12-2022

Machine Learning Solutions for Biomedical Applications

Madeleine St. Ville
mstvill@clemson.edu

Machine Learning Solutions for Biomedical Applications

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical and Statistical Sciences

---

by
Madeleine Elise St. Ville
December 2022

Accepted by:
Dr. Christopher McMahan, Committee Co-Chair
Dr. Joseph Bible, Committee Co-Chair
Dr. Deborah Kunkel
Dr. Xinyi Li

Abstract

This dissertation proposes three novel Bayesian modeling techniques to addresses the challenges arising from the complex, underlying features of biomedical data. These models are motivated by three different biomedical studies. The first is an analysis of data collected from six efficacy and safety clinical trials of buprenorphine maintenance treatment for opioid use disorder. The focus of this study is to overcome the problem of non-adherence by trial participants that, if left unaccounted for, obscures the true effect of buprenorphine on illicit opioid use. The second study is the assessment of hemodialysis cannulation skill through the use of a sensor-based simulator that provides objective metrics quantifying various facets of cannulation skill. The main objective of this study is to identify salient features from a high-dimensional feature space that influence multiple cannulation outcomes that are highly correlated, both implicitly and by design, while also addressing the presence of multicollinearity within the feature space. The third and final study focuses on modeling an individual's probability of disease from data collected on pooled specimens. The primary barrier of this study is measurement error: the individual disease statuses are likely to be obscured by the group testing protocol and the testing responses (on pools and individuals) are subject to misclassification due to imperfect testing. The key objective of this study is to develop a flexible model that can account for imperfect testing and can be used to analyze data arising from any group testing protocol. A key attribute of the proposed modeling techniques is that they scale easily to extremely large data sets. The scalability of the modeling strategies discussed here is accomplished by introducing carefully constructed latent random variables to develop Markov chain Monte Carlo (MCMC) sampling algorithms that consist primarily of Gibbs steps. This results in efficient computation of posterior estimates, especially in large data scenarios.

Dedication

To my parents, for the endless love and moral support they've given me throughout my Ph.D journey. And to my sweet dog Millie, for being by my side during the countless hours of research.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisors, Dr. Christopher McMahan and Dr. Joseph Bible, for the tremendous amount of support they continue to offer me. Their encouragement of my research has allowed me to grow as a statistician, and their enthusiasm made the completion of this dissertation a really enjoyable process. Secondly, I want to thank Dr. Alex Ewing of Prisma Health for his guidance, and for giving me the opportunity to collaborate with clinical professionals to conduct biomedical research. Finally, I would like to thank the other members of my dissertation committee, Dr. Deborah Kunkel and Dr. Xinyi Li, for their feedback and advice throughout my dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern advances in medicine and technology have resulted in the collection of large, complex datasets. Motivated by three biomedical applications, this dissertation proposes novel statistical modeling techniques to address the challenges that arise from the complicated, underlying features of these datatsets. The first motivating application is an analysis of data collected from six efficacy and safety clinical trials of buprenorphine maintenance treatment for opioid use disorder. The second study is the assessment of hemodialysis cannulation skill through the use of a sensor-based simulator that provides objective metrics that could be useful for assessing various aspects of skilled cannulation. The final study analyzes data collected on pooled specimens with the goal of modeling an individual's probability of disease.

There have been nearly 500,000 overdose deaths from opioids in the United States alone in the last 20 years, with associated annual costs exceeding $1 trillion [Kuehn, 2021]. To mitigate these issues arising from the opioid epidemic, it is essential to understand the effectiveness and safety of treatments. In substance abuse trials aimed at assessing the efficacy of various treatments and dosing protocols, the observational aspect of non-adherence can complicate the analysis. Buprenorphine is heavily used as a medication for opioid use disorder treatment (MOUD), and it has been shown that illicit opioid use and the number of weeks abstinent from illicit opioid use are significantly associated with daily buprenorphine adherence [Fiellin et al., 2006]. Many studies have been conducted to assess the effectiveness of various formulations and doses of buprenorphine on detoxification, retention in treatment, and on the elimination of illicit opioid use [Ling et al., 2010]. However, these studies do not acknowledge the problem of non-adherence by trial participants on buprenorphine. Non-adherence

is a primary barrier to being able to accurately assess the effectiveness of MOUD in clinical trials. When analyzing the relationship between buprenorphine and illicit opioid use, failing to account for the patterns in daily dose adherence will obscure the true effect of buprenorphine on opioid usage. The first project of this dissertation is motivated by and applied to publicly available data from six efficacy and safety clinical trials of buprenorphine maintenance treatment with detailed logs of patient buprenorphine dose. Trial participants were expected to attend weekly follow-up clinic visits with opiate urinalysis testing, and administration of buprenorphine doses varied across the six trials. This dissertation develops a functional generalized linear mixed model that views buprenorphine dose history as a time-varying covariate in order to estimate dose effect while accounting for lapses in adherence. The proposed model also makes use of random effects to account heterogeneity across trials, and to account for heterogeneity across subjects within studies. We cast our problem into the Bayesian paradigm to facilitate both parameter estimation and inference, and given the complexities of our problem, priors are chosen to regularize the estimation of the model parameters. The proposed methodology is used to re-assess the efficacy of buprenorphine as a MOUD, but it also demonstrates a modeling technique that can be used to directly acknowledge and account for the effect of non-adherence when assessing treatment effects and dosing protocols in medication assisted treatment trials.

End-stage kidney disease (ESKD) is the final stage of chronic kidney disease, leading to permanent kidney failure. With its prevalence increasing, ESKD is a leading public health problem [Wong et al., 2018]. Medicare costs associated with ESKD treatments account for approximately 7% of the total Medicare budget [Saran et al., 2020]. Hemodialysis is the most popular modality of dialysis treatment for ESKD [Thurlow et al., 2021], where a surgically created vascular access is cannulated so the patient's blood can be pumped through a dialysis machine in order to remove waste products and excess fluids. The hemodialysis cannulation procedure is critical for ESKD treatment, as patients' survival depends on successful cannulation of their vascular accesses at least three times a week. Unfortunately, hemodialysis cannulation is a notably challenging procedure for a variety of reasons including non-standard geometries of vascular accesses and lack of training opportunities for clinicians [Moist et al., 2013]. Consequently, accurate cannulation of vascular accesses for successful hemodialysis is a critical and complex skill, and lack of cannulation skill results in poor clinical outcomes due to miscannulation. One of the main reasons for miscannulation is infiltration, which occurs when the clinician punctures through the vascular access causing blood to leak out [Brouwer, 2011].

Infiltration can lead to adverse medical complications and even loss of a functioning vascular access - a catastrophic event for ESKD patients that would lead to death. Consequently, it is imperative that cannulation be performed by skilled clinicians, and learning how to cannulate vascular accesses for successful hemodialysis requires targeted training. Simulators have been successfully applied for assessment and training of clinical skills in a variety of medical specialties, with their ability to provide objective feedback being a key advantage [Noureldin et al., 2016; Zendejas et al., 2013]. Simulators provide trainees with the benefit of practicing clinical skills in a simulated, low-stakes, safe environment to instill confidence in skill prior to actual clinical practice, and to reduce any patient risks. The second project of this dissertation comes from the analysis of a simulator-based cannulation skill assessment study. A custom, state-of-the-art hemodialysis cannulation simulator, containing sensors that provide quantitative data measuring various facets of cannulation skill, was designed [Liu et al., 2020; Singapogu et al., 2021] and previous work has demonstrated its use to successfully quantify cannulation skill [Liu et al., 2021]. The sensor data results in a high-dimensional feature space of process metrics that allow the simulator to provide objective metrics used to measure outcomes of cannulation that are closely related to clinical outcomes. Building upon our previous research [Liu et al., 2021; Petersen et al., 2022], this project aims to identify salient process metrics from a high-dimensional feature space that influence multiple outcomes of interest; namely, the probability of successful cannulation *and* the quality of the cannulation task. This dissertation proposes a shared random parameter model under the Bayesian framework to jointly model two objective outcome metrics. These two outcomes are correlated both implicitly and by design, and we accommodate this dependence through the use of shared random effects that acknowledge subject-specific tendencies in cannulation performance. A two-stage data augmentation scheme is developed to construct a computationally efficient posterior sampling algorithm that is scalable to large datasets and easy to implement. The proposed approach has the ability to successfully identify salient features that influence cannulation performance, even under this high-dimensional setting with highly correlated outcomes and features. We apply a sparse principal component analysis (SPCA) technique that transforms the candidate feature space for the probability of success model in order to overcome the problem of multicollinearity and assist in identifying significant covariates.

The third and final project of this dissertation focuses on the development of a new modeling technique designed for the analysis of group testing data (i.e., data collected on pooled specimens). The concept of using pooling as a more cost effective data collection technique is becoming a main-

stream approach in a variety of applications such as infectious diseases [Westreich et al., 2008; Krajden et al., 2014; Lewis et al., 2012], animal disease testing [Dhand et al., 2010], environmental monitoring [Heffernan et al., 2014], and drug discovery [Hughes-Oliver, 2006]. In particular, pooled data is collected by first combining several specimens (e.g., blood, urine, etc.), collected from individuals, into a pooled sample, and this pooled sample is then measured for a characteristic of interest; e.g., in infectious disease studies, the pooled outcome is typically binary indicating disease status. With group testing, information on several individuals is obtained at the expense of a single diagnostic test, thus reducing the cost of data collection. However, the statistical analysis of measurements (either binary or continuous) taken on pools is often faced with many challenges; the individual measurements are obscured by a group testing protocol and the effect of imperfect assays. This problem of developing regression methods for group testing data with measurement error has been explored elsewhere: McMahan et al. [2017] proposed a Bayesian approach for the regression analysis of group testing data within a generalized linear model framework, and Liu et al. [2021] developed a novel Bayesian generalized additive model. Both of these methods can account for imperfect testing and can be used to analyze data collected according to any group testing process. The work presented here capitalizes on this previous research to develop a new methodology that is far more flexible. In particular, this dissertation proposes a Bayesian additive regression trees (BART) modeling framework to estimate regression models in potentially misclassified group testing data with individual-level covariate information. BART is a Bayesian, nonparametric approach to function estimation using regression trees [Chipman et al., 2010]. It is an ensemble, machine learning approach within the Bayesian paradigm, so uncertainty about both the functional form and the parameters will be accounted for in the posterior predictive distribution. BART accommodates nonlinear effects and high-order interactions without explicit specification, and overfitting is controlled by a regularization prior leading to increased accuracy and precision, and a better understanding of any complex effects.

The remainder of this dissertation is organized as follows. In Chapter 2, a functional generalized linear mixed effects model is developed to assess opioid use disorder treatments in trials subject to non-adherence. Chapter 3 develops generalized linear models with shared subject-specific random effects to jointly model the quality of cannulation and the probability of a successful cannulation using haptic feedback from a cannulation simulator. In Chapter 4, we develop a general Bayesian additive regression trees modeling approach for potentially misclassified group testing data with

individual-level covariate information. We illustrate the BART framework by applying the proposed approach to chlamydia test data from the State Hygienic Laboratory at the University of Iowa, which screens individuals for chlamydia using a group testing protocol. The modeling techniques introduced in these chapters are estimated within the Bayesian paradigm, with prior specifications chosen to regularize the estimation of model parameters and to aid in variable selection. In Chapter 5, we conclude with a summary discussion of the work presented in this dissertation.

Chapter 2

Assessing opioid use disorder treatments in trials subject to non-adherence via a functional generalized linear mixed effects model

2.1 Introduction

There have been nearly 500,000 overdose deaths from opioids in the United States alone in the last 20 years, with associated annual costs exceeding $1 trillion [Kuehn, 2021]. The treatment of opioid use disorder (OUD) is inherently complex, with clinician assessment of the patient, comorbidities, suitability for one of the three FDA-approved medications, psychosocial counseling and care for comorbidities [Kampman and Jarvis, 2015]. One of the two more heavily utilized medications in OUD treatment is buprenorphine, an opioid partial agonist. Studies have been conducted to assess the effectiveness of various formulations and doses of buprenorphine on detoxification, retention in treatment, and on the elimination of illicit opioid use [Ling et al., 2010]. A meta analysis of clinical

trials found that any dose over 2mg of buprenorphine was useful at retaining patients in treatment but only higher doses (16mg or more) reduced illicit opioid use [Mattick et al., 2014]. It has been shown that illicit opioid use and the number of weeks abstinent from illicit opioid use are significantly associated with daily buprenorphine adherence [Fiellin et al., 2006]. Given the potential interaction between buprenorphine dosing and adherence, further investigations aimed at better understanding these interactions are warranted and will support translational clinical research that seeks to optimize the overall effectiveness of medications for OUD treatment (MOUD).

A challenge to being able to accurately assess the effectiveness of MOUD in clinical trials is non-adherence. For example, a multicenter, randomized clinical trial (CSP-999) considered the effectiveness and safety of four buprenorphine dose levels (1, 4, 8, 16, or 32 mg/day), which were administered daily in clinic. Due to the mode of delivery, adherence for this study was directly observed, i.e., patients were either present or absent from the clinic visit. Figure 2.1 provides a depiction of dose history over a 50 day period for four randomly selected CSP-999 trial patients. In particular, three of the selected patients were assigned to a 16 mg/day dose while the remaining patient was assigned to a 8 mg/day dose. Days when patients missed a dose (i.e., were non-adherent) are represented by a dose level of 0 mg/day. Induction and re-induction after a lapse in dosing can be seen by increasing dose levels from 0 mg/day to the assigned levels. Failing to account for the patterns in adherence depicted in Figure 2.1 when trying to relate assigned dose to opioid use will obscure the true effect of buprenorphine on illicit opioid use. In particular, not accounting for these patterns will lead to underestimation of the effect (or log-odds) of dose on reducing illicit opioid use.

Given the challenge of non-adherence in substance abuse trials aimed at assessing efficacy of various treatments and dosing protocols, the goals of this paper are two-fold. First we seek to develop and demonstrate a methodology that can be used to directly acknowledge and account for the effects of adherence when assessing treatment effects. Second, we seek to use our proposed model to assess the effectiveness of buprenorphine as a MOUD. To this end, we compiled, aligned, and harmonized publicly available data from six efficacy and safety clinical trials of buprenorphine maintenance treatment with detailed logs of patient buprenorphine dose. For more on the combined data set and merging steps, see Bergen et al. [2022]. As a part of these trials, patients were expected to attend weekly follow-up clinic visits with opiate urinalysis testing. Dose administration varied across the 6 trials, with CSP-999 being the only trial requiring daily doses to be self-administered in clinic. Weekly dose adherence for the remaining trials were reconstructed using other available

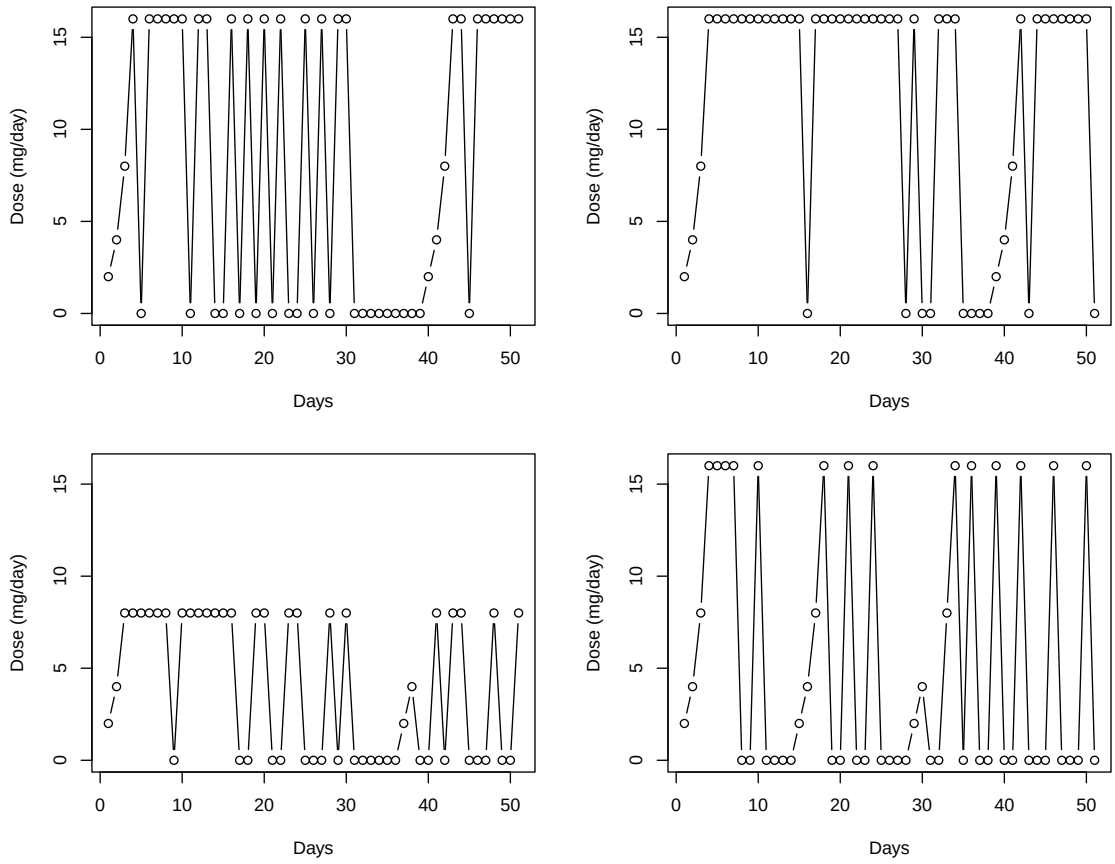

Figure 2.1: The four figures depict a time series of daily dose of buprenorphine taken by four randomly selected subjects, each coming from the CSP-999 trial, for the first 50 days of the study.

information, e.g., self reported non-adherence, returned pills, etc. Also available, were a collection of various sociodemographic and substance use variables that were included in the analysis to address potential confounding.

To analyze these data, we develop a generalized linear functional mixed effects model. The proposed model views daily dose level as a functional covariate whose value reflects the mg/day dose taken, with a value of 0 corresponding to days when the subject is non-adherent. By construction, our model has several key features. First, we can extract an estimate of, and conduct inference about, the dose effect for individuals with strict adherence (i.e., 100% compliance with the prescribed dosing protocol). Second, we can assess the effect of different types of dose self-administration or medication adherence patterns on illicit opioid use. Our model makes use of random effects to account for across trial heterogeneity, and across subject heterogeneity within studies. To complete

model fitting, we cast our model into the Bayesian paradigm and develop a custom Markov chain Monte Carlo (MCMC) posterior sampling algorithm.

2.2 Materials and Methods

2.2.1 Generalized Linear Functional Mixed Effects Model

In what follows, we outline the prominent features of our proposed model, which was designed to relate illicit opioid use to time-varying dose adherence while controlling for various demographic and drug use history factors purported to be related to the same. To this end, let Y_{ij}, for $j = 1, \ldots, n_i$, and $i = 1, \ldots, m$, be a binary indicator such that $Y_{ij} = 1$ denotes the event that the ith individual has a positive urinalysis test during the jth clinic visit with urinalysis and $Y_{ij} = 0$ otherwise. To relate the observed test data to the available covariates, we posit the following generalized linear functional mixed effects model

$$\nu_{ij} = g^{-1}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \int_{\mathcal{A}_{ij}} D_{ij}(s)\beta(s)ds + \boldsymbol{x}'_{ij}\boldsymbol{\alpha} + \gamma_{0i} + \gamma_{1k(i)}, \tag{2.1}$$

where $g(\cdot)$ is the logit link function which is used to relate the linear predictor, ν_{ij}, to the probability of relapse, $\pi_{ij} = P(Y_{ij} = 1)$. To elucidate the key feature of our model adopted to capture the effect of dose adherence, we note that the first term on the right hand side of (2.1) is the functional component which consists of the time varying buprenorphine dose curve ($D_{ij}(\cdot)$; e.g., see Figure 2.1), the functional coefficient ($\beta(\cdot)$), and the time frame ($\mathcal{A}_{ij}$) leading up to the urinalysis clinic visit over which the dose levels are allowed to impact the probability of relapse. The remaining components of the model consist of $\boldsymbol{x}_{ij}$ a $P-$dimensional vector of demographic and drug use history risk factors whose first entry is a one to allow for the usual intercept, $\boldsymbol{\alpha}$ the corresponding vector of regression coefficients, γ_{0i} a subject-specific random effect entered into the model to account for the heterogeneity across subjects, $\gamma_{1k(i)} = \gamma_{1k}$ if the ith subject is part of the kth trial, and γ_{1k} is a random effect specified to account for the heterogeneity across trials, $k = 1, \ldots, K$. Herein, the random effects distributions were taken to be

$$\begin{aligned} \gamma_{0i} &\overset{\text{iid}}{\sim} N(0, \sigma_0^2) \\ \gamma_{1k} &\overset{\text{iid}}{\sim} N(0, \sigma_1^2), \end{aligned} \tag{2.2}$$

and note, here the random effects are taken to be independent given the nesting of subjects within trials. A few comments on the form of the model are warranted. First, through adopting the functional regression framework, we are able to directly acknowledge and estimate the effect of time varying dose adherence, whereas more traditional variable aggregation techniques (e.g., average dose) fail to acknowledge key trends in dose adherence; e.g., waning adherence from the point of care or weekly patterning. Second, the time window (i.e., $\mathcal{A}_{ij}$) should be selected so that the upper bound is just before the jth clinic visit with urinalysis for the ith individual and that the length of the interval reflects the approximate elimination time for buprenorphine; i.e., buprenorphine doses taken prior to the lower bound are no longer present in the patient's system and therefore cannot impact the probability of opioid use. Generally speaking, it typically takes five half-lives for a drug to completely leave a subject's system. Thus, given that the elimination half-life of buprenorphine is 24 to 42 hours, we specified a time window consisting of 15 days to more than adequately capture the relevant dose history. Lastly, given the form of the proposed model, we can extract dose effect for individuals with strict adherence to their prescribed dosing regime. To see this, we note that if a subject adheres to the dosing regime, then $D_{ij}(s) = D_{ij}$ for all s. Thus, we would have that

$$\begin{aligned}\int_{\mathcal{A}_{ij}} D_{ij}(s)\beta(s)ds = \int_{\mathcal{A}_{ij}} D_{ij}\beta(s)ds &= D_{ij}\int_{\mathcal{A}_{ij}} \beta(s)ds \\ &= D_{ij}\beta^*\end{aligned}$$

where $\beta^* = \int_{\mathcal{A}_{ij}} \beta(s)ds$. Note, β^* represent the usual increase in log odds associated with a one unit increase in buprenorphine dose level. Thus, by estimating $\beta(\cdot)$ we can also estimate β^*.

Estimating the buprenorphine dose effect $\beta(\cdot)$ in model (2.1) is challenging from both a theoretical and computational perspective because of its infinite dimension. To reduce the number of unknown parameters needed to be estimated while also maintaining adequate modeling flexibility, we approximate $\beta(\cdot)$ using B-splines [Ramsay and Silverman, 2005]. This leads to the following representation of $\beta(\cdot)$:

$$\beta(\cdot) = \sum_{l=1}^{L} \eta_l b_l(\cdot), \tag{2.3}$$

where $b_l(\cdot)$ is a spline basis function and η_l is the corresponding spline coefficient, for $l = 1, \ldots, L$. The L spline basis functions are fully determined once the degree and knot set are specified, thus the

only unknown parameters in (2.3) are the spline coefficients. In specifying the basis functions, the degree controls the overall smoothness of the basis functions and the number of knots determines the overall modeling flexibility; for further discussion, see Ramsay and Silverman [2005]. We suggest selecting a relatively large knot set (e.g., 5-6 knots) and then regularize the estimation of the spline coefficients through the methodology outlined below.

Using the spline representation of $\beta(\cdot)$ depicted in (2.3), we can re-express the functional component in model (2.1) as follows

$$\begin{aligned}\int_{\mathcal{A}_{ij}} D_{ij}(s)\beta(s)ds &= \int_{\mathcal{A}_{ij}} D_{ij}(s)\left(\sum_{l=1}^{L}\eta_l b_l(s)\right)ds \\ &= \sum_{l=1}^{L}\left(\int_{\mathcal{A}_{ij}} D_{ij}(s)b_l(s)ds\right)\eta_l \\ &:= \boldsymbol{m}_{ij}'\boldsymbol{\eta},\end{aligned} \tag{2.4}$$

where $\boldsymbol{m}_{ij}$ is an L-dimensional vector whose lth element is $m_{ijl} = \int_{\mathcal{A}_{ij}} D_{ij}(s)b_l(s)ds$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)'$. Thus, the linear predictor of our model can be expressed as

$$\nu_{ij} = \boldsymbol{m}_{ij}'\boldsymbol{\eta} + \boldsymbol{x}_{ij}'\boldsymbol{\alpha} + \gamma_{0i} + \gamma_{1k(i)}. \tag{2.5}$$

2.2.2 Prior Specification

To facilitate both parameter estimation and inference, we cast our problem into the Bayesian paradigm. The first step in this process involves specifying prior distributions for all unknown parameters. Given the complexities of our problem, priors are chosen to regularize the estimation of the parameters. In particular, the prior for the spline coefficients is designed to encourage smoothness in the functional estimate while the prior for the regression coefficients is meant to "shrink" unimportant variables toward zero. In what follows, we briefly expand on these specifications.

To avoid overfitting issues and to encourage smooth functional estimates, herein we adopt a prior for the spline coefficients which leverages a covariance structure inspired by the usual roughness penalty [Hastie et al., 2009]. This common penalty encourages smoothness by penalizing for abrupt changes in the function through the following:

$$\int \left[\beta^{(2)}(s)\right]^2 ds = \boldsymbol{\eta}'\boldsymbol{R}\boldsymbol{\eta},$$

where $\beta^{(2)}(\cdot)$ denotes the second derivative of $\beta(\cdot)$ and $\boldsymbol{R}$ is an $L \times L$ matrix with entries $\boldsymbol{R}_{ll'} = \int b_l^{(2)}(s) b_{l'}^{(2)}(s) ds$ with $b_l^{(2)}(\cdot)$ being the second derivative of $b_l(\cdot)$. Note, the spline representation adopted for $\beta(\cdot)$ is key to being able to represent this penalty as the quadratic form depicted above; for details of this derivation, see Hastie et al. [2009]. Capitalizing on the structure of this penalty and the duality that exists between regularized estimation and prior distributions in the Bayesian paradigm, we specify the following smoothing penalty inspired prior distribution for $\boldsymbol{\eta}$:

$$\begin{aligned} \boldsymbol{\eta} &\sim N(\mathbf{0}, \lambda \boldsymbol{R}^{-1}) \\ \lambda &\sim \textit{Inv-Gamma}(a_\lambda, b_\lambda). \end{aligned}$$

In the prior specification above, the additional variance parameter λ governs the amount of smoothness and controls the trade off between over and underfitting the data.

To aid in variable selection, we adopt the generalized double Pareto shrinkage prior, proposed by Armagan et al. [2013], for all of the regression coefficients with the exception of the intercept; i.e., we specify

$$\begin{aligned} \alpha_0 &\sim N(0, \tau_0) \\ \alpha_p &\sim GDP\left(\psi = b_\delta / a_\delta, a_\delta\right), \text{ for } p = 1, \ldots P-1, \end{aligned}$$

where GDP(ψ, a_δ) refers to the generalized double Pareto distribution [Armagan et al., 2013]. Under these prior choices, setting τ_0 to be large provides a vague prior on α_0, while the hyperparameters $a_\delta > 0$ and $b_\delta > 0$ govern the amount of shrinkage. In particular, these parameters control the dispersion, with a_δ controlling the heaviness of the tails of the distribution. A typical default specification, and the one adopted herein, is to set $a_\delta = b_\delta = 1$ which leads to Cauchy-like tail behavior which is known to have desirable Bayesian robustness properties [Armagan et al., 2013].

Finally, we place inverse gamma priors on the variance components of the random effects; i.e., we specify

$$\sigma_q^2 \sim \textit{Inv-Gamma}(a_q, b_q), \quad q = 0, 1.$$

This specification is common among the literature and it leads to a proper posterior [Seltzer et al.,

1996]. Based on the prior specifications outlined above, we develop a Markov chain Monte Carlo (MCMC) sampling algorithm which facilitates both posterior estimation and inference. In what follows, we provide a brief overview of this algorithm and its construction.

2.2.3 Data Augmentation and Posterior Sampling

With ease of implementation and computational efficiency in mind, herein we outline the construction of a posterior sampling algorithm that consists solely of Gibbs steps [Gelman et al., 2013]. To accomplish this, we consider a two-stage data augmentation process. The first stage follows the work of Polson et al. [2013], and introduces carefully constructed Pólya-Gamma latent random variables so that the logistic function can be hierarchically expressed as a scale mixture of normals, where the mixing distribution is Pólya-Gamma; for further details see Polson et al. [2013]. The second stage decomposes the generalized double Pareto shrinkage prior as a scale mixture of normals; for further discussion see Armagan et al. [2013]. For the specific details of this two-stage data augmentation process, see Section A.1 of Appendix A.

The data augmentation scheme outlined above leads to the following full conditionals

$$
\begin{aligned}
\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{w},\boldsymbol{\eta},\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_1,\boldsymbol{\tau} &\sim N(\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha)\\
\boldsymbol{\eta}|\boldsymbol{Y},\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_1,\lambda &\sim N(\boldsymbol{\mu}_\eta,\boldsymbol{\Sigma}_\eta)\\
\lambda|\boldsymbol{\eta} &\sim \textit{Inv-Gamma}(a^*_\lambda,b^*_\lambda)\\
\sigma^2_q|\boldsymbol{\gamma}_q &\sim \textit{Inv-Gamma}(a^*_q,b^*_q)\\
w_{ij}|\boldsymbol{\alpha},\boldsymbol{\eta},\gamma_{0i},\boldsymbol{\gamma}_{1k(i)} &\sim PG(b^*_\delta/a^*_\delta,a^*_\delta)\\
\tau_p|\alpha_p,\delta_p &\sim \textit{Inv-Gaussian}(a^*_{\tau_p},b^*_{\tau_p})\\
\delta_p|\alpha_p &\sim \textit{Gamma}(a^*_{\delta_p},b^*_{\delta_p}),
\end{aligned}
$$

where the specific form of the parameters of these distributions are given in Section A.1 of Appendix A. These full conditionals were used to construct an MCMC algorithm in the usual manner; for further discussion see Gelman et al. [2013].

Table 2.1: Sociodemographic characteristics and drug use history for the individuals used in the analysis.

Demographics			*Sociodemographics*		
Age	Mean	SD	**Income**	Mean	SD
	36.14	9.85		20834	30025
Gender	N	%	**Employment History**	N	%
Male	2017	67	Skilled Manual	889	29
Female	1005	33	Never Gainfully	653	22
Race	N	%	Machine Operator	445	15
White	2001	66	Clerical/Sales	407	13
Hispanic	495	16	Administrative	239	8
Black	422	14	Unskilled	235	8
American Indian	50	2	Business Manager	101	3
Asian	48	2	Executive	53	2
Other	6	< 1	**Work Type**	N	%
			Fulltime	1758	58
Drug Use History			Unemployed	582	19
Years of Opiate Abuse	Mean	SD	Irregular PT	284	9
	8.23	8.41	Regular PT	232	8
Heroin Use	N	%	Retired	84	3
YES	2354	78	Student	64	2
NO	668	22	Controlled	17	< 1
Mode of Opiate Abuse	N	%	Military	1	< 1
IV	1710	57	**Education**	N	%
Snort	1089	36	High School	1456	48
Oral	122	4	Partial College	829	27
Smoking	74	2	Partial High School	304	10
Other	22	1	Standard College	213	7
Sublingual	5	< 1	Junior High School	116	4
Cocaine Use	N	%	Complete Graduate School	89	3
YES	1837	61	Less than 7th Grade	15	1
NO	1185	39	**Marital Status**	N	%
Meth Use	N	%	Married	1038	34
NO	2304	76	Never Married	1014	33
YES	718	24	Divorced	602	20
Alcohol Use	N	%	Separated	261	9
YES	1891	63	Widowed	87	3
NO	1131	37	Remarried	20	1
Tranquilizer Use	N	%	**Living Arr**	N	%
NO	1997	66	Partner & Child	1251	41
YES	1025	34	Partner Only	537	18
Marijuana Use	N	%	Parents	294	10
YES	1953	65	Family	263	9
NO	1069	35	Friends	255	8
PCP Use	N	%	Alone	190	6
NO	2545	84	Child Only	183	6
YES	477	16	Controlled	25	1
			No Stable	24	1

2.3 Secondary Data Analysis of Buprenorphine Efficacy

2.3.1 Trial Data

Clinical trial data for this analysis was sourced from the Clinical Trials Network (CTN) at NIDA's Data Share resource (`datashare.nida.nih.gov`). Using the search keyword *opiate*, we identified 10 efficacy and safety trials involving detoxification or maintenance treatment of DSM-IV opioid dependence. We selected six efficacy and safety trials focused on buprenorphine maintenance treatment for analysis. Detailed information on these trials are provided in Table 1 of Appendix A.2.

2.3.2 Patient Characteristics

The data consists of 55,739 urinalysis results from 3,022 subjects who participated in one of the six aforementioned clinical trials aimed at assessing the efficacy of buprenorphine for treating OUD. The number of urinalyses (i.e., urine drug screens for opioids) per subject ranged from 1 to 60 urinalyses, while the mean number of urinalyses per subject was 18.44 and the median was 18. The data was harmonized across the six trials and candidate predictors with a missingness greater than 25% were filtered out. This resulted in 18 demographic, sociodemographic, and substance use variables (excluding prescribed buprenorphine dose, handled by the functional component of the model). Missing demographic, sociodemographic, and substance use variables were imputed using the regularized iterative factorial analysis for mixed data (qualitative and quantitative variables) algorithm [Audigier et al., 2016], implemented by the `imputeFAMD` function in the `missMDA` R package. Summaries of the retained variables (with imputed values included) are given in Table 2.1. The daily dose of buprenorphine taken by each patient was either reported (CSP-999) or inferred from alternate information. Daily dose could vary throughout time for a variety of reasons; e.g., adherence, induction, re-induction after lapse in dosing, modification by a provider's clinical judgement, etc.

