beta}(a_x, b_x);$$

$$P_{0\psi|\theta} \sim \text{scaled Inv-}\chi^2(\nu_0, \tau_0^2) \times \text{N}(\mu_0, \tau^2/c),$$

for binary and continuous covariates, respectively. We use the notation $\mathbf{x}_i^*$ to indicate the vector $\mathbf{x}_i$ with time t_{ij} possibly added, as would be the case if splines were omitted.

We call clusters based on $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ θ-clusters and ψ-clusters, respectively. The nested clustering of the EDP prior means that it is possible that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ but $\boldsymbol{\psi}_i \neq \boldsymbol{\psi}_j$ for some $i \neq j$. This is typically not possible with a DP prior unless the centering measure $P_{0\psi|\theta}$ is discrete. Let $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*)$ denote all k unique values of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_j^*$ denote the parameter values for the j^{th} unique cluster. If all subjects share the same $\boldsymbol{\theta}$, then $\boldsymbol{\theta}^*$ is a vector of length 1. On the other hand, if $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j$ for all $i \neq j$, then $\boldsymbol{\theta}^*$ will be a vector of length n. We introduce a latent cluster membership parameter $s_i = (s_{i,y}, s_{i,x})$ in which $s_{i,y}$ refers to the corresponding θ-cluster to which the i^{th} subject belongs. If $s_{i,y} = k$, then $\boldsymbol{\theta}_i = \boldsymbol{\theta}_k^*$.

Causal Effects

Let Y_t^a denote the counterfactual outcome for Y at time t had, possibly contrary to fact, $A = a$ been the treatment administered. For binary treatment A, some common causal contrasts are given below:

- Average treatment effect: $E(Y_t^1 - Y_t^0)$
- Causal risk ratio: $E(Y_t^1)/E(Y_t^0)$
- Conditional treatment effect: $E(Y_t^1 - Y_t^0|V = v)$
- Treatment effect on the treated: $E(Y_t^1 - Y_t^0|A = 1)$.

In this paper, we focus on the average treatment effect and the conditional treatment effect for

continuous outcomes. To estimate these effects, we use the parametric g-formula, which involves modeling the observed data and then using that model to simulate outcomes under hypothetical treatments (Naimi, Cole, and Kennedy, 2017).

To identify the unobservable quantity $E(Y_t^a)$, we make the following assumptions:

1. Consistency: $Y_t = Y_t^a$ if $A = a$. This assumption asserts that the counterfactual outcome is equal to the observed outcome if $A = a$.

2. Exchangeability: $Y_t^a \perp A|X$. This assumption asserts that given confounders X, the treatment A can be thought of as randomly assigned.

3. Positivity: $P(A = a|X = x) > 0$ whenever $P(X = x) > 0$. This assumption asserts that there is a nonzero probability of treatment for every possible combination of covariates X.

Using these assumptions, we link the observed data (Y_t, A, X) to the counterfactual data through the g-formula, whose derivation is given in Naimi, Cole, and Kennedy, (2017), for example. We show how to use the g-formula in practice using our EDP model in the following section.

Computations

Full computations for the joint EDP longitudinal model are in the previous chapter and Appendix B. Corresponding code is available at https://www.github.com/zeldow/EDPlong. Our Gibbs sampler is based on algorithm 8 in Neal, (2000) and the algorithm in Wade et al., (2014). Briefly, the algorithm alternates between updating cluster membership and updating the parameter values. In this section, we describe computations for the parametric g-formula to estimate causal effects after the joint EDP model has been estimated and we have posterior distributions for all model parameters. The calculations for the g-formula are not required for fitting the joint EDP model so the g-formula may either be done in parallel or in a post-processing step using saved parameter values. Let n_k^θ be the number of subjects in the k^{th} unique θ-cluster and let $n_{j|k}^\psi$ denote the number of subjects in the j^{th} unique ψ-cluster nested within the k^{th} unique θ-cluster. Let ℓ_θ denote the number of unique θ-clusters and let $\ell_{\psi|k}$ be the number of unique ψ-clusters nested within the k^{th} θ-cluster.

For the g-formula, we simulate a dataset of subjects with the same covariate distribution as the tar-

get population. In this simulated population, we modify the treatment (and possibly other modifiable covariates) to represent a treatment strategy (e.g., everyone in the population is treated). Using the simulated/modified data, we then calculate the expected outcome given the covariates and treatment. Once we have this for everyone in the simulated population, we average the outcome over the the empirical distribution of the covariates.

Below are steps to estimate $E(Y_t^a)$ after having modeled the joint distribution for the observed data with the longitudinal EDP model. To calculate a risk difference for a binary treatment, perform the following steps separately at $A = 1$ and $A = 0$ and take the difference between the results. An advantage of using Bayesian methods for these calculations are the ease of getting posterior intervals as well as the ability to compute different quantities using the same posterior distribution. The data are simulated as follows:

For $m = 1, \ldots, M$,

1. Draw s_y^m from a multinomial distribution with probabilities $\left(\frac{n_1^\theta}{n+\alpha_\theta}, \ldots, \frac{n_{\ell_\theta}^\theta}{n+\alpha_\theta}, \frac{\alpha_\theta}{n+\alpha_\theta}\right)$.

2. If $s_y^m = k \leq \ell_\theta$, draw s_x^m from a multinomial distribution given $s_y^m = k$, with probabilities $\left(\frac{n_{1|k}^\psi}{n_k^\theta+\alpha_\psi}, \ldots, \frac{n_{\ell_{\psi|k}}^\psi}{n_k^\theta+\alpha_\psi}, \frac{\alpha_\psi}{n_k^\theta+\alpha_\psi}\right)$.

3. Draw X^m from $p(x|\boldsymbol{\psi}_{j|k}^*)$ whenever $s_y^m = k \leq \ell_\theta$ and $s_x^m = j \leq \ell_{\psi|k}$. Otherwise, subject m is part of a new cluster in which case we draw $\boldsymbol{\psi}_0$ from the base distribution $P_{0\psi|\theta}$, and then draw X^m from $p(x|\boldsymbol{\psi}_0)$.

4. Modify treatment $A^m = a$.

5. Calculate the probabilities for each $s_y^m = k$ for $k = 1, \ldots, \ell_\theta$ given X^m and A^m. In addition, calculate the probability that a subject is in a new θ-cluster. Draw s_y^m from a multinomial distribution with probability for $s_y = k \leq \ell_\theta$:

$$\Pr(s_y^m = k|A^m = a, X^m) = \frac{n_k^\theta}{\alpha_\theta + n} \times \left[\frac{\alpha_\psi}{\alpha_\psi + n_k^\theta} f_{x,0}(\mathbf{x}_i) + \sum_j \left(\frac{n_{j|k}^\psi}{\alpha_\psi + n_k^\theta} \prod_{l=1}^p f_{x,l}(x_{i,l}; \boldsymbol{\psi}_{j|k}^*)\right)\right],$$

and probability:

$$\text{Pr}(s_y^m = \ell_\theta + 1 | A^m = a, X^m) = \frac{\alpha_\theta}{\alpha_\theta + n} \times f_{x,0}(\mathbf{x}_i)$$

for $s_y = \ell_\theta + 1$.

6. Using the probabilities from the previous step, compute the weighted average of $E(Y^m|A^m, X^m, s_y^m)$. That is,

$$\begin{aligned}\gamma_m(a) &= E\left\{Y|A^m = a, X^m\right\} \\ &= \frac{1}{\ell_\theta + 1} \sum_{k=1}^{\ell_\theta+1} E\left\{Y|A^m = a, X^m, s_y^m = k\right\} \cdot \text{Pr}\left[s_y^m = k|A^m = a, X^m\right],\end{aligned}$$

where if $s_y^m = k$ for $1 \leq k \leq \ell_\theta$, we evaluate the expectation using parameters θ_k^*.

Once we compute the expected outcome under treatment $A = a$ for all m, average over all values M:

$$E(Y_t^a) = \frac{1}{M} \sum_{i=1}^{M} \gamma_i(a).$$

In step 3, the notation $f_{x,l}(x_{i,l}; \boldsymbol{\psi}_{j|k}^*)$ refers to the l^{th} covariate evaluated at the parameters for the j^{th} ψ-cluster nested within the k^{th} θ-cluster. The notation $f_{x,0}(\mathbf{x}_i)$ in step 5 refers to $\prod_{l=1}^{p} \int_\psi f_{x,l}(x_{i,l}) dP_{0\psi|\theta}$, the density of the covariates integrated over the base measure. This calculation for continuous and binary covariates with the distribution chosen in this EDP model appears in the Appendix B.

Simulations

We use simulation to measure the effectiveness and small sample properties of g-estimation with our joint model using an EDP prior. The quantity we are interested in for all simulations is the causal risk ratio $\omega = \omega_1 - \omega_0 = E(Y_{t^*}^1 - Y_{t^*}^0)$ for a time point of interest t^*. We use $n = 1000$ total subjects who are randomly assigned between 1 and 5 repeated measurements with time points generated randomly between 0 and 1. For all scenarios, the causal effect is assessed at $t^* = 0.67$.

Scenario 1: Simple functional forms, few covariates, two clusters

$$
\begin{aligned}
x_i &\overset{iid}{\sim} \text{N}(0,1);\\
v_i &\overset{iid}{\sim} \text{Bernoulli}(0.75);\\
a_i &\overset{iid}{\sim} \text{Bernoulli}(0.50);\\
u_i &\overset{iid}{\sim} N(0,\sigma_u^2);\\
y_i &\overset{ind}{\sim} N(\xi_i\mu_1 + (1-\xi_i)\mu_2, \sigma^2),
\end{aligned}
$$

where $\xi_i \sim \text{Bernoulli}(p_i)$ with $p_i = 0.7$ if $a_i = 1$ and $p_i = 0.2$ if $a_i = 0$. Also, $\mu_1 = \mathbb{1}_{t>0.75} \cdot (t - 0.75) + 0.25v_i$ and $\mu_2 = 2 \cdot \mathbb{1}_{t\leq 0.5} \cdot t + \mathbb{1}_{t>0.5} - 0.25v_i$.

