2021

Bayesian Nonparametric Models For Causal Inference And Clustering Under Dirichlet Process Priors

Arman Oganisian
University of Pennsylvania

Bayesian Nonparametric Models For Causal Inference And Clustering Under Dirichlet Process Priors

Abstract

This body of work develops new Bayesian nonparametric (BNP) models for estimating causal effects with observational data. Though broadly applicable, it is motivated by statistical complexities that frequently arise in health economics. Using potential outcomes, we formulate tailored causal estimands and determine the conditions under which they are identifiable from observed data. Once identified, flexible estimation follows from constructing models with high-dimensional sets of parameters that are allowed to grow with the sample size. We employ the Dirichlet Process (DP), and related stochastic processes, as priors over these high-dimensional spaces to do posterior causal inference. First, motivated by complexities in medical cost distributions, we construct a generative two-part model for zero-inflated outcomes under a DP prior. This model is able to capture structural zeros, skewness, and multimodality. We develop a Bayesian g-computation procedure for causal estimation and use the induced partitioning of the DP to detect latent clusters of patients with similar data distributions. Second, we extend this work to cost-effectiveness analyses, which requires jointly modeling a bivariate outcome under right-censoring. Posterior causal inference is done using a BNP joint model under the Enriched DP and Gamma Process priors. Finally, we tackle the difficulties of estimating causal effects in multiple sparse subgroups. Using an improper Hierarchical DP, we construct a new ``hierarchical Bayesian bootstrap" prior that partially pools confounder information across subgroups when performing g-computation. This allows for potential efficiency gains without imposing strong parametric assumptions on the confounder distributions. A key contribution throughout is the construction of Markov Chain Monte Carlo (MCMC) algorithms for efficient posterior sampling.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Nandita Mitra

Second Advisor

Jason A. Roy

Keywords

Bayesian nonparametrics, Bayesian Statistics, Causal Inference, Clustering, Dirichlet Process, Nonparametric Modeling

Subject Categories

Biostatistics

BAYESIAN NONPARAMETRIC MODELS FOR CAUSAL INFERENCE AND CLUSTERING UNDER DIRICHLET PROCESS PRIORS

Arman Oganisian

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Jason A. Roy

Professor of Biostatistics

Co-Supervisor of Dissertation

Nandita Mitra

Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Russell T. Shinohara, Associate Professor of Biostatistics

Dylan S. Small, Class of 1965 Wharton Professor of Statistics

Edward I. George, Universal Furniture Professor of Statistics

BAYESIAN NONPARAMETRIC MODELS FOR CAUSAL INFERENCE AND CLUSTERING UNDER DIRICHLET PROCESS PRIORS

ACKNOWLEDGEMENT

I would like to thank my dissertation supervisors, Dr. Jason Roy and Dr. Nandita Mitra, as well as my committee members for their guidance and mentorship throughout the years. I would also like to thank the members of the Center for Causal Inference at the University of Pennsylvania for providing an intellectually stimulating forum that benefited me greatly during my research.

This research was funded in part by the following grants: Grant R01GM112327 and Grant 124268-IRG-78-002-35-IRG from the American Cancer Society, the George and Emily McMichael Harrison Fund, Penn Presbyterian Harrison Fund of the University of Pennsylvania Hospital Obstetrics and Gynecology Department.

Analyses in Chapters 2 and 3 used the linked SEER-Medicare database and we acknowledge the efforts of the Applied Research Program; National Cancer Institute; Office of Research, Development and Information; Centers for Medicare and Medicaid Services; Information Management Services; and SEER program tumor registries in the creation of the SEER-Medicare database. We thank Dr. Emily Ko (Department of Obstetrics and Gynecology, University of Pennsylvania Health Systems) for data guidance. Chapter 4 used data from previous studies by Dr. James Metz and Dr. Justin Bekelman at the Department of Radiation Oncology, Perelman School of Medicine, University of Pennsylvania. We thank them for access to the data.

ABSTRACT

BAYESIAN NONPARAMETRIC MODELS FOR CAUSAL INFERENCE AND CLUSTERING UNDER DIRICHLET PROCESS PRIORS

Arman Oganisian

Jason A. Roy

Nandita Mitra

This body of work develops new Bayesian nonparametric (BNP) models for estimating causal effects with observational data. Though broadly applicable, it is motivated by statistical complexities that frequently arise in health economics. Using potential outcomes, we formulate tailored causal estimands and determine the conditions under which they are identifiable from observed data. Once identified, flexible estimation follows from constructing models with high-dimensional sets of parameters that are allowed to grow with the sample size. We employ the Dirichlet Process (DP), and related stochastic processes, as priors over these high-dimensional spaces to do posterior causal inference. First, motivated by complexities in medical cost distributions, we construct a generative two-part model for zero-inflated outcomes under a DP prior. This model is able to capture structural zeros, skewness, and multimodality. We develop a Bayesian g-computation procedure for causal estimation and use the induced partitioning of the DP to detect latent clusters of patients with similar data distributions. Second, we extend this work to cost-effectiveness analyses, which requires jointly modeling a bivariate outcome under right-censoring. Posterior causal inference is done using a BNP joint model under the Enriched DP and Gamma Process priors. Finally, we tackle the difficulties of estimating causal effects in multiple sparse subgroups. Using an improper Hierarchical DP, we construct a new "hierarchical Bayesian bootstrap" prior that partially pools confounder information across subgroups when performing g-computation. This allows for potential efficiency gains without imposing strong parametric assumptions on the confounder distributions. A key contribution throughout is the construction of Markov Chain Monte Carlo (MCMC) algorithms for efficient posterior sampling.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

This dissertation develops new Bayesian nonparametric (BNP) models for estimating causal effects with observational data. Bayesian approaches for causal inference date back to early work by Rubin, 1978 who framed the task as a type of missing data problem in a finite sample. Specifically, with a binary treatment we only observe one of two potential outcomes for each subject. From a Bayesian perspective, each subject's counterfactual is simply treated as an unknown/missing value which can be drawn from its posterior distribution under certain assumptions. This posterior then induces a posterior over functionals of the potential outcomes, such as sample average treatment effects. This missing-data perspective has been reviewed recently by Ding and Li, 2018.

Over the last few decades, the desire for more robust estimation of population-level causal effects has led to development of BNP methods. BNP models involve a high-dimensional set of parameters that is typically allowed to grow with the sample size - affording them a high degree of flexibility. Special priors over these high-dimensional spaces are used to define a suitably regularized posterior over the model space. Many well-known models such as Bayesian additive regression trees (Chipman, George, and McCulloch, 2010) and Gaussian processes (Rasmussen and Williams, 2005), for instance, belong to this class. An introduction and review of these methods in causal inference is provided by Oganisian and Roy, 2021.

In our view, BNP methods add value over classical nonparametric/machine learning methods in several ways. First, while most nonparametric methods from regression trees to kernel regression involve regularization, it is often done through ad-hoc penalties on the objective functions. In BNP methods, regularization is done in a more principled way via prior measures and uncertainty in the regularization flows through to the posterior automatically. Second, as opposed to classical machine learning methods where uncertainty estimation tends to be more complicated relative to point-estimation, BNP methods can do both quite easily since the output is a full posterior distribution. Summarizing the center and dispersion of the posterior in a number of ways is relatively easy once we have the posterior. Third, in causal inference targets of inference tend to be functionals (often integrals) over relevant models. An appealing property of fully Bayesian inference is that a

posterior over the model induces a posterior over these functionals and required integrals over the model can be performed with relative ease via Monte Carlo.

Here, we will primarily be developing Dirichlet Process (DP) mixtures which, at a high level, are BNP approaches that model outcome distributions as a data-adaptive mixtures of simpler distributions. At a low-level, the DP is a stochastic process that generates random discrete probability distributions. Due to conjugacy and overall tractability, it has become a canonical nonparametric prior over unknown probability distributions. It is a "nonparametric" prior in the sense that the distributions generated by the DP do not belong to any particular parametric family. DP mixtures follow from placing a DP prior over the mixing distribution of a mixture model. Over the years, the DP has been generalized and extended in several ways. For instance, the enriched DP (EDP) and the hierarchical DP (HDP) have unique properties which are useful for addressing important causal modeling challenges. Though broadly applicable to a wide range of applications, our BNP modeling solutions are particularly motivated by challenges in health economics.

In Chapter 2, we develop a generative BNP model equipped to handle outcomes that exhibit structural zeros, multi-modalities, and skewed tails. Such outcomes are quite common in analyses of medical costs in economics as some subjects may never accrue cost while others may accrue very high costs. However, applications include ecology (modeling rainfall: there can be many days/areas with no rain), actuarial sciences (modeling insurance payouts: some plans never paying out during a coverage period), and biomedical studies (modeling blood concentration of some biomarker: concentrations below limit of detection will return zero). Our approach specifies a joint model for the outcome, the propensity score, and the confounders with DP prior over the unknown distribution of the *subject-specific* parameters governing this joint. The DP probabilistically partitions the complex joint into more homogenous subregions/clusters and assigns each its own parameter vector. Importantly, the number of clusters that form is not pre-specified and is bounded only by the sample size. Thus, this model is capable of capturing complexities that models with a single set of finite-dimensional parameters cannot. We show how posterior output from this model can be (1) used in a g-computation algorithm for computing causal effects, (2) conduct posterior predictive checks around important causal assumptions, and (3) infer both soft and hard latent cluster assignments for subjects who share similar joint data distributions. An Markov chain Monte Carlo (MCMC) algorithm for posterior inference via auxiliary variables is outlined and simulations assessing frequentist

properties are performed. We apply the model to analyze medical costs accrued by patients diagnosed with endometrial cancer and subsequently assigned to post-operative chemotherapy and radiation therapy.

In Chapter 3 we extend the work of Chapter 2 to the setting of cost-effectiveness analyses (CEAs). CEAs are employed in health policy and economics to evaluate whether the benefits associated with a particular treatment outweigh the medical costs - with results often used to inform efficient resource allocation and production. In cancer studies, the most common efficacy measure is increased in survival time. This makes CEAs statistically challenging as we must now deal with modeling a bivariate outcome (cost and survival time) jointly under right-censoring. In this chapter we first define causal contrasts in terms of expected potential monetary value accrued under each treatment. We then define a causal net monetary benefit is the difference in the expected potential monetary value of two treatments and identify it under a set of causal assumptions. We then construct a BNP model for cost-efficacy based on an enriched DP (EDP) and Gamma Process priors and use it in a posterior g-computation procedure to obtain draws from the posterior of the causal net monetary benefit estimand. Simulations are performed, posterior MCMC computation is outlined, and the model is applied to a CEA of radiation versus chemotherapy for endometrial cancer.

Finally, in Chapter 4, we tackle the important challenge of heterogenous treatment effect (HTE) estimation. HTEs are causal effects within subpopulations defined by the strata of some variable of interest (e.g. within education levels, income brackets, phenotypes, genotypes, etc.). The primary interest lies in how these causal effects vary (display "heterogeneity") across strata. This often involves integrating stratum-specific outcome regressions over the stratum-specific confounder distributions. While some Bayesian pooling can be done across strata when estimating the regression models, usually the confounder distributions are estimated flexibly - but independently - via a Bayesian bootstrap (Rubin, 1981) or the empirical distribution. This, however, provides poor performance when strata are too sparse to yield such flexible estimates of the confounder distributions. We formate a hierarchical Bayesian bootstrap (HBB) prior using hierarchical DPs. The resulting posterior allows for borrowing of confounder information across strata, with more borrowing in sparse strata and less borrowing for populous strata. A conjugate MCMC update procedure is outlined and simulations are performed to assess frequentist operating characteristics. We show

that our approach limits to the empirical distribution and make connections to the smoothed Bootstrap. We apply the model to estimate the adverse event rates of proton versus photon therapy across cancer type strata.

In Chapter 5, we end with a brief discussion of our work within the broader context of causal inference and and BNP modeling. Namely, we discuss the role Bayesian nonparametrics occupies between classical statistical procedures that emphasize uncertainty estimation and more modern machine learning methods that emphasize flexible point estimation. We discuss a resurgence of BNP models in the computer science literature under the name of "probabilistic machine learning" and give a few reasons for optimism about the role of BNP methods in causal inference.

CHAPTER 2

BAYESIAN NONPARAMETRIC MODEL FOR ZERO-INFLATED OUTCOMES: PREDICTION, CLUSTERING, AND CAUSAL INFERENCE

2.1. Overview and Motivation

Researchers across many fields are often interested in outcome prediction, clustering analysis, and causal inference. For example, researchers in personalized medicine are broadly concerned with forming out-of-sample outcome predictions given a subject's covariates. Health economists are often interested in subgroup identification for resource allocation purposes and may turn to algorithms such as K-means. Policy researchers, on the other hand, focus on causality - estimating the average difference in outcomes that would have occurred under hypothetical policy interventions. All of these tasks become challenging in the presence of zero-inflated outcomes, multi-modality, and extreme skewness. Structural zeros often need to be modeled: if causal treatment effect estimation is the goal, failing to capture a difference in prevalence of zeros between treatment groups may bias effect estimates. For prediction purposes, it is necessary to capture outcomes at the skewed high-end of the distribution as well as predicting the structural zeros at the low-end. Failing to do so would tarnish predictions at both tails. For clustering analyses, having to pre-specify the number of clusters - typically an unknown quantity - poses a significant challenge.