Given the number of demographic and substance use variables considered, the reference group is specifically white men with a high school diploma who are employed full time doing skilled manual labor, married and living with a partner or child, and their primary mode of opioid use being intravenous, with a history of heroin, cocaine, alcohol and marijuana use and no history of methamphetamine, tranquilizer, or PCP use. The mean age, income, and years of opioid use are 36.14 years, $20,834 per year and 8.23 years, respectively (presented in Table 2.1), while the mean

Table 2.2: Treatment and outcome characteristics of individuals used in the analysis. Urinalysis is a binary indicator that takes value 1 to denote a positive opioid drug screen and 0 otherwise.

	Mean	Median	Range
Daily Dose	12.65	14	0-90
Days in Trial	112.70	87	1-527
Urinalysis	0.41	0	0-1

dose is 12.65 mg/day (presented in Table 2.2). When we discuss conditional probabilities of relapse, comparisons will be made with respect to this hypothetical individual in the reference group by changing specific variables as noted.

2.3.3 Functional General Linear Mixed Model

The outcome variable in this analysis was the urinalysis test result for illicit opioid use (1=positive drug screen vs 0=negative drug screen). Through the model in (2.1), we relate the daily dose patterns leading up to the clinic visit with urinalysis, while controlling for the 18 demographic, sociodemographic, and substance use variables detailed in Table 2.1. For the functional dose component in model (2.1), the time trajectory was chosen to be the 15 days leading up to the current urinalysis clinic visit and, for the B-spline basis expansion of the coefficient function in (2.3), we specify the degree to be 3 to construct cubic basis functions. Two interior knots were placed at the 33.33th and 66.67th percentiles of our 15-day time range. This leads to five fully determined spline basis functions and, hence, five spline basis coefficients to estimate. For the priors outlined in Section 2.2.2, we take $\tau_0 = 1000$ to specify a vague prior on the global intercept α_0 and let $a_0 = b_0 = 0.001$, $a_1 = b_1 = 0.005$, $a_\delta = 1, b_\delta = 1$, $a_\lambda = 1, b_\lambda = 0.005$. These hyperparameter values are chosen so to produce uninformative, proper prior specifications. For sampling, we retain 5,000 MCMC iterates after a burn-in of 5,000 samples. Convergence of the MCMC chains were assessed in the usual manner; i.e., trace plots. To summarize our analysis, we report the estimated posterior means (point estimates of the effects), estimated posterior standard deviations (measures of uncertainty), and 95% equal-tailed credible intervals.

Figure 2.2 summarizes the estimated functional coefficient $\hat{\beta}(t)$ (black solid line), which represents the buprenorphine daily dose effect for the 15 days leading up to a clinic visit with urinalysis. The dashed lines are the 95% credible interval limits. On the horizontal axis, if t is the day of the current urinalysis clinic visit, then $t - 15$ represents 15 days prior and $t - 1$ represents

one day prior. Table 2.3 reports the demographic and substance use variables that were found to be significant. Of the 54 fixed effects, four were deemed to be important by the model (i.e., their estimated credible intervals did not contain zero). Table 2.3 summarizes these significant factors by reporting the estimated posterior means (point estimate of the effect), estimated posterior standard deviations (measure of uncertainty), and 95% equal-tailed credible intervals. The analogous results for the full set of demographic and substance use variables are provided in Tables 2 and 3 in Section A.3 of Appendix A.

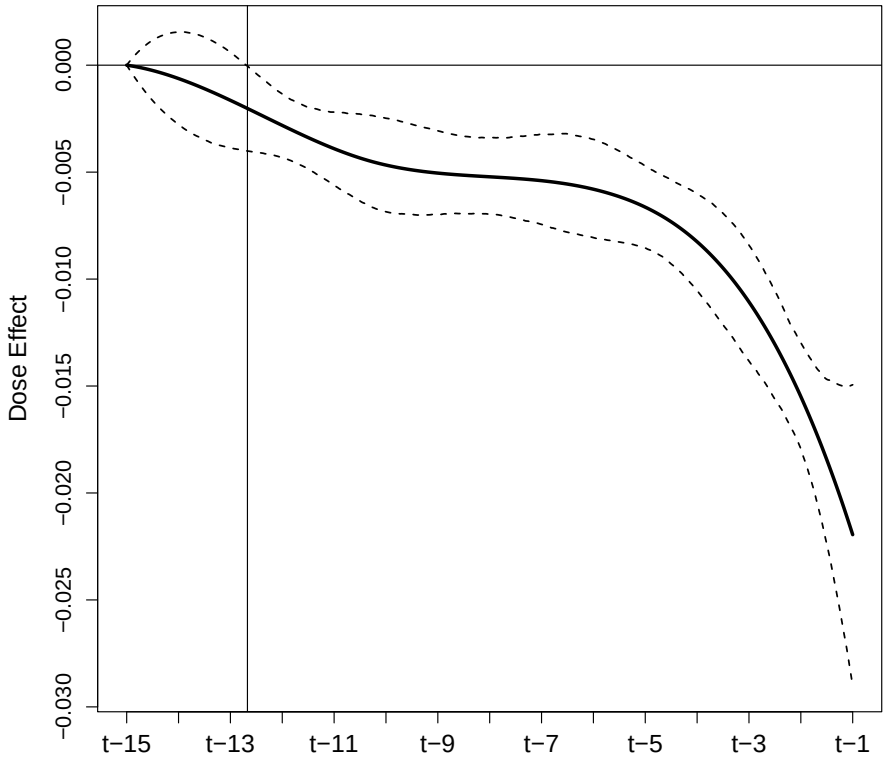

Figure 2.2: Estimated buprenorphine dose effect for the 15 days leading up to a urinalysis test, with 95% equal-tailed credible interval limits displayed as black dashed lines. The intersection of the vertical and horizontal lines is the point at which the credible interval is entirely below zero, marking the point where the dose effect becomes significant.

As previously stated, daily dose adherence was only directly recorded for patients in the CSP-999 trial. Specifically, while the assigned daily dose of buprenorphine was recorded as a part of the five other trials, adherence was not. For these trials, dose adherence was reconstructed using other available information, e.g., self reported non-adherence, returned pills, etc. To examine how

Table 2.3: Analysis results: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95) for the significant fixed effects.

Variable	**Est**	**ESE**	**CI95**
Intercept	2.05	0.27	(1.54, 2.61)
Age	-0.02	0.01	(-0.03, -0.01)
Work Type (Ref: Fulltime)			
Unemployed	0.33	0.14	(0.05, 0.59)
Heroine Use (Ref: YES)			
No Heroine Use	-0.57	0.21	(-1.00, -0.16)
Mode of Opioid Abuse (Ref: IV)			
Oral	-1.32	0.23	(-1.78, -0.88)

the buprenorphine dose reconstruction could have impacted our results, we reran our analysis on data from the CSP-999 trial only. A summary of these results can be found in Section A.3 of Appendix A.

To extract an estimate of dose effect for subjects that were strictly adherent to their assigned dosing regime, we compute the following integral for each realization $\beta(s)$, denoted $\beta^{(g)}(s)$, drawn from the posterior

$$\beta^{*(g)} = \int_{\mathcal{A}_{ij}} \beta^{(g)}(s) ds,$$

with $\beta^{*(g)}$ being a posterior realization of β^*. Table 2.4 provides a summary of these results for both the full and reduced (CSP-999) analysis to include the posterior mean estimate (point estimate of the effect), estimated standard deviation of the posterior (measure of uncertainty), and 95% equal-tailed credible interval.

Table 2.4: Analysis results: Summary includes the posterior mean estimate (Est), the estimated standard deviation of the posterior (ESE), and the estimated 95% equal-tailed credible interval (CI95) for the dose effect (i.e., β^*) for the full and reduced (CSP-999) analysis.

	Est	**ESE**	**CI95**
All Trials	-0.09	<0.01	(-0.09, -0.08)
CSP-999	-0.11	0.01	(-0.12, -0.10)

2.4 Discussion

The primary focus of our analysis is two-fold. First, we wish to demonstrate a novel approach to account for non-adherence that commonly arises in medication assisted treatment trials; especially those targeting substance use disorders. Second we wish to refine the understanding of the effectiveness of buprenorphine as a MOUD, while accounting for the potential non-adherence of study patients. To accomplish both of these tasks, we investigated the influence of multiple demographic, sociodemographic, drug use history, and treatment variables on the risk of illicit opioid use with publicly available individual patient data from six federally-sponsored buprenorphine efficacy and safety trials. To acknowledge and account for patterns of non-adherence, we conceptualized the daily dose histories of the study patients as a functional covariate and we estimated an associated functional effect. A summary of this estimated functional is provided in Figure 2.2. From these results, we identify several key findings. First, these results suggest that buprenorphine, as an MOUD, significantly reduces the risk of illicit opioid use. This can be seen from the fact that the point estimates, and associated credible intervals, are all below zero; i.e., the integral over the the product of this functional and $D_{ij}(\cdot) \geq 0$ results in a negative quantity. Second, we find that dose history extending to approximately 12.5 days prior to an opioid screening visit is significantly related to the risk of short term lapses. Third, the risk of illicit opioid use is related to dosing adherence patterns throughout the considered 15 day window leading up to the urinalysis, although, recent patterns have more influence. This can be seen from the decreasing nature of the functional estimate, especially for the five (approximately) days before the urinalysis test. Fourth, based on our estimated functional, we are able to extract an estimate of dose effect for subjects that were strictly adherent to their assigned dosing protocol. Based on this approach, we estimate that the log-odds of short-term lapse decreases by 0.09 with every 1 mg/day increase in dose; see Table 2.4. This new assessment of the efficacy of buprenorphine as an OUD treatment is unobscured by the effects of non-adherence and leverages six NIDA-sponsored efficacy and safety trials to render its conclusions.

When examining the association between risk of illicit opioid use and the other demographic, sociodemographic, and drug use history variables, four of the 54 were found to be significant. In particular, we find that increasing age is protective, while being unemployed, having a drug use history of using heroine and a drug use history of using opioids intravenously are associated with

an increased risk of illicit opioid use. A similar finding that increasing age is associated with no positive urine drug screen was recently reported in an analysis of Veterans Administration patients undergoing buprenorphine treatment [Crist et al., 2021]. The protective nature of employment for patients in recovery [Hser et al., 2015] is concordant with unemployment being identified as a risk factor for illicit opioid use. Older age, no heroine use history and no IV drug use have already been reported as protective with respect to successful opioid use outcomes (abstinent during week 24 and $\geq$ 2 of the previous 3 weeks) in a secondary data analysis of the Prescription Opioid Addiction Treatment Study (POATS or CTN-0300) [Weiss et al., 2010; Dreifuss et al., 2013], one of six CTN trials included in this study. Notably, the largest protective effect we observed was "primary mode of opioid abuse" with the log-odds of short-term lapse decreasing by 1.32 when the primary mode of abuse is oral. This could be attributable to the severity of the opioid use disorder, with intravenous use being a hallmark of more severe cases.

When examining the results of the sensitivity analysis (see Section A.3 of Appendix A) of the CSP-999 trial only, we note several similarities and differences. Importantly, the full and CSP-999 analysis came to virtually the same conclusions with regard to the efficacy of buprenorphine as an OUD treatment. In particular, the estimates of $\beta(\cdot)$ are not statistically different from each other. However, the estimate from the full analysis is slightly attenuated toward zero when compared to the CSP-999 only analysis. This feature can also be observed in the effect estimate reported in Table 2.4. A plausible explanation for this would be that our approach to reconstructing dose histories for the study patients, though effective, was not perfect, and therefore introduced "measurement error" into this variable. A hallmark of measurement error is the attenuation of effect estimates toward zero; e.g., see Stefanski [2000]. When comparing the estimated effects of demographic, sociodemographic, and substance use variables we find that most are not statistically different, yet there are differences in those deemed to be significant by the two analyses. These differences are likely attributable to increased precision due to larger sample sizes in the full analysis and differences in the demographic distribution across the full and reduced data.

We also acknowledge the lack of other risk factors that could be used to better understand/predict short-term lapse. Inclusion of time varying predictors such as current stress levels, occurrences of major life events (e.g., familial death, loss of job, etc.), and other psychological measures would undoubtedly enhance our model. However, the impact of not having these variables is mitigated by the adoption of subject specific random effects.

Importantly, this study was specifically aimed at estimating the effectiveness of recent buprenorphine treatment at reducing short term lapse. With that being said, this analysis did not consider adherence and its impact on illicit opioid use over longer periods of time and the subsequent associations with OUD related adverse outcomes, which is a far more complex problem. Studies aimed at these more long term outcomes could reveal OUD treatment strategies that would be poised to positively impact public health. That is, there have been nearly 500,000 overdose deaths from opioids in the United States alone in the last 20 years [Kuehn, 2021]. Further, OUD related mortality appears to be increasing. Specifically, the CDC estimates that overdose deaths from opioids increased to 75,673 in the 12-month period ending in April 2021, up from 56,064 in 2020 [Center for Disease Control and Prevention, 2021]. Moreover, less than one-third of patients enrolled in comprehensive health care with current OUD are being treated with one of three approved medications for OUD [Lapham et al., 2020]. Extended MOUD treatment (> 1 *vs* ≤ 1 year) appears to reduce mortality [Ma et al., 2019]. Thus, conducting more in depth studies relating MOUD treatment to long term outcomes has the potential to identify OUD treatment strategies that can be more effectively utilized to treat this epidemic and shift current clinical practice. Future research efforts will be aimed at studying these more complex topics related to dose, adherence and treatment outcomes and their association with OUD related mortality rates.

2.5 Conclusions

Inspired by the challenge of adherence in MOUD trial analysis, this work proposed a generalized linear functional mixed effects model that can acknowledge and account for the effects of adherence when assessing treatment effects. The proposed model was applied to six buprenorphine MOUD clinical trials in an effort to refine our understanding about time dependent effects that buprenorphine has on treating OUD. In particular, we find that buprenorphine dose history approximately 12.5 days prior to an opioid screening visit is significantly related to the risk of short term lapses, with the more recent history being more impactful. Further, we are able to extract an estimate of dose effect that is not obscured by adherence issues. That is, we estimate that the log-odds of short-term lapse decreases by 0.09 with every 1 mg/day increase in buprenorphine dose.

Chapter 3

High dimensional Bayesian joint modeling of skill and probability of successful simulated cannulation attempts

3.1 Introduction

End-stage kidney disease (ESKD) is the final stage of chronic kidney disease (CKD), leading to permanent kidney failure. CKD is a leading public health problem [Wong et al., 2018], with substantial associated costs and high morbidity and mortality rates [Collaboration, 2020; Saran et al., 2020], mainly attributable to ESKD treatments. Indeed, the Medicare costs associated with ESKD account for approximately 7% of the total Medicare budget [Saran et al., 2020]. The progression of CKD to ESKD results in the need for renal replacement therapy (RRT) [Walbaum et al., 2021]. The prevalence of ESKD is increasing [Wong et al., 2018], and the worldwide use of RRT for ESKD is expected to more than double by 2030 [Thurlow et al., 2021].

Dialysis is the leading form of RRT, and hemodialysis is the most popular modality [Thurlow et al., 2021]. During hemodialysis treatment, the patient's blood is pumped through a dialysis

machine to remove waste products and excess fluids. In order to perform a successful dialysis treatment, a vascular access is surgically created to the bloodstream, typically through the creation of an arteriovenous fistula (AVF) [Singapogu et al., 2021]. The hemodialysis cannulation (HDC) procedure requires the following steps: accurately locate where to insert the needle, needle insertion until blood enters the cannula (referred to as 'flashback'), advancing the needle forward to allow sustained blood flow (i.e., attaining 'stable blood flashback'), and securing the needle [Brouwer, 2011]. The HDC procedure is critical for ESKD treatment, as patients' survival depends on successful cannulation of their vascular accesses three times a week. This is a notably challenging procedure because vascular accesses (typically AVF) are patient-specific anatomical structures that are in non-standard geometries and have varying blood flow [Moist et al., 2013]. Consequently, accurate cannulation of AVFs for successful hemodialysis is a critical and complex skill, and it is imperative to avoid miscannulation [Lok et al., 2020; Brouwer, 2011]. One of the main reasons for miscannulation is infiltration, which occurs when the clinician punctures through the AVF causing blood to leak out [Brouwer, 2011]. Infiltration can lead to adverse medical complications and even loss of a functioning vascular access [Lee et al., 2006]. Thus, lack of cannulation skill can result in poor clinical outcomes, even morbidity and death. It is crucial cannulation is performed by skilled clinicians, and there is a pressing need to properly train clinicians to safely and effectively cannulate vascular accesses for hemodialysis.

Simulators have been successfully applied for assessment and training of clinical skills in a variety of medical specialties, with their ability to provide objective feedback being a key advantage [Noureldin et al., 2016; Zendejas et al., 2013]. Simulators provide trainees with the benefit of practicing clinical skills in a simulated, low-stakes, safe environment to instill confidence in skill prior to actual clinical practice, and to reduce any patient risks. A custom, state-of-the-art hemodialysis cannulation simulator was designed [Liu et al., 2020; Singapogu et al., 2021] and previous work has demonstrated its use to successfully quantify cannulation skill [Liu et al., 2021]. The simulator is comprised of four fistulas with various geometrical and physical characteristics, two different skin thicknesses, and two different motor vibration intensities. It contains four types of sensors that measure various facets of cannulation skill; for details, see Liu et al. [2020]. The quantitative data from these sensors result in over 400 descriptors quantifying summaries of time, force, motion smoothness, palpation, needle angle/location, etc. Based on these sensor data, this simulator has the ability to provide objective metrics used to measure outcomes of cannulation that are closely

related to clinical outcomes [Liu et al., 2020, 2021].

Liu et al. [2021] devised four objective metrics to measure cannulation outcomes based on recommendations from the recent Kidney Disease Outcomes Quality Initiative (KDOQI) guidelines, which defines skilled cannulation to be time-efficient with only one needle insertion attempt, stable flashback, no infiltration, and minimal patient pain [Lok et al., 2020]. In what follows, we briefly describe the four metrics; for further details, see Liu et al. [2021].

The first metric, *flash efficiency (eff)*, measures the quality of blood flashback. It is defined to be the ratio measuring efficiency of time spent inside the simulated vascular access obtaining flashback, relative to the time spent under the skin. The second metric is *number of attempts (atts)*, which counts the number of needle insertion attempts after initial insertion. Per the KDOQI guidelines, more than one insertion attempt is undesirable because it can lead to patient discomfort and damage to the vascular access. The third metric, *stb*, is a binary indicator of stable flashback attainment; i.e., *stb*=1(0) represents the ability (inability) to maintain stable flashback. The last metric is *number of infiltrations (infls)* estimates the number of infiltrations that occurs during cannulation by counting the number of times flashback occurs and then disappears. KDOQI guidelines stress the importance of avoiding infiltration because of the associated clinical complications [Lok et al., 2020].

The indicator of stable flashback, *stb*, signals sustained blood flow and completion of the cannulation task. It quantifies an outcome of cannulation as defined by the KDOQI guidelines, and it corresponds to the clinical scenario where successful cannulation is identified as sustained blood flow for hemodialysis. Petersen et al. [2022] used *stb* as an outcome metric to model the probability of successful cannulation and successfully identified salient palpation metrics. However, note that this metric does not account for any errors or inefficiencies before stable flashback was ultimately attained. Indeed, even with stable blood flashback resulting in a successful cannulation, adverse events such as infiltration or multiple insertion attempts can still occur and lead to patient discomfort along with clinical complications [Lok et al., 2020]. Liu et al. [2021] combined the four previously described metrics into a continuous metric, *ocScore*, quantifying overall quality of cannulation. This composite metric was formulated to facilitate precise measurement of cannulation performance based on cannulation outcomes as defined by the KDOQI guidelines. In particular,

ocScore is a penalized version of flash efficiency and is defined to be

$$ocScore = eff\,(1 - 0.25(\boldsymbol{I}[atts > 1] + \boldsymbol{I}[infils > 1] + \boldsymbol{I}[stb = 0])).$$

The penalties imposed are related to the occurrence of one or more of three distinct adverse events, based upon assessment of the assessment of the KDOQI guidelines' definition of quality cannulation; namely, infiltration, requiring multiple insertion attempts, and inability to attain stable flashback. The range of this metric is $[0, 1]$, and effective cannulation will produce ocScore values closer to 1; for further details, see Liu et al. [2021].

Building upon our previous research [Liu et al., 2021; Petersen et al., 2022], this work aims to identify salient process metrics from a high-dimensional feature space that influence the probability of successful cannulation (i.e., the probability of attaining stable flashback) *and* the quality of cannulation (i.e., ocScore). In this article, we develop a shared random parameter model under the Bayesian framework to jointly model the ocScore and stable flashback. These two outcomes are correlated both implicitly and by design, and we accommodate this dependence through the use of a shared random effect that acknowledge subject-specific tendencies in cannulation performance. The proposed methodology is motivated by and applied to a study of cannulation skill assessment from the sensor-based simulator data.

The remainder of this article is organized as follows. In Section 3.2, we develop the proposed methodology, including prior model specifications and data augmentation steps used to construct an efficient posterior sampling algorithm. Section 3.3 reports the results of a simulation study conducted to assess the performance of the proposed approach. Section 3.4 presents the analysis results for the motivating study. We conclude with a summary discussion in Section 3.5. Additional details and simulation results are provided in Appendix B.2.

3.2 Methodology

3.2.1 Model Notation

In what follows, we outline the prominent features of our proposed model. To this end, let $Y_{ij} \in [0, 1)$ be the ocScore for the ith subject during their jth trial (cannulation) after k_{ij} attempts, for $j = 1, \ldots, n_i$, and $i = 1, \ldots, m$. To relate the observed ocScore data to the available process

features, we posit the following linear mixed model:

$$Y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \gamma_i + \epsilon_{ij}, \tag{3.1}$$

where $\boldsymbol{x}_{ij} = (1, x_{ij1}, x_{ij2}, \ldots, x_{ijP})'$ is a $(P+1)$-dimensional vector of covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)'$ is the corresponding vector of regression coefficients with global intercept β_0, and ϵ_{ij} denotes random error. We assume that the errors are independent and identically distributed as normal random variables; i.e., $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$. The second term in model (3.1) is a subject-specific random effect, denoted by γ_i, entered into the model to account for the heterogeneity across subjects. Herein, the distribution for the random effects is taken to be $\gamma_i \overset{\text{iid}}{\sim} N(0, \sigma_\gamma^2)$.

To model the probability of success, let $S_{ij(k)}$ be the binary indicator of stable flashback for the ith individual's jth trial (cannulation) on their kth attempt, $k = 1, \ldots, k_{ij}$. To relate the observed stable flashback data to the available process features, we propose the following generalized linear mixed model:

$$\pi_{ij} \coloneqq P\left(S_{ij(k)} = 1 \mid \boldsymbol{z}_{ij}\right) = \left[1 + \exp(-\nu_{ij})\right]^{-1}, \quad \text{with } \nu_{ij} \coloneqq \boldsymbol{z}_{ij}'\boldsymbol{\alpha} + \zeta\gamma_i, \tag{3.2}$$

where we've used the logit link function to relate the linear predictor ν_{ij} to the probability of stable flashback. The first term of model (3.2) consists of a $(Q+1)$-dimensional vector of covariates, $\boldsymbol{z}_{ij} = (1, z_{ij1}, z_{ij2}, \ldots, z_{ijQ})'$, specific to the jth trial of the ith subject, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_Q)'$ is the corresponding vector of regression coefficients. The second term consists of the same subject-specific random effect, γ_i, that is shared with model (3.1), while ζ is the association parameter that can take into account the interdependence between ocScore and success. In particular, ζ allows the subject-specific propensity towards higher ocScores to contribute to the subject-specific propensity towards obtaining or failing to obtain stable flashback. We incorporate this association parameter because we expect the random effects of (3.2) to potentially have different magnitude, scale, and/or sign than the random effects of (3.1).

Within a prospective cannulation trial, participants are allowed to attempt cannulation as many times as needed or until they give up, resulting in k_{ij} attempts. Hence, success (i.e., attaining stable flashback) is observed as an attempt-based metric; that is, for each attempt during a trial, we observe success or failure. However, we model success as a trial-based metric that is a function of the

number of attempts. In particular, if the first stable flashback is obtained on the kth attempt, then $k_{ij} = k$ and $S_{ij(k)} = S_{ij} = 1$. If no stable flashback is obtained after k attempts and the participant gives up, then $k_{ij} = k$ and $S_{ij(k)} = S_{ij} = 0$.

3.2.1.1 Sparse Principal Component Analysis

The process features used in model (3.2) are highly correlated with one another. This presence of multicollinearity can lead to highly unreliable estimates for the regression coefficients and inflated standard errors, particularly when modeling a binary outcome. Principal component analysis (PCA) mitigates some of the issues associated with multicollinearity and the estimation of regression parameters by transforming the original feature space into a collection of orthogonal features; namely, principal components [Jolliffe, 2002]. The resulting principal components are linear combinations of *all* features in the original feature space. While coefficient estimation using principal components instead of the original feature space leads to biased estimates, the variability of the estimates themselves can be greatly reduced.

With each principal component being a linear combination of *all* features, it is difficult to interpret the estimation results and identify variables of importance. As a remedial measure, we will apply a sparse principal component analysis (SPCA) technique [Zou et al., 2006] to the success model (3.2). Zou et al. [2006] showed that PCA can be formulated as a regression-type optimization problem. SPCA produces modified principal components with sparse loadings by imposing the lasso constraint on the corresponding PCA regression coefficients, so that each principal component is a linear combination of only a subset of the covariates [Zou et al., 2006]. This leads to more interpretable principal component loadings that will assist in identifying significant covariates. The `spca` function in the `elasticnet` R package [Zou and Hastie, 2020] will output the modified principal component loadings, where sparsity can be enforced by specifying the number of principal components to use and the number of non-zero elements in each of the loadings. In what follows, we outline the reformulation of model (3.2) under SPCA.

For notational convenience, let $\boldsymbol{z}_{ij} \coloneqq (1, \boldsymbol{z}_{ij}^*)'$, where $\boldsymbol{z}_{ij}^* = (z_{ij1}, \ldots, z_{ijQ})'$ is the subset of $\boldsymbol{z}_{ij}$ that excludes the first entry associated with the global intercept. We assume that the predictor variables $\boldsymbol{z}_{ij}^*$ have been standardized. Define $\boldsymbol{V}^*$ to be the $Q \times L$ matrix whose columns are the sparse loadings for the first L modified principal components of the predictor variables, $\boldsymbol{z}_{ij}^*$, across all trials and participants, where $L \leq Q$. To accommodate the intercept term, define

$$\boldsymbol{V} = \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{V}^* \end{pmatrix}$$

to be the $(L+1) \times (L+1)$ block diagonal matrix. Then, let $\boldsymbol{c}_{ij} = \boldsymbol{z}_{ij}\boldsymbol{V} \coloneqq (1, \boldsymbol{c}^*_{ij})'$ denote the derived input vector, where $\boldsymbol{c}^*_{ij} = \boldsymbol{z}^*_{ij}\boldsymbol{V}^*$ is the vector of the first L modified principal component scores for the ith participant's jth trial. With this, we can replace the linear predictor expression of model (3.2) with the following (reduced) SPCA transformation:

$$\nu_{ij} = \boldsymbol{c}'_{ij}\boldsymbol{\theta} + \zeta\gamma_i, \tag{3.3}$$

where $\boldsymbol{\theta} = \boldsymbol{V'\alpha}$ is the vector of $(L+1)$ regression coefficients we wish to estimate. By replacing the original covariates with the modified principal components with sparse loadings, we can obtain much more stable estimates of the original model coefficients, $\boldsymbol{\alpha}$, without hindering our ability to identify significant predictors.

3.2.2 Prior Specifications

To facilitate both parameter estimation and inference, we cast our model into the Bayesian paradigm and develop a Markov chain Monte Carlo (MCMC) posterior sampling algorithm. The first step in this process involves specifying prior distributions for all unknown parameters. Given the large number of covariates, priors are chosen to regularize the estimation of parameters. In particular, the priors for the regression coefficients are designed to "shrink" unimportant variables towards zero. In what follows, we briefly expand on these specifications.

To aid in variable selection, we adopt the generalized double Pareto shrinkage prior [Armagan et al., 2013] for all regression coefficients with the exception of the intercepts; i.e., for the ocScore model coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)'$, we specify

$$\begin{aligned} \beta_0 &\sim N(0, \sigma^2_\epsilon \tau_0) \\ \beta_p &\sim GDP(\sigma_\epsilon b_\delta / a_\delta,\ a_\delta), \text{ for } p = 1, \ldots, P, \end{aligned}$$

and, for the stable flashback SPCA model coefficients $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_L)'$, we specify

$$\theta_0 \sim N(0, \rho_0)$$
$$\theta_l \sim GDP(b_\lambda/a_\lambda,\ a_\lambda), \text{ for } l = 1, \ldots, L,$$

where $GDP(\sigma_\epsilon b_\delta/a_\delta,\ a_\delta)$ and $GDP(b_\lambda/a_\lambda,\ a_\lambda)$ refer to generalized double Pareto distributions. Under these prior specifications, setting τ_0 and ρ_0 large provides vague priors for the intercepts β_0 and θ_0, respectively, while the hyperparameters $a_\lambda,\ b_\lambda > 0$ and $a_\delta,\ b_\delta > 0$ govern the amount of shrinkage. A typical default specification is to set both hyperparameters equal to 1, leading to Cauchy-like tail behavior which is known to have desirable Bayesian robustness properties [Armagan et al., 2013]. We adopt this default specification for the stable flashback model; i.e., $a_\lambda = b_\lambda = 1$. On the other hand, the ocScore model has a high-dimensional feature space. It could be very dense or very sparse, and setting the hyperparameters to their default specification could be very restrictive [Armagan et al., 2013]. As an alternative, we choose hyper-priors to allow the data to inform us about the values of a_δ and b_δ. In particular, we use $p(a_\delta) = 1/(1 + a_\delta)^2$ and $p(b_\delta) = 1/(1 + b_\delta)^2$ to correspond to generalized Pareto hyper-priors with location parameter 0, scale parameter 1 and shape parameter 1 [Armagan et al., 2013].

Next, we place an inverse gamma prior on the variance component of the subject-specific random effects γ_i; i.e., we specify $\sigma^2_\gamma \sim Inv - Gamma(a_\gamma, b_\gamma)$. This specification is common among the literature and it leads to a proper posterior [Seltzer et al., 1996]. For the random effects association parameter, ζ, in the SPCA model (3.3), we specify the following conjugate Normal prior: $\zeta \sim N(0, \sigma^2_\zeta)$, where σ^2_ζ is set large to provide a vague prior for ζ. Finally, we place a uniform prior on the variance component of the random error ϵ_{ij}; i.e., we specify $p(\sigma_\epsilon) \propto 1/\sigma_\epsilon$.

Based on the prior specifications outlined above, we develop a Markov chain Monte Carlo (MCMC) sampling algorithm which facilitates both posterior estimation and inference. In what follows, we provide a brief overview of this algorithm and its construction.

3.2.3 Data Augmentation and Posterior Sampling

With ease of implementation and computational efficiency in mind, herein we outline the construction of a posterior sampling algorithm that consists solely of Gibbs steps [Gelman et al., 2013]. To accomplish this, we consider a two-stage data augmentation process. The first stage

introduces carefully constructed Pólya-Gamma latent random variables so that the logistic function can be hierarchically expressed as a scale mixture of normals, where the mixing distribution is Pólya-Gamma [Polson et al., 2013]. The second stage decomposes the generalized double Pareto shrinkage prior as a scale mixture of normals [Armagan et al., 2013]. For the specific details of this two-stage data augmentation process, see Appendix B.1. This data augmentation scheme leads to the following full conditionals:

$$
\begin{aligned}
\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{\gamma}, \sigma_\epsilon^2, \boldsymbol{\tau} &\sim N(\boldsymbol{\mu_\beta},\ \boldsymbol{\Sigma_\beta}) \\
\boldsymbol{\theta} \mid \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \zeta, \boldsymbol{\rho} &\sim N(\boldsymbol{\mu_\theta},\ \boldsymbol{\Sigma_\theta}) \\
\zeta \mid \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\zeta^2 &\sim N(\mu_\zeta,\ \Sigma_\zeta) \\
\sigma_\gamma^2 \mid \boldsymbol{\gamma} &\sim Inv-Gamma(a_\gamma^*,\ b_\gamma^*) \\
\sigma_\epsilon^2 \mid \boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\gamma} &\sim Inv-Gamma(a_\epsilon^*,\ b_\epsilon^*) \\
\omega_{ij} \mid \boldsymbol{\theta}, \gamma_i, \zeta &\sim PG(k_{ij},\ \nu_{ij}) \\
\tau_p \mid \beta_p, \delta, \sigma_\epsilon^2 &\sim Inv-Gaussian(\mu_{\tau_p},\ \delta_p^2), \quad p = 1, \ldots, P \\
\delta_p \mid \beta_p, \sigma_\epsilon^2 &\sim Gamma(a_{\delta_p}^*,\ b_{\delta_p}^*), \quad p = 1, \ldots, P \\
\rho_l^{-1} \mid \theta_l, \lambda_l &\sim Inv-Gaussian(\mu_{\rho l},\ \lambda_l^2), \quad l = 1, \ldots, L \\
\lambda_l \mid \theta_l &\sim Gamma(a_{\lambda_l}^*,\ b_{\lambda_l}^*), \quad l = 1, \ldots, L,
\end{aligned}
$$

where the specific form of the parameters of these distributions are given in Appendix B.1. Recall the hyper-priors specified for a_δ and b_δ discussed in Section 3.2.2. The corresponding full conditional posterior distributions are given in Appendix B.1. For sampling, we use an embedded griddy Gibbs sampling scheme; for details, see Armagan et al. [2013].