In this scenario, the true value for ω is 0.3125. The mean estimated causal effect over 100 datasets was $\hat{\omega} = 0.3145$. Coverage for the 95% credible interval was 0.98 and the empirical mean square error was 0.0002. The true values for ω_1 and ω_0 were 0.625 and 0.3125, respectively, with estimates $\hat{\omega}_1 = 0.626$ and $\hat{\omega}_0 = 0.312$. Coverages for ω_1 and ω_0 were 94% and 99%.

Scenario 2: Simple functional forms, many covariates, two clusters

$$
\begin{aligned}
x_{i,1} - x_{i,15} &\overset{iid}{\sim} \text{Bernoulli}(0.50);\\
x_{i,16} - x_{i,20} &\overset{iid}{\sim} \text{Bernoulli}(0.75);\\
x_{i,21} - x_{i,30} &\overset{iid}{\sim} N(0,1);\\
v_i &\overset{iid}{\sim} \text{Bernoulli}(0.75);\\
a_i &\overset{ind}{\sim} \text{Bernoulli}(p_{a_i});\\
u_i &\overset{iid}{\sim} N(0,\sigma_u^2);\\
y_i &\overset{ind}{\sim} N(\xi_i\mu_1 + (1-\xi_i)\mu_2, \sigma^2),
\end{aligned}
$$

where $p_{a_i} = \text{expit}\left(-0.2 + 0.25x_{i,1} + 0.5x_{i,13} - 0.75x_{i,16} + 0.2x_{i,20} - 0.2x_{i,21}\right)$. Also, the latent parameter $\xi_i \sim \text{Bernoulli}(p_i)$ with $p_i = 0.7$ if $a_i = 1$ and $p_i = 0.2$ if $a_i = 0$. The mean functions are $\mu_1 = \mathbb{1}_{t>0.75} \cdot (t - 0.75) + 0.25v_i$ and $\mu_2 = 2 \cdot \mathbb{1}_{t\leq 0.5} \cdot t + \mathbb{1}_{t>0.5} - 0.25v_i$.

In this scenario, the true value for ω is 0.3125, with $\omega_1 = 0.625$ and $\omega_0 = 0.3125$. Our point estimates are $\hat{\omega} = 0.26$, $\hat{\omega}_1 = 0.59$, and $\hat{\omega}_0 = 0.33$, with 95% credible interval coverages of 47%, 73%, and 79%, respectively.

Scenario 3: Complex functional forms, many correlated covariates, two clusters

$$
\begin{aligned}
x_i &\overset{iid}{\sim} \text{MVN}(0, \Sigma); \\
a_i &\overset{ind}{\sim} \text{Bernoulli}(p_{a_i}); \\
u_i &\overset{iid}{\sim} N(0, \sigma_u^2); \\
y_i &\overset{ind}{\sim} N(\xi_i \mu_1 + (1 - \xi_i)\mu_2, \sigma^2)
\end{aligned}
$$

where Σ is a 30×30 AR(1) with $\rho = 0.5$ and the diagonal containing ones. The probability of receiving treatment is

$p_{a_i} = \text{expit}\left(-0.2 + 0.25x_{i,1} + 0.5x_{i,13} - 0.75x_{i,16} + 0.2x_{i,20}x_{i,21}\right)$. Here, $\xi_i \sim \text{Bernoulli}(p_i)$ with $p_i = \text{expit}\left(a_i + \frac{\sin(x_{i,1})}{4} - 4x_{i,21}\right)$. The mean functions for each cluster are $\mu_1 = \mathbb{1}_{t>0.75} \cdot (t - 0.75) + 0.5a_i + x_{i,15} - \sin(x_{i,16}/2)$ and $\mu_2 = 2 \cdot \mathbb{1}_{t \leq 0.5} \cdot t + \mathbb{1}_{t>0.5} - 0.2a_i + \exp(x_{i,15}) - \sin(x_{i,21}/2)$.

The parameter ρ does not alter the mean the causal effect but may affect estimation in terms of bias or efficiency. Here, the true value for $\omega = 0.036$. When $\rho = 0$, the mean causal effect over 100 datasets was 0.065. Coverage for the 95% credible interval was 0.92 with empirical mean square error (MSE) at 0.0022. When $\rho = 0.5$, the estimated causal effect over 100 datasets was $\hat{\omega} = 0.064$. Coverage for the 95% credible interval was 0.90 with empirical mean square error (MSE) at 0.0023. The true values for the ω_1 and ω_0 were 0.879 and 0.844. The estimates for these values were 0.899 and 0.836.

Scenario 4: Simple functional forms, many covariates, one cluster

$$
\begin{aligned}
x_i &\overset{iid}{\sim} \text{MVN}(0, \Sigma); \\
a_i &\overset{ind}{\sim} \text{Bernoulli}(p_{a_i}); \\
u_i &\overset{iid}{\sim} N(0, \sigma_u^2); \\
y_i &\overset{ind}{\sim} N(\mu_1, \sigma^2)
\end{aligned}
$$

where Σ is a 30×30 AR(1) with $\rho = 0.5$ and the diagonal containing ones. The probability of receiving treatment is

$p_{a_i} = \text{expit}\,(-0.2 + 0.25x_{i,1} + 0.5x_{i,13} - 0.75x_{i,16} + 0.2x_{i,20}x_{i,21})$. The mean function is $\mu_1 = \mathbb{1}_{t>0.75} \cdot (t - 0.75) + 0.5a_i + x_{i,15} - \sin(x_{i,16}/2)$.

Here, the true values are $\omega = 0.5$, $\omega_1 = 0.5$, and $\omega_0 = 0.0$. The posterior means for each of the three parameters were $\hat{\omega} = 0.497$, $\hat{\omega}_1 = 0.498$, and $\hat{\omega}_0 = 0.001$. The 95% coverage of the credible intervals were 95%, 97%, and 92%, respectively.

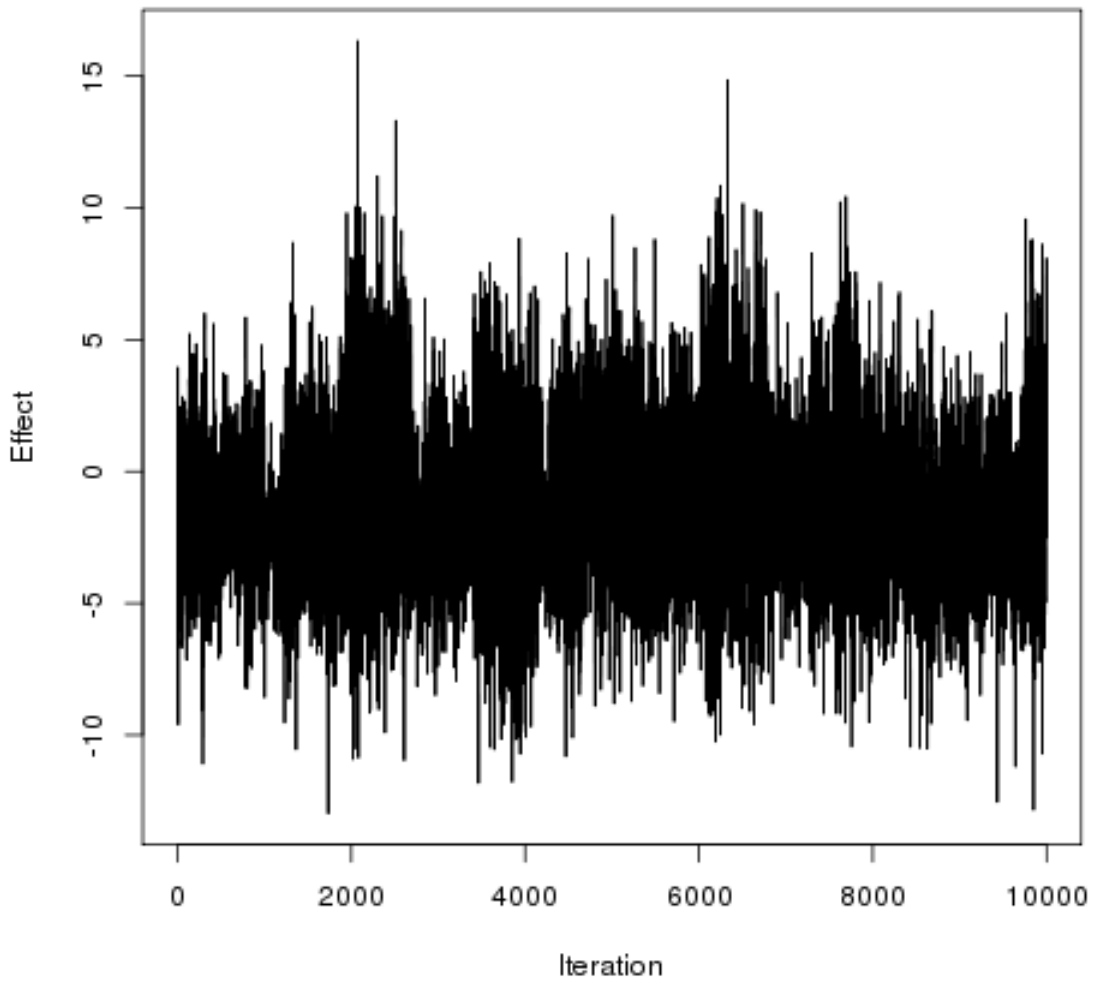
Data Analysis

To demonstrate implementation of the g-formula, we use data from the Sentinel Distributed Database (SDD) (*Sentinel*), which we used to esimate incidence rates of diabetes in the previous chapter. In short, the SDD is a collection of health records initiated by the US Food and Drug Administration with the aim of assessing safety in approved drugs and devices. Our analysis was restricted to adults who initiate a second generation antipsychotic (SGA), which are known to increase the risk of diabetes (De Hert et al., 2012; Newcomer, 2005). As part of its collection procedure, the SDD records laboratory tests (labs) over time. In this analysis, we look at fasting glucose labs for which elevated values (≥ 126 mg/dL) indicate diabetes. Our goal is to determine how much aripiprazole increases fasting glucose one year after initiation compared to other SGAs (olanzapine, quetiapine, and risperidone). That is, if we let Y denote fasting glucose, measure time t in years, and set $A = 1$ to denote use of aripiprazole, we are interested in the quantity:

$$\omega = E(Y^1_{t=1} - Y^0_{t=1}).$$

As this was not a planned study, there is no guarantee that fasting glucose is recorded at $t = 1$. In fact, there were only 2032 subjects contributing 4110 observations over a five year period. After 20,000 iterations with 10,000 burnin, the estimate for ω was -1.27 with a 95% credible interval of $(-6.80, 5.25)$. As the interval contains 0, there is no evidence of a causal effect of aripiprazole versus other SGAs on fasting glucose after one year. The trace plot for this effect is shown in Figure 4.1.