In this paper, we develop a Bayesian nonparametric (BNP) generative model that simultaneously predicts structural zeros as a function of covariates, captures skewness in both the outcome and continuous covariates, and induces a grouping of subjects into clusters with similar joint data distributions. The result is a flexible, multi-purpose model that is broadly applicable to the tasks described above. We demonstrate the ability of our model to produce robust causal effect estimates via standardization - a common method for computing marginal causal contrasts while adjusting for measured confounders. This fully Bayesian approach allows uncertainty to propagate through to the causal estimates, allowing point and interval estimation of various causal contrasts such as mean differences and quantile causal effects. Moreover, posterior predictive checks around positivity - a key causal identification assumption - can be readily conducted using the model output.

In particular, we propose a Dirichlet Process (DP) mixture of zero-inflated regressions. Each zero-inflated regression is a two-part model: a model for the probability of the outcome being zero and a regression for the continuous, non-zero outcomes. DP mixtures (Ferguson, 1973) are a class of BNP models that partition a complex joint distribution of the outcome and covariates into more homogeneous clusters. In our case, the cluster-specific conditional means are modeled using a zero-inflated regression. Unlike finite mixtures, DP mixtures assume there are infinitely many clusters in the population - removing the need to specify the number of clusters in advance. As many clusters are introduced as are needed to accommodate the complexity of the data. If the data are not complex and can be adequately fit with a parametric model, new clusters form less often. In this sense, our model is data adaptive - growing in proportion to the complexity of the data.

The flexibility and relative ease of constructing point and interval estimates for various types of contrasts are perhaps some of the reasons that BNP methods have been growing in popularity within the causal inference literature. For example, Bayesian additive regression trees (BART) (Chipman, George, and McCulloch, 2010; Hill, 2011) have been used to estimate causal treatment effects. Dependent Dirichlet process methods have been developed for estimating marginal structural models (Roy, Lum, and Daniels, 2017) and dynamic treatment regime models (Xu et al., 2016). Dirichlet process mixture approaches for mediation analysis (Kim et al., 2017) and Enriched Dirichlet process (Wade et al., 2014) mixture approaches to standardization have also been developed (Roy et al., 2018). However, these methods do not address the complications of zero-inflation discussed. We advance existing methodology by developing a BNP standardization approach that accounts for zero-inflation.

Several factors distinguish our approach from the existing zero-inflated models outside of the causal inference literature. As opposed to the parametric Bayesian approach of Ghosh, Mukhopadhyay, and Lu, 2006, our method is non-parametric and, therefore, better suited for complex data. Barcella et al., 2016 develop a DP mixture of poisson regressions. Though this provides a flexible fit to count data, it is inappropriate for semi-continuous data. Linero, Sinha, and Lipsitz, 2018 develop a semi-parametric Bayesian model for semi-continuous outcomes. They use a two-part model - a probit model for the probability of a zero and a parametric density for non-zero outcomes. The mean functions of both models are jointly estimated using a BART-based model. In contrast, our model is fully nonparametric, DP-based as opposed to BART-based, and generative as opposed to conditional.

That is, our model estimates the full joint data distribution rather than solely a conditional outcome distribution. The strength of DP-based procedures over BART-based procedures is that the former induces clustering - allowing us to capture multi-modalities. Using generative models as opposed to conditional models provides a framework for flexibly imputing missing data, as was demonstrated by Roy et al., 2018.

Though broadly applicable, we motivate our approach throughout the paper by the analysis of medical cost outcomes - an important use case of our method. Zero-inflation is the norm in cost data as patients may tend to have zero costs through mechanisms that depend on measured covariates and the assigned treatment. Medical costs also tend to be skewed by especially high-cost patients. Moreover, the joint distribution tends to be multi-modal with groups of patients that exhibit different cost-covariate relationships. Legislators and regulators often make use of economic analyses comparing costs associated with proposed policy interventions. These comparisons are causal in nature and require robust statistical modeling while adjusting for confounders.

2.2. Dirichlet Process Mixture of Zero-Inflated Regressions

2.2.1. A Generative Model

Consider observing data $D = (D_i)_{i=1:n} = (Y_i, A_i, L_i)_{i=1:n}$ from n independently sampled subjects. The $q \times 1$ covariate vector, L_i, contains both categorical and continuous covariates measured pre-treatment. The scalar $A_i \in \{0, 1\}$ denotes binary treatment assignment. The scalar outcome is Y_i - whose empirical distribution may exhibit excess zeros, skewness, and multimodality. We first define covariate vectors for subject i as $x_i = (1, A_i, L_i)'$ and $m_i = (1, L_i)'$. We specify a generative model - that is, a model for the full joint $p(D_i \mid \omega_i) = p(Y_i \mid A_i, L_i, \omega_i)p(A_i \mid L_i, \omega_i)p(L_i \mid \omega_i)$. Hierarchically this is given by,

$$
\begin{aligned}
Y_i \mid A_i, L_i, \beta_i, \gamma_i, \phi_i &\sim \pi(x_i'\gamma_i)\delta_0(y_i) + \big(1 - \pi(x_i'\gamma_i)\big) \cdot N(y_i \mid x_i'\beta_i, \phi_i) \\
A_i \mid L_i, \eta_i &\sim Ber\Big(\text{expit}(m_i'\eta_i)\Big) \\
L_i \mid \theta_i &\sim p(l_i \mid \theta_i) \\
\omega_i \mid G &\sim G \\
G \mid \alpha, G_0 &\sim DP(\alpha G_0)
\end{aligned}
\tag{2.1}
$$

Above, we define $\omega_i = (\beta_i, \phi_i, \gamma_i, \eta_i, \theta_i)$ for compactness. The conditional distribution of the outcome, Y_i, is modeled as a two-part mixture of a point-mass at 0, $\delta_0(y_i) = I(y_i = 0)$, and a Gaussian distribution with mean $x_i'\beta_i$ and variance ϕ_i. This allows for a positive probability of the outcome being zero, $P(Y_i = 0) = \pi(x_i'\gamma_i) = expit(x_i'\gamma_i)$. This probability is modeled as a function of treatment and confounders using a logistic regression with a $(q+2) \times 1$ parameter vector γ_i. Separately, the conditional mean of non-zero outcomes is modeled using a regression with a $(q+2) \times 1$ parameter vector β_i. Though we use logit links to model $\pi(x_i'\gamma_i)$ and $P(A_i = 1 \mid L_i, \eta_i)$, other links, such as the probit, could be used as well. In anticipation of subsequent application to causal estimation, we model treatment probability (i.e., the propensity score) as a function of confounders, L_i, using a logistic regression with a $(q+1) \times 1$ parameter vector η_i. Finally, a joint distribution over the confounders, $p(l_i|\theta_i)$, is specified and governed by a vector of parameters θ_i. Since specification is application-specific, we leave this distribution in general terms.

We assume subject-specific parameters are drawn from some distribution G. We place a DP prior with base distribution G_0 and concentration parameter α on G - denoted as $G \sim DP(\alpha G_0)$. The DP is a "distribution over distributions" Ferguson, 1973 and draws from a DP are discrete - implying a positive probability of ties among the subject-specific parameters. In other words, the DP prior induces a clustering of patients who are more homogeneous in terms of the parameters that govern the joint distribution of their data - including the conditional outcome distribution, structural zero distribution, propensity score distribution, and covariate distribution. The DP model allows the number of clusters and, therefore, parameters to grow with the sample size - making this a nonparametric model despite the seemingly rigid assumptions in Equation 2.1 (Hannah, Blei, and Powell, 2011). The model is data-adaptive in the sense that more clusters are introduced to capture data complexities (e.g. non-linear and non-additive covariate effects, multimodalities, skewness, etc). However, if the data are simple enough to be explained by a single cluster/set of parameters, then the model shrinks to a parametric one with linear, additive covariate effects given in Equation 2.1 (with ω_i being equal to some common ω^*, $\forall i$).

This model has several desirable properties when it comes to complex data such as cost outcomes, which we use as the motivating example throughout the paper. The clustering accounts for multi-modality in cost distributions. Structural zeros are modeled as a function of treatment and confounders. For causal inference, this model admits a flexible predictive distribution which,

as we will see, can be incorporated into a standardization procedure. Finally, explicit modeling of treatment assignment allows us to conduct posterior predictive checks assessing the validity of the positivity assumption. We note that in application areas where outcomes are non-negative and observed data are close to zero, a local Gaussian distribution that ignores the non-negative nature of the outcome is undesirable. In these instances, we can proceed with the model as presented after log-transforming non-zero values - essentially assuming a log-Normal distribution for these values. Appendix A.6 provides more details along with a proof-of-concept simulation.

2.2.2. Posterior Sampling and Hyperparameters

Using the Pólya Urn (Blackwell and MacQueen, 1973) representation of the DP, it can be shown that the conditional posterior of ω_i is given by (Muller et al., 2015),

$$p(\omega_i|\omega_{1:(i-1)}, D) \propto \frac{1}{\alpha + i - 1}\Big[\alpha p(D_i|\omega_i)G_0(\omega_i) + \sum_{j<i} p(D_i|\omega_j)\delta_{\omega_j}(\omega_i)\Big] \tag{2.2}$$

where $D_i = (Y_i, A_i, L_i)$ is the data vector for the i^{th} subject. The posterior clustering of patients is evident in Equation 2.2. Subject i's parameter, ω_i, can equal one of the previously drawn parameters, ω_j, with probability proportional to the subject's likelihood evaluation under ω_j, $\sum_{j<i} p(D_i|\omega_j) \cdot \delta_{\omega_j}(\omega_i)$. Or, with probability proportional to $\alpha p(D_i|\omega_i)$, ω_i can be a new, previously unseen parameter drawn from the prior $G_0(\omega_i)$.

If subject i is quite unique so that its likelihood evaluation is low under the $i-1$ existing parameters, then it is relatively more likely for this subject to be assigned its own set of parameters from the prior. Finally, note that as n gets large and i approaches n, the prior probability $\alpha/(\alpha + i - 1)$ of the i^{th} subject being assigned to a new cluster goes to zero. This property helps prevent overfitting.

The conditional posterior in Equation 2.2 forms the basis of a Metropolis-in-Gibbs sampler we use to sample $\omega_{1:n}$ from the full posterior, $p(\omega_{1:n}|D) = p(\omega_1|D)\prod_{i=2}^{n} p(\omega_i|\omega_{1:(i-1)}, D)$. The sampler proceeds in the spirit of Neal's Algorithm 8 (Neal, 2000) by introducing latent cluster membership indicators, $c_{1:n} = (c_1, c_2, \ldots, c_n)'$, for the subjects. We initialize the algorithm by partitioning subjects to one of K initial clusters. Each iteration t, with $K^{(t)}$ occupied clusters indexed by k, has two steps. First, conditional on $c_{1:n}^{(t)}$, we draw from the posterior of each model parameter based on the likelihood contributions of all subjects with $c_i^{(t)} = k$. Conditional on these updated parameter,

$\omega_{1:n}^{(t+1)}$, we update each assignment indicator

$$c_i^{(t+1)}|c_{1:(i-1)}^{(t)} \sim Cat\Big(\frac{1}{\alpha+i-1}p(D_i|\omega_1^{(t+1)}),\ldots,\frac{1}{\alpha+i-1}p(D_i|\omega_{i-1}^{(t+1)}),\frac{\alpha}{\alpha+i-1}p(D_i|\omega_0^{(t+1)})\Big)$$

Above, $Cat(\cdot)$ denotes the categorical distribution and $\omega_0^{(t+1)} \sim G_0$ is a draw from the prior taken at each iteration. Notice that in each iteration subject i has a $\frac{\alpha}{\alpha+i-1}p(D_i|\omega_0^{(t+1)})$ probability of being assigned to a new cluster.

The two hyperparameters of the model in Equation 2.1 are the choice of base distribution, G_0, and the concentration parameters α. A requirement for the base distribution is that it be over the space of the parameters $\omega_i = (\beta_i, \phi_i, \gamma_i, \eta_i, \theta_i)$. Prior independence is often assumed so that G_0 can be constructed as the product over parameter-specific priors. Conjugate priors for each parameter may be used, if possible, to simplify MCMC computation. The concentration parameter α governs how frequently new clusters appear in an MCMC run. It is often described as a prior sample size for a new cluster. Following previous analyses (Roy et al., 2018), we place a $\alpha \sim Gamma(1,1)$ prior on α rather than set it at a particular value. Examples of specifying G_0 is given in Appendices A.4, A.5, and A.6 for simulations, data analysis, and log-transform extension respectively.