These full conditionals were used to construct an MCMC algorithm in the usual manner [Gelman et al., 2013]. Recall from Section 3.2.1.1, $\boldsymbol{\theta}$ is the vector of regression coefficients associated with the SPCA model (3.3) for stable flashback. At each iteration, we can use the draw from the posterior of $\boldsymbol{\theta}$ to recover a draw from the posterior of $\boldsymbol{\alpha}$ in order to conduct posterior inference on the fixed effects from the stable flashback model (3.2). In the following section, we conduct a numerical study to evaluate the performance of our posterior sampling algorithm.

3.3 Simulation Study

We simulate data from the following shared random effects model:

$$Y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \gamma_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N(0, 1.25) \tag{3.4}$$

$$\pi_{ij} \coloneqq P\left(S_{ij(k)} = 1 \mid \boldsymbol{z}_{ij}\right) = \left[1 + \exp(-\nu_{ij})\right]^{-1}, \text{ with } \nu_{ij} \coloneqq \boldsymbol{z}_{ij}'\boldsymbol{\alpha} + \zeta\gamma_i, \tag{3.5}$$

where $\gamma_i \overset{\text{iid}}{\sim} N(0, .95)$ and $\zeta = 0.85$, for $j = 1, \ldots, 16$; $i = 1, \ldots, 50$; yielding 800 total observations of both outcome measures, Y_{ij} and $S_{ij(k)}$.

In model (3.4), we let $\boldsymbol{x}_{ij} = (1, x_{ij1}, \ldots, x_{ij400})'$, where the P=400 covariates of $\boldsymbol{x}_{ij}$ are independent, and each covariate is generated as normally distributed with mean zero and variance 0.10^2; i.e., $x_{ijk} \sim N(0, 0.10^2)$, for $k = 1, \ldots, 400$. The corresponding model coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{400})'$, are defined as follows. We set β_0=0.45, the first 80 fixed effects are non-zero, and the remaining 320 are zero. In particular, β_p=1 for $1 \leq p \leq 40$; β_p=-1 for $41 \leq p \leq 80$; and β_p=0 for $81 \leq p \leq 400$.

In model (3.5), we let $\boldsymbol{z}_{ij} = (1, z_{ij1}, \ldots, z_{ij10})'$, where the vector of the $Q = 10$ covariates $(z_{ij1}, \ldots, z_{ij10})'$ is generated as normally distributed with mean one and variance one, and a covariance structure specified as follows. The first 5 covariates $(z_{ij1}, \ldots, z_{ij5})'$ are correlated such that the pairwise correlations between the components of $(z_{ij1}, \ldots, z_{ij5})'$ are $r^{|k-l|}$ with r=0.90, for $k, l = 1, \ldots, 5$. The remaining 5 covariates $(z_{ij6}, \ldots, z_{ij10})'$ are specified to be independent and identically distributed, and are independent of the first 5 covariates. The corresponding, true model coefficients are defined to be $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{10})' = (1.05, -1, 0, 1, 1, 0, 0, 0, 0, 0, 0)'$. To address the multicollinearity that is present in the covariates $\boldsymbol{z}_{ij}$, we will conduct posterior estimation and inference using an SPCA model of form (3.3) in place of (3.5), where each of the L=10 modified principal component loadings $\boldsymbol{v_l}$ have 5 non-zero elements, for $l = 1, \ldots, 10$. This data simulation process was repeated for 500 independent data sets, and estimation results were averaged over these 500 data sets. We used our posterior sampling algorithm to draw 2,500 samples after a burn-in of 2,500 samples. Trace plots were used to assess the convergence of the MCMC chains.

Tables 7 - 10 in Appendix B.2 report the estimation results for the intercept and the 400 fixed effects associated with the covariates in model (3.4). For each parameter, the values of 'Bias' and 'SSD' are the empirical bias and standard deviation, respectively, of the 500 posterior mean

estimates; the value of 'ESE' is the averaged estimated posterior standard deviation; and the value of 'CP95' is the empirical coverage probability of the nominal 95% equal-tail credible interval. Table 3.1 summarizes how successful our approach was at correctly identifying significant predictors (i.e., truly non-zero regression coefficients) in model (3.4). It reports the average proportion of times the covariates, corresponding to truly zero and non-zero regression coefficients, were identified as statistically significant. The results in these tables indicate that our approach can identify the truly significant predictor variables in a high-dimensional setting (Table 3.1) while producing accurate, reliable estimates of these parameters. Moreover, the empirical bias is small relative to the true value of the corresponding parameter, and SSD and ESE values are relatively close in agreement for each parameter (Tables 7 - 10). Notice that the empirical coverage probabilities of the 95% equal-tail credible intervals are slightly below the nominal level for nonzero coefficients and slightly above for zero coefficients. This is to be expected from our shrinkage prior specification with the non-default hyperparameter setting. Indeed, this prior specification says that, *apriori*, we believe that the distribution is concentrated around zero. This 'drags' the posterior estimates towards zero and the shrinkage penalization pushes the credible intervals towards zero as we inject more information.

Table 3.1: Average proportion of times regression coefficients β_p in model (3.4) deemed important.

	Avg Proportion
$\boldsymbol{\beta} \neq \mathbf{0}$ (81)	0.998
$\boldsymbol{\beta} = \mathbf{0}$ (320)	0.011

Table 11 in Appendix B.2 reports the estimation results for the intercept and the 10 fixed effects associated with the covariates in model (3.5). For each parameter, the value of 'Bias' and 'SSD' is the empirical bias and standard deviation of the 500 posterior mean estimates; the value of 'ESE' is the averaged estimated posterior standard deviation; and the value of 'CP95' is the empirical coverage probability of the nominal 95% equal-tail credible interval. Table 3.2 summarizes how successful our approach (along with the proposed SPCA technique) was at correctly identifying significant predictors (i.e., truly non-zero regression coefficients) in model (3.5). It reports the average proportion of times the covariates, corresponding to truly zero and non-zero regression coefficients, were identified as statistically significant. The results in these two tables indicate that our approach with the SPCA technique can identify the truly significant predictor variables (Table 3.2) and produce stable and reliable estimates, even in the presence of multicollinearity. Moreover,

the empirical bias is small relative to the true value of the corresponding parameter, SSD and ESE values are relatively close in agreement for each parameter, and the empirical coverage probabilities of the 95% equal-tail credible intervals are roughly at the nominal level (Table 11).

Table 3.2: Average proportion of times regression coefficients α_q in model (3.5) deemed important.

	Avg. Proportion
$\boldsymbol{\alpha} \neq \mathbf{0}$ (4)	0.941
$\boldsymbol{\alpha} = \mathbf{0}$ (7)	0.042

Overall, the results of this numerical study suggest that our proposed approach performs well; our algorithm successfully identifies the truly influential features in a high-dimensional setting with the presence of multicollinearity, and has the ability to produce accurate, stable estimates of model parameters. Therefore, we conclude that the proposed approach is appropriate for analyzing the motivating data.

3.4 Analysis of Cannulation Skill Data

Ethics approval for this study was provided by the Institutional Review Boards (IRB) of Clemson University and Prisma Health (IRB number: Pro00064701). This study examine data collected from 52 healthcare professionals, with some degree of clinical experience in cannulation, who were recruited at a regional ESKD meeting. Upon providing informed consent to participate in the experiment, each participant was asked to perform 16 trials on the simulator to allow for different scenarios. Each of the four fistulas were presented four times, and the order of fistulas and their intensities were randomized. One of the two simulated skin thicknesses were used to conduct the first 8 trials, and the latter 8 trials were conducted using the other skin thickness. Each trial, or prospective cannulation, consists of two fundamental parts: the first is palpation, where subjects identify the location and orientation of the fistula; then, participants insert the needle to obtain blood flashback. Data on 83 of the trials were excluded due to testing purposes or unavailability of sensor data. Therefore, the dataset comprised of a total of 670 trials from 45 participants was identified for analysis. The median number of trials per participant is 15 and the typical number of insertion attempts per trial is one, as shown in Table 3.3. The average ocScore was 0.39, and stable flashback was attained in 84% of the cannulation trials (presented in Table 3.3).

Tables 12 - 14 in Appendix B.3 list the process metrics that will be used as covariates in the

Table 3.3: Trial and outcome characteristics of individuals used in the analysis. ocScore is a continuous outcome that takes values between 0 and 1. Stable flashback is a binary indicator that takes value 1 to denote a stable flashback and 0 otherwise.

	Mean	Median	Range
Trials	14.89	15	11-16
Attempts	1.23	1	1-7
ocScore	0.39	0.40	0-1
Stable Flashback	0.84	1	0-1

model for ocScore. This results in the estimation of P=449 fixed effects in the ocScore component of the model. Table 15 in Appendix B.3 lists the process metrics that will be used as covariates in the model for stable flashback. This leads to the estimation of Q=16 fixed effects in the stable flashback component of the model. To address the presence of multicollinearity among the features used in the prediction of stable flashback, we will conduct posterior estimation and inference using the SPCA model (3.3) in place of (3.2), where only the first L=5 modified principal components are incorporated. Sparsity is enforced as the number of non-zero elements in each of the loadings. In particular, the first two modified principal component loadings are restricted to having 5 non-zero elements, the third loading restricted to 4 non-zero elements, and the fourth and fifth loadings are restricted to having 3 and 2 non-zero elements, respectively. For the priors outlined in Section 3.2.2, we take $\tau_0 = \rho_0 = 1000$ to specify vague priors on the global intercepts β_0, θ_0, $\sigma^2_\zeta = 1000$ to specify a vague prior on the random effects association parameter ζ, and we set $a_\gamma = b_\gamma = 0.01$. These hyperparameter values were chosen so to produce uninformative, proper prior specifications. For our posterior sampling algorithm, we retain 2,500 MCMC iterates after a burn-in of 2,500 samples. Convergence of the MCMC chains were assessed in the usual manner; i.e., trace plots.

Table 3.4 summarizes the estimation results for the shared subject-specific random effect, and its association parameter. The association parameter ζ is significantly different from zero (i.e., its estimated credible interval does not contain zero) and the posterior mean estimate of ζ is a relatively large, positive value (Table 3.4). This indicates that there is a strong, positive association between a participant's ocScore and their cannulation success. In particular, as the value of ocScore increases by one unit, the log-odds of attaining stable flashback increases. This is to be expected because ocScore was devised as a function of the indicator of stable flashback in which ocScore decreases when stable flashback is not attained.

Table 3.5 reports the covariates of the ocScore component of the model that were deemed to

Table 3.4: Analysis results: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95) for subject-specific random effect and association parameter.

Param.	**Est**	**ESE**	**CI95**
σ^2_γ	0.01	0.00	(0.01, 0.02)
ζ	8.99	1.32	(6.50, 11.75)

be significant. Of the 449 fixed effects for ocScore, 22 were found to be important by the model (i.e., their estimated credible intervals did not contain zero). Table 3.6 reports the covariates of stable flashback that were deemed to be significant. Of the 16 fixed effects, 8 were deemed to be important by the stable flashback model. Both Tables 3.5 and 3.6 summarize the significant features by reporting the estimated posterior means (point estimate of the effect), estimated posterior standard deviations (measure of uncertainty), and 95% equal-tailed credible intervals.

Table 3.5: Analysis results: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95) for the significant fixed effects for ***ocScore***.

Variable	**Est**	**ESE**	**CI95**	**Variable**	**Est**	**ESE**	**CI95**
Intercept	0.39	0.02	(0.36, 0.43)	total_forcedctf2	0.02	0.01	(0.01, 0.03)
ldljV	0.06	0.01	(0.03, 0.09)	zdctf10	0.02	0.01	(0.00, 0.04)
avgV	-0.03	0.02	(-0.07, -0.00)	vdctf9	-0.02	0.01	(-0.03, -0.01)
zdctf7	0.03	0.01	(0.02, 0.05)	sparcV	0.02	0.01	(0.00, 0.03)
Nzfftf9	0.03	0.01	(0.01, 0.05)	beta_2	-0.02	0.01	(-0.03, -0.01)
Nzdctf8	-0.03	0.01	(-0.04, -0.01)	Nyfftf1	0.02	0.01	(0.00, 0.03)
Nxfftf5	0.03	0.01	(0.01, 0.04)	Nzfftf2	0.02	0.01	(0.00, 0.04)
Alphadctpow10	0.03	0.01	(0.00, 0.06)	total_forcefftpow2	-0.02	0.01	(-0.03, -0.00)
xfftf7	-0.02	0.01	(-0.04, -0.00)	dAlphafftf10	0.01	0.01	(0.00, 0.02)
Nzdctf9	-0.02	0.01	(-0.04, -0.01)	vfftf6	0.01	0.01	(0.00, 0.03)
bf_ldljV	-0.02	0.01	(-0.04, -0.01)	Nxdctf4	-0.01	0.01	(-0.03, -0.00)

The metrics *ldljV, sparcV*, and *LDLJ_S* are all measures of motion smoothness, where larger values are an indication of smoother motion or reduced "shakiness". From Tables 3.5 and 3.6, we can conclude that smooth, controlled motion increases the log-odds of success and improves the quality of cannulation. Moreover, as the average velocity of motion increases, the quality of cannulation lessens (as indicated by the negative sign for the effect of *avgV* in Table 3.5). In addition, the metrics *beta02, beta_0*, and *beta_2* describe the relative needle orientation (angular position) at various phases of the cannulation task. In particular, *beta_0* and *beta_2* are the needle angles measured instantaneously at the point of insertion and after flashback, respectively. *beta02* is the needle angle approximated

Table 3.6: Analysis results: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95) for the significant fixed effects for ***stable flashback***.

Variable	**Est**	**ESE**	**CI95**
Intercept	1.65	0.19	(1.29, 2.03)
sparcV	0.60	0.06	(0.49, 0.71)
ldljV	0.54	0.05	(0.44, 0.64)
LDLJ_S	0.22	0.04	(0.15, 0.30)
avg_alpha_S	-0.14	0.07	(-0.29, -0.00)
palp_force_range	-0.11	0.06	(-0.23, -0.00)
a2	-0.10	0.03	(-0.16, -0.05)
beta02	-0.08	0.03	(-0.14, -0.03)
beta_0	-0.08	0.03	(-0.14, -0.03)
beta_2	-0.08	0.03	(-0.13, -0.03)

by the needle tip location at two distinct points in time: point of insertion, and after flashback. From Table 3.5, as the angle of the needle increases after flashback (i.e., as *beta_2*) increases, the quality of cannulation depreciates. However, if the angular position of the needle orientation steepens during *any* phase of the cannulation task, the log-odds of successfully cannulating decreases (Table 3.6). Further, the hemodialysis cannulation procedure requires appropriate palpation for locating the fistula, and *palp_force_range* measures the range of forces applied during palpation. We see that, as palpation force increases, the log-odds of success decreases (Table 3.6). People who are not as skilled in cannulation tend to palpate with more force and, from our results, they are less likely to successfully complete cannulation compared to those who palpate with less force. Most of the remaining features summarized in Tables 3.5 and 3.6 do not have useful interpretations that help identify what trainees could improve upon.

3.5 Discussion

Motivated by and applied to a study of simulation-based cannulation skill assessment, we have developed a shared random parameter model to jointly model two objective outcome measures of simulation-based cannulation skill and identified salient process features from a high-dimensional feature space that influence the probability of successful cannulation (i.e., the probability of attaining stable flashback) and the quality of cannulation (i.e., ocScore). While the two outcomes are correlated by formulation (ocScore is functionally dependent on stable flashback), they are also correlated implicitly: subjects that consistently obtain stable flashback are inherently more likely to have a

better quality cannulation compared to those who do not. The shared random effects allow for an implicit correlation between a participant's ability to attain stable flashback and the quality of their cannulation.

There are a few limitations worth noting. Sensor-based metrics that are undefined if a participant fails to complete various phases of the cannulation task were excluded from the analysis. It is of interest to accommodate these metrics with potentially informative missingness so that we can examine the effects of all of the available features. In addition, the cannulation simulator is made up of artificial materials for various components such as skin, fistulas, tissue, etc., and thus has restrictions in realism and functionality.

To conclude, by identifying errors through the use of sensor-based metrics, simulators improve upon conventional skill assessment and reduce subjectivity. The main advantage of simulators is their ability to provide objective feedback to allow for a fine-grained assessment of skill and a more consistent, complete evaluation based on measurements unaffected by subjective biases. The results from this study suggest that the implementation of simulator-based training will lead to the improvement in end-stage kidney disease patient outcomes.

Chapter 4

Bayesian Additive Regression Trees for Group Testing Data

4.1 Introduction

When a high volume of specimens (such as blood, urine, swabs, etc.) need to be screened for the presence of a disease, it is not always feasible to do individual testing. Group testing, also known as pooled testing, pulls the individual specimens into groups or pools to get an overall positive or negative test response for each pool. Robert Dorfman [Dorfman, 1943] conceptualized the idea of group testing during World War II to screen US soldiers for syphilis. In most group testing protocols, if a pooled specimen tests negatively, then all contributing individuals are declared to be disease free at the expense of a single diagnostic test. In contrast, if a pooled specimen tests positively, the pool is resolved algorithmically to determine which individuals are positive. Dorfman's idea to pool individual specimens has since become a mainstream approach to screen large populations for multiple diseases because of its ability to provide substantial cost savings when compared to individual testing. It is used to screen a variety of infections including HIV, HCV, and HBV [Westreich et al., 2008; Krajden et al., 2014; Sarov et al., 2007; Kleinman et al., 2005], chlamydia and gonorrhea [Lewis et al., 2012], influenza [Van et al., 2012], the Zika virus [Saá et al., 2018], and COVID-19 [Bish et al., 2021; Torres et al., 2020]. Group testing also arises in other applications, such as animal disease testing [Dhand et al., 2010], environmental monitoring [Heffernan et al., 2014],

and drug discovery [Hughes-Oliver, 2006].

Statistical research in group testing focuses on either estimation or case identification problems. The former, which is the focus of this article, involves using pooled testing outcomes and individual-level covariates (e.g., age, race, presence of symptoms, etc.) to develop regression methods that model the probability of disease for individuals. Noteworthy research in the development of group testing regression methods include parametric approaches by Vansteelandt et al. [2000], Huang and Tebbs [2009], and Chen et al. [2009] as well as semiparametric and nonparametric approaches by Delaigle and Meister [2011], Delaigle et al. [2014], and Delaigle and Hall [2015]. A limitation of these regression methods is that only the initial (master) pool responses are used in the estimation of the corresponding models. That is, if individuals residing in positive master pools are retested, their subsequent responses are not utilized. If additional retesting responses are available, including them in the analysis can improve one's inference for the covariate effects.

Far fewer regression methods are available that can incorporate this extra information. Notably, McMahan et al. [2017] proposed a Bayesian approach for the regression analysis of group testing data within a generalized linear model (GLM) framework. The strengths of this approach are 3-fold. First, this approach can seamlessly incorporate all of the testing data collected from any group testing protocol (including retesting responses); second, it can estimate assay accuracy probabilities along with the regression coefficients (whereas previous regression methods require the assay accuracy probabilities to be known); and third, it can naturally incorporate historical information about disease prevalence and assay performance. More recently, Liu et al. [2021] expanded on the Bayesian methodology of McMahan et al. [2017] and developed a generalized additive regression model for group testing data that relaxes the linearity assumptions of conventional methods to allow for nonlinear covariate effects. However, the main limitation of Liu et al. [2021] is that the additivity assumption precludes interactions among covariates unless manually added and, with many variables, important interactions could be missed.

In this article, we propose a Bayesian additive regression trees (BART) modeling framework to estimate regression models using group testing data. BART is a powerful machine learning technique for predictive modeling that employs a nonparametric, tree-based approach. A major advantage of BART is its Bayesian structure and its ability to capture/quantify uncertainty in model estimates. The proposed framework retains the strengths of McMahan et al. [2017], and addresses the limitations of Liu et al. [2021] to allow for a more flexible and robust modeling approach.

Indeed, BART automates the detection of nonlinear relationships and interactions among predictor variables to reduce researchers' discretion [Chipman et al., 2010]. This leads to increased accuracy and precision, and allows for a better understanding of any complex, nonlinear effects.

The remainder of this article is organized as follows. In Section 4.2, we introduce the proposed BART model and describe modeling assumptions. In Section 4.3, we describe the data augmentation steps that facilitate our Bayesian framework and introduce our posterior sampling algorithm. In Section 4.4, we present the results of multiple simulations to assess the performance of our proposed method under a variety of settings for group testing protocols. In Section 4.5, we present analysis results for Iowa chlamydia data to illustrate the proposed technique. Finally, in Section 4.6, we conclude with a summary discussion and describe future research.

4.2 Notation and Model Formulation

Consider a setting in which group testing is used to screen N individuals for a binary characteristic, such as disease status. Let $\widetilde{Y}_i$, for $i = 1, ..., N$, denote the true disease status of the ith individual, with the usual convention that $\widetilde{Y}_i = 1$ denotes that the individual is truly positive and $\widetilde{Y}_i = 0$ otherwise. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iQ})'$ denote a vector of covariates observed for the ith individual. For ease of exposition, we aggregate the individuals' true infection statuses and covariates as $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_N)'$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$, respectively. For modeling purposes, we assume that the individuals' true disease statuses are conditionally independent given the individual-level covariate information, and that the relationship between $\widetilde{Y}_i$ and $\boldsymbol{x}_i$ is given by

$$\Phi^{-1}\left(P(\widetilde{Y}_i = 1 \big| \boldsymbol{x}_i)\right) = f(\boldsymbol{x}_i), \tag{4.1}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable (i.e., the probit link function), and $f(\cdot)$ is an unknown function. To model this relationship, we consider approximating $f(\cdot)$ by an ensemble of K regression trees; i.e., we approximate (4.1) by a sum-of-trees model, where

$$f(\boldsymbol{x}_i) \approx \eta(\boldsymbol{x}_i) := \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k, M_k). \tag{4.2}$$

The number of regression trees K is (typically) fixed. T_k is the kth regression tree structure consisting of a set of interior nodes, a set of b_k terminal nodes, and the decision rules connecting the interior nodes to the terminal nodes. The interior node decision rules are binary splits based on the single predictors in the form of $\{\boldsymbol{x} \in A\}$ versus $\{\boldsymbol{x} \notin A\}$, where A is a subset of the range of $\boldsymbol{x}$. These decision rules provide information on which covariate to split on and the associated cutoff value. $M_k = (\mu_{1k}, \ldots, \mu_{b_k k})'$ denotes the b_k-dimensional vector of parameters associated with the terminal nodes of T_k. Given T_k and M_k, the function $g(\boldsymbol{x}_i; T_k, M_k)$ outputs μ_{tk} if $\boldsymbol{x}_i$ is assigned to the t-th terminal node based on the interior node decision rules. Note that, based on the structure of the kth tree, $g(\boldsymbol{x}_i; T_k, M_k)$ could depend on a single component or multiple components of $\boldsymbol{x}_i$. Hence, each $\mu_{tk} \in M_k$ could represent a main effect or an interaction effect. In this way, $g(\boldsymbol{x}_i; T_k, M_k)$ can aptly account for many features; e.g., nonlinear effects and interactions of varying orders. As the number of trees K increases, the predictive performance of BART dramatically increases until leveling off [Chipman et al., 2010]. Thus, if BART is used for prediction or to estimate the unknown $f(\cdot)$, it is important to avoid choosing K too small. Chipman et al. [2010] recommends setting K=200 as the default number of trees, as it has been shown that BART yielded excellent predictive performance under this choice for K.

To illustrate the main idea of a sum-of-trees model, consider an example with K=2 trees and Q=3 covariates. Suppose we are given the two trees in Figure 4.1. Each tree uses two predictors to split the data into subgroups; the first tree of Figure 4.1 (k=1) uses x_{i1} and x_{i2}, while the second tree (k=2) uses x_{i3} and x_{i2}. For each tree, each $\boldsymbol{x}_i$ value is assigned to a single terminal node by following a sequence of decision rules at each interior node from top to bottom where it is then assigned a parameter value associated with that terminal node. Consider the hypothetical data from 5 subjects given in Table 4.1. We can see that the quantity that is being 'summed' in the final sum-of-trees model for the ith subject is the terminal node parameter value that each tree structure assigns to the ith subject.

Table 4.1: The values of $\sum_{k=1}^{2} g(\boldsymbol{x}_i; T_k, M_k)$ from the regression trees in Figure 4.1.

i	x_{i1}	x_{i2}	x_{i3}	$g(\boldsymbol{x}_i; T_1, M_1)$	$g(\boldsymbol{x}_i; T_2, M_2)$	$\sum_{k=1}^{2} g(\boldsymbol{x}_i; T_k, M_k)$
1	56	110	-13	μ_{31}	μ_{12}	$\mu_{31} + \mu_{12}$
2	27	173	-3	μ_{21}	μ_{32}	$\mu_{21} + \mu_{32}$
3	41	94	5	μ_{11}	μ_{22}	$\mu_{11} + \mu_{22}$
4	30	213	-9	μ_{21}	μ_{12}	$\mu_{21} + \mu_{12}$
5	48	168	39	μ_{31}	μ_{32}	$\mu_{31} + \mu_{32}$

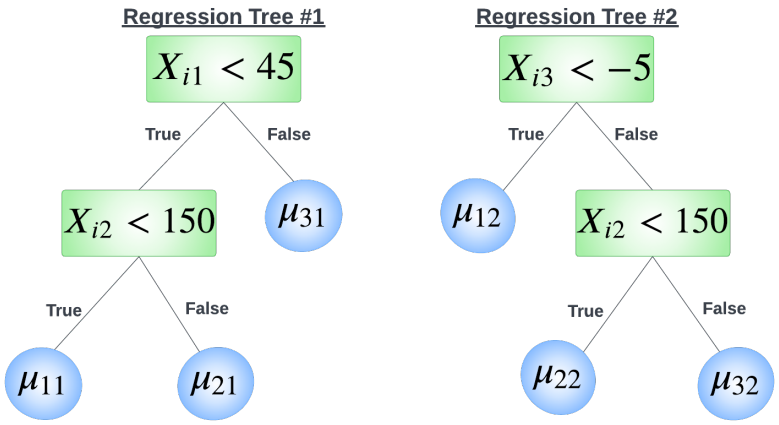

Figure 4.1: Illustrating the sum of regression trees using a simple two regression tree example.

If the individuals' true disease statuses were observed, we could fit the sum-of-trees model via standard statistical software; e.g., `Bayestree`, `bartMachine`, and `bart` [Chipman and McCulloch, 2016; Kapelner and Bleich, 2022; McCulloch et al., 2021]. However, in the group testing setting, the individual disease statuses are likely to be obscured by the testing protocol and the testing responses (on pools and individuals) are subject to misclassification due to imperfect assays. The observed data available for fitting of the sum-of-trees model consists of error contaminated test results that are taken on pools and/or individuals according to a group testing protocol. Further complicating the data structure, many group testing protocols require individuals to be tested in multiple, possibly overlapping, pools [Gastwirth and Johnson, 1994; Johnson and Gastwirth, 2000; Krajden et al., 2014]. Thus, to maintain generality, we track pool membership through the index sets $\mathcal{P}_j \subset \{1, 2, \ldots, N\}$, for $j = 1, ..., J$, where $\mathcal{P}_j$ consists of the indices of the individuals who contributed to the jth pool. Let Z_j denote the test outcome observed from assaying the jth pool, with the convention that $Z_j = 1$ denotes the event that the pool tested positively and $Z_j = 0$ otherwise. To relate the test outcomes to the individual level covariates and to allow for imperfect testing, we assume that $S_{ej} = P(Z_j = 1 \mid \widetilde{Z}_j = 1)$ and $S_{pj} = P(Z_j = 0 \mid \widetilde{Z}_j = 0)$, where S_{ej} and S_{pj} are the sensitivity and specificity of the assay when used to test the jth pool and $\widetilde{Z}_j$ is the true status of the pool. A few comments are warranted. First, the true status of a pool is said to be positive ($\widetilde{Z}_j = 1$) if it contains at least one truly positive individual and negative ($\widetilde{Z}_j = 0$) otherwise; i.e., $\widetilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0)$. Like the individuals' true statuses, the $\widetilde{Z}_j$'s are also unobserved due

to the effect of imperfect testing. Second, we consider pool-specific testing accuracies (i.e., S_{ej} and S_{pj}) to account for changes in these measures that are related to the use of different assays or other factors that could impact the assay's performance; e.g., specimen type, pool size (i.e., cardinality of $\mathcal{P}_j$).

In some settings, it may be reasonable to assume that the assay accuracies are known *a priori*. However, in others this may be an untenable assumption. When the pool-specific assay accuracies (i.e., S_{ej} and S_{pj}) are unknown, we would like to estimate them along with the sum-of-trees model parameters, following the approach of McMahan et al. [2017]. To do so, we first divide the test outcomes into L different strata based on relevant factors; e.g., pool size and specimen/assay type. Define the index set $\mathcal{M}(l) = \{j : \text{the } j\text{th test outcome is a part of the } l\text{th strata}\}$. We assume that the test accuracies vary across these strata, but are constant within strata. Thus, we define $S_{e(l)}$ and $S_{p(l)}$ to be the sensitivity and specificity of the assay associated with the lth strata; i.e., $S_{ej} = S_{e(l)}$ and $S_{pj} = S_{p(l)}$ if and only if $j \in \mathcal{M}(l)$. Proceeding in this fashion leads to a straightforward way of estimating these unknown quantities as well as a way to inject information about them through prior specifications; for further discussion, see Section 4.5. Based on the relations outlined above, and a few mild assumptions, the conditional distribution of the observed test data $\boldsymbol{Z} = (Z_1, ..., Z_J)'$ is given by

$$\begin{aligned}\pi(\boldsymbol{Z}\big|\boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}) = \sum_{\widetilde{\boldsymbol{Y}} \in \{0,1\}^N} \Bigg[\prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} \left\{ S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j} \right\}^{\widetilde{Z}_j} \\ \times \left\{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \right\}^{1-\widetilde{Z}_j} \\ \times \prod_{i=1}^{N} \{\Phi(\eta_i)\}^{\widetilde{Y}_i} \{1 - \Phi(\eta_i)\}^{1-\widetilde{Y}_i} \Bigg] \end{aligned} \tag{4.3}$$

where $\boldsymbol{S_e} = (S_{e(1)}, ..., S_{e(L)})'$, $\boldsymbol{S_p} = (S_{p(1)}, ..., S_{p(L)})'$, $\boldsymbol{T} = (T_1, ..., T_K)'$, $\boldsymbol{M} = (M_1, \ldots, M_K)'$, and $\eta_i = \eta(\boldsymbol{x}_i)$. A few comments regarding (4.3) are warranted. First, to derive (4.3), we assume that the observed testing responses $\boldsymbol{Z}$ are conditionally independent given their true statuses $\widetilde{\boldsymbol{Z}} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_J)'$, and that $\boldsymbol{Z} \mid \widetilde{\boldsymbol{Z}}$ does not depend on the covariates $\boldsymbol{X}$. These assumptions are common among the group testing literature; e.g., see Vansteelandt et al. [2000]; Xie [2001]. Second, evaluating the data model outlined in (4.3) requires taking the sum over the set $\{0, 1\}^N$, which denotes the collection of all 2^N possible realizations of $\widetilde{\boldsymbol{Y}}$. For this reason, directly evaluating (4.3) can be computationally burdensome if at all feasible. Admittedly, under specific group testing strategies

(e.g., master pool testing) simplifications are possible, yet not in the general case. To overcome this limitation, we make use of a data augmentation strategy, described in Section 4.3.1, to develop a posterior sampling algorithm that circumvents the need to directly evaluate this data model.

4.2.1 Prior specifications

To complete our Bayesian model, we specify priors for each of the unknown model parameters; i.e., the parameters governing the sum-of-trees model and the testing accuracies. Recall, the sum-of-trees model (4.2) is determined by the K trees $(T_1, M_1), \ldots, (T_K, M_K)$. Thus, we must impose priors on the kth tree structure, T_k, and the terminal node parameters given the tree structure, $M_k \mid T_k$, for $k = 1, \ldots, K$. Assuming that the trees and associated terminal node parameters, $(T_1, M_1), \ldots, (T_K, M_K)$, are independent of each other, we can write the prior distribution as

$$\begin{aligned} \pi\left\{(T_1, M_1), \ldots, (T_K, M_K)\right\} &= \prod_{k=1}^{K} \pi(T_k, M_k) \\ &= \prod_{k=1}^{K} \pi(M_k \mid T_k)\pi(T_k) \\ &= \prod_{k=1}^{K} \prod_{t=1}^{b_k} \pi(\mu_{tk} \mid T_k)\pi(T_k), \end{aligned} \tag{4.4}$$

where the last line of (4.4) follows from assuming that the terminal node parameters are conditionally independent given the tree structure. To elicit priors for each T_k and $\mu_{tk} \mid T_k$, we follow the work of Chipman et al. [2010]. In particular, we simplify prior specifications by using identical forms for all $\pi(T_k)$ and for all $\pi(\mu_{tk} \mid T_k)$, $t = 1, \ldots, b_k$; $k = 1, \ldots, K$.