Figure 4.1: Trace plot for causal effect on fasting glucose



The x-axis shows only post burn-in iterations.

Discussion

In this chapter, we provided an algorithm for the parametric g-formula in applications with a longitudinal outcome and baseline covariates. This builds on the work from the previous chapter in which we fit a joint model with an EDP prior. Using the joint model, we simulate a hypothetical dataset from the marginal distribution of the covariates. We modify treatment (and possibly other modifiable

covariates) to correspond to the causal estimand we seek to estimate. The mean counterfactuals are computed by averaging the expected outcomes for each individual, taking into account the probability of each individual being in each cluster.

In our simulations with many covariates, there was some bias in estimating the risk difference. The bias in the risk difference stemmed from some bias in the individual effects of $Y_{t^*}^1$ and $Y_{t^*}^0$. In general, the bias in these effects were in the opposite direction, indicating that the crux of the problem might be in the assignment of cluster probabilities in Step 5 of our algorithm.

Our model contains some limitations which we suggest directions for future work. Primarily, we would like to extend our joint model to handle time-varying treatment and covariates. Once our joint model is equipped for this, a natural extension of this algorithm for the g-formula is to handle more complicated scenarios such as those arising in dynamic treatment regimes (Taubman et al., 2009; Westreich et al., 2012). To incorporate time-varying covariates in our framework, we can introduce a dynamic distribution which can update over time. In a planned study with fixed observation points, the covariate distribution can be updated at fixed times (Rodriguez and Ter Horst, 2008). For EHR with covariates updated at irregular intervals, additional methods may be needed.

A final direction for future work combines functional clustering and causal inference. Causal inference on continuous treatment effects have been developed (Kennedy et al., 2017), yielding continuous effect curves, but there are no instances in the literature in which functional clustering is the target of inference in causal settings. In the previous chapter, we showed how our joint EDP longitudinal model also serves as a functional clustering algorithm. An interesting extension of this work would be to apply this functional clustering algorithm to hypothetical treatments in a causal inference setting.

CHAPTER 5

CONCLUSION

Summary

In this thesis, we developed novel Bayesian nonparametric methods for common statistical procedures, including generalized linear models and linear mixed models. While all Bayesian procedures require full specification of the observed data likelihood, Bayesian nonparametrics relax modeling assumptions by introducing either an infinite dimensional parameter, or parameters whose dimension increases as the sample size grows. These methods have gained popularity in recent years in conjunction with increased computational power, making the nonparametric procedures more attractive for their robustness to modeling assumptions while being computationally feasible.

One of the first statistical models discussed in an introductory statistics course is linear regression (and generalized linear models). In Chapter 2, we inserted Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch, 2010) into a generalized linear model framework where a large subset of covariates are modeled nonparametrically and a small subset are allowed to have standard parametric form. In practice, researchers will often fit a large linear regression model where only a few covariates are of scientific interest–such as a treatment effect and its effect modifiers–but there are a number of additional variables that are necessary to control for confounding. In a linear model, these effects are likely to be misspecified and may yield inefficient estimates of the effects that are of interest. In response to this, we propose semi-BART, which allows for these confounders to be modeled with BART, while the covariates of interest still have parametric form. We showed in simulation that when the covariates have a complex relationship with the outcome, using our semi-BART model results in increased efficiency for the parameter estimates of the effects of interest.

In the same chapter, we show that under the typical causal assumptions of consistency, exchangeability, and positivity, this model can be interpretted as a structural mean model (SMM), the first such Bayesian implementation. This is particularly useful when the outcome is binary as g-estimation (the usual estimation procedure for the parameters of a SMM with a continuous outcome) is not possible. In fact, our model fits within the framework laid out by Vansteelandt and Goetghebeur,

2003. We demonstrated our Bayesian SMM on subjects with HIV and Hepatitis C (HCV) coinfection who initiatied highly active antiretroviral treatment (HAART). Some nucleoside reverse transcriptase inhibitors (NRTIs), a class of antiretrovirals used in many HAART regimens, are known to be to be mitochondrial toxic (mtNRTI). These mtNRTIs, which include didanosine, stavudine, zidovudine, and zalcitabine, may exacerabate liver injury. Further, it may be that this effect is worsened whenever Fibrosis-4 (FIB-4), a marker of liver injury, is high. Our goal was to determine if a HAART regimen containing a mtNRTI increased the risk of death compared to a HAART regimen with a NRTI that was not a mtNRTI. In addition, we quantified to what extent FIB-4 modified this effect. To this end, we fit three models: one with a treatment effect and no effect modifier, one with effect modification by continuous FIB-4, and one with effect modification with a dichotomized version of FIB-4. In the first model with no effect modifiers, the effect estimate for mtNRTI was positive, indicating increased harm from mtNRTIs, but the 95% credible interval contained zero. In the last two models with effect modification of FIB-4, we found that high values of FIB-4 increased the risk of death from a regimen with a mtNRTI.

In Chapter 3, we developed a novel model for the joint distribution of a longitudinal continuous outcome and baseline covariates. This model decomposes into the product of marginal distributions for the covariates and a linear mixed model for the outcome given the covariates. The parameters governing the covariates and the regression parameters were given an enriched DP (EDP) prior (Wade, Mongelluzzo, and Petrone, 2011), which is the first time the prior has been used in an analysis of repeated measurements. Like the DP prior, the EDP induces a partitioning on the parameters so that subjects with similar regression patterns and similar covariate parameters are part of the same cluster. With a DP prior, the clustering occurs jointly on the regression and covariates. In contrast an EDP prior allows for this clustering to occur separately. In fact, clusters for the covariate parameters are nested within the clusters for the regression parameters. Because of this, the EDP is prefered for preditive models when the covariate space is large and clustering from the DP is dominated by the covariates rather than the regression(Wade et al., 2014). Thus, using a DP leads to many small clusters which degrades predictive performance as demonstrated by a simulation study. When data were generated from a complex scenario with nested clustering, the EDP mixture model outperformed both the DP mixture and the standard linear mixed model. Even when the simulated data arose from one cluster, meaning the linear mixed model was correctly specified, using an EDP model did not do substantially worse than the linear mixed model

in most situations. This indicates that opting for more flexibility using nonparametric priors did not hurt prediction compared to the correct model.

This methodolodgy was motivated by a study of electronic health records (EHR) in which we sought to calculate incidence rates of Type 2 Diabetes among subjects within a year of initiating a second generation antipsychotic (SGA), which is known to increase the risk of Type 2 Diabetes (De Hert et al., 2012; Newcomer, 2005). Unlike in planned observational studies which have scheduled study visits, there is no guarantee that data are available in EHR or, even if data are available, that the data are recorded at the desired times. A previous analysis looked at the differences in incidence rates between an outcome O1 which was defined by solely diagnosis codes and antidiabetic medication dispensements versus an outcome O2 defined by diagnosis codes, antidiabetic medication dispensements, and observed elevated lab values with one year of initiating an SGA (Flory et al., 2017). For the outcome O2, Flory et al., 2017 classified a subject as having diabetes through lab values if any one of the three lab values were elevated before 365 days of SGA initiation. We identified a few potential issues with this system of classifying outcomes. First, of all the lab values recorded, only about 30% were recorded within the first year, so 70% of the data is ignored. Second, outcome identification does not account for any uncertainty. That is, a subject with 100 lab values within one year with only one elevated lab would count as diabetic. Lastly, many subjects did not have any lab data recorded within one year and so were not eligible to contribute to the outcome O2 from lab data. To account for all this uncertainty, we used our EDP mixture to predict each of the three labs at one year for each subject, as if data were recorded in a planned study. Using these predictions, we calculated the incidence rate of diabetes supplemented with elevated predicted labs. Using predicted labs, we found similar incidence rates with wider confidence intervals than was reported in Flory et al., 2017, reflecting that our method captured greater uncertainty in outcome classification.

In Chapter 4, we extended the model developed in Chapter 3 to causal inference settings using the parametric g-formula (Robins, 1986). For the parametric g-formula, we simulate a hypothetical dataset from the marginal distributions of the covariates. We then modify treatment to represent a treatment strategy, such as everyone is treated or everyone is not treated. Using the joint model, we calculate the conditional mean of the outcome among the hypothetical dataset and average these values to obtain the marginal distribution of the expected potential outcome. Doing this separately,

for example, for a hypothetical dataset where everyone is treated and another where everyone is untreated yields the causal risk difference after subtracting the two results. This technique was demonstrated in simulation and in a data analysis using the same data as in Chapter 3. For this method, we sought to quantify the causal risk difference in fasting glucose value for subjects initiating aripiprazole versus another SGA. After one year, we found no evidence of a causal effect by type of SGA on fasting glucose.

Future Directions

Structural Nested Mean Models

The generalization of SMMs to time-varying treatments is called a structural nested mean model (SNMM). SNMMs were originally developed by Robins, 1986 to adjust for time-varying confounding for a time-varying treatment. Extending our semi-BART model for use as a SNMM would require developing a BART model suitable for correlated observations found in longitudinal data. A handful of researchers have done this in applied settings (Low-Kam et al., 2015; Tan, Flannagan, and Elliott, 2016; Zhang, Shih, Müller, et al., 2007), but would be an interesting topic of future research to tie in with causal models such as the SNMM.