2.2.3. Posterior Mode Clustering in the Presence of Label Switching

Often we may like to cluster patients using the posterior mode - allowing us to identify and summarize distinct groups in terms of observed characteristics. In mixture models, posterior mode inference on cluster assignment is complicated by label switching (Rodrguez and Walker, 2014) - the fact that cluster labels $c_{1:n}$ do not have consistent meanings across Gibbs iterations. For example, at iteration t, a new cluster, labeled cluster 2, may be proposed and all subjects previously in, say, cluster 1 may be re-assigned to this new cluster. Even though the cluster label has changed from 1 to 2, the cluster still contains the same subjects. Therefore, naively using the mode of the T cluster indicators, $c_i^{(1)},\ldots,c_i^{(T)}$, as each subject as the mode assignment is problematic. To meaningfully cluster subjects based on posterior mode, we perform a deterministic relabeling of cluster indicators after posterior sampling (Dahl, 2006; Stephens, 2000). We compute for each iteration t an $n \times n$ adjacency matrix with a one in the $(i,j)^{th}$ entry indicating patients i and j were clustered together and zero indicating otherwise. The element-wise mean of this matrix across the t iterations gives us a posterior mode matrix with $(i,j)^{th}$ entry being the posterior probability of patients i

and j being clustered together. To obtain cluster assignments, we select the adjacency matrix that is closest in the L_2 sense to the posterior mode matrix. More details regarding the relabeling is provided in Appendix A.2.

2.3. Counterfactual Prediction and Estimating Causal Contrasts

2.3.1. Review of Counterfactuals and Causal Estimation

We first provide a motivating review of causal estimation before discussing our BNP standardization procedure. Consider observing $D = (Y_i, A_i, L_i)_{i=1:n}$ as defined in Section 2.2.1 from some target population we wish to make inference about. Using potential outcome notation, let the random variable $Y^{A=a}$ represent the potential outcome under treatment $A = a$. The marginal causal effect of treatment on the outcome, $\Psi = E\Big[Y^{A=1} - Y^{A=0}\Big]$, can be computed via the method of standardization under the standard causal identification assumptions (Rubin, 1978) of ignorability, consistency, no interference, and positivity. Briefly, and in order, these assumptions require that all confounders are controlled for, that there is only one form of each treatment, that each patient's outcome is independent of others' treatment assignments, and that treatment assignment is not deterministic for any individual in the population. We provide a formal statement in Appendix A.3.

In the Bayesian framework, standardization is conducted using the posterior predictive distribution of the outcome (Keil et al., 2017). Throughout, we use tildes to denote posterior predictive draws. Let $\tilde{Y}^a$ denote the posterior predictive outcome under intervention $A = a$ with predictive distribution $p(\tilde{Y}^a|D)$. Also, let $\tilde{L}$ denote a posterior predictive draw of confounders. If the causal assumptions hold, standardization under intervention $A = a$ is given by

$$E(\tilde{Y}^a|D) = \int_\theta \int_\beta \int_{\tilde{L}} E(\tilde{Y}|A = a, \tilde{L}, \beta)p(\tilde{L}|\theta)p(\beta, \theta|D)\; d\tilde{L}\; d\beta\; d\theta \tag{2.3}$$

Above, β and θ are parameter vectors that govern the conditional distribution of the outcome and the distribution of the confounders, respectively. This slightly differs from frequentist standardization by additionally averaging a prediction model for the outcome, $E(\tilde{Y} \mid A = a, \tilde{L}, \beta)$, over the posterior distribution of the parameters, $p(\beta, \theta|D)$. Given T draws from the posterior $(\theta^{(t)}, \beta^{(t)})_{t=1:T}$, we can compute Equation 2.3 by first drawing $\tilde{l}^{(t)} \sim p(\tilde{L}|\theta^{(t)})$, then computing $E[\tilde{Y}|A = a, \tilde{L} = \tilde{l}^{(t)}, \beta^{(t)}]$. This yields a posterior distribution for the difference $\{\delta^{(t)}\}_{1:T} = \{E[\tilde{Y}|A =$

$1, \tilde{L} = \tilde{l}^{(t)}, \beta^{(t)}] - E[\tilde{Y}|A = 0, \tilde{L} = \tilde{l}^{(t)}, \beta^{(t)}]\}_{1:T}$. The posterior mean $\hat{\Psi} = T^{-1} \sum_t \delta^{(t)}$ can be taken as a point estimate of Ψ, while percentiles can be used for interval estimation. Standardization in Equation 2.3 crucially requires both a correctly specified regression for the outcome as well as an accurate estimate of the marginal confounder distribution. As correct specification is unlikely, robust estimation requires nonparametric modeling. This is especially the case in medical cost data - where multimodality, zero-inflation, and skewness are unlikely to be captured by simple parametric models.

2.3.2. Sampling from the Posterior Predictive Distribution

The model outlined in Equation 2.1 yields a flexible predictive distribution, which in turn yields robust causal effect estimates. Under standard causal identification assumptions, the posterior predictive distribution of potential outcome $\tilde{Y}^a$ is given by

$$\begin{aligned} p(\tilde{Y}^a|D) = \frac{\alpha}{\alpha + n} \int_{\tilde{\omega}} \int_{\tilde{L}} p(\tilde{Y}|A = a, \tilde{L}, \tilde{\omega}) dP(\tilde{L}|\tilde{\omega}) dG_0(\tilde{\omega}) \\ + \frac{1}{\alpha + n} \int_{\omega_{1:n}} \Big[\int_{\tilde{L}} \sum_{i=1}^{n} p(\tilde{Y}|A = a, \tilde{L}, \omega_i) dP(\tilde{L}|\omega_i) \Big] dP(\omega_{1:n}|D) \end{aligned} \tag{2.4}$$

A derivation is provided in Appendix A.1. Note that the particular forms of $p(\tilde{Y}|A, \tilde{L}, \omega)$ and $p(\tilde{L}|\omega)$ are specified in Equation 2.1. We can draw from this distribution via Monte Carlo. For each of the T posterior draws, $(\omega_{1:n}^{(t)})_{t=1:T}$ from $p(\omega_{1:n}|D)$, draw from the conditional distribution $l \sim p(\tilde{L}|\omega_i^{(t)})$. Then, under intervention $A = a$, we can draw from $\tilde{y}_a^{(t)} \sim p(\tilde{Y}|A = a, \tilde{L} = l, \omega_i^{(t)})$. For the t^{th} draw, the inner integral over $\tilde{\omega}$ in Equation 2.4 can be evaluated numerically by drawing from the prior $\tilde{\omega}_0 \sim G_0$, then drawing confounders conditional on this prior draw $\tilde{l} \sim p(\tilde{l}|\tilde{\omega}_0)$. This procedure yields predictive draws $\{\tilde{y}_a^{(t)}\}_{t=1:T}$.

After obtaining draws $\{\tilde{y}_1^{(t)}\}_{1:T}$ and $\{\tilde{y}_0^{(t)}\}_{1:T}$, we can compute T draws of the difference $\{\delta_t\}_{t=1:T}$, where $\delta_t = \tilde{y}_1^{(t)} - y_0^{(t)}$. Thus, a Bayesian nonparametric point estimate of Ψ is given by $\hat{\Psi}_{BNP} = E[\tilde{Y}^1|D] - E[\tilde{Y}^0|D] \approx \frac{1}{T} \sum_t \delta_t$. Intervals can be constructed using percentiles of $\{\delta_t\}_{1:T}$. Quantile causal effects and counterfactuals (Xu, Daniels, and Winterstein, 2018) may also be computed from Equation 2.4. We estimate the posterior predictive CDF of the potential outcome under intervention a using the posterior predictive draws, $F_a(v) = P(Y^a \leq v|D) \approx \frac{1}{T} \sum_{t=1}^{T} I(\tilde{y}_a^{(t)} \leq v)$. The inverse of this estimated CDF can be used to estimate quantile causal effects. For instance, the median causal

effect can be estimated as the difference in median cost under two interventions $F_1^{-1}(.5) - F_0^{-1}(.5)$. This contrast may be preferable to Ψ for skewed outcomes.

2.3.3. Assessing the Positivity Assumption

Positivity is the only identification assumptions that can be assessed empirically. The assumption requires that the probability of treatment is bounded $0 < P(A = 1|L) < 1, \ \forall L$. Violations of positivity (e.g. $P(A = 1|L) = 1$) imply that there are subgroups of the data for which no comparator patients exist, thus forcing the model to extrapolate when computing causal contrasts. Incorrect extrapolation in these regions will bias causal effect estimates. There are many methods of handling violations once they are identified (Petersen et al., 2012), but these are out of scope for this paper. Here we simply provide a framework for assessing this assumption within the unique context of our zero-infalted DP model. Note in Equation 2.1, we have explicitly modeled treatment probability as a function of covariates. This allows us to predict treatment probability for each patient, given posterior draws $(\omega_{1:n}^t)_{t=1:T}$ via Monte Carlo:

$$P(\tilde{A} = 1|l, D) \approx \frac{1}{T} \sum_{t=1}^{T} \frac{\frac{\alpha}{\alpha+n} \int_{\tilde{\omega}} p(\tilde{A}|l, \tilde{\omega}) p(l|\tilde{\omega}) dG_0(\tilde{\omega}) + \frac{1}{\alpha+n} \sum_{i=1}^{n} p(\tilde{A}|l, \omega_i^{(t)}) p(l|\omega_i^{(t)})}{\frac{\alpha}{\alpha+n} \int_{\tilde{\omega}} p(l|\tilde{\omega}) dG_0(\tilde{\omega}) + \frac{1}{\alpha+n} \sum_{i=1}^{n} p(l|\omega_i^{(t)})} \tag{2.5}$$

A derivation is provided in Appendix A.1. Using the above, we can compute $P(\tilde{A}_i = 1 \mid L_i, D)$ for each subject in our sample. Typically, histograms of these probabilities are plotted for treated and untreated patients separately. Separated distributions indicate a lack of overlap and, therefore, high posterior belief of a positivity violation.

2.4. Simulation Study

In this section, we evaluate bias of $\hat{\Psi}_{BNP}$, coverage of the credible interval estimates, and precision of the estimate as measured by interval width. We compare results to existing methods that may be considered by researchers faced with zero-inflated outcomes - namely BART, a non-Bayesian doubly robust estimator, and two parametric Bayesian Gamma models. BART is a Bayesian non-parametric, tree-based ensemble for the conditional mean of the outcome. The doubly robust estimator is a two-part model for treatment assignment and the outcome. We use a boosted frequentist logistic regression for the treatment model and a frequentist Gaussian model for the outcome. The first parametric model is a Bayesian Gamma hurdle model. This is a two-part model that explicitly

models the probability of the outcome being zero with a logistic regression, while modeling positive outcomes with a Gamma regression. The second parametric model is a naive, yet somewhat common, approach of adding .01 to zero outcome values and modeling this transformed outcome using a Bayesian Gamma regression. We refer to this as the Gamma $+.01$ model.

We simulate from two data generating processes (DGPs). In the clustered DGP, we simulate data from three distinct clusters - each with its own set of parameters that govern confounder distributions, binary treatment assignment, zero-inflation, and Gamma-distributed positive outcomes. The Gamma distribution is used to simulate realistic cost data that are non-negative and skewed within each cluster. Thus, the *local* conditional outcome distribution assumed in Equation 2.1 is deliberately misspecified. In the parametric DGP, we simulate data from a single cluster with a common covariate distribution, treatment assignment model, zero-inflation, and Gamma-distributed positive outcomes. The data are skewed, but not multimodal. Other simulation details regarding hyperparameter settings and sampling are given in Appendix A.4.

Table 2.1: Results across 1000 simulated datasets with 3000 subjects each. Average bias of the posterior mean is reported as a proportion of the true value ($\Psi = -9740.3$ in the clustered setting and $\Psi = -10184.1$ in the parametric setting). Mean credible interval widths are presented for the Zero-inflated DP model, the BART model, and two Gamma models. Confidence intervals are given for the doubly robust method. In the parametric setting, we have 45% in the clustered setting and 55% in the parametric setting. We simulate with one continuous covariate and four binary covariates, all of which affect zero-inflation, treatment probability, and the outcome. All models condition on the simulated confounders so that ignorability holds.

DGP	Model	Bias	Coverage	Interval Width
Clustered	Zero-Inflated DP	-.081	94.3%	21612.2
	BART	-.746	76.2%	26374.2
	Doubly Robust	.795	87.1%	33449.3
	Gamma Hurdle	-.509	79.8%	19692.2
	Gamma +.01	1.817	4.7%	27358.1
Parametric	Zero-Inflated DP	.097	95.1%	22034.1
	BART	-.054	96.1%	23825.3
	Doubly Robust	-.027	95.9%	23339.1
	Gamma Hurdle	-.014	95.1%	21778.7
	Gamma +.01	-.489	100%	50580.3

In the clustered setting, the zero-inflated DP model produces effect estimates with the smallest bias - -8.1% of the true value. The 95% credible interval has close to nominal coverage of 94.2%. Though the Gamma hurdle model cannot handle multimodality, it outperforms both BART and the doubly robust estimators due to its explicit modeling of structural zeros and skewness. The latter two models capture neither zero-inflation nor multimodality and consequently perform poorly.