Following the work of Chipman et al. [2010], the prior specification for the tree structure, $\pi(T_k)$, is based on three probabilistic rules that control the size (i.e., number of terminal nodes) of the tree, the variables selected to split on, and the location of the splits. The size of the tree is determined based on the depth of the terminal nodes, where a node at depth $d \in \{0, 1, 2, \ldots\}$ is nonterminal (i.e., an interior node) with probability $\alpha(1 + d)^{-\beta}$, where $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$. The default values of the hyperparameters recommended by Chipman et al. [2010], and used herein, are $\alpha = 0.95$ and $\beta = 2$. This default specification tends to *a priori* favor smaller trees; i.e., trees having 2 to 3 terminal nodes. For nonterminal nodes, the variable to split on is randomly selected from the set of available covariates; and the location of the split, given the selected splitting variable, is randomly selected from the discrete set of observed values of that variable.

Attention is now turned to the prior for the terminal node parameters given the tree structure; i.e., $\pi(M_k \mid T_k)$. Following Chipman et al. [2010], we assume that the interval $(\Phi[-3.0], \Phi[3.0])$ contains most of the classification probability values of interest, a case which will often be of practical pertinence. The following conjugate normal prior is specified for each terminal node parameter, given the tree structure: $\mu_{tk} \sim N(0, \sigma_\mu^2)$, where $\sigma_\mu = 3.0/\left(H\sqrt{K}\right)$, and H is such that $\eta(\boldsymbol{x}_i)$ will be in the interval $(-3.0, 3.0)$ with high probability. The aim of this prior is to provide model regularization; it has the ability to shrink the terminal node parameters, limiting the effect of the individual tree components. As H or the number of trees K is increased, greater shrinkage will be applied to the terminal node parameters. The recommended default hyperparemeter setting is $H = 2$, which is used herein. For further details about the sum-of-trees prior specifications, see Chipman et al. [2010].

Finally, to acknowledge uncertainty in the assay accuracies, we need to elicit prior distributions for $S_{e(l)}$ and $S_{p(l)}$, for $l = 1, \ldots, L$. Given the form of (4.3), we naturally specify the following independent Beta priors:

$$\begin{aligned} S_{e(l)} &\sim Beta(a_{e(l)}, b_{e(l)}) \\ S_{p(l)} &\sim Beta(a_{p(l)}, b_{p(l)}), \text{ for } l = 1, \ldots, L. \end{aligned} \tag{4.5}$$

When historical information about assay performance is available (e.g., from pilot studies used to validate the testing assay), we can incorporate it into the model by choosing hyperparameter values that reflect our prior belief about assay accuracy through the use of informative priors. We illustrate the use of informative priors for the assay accuracies in Section 4.5.

4.3 Posterior Inference

4.3.1 Data augmentation

Recall that evaluating (4.3) is computationally infeasible. To facilitate the development of an efficient posterior sampling algorithm and to avoid having to directly evaluate the data model outlined in (4.3), we propose a two-stage data augmentation strategy. In the first stage, we introduce the individuals' true disease statuses $\widetilde{Y}_i$ as latent random variables and instead consider the joint

conditional distribution

$$\begin{aligned}\pi(\boldsymbol{Z}, \widetilde{\boldsymbol{Y}} | \boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}) = & \prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} \{S_{e(l)}^{Z_j}(1 - S_{e(l)})^{1-Z_j}\}^{\widetilde{Z}_j} \\ & \times \{(1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j}\}^{1-\widetilde{Z}_j} \\ & \times \prod_{i=1}^{N} \{\Phi(\eta_i)\}^{\widetilde{Y}_i} \{1 - \Phi(\eta_i)\}^{1-\widetilde{Y}_i} \end{aligned} \tag{4.6}$$

Making use of the fact that our data model uses the probit link function, the second stage of our data augmentation strategy introduces a carefully constructed latent random variable, ω_i, for each individual, for $i = 1, \ldots, N$. These random variables are structured to be mutually independent and normally distributed such that $\omega_i > 0$ if $\widetilde{Y}_i = 1$ and $\omega_i \leq 0$ if $\widetilde{Y}_i = 0$; for details, see Albert and Chib [1993]. This stage of our data augmentation procedure yields the following augmented likelihood:

$$\begin{aligned}\pi(\boldsymbol{Z}, \widetilde{\boldsymbol{Y}}, \boldsymbol{\omega} | \boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}) = & \prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} \{S_{e(l)}^{Z_j}(1 - S_{e(l)})^{1-Z_j}\}^{\widetilde{Z}_j} \\ & \times \{(1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j}\}^{1-\widetilde{Z}_j} \\ & \times \prod_{i=1}^{N} \phi(\omega_i - \eta_i) \xi(\omega_i), \end{aligned} \tag{4.7}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)'$, $\phi(\cdot)$ denotes the standard normal PDF, and $\xi(\omega_i) = I(\widetilde{Y}_i = 1, \omega_i > 0) + I(\widetilde{Y}_i = 0, \omega_i \leq 0)$. This two-stage data augmentation procedure, together with the proposed prior specifications, allows for the construction of an easy-to-implement, fully Gibbs sampling algorithm to be used for posterior inference.

4.3.2 Posterior sampling algorithm

In this section, we briefly describe the full conditional posterior distributions used in this algorithm. A complete, description of the posterior sampling algorithm is provided in Appendix C.

Attention is first turned to the latent random variables introduced through the data augmentation procedure; i.e., $\widetilde{\boldsymbol{Y}}$ and $\boldsymbol{\omega}$. It follows from the conditional distribution in equation (4.6) that the full conditional posterior of $\widetilde{Y}_i$ is Bernoulli; i.e., $\widetilde{Y}_i | \boldsymbol{Z}, \widetilde{\boldsymbol{Y}}_{-i}, \boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{T}, \boldsymbol{M} \sim Bernoulli\{p_{i1}^*/(p_{i0}^* + p_{i1}^*)\}$,

where $\widetilde{\boldsymbol{Y}}_{-i}$ is the vector $\widetilde{\boldsymbol{Y}}$ with the ith element removed, and

$$p_{i1}^* = \Phi(\eta_i) \prod_{l=1}^{L} \prod_{j \in \mathcal{I}_i(l)} S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j}$$
$$p_{i0}^* = \{1 - \Phi(\eta_i)\} \prod_{l=1}^{L} \prod_{j \in \mathcal{I}_i(l)} \left\{ S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j} \right\}^{I(s_{ij}>0)} \left\{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \right\}^{I(s_{ij}=0)}.$$

In the above expressions, $s_{ij} = \sum_{i' \in \mathcal{P}_j : i' \neq i} \widetilde{Y}_{i'}$; and the index set $\mathcal{I}_i(l) = \{j \in \mathcal{M}(l) : i \in \mathcal{P}_j\}$ keeps track of the pools that belong to the lth strata to which the ith individual was a member of. It follows from (4.7) that the full conditional posterior of ω_i is truncated normal, where the truncation depends on the ith latent disease status $\widetilde{Y}_i$; that is,

$$\omega_i \mid \widetilde{Y}_i, \boldsymbol{T}, \boldsymbol{M} \sim \begin{cases} TN\{\eta_i, 1, (0, \infty)\}, & \text{if } \widetilde{Y}_i = 1 \\ TN\{\eta_i, 1, (-\infty, 0)\}, & \text{if } \widetilde{Y}_i = 0, \end{cases} \tag{4.8}$$

for $i = 1, \ldots, N$, where $TN\{\mu, \sigma^2, (a, b)\}$ denotes a truncated normal distribution with mean μ, variance σ^2, and support over the interval (a, b); see Albert and Chib [1993].

Given the carefully constructed latent random variables and the form of the augmented likelihood (4.7), sampling the sum-of-trees model parameters is straightforward following the Bayesian backfitting algorithm of Chipman et al. [2010]. For details and complete expressions for the posteriors of the sum-of-trees model parameters, refer to Appendices C.1, C.2, and C.3.

Under the prior specifications in (4.5), the full conditional distributions for the assay accuracies $\boldsymbol{S_e}$ and $\boldsymbol{S_p}$ are also Beta; that is,

$$S_{e(l)} \mid \boldsymbol{Z}, \widetilde{\boldsymbol{Y}} \sim Beta(a_{e(l)}^*, b_{e(l)}^*)$$
$$S_{p(l)} \mid \boldsymbol{Z}, \widetilde{\boldsymbol{Y}} * \sim Beta(a_{p(l)}^*, b_{p(l)}^*), \text{ for } l = 1, \ldots, L,$$

where $a_{e(l)}^* = a_{e(l)} + \sum_{j \in \mathcal{M}(l)} Z_j \widetilde{Z}_j$, $b_{e(l)}^* = b_{e(l)} + \sum_{j \in \mathcal{M}(l)} (1 - Z_j) \widetilde{Z}_j$, $a_{p(l)}^* = a_{p(l)} + \sum_{j \in \mathcal{M}(l)} (1 - Z_j)(1 - \widetilde{Z}_j)$, and $b_{p(l)}^* = b_{p(l)} + \sum_{j \in \mathcal{M}(m)} Z_j (1 - \widetilde{Z}_j)$.

A complete, step-by-step description of the posterior sampling algorithm is provided in Appendix C.4.

4.3.3 Variable Selection

After running the posterior sampling algorithm long enough after an appropriate burn-in period, we obtain a sequence of S successive sum-of-trees model draws. These simulated sum-of-trees models can be used to assess variable importance via the 'model-free' variable selection approach of Chipman et al. [2010], which selects those variables that appear most often in the sum-of-trees model draws. For the sth MCMC iterate's simulated sum-of-trees model, let z_{sq} be the proportion of all splitting rules that use the qth covariate component. With this, we define

$$v_q = \frac{1}{S} \sum_{s=1}^{S} z_{sq} \tag{4.9}$$

to be the average use per splitting rule for the qth covariate component, for $q = 1, \ldots, Q$. The covariates with larger values of v_q contribute the most information for predicting the outcome.

While a large number of regression trees is needed for prediction and estimation, this variable selection strategy is actually much more effective when the number of trees is small, because predictors are forced to compete with each other to improve the fit [Chipman et al., 2010]. We illustrate this variable selection strategy and compare its performance with a small and large number of trees in Sections 4.4 and 4.5.

4.4 Simulation Studies

In this section, we conduct numerical studies to examine the performance of our estimation method. We consider two population-level models, both of which following the form of (4.1), for $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})'$, where $x_{i1}, x_{i2} \overset{\text{ind.}}{\sim} \text{Uniform}(0, 10)$, and $x_{i3} \sim Bernoulli(0.5)$.
In the first model (M1),

$$f(\boldsymbol{x}_i) = \sin(\pi \cdot x_{i1}) - 1.25,$$

while in the second model (M2),

$$f(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)' = (-0.85, 0.55, -1.25, -0.35)'$. The first model (M1) was chosen to examine BART's performance when the data structure exhibits nonlinear patterns. Further, only one of the covariates (x_{i1} in particular) was chosen to be truly related to the individual statuses $\widetilde{Y}_i$ so we can explore BART's variable selection performance. The second model (M2) is linear and was chosen so we can explore what is potentially lost or gained from using the BART approach compared to conventional methods (e.g., generalized linear regression methods).

To closely mimic the features of the motivating Iowa chlamydia data analyzed in Section 4.5, we generated $N = 5000$ individual true statuses $\widetilde{Y}_i$ from both models (M1 and M2) to induce a relatively low population prevalence around 10% - 15%. This sample siz is chosen to be roughly one third of the motivating data's sample size. The majority of master pools in the motivating data were of size four, so we randomly assigned the generated individual statuses to pools of size four. For the simulation of the observed testing responses Z_j, we consider two group testing protocols: master pool testing (MPT) and Dorfman testing (DT). With MPT, only the non-overlapping, initial master pools are tested and no further testing is performed, regardless of the outcome. DT is a two-stage hierarchical testing procedure where the master pools are tested in the first stage (like MPT) and in the second stage, positive pools are resolved by retesting each individual separately. Even if the same assay type is used, there could be differences in its accuracy when testing master pools as opposed to individuals. Thus, the testing outcomes can be divided into $L = 2$ strata: master pool test outcomes with assay accuracies $S_{e(1)}$, $S_{p(1)}$ and individual retest outcomes with assay accuracies $S_{e(2)}$, $S_{p(2)}$. For each model (M1 and M2) and protocol (MPT and DT), we examine two settings for the assay accuracy probabilities. In the first, sensitivity and specificity are assumed to be known and we set $S_{ej} = 0.95$ and $S_{pj} = 0.98$ for all $j = 1, \ldots, J$. In the second setting, we assume that assay accuracies are unknown and are estimated simultaneously along with the sum-of-tress model parameters. Only DT is implemented under this setting, and assay accuracies vary across the $L = 2$ strata: master pool tests have accuracies $S_{e(1)} = 0.95$, $S_{p(1)} = 0.98$, and individual retests have accuracies $S_{e(2)} = 0.98$, $S_{p(2)} = 0.99$. Note that MPT is for estimation purposes only, as positive pools are not resolved further, and is not implemented in the second setting.

To examine BART's overall performance when a small or large ensemble of trees is used, we fit two BART models with K=20 and K=200 trees, respectively. For the BART model parameters, we use the default prior specifications of Chipman et al. [2010], as described in Section 4.2.1. Under the setting with unknown accuracies, we assume that no prior knowledge about test performance

is available and specify flat, uninformative Beta priors for the assay accuracy probabilities. That is, $S_{e(l)}, S_{p(l)} \sim Beta(1,1)$. For purposes of comparison, we also fit the Bayesian generalized linear model (GLM) described in McMahan et al. [2017], where flat priors were placed on all GLM parameters.

We used our posterior sampling algorithm for the BART models (and wthe posterior sampling algorithm outlined in McMahan et al. [2017] for GLM) to draw 2500 samples after a burn-in of 2500 samples. Trace plots were used to assess convergence. All results are based on 500 independent group testing data sets. To examine classification accuracy, we conducted a receiver operating characteristic (ROC) curve analysis which was summarized by using the area under the curve (AUC). To assess out-of-sample classification accuracy for each model fit, we simulated 1,000 new individuals using the process outlined above and then used our model fits to predict their infection probabilities and compute the associated AUC scores. BART can also be used to screen for variable selection, as described in Section 4.3.3. To illustrate this strategy and examine the variable selection performance for both of the BART model fits, the average use per splitting rule, v_q, defined in (4.9), is recorded for each component of $\boldsymbol{x}_i$ over the 2500 MCMC samples.

Figure 4.2 shows the in-sample data results when estimating $f(\cdot)$ in Model 1 (M1), assuming assay accuracy probabilities are known, by using the three model fits: BART with K=20 trees (left), BART with K=200 trees (middle), and GLM (right). In each subfigure, we display the mean of the 500 estimated functions - i.e., posterior means - from each simulation (solid red curves) along with the 0.025 & 0.975 quantiles of the 500 posterior means (dashed red curves). The black solid curve in each subfigure is the true function $f(\cdot)$ in model M1. The mean estimated functions from both of the BART fits are in agreement with the true regression function of model M1, indicated by Figure 4.2. This showcases BART's ability to model nonlinear effects between response and predictor variables, while also illustrating the limitations of conventional linear methods.

Next, we summarize the ROC analysis for both models M1 and M2 under the two group testing protocols (MPT and DT), when the assay accuracy probabilities are known. Table 4.2 reports the average (and sample standard deviation in parentheses) of 500 AUC scores for in-sample and out-of-sample predictions. For model M1, this table shows that the average AUC scores for the two BART models are significantly larger than that for the conventional GLM. When nonlinear effects between response and predictor variables are present, our BART approach has notably better classification accuracy. When the effects between response and predictor variables are truly linear,

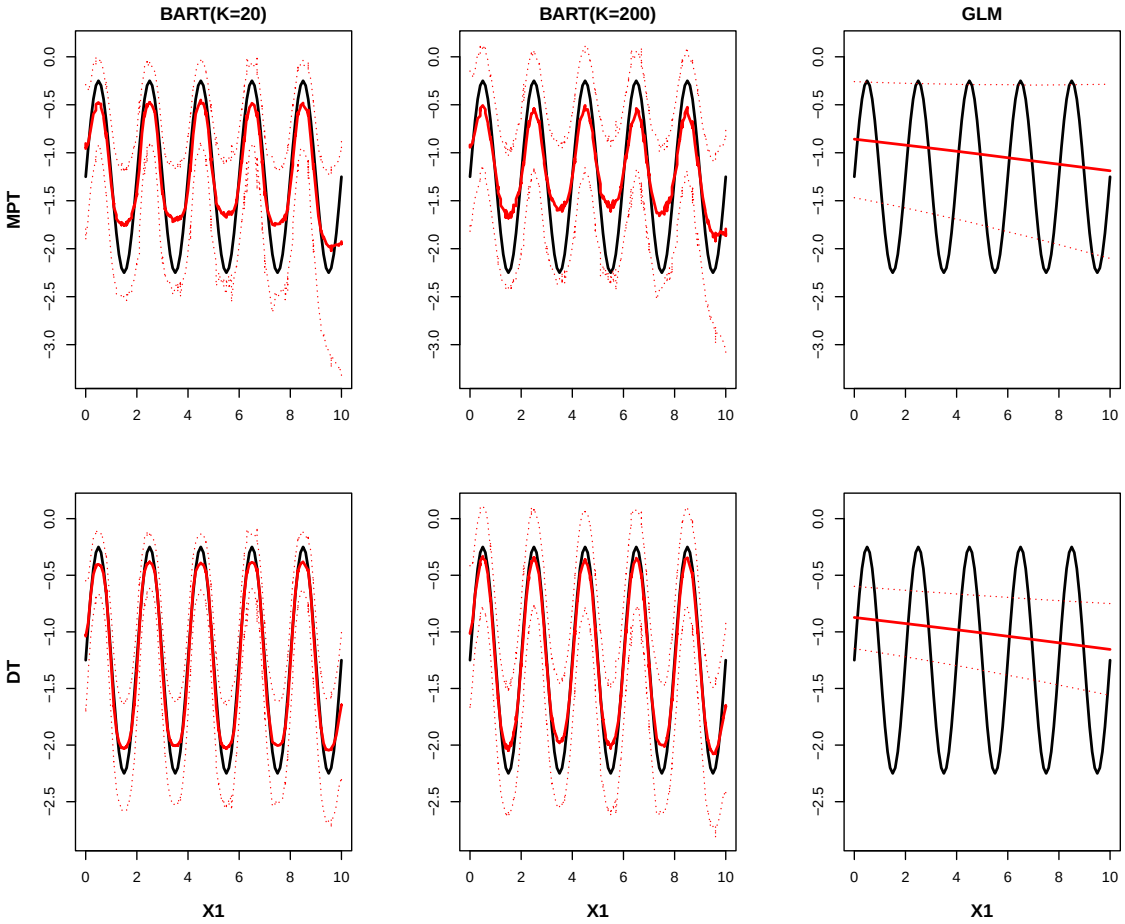

Figure 4.2: In-sample simulation results for MPT (top row) and DT (bottom row) when assay accuracy probabilities are **known** for the three model fits BART with K=20 trees (left), BART with K=200 trees (middle), and GLM (right). The black solid curve in each subfigure is the true function $f(\cdot)$ in model M1. In each subfigure the following are displayed as red curves: the average of 500 posterior mean estimates (solid curves) and the .025 & .975 posterior mean quantiles (dashed curves).

BART and GLM have similar predictive accuracy (indicated by the AUC scores for model M2 in Table 4.2), and we can conclude that the proposed BART methodology performs just as well as the conventional GLM.

Figure 4.3 plots the average use per splitting rule measures (4.9) for the three covariates for MPT (top row) and DT (bottom row) for models M1 (left) and M2 (right) under the BART fits with K=20 trees (blue lines) and K=200 trees (red lines), when the assay accuracy probabilities are known. Over the 2500 MCMC iterations for model M1, the fitted sum-of-trees models increasingly incorporate the covariates that are truly important for prediction as the number of trees K decreases

Table 4.2: Average estimated AUC and sample standard deviation (in parentheses) for the three model fits (BART with K=20, BART with K=200 trees, and GLM) when the assay accuracy probabilities are **known**.

Model	GT Protocol		BART(K=20)	BART(K=200)	GLM
M1	MPT	In-Sample	0.758 (0.013)	0.773 (0.013)	0.537 (0.015)
		Out-of-Sample	0.743 (0.023)	0.750 (0.020)	0.523 (0.025)
	DT	In-Sample	0.799 (0.007)	0.816 (0.007)	0.544 (0.010)
		Out-of-Sample	0.773 (0.018)	0.777 (0.017)	0.527 (0.023)
M2	MPT	In-Sample	0.983 (0.002)	0.986 (0.001)	0.985 (0.001)
		Out-of-Sample	0.974 (0.004)	0.976 (0.004)	0.980 (0.003)
	DT	In-Sample	0.986 (0.001)	0.988 (0.001)	0.985 (0.001)
		Out-of-Sample	0.977 (0.004)	0.977 (0.004)	0.980 (0.003)

from K=200 to K=20. This is particularly true for DT protocol compared to MPT. For model M1, x_{i1} has a significantly larger average use value compared to the other two covariates, suggesting that x_{i1} is important for predicting the outcome. For model M2, the average use measures for the three covariates are not drastically different in value, suggesting that all three are useful for prediction.

We now turn our attention to the simulation results for models M1 and M2 when the assay accuracy probabilities are unknown (and estimated simultaneously with other model parameters), under the DT protocol. Summaries of the results are provided in Appendix C.5. Table 16 summarizes the estimation results for the unknown assay accuracy probabilities; namely, the screening accuracies of the pools, $S_{e(1)}$ and $S_{p(1)}$, and the confirmatory accuracies for the individuals, $S_{e(2)}$ and $S_{p(2)}$. Estimates of the accuracies in Table 16 exhibit little (if any) average bias, there is close agreement between SSD and ESE, and credible intervals attain their nominal level. We can conclude that BART provides reliable inference for the assay accuracies, even when providing no information in the prior distributions. Figure 2 shows the in-sample data results under DT when estimating $f(\cdot)$ in Model M1, assuming assay accuracy probabilities are unknown, using the three model fits BART with K=20 trees (left), BART with K=200 trees (middle), and GLM (right). The estimated functions appear to be analogous to that with known accuracies under DT (Figure 4.2). Further, the ROC analysis results reported in Table 17 are identical to that of the BART fits with known accuracies under DT (Table 4.2). Finally, from the average variable use measures plotted in Figure 3, we see that BART still correctly identifies the truly influential variables with unknown assay accuracies. Thus, we can conclude from these findings that BART's overall performance is unaffected by the estimation of unknown assay accuracies. Overall, the simulation results outlined above suggest the

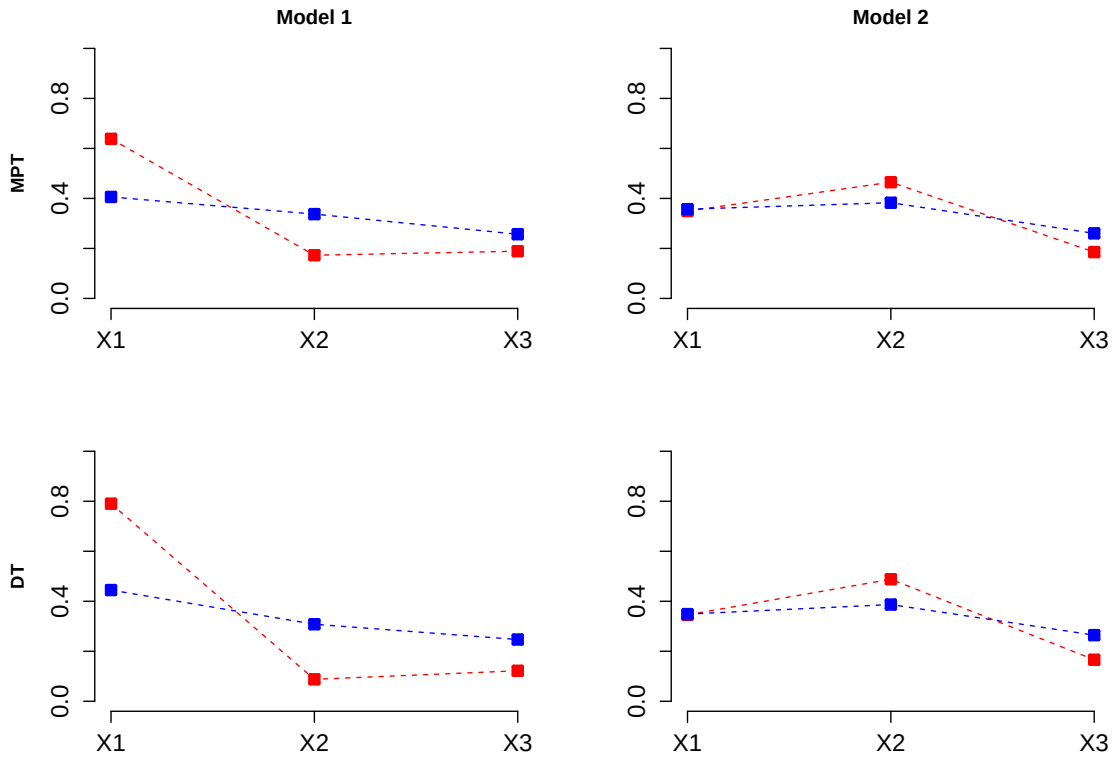

Figure 4.3: Simulation results for MPT (top row) & DT (bottom row) for model M1 (left) and model M2 (right) when assay accuracy probabilities are **known**. For each covariate, the average use proportion (averaged over the 500 simulations) is plotted for the two BART fits with K=20 trees (red) and K=200 trees (blue).

our approach outperforms conventional methods when nonlinear effects are present, and performs just as well as conventional methods when only linear effects are present. The unknown assay accuracy probabilities do not impact BART's predictive accuracy nor its ability to provide reliable inference. These numerical studies provide us with the confirmation that our proposed approach can be used in group testing data estimation.

4.5 Iowa Chlamydia Data Analysis

The State Hygienic Laboratory (SHL) at the University of Iowa is the largest public health laboratory in Iowa. Each year the lab tests thousands of Iowa residents for chlamydia and gonorrhea as part of federally sponsored STD assessment & prevention programs. The SHL receives both endocervical swab and urine specimens each day; their current protocol is to use Dorfman testing

(DT) for all endocervical swab specimens collected from females, usually in master pools of size four, and to use individual testing for all other specimens (i.e., urine specimens). Individual swab specimens residing in master pools which test positively are retested immediately in order to provide final diagnoses to patients in a timely manner. To test both the urine and endocervical swab specimens, the SHL uses the Aptima Combo 2 Assay (AC2A). Pilot data describing the accuracy of the AC2A for individual testing are summarized in the product literature, available at `www.fda.com`; see also Gaydos et al. [2003]. We also summarize these pilot data in Table 18 in Appendix C.6.

To illustrate the BART methodology described in this paper, our analysis specifically examines the chlamydia data collected on $N = 13,862$ female subjects during the 2014 calendar year. The available data consists of test results for 2286 swab master pools (1 of size 2; 12 of size 3; and 2273 of size 4), 416 individual swab specimens, and 4316 individual urine specimens. Dorfman retesting results on positive swab master pools are also included. Additionally, six covariates to be included in the model were collected on each individual: age (in years, denoted by x_{i1}), a race indicator ($x_{i2} = 1$ if Caucasian and $x_{i2} = 0$ otherwise), an indicator denoting whether the patient reported a new sexual partner in the last 90 days ($x_{i3} = 1$ if affirmative and $x_{i3} = 0$ otherwise), an indicator denoting whether the patient reported having multiple sexual partners in the last 90 days ($x_{i4} = 1$ if affirmative and $x_{i4} = 0$ otherwise), an indicator denoting whether the patient reported sexual contact with an STD-positive partner in the previous year ($x_{i5} = 1$ if affirmative and $x_{i5} = 0$ otherwise), and an indicator denoting whether the patient presented with symptoms ($x_{i6} = 1$ if affirmative and $x_{i6} = 0$ otherwise). To relate an individual's chlamydia disease status to the available covariates, we consider the following BART model

$$\Phi^{-1}\left[P(\widetilde{Y}_i = 1 \big| \boldsymbol{x}_i)\right] = \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k, M_k)$$

under two specifications, namely with K=20 trees and K=200 trees, for $i = 1, 2, \ldots, 13,862$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i6})'$.

We use the priors outlined in Section 4.2.1. For model parameters associated with the K regression trees, we use the default prior specifications of Chipman et al. [2010]. Although the same testing assay (AC2A) was used on all specimen types, it is important to acknowledge differences in how it may perform on swab versus urine specimens [Gaydos et al., 2003], and to recognize that there could be differences in its performance when testing pools as opposed to individuals. With

this in mind, we posited three ($L = 3$) sensitivity and specificity parameter pairs: $S_{e(1)}$ and $S_{p(1)}$ for swab specimens tested individually, $S_{e(2)}$ and $S_{p(2)}$ for urine specimens tested individually, and $S_{e(3)}$ and $S_{p(3)}$ for swab specimens tested in pools. For these six parameters, we chose very informative Beta priors based on the individual AC2A pilot data; for further details, refer to Appendix C.6. For purposes of comparison, we also consider a Bayesian generalized linear model (GLM), following McMahan et al. [2017].

First, we seek to compare the predictive performance of a BART model with K=20 trees (a small number of trees for variable selection), a BART model with K=200 trees (a large number of trees for flexible prediction), and the Bayesian GLM fit of McMahan et al. [2017]. To do so, we randomly split the data into a training and test set where 85% of the data was used to train the model and the remaining 15% was allocated to the test set. Note that the true responses (individual disease statuses) are obscured by the assay testing errors. Therefore, it is not appropriate to conduct an ROC curve analysis as was done in the simulation studies of Section 4.4. Instead, we will examine the predictive error through the log-likelihood. For both BART and GLM, using the posterior mean parameter estimates, we computed the log-likelihood as a measure of overall fit. Table 4.3 reports the calculated log-likelihood from both the in-sample and out-of-sample data. The BART models results in a larger log-likelihood for in- and out-of-sample, implying that they fit the data better than the GLM model. For confirmation, we fit a 'age-only' model (i.e., only the age covariate x_{i1} was included) to the data to compare the regression function fits for GLM, BART with 20 trees, and BART with 200 trees. Figure 4.4 displays the posterior mean estimated functions against the age covariate for the model fits. This figure confirms the findings in Liu et al. [2021] regarding the nonlinear effect of age, particularly for specific subsets of age. Our BART approach extends the methodology of Liu et al. [2021] by accommodating the nonlinearity of age, as well as allowing for potential nonlinear interactions of multiple covariates without having to explicitly specify them.

Table 4.3: In- and out-of-sample log likelihood calculated with posterior mean estimates of the assay accuracy probabilities (sensitivity and specificity) and the individual probabilities of being truly positive for chlamydia.

	BART(K=20)	BART(K=200)	GLM
In-Sample	-3329.85	-3320.62	-4379.05
Out-of-Sample	-595.75	-594.95	-802.07

Table 4.4 reports the posterior mean estimates, estimated posterior standard deviations,

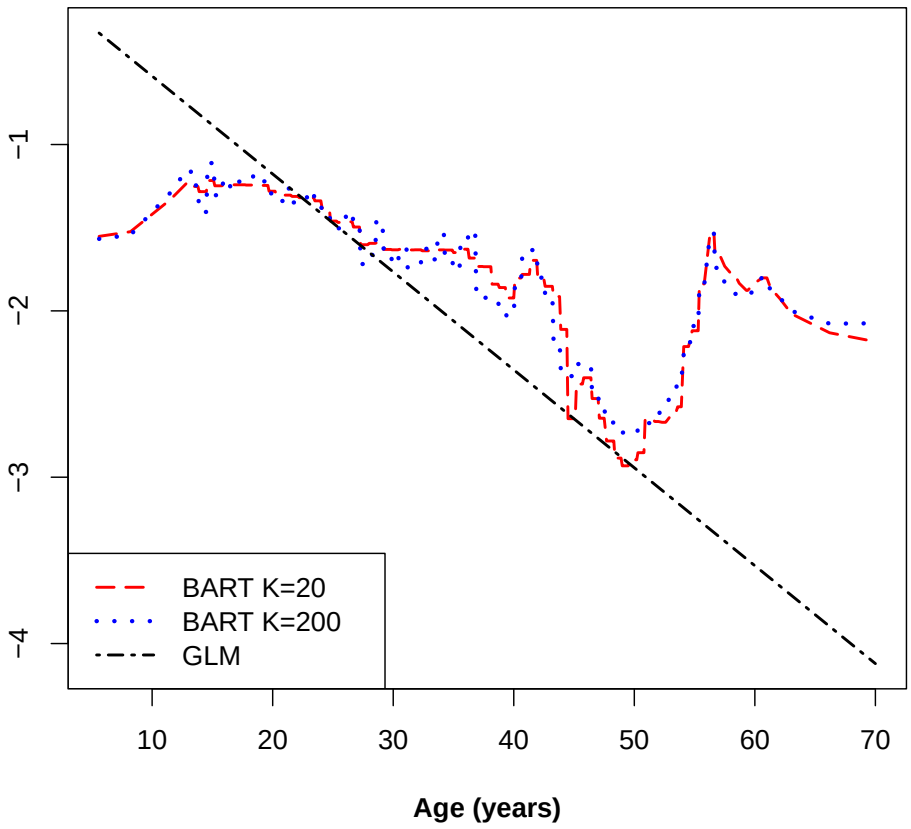

Figure 4.4: Estimation results from the age-only model from GLM (dot-dashed black curve), BART with K=20 trees (dashed red curve), and BART with K=200 trees (dotted blue curve).

and 95% equal-tail credible intervals for the six assay accuracy probabilities previously described. Note that the BART models produces specificity estimates that are similar to those produced by the GLM fit, and the amount of variability in these estimates is also similar. On the other hand, the BART model produces slightly larger sensitivity estimates than GLM, and the variability in these estimates is notably larger for GLM.