Time-varying Treatment

One of the difficulties of using Bayesian nonparametric priors is incorporating time-varying treatments and confounders. The challenge is defining a nonparametric prior that can evolve over time as the distributions of a covariate may be different at baseline than its distribution at later time points. There has been at least one attempt at defining such a prior called the dynamic DP (Rodriguez and Ter Horst, 2008). However, the dynamic DP is designed to update at fixed time intervals so it is not clear how to apply such a prior to EHR data, whose data are collected sporadically. Finding such a prior to use with sporadically collected EHR data would make the EDP mixture model in Chapter 3 stronger, and our subsequent analysis of incidence rates of diabetes could account for possible changes in SGA use over time.

Functional Clustering

The EDP mixture model developed in Chapter 3 also serves as a functional clustering algorithm in which one does not have to specify the number of clusters present in the data beforehand. While this is a distinct advantage over parametric mixture models, it makes summarizing clusters more difficult as both the number of clusters and cluster membership change. An interesting direction for future research would be identifying the best way to summarize this clustering. In this paper, we used an *ad hoc* method proposed by Medvedovic and Sivaganesan, 2002 that worked well in our application, but whether there are better ways of summarizing clusters is an open problem.

An additional direction for future research is to use the functional clustering within a causal inference framework. To my knowledge, there is no literature in causal inference where the outcome is a cluster or the probability of being in a cluster. For example, in Chapter 4 we examined whether or not fasting glucose levels differed by type of SGA at a specific time point (one year post initiation of a SGA). Using functional clustering, it may be possible to examine if the entire trajectory of fasting glucose up to one year differs by SGA type.

APPENDIX A

CHAPTER 2 SUPPLEMENTARY MATERIALS

Simulation Setup

$$
\begin{aligned}
x_1 &\overset{iid}{\sim} Bern(0.25) \\
x_2 &\overset{iid}{\sim} Bern(0.50) \\
x_3 &\overset{iid}{\sim} Bern(0.50) \\
x_4 &\overset{iid}{\sim} Bern(0.75) \\
x_5 &\overset{iid}{\sim} Bern(0.75) \\
\begin{pmatrix} x_6 \\ \vdots \\ x_{25} \end{pmatrix} &\overset{ind}{\sim} MVN(\mu, \Sigma),
\end{aligned}
$$

where

$$\mu = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 1.5 \\ 1.5 \\ 1.5 \\ 1.5 \\ 1.5 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and

$$\Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_4 \end{bmatrix},$$

where Σ is the 20×20 covariance matrix with

$$\Sigma_1 = \begin{bmatrix} 1 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1 \end{bmatrix},$$

$$\Sigma_2 = \begin{bmatrix} 1 & 0.15 & 0.15 & 0.15 & 0.15 \\ 0.15 & 1 & 0.15 & 0.15 & 0.15 \\ 0.15 & 0.15 & 1 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.15 & 1 & 0.15 \\ 0.15 & 0.15 & 0.15 & 0.15 & 1 \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} 1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 1 \end{bmatrix},$$

$$\Sigma_4 = \begin{bmatrix} 1 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 1 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 1 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 1 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 1 \end{bmatrix},$$

,

and $\mathbf{0}$ the 5×5 matrix with zeroes.

Additional Simulation Results

Table A.1: Efficiency of Semi-BART for a continuous outcome (standard deviation = 0.01) with no effect modification.

		Semi-BART			Linear Regression			g-estimation			BART		
Function	n	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	-0.01	0.96	0.011	0.00	0.95	0.000	0.00	1.00	0.000	-0.00	0.82	0.003
	1000	0.00	0.92	0.000	0.00	0.94	0.000	0.00	1.00	0.000	0.00	0.95	0.000
	5000	-0.00	0.95	0.000	-0.00	0.94	0.000	-0.00	1.00	0.000	-0.00	0.96	0.000
Non-linear	250	0.01	0.98	0.067	0.02	0.95	0.292	0.02	0.95	0.298	-0.01	0.96	0.026
	1000	-0.00	0.96	0.001	0.01	0.94	0.073	0.01	0.96	0.073	-0.00	0.97	0.001
	5000	0.00	0.97	0.000	0.01	0.95	0.012	0.01	0.97	0.012	-0.00	0.97	0.000

Table A.2: Efficiency of Semi-BART for a continuous outcome (standard deviation = 2) with no effect modification.

		Semi-BART			Linear Regression			g-estimation			BART		
Function	n	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	0.01	0.94	0.119	0.01	0.94	0.103	0.01	0.92	0.104	-0.10	0.93	0.112
	1000	0.00	0.95	0.023	0.00	0.94	0.022	0.00	0.94	0.022	-0.02	0.95	0.023
	5000	-0.00	0.95	0.004	0.00	0.94	0.004	0.00	0.95	0.004	-0.00	0.95	0.004
Non-linear	250	0.04	0.95	0.224	0.06	0.94	0.381	0.06	0.95	0.381	-0.00	0.93	0.172
	1000	0.00	0.94	0.028	-0.00	0.95	0.089	-0.00	0.96	0.089	0.00	0.95	0.029
	5000	0.01	0.94	0.005	0.01	0.95	0.016	0.01	0.97	0.016	0.00	0.95	0.005

Table A.3: Efficiency of Semi-BART for a continuous outcome (standard deviation = 3) with no effect modification.

		Semi-BART			Linear Regression			g-estimation			BART		
Function	n	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	0.00	0.95	0.230	-0.00	0.95	0.216	-0.00	0.93	0.217	-0.17	0.93	0.224
	1000	0.01	0.95	0.053	0.01	0.95	0.048	0.01	0.94	0.048	-0.02	0.95	0.050
	5000	0.00	0.96	0.009	0.00	0.96	0.009	0.00	0.96	0.009	-0.00	0.95	0.009
Non-linear	250	0.01	0.96	0.292	0.05	0.96	0.442	0.05	0.96	0.439	-0.06	0.94	0.263
	1000	-0.00	0.94	0.060	-0.01	0.95	0.116	-0.02	0.96	0.117	-0.02	0.94	0.060
	5000	0.01	0.96	0.009	0.02	0.95	0.023	0.02	0.96	0.023	0.01	0.96	0.009

Table A.4: Efficiency of Semi-BART for a continuous outcome (standard deviation = 0.01) with effect modification.

			Semi-BART			Linear Regression			g-estimation		
Function	n		Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	ψ_1	-0.00	0.94	0.023	0.00	0.96	0.001	-0.00	0.96	0.005
		ψ_2	-0.00	0.91	0.006	0.00	0.95	0.000	0.00	0.95	0.001
	1000	ψ_1	-0.00	0.94	0.000	-0.00	0.96	0.000	-0.00	0.98	0.001
		ψ_2	0.00	0.94	0.000	0.00	0.96	0.000	0.00	0.97	0.000
	5000	ψ_1	0.00	0.93	0.000	0.00	0.94	0.000	0.00	0.96	0.000
		ψ_2	-0.00	0.94	0.000	-0.00	0.94	0.000	-0.00	0.95	0.000
Non-linear	250	ψ_1	0.01	0.98	0.284	-0.02	0.94	1.499	-0.00	0.93	1.939
		ψ_2	-0.00	0.96	0.070	0.02	0.95	0.305	0.01	0.93	0.419
	1000	ψ_1	0.00	0.98	0.002	0.02	0.95	0.326	0.03	0.96	0.368
		ψ_2	-0.00	0.98	0.000	-0.00	0.96	0.070	-0.01	0.96	0.081
	5000	ψ_1	-0.00	0.96	0.000	-0.00	0.95	0.060	-0.01	0.97	0.067
		ψ_2	-0.00	0.96	0.000	0.01	0.96	0.013	0.01	0.97	0.015

Table A.5: Efficiency of Semi-BART for a continuous outcome (standard deviation = 2) with no effect modification.

			Semi-BART			Linear Regression			g-estimation		
Function	n		Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	ψ_1	-0.07	0.96	0.537	-0.03	0.95	0.488	-0.02	0.95	0.507
		ψ_2	0.03	0.96	0.106	0.01	0.95	0.097	0.01	0.95	0.103
	1000	ψ_1	-0.03	0.97	0.106	-0.01	0.96	0.102	-0.02	0.95	0.105
		ψ_2	0.01	0.97	0.022	0.01	0.96	0.020	0.01	0.96	0.021
	5000	ψ_1	-0.01	0.96	0.022	-0.01	0.96	0.021	-0.01	0.96	0.021
		ψ_2	0.00	0.96	0.005	0.00	0.95	0.004	0.00	0.95	0.004
Non-linear	250	ψ_1	-0.04	0.96	1.046	0.01	0.95	1.991	0.03	0.95	2.213
		ψ_2	0.01	0.95	0.216	-0.01	0.94	0.435	-0.01	0.95	0.511
	1000	ψ_1	0.01	0.97	0.128	-0.01	0.96	0.414	0.00	0.96	0.457
		ψ_2	-0.00	0.97	0.025	0.01	0.93	0.093	-0.00	0.94	0.107
	5000	ψ_1	0.00	0.95	0.023	-0.00	0.95	0.083	0.01	0.96	0.087
		ψ_2	-0.00	0.95	0.005	0.00	0.94	0.018	0.00	0.95	0.020

Table A.6: Efficiency of Semi-BART for a continuous outcome (standard deviation = 3) with no effect modification.

			Semi-BART			Linear Regression			g-estimation		
Function	n		Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Linear	250	ψ_1	-0.05	0.96	1.176	-0.00	0.95	1.117	0.01	0.93	1.254
		ψ_2	0.03	0.96	0.228	0.01	0.96	0.220	0.00	0.95	0.252
	1000	ψ_1	0.03	0.94	0.277	0.04	0.94	0.263	0.05	0.94	0.270
		ψ_2	-0.01	0.95	0.053	-0.02	0.95	0.050	-0.02	0.95	0.052
	5000	ψ_1	0.00	0.95	0.049	0.00	0.94	0.049	-0.00	0.94	0.049
		ψ_2	-0.00	0.96	0.010	-0.00	0.95	0.010	0.00	0.95	0.010
Non-linear	250	ψ_1	-0.08	0.97	1.603	0.04	0.97	2.311	0.11	0.94	2.802
		ψ_2	0.04	0.96	0.335	0.00	0.96	0.500	-0.03	0.94	0.621
	1000	ψ_1	0.02	0.95	0.279	-0.03	0.96	0.540	-0.00	0.96	0.576
		ψ_2	-0.00	0.97	0.056	0.02	0.95	0.118	0.00	0.96	0.125
	5000	ψ_1	-0.00	0.95	0.051	0.00	0.95	0.109	0.01	0.95	0.119
		ψ_2	0.00	0.94	0.010	0.00	0.95	0.022	-0.00	0.95	0.025

Trace Plots

Figure A.1: Trace plot for analysis without effect modification.