In the parametric setting, the zero-inflated DP model again exhibits low bias and close to nominal coverage. BART and the doubly robust models have lower bias, but exhibit slight overcoverage (about 96%) in the interval estimates. The Gamma hurdle model is correctly specified in the parametric DGP and so performs the best - exhibiting the lowest bias of 1.4%, 95.1% coverage, and yielding the shortest interval. Relative to this correctly specified hurdle model, the Zero-inflated DP has only a slightly wider interval length on average (22034.1 versus 21778.7), suggesting little efficiency loss. BART and the doubly robust estimators both have wider intervals than the DP on average.

The particularly bad performance of the naive Gamma +.01 model - under both DGPs - should be noted. While it is a simple, seemingly harmless trick, adding a small constant severely degrades the accuracy and precision of treatment effect estimates. Unlike the hurdle model and DP model, it does not model structural zeros, and so ignores the effect of treatment that generates these zeros. The zero-inflated DP mixture captures multi-modality, skewness, and the treatment's effect on structural zeros. This allows for good treatment effect estimates under both simple and pathological data distributions with minimal efficiency loss if the parametric model is correct.

2.5. Application: Inpatient Medical Costs of Endometrial Cancer Treatments

In this section, we use the proposed DP mixture of zero-inflated regressions to analyze inpatient medical costs among patients with endometrial cancer. Patients who were diagnosed with endometrial cancer between 2000 and 2014 were identified in the SEER Medicare database. Those assigned to either radiation or chemotherapy post-hysterectomy were followed for a maximum of two years after initial treatment. The total inpatient costs, measured in 2018 US dollars, accrued over the followup period was recorded and is our primary outcome of interest. Inpatient costs are costs that accrue during overnight hospitalizations and do not include costs such as prescription treatment costs, outpatient costs, or hospice care costs.

Table 2.2: Baseline Characteristics: Means and standard deviations are reported for continuous variables. Counts and percentages are reported for categorical variables. Standardized mean differences (SMD) are provided. All monetary amounts are in thousands of 2018 U.S. Dollars.

	Chemotherapy (N=92)	Radiation (N=952)	SMD
Total Inpatient Costs ($)	22.1 (28.6)	23.4 (34.5)	.039
Zero Costs	14 (15.2%)	75 (7.9%)	
Age (years)	73.68 (6.98)	73.25 (5.98)	.066
Household Income ($)	64.4 (32.4)	56.8 (26.2)	.257
White	76 (82.6%)	835 (87.8%)	.147
Diabetic	20 (21.7%)	197 (20.7%)	.026
CCI			.350
0	49 (53.3%)	529 (55.6%)	
1	22 (23.9%)	260 (27.3%)	
$\geq$2	21 (22.8%)	131 (13.8%)	
Grade = 1	28 (30.4%)	208 (21.8%)	.196
FIGO Stage I-N0 or I-A	63 (68.5%)	357 (37.5%)	.653

Table 2.2 presents baseline characteristics of the two treatment groups. There is a significant proportion of zero costs - 15.2% in the chemotherapy arm versus 7.9% in the radiation arm. Chemotherapy subjects have lower inpatient costs over the follow up period. However, there may be several confounding factors. For example, the primary determinants of post-hysterectomy treatment are the stage and grade of the cancer, with consideration for patient commorbidity and age. These factors, which are measured pre-treatment, likely also affect inpatient costs. The standardized mean difference for stage, grade, and CCI are all $> .1$.

In the following subsections, we demonstrate how our method can be used to model the data from several angles. All results are from posterior sampling of the model in Equation 2.1. We control for race, CCI, household income, cancer grade and stage in both the positive outcome model and the zero-probability model. We model treatment assignment as a function of these confounders as well. We assume local Gaussian distributions for CCI and household income and Bernoulli distributions for binary covariates. Details about hyperparameter settings, priors, and sampling

results are provided in Appendix A.5.

2.5.1. Multi-modality and Clustering Results

The patients in this study are heterogeneous in terms of their observed costs and covariates. Some have extremely high costs, more comorbidities, and come from varying socio-economic backgrounds. Clustering can be useful for both describing these groups in terms of observed characteristics or motivating new research. There is a vast literature on clustering methods and we do not claim the DP-induced method is superior, but it does have several advantages. First, since the DP mixture assumes there are infinitely many clusters in the population (though in a particular analysis, the number of clusters is bounded by n), we need not specify the number of clusters beforehand. Second, this method allows for uncertainty quantification around the posterior mode data partition.

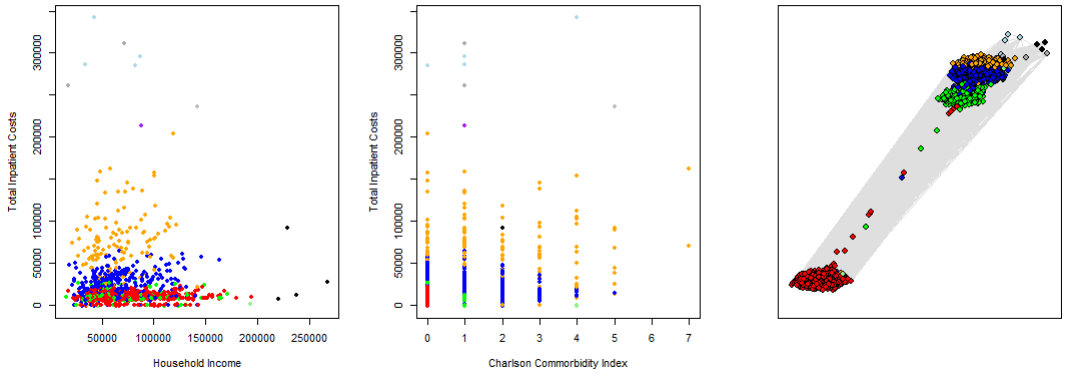


Figure 2.1: Clustering results from the zero-inflated DP mixture. Colors indicate posterior mode cluster assignment. The first panel projects clustering results onto the cost-income space and the second projects the results onto the cost-CCI space - both relevant dimensions for understanding costs. The third panel visualizes the full posterior mode matrix discussed in Section 2.2.3 with a network diagram. Each node represents a patient and the lengths of vertices connecting any two nodes are inversely proportional to the posterior probability of being clustered together. The position of the nodes in x-y space have no meaning (hence the absence of axis labels), only the relative distance between nodes are relevant.

Two potentially important confounders of costs and treatment assignment are household income and CCI. The first two panels of Figure 2.1 visualize cost along these dimensions. While we initialize the model with five clusters, the model identified ten clusters in the posterior - introducing five additional clusters to accommodate the complexity of the data. In the first panel, we see the orange cluster has very high costs, the blue cluster has moderately high costs, while the green and red clusters have lower costs. There are two results worth noting. First, the light blue and gray clusters

represent patients who have such distinctly high costs that the DP model places them in their own cluster. Second, the black points represent patients who, while having similar costs to most patients, have distinctly high household income. Thus the DP model places them in their own cluster. From this we can see that the clustering is happening in multiple dimensions rather than only on the cost space.

Similarly, we cannot see much difference between the green and red clusters on the cost-household income space. However, the second panel shows that these patients occupy distinct places on the cost-CCI space, where the red cluster ranks lower than green on CCI. It may be clear at this point that visualizing clustering in two-dimensions is limited by the need to choose the variables on each dimension. The third panel solves this issue by visualizing the entire posterior mode matrix discussed in Section 2.2.3 as a network diagram. We can use this diagram to get a sense of the uncertainty around the mode cluster assignment/partition. For example, the nodes between the red and green cluster have very uncertain assignment. About half the time, they were clustered with the red patients and the other half they were clustered with the green patients - indicating we should not have very high confidence in their posterior mode assignment. This type of uncertainty characterization is absent in many classical clustering algorithms, like K-means. We can summarize observed characteristics of patients by posterior mode assignment. In the orange cluster, average cost in this cluster is $71,139. The distribution of CCI in this group is skewed much higher, suggesting a possible positive association between cost and CCI. On the other hand, we can see from Figure 1 that the light blue cluster has much higher costs (first panel), yet these subjects are relatively low on the CCI scale (second panel). The relationship between cost and CCI seems unclear - and perhaps this motivates future research targeted at learning this relationship.

2.5.2. Cost Prediction in Presence of Zero-Inflation

Induced clustering is the core strength of DP mixtures: a single parametric model estimated using heterogeneous data will have worse fit than an ensemble of locally parametric models fit on more homogeneous partitions. Figure 2.2 demonstrates the proposed model's effectiveness at capturing the cost distribution. The predictive cost distributions are quite similar to the observed distribution. This is not the case for the BART and hurdle models - which fail to capture the high end of the distribution and, therefore, consistently under-predict costs.

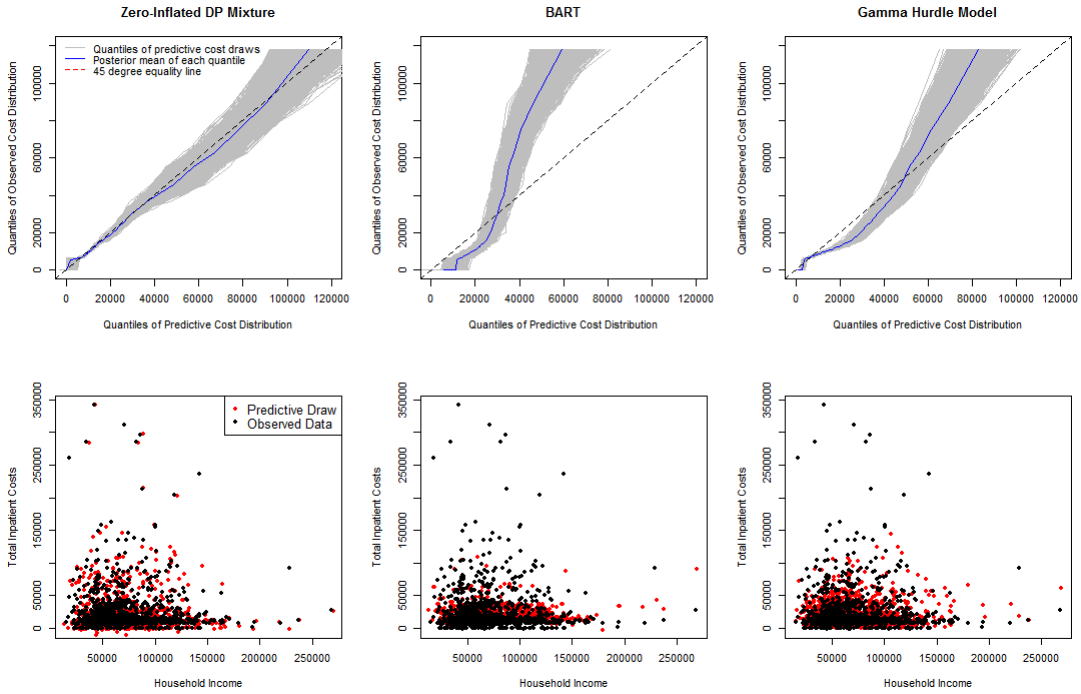


Figure 2.2: Top row: QQ plots of percentiles (.02 - .98 in increments of .02) of the observed cost distribution against predictive cost distributions. Each gray line is a draw of the same size as the data from the predictive cost distribution. The blue line indicates the mean of each percentile across these predictive draws while the dashed line indicates equality (a perfect fit). The DP mixture opens new cluster to capture skewness - resulting in a predictive distribution closely matching the observed data. The BART model and hurdle model cannot capture this extreme skewness. This is also demonstrated in the bottom row: the DP model occasionally predicts very high costs, while predictions from BART and hurdle models hardly ever predict such high costs.

In the second row of plots, we see the DP model occasionally predicts very high costs, while having the bulk of the predictions at $< \$50,000$. Both BART and the Hurdle model capture the lower end of the cost distribution well - also predicting the bulk of the costs at $< \$50,000$. However, they rarely predict costs at the high end - thus, failing to capture skewness.

2.5.3. Estimating Causal Contrasts and Assessing Overlap

Finally, we use our method to estimate differences in costs that would have accumulated over two years under hypothetical interventions where everyone received radiation versus everyone received chemotherapy as their first post-hysterectomy treatment. After applying the method of Section 2.3.3, Figure 2.3 shows there is adequate overlap between the two treatment groups, reducing concerns about positivity violations.