Finally, BART can also be used to assess variable importance by calculating the average variable use per splitting rule for each covariate, as discussed in Section 4.3.3. Figure 4.5 plots the average use for the 6 covariates under the two BART fits. It appears that all 6 covariates are important in the prediction of disease status. As the number of trees decreases from K=200 to K=20, the fitted sum-of-trees models increasingly incorporate the variable age but the multiple partners variable is incorporated less often. From this, we can conclude that age is one of the more

Table 4.4: Iowa Chlamydia Data. Results from estimating the assay accuracy probabilities $S_{e(l)}$ and $S_{p(l)}$, for $l = 1, 2, 3$. Posterior mean estimates (Est), estimated posterior standard deviations (ESE), and 95% equal-tail credible intervals (CI95) are provided.

		BART(K=20)			**BART(K=200)**			**GLM**		
Param.	Descrip.	Est	ESE	CI95	Est	ESE	CI95	Est	ESE	CI95
$S_{e(1)}$	Swab Ind.	0.98	0.01	(0.97, 0.99)	0.98	0.01	(0.97, 0.99)	0.97	0.04	(0.84, 0.99)
$S_{e(2)}$	Urine Ind.	0.95	0.02	(0.91, 0.97)	0.95	0.02	(0.91, 0.97)	0.90	0.10	(0.56, 0.97)
$S_{e(3)}$	Swab Pool	0.94	0.02	(0.91, 0.97)	0.94	0.02	(0.91, 0.97)	0.91	0.10	(0.57, 0.97)
$S_{p(1)}$	Swab Ind.	0.97	0.00	(0.97, 0.98)	0.97	0.00	(0.97 0.98)	0.97	0.00	(0.96, 0.98)
$S_{p(2)}$	Urine Ind.	0.99	0.00	(0.99, 0.99)	0.99	0.00	(0.99, 0.99)	0.99	0.00	(0.98, 0.99)
$S_{p(3)}$	Swab Pool	0.99	0.00	(0.99, 0.99)	0.99	0.00	(0.99, 0.99)	0.99	0.00	(0.98, 0.99)

influential predictors of chlamydia infection status and having multiple sexual partners is the least influential predictor of chlamydia infection status.

4.6 Discussion

BART is an attractive approach for developing flexible predictive models and, in particular, it offers the ability to provide uncertainty in estimates. In this article, we have developed a general Bayesian additive regression trees (BART) approach with potentially misclassified group testing data with individual-level covariate information. The proposed method extends the methodology described in McMahan et al. [2017] and Liu et al. [2021] to allow for a more flexible estimation framework that has the ability to handle nonlinear main effects and multi-way interaction effects without any input from the researcher. It also has the ability to assess variable importance using a 'model-free' approach.

Several modeling extensions could be of interest. Our proposed BART approach inspires the exploration of other advanced machine learning techniques that could be used for estimation in the group testing setting. One possible extension would be the development of regression techniques used to analyze data that incorporates the testing responses from multiplex assays; i.e., assays that test specimens for multiple diseases simultaneously. Another useful modeling extension would be incorporating the 'dilution effect', and common concern that arises in group testing. This occurs if the signal from a positive individual's specimen is diluted past an assay's threshold of detection when it is pooled with multiple negative specimens.

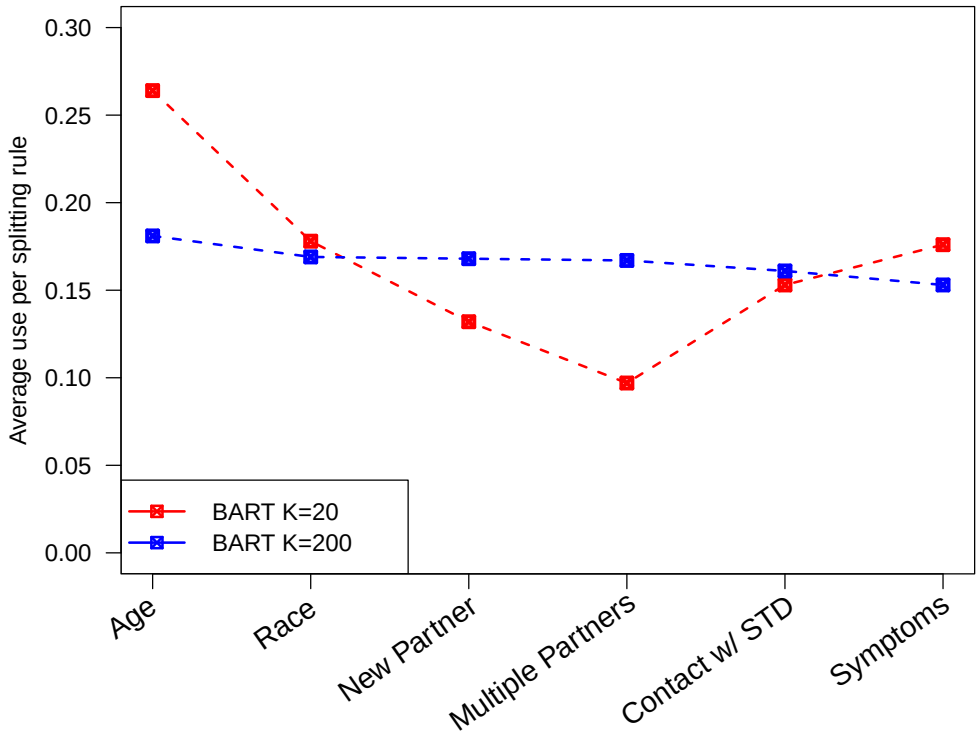

Figure 4.5: Average variable use for the BART models with 20 (red) and 200 (blue) trees.

Chapter 5

Discussion

In this dissertation we developed three models that were directly inspired by complex data collected as a part of several biomedical studies. The three models were all Bayesian in nature and made use of regularizing priors as a means to smooth functional estimates and perform variable selection. For all models, fitting was facilitated through the development of custom MCMC routines that consisted entirely of Gibbs steps which involve sampling from common distributions. For this reason, the proposed algorithms are computationally efficient, easy to implement, and scale well to larger data sets.

Appendices

Appendix A Supplementary Material for Chapter 2

A.1 Posterior Distributions

We assume conditional independence given the covariate effects and random effects and observe that Y_{ij} depends on the model parameters only through the linear predictor, ν_{ij}. Hence, the likelihood can be expressed as

$$p(\boldsymbol{Y}|\boldsymbol{\nu}) \propto \prod_{i,j} g(\nu_{ij})^{Y_{ij}} \{1 - g(\nu_{ij})\}^{1-Y_{ij}},$$

where $g(\cdot)$ is defined to be the logit link function.

We develop a two-stage data augmentation process to construct a posterior sampling algorithm consisting only of Gibbs steps. In the first stage, we exploit a hierarchical representation of the proposed data model by introducing Póyla - Gamma latent random variables w_{ij}; for further details see Polson et al. [2013]. Under this specification, the joint density of the observed and latent data for the ith individual is given by

$$p(\boldsymbol{Y}_i, \boldsymbol{w}_i|\boldsymbol{\nu}_i) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{h}_i - \boldsymbol{\nu}_i)'\boldsymbol{W}_i(\boldsymbol{h}_i - \boldsymbol{\nu}_i)\right\} \times \prod_j \xi(w_{ij}),$$

where $\boldsymbol{h}_i = (\kappa_{i1}/w_{i1}, \ldots, \kappa_{in_i}/w_{in_i})'$ are synthetic responses with $\kappa_{ij} = Y_{ij} - 1/2$, $\boldsymbol{W}_i = \text{diag}(\boldsymbol{w}_i)$, $\xi(w_{ij}) = f(w_{ij}|1,0)\exp\{\kappa_{ij}^2/(2w_{ij})\}$, and $f(w_{ij}|a,b)$ denotes the Pólya - Gamma density with parameters (a,b); for further details, see Polson et al. [2013].

Attention is now turned to the second stage of the data augmentation process and the construction of the hierarchical representation of the joint posterior distribution. Recall from Chapter 2, we specify a generalized double Pareto shrinkage prior for all of the regression coefficients with the exception of the intercept; i.e.,

$$\begin{aligned}
&\alpha_0 \sim N(0, \tau_0), \\
&\alpha_p \sim GDP\left(\psi = b_\delta/a_\delta, a_\delta\right), \text{ for } p = 1, \ldots P-1.
\end{aligned}$$

As noted by Proposition 1 of Armagan et al. [2013], the generalized double Pareto shrinkage prior can be represented as a scale mixture of normal distributions. Thus, for the regression coefficients,

the following hierarchical representation provides the same prior specifications as those given above:

$$\begin{aligned}
&\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{T}), \text{ where } \boldsymbol{T} = \text{diag}(\boldsymbol{\tau}),\ \boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{P-1})' \\
&\tau_p \overset{\text{ind.}}{\sim} Exponential(\delta_p^2/2), \text{ for } p = 1, \ldots, P-1 \\
&\delta_p \overset{\text{iid}}{\sim} Gamma(a_\delta, b_\delta), \text{ for } p = 1, \ldots, P-1.
\end{aligned}$$

The δ_p's are the global shrinkage parameters, while the τ_p's are the local shrinkage parameters and override the impact of the global shrinkage components for the variable fixed effects that are not near zero [Armagan et al., 2013].

In deriving the full conditional distributions, for notational convenience, a dot $\cdot$ is used as shorthand for all the parameters one is conditioning on; e.g., we may write the posterior $p(\boldsymbol{\alpha}\big|\boldsymbol{Y}, \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\tau})$ as $p(\boldsymbol{\alpha}\big|\cdot)$.

We begin by deriving the full conditional distribution for the spline coefficients $\boldsymbol{\eta}$, based on the smoothing penalty inspired prior distribution outlined in Chapter 2. Letting $\boldsymbol{h}^\eta = (\boldsymbol{h}_1^\eta, \ldots, \boldsymbol{h}_n^\eta)'$, where $\boldsymbol{h}_i^\eta := \boldsymbol{h}_i - \boldsymbol{X}_i\boldsymbol{\alpha} - \mathbf{1}_{n_i}\gamma_{0i} - \mathbf{1}_{n_i}\gamma_{1k(i)}$, and $\boldsymbol{M} = (\boldsymbol{M}_1, \ldots \boldsymbol{M}_n)'$, $\boldsymbol{W} = \text{diag}(\boldsymbol{w})$,

$$\begin{aligned}
p(\boldsymbol{\eta}|\cdot) &\propto p(\boldsymbol{Y}, \boldsymbol{w}|\boldsymbol{\nu})p(\boldsymbol{\eta}|\boldsymbol{\tau}) \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{h}^\eta - \boldsymbol{M}\boldsymbol{\eta}\right)'\boldsymbol{W}\left(\boldsymbol{h}^\eta - \boldsymbol{M}\boldsymbol{\eta}\right)\right\} \times \exp\left\{-\frac{1}{2}\boldsymbol{\eta}'\left(\lambda^{-1}\boldsymbol{R}\right)\boldsymbol{\eta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\eta} - \boldsymbol{\mu}_\eta\right)'\left(\boldsymbol{\Sigma}_\eta\right)^{-1}\left(\boldsymbol{\eta} - \boldsymbol{\mu}_\eta\right)\right\},
\end{aligned}$$

where $\boldsymbol{\Sigma}_\eta = \left(\boldsymbol{M'WM} + \lambda^{-1}R\right)^{-1}$ and $\boldsymbol{\mu}_\eta = \boldsymbol{\Sigma}_\eta\boldsymbol{M'Wh}^\eta$. Recognizing this as the kernel of a normal density, we have that

$$\boldsymbol{\eta}\big|\cdot \sim N(\boldsymbol{\mu}_\eta, \boldsymbol{\Sigma}_\eta).$$

Further, the full conditional distribution for the variance parameter λ associated with the smoothing prior for the spline coefficients $\boldsymbol{\eta}$ is derived as follows.

$$\begin{aligned}
p(\lambda\big|\cdot) &\propto p(\boldsymbol{\eta}\big|\lambda)p(\lambda) \\
&\propto (\lambda^{-1})^{L/2}\exp\left\{-\frac{\lambda^{-1}}{2}\boldsymbol{\eta'R\eta}\right\} \times (\lambda^{-1})^{a_\lambda - 1}\exp\left\{-\lambda^{-1}b_\lambda\right\}
\end{aligned}$$

$$\propto (\lambda^{-1})^{a_\lambda^*-1} \exp\left\{-\lambda^{-1} b_\lambda^*\right\},$$

where $a_\lambda^* = a_\lambda + L/2$ and $b_\lambda^* = b_\lambda + 0.5 \cdot \boldsymbol{\eta}'\boldsymbol{\eta}$. Recognizing this as the kernel of a Gamma density, we find that

$$\lambda\big|\cdot \sim \textit{Inv-Gamma}(a_\lambda^*, b_\lambda^*).$$

Given the hierarchical representation of the priors placed on the regression coefficients, let $\boldsymbol{h}^\alpha = (\boldsymbol{h}_1^\alpha, \ldots, \boldsymbol{h}_n^\alpha)'$ where $\boldsymbol{h}_i^\alpha := \boldsymbol{h}_i - \boldsymbol{M}_i\eta - \mathbf{1}_{n_i}\gamma_{0i} - \mathbf{1}_{n_i}\boldsymbol{\gamma}_{1k(i)}$, and let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$. We derive the full conditional distribution for the regression coefficients $\boldsymbol{\alpha}$ as follows.

$$\begin{aligned} p(\boldsymbol{\alpha}\big|\cdot) &\propto p(\boldsymbol{Y}, \boldsymbol{w}\big|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}\big|\boldsymbol{\tau}) \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{h}^\alpha - \boldsymbol{X}\boldsymbol{\alpha}\right)' \boldsymbol{W}\left(\boldsymbol{h}^\alpha - \boldsymbol{X}\boldsymbol{\alpha}\right)\right\} \times \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{T}^{-1}\boldsymbol{\alpha}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha\right)' \left(\boldsymbol{\Sigma}_\alpha\right)^{-1}\left(\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha\right)\right\}, \end{aligned}$$

where $\boldsymbol{\Sigma}_\alpha = \left(\boldsymbol{X'W X} + \boldsymbol{T}^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_\alpha = \boldsymbol{\Sigma}_\alpha \boldsymbol{X'W h}^\alpha$. Recognizing this as the kernel of a Normal density, we find that

$$\boldsymbol{\alpha}\big|\cdot \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha).$$

Next, we derive the full conditional of the local shrinkage parameters τ_p, for $p = 1, \ldots, P-1$:

$$\begin{aligned} p(\tau_p\big|\cdot) &\propto p(\alpha_p\big|\tau_p) p(\tau_p\big|\delta_p) \\ &\propto \exp\left\{-\frac{\tau_p^{-1}}{2}\alpha_p^2\right\} \times \exp\left\{-\tau_p\frac{\delta_p^2}{2}\right\} \\ &\propto \exp\left\{-\frac{\delta_p^2\left(\tau_p^{-1} - \mu_p\right)^2}{2\left(\mu_p\right)^2\tau_p^{-1}}\right\}, \end{aligned}$$

where $\mu_p = \sqrt{\delta_p^2/\alpha_p^2}$. Recognizing this as the kernel of an inverse-Gaussian density, we find that

$$\tau_p^{-1}\big|\cdot \sim \textit{Inv-Gaussian}(\mu_p, \delta_p^2), \text{ for } p = 1, \ldots, P-1.$$

Moreover, the full conditional distribution for the global shrinkage parameters δ_p is derived as follows. Exploiting the fact that the Laplace density is a scale mixture of normals with an exponential mixing density [Park and Casella, 2008]:

$$\begin{aligned} p(\delta_p|\cdot) &\propto p(\alpha_p|\tau_p)p(\tau_p|\delta_p)p(\delta_p) \\ &\propto \frac{\delta_p}{2}\exp\{-\delta_p|\alpha_p|\} \times (\delta_p)^{a_\delta - 1}\exp\{-\delta_p b_p\} \\ &\propto (\delta_p)^{a^*_{\delta_p}-1}\exp\left\{-\delta_p(b^*_{\delta_p})\right\}, \end{aligned}$$

where $a^*_{\delta_p} = a_\delta + 1$ and $b^*_{\delta_p} = b_\delta + |\alpha_p|$. Recognizing this as the kernel of a Gamma density, we find that

$$\delta_p|\cdot \sim Gamma(a^*_{\delta_p}, b^*_{\delta_p}), \text{ for } p = 1, \ldots, P-1.$$

We turn our attention to the random effects and derive the full conditional for the subject-specific random effects, $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0N})'$, where N is the number of participants in the study. Let $\boldsymbol{h}^0 = (\boldsymbol{h}^0_1, \ldots, \boldsymbol{h}^0_N)'$ where $\boldsymbol{h}^0_i := \boldsymbol{h}_i - \boldsymbol{M}_i\boldsymbol{\eta} - \boldsymbol{X}_i\boldsymbol{\alpha} - \mathbf{1}_{n_i}\gamma_{1k(i)}$, and let $\boldsymbol{Z}_0 = \text{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_N})$, where n_i is the number of observations from the ith individual, $i = 1, \ldots, N$.

$$\begin{aligned} p(\boldsymbol{\gamma}_0|\cdot) &\propto p(\boldsymbol{Y}, \boldsymbol{w}|\boldsymbol{\nu})p(\boldsymbol{\gamma}_0|\sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{h}^0 - \boldsymbol{Z}_0\boldsymbol{\gamma}_0\right)'\boldsymbol{W}\left(\boldsymbol{h}^0 - \boldsymbol{Z}_0\boldsymbol{\gamma}_0\right)\right\} \times \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_0'\left(\sigma_0^{-2}\boldsymbol{I}_n\right)\boldsymbol{\gamma}_0\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}_0 - \boldsymbol{\mu}_0\right)'\left(\Sigma_0\right)^{-1}\left(\boldsymbol{\gamma}_0 - \boldsymbol{\mu}_0\right)\right\}, \end{aligned}$$

where $\boldsymbol{\Sigma}_0 = \left(\boldsymbol{Z}_0'\boldsymbol{W}\boldsymbol{Z}_0 + \sigma_0^{-2}\boldsymbol{I}_n\right)^{-1}$ and $\boldsymbol{\mu}_0 = \boldsymbol{\Sigma}_0\boldsymbol{Z}_0'\boldsymbol{W}\boldsymbol{h}^0$. Recognizing this as the kernel of a normal density, we find that

$$\boldsymbol{\gamma}_0|\cdot \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

The full conditional distribution for the variance component of the subject-specific random effects, σ_0^2, is derived as follows.

$$p(\sigma_0^2|\cdot) \propto p(\boldsymbol{\gamma}_0|\sigma_0^2)p(\sigma_0^2)$$

$$\begin{aligned}
&\left(\sigma_0^{-2}\right)^{n/2} \exp\left\{-\frac{\sigma_0^{-2}}{2}\boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0\right\} \times \left(\sigma_0^{-2}\right)^{a_0-1} \exp\left\{-\sigma_0^{-2} b_0\right\} \\
&\propto \left(\sigma_0^{-2}\right)^{a_0^*-1} \exp\left\{-\sigma_0^{-2}(b_0^*)\right\},
\end{aligned}$$

where $a_0^* = a_0 + N/2$ and $b_0^* = b_0 + 0.5 \cdot \boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0$. Recognizing this as the kernel of a Gamma density, we find that

$$\sigma_0^2 \sim \textit{Inv-Gamma}(a_0^*, b_0^*).$$

Next, we derive the full conditional for the trial-specific random effects $\boldsymbol{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{1K})'$, where K is the number of trials in the study. Let $\boldsymbol{h}^1 = \boldsymbol{h} - \boldsymbol{M}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{Z}_0\boldsymbol{\gamma}_0$, and let $\boldsymbol{Z}_1 = \text{diag}(\mathbf{1}_{m_1}, \ldots, \mathbf{1}_{m_K})'$ where m_k is the number of observations coming from trial k, $k = 1, \ldots, K$. Then,

$$\begin{aligned}
p(\boldsymbol{\gamma}_1|\cdot) &\propto p(\boldsymbol{Y}, \boldsymbol{w}|\boldsymbol{\nu}) p(\boldsymbol{\gamma}_1|\sigma_1^2) \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{h}^1 - \boldsymbol{Z}_1\boldsymbol{\gamma}_1\right)' \boldsymbol{W}\left(\boldsymbol{h}^1 - \boldsymbol{Z}_1\boldsymbol{\gamma}_1\right)\right\} \times \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_1'\left(\sigma_1^{-2}\boldsymbol{I}_K\right)\boldsymbol{\gamma}_1\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}_1 - \boldsymbol{\mu}_1\right)'\left(\Sigma_1\right)^{-1}\left(\boldsymbol{\gamma}_1 - \boldsymbol{\mu}_1\right)\right\},
\end{aligned}$$

where $\boldsymbol{\Sigma}_1 = \left(\boldsymbol{Z}_1'\boldsymbol{W}\boldsymbol{Z}_1 + \sigma_1^{-2}\boldsymbol{I}_K\right)^{-1}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1\boldsymbol{Z}_1'\boldsymbol{W}\boldsymbol{h}^1$. Recognizing this as the kernel of a normal density, we find that

$$\boldsymbol{\gamma}_1|\cdot \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1).$$

The full conditional distribution for the variance component of the trial-specific random effects, σ_1^2, is derived as follows.

$$\begin{aligned}
p(\sigma_1^2|\cdot) &\propto p(\boldsymbol{\gamma}_1|\sigma_1^2) p(\sigma_1^2) \\
&\left(\sigma_1^{-2}\right)^{K/2} \exp\left\{-\frac{\sigma_1^{-2}}{2}\boldsymbol{\gamma}_1'\boldsymbol{\gamma}_1\right\} \times \left(\sigma_1^{-2}\right)^{a_1-1} \exp\left\{-\sigma_1^{-2} b_1\right\} \\
&\propto \left(\sigma_1^{-2}\right)^{a_1^*-1} \exp\left\{-\sigma_1^{-2}(b_1^*)\right\},
\end{aligned}$$

where $a_1^* = a_1 + K/2$ and $b_1^* = b_1 + 0.5 \cdot \boldsymbol{\gamma}_1'\boldsymbol{\gamma}_1$. Recognizing this as the kernel of a Gamma density,

we find that

$$\sigma_1^2 \sim Inv\text{-}Gamma(a_1^*, b_1^*).$$

These full conditionals were used to construct an MCMC algorithm in the usual manner. To validate the proposed model and MCMC algorithm, an in depth numerical study was conducted. The study was designed to emulate the primary features of the opioid data under study. The results (not shown) of this numerical study suggest that our proposed approach performs well and is appropriate for analyzing the motivating data.

A.2 Efficacy and Safety Clinical Trials of Buprenorphine Maintenance Treatment

We selected six efficacy and safety trials focused on buprenorphine maintenance treatment for analysis from the Clinical Trials Network (CTN) at NIDA's Data Share resource (`datashare.nida.nih.gov`). Trial information is provided in Table 1.

Table 1: Studies with Individual Patient Data Analyzed

Division (Study ID)	Title	Investigators	Release Date
DTMC (CSP-999)	A Multicenter Clinical Trial of Buprenorphine in Treatment of Opiate Dependence	Walter Ling, M.D., Donald R. Wesson, M.D., C. James Klett, Ph.D.	Sep 02, 2015
DTMC (CSP-1008A)	A Multicenter Efficacy/Safety Trial of Buprenorphine/Naloxone for the Treatment of Opiate Dependence	Peter Bridge, M.D., Paul J. Fudala, Ph.D.	Dec 04, 2014
DTMC (CSP-1008B)	A Multicenter Safety Trial of Buprenorphine/Naloxone for the Treatment of Opiate Dependence	Peter Bridge, M.D.	Dec 04, 2014
DTMC (CSP-1018)	A Multicenter Safety Trial of Buprenorphine/Naloxone for the Treatment of Opiate Dependence	Walter Ling, M.D., Paul J. Fudala, Ph.D., Paul Casadonte, M.D.	Sep 02, 2015
CTN (CTN-0027)	Starting Treatment with Agonist Replacement Therapies (START)	Walter Ling, M.D.	Jul 30, 2009
CTN (CTN-0030)	A Two-Phase Randomized Controlled Clinical Trial of Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling for Opioid Analgesic Dependence	Walter Ling, M.D., Roger Weiss, M.D.	Jun 22, 2011

A.3 Model Analysis for CSP-999 Trial

A.3.1 Patient Characteristics

The data consists of 15,983 urinalysis results from 654 subjects who participated in the CSP-999 trial. The number of urinalyses per subject ranged from 1 to 59, while the mean number of urinalyses per subject was 24.44 and the median was 17. Summaries of the demographic, sociodemographic, and substance use variables are given in Table 5. Several of the variables summarized in Table 2.1 are no longer available when only examining CSP-999 trial In particular, CSP-999 trial patients are either white, Hispanic, black, American Indian, or Asian and hence, the race categorized as "Other" was removed. None of the patients work for the military, so the work type category "Military" was removed. All of the patients have a stable living arrangement, so the "No Stable" living arrangement category was removed. All CSP-999 trial patients use heroine, so the "Heroine Use" categorical variable was removed. Finally, none of the patients' chosen mode of opioid abuse was sublingual, so the "Sublingual" category for the mode of opiate abuse variable was removed. For ease of comparison, the reference group is the same as the one used for the full analysis. The mean age, income, and years of opioid use are 36.16 years, $19,454 per year and 11.56 years, respectively (presented in Table 5), while the mean dose is 7.49 mg/day (presented in Table 4).

A.3.2 Functional Generalized Linear Mixed Model

Through the model in (2.1), we relate the daily dose patterns leading up to the clinic visit with urinalysis, while controlling for the 17 demographic, sociodemographic, and substance use variables detailed in Table 5. Note that, because all of our data comes from the CSP-999 trial, the trial-specific random effects $\gamma_{1k(i)}$ are removed from the model.

Figure 1 summarizes the estimated coefficient function $\hat{\beta}(t)$ (solid black line), which represents the buprenorphine daily dose effect for the 15 days leading up to a clinic visit with urinalysis. For comparison purposes, the coefficient function estimated from the full (all trials) analysis is also plotted (solid red line). The 95% credible intervals estimated from the full and reduced analyses are also plotted (red and black dashed line, respectively). Table 6 reports the demographic and substance use variables that were found to be significant. Of the 49 variable fixed effects, 5 were deemed to be important by the model (i.e., their estimated credible intervals did not contain zero). Table 6 summarizes these significant factors by reporting the estimated posterior mean (point estimate of

the effect), estimated standard deviation of the posterior (measure of uncertainty), and 95% equal-tailed credible interval for each parameter. The analogous results for the full set of demographic and substance use variables are provided in Tables 2 and 3.

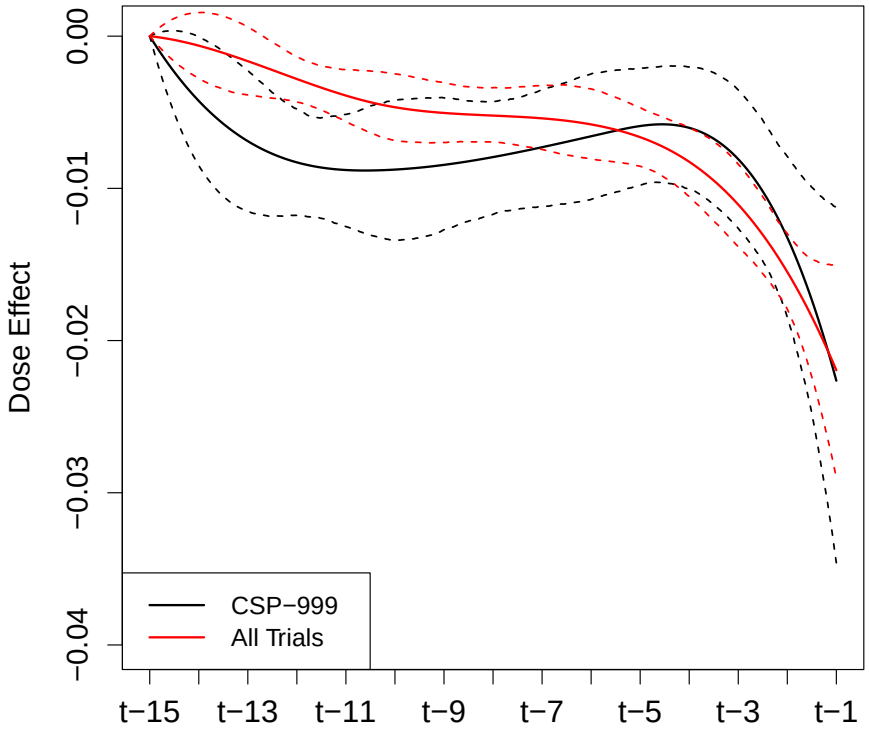

Figure 1: Estimated buprenorphine dose effect for the 15 days leading up to a urinalysis test, with 95% equal-tailed credible interval limits.

Table 2: Analysis results: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95).

	All Trials			CSP-999		
Variable	**Est**	**ESE**	**CI95**	**Est**	**ESE**	**CI95**
Intercept	**2.05**	**0.27**	**(1.54, 2.61)**	1.49	0.77	(-0.07, 2.96)
Age	**-0.02**	**0.01**	**(-0.03, -0.01)**	-0.02	0.02	(-0.05, 0.03)
Gender (Ref: Male)						
Female	0.13	0.08	(-0.03, 0.30)	0.43	0.26	(-0.03, 0.94)
Race (Ref: White)						
Black	-0.16	0.14	(-0.43, 0.12)	**-1.27**	**0.32**	**(-1.89, -0.64)**
American Indian	0.14	0.27	(-0.39, 0.68)	-0.62	0.88	(-2.83, 0.83)
Asian	-0.14	0.27	(-0.72, 0.35)	-1.31	1.44	(-4.83, 0.67)
Hispanic	-0.22	0.12	(-0.46, 0.02)	0.02	0.24	(-0.47, 0.52)
Other	-0.32	0.67	(-1.79, 0.93)			
Education (Ref: High School)						
Graduate School	-0.34	0.24	(-0.81, 0.11)	-1.27	0.86	(-2.98, 0.21)
Standard College	0.05	0.16	(-0.25, 0.37)	-0.11	0.34	(-0.79, 0.59)
Partial College	-0.18	0.10	(-0.39, 0.02)	-0.04	0.21	(-0.47, 0.37)
Partial High School	-0.27	0.15	(-0.58, 0.01)	-0.03	0.27	(-0.63, 0.44)
Junior High School	-0.20	0.21	(-0.63, 0.18)	-0.18	0.34	(-0.98, 0.44)
Less than 7th Grade	-0.35	0.47	(-1.39, 0.47)	-0.48	0.84	(-2.53, 0.88)
Emp. History (Ref: Skilled)						
Never Gainfully	0.07	0.14	(-0.20, 0.34)	0.02	0.32	(-0.62, 0.70)
Unskilled	-0.20	0.16	(-0.52, 0.09)	-0.22	0.36	(-0.97, 0.44)
Machine Operator	-0.01	0.13	(-0.26, 0.24)	-0.18	0.28	(-0.76, 0.34)
Clerical/Sales	0.11	0.12	(-0.14, 0.35)	0.07	0.28	(-0.51, 0.64)
Administrative	-0.04	0.15	(-0.35, 0.24)	0.04	0.32	(-0.55, 0.73)
Business Manager	-0.24	0.24	(-0.72, 0.22)	-0.93	0.93	(-3.03, 0.48)
Executive	-0.18	0.26	(-0.71, 0.29)	0.05	1.04	(-2.03, 2.47)
Work Type (Ref: Fulltime)						
Regular PT	-0.10	0.15	(-0.39, 0.18)	-0.42	0.34	(-1.13, 0.21)
Irregular PT	0.07	0.13	(-0.19, 0.35)	0.29	0.30	(-0.25, 0.91)
Student	0.06	0.26	(-0.44, 0.61)	0.17	0.72	(-1.38, 1.67)
Military	-0.28	1.22	(-3.25, 2.02)			
Retired	0.23	0.24	(-0.19, 0.71)	0.21	0.79	(-1.13, 2.08)
Unemployed	**0.33**	**0.14**	**(0.05, 0.59)**	0.35	0.34	(-0.27, 1.05)
Controlled	0.84	0.59	(-0.2, 2.11)	1.18	0.97	(-0.30, 3.30)
Income	0.00	0.00	(0.00, 0.00)	0.00	0.00	(0.00, 0.00)
Marital Status (Ref: Married)						
Remarried	-0.17	0.39	(-1.01, 0.59)	0.56	0.93	(-1.07, 2.67)
Widowed	0.19	0.25	(-0.25, 0.73)	0.30	0.52	(-0.61, 1.38)
Separated	0.19	0.17	(-0.10, 0.55)	0.15	0.33	(-0.46, 0.81)
Divorced	0.04	0.12	(-0.20, 0.28)	0.09	0.28	(-0.44, 0.66)
Never Married	0.19	0.12	(-0.05, 0.42)	0.35	0.26	(-0.08, 0.89)
Living Arr (Ref: Partner & Child)						
Partner Only	0.24	0.13	(-0.02, 0.49)	**0.79**	**0.30**	**(0.17, 1.38)**
Child Only	-0.03	0.16	(-0.36, 0.29)	-0.19	0.31	(-0.82, 0.42)
Parents	0.21	0.17	(-0.10, 0.53)	0.28	0.29	(-0.23, 0.87)
Family	0.05	0.16	(-0.27, 0.36)	-0.22	0.34	(-0.96, 0.38)
Friends	-0.14	0.16	(-0.44, 0.17)	-0.07	0.30	(-0.74, 0.50)
Alone	-0.03	0.18	(-0.42, 0.31)	0.10	0.70	(-1.40, 1.60)
Controlled	-0.01	0.37	(-0.77, 0.75)	0.98	1.02	(-0.63, 3.24)
No Stable	-0.15	0.36	(-0.93, 0.52)			

Table 3: Analysis results: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95).