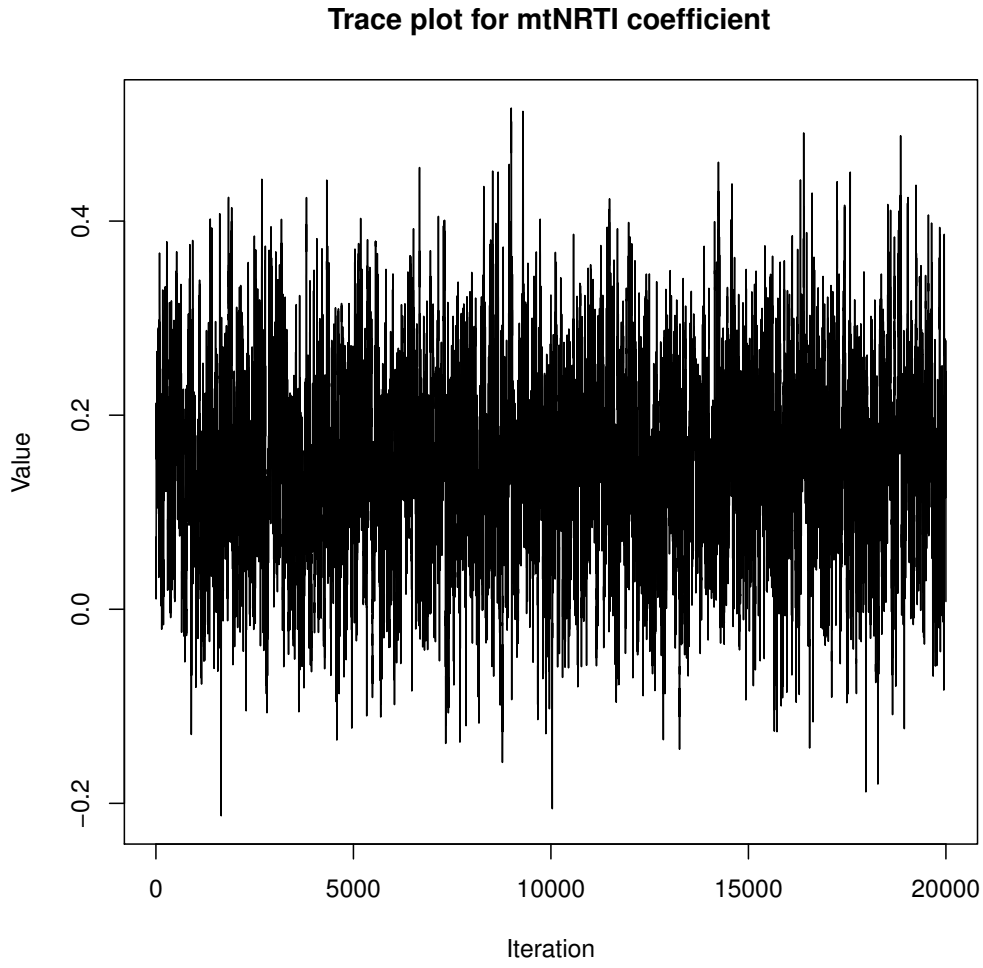


Figure A.2: Trace plots for analysis with effect modification for continuous FIB-4 (centered around 3.25).

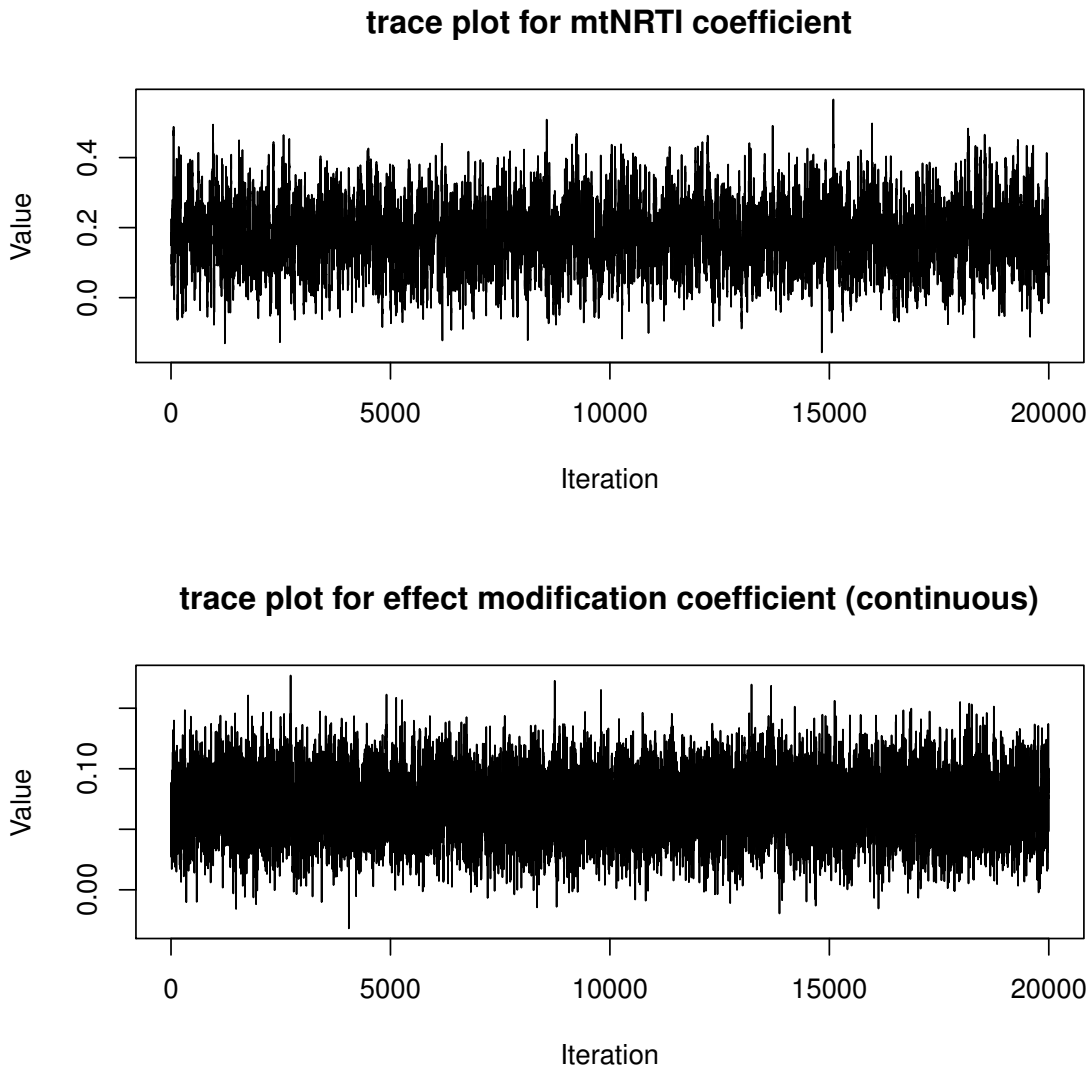
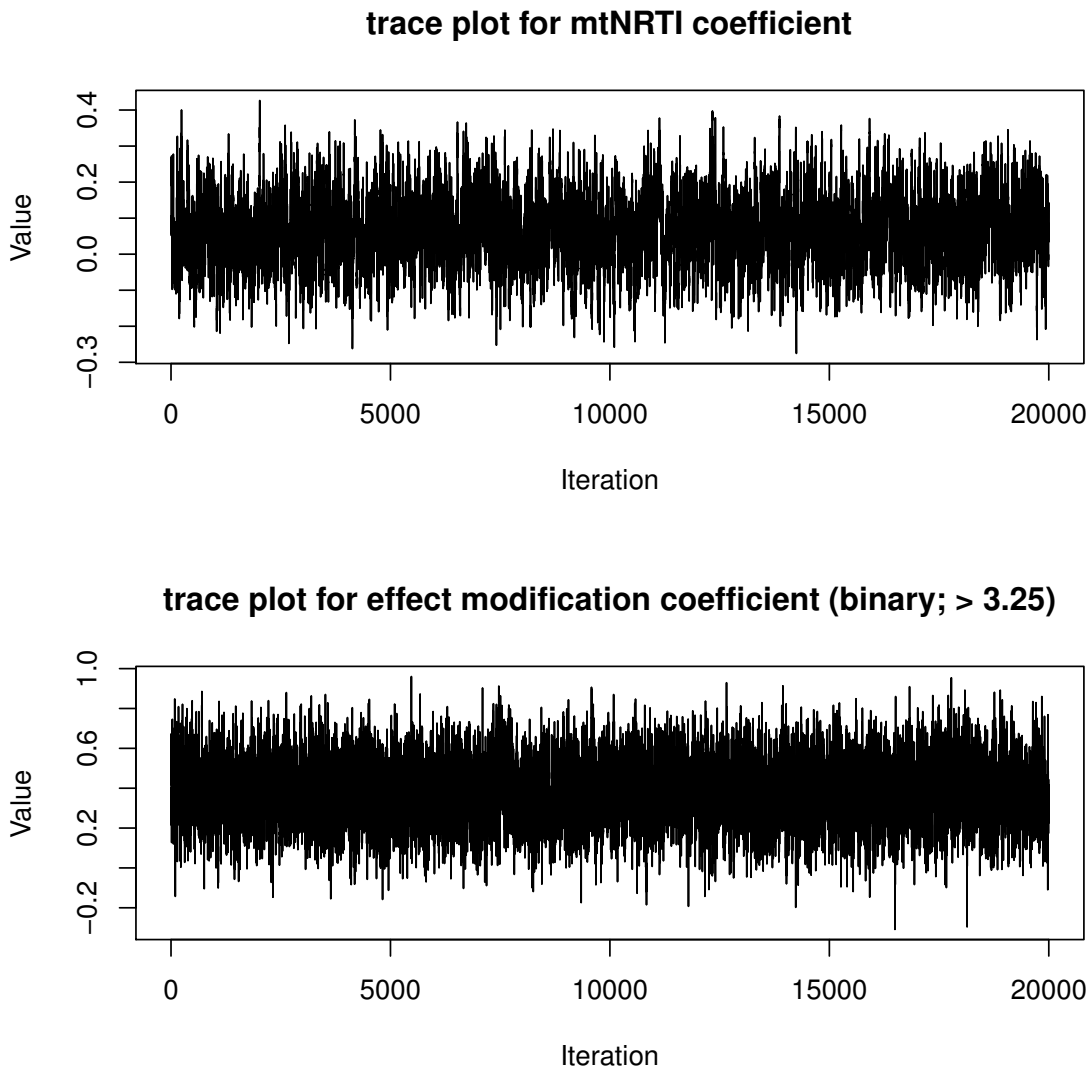


Figure A.3: Trace plots for analysis with effect modification for binary FIB-4 (cutpoint = 3.25).