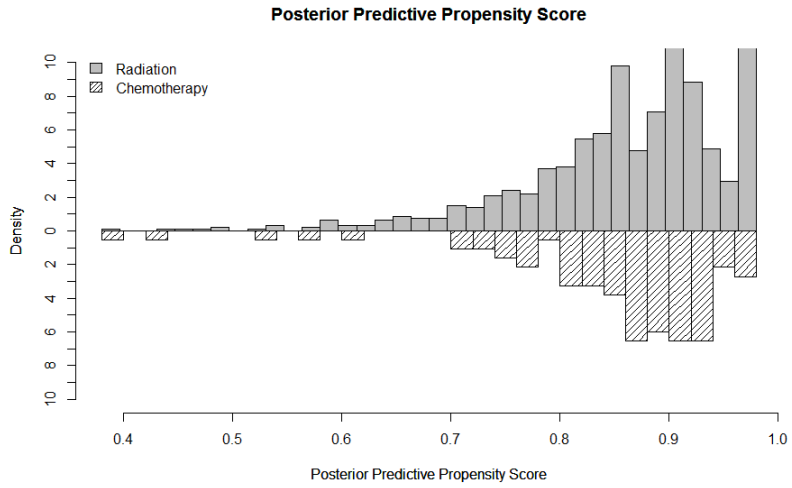


Figure 2.3: Posterior propensity scores calculated using Equation 2.5 for both groups indicate adequate overlap - suggesting no evidence of positivity violations.

We use the method of Section 2.3 to compute a marginal causal effect, a median causal effect, and a risk ratio contrasting the probability of zero cost under radiation versus chemotherapy. Posterior means and credible intervals are displayed in Table 4. Under standard causal identification assumptions, we estimate the causal difference in costs to be $\$1672$ (CI: $-2566, 5722$), showing radiation therapy to be more expensive. We estimate a median causal difference to be $\$872$ (CI: $-833, 2790$). Finally, we estimate that the probability of having zero costs under radiation therapy is 50% (CI: $0.31, 0.78$) lower than under chemotherapy. These results are consistent with unadjusted results (see Table 2.2).

Table 2.3: Posterior means and credible interval from standardization procedure using DP mixture along with posterior marginal causal effect estimates and credible intervals from BART, a Gamma hurdle model, and a Gamma model with zeros replaced by .01. All monetary amounts are in thousands of 2018 U.S. Dollars.

	Marginal Causal Effect	Median Causal Effect	Causal Risk Ratio of Zero Cost
Zero-Inflated DP	1.7	.9	0.498
	(-2.6, 5.7)	(-.8, 2.8)	(0.31, 0.78)
BART	1.8	-	-
	(-6.1, 9.8)	-	-
Gamma Hurdle	2	-	.505
	(-1.5, 5.6)	-	(.34, .76)
Gamma +.01	4.9	-	-
	(1, 8.8)		

marginal causal effects from BART and the Gamma hurdle model are roughly in-line with the DP mixture estimates but suffer from relative ineffectiveness at predicting high costs, as explained in the previous section. The risk ratio estimate from the hurdle model is similar to the DP estimate. We note that, consistent with simulation results, the marginal causal effect estimate from the Gamma +.01 model differs greatly from the other three models.

2.6. Discussion and Future work

The proposed DP mixture is ideal for capturing joint distributions with continuous, zero-inflated outcomes. It is multipurpose: simultaneously modeling structural zeros, inducing clustering to handle multi-modality, and accommodating skewness in the outcome and covariates. As we show in our simulation studies, these traits allow our proposed DP mixture to both produce high-quality causal effects estimates as well as capture the entire outcome distribution. At the same time, posterior draws from the model can be used to perform posterior checks evaluating the validity of positivity.

One might expect BART to perform better than it did in our simulations and in our analysis of cancer data. After all, BART is said to be “effectively nonparametric” (Chipman, George, and McCulloch, 2010). However, while it flexibly models the conditional outcome mean, BART still assumes that the outcome distribution is Gaussian - yielding biased estimates in simulation settings where the

data are drawn from a skewed distribution like the Gamma. Moreover, it does not account for multi-modality as it assumes the data are generated from a single mean function and single error variance. The DP mixture makes no such assumptions - yielding better estimation of the entire outcome distribution, as was shown in Figure 2.2. We note that very recently George et al., 2018 proposed extending BART by modeling the error term nonparametrically using a DP mixture. This may better equip BART to handle skewness, though multi-modality will likely remain challenging.

Finally, we consider several extensions of our model for future work. First, while our model provides a framework for assessing positivity, designing a solution within the framework of our model is an important extension. Second, unmeasured confounding is always a concern in observational studies. In our application this concern is mitigated by the fact that post-hysterectomy treatment assignment mechanism is well defined by American Cancer Society guidelines to be mostly driven by patient comorbidity, age, cancer stage, and grade. In many settings, the treatment mechanism may be less well-understood, necessitating sensitivity analyses. Third, standardization provides valid causal estimates only in scenarios with time-constant treatment and confounding. Extending this model to a setting with time-varying confounding would be a worthwhile endeavor.

CHAPTER 3

BAYESIAN NONPARAMETRIC COST-EFFECTIVENESS ANALYSES VIA ENRICHED DIRICHLET PROCESS PRIOR

3.1. Introduction

Cost-effectiveness analyses (CEAs) are ubiquitous in public health policy and health economics research, with use-cases ranging from treatment comparison to determining drug coverage and informing policy. However, they remain statistically challenging for several reasons. First, cost and effectiveness are often correlated, with joint distributions typically exhibiting extreme skewness and multimodality. In these settings, parametric models that impose strong distributional, linearity, and additivity assumptions are not tenable. Second, in many cases effectiveness is operationalized as gains in survival time - which is prone to right-censoring if subjects drop out before the end of the study. For such patients, we only observe a lower bound on their survival time and accumulated costs. Third, CEAs are often conducted using observational data which are less expensive and more readily available, but are prone to confounding. Valid estimation of CEA contrasts therefore requires adjustment so that differences in cost-effectiveness due to treatment can be disentangled from differences due to confounders.

Early statistical literature (Bang and Tsiatis, 2000; Lin, 2003; Lin, 2000; Lin et al., 1997) focused on cost estimation, while assuming efficacy was constant between treatments. Cost estimation alone is challenging due to the pathological nature of costs (censoring, skewness, zero-inflation, etc). Our work enhances this literature by developing a joint model for cost and survival time, rather than solely focusing on cost. Previous work decomposed the joint distribution into a product of a marginal survival time distribution and a cost distribution conditional on survival time. Huang, 2002 refer to this as a “calibration regression” approach. Handorf et al., 2019 and Huang, 2002 approach the modeling from a frequentist point of view. While the former uses a fully parametric approach, the latter uses a semi-parametric approach - making only first and second moment assumptions. Baio, 2014 took a fully parametric Bayesian approach to joint modeling that did not allow for full covariate adjustment since the data application of interest was from a randomized trial. In contrast, our Bayesian joint modeling approach makes neither strong distributional assumptions nor functional

form (e.g., linearity, additivity) assumptions and allows for covariate adjustment.

Li et al., 2018 took a significant step toward robust causal inference in cost-effectiveness. They formulate causal CEA contrasts in terms of potential outcomes and develop a doubly-robust estimation approach that combines separate conditional mean models for cost and survival with a treatment propensity score model. They show that CEA contrasts can be estimated consistently if either the propensity score or cost/survival regressions are correct. We build on this work in several ways. We also formulate CEA contrasts in terms of potential outcomes - endowing these contrasts with explicitly causal interpretations. However, our modeling approach is fully nonparametric and, therefore, more flexible than the doubly-robust estimator. While, the doubly-robust approach only uses data on uncensored subjects (weighted by the inverse probability of being uncensored) our approach uses data from both censored and uncensored subjects potentially generating efficiency gains. Moreover, our approach is a Bayesian model for the full joint cost-effectiveness distribution - not a weighted combination of separate conditional mean models. This in principle allows for full posterior inference for any function of the joint distribution. Finally, our approach allows for covariate-dependent censoring. Though Li et al., 2018 mention an extension to covariate-dependent censoring, the method proposed and analyzed in their paper relies on randomly censored survival times.

Specifically, our proposed method decomposes the full joint cost-effectiveness distribution into a survival distribution, and a cost model conditional on time. We specify a "local" parametric cost model and a proportional hazard survival model. A Gamma process (GP) prior is placed on the baseline hazard of the survival time distribution while an enriched Dirichlet process (EDP) prior is placed on the cost and survival covariate effects of the local models, jointly. A key property of the EDP is its induced posterior clustering. The EDP probabilistically partitions the dataset into clusters with similar cost-effectiveness covariate effects and associates different "local" models with each cluster. Thus, the joint posterior model for cost-effectiveness is an adaptive mixture of locally parametric models. It is adaptive in the sense that the number of clusters need not be pre-specified. More or less clusters are introduced depending on the complexity of the cost-effectiveness distribution.

Our work also advances the literature in Bayesian nonparametric (BNP) causal inference. An array of nonparametric priors have been successfully applied to causal inference problems (Hill, 2011;

Kim et al., 2017; Roy, Lum, and Daniels, 2017; Xu, Daniels, and Winterstein, 2018; Xu et al., 2016). For instance, Roy et al., 2018 use an EDP prior to model joint outcome-covariate distributions and apply the model to causal estimation with missing-at-random covariates. However, modeling of bivariate counterfactual outcomes using the EDP and GP has not been explored. In CEAs, heterogeneity in cost-efficacy is typically either ignored in favor of a single, marginal effect estimate or is explored along pre-defined subgroups (e.g. hispanic males). Methods in the heterogenous treatment effects literature such as Bayesian Additive Regression Trees (BART)-based procedures (Hahn, Murray, and Carvalho, 2020; Henderson et al., 2018) and Causal Forests (Athey and Wager, 2019) are distinct from our approach as they focus on estimating individual-level treatment effects. Moreover, these methods cannot be readily applied to the joint outcome setting with censoring. Instead, we use the induced clustering of the EDP to *propose* subgroups in a probabilistically principled way. We can then describe each subgroup of the joint in terms of its covariate, cost, and efficacy distributions and use these to motivate future, targeted studies. We propose a "Differential Subgroup Index" which measures how much of the cost-efficacy heterogeneity is explained by the EDP's partitioning of the joint distribution. This helps us assess the meaningfulness of the clusters.

We begin by providing a brief overview of cost-effectiveness and the desirability of causal estimands. We then present our model along with a Markov Chain Monte Carlo (MCMC) algorithm for posterior inference. We incorporate our model into a g-computation framework for posterior causal effect estimation under specified identification assumptions. Finally, we outline how the induced clustering of the EDP can be used to explore heterogeneity. Simulation studies assessing frequentist properties of our causal effect estimates under various censoring scenarios and generating models are conducted. We end with a cost-effectiveness analysis of chemotherapy and radiation therapy treatments for endometrial cancer using SEER-Medicare claims data.

3.2. Overview of Relevant Cost-Effectiveness Contrasts

In this paper, we consider a binary treatment setting where assignment is indicated by $A \in \{0, 1\}$. The goal of CEAs is to characterize the relative cost-effectiveness of these two treatments - necessitating both a cost and efficacy measure. In many settings, the total cost, Y, includes all costs accumulated under this treatment - e.g., hospitalization and medication costs incurred due to adverse events. Moreover, costs are typically measured from the *payer's* perspective, not the patient's

perspective. In single-payer systems like that of the United Kingdom, this would be the National Health Service (NHS). For older patients in the United States, as in our data analysis to follow, the payer of interest is typically Medicare. Though lifetime costs is often of interest, many CEAs set a duration for cost accrual (e.g. 2-year costs) due to follow-up constraints. In this paper, we consider a survival time effectiveness measure, D. This is the dominant effectiveness measure in cancer CEAs, the motivating data application of our paper.

A typical observational CEA study follows diagnosed patients after assignment to one of two treatment regimes. After some follow-up period, everyone's (possibly censored) cost and survival time, are recorded and various cost-effectiveness contrasts can then be computed. For instance, the incremental cost effectiveness ratio (ICER) is given as $ICER = \frac{E[Y|A=1]-E[Y|A=0]}{E[D|A=1]-E[D|A=0]}$. This measures the average cost per unit of effectiveness (increase in survival time). We can also define a monetary value under each treatment, $MV(\kappa) = D\kappa - Y$. Here, κ is the "willingness-to-pay" parameter. It is interpreted as the maximum dollar value the payer is willing to give for a one unit increase in effectiveness. It is considered a fixed, user-specified value. Here, we will suppress notational dependence on κ by simply writing MV where there is no ambiguity. A treatment with positive MV suggests that accrued gains in life value, κD, are greater than accrued costs. Health economists often assess cost-effectiveness via the average net monetary benefit, $E[NMB] = E[MV \mid A = 1] - E[MV \mid A = 0]$, where we have again suppressed dependence of NMB on κ. This contrast is closely related to $ICER$ and can be interpreted as the average difference in monetary value between treatment groups. Note that average NMB can also be written equivalently as $E[NMB] = (E[D \mid A = 1] - E[D \mid A = 0])\kappa - (E[Y \mid A = 1] - E[Y \mid A = 0])$. This is linear function of κ with the efficacy differential as the slope and the cost differential as the intercept. Another related quantity is the Cost Effectiveness Acceptability Curve (CEAC), which is a curve comprised of $P(NMB > 0)$ plotted for various κ.