	All Trials			CSP-999		
Variable	**Est**	**ESE**	**CI95**	**Est**	**ESE**	**CI95**
Heroine Use (Ref: YES)						
NO	**-0.57**	**0.21**	**(-1.00, -0.16)**			
Years of Opiate Use	0.01	0.01	(-0.01, 0.02)	0.01	0.02	(-0.02, 0.04)
Mode of Opioid Abuse (Ref: IV)						
Oral	**-1.32**	**0.23**	**(-1.78, -0.88)**	**-1.21**	**0.48**	**(-2.11, -0.21)**
Snorting	-0.14	0.10	(-0.35, 0.06)	-0.15	0.23	(-0.60, 0.29)
Smoking	-0.40	0.28	(-0.94, 0.09)	-0.57	0.78	(-2.37, 0.69)
Sublingual	-1.34	0.97	(-3.42, 0.16)			
Other	0.02	0.41	(-0.82, 0.89)	-0.42	0.72	(-2.08, 0.74)
Cocaine Use (Ref: YES)						
NO	-0.04	0.09	(-0.22, 0.12)	-0.17	0.28	(-0.69, 0.37)
Meth Use (Ref: NO)						
YES	0.13	0.09	(-0.05, 0.31)	**0.77**	**0.28**	**(0.24, 1.31)**
Alcohol Use (Ref: YES)						
NO	-0.12	0.10	(-0.33, 0.07)	**-0.49**	**0.23**	**(-0.97, -0.04)**
Tranquilizer Use (Ref: NO)						
YES	0.07	0.10	(-0.14, 0.25)	0.38	0.24	(-0.06, 0.85)
Marijuana Use (Ref: YES)						
NO	0.05	0.10	(-0.16, 0.25)	0.20	0.23	(-0.27, 0.64)
PCP Use (Ref: NO)						
YES	-0.01	0.11	(-0.22, 0.21)	-0.07	0.28	(-0.60, 0.54)

Table 4: Treatment and outcome characteristics of individuals used in the trial CSP-999 analysis.

	Mean	Median	Range
Daily Dose	7.49	4	0-64
Time in Trial	148.76	133	3-527
Urinalysis (Yes=1)	0.51	1	0-1

Table 5: Sociodemographic characteristics and drug use history for the individuals used in the CSP-999 trial analysis.

Demographics		
Age	Mean	SD
	36.16	7.78
Gender	N	%
Male	453	69
Female	201	31
Race	N	%
White	309	47
Hispanic	184	28
Black	152	23
American Indian	5	1
Asian	4	1
Other	0	0

Drug Use History		
Years of Opiate Abuse	Mean	SD
	11.56	8.70
Heroin Use	N	%
YES	654	100
NO	0	0
Mode of Opiate Abuse	N	%
IV	406	62
Snort	194	30
Oral	37	6
Smoking	9	1
Other	8	1
Sublingual	0	0
Cocaine Use	N	%
YES	540	83
NO	114	17
Meth Use	N	%
NO	436	67
YES	218	33
Alcohol Use	N	%
YES	446	68
NO	208	32
Tranquilizer Use	N	%
NO	353	54
YES	301	46
Marijuana Use	N	%
YES	482	74
NO	172	26
PCP Use	N	%
NO	533	82
YES	121	18

Sociodemographics		
Income	Mean	SD
	19454	21365
Employment History	N	%
Skilled Manual	151	23
Never Gainfully	168	26
Machine Operator	83	13
Clerical/Sales	124	19
Administrative	54	8
Unskilled	64	10
Business Manager	7	1
Executive	3	< 1
Work Type	N	%
Fulltime	294	45
Unemployed	198	30
Irregular PT	81	12
Regular PT	58	9
Retired	7	1
Student	7	1
Controlled	9	1
Military	0	0
Education	N	%
High School	213	33
Partial College	201	31
Partial High School	119	18
Standard College	48	7
Junior High School	56	8
Complete Graduate School	11	2
Less than 7th Grade	6	1
Marital Status	N	%
Married	170	26
Never Married	256	39
Divorced	142	22
Separated	64	10
Widowed	18	3
Remarried	4	< 1
Living Arr	N	%
Partner & Child	163	25
Partner Only	126	19
Parents	116	18
Family	55	8
Friends	97	15
Alone	8	1
Child Only	82	13
Controlled	7	1
No Stable	0	0

Table 6: Sensitivity analysis results for CSP-999: Summary includes the posterior mean estimates (Est), the estimated posterior standard deviations (ESE), and the estimated 95% equal-tailed credible intervals (CI95) for the significant fixed effects.

Variable	**Est**	**ESE**	**CI95**
Race (Ref: White)			
Black	-1.27	0.32	(-1.89, -0.64)
Living Arr (Ref: Partner & Child)			
Partner Only	0.79	0.30	(0.17, 1.38)
Mode of Opioid Abuse (Ref: IV)			
Oral	-1.21	0.48	(-2.11, -0.21)
Meth Use (Ref: NO)			
YES	0.77	0.28	(0.24, 1.31)
Alcohol Use (Ref: YES)			
NO	-0.49	0.23	(-0.97, -0.04)

Appendix B Supplementary Material for Chapter 3

B.1 Posterior Distributions

To begin, we introduce some notation. Let $N = \sum_{i=1}^{m} n_i$ be the total number of observations in the data set. Aggregate ocScores, Y_{ij}, and stable flashback indicators, S_{ij}, into the N-dimensional vectors denoted by $\boldsymbol{Y}$ and $\boldsymbol{S}$, respectively; and aggregate the subject-specific random effects, γ_i, into the m-dimensional vector denoted by $\boldsymbol{\gamma}$. Let $\boldsymbol{X}$ be the $N \times (P+1)$ covariate matrix aggregating the covariate vectors $\boldsymbol{x}_{ij}$ used in the ocScore model; let $\boldsymbol{C}$ be the $N \times (L+1)$ input matrix derived from the modified principal components with sparse loadings used in the stable flashback SPCA model, as described in Section 3.2.1.1; and let $\boldsymbol{U}$ be the $N \times m$ block diagonal, design matrix associated with the random effects, $\boldsymbol{\gamma}$.

For both ocScore Y_{ij} and stable flashback indicator S_{ij}, we assume conditional independence given their respective covariate effects, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and random effects, γ_i. Hence, the likelihood can be expressed as

$$p(\boldsymbol{Y}, \boldsymbol{S} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto (2\pi\sigma_\epsilon^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})\right\} \times \prod_{i,j} \pi_{ij}^{S_{ij}} \{1 - \pi_{ij}\}^{k_{ij} - S_{ij}},$$

where $\pi_{ij} = P(S_{ij(k)} = 1 \mid \boldsymbol{z}_{ij}) = [1 + \exp(-\nu_{ij})]^{-1}$, and $\nu_{ij} = \boldsymbol{c}_{ij}'\boldsymbol{\theta} + \zeta\gamma_i$ under the SPCA model (3.3).

We develop a two-stage data augmentation process to construct a posterior sampling algorithm consisting only of Gibbs steps. In the first stage, we exploit a hierarchical representation of the proposed data model by introducing Póyla - Gamma latent random variables ω_{ij}; for further details see Polson et al. [2013]. Under this specification, the joint density of the observed and latent data is given by

$$p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\omega} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto (2\pi\sigma_\epsilon^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})\right\} \times \exp\left\{-\frac{1}{2}(\boldsymbol{h} - \boldsymbol{\nu})'\boldsymbol{\Omega}(\boldsymbol{h} - \boldsymbol{\nu})\right\} \times \prod_{i,j} \xi(\omega_{ij}),$$

where $\boldsymbol{h}$ is the N-dimensional vector aggregating the synthetic responses defined as $h_{ij} = \eta_{ij}/\omega_{ij}$, with $\eta_{ij} = S_{ij} - k_{ij}/2$; where $\boldsymbol{\omega}$ is the N-dimensional vector aggregating the latent variables ω_{ij}

and $\Omega = \text{diag}(\boldsymbol{\omega})$; where $\xi(\omega_{ij}) = f(\omega_{ij} \mid k_{ij}, 0) \exp\{\eta_{ij}^2/(2\omega_{ij})\}$, and $f(\omega_{ij} \mid a, b)$ denotes the Póyla-Gamma density with parameters (a, b) (for further details, see Polson et al. [2013]); and where $\boldsymbol{\nu}$ is the N-dimensional vector aggregating the linear predictors ν_{ij}; that is, $\boldsymbol{\nu} = \boldsymbol{C\theta} + \boldsymbol{U}(\zeta\boldsymbol{\gamma})$.

We now highlight the second stage of the data augmentation process and the construction of the hierarchical representation of the joint posterior distribution. Recall from Section 3.2.2, we specify a generalized double Pareto shrinkage prior for all of the regression coefficients with the exception of the intercept; i.e., for the coefficients $\boldsymbol{\beta}$, we specify

$$\begin{aligned}
&\beta_0 \sim N(0,\ \sigma_\epsilon^2 \tau_0) \\
&\beta_p \sim GDP(\sigma_\epsilon b_\delta / a_\delta,\ a_\delta), \text{ for } p = 1, \ldots, P,
\end{aligned}$$

and, for the coefficients $\boldsymbol{\theta}$, we specify

$$\begin{aligned}
&\theta_0 \sim N(0,\ \rho_0) \\
&\theta_l \sim GDP(b_\lambda / a_\lambda,\ a_\lambda), \text{ for } l = 1, \ldots, L.
\end{aligned}$$

For computational simplifications, the generalized double Pareto shrinkage prior can be represented as a scale mixture of normal distributions Armagan et al. [2013]. Thus, for the regression coefficients $\boldsymbol{\beta}$, the following hierarchical representation provides the same prior specifications as those given above for $\boldsymbol{\beta}$:

$$\begin{aligned}
&\boldsymbol{\beta} \sim N(\mathbf{0},\ \sigma_\epsilon^2 \boldsymbol{T}), \text{ where } \boldsymbol{T} = \text{diag}(\boldsymbol{\tau}),\ \ \boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_P)' \\
&\tau_p \overset{\text{ind.}}{\sim} Exponential(\delta_p^2/2), \text{ for } p = 1, \ldots, P \\
&\delta_p \overset{\text{ind.}}{\sim} Gamma(a_\delta,\ b_\delta), \text{ for } p = 1, \ldots, P.
\end{aligned}$$

Similarly, for the regression coefficients $\boldsymbol{\theta}$, the following hierarchical representation provides the same prior specifications as those given above for $\boldsymbol{\theta}$:

$$\begin{aligned}
&\boldsymbol{\theta} \sim N(\mathbf{0},\ \boldsymbol{R}), \text{ where } \boldsymbol{R} = \text{diag}(\boldsymbol{\rho}),\ \ \boldsymbol{\rho} = (\rho_0, \rho_1, \ldots, \rho_L)' \\
&\rho_l \overset{\text{ind.}}{\sim} Exponential(\lambda_l^2/2) \text{ , for } l = 1, \ldots, L \\
&\lambda_l \overset{\text{ind.}}{\sim} Gamma(a_\lambda,\ b_\lambda), \text{ for } l = 1, \ldots, L.
\end{aligned}$$

Attention is now turned to deriving the full conditional distributions based on the two-stage

data augmentation scheme outlined above. For notational convenience, a dot $\cdot$ is used as shorthand for all the parameters one is conditioning on; e.g., we may write the posterior $p(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{\gamma}, \sigma_\epsilon^2, \boldsymbol{\tau})$ as $p(\boldsymbol{\beta} \mid \cdot)$.

We begin with the derivation of the full conditional posterior distribution of $\boldsymbol{\beta}$ and the associated hyperparameters. Define $\boldsymbol{\Sigma_\beta} = \left(\boldsymbol{X'X} + \boldsymbol{T}^{-1}\right)^{-1}$ and $\boldsymbol{\mu_\beta} = \boldsymbol{\Sigma_\beta X'}\left(\boldsymbol{Y} - \boldsymbol{U}(\zeta\boldsymbol{\gamma})\right)$. Then,

$$\boldsymbol{\beta} \mid \cdot \sim N(\boldsymbol{\mu_\beta},\ \sigma_\epsilon^2 \boldsymbol{\Sigma_\beta}).$$

Defining $\mu_{\tau_p} = \sqrt{\sigma_\epsilon^2 \delta_p^2 / \beta_p^2}$, we have that

$$\tau_p^{-1} \mid \cdot \sim Inv - Gaussian(\mu_{\tau_p},\ \delta_p^2), \text{ for } p = 1, \ldots, P.$$

Define $a^*_{\delta_p} = a_\delta + 1$ and $b^*_{\delta_p} = b_\delta + |\beta_p|/\sigma_\epsilon$. Then,

$$\delta_p \mid \cdot \sim Gamma(a^*_{\delta_p},\ b^*_{\delta_p}), \text{ for } p = 1, \ldots, P.$$

Finally, let $\widetilde{a}_\delta = 1/(1 + a_\delta)$ and $\widetilde{b}_\delta = 1/(1 + b_\delta)$ be transformations of a_δ and b_δ, respectively. Given the generalized Pareto hyper-priors on a_δ and b_δ from Section 3.2.2, these transformations suggest uniform priors on $\widetilde{a}_\delta$ and $\widetilde{b}_\delta$ in (0, 1). As a result, the conditional posteriors for $\widetilde{a}_\delta$ and $\widetilde{b}_\delta$ are

$$\begin{aligned} p\left(\widetilde{a}_\delta \mid \boldsymbol{\beta}, b_\delta\right) &\propto \left(\frac{1 - \widetilde{a}_\delta}{\widetilde{a}_\delta}\right)^P \prod_{p=1}^{P} \left(1 + \frac{|\beta_p|}{\sigma_\epsilon b_\delta}\right)^{-1/\widetilde{a}_\delta} \\ p\left(\widetilde{b}_\delta \mid \boldsymbol{\beta}, a_\delta\right) &\propto \left(\frac{\widetilde{b}_\delta}{1 - \widetilde{b}_\delta}\right)^P \prod_{p=1}^{P} \left\{1 + \widetilde{b}_\delta \frac{|\beta_p|}{\sigma_\epsilon(1 - \widetilde{b}_\delta)}\right\}^{-(a_\delta + 1)} \end{aligned}.$$

To continue, we turn our attention to the posterior distribution of $\boldsymbol{\theta}$ and the associated hyperparameters. Define $\boldsymbol{\Sigma_\theta} = \left(\boldsymbol{C'\Omega C} + \boldsymbol{R}^{-1}\right)^{-1}$ and $\boldsymbol{\mu_\theta} = \boldsymbol{\Sigma_\theta C'\Omega}(\boldsymbol{h} - \boldsymbol{U\gamma})$. Then,

$$\boldsymbol{\theta} \mid \cdot \sim N(\boldsymbol{\mu_\theta},\ \boldsymbol{\Sigma_\theta}).$$

Defining $\mu_{\rho_l} = \sqrt{\lambda_l^2 / \theta_l^2}$, we have that

$$\rho_l^{-1} \mid \cdot \sim Inv - Gaussian(\mu_{\rho_l},\ \lambda_l^2), \text{ for } l = 1, \ldots, L.$$

Finally, define $a^*_{\lambda_l} = a_\lambda + 1$ and $b^*_{\lambda_l} = b_\lambda + |\theta_l|$. Then,

$$\lambda_l \mid \cdot \sim Gamma(a^*_{\lambda_l},\ b^*_{\lambda_l}), \text{ for } l = 1, \ldots, L.$$

Next, we turn our focus to the posterior distribution of $\boldsymbol{\gamma}$ and its variance component σ^2_γ. First, define

$$\boldsymbol{\Sigma_\gamma} = \left\{\sigma_\epsilon^{-2}\boldsymbol{U}'\boldsymbol{U} + (\boldsymbol{U}\zeta)'\boldsymbol{\Omega}(\boldsymbol{U}\zeta) + \sigma_\gamma^{-2}\boldsymbol{I}_m\right\}^{-1},$$

where $\boldsymbol{I}_m$ refers to the identity matrix of dimension m, and define

$$\boldsymbol{\mu_\gamma} = \boldsymbol{\Sigma_\gamma}\left\{\sigma_\epsilon^{-2}\boldsymbol{U}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{U}\zeta)'\boldsymbol{\Omega}(\boldsymbol{h} - \boldsymbol{C}\boldsymbol{\theta})\right\}.$$

Then, the full conditional posterior distribution of $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma} \mid \cdot \sim N(\boldsymbol{\mu_\gamma},\ \boldsymbol{\Sigma_\gamma}).$$

Define $a^*_\gamma = a_\gamma + m/2$ and $b^*_\gamma = b_\gamma + 0.5\boldsymbol{\gamma}'\boldsymbol{\gamma}$. Then,

$$\sigma^2_\gamma \mid \cdot \sim Inv - Gamma(a^*_\gamma,\ b^*_\gamma).$$

We now turn our attention to the posterior distribution of the random effects' association parameter, ζ. Define $\Sigma_\zeta = \left((\boldsymbol{U}\boldsymbol{\gamma})'\boldsymbol{\Omega}(\boldsymbol{U}\boldsymbol{\gamma}) + \sigma_\zeta^{-2}\right)^{-1}$, and $\mu_\zeta = \Sigma_\zeta(\boldsymbol{U}\boldsymbol{\gamma})'\boldsymbol{\Omega}(\boldsymbol{h} - \boldsymbol{C}\boldsymbol{\theta})$. Then,

$$\zeta \mid \cdot \sim N(\mu_\zeta,\ \Sigma_\zeta).$$

We end our discuss by describing the posterior distribution of the variance component for the random error in model (3.1); i.e., the posterior of σ^2_ϵ. Define $a^*_\epsilon = (N + P + 1)/2$ and $b^*_\epsilon = 0.5(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})$. Then,

$$\sigma^2_\epsilon \mid \cdot \sim Inv - Gamma(a^*_\epsilon,\ b^*_\epsilon).$$

B.2 Additional Simulation Results

Here is an outline of the material in this section of the appenidx:

Pages 78 - 81: Tables 7 - 10. Simulation results for regression coefficients, $\boldsymbol{\beta}$, in the ocScore model (3.4).

Page 82: Table 11. Simulation results for regression coefficients, $\boldsymbol{\alpha}$, in the stable flashback model (3.5).

Table 7: Simulation results for regression coefficients, $\boldsymbol{\beta}$, corresponding to the intercept and the 80 truly significant covariates.

True	**Bias (CP95)**	**SSD (ESE)**	**True**	**Bias (CP95)**	**SSD (ESE)**
0.45	-0.006 (0.932)	0.157 (0.148)	1.00	-0.059 (0.922)	0.159 (0.151)
1.00	-0.055 (0.922)	0.160 (0.151)	-1.00	0.055 (0.926)	0.157 (0.151)
1.00	-0.053 (0.940)	0.154 (0.151)	-1.00	0.044 (0.926)	0.165 (0.151)
1.00	-0.053 (0.938)	0.146 (0.151)	-1.00	0.053 (0.924)	0.167 (0.151)
1.00	-0.054 (0.930)	0.156 (0.151)	-1.00	0.044 (0.904)	0.162 (0.151)
1.00	-0.052 (0.920)	0.161 (0.151)	-1.00	0.061 (0.942)	0.148 (0.151)
1.00	-0.063 (0.912)	0.157 (0.151)	-1.00	0.059 (0.920)	0.163 (0.151)
1.00	-0.053 (0.904)	0.158 (0.151)	-1.00	0.057 (0.928)	0.155 (0.151)
1.00	-0.063 (0.886)	0.167 (0.151)	-1.00	0.053 (0.944)	0.149 (0.151)
1.00	-0.051 (0.962)	0.144 (0.151)	-1.00	0.054 (0.918)	0.156 (0.151)
1.00	-0.061 (0.930)	0.151 (0.151)	-1.00	0.054 (0.928)	0.161 (0.151)
1.00	-0.039 (0.946)	0.153 (0.151)	-1.00	0.054 (0.924)	0.159 (0.151)
1.00	-0.044 (0.920)	0.163 (0.151)	-1.00	0.060 (0.906)	0.168 (0.152)
1.00	-0.063 (0.922)	0.157 (0.151)	-1.00	0.064 (0.916)	0.165 (0.151)
1.00	-0.052 (0.926)	0.158 (0.151)	-1.00	0.054 (0.936)	0.156 (0.151)
1.00	-0.050 (0.930)	0.157 (0.151)	-1.00	0.052 (0.930)	0.157 (0.151)
1.00	-0.055 (0.910)	0.162 (0.151)	-1.00	0.053 (0.930)	0.153 (0.151)
1.00	-0.054 (0.910)	0.165 (0.151)	-1.00	0.054 (0.934)	0.155 (0.151)
1.00	-0.061 (0.924)	0.157 (0.151)	-1.00	0.044 (0.934)	0.149 (0.151)
1.00	-0.043 (0.934)	0.156 (0.151)	-1.00	0.062 (0.926)	0.162 (0.151)
1.00	-0.051 (0.902)	0.167 (0.151)	-1.00	0.048 (0.930)	0.162 (0.150)
1.00	-0.044 (0.922)	0.160 (0.151)	-1.00	0.053 (0.936)	0.153 (0.151)
1.00	-0.055 (0.912)	0.165 (0.151)	-1.00	0.057 (0.926)	0.154 (0.150)
1.00	-0.048 (0.926)	0.156 (0.151)	-1.00	0.042 (0.926)	0.156 (0.151)
1.00	-0.062 (0.898)	0.164 (0.151)	-1.00	0.062 (0.922)	0.159 (0.151)
1.00	-0.059 (0.910)	0.150 (0.151)	-1.00	0.056 (0.928)	0.154 (0.151)
1.00	-0.054 (0.926)	0.148 (0.151)	-1.00	0.045 (0.928)	0.157 (0.151)
1.00	-0.054 (0.924)	0.156 (0.151)	-1.00	0.056 (0.922)	0.165 (0.151)
1.00	-0.059 (0.948)	0.147 (0.151)	-1.00	0.051 (0.932)	0.156 (0.151)
1.00	-0.062 (0.912)	0.162 (0.151)	-1.00	0.062 (0.930)	0.149 (0.151)
1.00	-0.066 (0.920)	0.158 (0.151)	-1.00	0.051 (0.932)	0.154 (0.151)
1.00	-0.038 (0.932)	0.161 (0.150)	-1.00	0.055 (0.918)	0.162 (0.151)
1.00	-0.056 (0.920)	0.154 (0.151)	-1.00	0.058 (0.944)	0.149 (0.151)
1.00	-0.064 (0.918)	0.157 (0.151)	-1.00	0.049 (0.934)	0.151 (0.151)
1.00	-0.049 (0.930)	0.149 (0.151)	-1.00	0.051 (0.932)	0.157 (0.151)
1.00	-0.054 (0.932)	0.157 (0.151)	-1.00	0.051 (0.928)	0.159 (0.151)
1.00	-0.061 (0.912)	0.160 (0.151)	-1.00	0.052 (0.936)	0.156 (0.151)
1.00	-0.047 (0.936)	0.155 (0.151)	-1.00	0.045 (0.918)	0.160 (0.151)
1.00	-0.054 (0.922)	0.164 (0.151)	-1.00	0.048 (0.916)	0.158 (0.151)
1.00	-0.049 (0.920)	0.165 (0.151)	-1.00	0.058 (0.940)	0.157 (0.151)
-1.00	-.053 (0.940)	0.152 (0.151)			

Table 8: Simulation results for regression coefficients, $\boldsymbol{\beta}$.

True	**Bias (CP95)**	**SSD (ESE)**	**True**	**Bias (CP95)**	**SSD (ESE)**	**True**	**Bias (CP95)**	**SSD (ESE)**
0.00	-0.001 (0.996)	0.080 (0.110)	0.00	-0.001 (0.998)	0.081 (0.110)	0.00	-0.006 (0.980)	0.093 (0.111)
0.00	0.003 (0.984)	0.089 (0.111)	0.00	0.007 (0.990)	0.089 (0.111)	0.00	0.005 (0.986)	0.093 (0.112)
0.00	0.005 (0.984)	0.088 (0.110)	0.00	0.002 (0.986)	0.092 (0.112)	0.00	-0.009 (0.988)	0.088 (0.111)
0.00	0.004 (0.970)	0.098 (0.112)	0.00	0.003 (0.978)	0.093 (0.112)	0.00	0.002 (0.990)	0.083 (0.110)
0.00	0.003 (0.988)	0.087 (0.111)	0.00	0.006 (0.980)	0.095 (0.111)	0.00	0.002 (0.998)	0.074 (0.109)
0.00	0.001 (0.988)	0.089 (0.111)	0.00	-0.006 (0.984)	0.093 (0.111)	0.00	-0.001 (0.982)	0.096 (0.111)
0.00	-0.008 (0.988)	0.093 (0.112)	0.00	-0.004 (0.988)	0.090 (0.110)	0.00	0.002 (0.994)	0.084 (0.111)
0.00	-0.000 (0.986)	0.090 (0.111)	0.00	-0.001 (0.986)	0.091 (0.112)	0.00	-0.002 (0.984)	0.093 (0.112)
0.00	-0.001 (0.984)	0.089 (0.111)	0.00	-0.002 (0.978)	0.092 (0.111)	0.00	0.001 (0.992)	0.082 (0.110)
0.00	0.002 (0.986)	0.090 (0.112)	0.00	-0.004 (0.978)	0.095 (0.111)	0.00	0.000 (0.982)	0.088 (0.111)
0.00	-0.006 (0.992)	0.090 (0.112)	0.00	0.003 (0.998)	0.082 (0.111)	0.00	-0.004 (0.990)	0.086 (0.111)
0.00	-0.006 (0.988)	0.091 (0.112)	0.00	0.005 (0.978)	0.096 (0.111)	0.00	-0.004 (0.982)	0.096 (0.112)
0.00	-0.002 (0.996)	0.086 (0.111)	0.00	0.002 (0.988)	0.087 (0.111)	0.00	-0.004 (0.990)	0.083 (0.110)
0.00	0.001 (0.992)	0.093 (0.112)	0.00	0.003 (0.992)	0.090 (0.111)	0.00	-0.001 (0.984)	0.089 (0.111)
0.00	-0.003 (0.984)	0.093 (0.112)	0.00	-0.003 (0.986)	0.089 (0.111)	0.00	-0.002 (0.986)	0.088 (0.111)
0.00	0.002 (0.990)	0.088 (0.111)	0.00	-0.005 (0.984)	0.090 (0.111)	0.00	0.001 (0.988)	0.089 (0.111)
0.00	0.005 (0.994)	0.084 (0.111)	0.00	-0.006 (0.988)	0.092 (0.112)	0.00	0.001 (0.978)	0.094 (0.112)
0.00	0.004 (0.990)	0.090 (0.111)	0.00	0.001 (0.994)	0.083 (0.110)	0.00	0.002 (0.992)	0.088 (0.111)
0.00	-0.002 (0.976)	0.091 (0.111)	0.00	0.001 (0.992)	0.090 (0.111)	0.00	-0.007 (0.996)	0.087 (0.112)
0.00	0.003 (0.992)	0.088 (0.111)	0.00	-0.006 (0.984)	0.090 (0.112)	0.00	0.004 (0.986)	0.091 (0.111)
0.00	-0.002 (0.992)	0.090 (0.112)	0.00	0.002 (0.986)	0.091 (0.111)	0.00	-0.004 (0.990)	0.090 (0.111)
0.00	-0.006 (0.990)	0.088 (0.111)	0.00	0.005 (0.988)	0.092 (0.112)	0.00	-0.008 (0.986)	0.089 (0.111)
0.00	0.001 (0.986)	0.089 (0.111)	0.00	-0.006 (0.990)	0.087 (0.111)	0.00	0.003 (0.988)	0.086 (0.110)
0.00	0.000 (0.990)	0.090 (0.111)	0.00	-0.001 (0.994)	0.082 (0.110)	0.00	0.002 (0.984)	0.093 (0.112)
0.00	-0.003 (0.998)	0.082 (0.111)	0.00	-0.002 (0.992)	0.083 (0.110)	0.00	-0.002 (0.990)	0.087 (0.111)
0.00	0.004 (0.986)	0.094 (0.112)	0.00	-0.003 (0.992)	0.086 (0.111)	0.00	0.002 (0.986)	0.083 (0.110)
0.00	0.002 (0.986)	0.088 (0.111)	0.00	-0.004 (0.986)	0.090 (0.111)	0.00	0.000 (0.988)	0.091 (0.111)
0.00	0.002 (0.988)	0.088 (0.112)	0.00	-0.002 (0.982)	0.091 (0.111)	0.00	0.002 (0.990)	0.089 (0.111)
0.00	0.006 (0.996)	0.081 (0.110)	0.00	0.005 (0.988)	0.086 (0.111)	0.00	-0.001 (0.994)	0.090 (0.112)
0.00	-0.003 (0.986)	0.089 (0.111)	0.00	-0.001 (0.992)	0.084 (0.111)	0.00	0.007 (0.986)	0.090 (0.111)
0.00	0.001 (0.980)	0.095 (0.112)	0.00	0.001 (0.996)	0.081 (0.110)	0.00	-0.000 (0.996)	0.084 (0.111)
0.00	-0.004 (0.988)	0.090 (0.111)	0.00	0.000 (0.990)	0.087 (0.111)	0.00	0.006 (0.990)	0.090 (0.111)
0.00	-0.003 (0.990)	0.088 (0.111)	0.00	0.003 (0.994)	0.083 (0.110)	0.00	-0.002 (0.994)	0.077 (0.110)
0.00	-0.005 (0.988)	0.089 (0.111)	0.00	-0.004 (0.988)	0.086 (0.110)	0.00	0.001 (0.990)	0.092 (0.112)
0.00	-0.003 (0.984)	0.089 (0.111)	0.00	0.001 (0.990)	0.087 (0.111)	0.00	0.001 (0.990)	0.097 (0.113)
0.00	-0.001 (0.992)	0.088 (0.111)	0.00	-0.000 (0.984)	0.088 (0.111)	0.00	-0.002 (0.990)	0.090 (0.111)
0.00	0.007 (0.986)	0.094 (0.112)	0.00	0.003 (0.988)	0.089 (0.111)	0.00	0.002 (0.980)	0.092 (0.111)
0.00	0.007 (0.990)	0.090 (0.111)	0.00	-0.000 (0.996)	0.085 (0.111)	0.00	0.002 (0.992)	0.089 (0.111)
0.00	0.001 (0.990)	0.084 (0.110)	0.00	-0.003 (0.984)	0.096 (0.112)	0.00	0.003 (0.992)	0.084 (0.111)
0.00	0.002 (0.994)	0.085 (0.111)	0.00	0.003 (0.984)	0.089 (0.111)	0.00	0.002 (0.994)	0.082 (0.110)

Table 9: Simulation results for regression coefficients, $\boldsymbol{\beta}$.