APPENDIX B

CHAPTER 3 SUPPLEMENTARY MATERIALS

Simulation Setup

See Figure 3.2 for details on clustering structure. Each subject i has a random number of time points observed drawn from a discrete uniform distribution on $[1, \ldots, 5]$, call this n_i. We now randomly draw n_i time points from a uniform distribution on $[0, 1]$ and order them as $t_{i1} < \cdots < t_{in_i}$. For each subject we draw a random intercept from a $N(0, \sigma_u^2)$ distribution. The outcome is generated from independent normal distribution (given u_i) with variance σ^2 and the mean μ depending on which θ-cluster the subject was randomly assigned to. For θ_1, the outcome has mean $\mu = 2 + 7t - 2x_1 - 0.5 + 2\cos(x_4) + u_i$. For θ_2, $\mu = 7 - 20(t - 0.4)^2 + 1.1x_2 - 0.8x_3 + 0.5x_4^2 + u_i$. For θ_3, $\mu = 6 - 8(t - 0.75)^2 - 3x_1 - x_4 + x_5 + u_i$. For all the above t represents the randomly generated time points for each subject.

There were 20 covariates x were generated as:

$\boldsymbol{\psi}_{1\|1}$:	$\boldsymbol{\psi}_{2\|1}$:	$\boldsymbol{\psi}_{3\|1}$:
$x_1 \sim Bern(0.5)$	$x_1 \sim Bern(0.3)$	$x_1 \sim Bern(0.5)$
$x_2 \sim Bern(0.75)$	$x_2 \sim Bern(0.5)$	$x_2 \sim Bern(0.5)$
$x_3 \sim Bern(0.2)$	$x_3 \sim Bern(0.5)$	$x_3 \sim Bern(0.8)$
$x_4 \sim N(0, 1)$	$x_4 \sim N(0.5, 0.5)$	$x_4 \sim N(0.5, 2)$
$x_5 \sim N(\sqrt{2}, \sqrt{2})$	$x_5 \sim N(1, 2)$	$x_5 \sim N(0, 1)$
$x_6 - x_{20} \sim N(0, 1)$	$x_6 - x_{20} \sim N(-0.5, 1)$	$x_6 - x_{20} \sim N(0.5, 1)$

$$
\begin{aligned}
\boldsymbol{\psi}_{1|2}: & \\
x_1 &\sim Bern(0.75) \\
x_2 &\sim Bern(0.5) \\
x_3 &\sim Bern(0.35) \\
x_4 &\sim N(2,1) \\
x_5 &\sim N(0,1) \\
x_6 - x_{20} &\sim N(-0.5,1)
\end{aligned}
\qquad
\begin{aligned}
\boldsymbol{\psi}_{2|2}: & \\
x_1 &\sim Bern(0.5) \\
x_2 &\sim Bern(0.5) \\
x_3 &\sim Bern(0.5) \\
x_4 &\sim N(1,2) \\
x_5 &\sim N(-1,1) \\
x_6 - x_{20} &\sim N(0.5,1)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\psi}_{1|3}: & \\
x_1 &\sim Bern(0.75) \\
x_2 &\sim Bern(0.1) \\
x_3 &\sim Bern(0.3) \\
x_4 &\sim N(0.5,1.5) \\
x_5 &\sim N(0,1) \\
x_6 - x_{20} &\sim N(0.5,1)
\end{aligned}
\qquad
\begin{aligned}
\boldsymbol{\psi}_{2|3}: & \\
x_1 &\sim Bern(0.5) \\
x_2 &\sim Bern(0.3) \\
x_3 &\sim Bern(0.5) \\
x_4 &\sim N(-0.5,1) \\
x_5 &\sim N(0,0.5) \\
x_6 - x_{20} &\sim N(-0.5,1)
\end{aligned}
\qquad
\begin{aligned}
\boldsymbol{\psi}_{3|3}: & \\
x_1 &\sim Bern(0.5) \\
x_2 &\sim Bern(0.7) \\
x_3 &\sim Bern(0.5) \\
x_4 &\sim N(0,2) \\
x_5 &\sim N(-1,2) \\
x_6 - x_{20} &\sim N(0,1)
\end{aligned}
$$

Computations

The R/C++ code is available at https://github.com/zeldow/EDPlong. We give some further details on updating some of the parameters in our model below.

MCMC program:

Step 0: Let n be the total number of subjects and N denote the total number of observations. Initialize all parameter values including s, the partitioning variable.

Step 1: Update s_i for $i = 1, \ldots, n$.

Let n_θ be the number of unique θ-clusters in $s_{y,i}$.

Step 2: Iterate through $k = 1, \ldots, n_\theta$.

Restrict to subjects with $s_{i,y} = k$. Let n_k be the number of subjects in the cluster and N_k be the total number of observations in this cluster. Below y_i, y, X, x_i, Z, and z_i will refer to subjects within the given cluster.

Step 2a: Update σ_k^2 and β_k^*:

First, calculate residuals: $y_i^* = y_i - z_i\eta_i - u_i$. We specify priors $P(\sigma^2) \sim$ Inv-Ga(a_β, b_β) and $P(\boldsymbol{\beta}) \sim$ N$(\boldsymbol{\beta_0}, \Sigma)$. Define $\Sigma_n = X^\intercal X + \Sigma$ and $\beta_n = \Sigma_n^{-1}(\Sigma\beta_0 + X^\intercal y^*)$. The posteriors (within clusters) are given by $P(\sigma^2|\text{rest}) \sim$ Inv-Ga$(a_\beta + \frac{N_k}{2}, b_\beta + \frac{1}{2}(y^{*\intercal}y^* + \beta_0^\intercal\Sigma\beta_0 - \beta_n^\intercal\Sigma_n\beta_n)$ and $P(\beta|rest) \sim N(\beta_n, \sigma_k^2\Sigma_n)$.

Step 2b: Update $\sigma_{b,k}^2$ and η_k^*

Now, calculate residuals: $y_i^* = y_i - x_i\beta_i - u_i$. Given prior distributions $P(\sigma_b^2) \sim$ Inv-Ga(a_β, b_β) and $P(\boldsymbol{\eta}) \sim$ N$(\mathbf{0}, \sigma_b^2\mathbf{I})$, define $\Sigma_{b,n} = Z^\intercal Z/\sigma_k^2 + \mathbf{I}/\sigma_{b,k}^2$ and $\mu_n = \left[\sigma_{b,n}^2\right]^{-1} Z^\intercal y^*/\sigma_k^2$. The posteriors are given by $P(\sigma_b^2|\text{rest}) \sim$ Inv-Ga$(a_\beta + N/2, b_\beta + \frac{1}{2}\boldsymbol{\eta}^\intercal\boldsymbol{\eta})$ and $P(\boldsymbol{\eta}|rest) \sim N(\mu_n, [\Sigma_{b,n}]^{-1})$.

Step 2c: Iterate through $k = 1, \ldots, n_{\psi,k}$, where $n_{\psi,k}$ is the number of ψ-clusters nested within the k^{th} θ-cluster. Now, we update covariate parameters ψ, further restricting to subjects with $s_{i,2} = k$:

For binary covariates, the prior is $P(p) \sim$ Beta(a_x, b_x) and the posterior is given by $P(p|\text{rest}) \sim$ Beta$(\sum_{s=(j,k)} x_{i,l} + a_x, n_{j|k} - \sum_{s=(j,k)} x_{i,l} + b_x)$..

For continuous covariates, prior: $P(\mu, \sigma) \sim$ scaled Inv-$\chi^2(\nu_0, \tau_0^2) \times$ N$(\mu_0, \tau^2/c)$ with posteriors $P(\sigma^2|\text{rest}) \sim$ scaled Inv-$\chi^2(\nu_0 + n_{\psi,k}, \nu_0\tau_0 + n_{j|k} * \text{var}(x) + c_0 n_{j|k}/(c_0 + n_{j|k}) * (\bar{x} - \mu_0)^2)$ and $P(\mu|\text{rest}) \sim N(\mu_n, \sigma_n^2)$, where $\sigma_n^2 = \frac{1}{c_0/\tau + n_{j|k}/\tau}$ and $\mu_n = \sigma_n^2(\mu_0 * c_0/\tau + \bar{x}n_{j|k}/\tau)$.

This marks the end of the within-cluster updates.

Step 3: Update random intercept u_i:

Iterate through $i = 1, \ldots, n$. Calculate residuals $y_i^* = y_i - x_i\beta_i - z_i\eta_i$. Draw u_i from $N(\mu_u, \sigma^2_{\text{new}})$ where $\sigma^2_{\text{new}} = \frac{\sigma_u^2\sigma_i^2}{n_i\sigma_u^2+\sigma_i^2}$ and $\mu_u = \sigma_u^2 \sum_{j=1}^{n_i} y_{ij}/(n_i\sigma_u^2 + \sigma_i^2)$.