However, note that MV and NMB presented above have no causal meaning as treated and untreated subjects may differ systematically in observational studies. This is undesirable because many policy questions are inherently causal with the goal being to estimate the average cost-effectiveness that *would have* accrued had everyone taken a particular treatment, possibly counter to fact. Estimation of MV with causal meaning requires (1) an estimate of the joint distribution of cost and survival time while adjusting for confounders and (2) causal identification assumptions.

Even if all relevant confounders are measured and included in the model, misspecification of the adjustment model may yield biased estimates of cost-effectiveness contrasts - motivating the need for robust, nonparametric modeling of the joint. In the following sections we first describe a Bayesian nonparametric model for the joint outcome conditional on confounders and treatment. We then define a causal NMB as the difference in average *potential* monetary value that would have accrued under each treatment. We go on to formulate the identification assumptions required to estimate these causal quantities using our nonparametric joint model.

3.3. Joint Nonparametric Model for Cost and Survival Time

We consider a binary treatment setting in which n patients are assigned to treatment $A_i \in \{0, 1\}$ at baseline. Suppose we are interested in contrasting cost-effectiveness over τ periods (e.g. $\tau = 2$ year cost-effectiveness). We observe data $\mathcal{D} = \{Y_i, T_i, X_i, \delta_i\}_{i=1:n}$ from this study. Here, $X_i = (A_i, L_i)$ is a covariate vector that contains the treatment indicator and a vector of q categorical or continuous pre-treatment confounders, L_i. For notational convenience, we proceed without an intercept, but a 1 can be included as the first entry of X_i. We let $T_i = min(D_i, C_i, \tau)$ be the observed time under study (the minimum of a random right-censoring time C_i, end of study τ, and death time D_i). Define a censoring indicator as $\delta_i = I(D_i > min(C_i, \tau))$ Finally, $Y_i \in \mathcal{Y}$ denotes cost accumulated through time T_i. The joint distribution can be factored into a distribution for observed time and cost distribution conditional on time. A joint model follows from specifying "local" models for each of these two distributions:

$$\begin{aligned} Y_i \mid T_i, \delta_i, X_i, \omega_i &\sim p(Y_i \mid T_i, \delta_i, X_i, \omega_i) \\ T_i \mid \delta_i, X_i, \theta_i, \lambda_0 &\sim \lambda_0(t) \exp(X_i' \theta_i) \\ \omega_i, \theta_i \mid G &\sim G. \end{aligned} \tag{3.1}$$

At a particular time, T, cost follows some local distribution $p(Y_i \mid T_i, \delta_i, X_i, \omega_i)$ governed by parameters ω_i. Survival time follows some local hazard function which is parameterized as having some baseline hazard, λ_0 with covariate effects, θ_i, multiplying this baseline hazard. Lastly, ω_i and θ_i - the covariate effects of the cost and effectiveness model - both follow some joint prior distribution G, which is unknown. Choice of the local models are application-specific but are not crucial for model fit, as will become apparent when we discuss the nonparametric priors used for G and λ_0.

One consideration when choosing the local model is desired predictive support. For instance, if costs are sufficiently far from zero, we may be willing to set $p(Y_i \mid T_i, \delta_i, X_i, \omega_i)$ to a Gaussian over $\mathcal{Y} = \mathbb{R}$ with mean and variance $\omega_i = (\mu_i, \phi_i)$. The corresponding regression could be specified as $\mu_i = (T_i, \delta_i, X_i)'\beta_i$. If the non-negative nature of costs must be respected, we could instead specify a log-normal distribution over $\mathcal{Y} = \mathbb{R}^+$. For applications with zero-inflated costs, we may wish to explicitly put positive measure on zero - i.e. setting $\mathcal{Y} = \{0\} \cup \mathbb{R}^+$. This can be done by specifying a two-part model $Y_i \mid T_i, \delta_i, X_i, \omega_i \sim \pi_i \delta_0(Y_i) + (1 - \pi_i) f(Y_i \mid T_i, \delta_i, X_i, \beta_i)$, where $\pi_i = P(Y_i = 0 \mid T_i, \delta_i, X_i, \gamma_i)$ is a covariate-dependent model for the probability of cost being zero (e.g. a local logistic regression) and δ_0 is the point mass distribution at 0. In this case, the cost parameter vector is $\omega_i = (\gamma_i, \beta_i)$. Oganisian, Mitra, and Roy, 2020 developed a nonparametric Bayesian estimation procedure for such a two-part model, where f could be either log-Normal or Normal, using a Dirichlet Process prior.

In (3.1), censored patients contribute to the likelihood through both the cost and survival time models. In the survival model, they contribute to the likelihood through the survival function in the usual way, provided that, conditional on covariates, censoring times are independent of survival times. Specifically, the probability distribution of death times can be expressed in terms of the hazard above, $\lambda(d; \theta_i, X) = \lambda_0(d) \exp(X_i' \theta_i)$, as $p(d \mid \theta_i, \lambda_0, X_i) = \lambda(d; \theta_i, X_i) \exp\Big(- \int_0^d \lambda(u; \theta_i, X_i) du \Big)$. Similarly, the survival distribution can be expressed as $P(D > d \mid X_i, \theta_i) = \exp\Big(- \int_0^d \lambda(u; X_i, \theta_i) du \Big)$. Censored subjects ($\delta_i = 0$) contribute to the likelihood via $P(D > C_i \mid \theta_i, \lambda_0, X_i)$. Failed patients ($\delta_i = 1$) contribute via $p(D_i \mid \theta_i, \lambda_0, X_i)$. In the cost model, dead patients provide information about the cost distribution at death time $p(Y_i \mid T = D_i, \delta = 0, X_i, \omega_i)$, while censored subjects inform the model at time of censoring $p(Y_i \mid T = C_i, \delta = 1, X_i, \omega_i)$.

3.3.1. Nonparametric Priors

We specify the following nonparametric priors on the unknown model quantities, G and λ_0.

$$
\begin{aligned}
G \mid \alpha_\omega, \alpha_\theta &\sim EDP(\alpha_\omega, \alpha_\theta, G_0) \\
\lambda_0 \mid b, \lambda_0^*, \xi &\sim GP(b\lambda_0^*, b, \xi),
\end{aligned}
\tag{3.2}
$$

Above, EDP denotes the Enriched Dirichlet Process (Wade et al., 2014) prior on G and GP denotes the dependent Gamma Process prior (Nieto-Barajas and Walker, 2002) on the baseline hazard

λ_0. These priors are nonparametric in the sense that they are probability measures on infinite-dimensional objects - the former over probability distributions and the latter over hazard functions. Realizations, G, from the EDP are discrete probability distributions centered around a base distribution $G_0(\omega_i, \theta_i) = G_{0\omega}(\omega_i)G_{0\theta|\omega}(\theta_i|\omega_i)$ with two concentration parameters, α_ω and α_θ. Some prior realizations are visualized in the left panel of Figure 3.1. Just as with the Dirichlet Process (DP), this discreteness induces a posterior clustering of patients. Unlike the DP, the clustering induced by the EDP is nested. *A posteriori*, patients with similar cost parameters are clustered together into what we call ω-clusters. Within each ω-cluster, patients with similar effectiveness parameters are clustered together (θ-clusters). The EDP prior does not require pre-specification of the number of clusters. The clustering is data-adaptive, with more clusters being introduced to capture more complex cost-effectiveness distribution. The posterior model for the joint distribution is an adaptive nested mixture of cost-effectiveness models - with each component model having the form of the local model in (3.1), but with different component-specific parameters. In the machine learning literature, these models are often referred to as "mixture of experts" learners: the data space are partitioned into homogenous regions, each having its own model that develops "expertise" in that region. This is in contrast to ensemble learners (e.g. BART and Random Forests), which apply multiple models to the *entire* data and combine the results post-hoc.

The GP can be thought of as a prior over the space of hazard functions. Each realization λ_0 from the GP is a hazard function centered around a mean function λ_0^* with concentration parameter b. Some prior realizations are visualized in the right panel of Figure 3.1. The process is "dependent" in that it induces a prior AR(1) autocorrelation structure on λ_0: the hazard at time point t is a weighted average of the hazard at the previous time point and the prior hazard, λ_0. The resulting shrinkage/smoothness, controlled by hyperparameter ξ, regularizes the empirical estimate of the baseline hazard - which can be erratic at later time points when the at-risk set becomes small.

These prior choices are motivated by the shortcomings of the standard DP. A potential issue with specifying $G \sim DP(\alpha G_0)$ is that it imposes a single layer of clustering for both cost and effectiveness. Many clusters may be introduced to fit the joint of Y and T if one of these dimensions is more complex - even if the other is very simple. This makes estimates needlessly variable. The nested nature of the EDP avoids this by allowing varying number of clusters on each dimension controlled by separate concentration parameters. Thus, it is possible to introduce a single cost cluster that

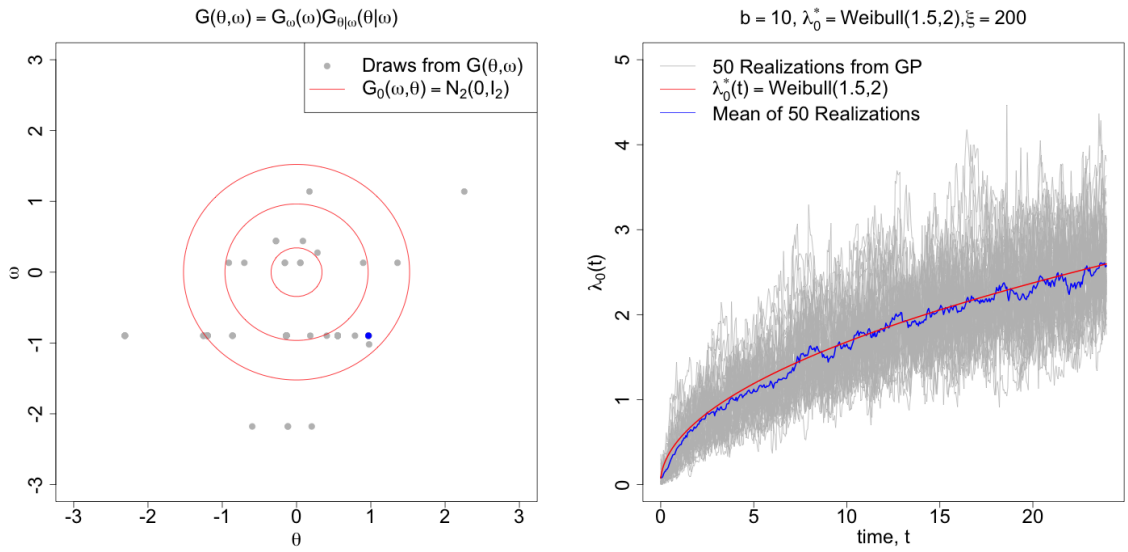


Figure 3.1: Realizations of the Enriched Dirichlet and Gamma Processes. (a) 100 draws of $(\theta, \omega) \sim G$ where $G \sim EDP(10, 10, N_2(0, I_2))$. Note the nested discreteness of G causes ties (i.e. clustering) among the draws: there are 80 other draws with the same ω value as the blue point, but with different θ values. Twenty three of those 80 also have the same θ value. (b) Gray lines show 50 hazard realizations from a gamma process centered around the hazard of a $Weibull(1.5, 2)$ distribution. The blue line shows the mean of the 50 realizations.

has many survival time subclusters. Similarly, modeling the baseline hazard separately avoids introduction of excess clusters to fit a potentially complicated function which, for causal estimation purposes, is just a nuisance parameter. This is also the reason why we opt for a proportional hazard (PH) formulation rather than an accelerated failure time (AFT) approach: PH models clearly separate the covariate effects from the baseline risk, which we do not want influencing the EDP mixture.