True	**Bias (CP95)**	**SSD (ESE)**	**True**	**Bias (CP95)**	**SSD (ESE)**	**True**	**Bias (CP95)**	**SSD (ESE)**
0.00	-0.005 (0.982)	0.093 (0.111)	0.00	0.007 (0.992)	0.084 (0.111)	0.00	-0.007 (0.992)	0.085 (0.111)
0.00	-0.004 (0.992)	0.089 (0.111)	0.00	0.002 (0.998)	0.084 (0.111)	0.00	0.000 (0.992)	0.084 (0.110)
0.00	0.001 (0.990)	0.088 (0.111)	0.00	-0.001 (0.984)	0.094 (0.112)	0.00	0.003 (0.990)	0.085 (0.111)
0.00	0.000 (0.996)	0.075 (0.109)	0.00	0.000 (0.992)	0.086 (0.110)	0.00	-0.003 (0.994)	0.082 (0.110)
0.00	-0.001 (0.990)	0.085 (0.111)	0.00	0.001 (0.984)	0.093 (0.111)	0.00	0.001 (0.988)	0.093 (0.111)
0.00	0.004 (0.988)	0.089 (0.111)	0.00	-0.003 (0.990)	0.090 (0.111)	0.00	-0.003 (0.986)	0.094 (0.112)
0.00	0.001 (0.982)	0.095 (0.112)	0.00	0.000 (0.976)	0.095 (0.111)	0.00	0.006 (0.982)	0.091 (0.111)
0.00	-0.001 (0.988)	0.089 (0.111)	0.00	0.004 (0.990)	0.084 (0.110)	0.00	-0.000 (0.980)	0.097 (0.112)
0.00	0.000 (0.980)	0.096 (0.112)	0.00	-0.001 (0.992)	0.086 (0.111)	0.00	-0.004 (0.986)	0.086 (0.111)
0.00	-0.004 (0.992)	0.081 (0.110)	0.00	-0.006 (0.988)	0.081 (0.110)	0.00	-0.002 (0.988)	0.095 (0.112)
0.00	0.004 (0.990)	0.088 (0.111)	0.00	-0.000 (0.990)	0.085 (0.111)	0.00	0.006 (0.988)	0.089 (0.111)
0.00	0.003 (0.992)	0.085 (0.110)	0.00	0.003 (0.982)	0.097 (0.112)	0.00	0.001 (0.988)	0.087 (0.111)
0.00	0.002 (0.994)	0.090 (0.111)	0.00	-0.005 (0.990)	0.089 (0.111)	0.00	0.003 (0.996)	0.082 (0.110)
0.00	0.001 (0.996)	0.081 (0.110)	0.00	-0.001 (0.990)	0.086 (0.111)	0.00	0.001 (0.984)	0.092 (0.111)
0.00	-0.001 (0.990)	0.082 (0.110)	0.00	-0.002 (0.992)	0.088 (0.111)	0.00	0.001 (0.996)	0.080 (0.110)
0.00	-0.008 (0.990)	0.077 (0.109)	0.00	0.007 (0.982)	0.092 (0.111)	0.00	0.003 (0.992)	0.093 (0.111)
0.00	-0.004 (0.996)	0.089 (0.112)	0.00	-0.002 (0.998)	0.080 (0.110)	0.00	0.006 (0.990)	0.086 (0.110)
0.00	-0.017 (0.990)	0.088 (0.111)	0.00	0.009 (0.986)	0.090 (0.112)	0.00	0.004 (0.984)	0.095 (0.112)
0.00	-0.001 (0.978)	0.092 (0.111)	0.00	-0.002 (0.988)	0.087 (0.111)	0.00	0.002 (0.994)	0.084 (0.111)
0.00	-0.001 (0.992)	0.089 (0.111)	0.00	0.001 (0.988)	0.085 (0.111)	0.00	0.000 (0.994)	0.083 (0.110)
0.00	-0.002 (0.990)	0.092 (0.111)	0.00	-0.005 (0.984)	0.093 (0.111)	0.00	-0.002 (0.986)	0.088 (0.111)
0.00	0.008 (0.986)	0.087 (0.110)	0.00	0.006 (0.988)	0.085 (0.110)	0.00	0.001 (0.994)	0.087 (0.111)
0.00	0.001 (0.996)	0.087 (0.111)	0.00	-0.002 (0.996)	0.083 (0.111)	0.00	0.002 (0.986)	0.086 (0.110)
0.00	0.001 (0.984)	0.090 (0.111)	0.00	0.006 (0.988)	0.089 (0.111)	0.00	-0.003 (0.988)	0.084 (0.110)
0.00	-0.001 (0.996)	0.085 (0.111)	0.00	0.000 (0.992)	0.091 (0.111)	0.00	0.005 (0.990)	0.083 (0.109)
0.00	-0.006 (0.978)	0.093 (0.111)	0.00	0.001 (0.988)	0.086 (0.110)	0.00	-0.003 (0.992)	0.089 (0.112)
0.00	-0.003 (0.988)	0.092 (0.111)	0.00	0.001 (0.984)	0.094 (0.112)	0.00	-0.001 (0.988)	0.086 (0.111)
0.00	-0.001 (0.988)	0.095 (0.112)	0.00	-0.007 (0.992)	0.092 (0.112)	0.00	-0.003 (0.992)	0.089 (0.112)
0.00	-0.003 (0.990)	0.087 (0.111)	0.00	-0.001 (0.992)	0.084 (0.111)	0.00	-0.004 (0.980)	0.092 (0.111)
0.00	-0.004 (0.990)	0.087 (0.111)	0.00	-0.004 (0.990)	0.088 (0.111)	0.00	-0.001 (0.994)	0.088 (0.111)
0.00	-0.004 (0.994)	0.090 (0.111)	0.00	-0.003 (0.988)	0.090 (0.111)	0.00	-0.000 (0.988)	0.090 (0.111)
0.00	0.009 (0.984)	0.088 (0.111)	0.00	-0.003 (0.988)	0.090 (0.111)	0.00	-0.001 (0.982)	0.090 (0.111)
0.00	0.006 (0.990)	0.087 (0.111)	0.00	0.003 (0.988)	0.089 (0.111)	0.00	0.004 (0.984)	0.086 (0.111)
0.00	-0.000 (0.992)	0.083 (0.111)	0.00	-0.001 (0.986)	0.083 (0.109)	0.00	0.006 (0.984)	0.087 (0.111)
0.00	-0.004 (0.984)	0.094 (0.112)	0.00	-0.002 (0.992)	0.085 (0.111)	0.00	0.004 (0.990)	0.089 (0.111)
0.00	-0.000 (0.988)	0.088 (0.111)	0.00	0.005 (0.994)	0.085 (0.111)	0.00	-0.004 (0.980)	0.091 (0.111)
0.00	0.003 (0.994)	0.083 (0.110)	0.00	0.001 (0.982)	0.094 (0.111)	0.00	-0.003 (0.988)	0.086 (0.110)
0.00	0.005 (0.996)	0.085 (0.111)	0.00	0.003 (0.976)	0.092 (0.111)	0.00	0.004 (0.984)	0.090 (0.111)
0.00	-0.001 (0.988)	0.093 (0.112)	0.00	0.004 (0.982)	0.089 (0.111)	0.00	-0.001 (0.990)	0.082 (0.110)
0.00	0.002 (0.990)	0.088 (0.111)	0.00	0.003 (0.986)	0.095 (0.112)	0.00	0.002 (0.982)	0.089 (0.111)

Table 10: Simulation results for regression coefficients, $\boldsymbol{\beta}$.

True	**Bias (CP95)**	**SSD (ESE)**	**True**	**Bias (CP95)**	**SSD (ESE)**
0.00	-0.001 (0.992)	0.080 (0.110)	0.00	0.002 (0.990)	0.087 (0.111)
0.00	0.001 (0.992)	0.090 (0.111)	0.00	0.003 (0.988)	0.080 (0.110)
0.00	-0.004 (0.996)	0.079 (0.110)	0.00	-0.001 (0.994)	0.086 (0.111)
0.00	-0.004 (0.990)	0.086 (0.111)	0.00	-0.007 (0.990)	0.093 (0.112)
0.00	0.001 (0.988)	0.087 (0.111)	0.00	0.003 (0.992)	0.089 (0.111)
0.00	-0.001 (0.988)	0.091 (0.112)	0.00	0.002 (0.996)	0.082 (0.110)
0.00	0.006 (0.992)	0.087 (0.111)	0.00	0.003 (0.986)	0.089 (0.111)
0.00	0.002 (0.986)	0.095 (0.112)	0.00	0.000 (0.990)	0.093 (0.112)
0.00	-0.000 (0.990)	0.093 (0.112)	0.00	0.001 (0.990)	0.091 (0.111)
0.00	0.004 (0.994)	0.084 (0.111)	0.00	-0.001 (0.992)	0.090 (0.111)
0.00	0.005 (0.992)	0.082 (0.110)	0.00	-0.000 (0.988)	0.094 (0.112)
0.00	-0.003 (0.994)	0.084 (0.111)	0.00	0.000 (0.994)	0.083 (0.111)
0.00	-0.005 (0.984)	0.093 (0.112)	0.00	0.007 (0.988)	0.088 (0.111)
0.00	0.004 (0.986)	0.087 (0.111)	0.00	0.006 (0.984)	0.091 (0.111)
0.00	-0.007 (0.982)	0.095 (0.112)	0.00	-0.005 (0.992)	0.088 (0.111)
0.00	-0.001 (0.990)	0.087 (0.111)	0.00	-0.002 (0.996)	0.083 (0.111)
0.00	-0.007 (0.986)	0.087 (0.111)	0.00	-0.002 (0.990)	0.091 (0.112)
0.00	-0.002 (0.984)	0.090 (0.111)	0.00	-0.004 (0.978)	0.091 (0.111)
0.00	-0.001 (0.982)	0.093 (0.112)	0.00	-0.001 (0.992)	0.088 (0.111)
0.00	0.003 (0.994)	0.081 (0.110)	0.00	-0.000 (0.988)	0.087 (0.111)
0.00	0.008 (0.980)	0.099 (0.112)	0.00	-0.005 (0.990)	0.087 (0.111)
0.00	-0.000 (0.994)	0.083 (0.110)	0.00	-0.002 (0.988)	0.084 (0.110)
0.00	0.005 (0.980)	0.092 (0.111)	0.00	0.000 (0.992)	0.089 (0.111)
0.00	-0.001 (0.996)	0.083 (0.110)	0.00	-0.005 (0.992)	0.082 (0.110)
0.00	0.001 (0.982)	0.090 (0.111)	0.00	0.000 (0.992)	0.084 (0.111)
0.00	0.005 (0.984)	0.092 (0.111)	0.00	0.002 (0.984)	0.090 (0.111)
0.00	0.004 (0.982)	0.095 (0.112)	0.00	-0.001 (0.992)	0.089 (0.111)
0.00	-0.006 (0.990)	0.083 (0.110)	0.00	0.003 (0.992)	0.086 (0.111)
0.00	-0.001 (0.992)	0.084 (0.111)	0.00	0.009 (0.986)	0.091 (0.112)
0.00	0.000 (0.990)	0.092 (0.112)	0.00	-0.005 (0.986)	0.085 (0.110)
0.00	-0.002 (0.990)	0.090 (0.111)	0.00	0.006 (0.982)	0.097 (0.112)
0.00	0.002 (0.990)	0.092 (0.112)	0.00	0.004 (0.982)	0.089 (0.111)
0.00	0.002 (0.992)	0.086 (0.111)	0.00	0.000 (0.994)	0.084 (0.110)
0.00	0.004 (0.990)	0.087 (0.111)	0.00	0.003 (0.986)	0.091 (0.111)
0.00	-0.006 (0.992)	0.090 (0.111)	0.00	-0.002 (0.988)	0.083 (0.110)
0.00	-0.005 (0.990)	0.085 (0.110)	0.00	0.006 (0.994)	0.084 (0.111)
0.00	-0.004 (0.990)	0.088 (0.111)	0.00	0.004 (0.992)	0.088 (0.111)
0.00	-0.006 (0.988)	0.088 (0.111)	0.00	-0.006 (0.994)	0.084 (0.110)
0.00	0.001 (0.984)	0.098 (0.112)	0.00	-0.007 (0.988)	0.088 (0.111)
0.00	-0.005 (0.990)	0.086 (0.111)	0.00	0.001 (0.988)	0.086 (0.111)

Table 11: Simulation results for regression coefficients $\boldsymbol{\alpha}$.

True	**Bias (CP95)**	**SSD (ESE)**
1.05	0.014 (0.952)	0.303 (0.301)
-1.00	-0.000 (0.956)	0.238 (0.243)
0.00	0.034 (0.962)	0.273 (0.293)
1.00	0.009 (0.946)	0.307 (0.309)
1.00	-0.041 (0.942)	0.291 (0.299)
0.00	0.038 (0.952)	0.230 (0.235)
0.00	0.004 (0.962)	0.100 (0.103)
0.00	0.005 (0.960)	0.096 (0.103)
0.00	0.007 (0.958)	0.101 (0.103)
0.00	-0.005 (0.956)	0.099 (0.103)
0.00	-0.005 (0.958)	0.102 (0.103)

B.3 Tables of Covariates from Section 3

Here is an outline of the material in this section of the appendix:

Table 12: Covariates included in the **ocScore** model.

Frequency Domain Metrics				
xfftf1	yfftf1	zfftf1	Nxfftf1	Nyfftf1
xfftf2	yfftf2	zfftf2	Nxfftf2	Nyfftf2
xfftf3	yfftf3	zfftf3	Nxfftf3	Nyfftf3
xfftf4	yfftf4	zfftf4	Nxfftf4	Nyfftf4
xfftf5	yfftf5	zfftf5	Nxfftf5	Nyfftf5
xfftf6	yfftf6	zfftf6	Nxfftf6	Nyfftf6
xfftf7	yfftf7	zfftf7	Nxfftf7	Nyfftf7
xfftf8	yfftf8	zfftf8	Nxfftf8	Nyfftf8
xfftf9	yfftf9	zfftf9	Nxfftf9	Nyfftf9
xfftf10	yfftf10	zfftf10	Nxfftf10	Nyfftf10
xfftpow1	yfftpow1	zfftpow1	Nxfftpow1	Nyfftpow1
xfftpow2	yfftpow2	zfftpow2	Nxfftpow2	Nyfftpow2
xfftpow3	yfftpow3	zfftpow3	Nxfftpow3	Nyfftpow3
xfftpow4	yfftpow4	zfftpow4	Nxfftpow4	Nyfftpow4
xfftpow5	yfftpow5	zfftpow5	Nxfftpow5	Nyfftpow5
xfftpow6	yfftpow6	zfftpow6	Nxfftpow6	Nyfftpow6
xfftpow7	yfftpow7	zfftpow7	Nxfftpow7	Nyfftpow7
xfftpow8	yfftpow8	zfftpow8	Nxfftpow8	Nyfftpow8
xfftpow9	yfftpow9	zfftpow9	Nxfftpow9	Nyfftpow9
xfftpow10	yfftpow10	zfftpow10	Nxfftpow10	Nyfftpow10
xdctf1	ydctf1	zdctf1	Nxdctf1	Nydctf1
xdctf2	ydctf2	zdctf2	Nxdctf2	Nydctf2
xdctf3	ydctf3	zdctf3	Nxdctf3	Nydctf3
xdctf4	ydctf4	zdctf4	Nxdctf4	Nydctf4
xdctf5	ydctf5	zdctf5	Nxdctf5	Nydctf5
xdctf6	ydctf6	zdctf6	Nxdctf6	Nydctf6
xdctf7	ydctf7	zdctf7	Nxdctf7	Nydctf7
xdctf8	ydctf8	zdctf8	Nxdctf8	Nydctf8
xdctf9	ydctf9	zdctf9	Nxdctf9	Nydctf9
xdctf10	ydctf10	zdctf10	Nxdctf10	Nydctf10
xdctpow1	ydctpow1	zdctpow1	Nxdctpow1	Nydctpow1
xdctpow2	ydctpow2	zdctpow2	Nxdctpow2	Nydctpow2
xdctpow3	ydctpow3	zdctpow3	Nxdctpow3	Nydctpow3
xdctpow4	ydctpow4	zdctpow4	Nxdctpow4	Nydctpow4
xdctpow5	ydctpow5	zdctpow5	Nxdctpow5	Nydctpow5
xdctpow6	ydctpow6	zdctpow6	Nxdctpow6	Nydctpow6
xdctpow7	ydctpow7	zdctpow7	Nxdctpow7	Nydctpow7
xdctpow8	ydctpow8	zdctpow8	Nxdctpow8	Nydctpow8
xdctpow9	ydctpow9	zdctpow9	Nxdctpow9	Nydctpow9
xdctpow10	ydctpow10	zdctpow10	Nxdctpow10	Nydctpow10

Table 13: Covariates included in the **ocScore** model.

Frequency Domain Metrics Cont'd				
Nzfftf1	vfftf1	Alphafftf1	dAlphafftf1	total_forcefftf1
Nzfftf2	vfftf2	Alphafftf2	dAlphafftf2	total_forcefftf2
Nzfftf3	vfftf3	Alphafftf3	dAlphafftf3	total_forcefftf3
Nzfftf4	vfftf4	Alphafftf4	dAlphafftf4	total_forcefftf4
Nzfftf5	vfftf5	Alphafftf5	dAlphafftf5	total_forcefftf5
Nzfftf6	vfftf6	Alphafftf6	dAlphafftf6	total_forcefftf6
Nzfftf7	vfftf7	Alphafftf7	dAlphafftf7	total_forcefftf7
Nzfftf8	vfftf8	Alphafftf8	dAlphafftf8	total_forcefftf8
Nzfftf9	vfftf9	Alphafftf9	dAlphafftf9	total_forcefftf9
Nzfftf10	vfftf10	Alphafftf10	dAlphafftf10	total_forcefftf10
Nzfftpow1	vfftpow1	Alphafftpow1	dAlphafftpow1	total_forcefftpow1
Nzfftpow2	vfftpow2	Alphafftpow2	dAlphafftpow2	total_forcefftpow2
Nzfftpow3	vfftpow3	Alphafftpow3	dAlphafftpow3	total_forcefftpow3
Nzfftpow4	vfftpow4	Alphafftpow4	dAlphafftpow4	total_forcefftpow4
Nzfftpow5	vfftpow5	Alphafftpow5	dAlphafftpow5	total_forcefftpow5
Nzfftpow6	vfftpow6	Alphafftpow6	dAlphafftpow6	total_forcefftpow6
Nzfftpow7	vfftpow7	Alphafftpow7	dAlphafftpow7	total_forcefftpow7
Nzfftpow8	vfftpow8	Alphafftpow8	dAlphafftpow8	total_forcefftpow8
Nzfftpow9	vfftpow9	Alphafftpow9	dAlphafftpow9	total_forcefftpow9
Nzfftpow10	vfftpow10	Alphafftpow10	dAlphafftpow10	total_forcefftpow10
Nzdctf1	vdctf1	Alphadctf1	dAlphadctf1	total_forcedctf1
Nzdctf2	vdctf2	Alphadctf2	dAlphadctf2	total_forcedctf2
Nzdctf3	vdctf3	Alphadctf3	dAlphadctf3	total_forcedctf3
Nzdctf4	vdctf4	Alphadctf4	dAlphadctf4	total_forcedctf4
Nzdctf5	vdctf5	Alphadctf5	dAlphadctf5	total_forcedctf5
Nzdctf6	vdctf6	Alphadctf6	dAlphadctf6	total_forcedctf6
Nzdctf7	vdctf7	Alphadctf7	dAlphadctf7	total_forcedctf7
Nzdctf8	vdctf8	Alphadctf8	dAlphadctf8	total_forcedctf8
Nzdctf9	vdctf9	Alphadctf9	dAlphadctf9	total_forcedctf9
Nzdctf10	vdctf10	Alphadctf10	dAlphadctf10	total_forcedctf10
Nzdctpow1	vdctpow1	Alphadctpow1	dAlphadctpow1	total_forcedctpow1
Nzdctpow2	vdctpow2	Alphadctpow2	dAlphadctpow2	total_forcedctpow2
Nzdctpow3	vdctpow3	Alphadctpow3	dAlphadctpow3	total_forcedctpow3
Nzdctpow4	vdctpow4	Alphadctpow4	dAlphadctpow4	total_forcedctpow4
Nzdctpow5	vdctpow5	Alphadctpow5	dAlphadctpow5	total_forcedctpow5
Nzdctpow6	vdctpow6	Alphadctpow6	dAlphadctpow6	total_forcedctpow6
Nzdctpow7	vdctpow7	Alphadctpow7	dAlphadctpow7	total_forcedctpow7
Nzdctpow8	vdctpow8	Alphadctpow8	dAlphadctpow8	total_forcedctpow8
Nzdctpow9	vdctpow9	Alphadctpow9	dAlphadctpow9	total_forcedctpow9
Nzdctpow10	vdctpow10	Alphadctpow10	dAlphadctpow10	total_forcedctpow10

Table 14: Covariates included in the **ocScore** model.

Economy of Motion		***Palpation Force-Based***
avgV	sparcV	palp_time
sdV	ldljV	touchpoints
aadV	bf_avgV	tot_dwell_time
arsV	bf_sdV	palp_force_range
ardV	bf_aadV	norm_tot_palp_force
avgF	bf_arsV	***Membership of Insertion Curve***
sdF	bf_ardV	clt0
aadF	bf_avgF	clt1
arsF	bf_sdF	clt2
ardF	bf_aadF	***Force Roughness***
t_underS	bf_arsF	Frgh
PL_underS	bf_ardF	***Needle Location***
force_underS	bf_sparcV	a0
avgAngle_underS	bf_ldljV	a2
Needle Angle		beta02
avg_alpha_S	avg_alphaDot_S	beta_0
LDLJ_S		beta_2

Table 15: Covariates included in the **stable flashback** model.

Palpation Force-Based	***Economy of Motion***
palp_time	sparcV
touchpoints	ldljV
tot_dwell_time	***Needle Angle***
palp_force_range	avg_alpha_S
norm_tot_palp_force	avg_alphaDot_S
Needle Location	LDLJ_S
a0	
a2	
beta02	
beta_0	
beta_2	

Appendix C Supplementary Material for Chapter 4

C.1 BART backfitting algorithm details

In this section, we describe the details of the Bayesian backfitting MCMC algorithm, outlined in Chipman et al. [2010], to sample from the posterior distribution of the regression trees $(T_1, M_1), (T_2, M_2), \ldots, (T_K, M_K)$. In general, the algorithm is simply a Gibbs sampler and can be thought of as a tailored version of Bayesian backfitting MCMC [Hastie and Tibshirani, 2000]. Our algorithm repeatedly resamples the parameters of each learner in the ensemble. To estimate the posterior distribution, we first obtain a draw of the latent random variables ω_i that were introduced in the second stage of our data augmentation procedure:

$$\omega_i \sim \begin{cases} TN\left[\eta_i, 1, (0, \infty)\right], & \text{if } \widetilde{Y}_i = 1 \\ TN\left[\eta_i, 1, (-\infty, 0)\right], & \text{if } \widetilde{Y}_i = 0, \end{cases}$$

where $TN\left[\mu, \sigma^2, (a, b)\right]$ denotes a truncated normal distribution with mean μ and variance σ^2, and support over the interval (a, b). Then, we can treat ω_i as a continuous outcome and recast our BART model as

$$\omega_i = \eta(\boldsymbol{x}_i) + \epsilon_i, \tag{1}$$

where $\epsilon \overset{\text{iid}}{\sim} N(0, 1)$ because we've employed a probit link.

For notational convenience, let $\boldsymbol{T_{-k}}$ be the set of all tree structures excluding T_k, and define $\boldsymbol{M_{-k}}$ in a similar manner, such that $\boldsymbol{T_{-k}}$ is a set of K-1 tree structures and $\boldsymbol{M_{-k}}$ are the associated terminal node parameters. An iteration of the backfitting algorithm entails K successive draws of (T_k, M_k) conditioning on $(\boldsymbol{T_{-k}}, \boldsymbol{M_{-k}})$:

$$(T_k, M_k) \mid \boldsymbol{T_{-k}}, \boldsymbol{M_{-k}}, \boldsymbol{\omega}, \tag{2}$$

for $k = 1, \ldots, K$. To obtain a draw from (2), note that the conditional distribution of $T_k, M_k \mid \boldsymbol{T_{-k}}, \boldsymbol{M_{-k}}, \boldsymbol{\omega}$ depends on $(\boldsymbol{T_{-k}}, \boldsymbol{M_{-k}}, \boldsymbol{\omega})$ through the kth vector set of partial residuals $\boldsymbol{R_k} =$

$(R_{1k}, \ldots, R_{Nk})'$, where the ith element of $\boldsymbol{R_k}$ is given by

$$R_{ik} = \omega_i - \sum_{u \neq k}^{K} g(\boldsymbol{x}_i; T_u, M_u),$$

for $i = 1, \ldots, N$. Since the MCMC update for (T_k, M_k) conditions on all other remaining trees and associated terminal node parameters, the model can be temporarily reparameterized in terms of these partial residuals. Under model (1),

$$R_{ik} \sim N\left(g(\boldsymbol{x}_i; T_k, M_k), 1\right).$$

So, to update (T_k, M_k), we can adopt any of the single-tree MCMC updates, treating the partial residuals as the data. Thus, (2) is equivalent to the posterior draw from a single regression tree $R_{ik} = g(\boldsymbol{x}_i; T_k, M_k) + \epsilon_i$, or

$$(T_k, M_k) \mid \boldsymbol{R_k}, \tag{3}$$

for $k = 1, \ldots, K$. We can obtain a draw from (3) in two successive steps. Since a conjugate normal prior on μ_{gk} was employed, for $g = 1, \ldots, b_k$, we can first integrate out M_k to obtain $T_k \mid \boldsymbol{R_k}$. Then, we can obtain a draw from $M_k \mid T_k, \boldsymbol{R_k}$.

We draw $T_k \mid \boldsymbol{R_k}$ using the Metropolis-Hastings (MH) algorithm of Chipman et al. [1998] where we first generate a candidate tree T_k^* with probability distribution $q(T_k, T_k^*)$ and then we accept T_k^* with probability

$$\alpha(T_k, T_k^*) = \min\left\{1, \frac{q(T_k^*, T_k)}{q(T_k, T_k^*)} \frac{p(\boldsymbol{R_k} \mid T_k^*, M_k)}{p(\boldsymbol{R_k} \mid T_k, M_k)} \frac{\pi(T_k^*)}{\pi(T_k)}\right\}, \tag{4}$$

where $\frac{q(T_k^*, T_k)}{q(T_k, T_k^*)}$ is the transition ratio, $\frac{p(\boldsymbol{R_k}|T_k^*, M_k)}{p(\boldsymbol{R_k}|T_k, M_k)}$ is the likelihood ratio, and $\frac{\pi(T_k^*)}{\pi(T_k)}$ is the tree structure ratio. A new tree T_k^* can be proposed given the current tree T_k using one of four moves: growing a terminal node; pruning a pair of terminal nodes; swapping the splitting criteria of two non-terminal nodes; and changing the splitting criteria of a non-terminal node. We derive equation (4) for the grow, prune, and change steps in Section C.2. For further details, see Chipman et al. [1998, 2010].

Once we have the draw of $T_k \mid \boldsymbol{R_k}$, the draw of $M_k \mid T_k, \boldsymbol{R_k}$ is a set of independent draws of

the terminal node parameters μ_{gk} from a normal distribution. Refer to Section C.3 for the complete expression and its derivation.

C.2 Metropolis-Hastings acceptance probabilities

Here we present the explicit formula for each ratio in the acceptance probability (4) under the grow, prune, and change proposal. This section is modified from Appendix A of Kapelner and Bleich [2013] and Appendix C of Tan and Roy [2019]. The parameter we are sampling is the kth tree structure, T_k, and the data is the residual responses (from the other K-1 trees) that remain unfitted, $\boldsymbol{R_{-k}}$. For notational convenience, let $T := T_k$, $M := M_k$, $\boldsymbol{R} := \boldsymbol{R_{-k}}$, and $\mu_g := \mu_{gk}$.

C.2.1 Grow proposal

Transition ratio

$q(T^*, T)$ is the probability of moving from T to T^*; i.e., selecting a terminal node from T to split into two new child nodes. Hence,

$$\begin{aligned} P(T^* \mid T) =& P(grow)\, P(\text{selecting terminal node to grow from}) \times \\ & P(\text{selecting covariate to split from}) \times \\ & P(\text{selecting value to split on}) \\ =& P(grow)\, \frac{1}{b} \frac{1}{p} \frac{1}{\eta}, \end{aligned}$$

where b is the number of available terminal nodes to split on in T, p is the number of variables left available to split on, and η is the number of unique values left in the chosen variable after adjusting for the parent nodes' splits. The default value for the 'grow' proposal probability $P(grow)$ is 0.25.

On the other hand, $q(T, T^*)$ involves the probability of selecting the correct internal node to prune on such that T^* becomes T, which indicates a pruning move. This is given as

$$\begin{aligned} P(T \mid T^*) &= P(prune)\, P(\text{selecting the correct internal node to prune}) \\ &= P(prune) \frac{1}{w_2^*}, \end{aligned}$$

where w_2^* denotes the number of internal nodes which only have two children terminal nodes. The default value for the 'prune' proposal probability $P(prune)$ is also 0.25.

With this, the transition ratio is given as

$$\frac{q(T^*, T)}{q(T, T^*)} = \frac{P(T^* \mid T)}{P(T \mid T^*)} = \frac{P(prune)}{P(grow)} \frac{bp\eta}{w_2^*}.$$

If there are no variables with two or more unique values, this transition ratio will be set equal to 0, and the grow proposal will be automatically rejected.

Likelihood ratio

To calculate the likelihood, note that the tree structure, T, determines which responses, R_i, fall into which of the b terminal nodes. Hence,

$$P\left(\boldsymbol{R} \mid T, M\right) = \prod_{\ell=1}^{b} P\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}}\right) = \prod_{\ell=1}^{b} P\left(\boldsymbol{R_\ell}\right),$$

where $\boldsymbol{R_\ell} = (R_{\ell_1}, \ldots, R_{\ell_{n_\ell}})'$ are the data in the ℓth terminal node, and n_ℓ denotes how many observations are in the ℓth terminal node, such that $N = \sum_{\ell=1}^{b} n_\ell$.

Recall that $R_{\ell_j} \mid \mu_\ell \overset{\text{iid}}{\sim} N(\mu_\ell, 1)$, for $j = 1, \ldots, n_\ell$, and a normal prior with mean zero is specified for μ_ℓ; i.e., $\mu_\ell \overset{\text{iid}}{\sim} N(0, \sigma_\mu^2)$. Thus, is can be shown [Kapelner and Bleich, 2013] that

$$\begin{aligned} P\left(\boldsymbol{R_\ell}\right) &= \int_{\mathbb{R}} P\left(\boldsymbol{R_\ell} \mid \mu_\ell\right) P(\mu_\ell) d\mu_\ell \\ &= \frac{1}{(2\pi)^{n_\ell/2}} \sqrt{\frac{1}{1 + n_\ell \sigma_\mu^2}} \times \\ &\quad \exp\left\{ -\frac{1}{2} \left(\sum_{j=1}^{n_\ell} (R_{\ell_j} - \overline{R}_\ell)^2 - \frac{\overline{R}_\ell^2 n_\ell^2}{n_\ell + \frac{1}{\sigma_\mu^2}} + n_\ell \overline{R}_\ell^2 \right) \right\}. \end{aligned} \tag{5}$$

Let ℓ be the terminal node of T selected to be grown by the proposal tree. Then, the proposal tree structure T^* is the same as T except for the terminal node ℓ where two children are grown, which we denote by ℓ_L and ℓ_R. Note that the likelihoods are determined by the terminal nodes. Using equation (5), the likelihood ratio becomes

$$\begin{aligned} \frac{P\left(\boldsymbol{R} \mid T^*, M\right)}{P\left(\boldsymbol{R} \mid T, M\right)} &= \frac{P(\boldsymbol{R_{\ell_L}}) P(\boldsymbol{R_{\ell_R}})}{P(\boldsymbol{R_\ell})} \\ &= \sqrt{\frac{(1 + n_\ell \sigma_\mu^2)}{(1 + n_{\ell_L} \sigma_\mu^2)(1 + n_{\ell_R} \sigma_\mu^2)}} \end{aligned}$$

$$\times \exp\left\{\frac{\sigma_\mu^2}{2}\left[\frac{(\sum_{j=1}^{n_{\ell_L}} R_{\ell_{Lj}})^2}{1+n_{\ell_L}\sigma_\mu^2}+\frac{(\sum_{j=1}^{n_{\ell_R}} R_{\ell_{Rj}})^2}{1+n_{\ell_R}\sigma_\mu^2}-\frac{(\sum_{j=1}^{n_\ell} R_{\ell_j})^2}{1+n_\ell\sigma_\mu^2}\right]\right\},$$

where n_{ℓ_L} and n_{ℓ_R} denote the number of data points in the newly grown left and right child nodes, respectively.

Tree structure ratio

Recall from Section 4.2.1 that the tree structure T is made up of the following aspects: its depth, and its decision rules. Let $P_{SPLIT}(\theta)$ denote the probability that a selected node θ will split, and let $P_{RULE}(\theta)$ denote the probability that a variable and and value is selected as the splitting rule. Then, for the entire tree,

$$P(T) = \prod_{\theta\in H_{terminals}}(1-P_{SPLIT}(\theta)) \prod_{\theta\in H_{internals}} P_{SPLIT}(\theta) \prod_{\theta\in H_{internals}} P_{RULE}(\theta),$$

where $H_{terminals}$ and $H_{internals}$ denote the sets of terminal and interior nodes, respectively.

Again, recall from Section 4.2.1 that $P_{SPLIT}(\theta) = \alpha/(1+d_\theta)^\beta$ where d_θ is the depth of node θ, and α and β are the hyperparameters that control the probability. Using the notation from the transition ratio under the grow proposal, $P_{RULE}(\theta) = (1/p)\times(1/\eta)$.

Let θ denote the node on the original tree that was selected to be grown. Then, the proposal tree structure T^* only differs from the current tree T with two child nodes denoted by θ_L and θ_R.

With this, we can now form the tree structure ratio:

$$\begin{aligned}\frac{P(T^*)}{P(T)} &= \frac{\prod_{\theta\in H^*_{terminals}}(1-P_{SPLIT}(\theta))\prod_{\theta\in H^*_{internals}} P_{SPLIT}(\theta)\prod_{\theta\in H^*_{internals}} P_{RULE}(\theta)}{\prod_{\theta\in H_{terminals}}(1-P_{SPLIT}(\theta))\prod_{\theta\in H_{internals}} P_{SPLIT}(\theta)\prod_{\theta\in H_{internals}} P_{RULE}(\theta)} \\ &= \frac{(1-\frac{\alpha}{(1+d_{\theta_L})^\beta})(1-\frac{\alpha}{(1+d_{\theta_R})^\beta})\frac{\alpha}{(1+d_\theta)^\beta}\frac{1}{p}\frac{1}{\eta}}{\frac{\alpha}{(1+d_\theta)^\beta}} \\ &= \alpha\frac{(1-\frac{\alpha}{(2+d_\theta)^\beta})^2}{[(1+d_\theta)^\beta-\alpha]p\eta},\end{aligned}$$

where the last line uses the fact that the depth of the two grown child nodes, d_{θ_L} and d_{θ_R}, are simply the depth of the parent node, d_θ, incremented by one ($d_{\theta_L} = d_{\theta_R} = d_\theta + 1$).

C.2.2 Prune proposal

A prune proposal is the opposite of a grow proposal; it selects an internal node whose children are both terminal and removes both of its children. Thus, ratios will be approximately the inverse of the ratios found in Section C.2.1 for the grow proposal.

Transition ratio

$$\begin{aligned} P(T^* \mid T) &= P\left(prune\right) P\left(\text{selecting the correct internal node to prune}\right) \\ &= P\left(prune\right) \frac{1}{w_2}, \end{aligned}$$

where w_2 is the number of internal parent nodes with two terminal children. On the other hand,

$$P(T \mid T^*) = P\left(grow\right) \frac{1}{(b-1)} \frac{1}{p^*} \frac{1}{\eta^*},$$

which is similar to $P(T^* \mid T)$ for the growth proposal in Section C.2.1 except for the fact that the number of available terminal nodes to split on in T^* is one less than the original tree due to the pruning. Here, p^* is the number of variables left available to split on, and η^* is the number of unique values left in the chosen variable after adjusting for the parent nodes' splits.