Step 4: Update random intercept variance σ_u^2:

Given prior $P(\sigma_u^2) \sim$ Inv-Ga(a_u, b_u) the posterior is $P(\sigma_u^2|\text{rest}) \sim$ Inv-Ga$(a_u + \frac{n}{2}, b_u + \frac{1}{2}\mathbf{u}^\intercal\mathbf{u})$.

Step 5: Update α_θ (from Escobar and West (1995)):

Let α_θ be the current value. Draw $\gamma \sim Beta(\alpha_\theta, n)$. Define $\pi = \frac{n_\theta/(n(1-\log(\gamma)))}{1+n_\theta/(1-\log(\gamma))}$. Draw $p \sim Bern(\pi)$. Update α_θ from Gamma$(a_\alpha + n_\theta, b_\alpha - \log(\gamma))$ with probability p and from $Gamma(a_\alpha + n_\theta - 1, b_\alpha - \log(\gamma))$ with probability $1 - p$.

Step 6: Update α_ψ:

Update α_ψ with Metropolis-Hastings step. Our proposal distribution is Gamma(a_0, b_0). Draw α_{prop} from proposal distribution. Define

$$p_1 = \text{dGamma}(\alpha_\psi; a_\alpha, b_\alpha)\alpha_\psi^{n_\theta} \prod_{j=1}^{n_\theta} \left[(\alpha_\psi + n_j)\text{Beta}(\alpha_\psi + 1, n_j)\right].$$

Let

$$p_2 = \text{dGamma}(\alpha_{\text{prop}}; a_\alpha, b_\alpha)\alpha_{\text{prop}}^{n_\theta} \prod_{j=1}^{n_\theta} \left[(\alpha_{\text{prop}} + n_j)\text{Beta}(\alpha_{\text{prop}} + 1, n_j)\right].$$

Note dGamma$(x; a, b)$ denotes the density of function of a Gamma distribution with parameters a and b evaluated at x. Beta(u, v) denotes the Beta function evaluated at u and v. Set $\alpha_\psi = \alpha_{\text{prop}}$ with probability $p = \frac{p_2}{p_1}$. Otherwise, use previous α_ψ.

Return to Step 1 and repeat until convergence and posteriors are well approximated.

Additional Simulation Results

Table B.1: Simulation results for $n = 1000$ showing mean l_1 and l_2 errors over 100 datasets for predictions at $t = 0.75$ using cubic B-splines.

	EDP		DP		ME	
	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$
$\sigma^2 = 1; \sigma_u^2 = 0.15$	0.66	0.87	0.91	1.46	1.10	1.83
$\sigma^2 = 1; \sigma_u^2 = 0.5$	0.84	1.25	1.09	1.96	1.11	1.85
$\sigma^2 = 4; \sigma_u^2 = 0.15$	0.91	1.52	1.09	2.03	1.22	2.28
$\sigma^2 = 4; \sigma_u^2 = 0.5$	1.06	1.92	1.18	2.33	1.23	2.33

σ^2 indicates the simulated regression variance and σ_u^2 indicates the simulated random intercept variance. EDP indicates the longitudinal model with an enriched Dirichlet process prior. DP indicates the longitudinal model with a Dirichlet process prior. ME indicates a mixed effects model fit using the lmer package in R. Fit with cubic B-splines with 2 knots.

Table B.2: Simulation results for $n = 5000$ showing mean l_1 and l_2 errors over 100 datasets for predictions at $t = 0.75$ using cubic B-splines.

	EDP		DP		ME	
	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$
$\sigma^2 = 1; \sigma_u^2 = 0.15$	0.62	0.76	0.92	1.49	1.10	1.81
$\sigma^2 = 1; \sigma_u^2 = 0.5$	0.77	1.07	1.10	2.02	1.11	1.85
$\sigma^2 = 4; \sigma_u^2 = 0.15$	0.72	1.03	1.04	1.90	1.21	2.26
$\sigma^2 = 4; \sigma_u^2 = 0.5$	0.85	1.28	1.15	2.26	1.23	2.33

σ^2 indicates the simulated regression variance and σ_u^2 indicates the simulated random intercept variance. EDP indicates the longitudinal model with an enriched Dirichlet process prior. DP indicates the longitudinal model with a Dirichlet process prior. ME indicates a mixed effects model fit using the lmer package in R. Fit with cubic B-splines with 2 knots.

Table B.3: Simulation results for $n = 5000$ showing mean l_1 and l_2 errors over 100 datasets for predictions at $t = 0.75$ when the standard mixed effects model is correctly specified.

	EDP		DP		ME	
	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$	$\bar{\ell}_1$	$\bar{\ell}_2$
$\sigma^2 = 1; \sigma_u^2 = 0.15$	0.31	0.15	0.31	0.15	0.26	0.11
$\sigma^2 = 1; \sigma_u^2 = 0.5$	0.56	0.50	0.56	0.50	0.37	0.22
$\sigma^2 = 4; \sigma_u^2 = 0.15$	0.32	0.16	0.32	0.16	0.31	0.15
$\sigma^2 = 4; \sigma_u^2 = 0.5$	0.57	0.50	0.57	0.50	0.49	0.38

σ^2 indicates the simulated regression variance and σ_u^2 indicates the simulated random intercept variance. EDP indicates the longitudinal model with an enriched Dirichlet process prior. DP indicates the longitudinal model with a Dirichlet process prior. ME indicates a mixed effects model fit using the lmer package in R. Fit with cubic B-splines with 2 knots.

R Code for Choosing Number of Clusters

```
## function for calculating n x n matrix of how many times
## subjects in same cluster
adjmatrix <- function(s) {
  n <- nrow(s[[1]])
  mat <- matrix(0, n, n)
  nelem <- length(s)
  for(i in 1:nelem){
    temp.mat <- as.integer(outer( s[[i]][ ,1], s[[i]][ ,1], FUN = "==" ) )
    mat <- mat + temp.mat
  }
  return(mat)
}

## a1c.res$s is a list of cluster memberships
## for successive MCMC iterations
## each element is a n x 2 matrix
## the first column is the theta-cluster membership
## the second column is the psi-cluster subcluster membership
a1c.adj <- adjmatrix(a1c.res$s)
a1c.dist <- dist(a1c.adj, method = "maximum")
hi <- hclust(a1c.dist, method = "ward.D2")
clust.a1c <-cutree(hi, k = 2)
```

Additional Computations

Discrete covariates:

$$\begin{aligned}\int p^x(1-p)^{1-x}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathsf{Be}(\alpha,\beta)}dp &= \frac{1}{\mathsf{Be}(\alpha,\beta)}\int p^{\alpha+x-1}(1-p)^{\beta-x}dp \\ &= \frac{\mathsf{Be}(\alpha+x,\beta-x+1)}{\mathsf{Be}(\alpha,\beta)}\end{aligned}$$

Continuous covariates:

Prior: Normal-inverse-χ-squared:

$$\begin{aligned}p(\mu,\sigma^2) &= \frac{\sqrt{c_0}}{\sqrt{2\pi}\sqrt{\tau_0}}\exp\left(\frac{-(\mu-\mu_0)^2}{2\tau_0/c_0}\right)\frac{(\tau_0\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\frac{\exp\left(\frac{-\nu_0\tau_0}{2\sigma^2}\right)}{(\sigma^2)^{1+\nu_0/2}} \\ &= \frac{\sqrt{c_0}}{\sqrt{2\pi}\sqrt{\tau_0}}\frac{(\tau_0\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\exp\left(\frac{-(\mu-\mu_0)^2}{2\tau_0/c_0}\right)\frac{\exp\left(\frac{-\nu_0\tau_0}{2\sigma^2}\right)}{(\sigma^2)^{1+\nu_0/2}} \\ &\propto \exp\left(\frac{-(\mu-\mu_0)^2}{2\tau_0/c_0}\right)\frac{\exp\left(\frac{-\nu_0\tau_0}{2\sigma^2}\right)}{(\sigma^2)^{1+\nu_0/2}}\end{aligned}$$

Data:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Posterior:

$$p(\mu,\sigma^2|x) \propto \sigma^{-3}(\sigma^2)^{-(\nu_n/2)}\exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2 + c_n(\mu_n-\mu)^2]\right)$$

$$
\begin{aligned}
h(x) &= \int\int p(\mu, \sigma^2|x) d\mu \; d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{c_0}}{\sqrt{2\pi}\sqrt{\tau_0}} \frac{(\tau_0\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \frac{\sqrt{2\pi}\sqrt{\tau_n}}{\sqrt{c_n}} \frac{\Gamma(\nu_n/2)}{(\tau_n\nu_n/2)^{\nu_n/2}} \\
&= \frac{1}{\sqrt{2\pi}} \frac{c_0}{c_n} \frac{\tau_n}{\tau_0} \frac{(\tau_0\nu_0/2)^{\nu_0/2}}{(\tau_n\nu_n/2)^{\nu_n/2}} \frac{\Gamma(\nu_n/2)}{\Gamma(\nu_0/2)}
\end{aligned}
$$

where

$$
\begin{aligned}
c_n &= c_0 + 1 \\
\nu_n &= \nu_0 + 1 \\
\tau_n &= \frac{1}{\nu_n} \left(\nu_0\tau_0 + \frac{c_0}{c_n} (\mu_0 - x)^2 \right)
\end{aligned}
$$

BIBLIOGRAPHY

Albert, JH and Chib, S (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* 88.422, 669–679.

Association, AD (2014). Diagnosis and classification of diabetes mellitus. *Diabetes care* 37.Supplement 1, S81–S90.

Bates, D, Maechler, M, Bolker, B, and Walker, S (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version* 1.7.

Bigelow, JL and Dunson, DB (2009). Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association* 104.485, 26–36.

Biller, C (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics* 9.1, 122–140.

Biller, C and Fahrmeir, L (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling* 1.3, 195–211.

Black, N (1996). Why we need observational studies to evaluate the effectiveness of health care. *BMJ: British Medical Journal* 312.7040, 1215.

Brezger, A and Lang, S (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis* 50.4, 967–991.

Chipman, HA, George, EI, and McCulloch, RE (1998). Bayesian CART Model Search. *Journal of the American Statistical Association* 93, 935–948.