3.3.2. Posterior Inference using Markov Chain Monte Carlo

Inference for (3.1) is done via MCMC. We follow the general scheme of Neal's algorithm 8 (Neal, 2000), which introduces auxiliary parameters to sample from the DP posteriors. Roy et al., 2018 used this approach to sample EDP posteriors, though without a Gamma Process update and no joint outcome considerations. The idea is to introduce latent cluster indicators (the auxiliary parameters) for each subject. Conditional on draws in the previous iteration, each MCMC iteration then updates clustering indicators conditional on parameters and before updating cluster-specific parameters conditional on these newly updated indicators. At iteration m, we may have $J^{(m)}$ oc-

cupied ω-clusters indexed by $j \in \{1, \ldots, J^{(m)}\}$ and, within the j^{th} ω-cluster, we may have $K_j^{(m)}$ occupied θ-clusters indexed by $k_j \in \{1, \ldots, K_j^{(m)}\}$. Let $c_{1:n} = (c_1, \ldots, c_n)$ be cluster assignment indicators where each c_i is a length two vector with first and second entry indicating membership to an ω-cluster and θ-subcluster, respectively. Throughout, we use the notation $v_{a:b}$, where $a < b$ are integers, to denote the collection $(v_a, v_{a+1}, \ldots, v_b)$. Let $\omega_{[j]}$ represent the cost parameter associated with cluster j and $\theta_{[j,k]}$ represent the effectiveness parameter associated with the k^{th} subcluster of ω-cluster j. We should strictly denote $\theta_{[j,k]}$ as $\theta_{[j,k_j]}$ but suppress the subscript throughout wherever reference is clearly made to the k^{th} subcluster of ω-cluster j. Moreover, define n_j^{-i} and $n_{j,k}^{-i}$ as the number of subjects (excluding subject i) currently occupying ω-cluster j and ω-θ cluster (j,k), respectively, at the current iteration, m. At each iteration m we conduct the following sequence of conditional posterior updates:

- *Update cluster membership*:
  - Propose parameters for a new θ-subcluster for each existing ω-cluster, $\{\theta_{[j,K_j^{(m)}+1]} : j \in 1, \ldots, J^{(m)}\}$ by drawing from the prior G_0.
  - Similarly, propose parameters for new ω-cluster with θ subcluster, $\{\omega_{[J^{(m)}+1]}, \theta_{[J^{(m)}+1,1]}\}$.
  - Conditional on current draws of all cost-effectiveness parameters and $\lambda_0^{(m)}$ (indicated by "$-$" for compactness), update $c_i^{(m)}$ according to the following probabilities:

$$P(c_i^{(m+1)} \mid -, \mathcal{D}) \propto \begin{cases} \frac{n_j^{-i} n_{j,k}^{-i}}{n_j^{-i} + \alpha_\theta} p(Y_i, T_i \mid X_i, \delta_i, \omega_{[j]}^{(m)}, \theta_{[j,k]}^{(m)}, \lambda_0^{(m)}) & \text{existing } j,k \\ \frac{n_j^{-i} \alpha_\theta}{n_j^{-i} + \alpha_\theta} p(Y_i, T_i \mid X_i, \delta_i, \omega_{[j]}^{(m)}, \theta_{[j,K_j^{(m)}+1]}^{(m)}, \lambda_0^{(m)}) & \text{existing } j, \text{ new } k \\ \alpha_\omega p(Y_i, T_i \mid X_i, \delta_i, \omega_{[J^{(m)}+1]}^{(m)}, \theta_{[J^{(m)}+1,K_j^{(m)}+1]}^{(m)}, \lambda_0^{(m)}) & \text{new } j,k \end{cases}$$

- *Update cluster parameters*: These require Metropolis-Hastings steps if $G_0\omega$ or $G_{0\theta|\omega}$ are not conjugate.
  - Update each cluster's cost parameter, $\omega_{[j]}$, by drawing from conditional posterior

$$\omega_{[j]}^{(m+1)} \sim p(\omega_{[j]} | c_{1:n}^{(m+1)}, \mathcal{D}) \propto G_{0\omega}(\omega_{[j]}) \prod_{i | c_i^{(m+1)} \in (j,\cdot)} p(Y_i | T_i, X_i, \delta_i, \omega_{[j]})$$

  - For each j, update all $\theta_{[j,k_j]}$ by drawing from conditional posterior

$$\theta_{[j,k]}^{(m+1)} \sim p(\theta_{[j,k]} | c_{1:n}^{(m)}, \lambda_0^{(m)}, \mathcal{D}) \propto G_{0\theta|\omega}(\theta_{[j,k]}) \prod_{i | c_i^{(m)} \in (j,k)} p(T_i \mid X_i, \delta_i, \lambda_0^{(m)}, \theta_{[j,k]})$$

- *Update baseline hazard,* $\lambda_0^{(m+1)}$: This is a multi-step update involving a discretization of the

time interval $[0, \tau]$ into increments, then modeling the hazard rate in each increment. This is motivated by the fact that if λ_0 follows a Gamma Process, then the hazard rates in any finite partition of the time interval have Gamma distributions (Nieto-Barajas and Walker, 2002). Additionally, the latent parameters inducing the AR(1) smoothness across increments are also updated with a mix of grid sampling and adaptive Metropolis steps. Details are provided in Appendix B.2.

Note that the induced nested clustering of the EDP is explicitly encoded into this sampler. In the cluster-update step, a given subject is most likely to be assigned to the cluster with parameters that yield the highest joint-distribution evaluation (i.e. fit their data the best). Moreover, each subject can possibly be assigned to a new cost cluster, new effectiveness cluster within an existing cost cluster, or a new cost-effectiveness cluster. This last event is likely to occur if, for example, the subject is so unique that random parameter draws from the prior fit that subject's data better than any of the existing cluster-specific parameters. Furthermore, note that each term for an existing cluster in $P(c_i^{(m+1)} = (j,k) \mid -, D)$ is an increasing function of the number of patients already assigned to that cluster. This is the well-known "rich-get-richer" property of the EDP - the *a priori* favoring of assignment to larger clusters. This prevents over-fitting by penalizing small clusters. After every cycle, $c_i^{(m+1)}$ maps each subject to a set of updated parameters $(\omega_i^{(m+1)}, \theta_i^{(m+1)}, \lambda_0^{(m+1)})$. After a sufficient burn-in period this algorithm produces M draws from the posterior $\{\omega_{1:n}^{(m)}, \theta_{1:n}^{(m)}, \lambda_0^{(m)}, c_{1:n}^{(m)}\}_{1:M}$. These can be used to do full posterior inference on any functional of the joint including, as we will see, causal estimands.

3.3.3. Priors and Hyperparameter Choice

The hyperparameters for the EDP are the base distribution $G_0(\omega_i, \theta_i) = G_{0\omega}(\omega_i) G_{0\theta|\omega}(\theta_i|\omega_i)$ and the concentration parameters α_θ and α_ω. Following previous papers (Oganisian, Mitra, and Roy, 2020; Roy et al., 2018), we use prior independence so that $G_0(\omega_i, \theta_i) = G_{0\omega}(\omega_i) G_{0\theta}(\theta_i)$ and set $G_{0\theta}(\theta_i) = N(\hat{\theta}_{PH}, \nu_\theta \hat{C}_{PH})$. Here, we are centering the cluster-specific covariate effects around the Cox proportional hazard estimate, $\hat{\theta}_{PH}$. The prior covariance matrix, $\hat{C}_{PH}$, is diagonal with the square of the Cox proportional hazard standard error estimates along the diagonal. The parameter $\nu_\theta > 0$ is a user-specified scalar that controls how tightly or widely dispersed the cluster-specific effects are around the Cox estimates.

The choice of $G_{0\omega}(\omega_i)$ depends on the choice of local cost model. Suppose our local model, $p(Y_i T_i, \delta_i, X_i, \omega_i)$ is Gaussian, $N(\mu_i, \phi_i)$ with regression $\mu_i = E[Y_i \mid T_i, \delta_i, X_i, \omega_i] = (\delta_i, T_i, X_i)' \beta_i$

and variance ϕ_i, where β_i is the vector of covariate effects. The full cost parameter vector is $\omega_i = (\beta_i, \phi_i)$ and we could set $G_{0\omega}(\beta_i, \phi_i) = N(\beta_i; \hat{\beta}, \nu_\omega \hat{\Sigma}) IG(\phi_i; shape = a_0, scale = \hat{s}^2(a_0 - 1)$. The vector $\hat{\beta}$ is the MLE estimate of the cost regression and $\hat{\Sigma}$ is a diagonal matrix with the square of the standard error estimates along the diagonal. The parameter $\nu_\omega > 0$ is user-specified and controls the tightness of the prior around $\hat{\beta}$. Similarly, the Inverse Gamma prior for ϕ_i having mean equal to the empirical outcome variance, $\hat{s}^2 = \frac{1}{n-1}(Y_i - \bar{Y})^2$. The user-specified parameter, a_0, controls how widely the cluster-specific variances are dispersed around the empirical variance, with higher values corresponding to a tight prior around the empirical estimate. Finally, we follow previous approaches (Oganisian, Mitra, and Roy, 2020; Roy et al., 2018) and set $Gam(1, 1)$ (i.e. flat, uninformative) priors on each of the concentration parameters. These parameters can be interpreted as prior sample sizes for the cost and effectiveness clusters - higher values on average lead to more occupied clustering. Thus, this Gamma prior penalizes many occupied clusters, but has a long tail to allow posterior deviations if demanded by the data.

Finally, we center the Gamma Process prior around a constant hazard function. Specifically, we compute the Nelson-Aalen estimate of the baseline cumulative hazard, then take the difference between each point on this curve to obtain the baseline hazard estimate at each time point. We then compute the average of these hazard rates across time, $\hat{\lambda}$. Then, in $GP(b\lambda_0^*, b, \xi)$ we can set λ_0^* to be exponential with rate $\hat{\lambda}$. Intuitively, this expresses the prior belief of a constant hazard (with rate in the range of the observed rates). However, if the data disagrees, the posterior will move us to a richer estimate governed by the data. The parameters ξ and b can be used to calibrate degrees of informativeness. For example ξ near zero and large b corresponds to an informative prior belief of a constant hazard. Conversely, values of b near 0 correspond to an uninformative prior.

3.4. Posterior Causal Estimation via g-Computation

Here we describe full posterior inference for various causal estimands expressed in terms of potential outcomes (Rubin, 1978). In scenarios with censored outcomes, causal estimands are typically formulated under a hypothetical "joint intervention" (Robins, Hernán, and Brumback, 2000) on *both* treatment and censoring. Let $MV^{A=a,\delta=0} = D^{A=a,\delta=0}\kappa - Y^{A=a,\delta=0}$ be the monetary value that would have accrued over τ periods had the patient received treatment a and not been censored. The components $D^{a,0}$ and $Y^{a,0}$ are the survival time and costs, respectively, that would have been

observed under treatment $A = a$ had the subject not been censored. The *population-level* estimand of interest is $\Psi = E[NMB] = E[MV^{1,0}] - E[MV^{0,0}]$. This is the average difference in monetary value that would have accrued over τ periods had everyone in the target population been assigned to treatment 1 versus treatment 0, and not been censored. In general, interventions in observational CEAs are not random. Instead, they are driven by confounders - factors which both influence treatment and cost-effectiveness. Thus, $E[MV^{a,0}] \neq E[MV \mid A = a, \delta = 0]$ in general, since those who actually received treatment and remained uncensored may not be representative of the target population. Suppose, however, that we observe a set of pre-treatment confounders, L. Under the following extensions of the usual causal identification assumptions, we can identify Ψ:

IA.1 *Joint ignorability*: $(Y^{a,\delta}, D^{a,\delta}) \perp (A, \delta) \mid L$. Conditional on L, censoring and treatment should be as good as random - being completely independent of the death and costs that would have accrued under a particular treatment. Omission of unmeasured drivers of both the joint intervention or cost-effectiveness would result in a violation of this assumption.

IA.2 *Joint Consistency*: $(Y^{a,0}, D^{a,0}) = (Y, D) \mid A = a, \delta = 0$. This requires that cost and death time observed for an uncensored ($\delta = 0$) subject assigned treatment $A = a$ is actually $(Y^{a,0}, D^{a,0})$. This could be be violated if, for instance, we had non-compliance to the treatment. Then, a subject assigned a may not have actually taken a and thus we would not observe $Y^{a,0}$.

IA.3 *Joint Positivity*: $0 < P(A = a, \delta = 0 \mid L) < 1$. The joint intervention cannot be deterministic at any level of L. This could be violated if, for example, all uncensored males received treatment $A = 1$ - leaving us with no information on how well uncensored males with treatment $A = 0$ faired. In these cases, the model may extrapolate the outcome under treatment $A = 0$ learned from females onto males. Poor extrapolation could lead to bias.

IA.4 *No Joint Interference*: $(Y_i^{a_{1:n},\delta_{1:n}}, D_i^{a_{1:n},\delta_{1:n}}) = (Y_i^{a_i,\delta_i}, D_i^{a_i,\delta_i})$. Here, $a_{1:n}$ and $\delta_{1:n}$ are n-dimensional vectors containing each subject's treatment and censoring status. This assumption requires that one person's joint treatment-censoring intervention cannot impact another's cost-effectiveness. It allows us to drop all but the i^{th} element of $a_{1:n}$ and $\delta_{1:n}$. Usually this assumption would be violated in infectious disease exposures or other settings where subjects cannot be reasonably viewed as exchangeable (one person's infection status may impact another's infection probability).