Thus, the transition ratio is:

$$\frac{q(T^*, T)}{q(T, T^*)} = \frac{P(T^* \mid T)}{P(T \mid T^*)} = \frac{P\left(grow\right)}{P\left(prune\right)} \frac{w_2}{(b-1)p^*\eta^*}.$$

Likelihood ratio

The likelihood ratio is simply the inverse of the likelihood ratio for the grow proposal:

$$\begin{aligned} \frac{P\left(\boldsymbol{R} \mid T^*, M\right)}{P\left(\boldsymbol{R} \mid T, M\right)} =& \sqrt{\frac{(1+n_{\ell_L}\sigma_\mu^2)(1+n_{\ell_R}\sigma_\mu^2)}{1+n_\ell\sigma_\mu^2}} \\ & \times \exp\left\{\frac{\sigma_\mu^2}{2}\left(\frac{(\sum_{j=1}^{n_\ell} R_{\ell_j})^2}{1+n_\ell\sigma_\mu^2} - \frac{(\sum_{j=1}^{n_{\ell_L}} R_{\ell_{L_j}})^2}{1+n_{\ell_L}\sigma_\mu^2} - \frac{(\sum_{j=1}^{n_{\ell_R}} R_{\ell_{R_j}})^2}{1+n_{\ell_R}\sigma_\mu^2}\right)\right\}. \end{aligned}$$

Tree structure ratio

This is also the inverse of the tree structure ratio for the grow proposal:

$$\frac{P(T^*)}{P(T)} = \frac{[(1+d_\theta)^\beta - \alpha]p^*\eta^*}{\alpha(1-\frac{\alpha}{(2+d_\theta)})^2}.$$

C.2.3 Change proposal

A change proposal involves selecting an internal node and changing its decision rule by selected a new available variable to split on and a new valid splitting value among available values of the selected variable. For simplicity, we will limit the implementation of the change proposal to an internal node that was two terminal child nodes.

Transition ratio

The transition to a proposal tree, $q(T^*, T)$ under a change proposal is given as

$$\begin{aligned} q(T^*, T) = P(T^* \mid T) = & P\left(change\right) P\left(\text{selecting node to change}\right) \times \\ & P\left(\text{selecting new variable to split on}\right) \times \\ & P\left(\text{selecting new value to split on}\right). \end{aligned}$$

When calculating the transition ratio, the first three terms are shared in both the numerator and denominator. The last term will differ as different splitting variables have different numbers of unique values available. Thus, the transition ratio is given as

$$\frac{q(T^*, T)}{q(T, T^*)} = \frac{P(T^* \mid T)}{P(T \mid T^*)} = \frac{\eta^*}{\eta},$$

where η^* and η are the number of splitting values available under the proposal tree's and the original tree's splitting rules, respectively.

Likelihood ratio

The proposal tree structure T^* differs from the original tree structure T only in the two child nodes of the selected change node. These two terminal nodes have the data apportioned differently. Define $\boldsymbol{R_1} = (R_{1,1}, \ldots, R_{1,n_1})'$ as the residual response data in the first child node in the original tree

and $\boldsymbol{R_2} = (R_{2,1}, \ldots, R_{2,n_2})'$ as the residual response data in the second child node in the original tree. Define $\boldsymbol{R_1^*} = (R_{1^*,1}, \ldots, R_{1^*,n_1^*})'$ and $\boldsymbol{R_2^*} = (R_{2^*,1}, \ldots, R_{2^*,n_2^*})'$ similarly for the proposal tree. Thus,

$$\frac{P\left(\boldsymbol{R} \mid T^*, M\right)}{P\left(\boldsymbol{R} \mid T, M\right)} = \frac{P(\boldsymbol{R_1^*})P(\boldsymbol{R_2^*})}{P(\boldsymbol{R_1})P(\boldsymbol{R_2})}.$$

Using equation (5), the following expression is obtained for the likelihood ratio:

$$\frac{P\left(\boldsymbol{R} \mid T^*, M\right)}{P\left(\boldsymbol{R} \mid T, M\right)} = \sqrt{\frac{\left(\frac{1}{\sigma_\mu^2} + n_1\right)\left(\frac{1}{\sigma_\mu^2} + n_2\right)}{\left(\frac{1}{\sigma_\mu^2} + n_1^*\right)\left(\frac{1}{\sigma_\mu^2} + n_2^*\right)}} \times \exp\left\{\frac{1}{2}\left(\frac{(\sum_{j=1}^{n_1^*} R_{1_j^*})^2}{n_1^* + \frac{1}{\sigma_\mu^2}} + \frac{(\sum_{j=1}^{n_2^*} R_{2_j^*})^2}{n_2^* + \frac{1}{\sigma_\mu^2}} - \frac{(\sum_{j=1}^{n_1} R_{1_j})^2}{n_1 + \frac{1}{\sigma_\mu^2}} - \frac{(\sum_{j=1}^{n_2} R_{2_j})^2}{n_2 + \frac{1}{\sigma_\mu^2}}\right)\right\}.$$

Tree structure ratio

The proposal tree has the same structure as the original tree. Thus, we only need to take into account the changed node's children:

$$\frac{P(T^*)}{P(T)} = \frac{(1 - P_{SPLIT}(\theta_1^*))(1 - P_{SPLIT}(\theta_2^*))P_{SPLIT}(\theta^*)P_{RULE}(\theta^*)}{(1 - P_{SPLIT}(\theta_1))(1 - P_{SPLIT}(\theta_2))P_{SPLIT}(\theta)P_{RULE}(\theta)}.$$

The probability of splitting remains the same because the child nodes are at the same depths. Thus, we only need to consider the ratio of the probability of the rules. As previously stated, the probability of selecting the new splitting value will differ as different splitting variables have different numbers of unique values available. Hence, we are left with

$$\frac{P(T^*)}{P(T)} = \frac{\eta}{\eta^*},$$

which is the inverse of the transition ratio. Therefore, for the change proposal, only the likelihood ratio needs to be computed to determine the acceptance probability $\alpha(T, T^*)$.

C.3 Posterior distribution for μ_{gk}

Let $\boldsymbol{R_{gk}} = (R_{gk1}, \ldots, R_{gkn_g})'$ be a subset of $\boldsymbol{R_{-k}}$, where n_g is the number of R_{gkh}'s allocated to the terminal node with parameter μ_{gk} and h indexes the subjects allocated to the terminal node with parameter μ_{gk}. Note that $R_{gkh} \mid T_k, M_k \sim N(\mu_{gk}, 1)$ and $\mu_{gk} \mid T_k \sim N(0, \sigma_\mu^2)$. Then, the posterior distribution of μ_{gk} is given by

$$\begin{aligned}
P(\mu_{gk} \mid T_k, \boldsymbol{R_{-k}} &\propto P\left(\boldsymbol{R_{gk}} \mid T_k, \mu_{gk}\right) P(\mu_{gk} \mid T_k) \\
&\propto \exp\left\{-\frac{\sum_h (R_{gkh} - \mu_{gk})^2}{2}\right\} \exp\left\{-\frac{\mu_{gk}^2}{2\sigma_\mu^2}\right\} \\
&\propto \exp\left\{-\frac{(n_g\sigma_\mu^2 + 1)\mu_{gk}^2 - 2(\sigma_\mu^2 \sum_h R_{gkh})\mu_{gk}}{2\sigma_\mu^2}\right\} \\
&\propto \exp\left\{-\frac{\left(\mu_{gk} - \frac{\sigma_\mu^2 \sum_h R_{gkh}}{n_g\sigma_\mu^2+1}\right)^2}{2\frac{\sigma_\mu^2}{n_g\sigma_\mu^2+1}}\right\}.
\end{aligned}$$

C.4 Posterior sampling algorithm

1. Initialize $(T_1^{(0)}, M_1^{(0)}), \ldots, (T_K^{(0)}, M_K^{(0)})$ and $\widetilde{Y}_i^{(0)}$ for $i = 1, \ldots, N$. If estimating assay accuracy probabilities, then also initialize $\boldsymbol{S_e^{(0)}}$ and $\boldsymbol{S_p^{(0)}}$. Set $s = 1$.

   If not estimating assay accuracy probabilities, set $\boldsymbol{S_e^{(s)}} = \boldsymbol{S_e}$ and $\boldsymbol{S_p^{(s)}} = \boldsymbol{S_p}$ for all s.

2. For $i = 1, \ldots, N$, sample

$$\omega_i^{(s)} \sim \begin{cases} TN\left[\eta_i, 1, (0, \infty)\right], & \text{if } \widetilde{Y}_i^{(s-1)} = 1 \\ TN\left[\eta_i, 1, (-\infty, 0)\right], & \text{if } \widetilde{Y}_i^{(s-1)} = 0, \end{cases}$$

   where $\eta_i = \eta(\boldsymbol{x}_i) = \sum_{k=1}^K g(\boldsymbol{x}_i; T_k^{(s-1)}, M_k^{(s-1)})$. Aggregate $\boldsymbol{\omega}^{(s)} = (\omega_1^{(s)}, \ldots, \omega_N^{(s)})'$.

3. For $k = 1, \ldots, K$, sample $(T_k^{(s)}, M_k^{(s)})$ from $\pi\left((T_k, M_k) \mid (\boldsymbol{T_{-k}^{(s)}}, \boldsymbol{M_{-k}^{(s)}}), \boldsymbol{\omega}^{(s)}\right)$, where

$$(\boldsymbol{T_{-k}^{(s)}}, \boldsymbol{M_{-k}^{(s)}}) = \left((T_1^{(s)}, M_1^{(s)}), \ldots, (T_{k-1}^{(s)}, M_{k-1}^{(s)}), (T_{k+1}^{(s-1)}, M_{k+1}^{(s-1)}), \ldots, (T_K^{(s-1)}, M_K^{(s-1)})\right)',$$

   to obtain $\eta_i = \sum_{k=1}^K g(\boldsymbol{x}_i; T_k^{(s)}, M_k^{(s)})$ for $i = 1, \ldots, N$.

4. If estimating assay accuracy probabilities, then for $l = 1, \ldots, L$, sample $S_{e(l)}^{(s)} \sim Beta(a_{e(l)}^*, b_{e(l)}^*)$

and $S^{(s)}_{p(l)} \sim Beta(a^*_{p(l)}, b^*_{p(l)})$, where $a^*_{e(l)}$, $b^*_{e(l)}$, $a^*_{p(l)}$, and $b^*_{p(l)}$ are evaluated at $\widetilde{\boldsymbol{Y}}^{(s-1)}$. Aggregate $\boldsymbol{S}^{(s)}_{\boldsymbol{e}} = \left(S^{(s)}_{e(1)}, \ldots, S^{(s)}_{e(L)}\right)'$ and $\boldsymbol{S}^{(s)}_{\boldsymbol{p}} = \left(S^{(s)}_{p(1)}, \ldots, S^{(s)}_{p(L)}\right)'$.

5. For $i = 1, \ldots, N$, sample

$$\widetilde{Y}^{(s)}_i \sim Bernoulli\left(\frac{p^*_{i1}}{p^*_{i0} + p^*_{i1}}\right),$$

where p^*_{i0} and p^*_{i1} are evaluated at $\widetilde{\boldsymbol{Y}}^{(s)}_{-\boldsymbol{i}} = \left(\widetilde{Y}^{(s)}_1, \ldots, \widetilde{Y}^{(s)}_{i-1}, \widetilde{Y}^{(s-1)}_{i+1}, \ldots, \widetilde{Y}^{(s-1)}_N\right)'$, $\boldsymbol{S}^{(s)}_{\boldsymbol{e}}$, $\boldsymbol{S}^{(s)}_{\boldsymbol{p}}$, and $\eta_i = \sum_{k=1}^{K} g(\boldsymbol{x}_i; T^{(s)}_k, M^{(s)}_k)$.

6. Increment $s = s + 1$ and return to Step 2.

C.5 Additional simulation results

Here is an outline of the material in this section of the appendix:

Page 96: Table 16. Simulation results for DT for models M1 and M2 when assay accuracy probabilities are unknown.

Page 96: Table 16. Estimation results for unknown assay accuracy probabilities under DT for models M1 and M2 when assay accuracy probabilities are unknown.

Page 96: Table 17. ROC analysis results for models M1 and M2 under DT when assay accuracy probabilities are unknown.

Page 97: Figure 2. Estimation results for unknown function $f(\cdot)$ for models M1 and M2 under DT when assay accuracy probabilities are unknown.

Page 97: Figure 3. Estimation results for average variable use per splitting rule for models M1 and M2 under DT when assay accuracy probabilities are unknown.

Table 16: Simulation results for DT for models M1 and M2 when assay accuracy probabilities are **unknown**. Average bias of 500 posterior mean estimates (Bias), sample standard deviation of 500 posterior mean estimates (SSD), average of 500 estimated of the posterior standard deviation (ESE), and empirical coverage probability (CP95) of nominal 95% equal-tail credible intervals are reported. Note that close agreement between SSD and ESE is preferred.

Model		$S_{e(1)} = 0.95$	$S_{p(1)} = 0.98$	$S_{e(2)} = 0.98$	$S_{p(2)} = 0.99$
M1/BART(K=20)	Bias (CP95)	-0.021 (0.956)	-0.000 (0.992)	-0.005 (0.988)	-0.002 (0.976)
	SSD (ESE)	0.031 (0.037)	0.008 (0.010)	0.009 (0.013)	0.005 (0.006)
M1/BART(K=200)	Bias (CP95)	-0.017 (0.970)	-0.000 (0.996)	-0.004 (0.992)	0.001 (1.000)
	SSD (ESE)	0.028 (0.036)	0.007 (0.010)	0.009 (0.013)	0.004 (0.006)
M1/GLM	Bias (CP95)	-0.026 (0.996)	0.000 (1.000)	-0.007 (1.000)	-0.003 (1.000)
	SSD (ESE)	0.025 (0.047)	0.004 (0.012)	0.006 (0.016)	0.004 (0.008)
M2/BART(K=20)	Bias (CP95)	-0.005 (0.928)	-0.001 (0.946)	-0.002 (0.928)	-0.001 (0.948)
	SSD (ESE)	0.014 (0.014)	0.006 (0.006)	0.007 (0.007)	0.003 (0.003)
M2/BART(K=200)	Bias (CP95)	-0.011 (0.864)	-0.001 (0.964)	-0.005 (0.916)	-0.001 (0.942)
	SSD (ESE)	0.015 (0.014)	0.006 (0.006)	0.007 (0.008)	0.003 (0.003)
M2/GLM	Bias (CP95)	-0.003 (0.950)	-0.002 (0.968)	-0.002 (0.954)	-0.001 (0.952)
	SSD (ESE)	0.013 (0.013)	0.006 (0.006)	0.007 (0.007)	0.003 (0.003)

Table 17: Average estimated AUC (and sample standard deviation in parentheses) for the three model fits (BART with K=20 trees, BART with K=200 trees, and GLM) when the assay accuracy probabilities are **unknown**.

Model		BART(K=20)	BART(K=200)	GLM
M1	In-Sample	0.800 (0.008)	0.818 (0.007)	0.544 (0.010)
	Out-of-Sample	0.773 (0.017)	0.779 (0.018)	0.527 (0.023)
M2	In-Sample	0.986 (0.001)	0.988 (0.001)	0.985 (0.001)
	Out-of-Sample	0.977 (0.003)	0.977 (0.003)	0.979 (0.003)

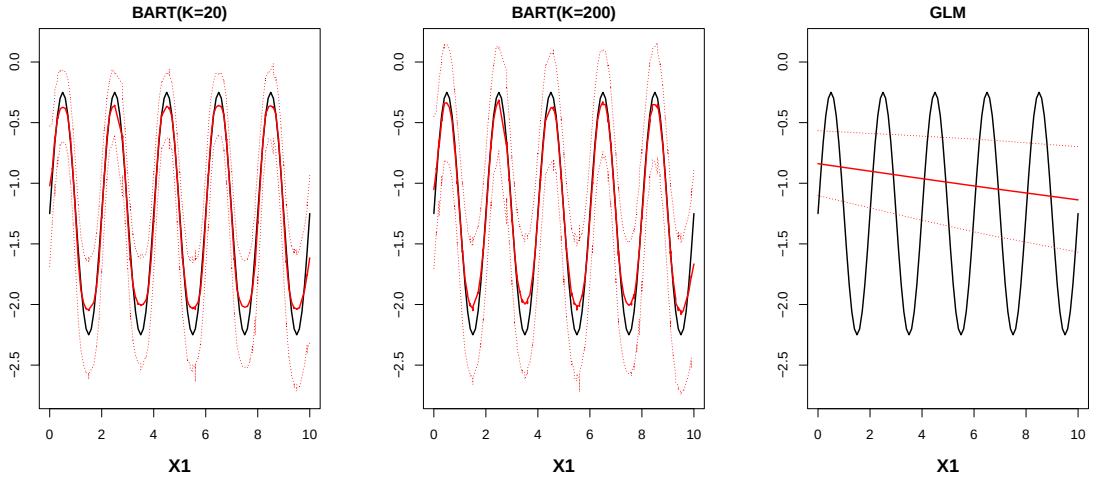

Figure 2: In-sample estimation results for DT when assay accuracy probabilities are **unknown** for the three model fits BART K=20 (left), BART K=200 (middle), and GLM (right). The black solid curve in each subfigure is the true function $f(\cdot)$ in model M1. In each subfigure the following are displayed as red curves: the average of 500 posterior mean estimates (solid curves) and the .025 & .975 posterior mean quantiles (dashed curves).

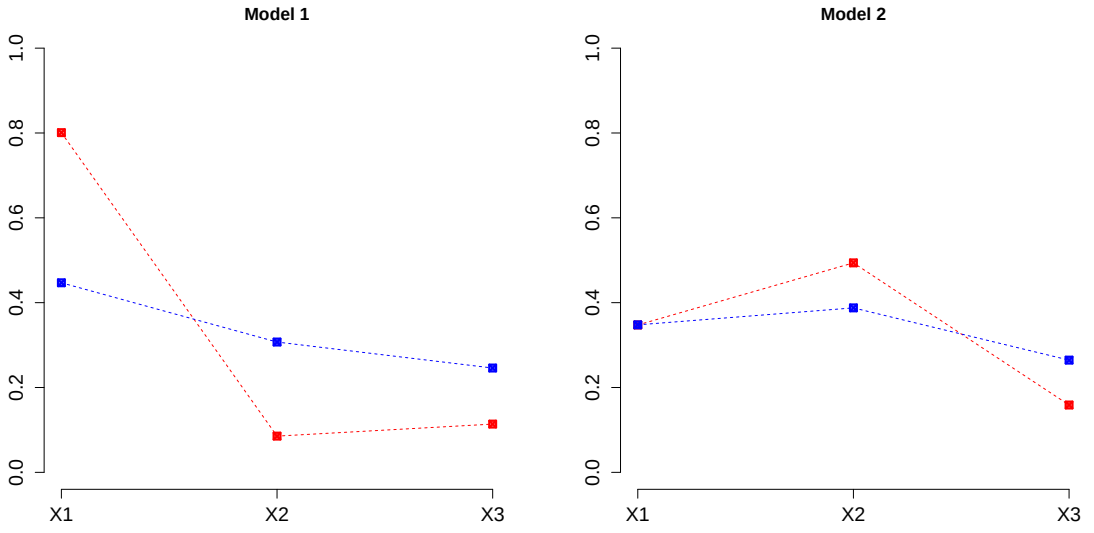

Figure 3: Simulation results for DT for model M1 (left) and model M2 (right) when assay accuracy probabilities are **unknown**. For each covariate, the average use proportion (averaged over the 500 simulations) is plotted for the two BART fits with K=20 trees (red) and K=200 trees (blue).

C.6 Aptima Combo 2 Assay (AC2A) accuracy

Here we summarize the AC2A accuracy based on data from a pilot study and describe how to incorporate this information into our model used for the data application in Section 4.5, when assay accuracy is unknown and to be estimated. This section is modified from Web Appendix D of McMahan et al. [2017].

Table 18: AC2A pilot data.

Stratum	TP	FN	TN	FP	Sensitivity	Specificity
Female/Swab	195	12	1154	28	$S_{e(1)} = 0.942$	$S_{p(1)} = 0.976$
Female/Urine	197	11	1170	13	$S_{e(2)} = 0.947$	$S_{p(2)} = 0.989$
Male/Swab	260	11	774	20	$S_{e(3)} = 0.959$	$S_{p(3)} = 0.975$
Male/Urine	276	6	801	12	$S_{e(4)} = 0.979$	$S_{p(4)} = 0.985$

The notation used in Table 18 is defined below.

$$\begin{aligned} \text{TP} &= \text{ number of true positive individual test results} \\ \text{FN} &= \text{ number of false negative individual test results} \\ \text{TN} &= \text{ number of true negative individual test results} \\ \text{FP} &= \text{ number of false positive individual test results} \end{aligned}$$

Recall from Section 4.2.1 that we place independent Beta priors on the assay accuracies, chosen for computational convenience. We create informative priors by choosing Beta prior hyper-parameter values that incorporate our prior belief about the assay sensitivity and specificity based on the pilot data:

$$\begin{aligned} S_e &\sim Beta(\text{TP} + 1, \text{FN} + 1) \\ S_p &\sim Beta(\text{TN} + 1, \text{FP} + 1). \end{aligned}$$

With this, the prior distributions for S_e and S_p are concentrated around TP/(TP + FN) and TN/(TN + FP), respectively. In particular, for swab specimen we specify $S_e \sim Beta(196, 123)$ and $S_p \sim Beta(1156, 29)$; for urine specimen, we specify $S_e \sim Beta(198, 12)$ and $S_p \sim Beta(1171, 13)$.

Bibliography

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Armagan, A., Dunson, D., and Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica*, 23:119–143.

Audigier, V., Husson, F., and Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10:5–26.

Bergen, A. W., Baurley, J. W., Ervin, C. M., McMahan, C. S., Bible, J., Stafford, R. S., Mudumbai, S. C., and Saxon, A. J. (2022). Effects of buprenorphine dose and therapeutic engagement on illicit opiate use in opioid use disorder treatment trials. *Int. J. Environ. Res. Public Health*, 19(7):4106.

Bish, D. R., Bish, E. K., El-Hajj, H., and Aprahamian, H. (2021). A robust pooled testing approach to expand COVID-19 screening capacity. *PLoS One*, 16:e0246285.

Brouwer, D. J. (2011). Cannulation camp: basic needle cannulation training for dialysis staff. *Dialysis and Transplantation*, 40:434–439.

Center for Disease Control and Prevention (2021). Drug overdose deaths in the u.s. top 100,000 annually. *https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2021/20211117.htm*, Accessed: 2022-04-19.

Chen, P., Tebbs, J., and Bilder, C. (2009). Group testing regression models with fixed and random effects. *Biometrics*, 65:1270–1278.

Chipman, H., George, E., and McCulloch, R. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4:266–298.

Chipman, H. and McCulloch, R. (2016). *BayesTree: Bayesian Additive Regression Trees*. R package version 0.3-1.4.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93:935–948.

Collaboration, G. C. K. D. (2020). Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 395:709–733.

Crist, R. C., Vickers-Smith, R., Kember, R. L., Rentsch, C. T., Xu, H., Edelman, E. J., Hartwell, E. E., Kampman, K. M., and Kranzler, H. R. (2021). Analysis of genetic and clinical factors associated with buprenorphine response. *Drug Alcohol Depend.*, 227:109013.

Delaigle, A. and Hall, P. (2015). Nonparametric methods for group testing data, taking dilution into account. *Biometrika*, 102:871–887.

Delaigle, A., Hall, P., and Wishart, J. (2014). New approaches to non- and semi-parametric regression for univariate and multivariate group testing data. *Biometrika*, 101:567–585.

Delaigle, A. and Meister, A. (2011). Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*, 106:640–650.

Dhand, N., Johnson, W., and Toribio, J. (2010). A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics*, 15:452–473.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440.

Dreifuss, J. A., Griffin, M. L., Frost, K., Fitzmaurice, G. M., Potter, J. S., Fiellin, D. A., Selzer, J., Hatch-Maillette, M., Sonne, S. C., and Weiss, R. D. (2013). Patient characteristics associated with buprenorphine/naloxone treatment outcome for prescription opioid dependence: Results from a multisite study. *Drug Alcohol Depend.*, 131(1-2):112–118.

Fiellin, D. A., Pantalon, M. V., Chawarski, M. C., Moore, B. A., Sullivan, L. E., O'Connor, P. G., and Schottenfeld, R. S. (2006). Counseling plus buprenorphine-naloxone maintenance therapy for opioid dependence. *N. Engl. J. Med.*, 355(4):365–374.

Gastwirth, J. and Johnson, W. (1994). Screening with cost effective quality control: Potential applications to hiv and drug testing. *Journal of the American Statistical Association*, 89:972–981.

Gaydos, C., Quinn, T., Willis, D., Weissfeld, A., Hook, E., Martin, D., Ferrero, D., and Schachter, J. (2003). Performance of the APTIMA combo 2 assay for detection of Chlamydia trachomatis and Neisseria gonorrhoeae in female urine and endocervical swab specimens. *Journal of Clinical Microbiology*, 41:304–309.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition.

Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science*, 15:196–213.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elemnts of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Heffernan, A., Aylward, L., Toms, L., Sly, P., Macleod, M., and Mueller, J. (2014). Pooled biological specimens for human biomonitoring of environmental chemicals: Opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology*, 24:225–232.

Hser, Y.-I., Evans, E., Grella, C., Ling, W., and Anglin, D. (2015). Long-term course of opioid addiction. *Harv. Rev. Psychiatry*, 23(2):76–89.

Huang, X. and Tebbs, J. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics*, 65:710–718.

Hughes-Oliver, J. M. (2006). Pooling experiments for blood screening and drug discovery. In Dean, A. and Lewis, S., editors, *Screening*, pages 48–68. Springer, New York.

Johnson, W. and Gastwirth, J. (2000). Dual group screening. *Journal of Statistical Planning and Inferences*, 83:449–473.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, 2 edition.

Kampman, K. and Jarvis, M. (2015). American society of addiction medicine (ASAM) national practice guideline for the use of medications in the treatment of addiction involving opioid use. *J. Addict. Med.*, 9(5):358–367.

Kapelner, A. and Bleich, J. (2013). bartMachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*.

Kapelner, A. and Bleich, J. (2022). *bartMachine: Bayesian Additive Regression Trees*. R package version 1.3.2.

Kleinman, S. H., Strong, D. M., Tegtmeier, G. G. E., Holland, P. V., Gorlin, J. B., Cousins, C., Chiacchierini, R. P., and Pietrelli, L. A. (2005). Hepatitis B virus (HBV) DNA screening of blood donations in minipools with the COBAS AmpliScreen HBV test. *Transfusion*, 45:1247–1257.

Krajden, M., Cook, D., Mak, A., Chu, K., Chahil, N., Steinberg, M., Rekart, M., and Gilbert, M. (2014). Pooled nucleic acid testing increases the diagnositc yield of acute hiv infections in high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. *Journal of Clinical Virology*, 61:132–137.

Kuehn, B. M. (2021). Massive costs of the us opioid epidemic in lives and dollars. *JAMA*, 325:2040.

Lapham, G., Boudreau, D. M., Johnson, E. A., Bobb, J. F., Matthews, A. G., McCormack, J., Liu, D., Samet, J. H., Saxon, A. J., Campbell, C. I., Glass, J. E., Rossom, R. C., Murphy, M. T., Binswanger, I. A., Yarborough, B. J. H., Bradley, K. A., and PROUD Collaborative Investigators (2020). Prevalence and treatment of opioid use disorders among primary care patients in six health systems. *Drug Alcohol Depend.*, 207:107732.

Lee, T., Barker, J., and Allon, M. (2006). Needle infiltration of arteriovenous fistulae in hemodialysis: risk factors and consequences. *American Journal of Kidney Diseases*, 47:1020–1026.

Lewis, J., Lockary, V., and Kobic, S. (2012). Cost savings increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae. *Sexually Transmitted Diseases*, 39:46–48.

Ling, W., Jacobs, P., Hillhouse, M., Hasson, A., Thomas, C., Freese, T., Sparenborg, S., McCarty, D., Weiss, R., Saxon, A., Cohen, A., Straus, M., Brigham, G., Liu, D., McLaughlin, P., and Tai, B. (2010). From research to the real world: buprenorphine in the decade of the clinical trials network. *J. Subst. Abuse Treat.*, 38 Suppl 1:S53–60.

Liu, Y., McMahan, C., Tebbs, J., Gallagher, C., and Bilder, C. (2021). Generalized additive regression for group testing data. *Biostatistics*, 22:873–889.

Liu, Z., Petersen, L., Zhang, Z., and Singapogu, R. (2020). A method for segmenting the process of needle insertion during simulated cannulation using sensor data. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6090–6094.

Lok, C. E., Huber, T. S., Lee, T., Shenoy, S., Yevzlin, A., S., Abreo, K., Allon, M., Asif, A., Astor, B. C., Glickman, M. H., Graham, J., Moist, L. M., Rajan, D. K., Roberts, C., Vachharajani, T. J., and Valentini, R. P. (2020). KDOQI clinical practice guideline for vascular access: 2019 update,. *American Journal of Kidney Diseases*, 75:S1–S164.

Ma, J., Bao, Y.-P., Wang, R.-J., Su, M.-F., Liu, M.-X., Li, J.-Q., Degenhardt, L., Farrell, M., Blow, F. C., Ilgen, M., Shi, J., and Lu, L. (2019). Effects of medication-assisted treatment on mortality among opioids users: a systematic review and meta-analysis. *Mol. Psychiatry*, 24(12):1868–1883.

Mattick, R. P., Breen, C., Kimber, J., and Davoli, M. (2014). Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. *Cochrane Database of Systematic Reviews*, (2).

McCulloch, R., Sparapani, R., Spanbauer, C., Gramacy, R., and Pratola, M. (2021). *BART: Bayesian Additive Regression Trees*. R package version 2.9.

McMahan, C., Tebbs, J., Hanson, T., and Bilder, C. (2017). Bayesian regression for group testing data. *Biometrics*, 73:1443–1452.

Moist, L. M., Lee, T. C., Lok, C. E., Al-Jaishi, A., Xi, W., Campbell, V., Graham, J., Wilson, B., and Vachharajani, T. J. (2013). Education in vascular access. *Seminars in Dialysis*, 26:148–153.

Noureldin, Y. A., Fahmy, N., Anidjar, M., and Andonian, S. (2016). Is there a place for virtual reality simulators in assessment of competency in percutaneous renal access? *World Journal of Urology*, 34:733–739.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

Petersen, L., Liu, Z., Bible, J., Shukla, D., and Singapogu, R. (2022). Simulator-based metrics for quantifying vascular palpation skill for cannulation. *IEEE Access*, 10:66862–66873.

Polson, N., Scott, J., and Windle, J. (2013). Bayesian inference for logistic models using pólya-gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, 2 edition.

Saá, P., Proctor, M., Foster, G., Krysztof, D., Winton, C., Linnen, J., Gao, K., Brodsky, J., Limberger, R., Dodd, R., and Stramer, S. (2018). Investigational testing for Zika virus among us blood donors. *New England Journal of Medicine*, 378:1778–1788.

Saran, R., Robinson, B., and Abbott, K. C. (2020). US Renal Data System 2019 Annual Data Report: epidemiology of kidney disease in the United States. *American Journal of Kidney Diseases*, 75:A6–A7.

Sarov, B., Novack, L., Beer, N., Saf, J., Soliman, H., Pliskin, J. S., Lit-vak, E., Yaari, A., and Shinar, E. (2007). Feasibility and cost–benefit of implementing pooled screening for HCVAg in small blood bank settings. *Transfusion Medicine*, 17:479–487.

Seltzer, M. H., Wong, W. H., and Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: issues and methods. *Journal of Educational and Behavioral Statistics*, 21:131–167.

Singapogu, R., Chowdhury, A., Roy-Chaudhury, P., and Brouwer-Maier, D. (2021). Simulator-based hemodialysis cannulation skills training: a new horizon? *Clincal Kidney Journal*, 14:465–470.

Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association*, 95:1353–1358.

Tan, Y. and Roy, J. (2019). Bayesian additive regression trees and the general BART model. *Statistics in Medicine*, 38:5048–5069.

Thurlow, J. S., Joshi, M., Yan, G., Norris, K. C., Agodoa, L. Y., Yuan, C. M., and Nee, R. (2021). Global epidemiology of end-stage kidney disease and disparities in kidney replacement therapy. *American Journal of Nephrology*, 52:98–107.

Torres, I., Albert, E., and Navarro, D. (2020). Pooling of nasopharyngeal swab specimens for SARS-CoV-2 detection by RT-PCR. *Journal of Medical Virology*, 92:2306–2307.

Van, T., Miller, J., Warshauer, D., Reisdorf, E., Jerrigan, D., Humes, R., and Shult, P. (2012). Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology*, 50:891–896.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools and serum samples. *Biometrics*, 56:1126–1133.

Walbaum, M., Scholes, S., Rojas, R., Mindell, J. S., and Pizzo, E. (2021). Projection of the health impacts of chronic kidney disease in the Chilean population. *PLOS ONE*, 16:1–18.

Weiss, R. D., Potter, J. S., Provost, S. E., Huang, Z., Jacobs, P., Hasson, A., Lindblad, R., Connery, H. S., Prather, K., and Ling, W. (2010). A multi-site, two-phase, prescription opioid addiction treatment study (POATS): rationale, design, and methodology. *Contemp. Clin. Trials*, 31(2):189–199.

Westreich, D., Hudgens, M., Fiscus, S., and Pilcher, C. (2008). Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. *Journal of Clinical Microbiology*, 46:1785–1792.

Wong, L. Y., Liew, A. S. T., Weng, W. T., Lim, C. K., Vathsala, A., and Toh, M. P. H. S. (2018). Projecting the burden of chronic kidney disease in a developed country and its implications on public health. *International Journal of Nephrology*, 2018.

Xie, M. (2001). Regression analysis of group testing samples. *Statistics in Medicine*, 20:1957–1969.

Zendejas, B., Brydges, R., Hamstra, S. J., and Cook, D. A. (2013). State of the evidence on simulation-based training for laparoscopic surgery: a systematic review. *Annals of Surgery*, 257:586–593.

Zou, H. and Hastie, T. (2020). *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.3.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286.