Chipman, HA, George, EI, and McCulloch, RE (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics* 4, 266–298.

Crainiceanu, CM, Ruppert, D, and Wand, MP (2005). Bayesian analysis for penalized spline regression using Win BUGS. *Journal of Statistical Software* 14.14.

Cruz-Mesía, RDl, Quintana, FA, and Müller, P (2007). Semiparametric Bayesian classification with longitudinal markers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 56.2, 119–137.

Das, K, Li, R, Sengupta, S, and Wu, R (2013). A Bayesian semiparametric model for bivariate sparse longitudinal data. *Statistics in medicine* 32.22, 3899–3910.

De Hert, M, Detraux, J, Van Winkel, R, Yu, W, and Correll, CU (2012). Metabolic and cardiovascular adverse effects associated with antipsychotic drugs. *Nature Reviews Endocrinology* 8.2, 114–126.

Denison, DGT, Mallick, BK, and Smith, AFM (1998a). A bayesian CART algorithm. *Biometrika* 85, 363–377.

Denison, DG, Mallick, BK, and Smith, AF (1998b). Bayesian mars. *Statistics and Computing* 8.4, 337–346.

Denison, D, Mallick, B, and Smith, A (1998c). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2, 333–350.

Eilers, PH and Marx, BD (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 89–102.

Escobar, MD and West, M (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association* 90.430, 577–588.

Ferguson, TS (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Ferguson, TS (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics* 24.1983, 287–302.

Flory, JH, Roy, J, Gagne, JJ, Haynes, K, Herrinton, L, Lu, C, Patorno, E, Shoaibi, A, and Raebel, MA (2017). Missing laboratory results data in electronic health databases: implications for monitoring diabetes risk. *Journal of Comparative Effectiveness Research* 6.1, 25–32.

Friedman, JH (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

Gelman, A, Carlin, JB, Stern, HS, and Rubin, DB (2014). *Bayesian data analysis*. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Günthard, HF, Saag, MS, Benson, CA, Del Rio, C, Eron, JJ, Gallant, JE, Hoy, JF, Mugavero, MJ, Sax, PE, Thompson, MA, et al. (2016). Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2016 recommendations of the International Antiviral Society–USA panel. *Jama* 316.2, 191–210.

Hannah, LA, Blei, DM, and Powell, WB (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research* 12.Jun, 1923–1953.

Hanson, T and Johnson, WO (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics* 13.2, 341–361.

Hastie, T, Tibshirani, R, et al. (2000). Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science* 15.3, 196–223.

Hastie, TJ and Tibshirani, RJ (1990). *Generalized additive models*. Vol. 43. CRC press.

Hernán, MA and Robins, JM (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology* 183.8, 758–764.

Hernán, MA and Robins, JM (2018). *Causal Inference*. Chapman & Hall/CRC.

Hernán, MÁ, Brumback, B, and Robins, JM (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 11.5, 561–570.

Hill, JL (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20.1, 217–240.

Holmes, C and Mallick, B (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.1, 3–17.

Jacques, J and Preda, C (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification* 8.3, 231–255.

Kennedy, EH (2016). "Semiparametric Theory and Empirical Processes in Causal Inference". In: *Statistical Causal Inferences and Their Applications in Public Health Research*. Ed. by H He, P Wu, and DGD Chen. Cham: Springer International Publishing, 141–167. ISBN: 978-3-319-41259-7. DOI: 10.1007/978-3-319-41259-7_8. URL: https://doi.org/10.1007/978-3-319-41259-7_8.

Kennedy, EH, Ma, Z, McHugh, MD, and Small, DS (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, 1229–1245. ISSN: 1467-9868. DOI: 10.1111/rssb.12212. URL: http://dx.doi.org/10.1111/rssb.12212.

Li, Y, Lin, X, and Müller, P (2010). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics* 66.1, 70–78.

Low-Kam, C, Telesca, D, Ji, Z, Zhang, H, Xia, T, Zink, JI, Nel, AE, et al. (2015). A Bayesian regression tree approach to identify the effect of nanoparticles properties on toxicity profiles. *The Annals of Applied Statistics* 9.1, 383–401.

Matsouaka, RA and Tchetgen Tchetgen, EJ (2014). Likelihood Based Estimation of Logistic Structural Nested Mean Models with an Instrumental Variable.

McCullagh, P (1984). Generalized linear models. *European Journal of Operational Research* 16.3, 285–292.

Medvedovic, M and Sivaganesan, S (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* 18.9, 1194–1206.

Müller, P and Rosner, GL (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* 92.440, 1279–1292.

Müller, P, Quintana, FA, Jara, A, and Hanson, T (2015). *Bayesian nonparametric data analysis*. Springer.

Murtagh, F and Legendre, P (2014). Wards hierarchical agglomerative clustering method: which algorithms implement Wards criterion? *Journal of Classification* 31.3, 274–295.

Naimi, AI, Cole, SR, and Kennedy, EH (2017). An introduction to g methods. *International journal of epidemiology* 46.2, 756–762.

Neal, RM (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* 9.2, 249–265.

Newcomer, JW (2005). Second-generation (atypical) antipsychotics and metabolic effects. *CNS drugs* 19.1, 1–93.

Quintana, FA, Johnson, WO, Waetjen, LE, and B. Gold, E (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association* 111.515, 1168–1181.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Raebel, MA, Shetterly, S, Paolino, AR, Lu, CY, Gagne, JJ, Haynes, K, Flory, J, Patorno, E, Smith, DH, Selvan, M, Herrinton, LJ, Harrell Jr, FE, Shoaibi, A, and Roy, J. *Analytic Methods for Using Laboratory Test Results In Active Database Surveillance*. https://www.sentinelinitiative.org/sentinel/methods/analytic-methods-using-laboratory-test-results-active-database-surveillance. Accessed: 2017-10-13.

Rasmussen, CE (2006). Gaussian processes for machine learning.

Ray, S and Mallick, B (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2, 305–332.

Robins, J (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling* 7.9-12, 1393–1512.

Robins, J and Rotnitzky, A (2004). Estimation of treatment effects in randomised trials with noncompliance and a dichotomous outcome using structural mean models. *Biometrika* 91.4, 763–783.

Robins, JM (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods* 23.8, 2379–2412.

Robins, JM (2000). Marginal structural models versus structural nested models as tools for causal inference. In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 95–133.

Robins, JM, Hernán, MA, and Wasserman, L (2015). On Bayesian estimation of marginal structural models. *Biometrics* 71.2, 296.

Robins, JM, Hernan, MA, and Brumback, B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 550–560.

Rodriguez, A and Ter Horst, E (2008). Bayesian dynamic density estimation. *Bayesian Analysis* 3.2, 339–365.

Roy, J, Lum, KJ, and Daniels, MJ (2016). A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, kxw029.

Roy, J, Lum, KJ, Daniels, MJ, Zeldow, B, Dworkin, J, and Re III, VL (2017). Bayesian nonparametric generative models for causal inference with missing at random covariates. *arXiv preprint arXiv:1702.08496*.

Rubin, DB (2004). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

Saarela, O, Stephens, DA, Moodie, EE, and Klein, MB (2015). On Bayesian estimation of marginal structural models. *Biometrics* 71.2, 279–288.

Scarpa, B and Dunson, DB (2014). Enriched stick-breaking processes for functional data. *Journal of the American Statistical Association* 109.506, 647–660.

Schoen, C, Osborn, R, Squires, D, Doty, M, Rasmussen, P, Pierson, R, and Applebaum, S (2012). A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. *Health affairs* 31.12, 2805–2816.

Sciences, Engineering, and Medicine, NA of (2017). *Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop*. National Academies Press.

Scourfield, A, Jackson, A, Waters, L, Gazzard, B, and Nelson, M (2011). The value of screening HIV-infected individuals for didanosine-related liver disease? *Antiviral therapy* 16, 941–942.

Sentinel. `https://www.sentinelinitiative.org/`. Accessed: 2017-10-13.

Shahbaba, B and Neal, R (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research* 10.Aug, 1829–1850.

Soriano, V, Puoti, M, Garcia-Gasco, P, Rockstroh, JK, Benhamou, Y, Barreiro, P, and McGovern, B (2008). Antiretroviral drugs and liver injury. *Aids* 22.1, 1–13.

Sterling, RK, Lissen, E, Clumeck, N, Sola, R, Correa, MC, Montaner, J, S Sulkowski, M, Torriani, FJ, Dieterich, DT, Thomas, DL, et al. (2006). Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology* 43.6, 1317–1325.

Tan, YV, Flannagan, CA, and Elliott, MR (2016). Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian Additive Regression Trees. *arXiv preprint*.

Taubman, SL, Robins, JM, Mittleman, MA, and Hernán, MA (2009). Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International journal of epidemiology* 38.6, 1599–1611.

Teh, YW, Jordan, MI, Beal, MJ, and Blei, DM (2004). "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes." In: *NIPS*, 1385–1392.

Tsiatis, A (2006). *Semiparametric theory and missing data*. Springer Science & Business Media.

Vansteelandt, S and Joffe, M (2014). Structural nested models and G-estimation: The partially realized promise. *Statistical Science* 29, 707–731.

Vansteelandt, S and Goetghebeur, E (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.4, 817–835.

Wade, S, Mongelluzzo, S, and Petrone, S (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis* 6.3, 359–385.

Wade, S, Dunson, DB, Petrone, S, and Trippa, L (2014). Improving prediction from dirichlet process mixtures via enrichment. *Journal of Machine Learning Research* 15.1, 1041–1071.

Westreich, D, Cole, SR, Young, JG, Palella, F, Tien, PC, Kingsley, L, Gange, SJ, and Hernán, MA (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in medicine* 31.18, 2000–2009.

Zeger, SL and Karim, MR (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association* 86.413, 79–86.

Zhang, S, Shih, YCT, Müller, P, et al. (2007). A spatially-adjusted Bayesian additive regression tree model to merge two datasets. *Bayesian Analysis* 2.3, 611–633.

Zhao, Y, Staudenmayer, J, Coull, BA, and Wand, MP (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, 35–51.