Under these assumptions, Ψ is identified via Robins' g-formula (Robins, 1986)

$$\Psi(\omega_{1:n}, \theta_{1:n}, \lambda_0) = \int_{\mathcal{L}} \Big(E[MV \mid A = 1, \delta = 0, L, \omega_{1:n}, \theta_{1:n}, \lambda_0] \\ - E[MV \mid A = 0, \delta = 0, L, \omega_{1:n}, \theta_{1:n}, \lambda_0] \Big) dP(L) \tag{3.3}$$

Details are provided in the Appendix B.1. Above, we have explicitly written $\Psi = \Psi(\omega_{1:n}, \theta_{1:n}, \lambda_0)$ as a function of the parameters governing the joint cost-effectiveness distribution. This is to highlight that a posterior distribution over these parameters induces a posterior on the the causal estimand Ψ. Let each expectation in (3.3) be denoted as $\mu(a, 0) = E[MV \mid A = a, \delta = 0, L, \omega_{1:n}, \theta_{1:n}, \lambda_0]$. Then,

$$\mu(a, 0) = \int_0^{\tau} \int_0^{\infty} (D\kappa - Y) p(Y, T \mid L, A = a, \delta = 0, \omega_{1:n}, \theta_{1:n}, \lambda_0) dY dD \tag{3.4}$$

Where this inner integration is over the joint model we presented in (3.1) with $X_i = (A_i, L_i)$. Note that conditional on $\delta = 0$, $T = D$ in the joint model and we integrate along the time up until τ - resulting in τ-period monetary value. This integration can be done efficiently via Monte Carlo (see Appendix B.2).

The outer integration over $\mathcal{L}$ in (3.3) requires an estimate of $P(L)$. To avoid strong parametric assumptions, we use a Bayesian bootstrap (Rubin, 1981). That is, we express $p(L)$ as a discrete distribution with mass p_i at the i^{th} observed confounder vector L_i. Specifically, $p(L = l) = \sum_{i=1}^{n} p_i \cdot \delta_{L_i}(l)$. Here δ_{L_i} is a point-mass at L_i. The Bayesian bootstrap follows from an improper Dirichlet prior on the weights, $p_{1:n} = (p_1, \ldots, p_n) \sim Dir(0, \ldots, 0)$. This yields a conjugate posterior $p_{1:n} \mid L \sim Dir(1, \ldots, 1)$ with $n-$dimensional posterior mean vector $E[p_{1:n} \mid L] = (1/n, 1/n, \ldots, 1/n)$.

At the end of the m^{th} iteration of updates from Section 3.3.2, we have a set of parameter draws $\{\omega_{1:n}^{(m)}, \theta_{1:n}^{(m)}, \lambda_0^{(m)}\}$, which we can use to construct a posterior draw of monetary value $\mu_i^{(m)}(a, 0) = E[MV \mid A = a, \delta = 0, L_i, \omega_i^{(m)}, \theta_i^{(m)}, \lambda_0^{(m)}]$. We then take a draw $p_{1:n}^{(m)}$ from the Dirichlet posterior and construct a draw of the confounder distribution $p^{(m)}(L = l) = \sum_{i=1}^{n} p_i^{(m)} \cdot \delta_{L_i}(l)$. Substituting both of these into (3.3), yields a draw from the posterior of Ψ

$$\Psi^{(m)} \approx \sum_{i=1}^{n} p_i^{(m)} \Big(\mu_i^{(m)}(1,0) - \mu_i^{(m)}(0,0) \Big) \tag{3.5}$$

Repeating for iterations $m = 1, \ldots, M$ yields M draws from the posterior of the causal τ-period NMB: $\{\Psi^{(m)}\}_{1:M}$. The mean of these draws can serve as a Bayesian nonparametric point estimate of Ψ and percentiles of the M draws can be used to form credible intervals.

The posterior draws can also be used to compute a point on the CEAC for each κ, $P(NMB > 0 \mid D) \approx \frac{1}{M} \sum_m I(\Psi^{(m)} > 0)$. We note that, from this Bayesian perspective, each point on the CEAC is a posterior p-value or tail-area probability. If individual-specific estimates are required, Equation (3.4) can be evaluated for particular L_i under both treatments using each of the m posterior parameter draws. The difference would be a draw from the posterior of $\Psi_i = NMB_i(\kappa)$, denoted $\Psi_i^{(m)} = \mu_i^{(m)}(1,0) - \mu_i^{(m)}(0,0)$. In the causal literature, these are variously referred to as conditional average treatment effects (CATEs) or individual treatment effects (ITEs). Across M iterations, we would also have subject-level credible intervals for Ψ_i. The left panel of Figure 3.2 visualizes posterior mean and intervals for each Ψ_i using an illustrative synthetic example.

3.5. Adaptive Subgroup Discovery

The MCMC scheme of Section 3.3.2 yields posterior draws of latent cost-effectiveness cluster membership, $\{c_i^{(m)}\}_{1:M}$. In this section, we propose using these draws to adaptively discover subgroups of patients with different cost-effectiveness profiles. This is "adaptive" in the sense that the number of clusters is not pre-specified, but grows or shrinks as the model adapts to the data complexity. Subgroup discovery is a policy-relevant endeavor since current CEA practice tends to focuses on marginal, population-level analyses - even if there is significant variation in the target population. Existing approaches to heterogeneity Athey and Wager, 2019; Hahn, Murray, and Carvalho, 2020; Henderson et al., 2018 focus on computing ITEs and use post-hoc heuristic procedures to characterize this heterogeneity across pre-defined subgroups - rather than proposing subgroups adaptively.

Using the given MCMC outputs for subgroup discovery is challenging for two reasons. First, the vector of cluster assignment labels, $c_{1:n}^{(m)}$, have no meaning across MCMC iterations - making it difficult to determine the posterior mode partition. This is known as label switching (Stephens,

2000). To illustrate, consider that a new cost-effectiveness cluster forms in iteration $m+1$ and all subjects previously in another cluster are re-assigned to this new cluster. In this case, even though the assignment has changed, the underlying composition of the cluster did not. As a solution, we propose to keep track of the $n \times n$ adjacency matrix $\mathcal{C}^{(m)}$, where the ij^{th} element, $\mathcal{C}_{ij}^{(m)}$, is a binary indicator of subject i and j being in the same cost-effectiveness cluster at iteration m. Note that this is just the vector $c_{1:n}^{(m)}$ re-arranged into a matrix. Taking the elementwise mean of this matrix across the m posterior draws yields a probability matrix $\mathcal{P} = (1/M)\sum_m \mathcal{C}^{(m)}$ where ij^{th} element, $\mathcal{P}_{ij}$, is the posterior probability of subject i and j being in the same cost-effectiveness cluster. To get a hard clustering assignment, we then search draws, $\{c_{1:n}^{(m)}\}_{1:M}$, for the assignment that is "closest" to $\mathcal{P}$. That is, we search for $c_{1:n}^{*} = \arg\min_m ||\mathcal{C}^{(m)} - \mathcal{P}||$, where $||\cdot||$ is some matrix norm. As in earlier papers on Bayesian clustering, here we adopt "Binder's Loss" $||\cdot|| = \sum_{i,j}(\mathcal{C}_{ij}^{(m)} - \mathcal{P}_{ik})^2$ (Binder, 1978; Dahl, 2006). This essentially approximates the posterior mode of the EDP-induced partition, $\mathcal{P}$. The middle panel of 3.2 visualizes $\mathcal{P}$ from an illustrative synthetic example as a weighted graph where each subject is a node and the length of vertices connecting two nodes are inversely proportional to $\mathcal{P}_{ij}$. Subjects with low posterior probability of being in the same cost-effectiveness cluster are far apart on the graph. Such figures are good tools for assessing uncertainty in posterior mode assignments, c_i^*. For instance, the points between the group of dark red and blue clusters represent subjects with highly uncertain mode assignments. The covariate effects of these subjects look just as similar to the well-separated dark blue points as they do to the well-separated dark red points.

A second challenge with using the assignments for subgroup discovery is that the EDP clusters are not explicitly designed to cluster on NMB. The clustering is driven by the complexity of the joint cost-effectiveness distribution. This is necessary for a flexible joint distribution estimate, but may not translate into meaningful NMB clusters. For instance, consider a bimodal cost-effectiveness distribution with two groups having very different mean costs. However, the difference in costs between treatment groups in both clusters may be the same. In this case, the EDP will likely introduce two clusters with similar $NMBs$. This begs the question: are the clustering results detecting subgroups with different cost-effectiveness profiles? To answer this question, we propose a posterior Differential Subgroup Index (DSI) that, at each MCMC iteration, computes the proportion of the total variation in the ITEs, $\Psi_i^{(m)}$, that is explained by the cluster partition in that iteration. First, de-

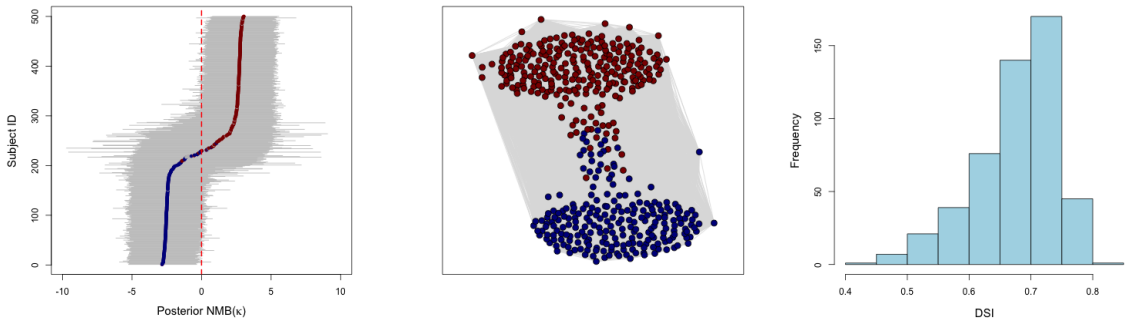


Figure 3.2: Clustering results from EDP-GP fit using synthetic data with two latent cost-effectiveness clusters. Here, EDP-induced clusters on the joint distribution capture differences in NMB. The left panel shows posterior point and 95% interval estimates Ψ_i (with $\kappa = 1$). Colors indicate posterior model cluster assignment, $c^*_{1:n}$. The middle panel visualizes the posterior probability matrix $\mathcal{P}$. The right panel is the posterior distribution of DSI - indicating that about 70% of the variation in subject-level Ψ_i is explained by the EDP clustering. However, this need not be the case. The EDP clusters may be capturing complexities unrelated to NMB. While this is desirable to obtain a good fit to a complex distribution, it means the clusters have no substantive meaning. The DSI is necessary to distinguish between these scenarios.

fine the mean NMB in subject i's cluster at iteration m: $\bar{\Psi}_i^{(m)} = \frac{1}{\sum_j I(c_j^{(m)}=c_i^{(m)})} \sum_{j|c_j^{(m)}=c_i^{(m)}} \Psi_j^{(m)}$. Then the DSI measure is,

$$DSI^{(m)} = \frac{\sum_i \left(\bar{\Psi}_i^{(m)} - \Psi^{(m)}\right)^2}{\sum_i (\Psi_i^{(m)} - \Psi^{(m)})^2} \tag{3.6}$$

This intuitively plays the same role as a regression R^2 statistic. Across the m iterations, we have a set of draws for this statistic, $\{DSI^{(m)}\}_{1:M}$, which reflects our uncertainty about how well the clustering is capturing heterogeneity in NMB. A posterior distribution for DSI concentrated near 1 suggests that the EDP-induced clustering explains nearly all of the variation in the subject-specific $NMBs$. This implies that the EDP-induced clustering at the joint cost-effectiveness level is capturing variation at the NMB level. The right panel of 3.2 plots the posterior distribution for DSI for an illustrative synthetic example generated with two cost-effectiveness clusters. We can then summarize our data along the mode partition, $c^*_{1:n}$. For instance, in the synthetic example, we can create a table summarizing the observed costs, survival, and covariate distributions of the two identified clusters. These can be used to motivate future cost-effectiveness studies targeting these subgroups. The DSI also provides context for our marginal posterior estimate, Ψ. A high DSI indicates that a marginal estimate is not capturing substantial treatment effect heterogeneity detected by the EDP posterior.

3.6. Assessing Frequentist Properties via Simulation

In this section we report results of several simulation experiments exploring the frequentist properties (i.e. bias, coverage, and precision) of our posterior mean and interval estimates for Ψ under a variety of settings. These results are reported in Table 3.1. We simulate data with one continuous confounder, four binary confounders, and a binary treatment. We simulate survival times conditional on treatment and confounders from a Weibull distribution. Survival times are censored by censoring times that also follow a covariate-dependent Weibull distribution. We simulate an outcome from a true Y distribution of either a Gaussian or Log-Normal, with confounder- and treatment-dependent means. Data were simulated under low (5%) and high (20%) covariate-dependent censoring. For each of these, we simulate under a parametric and bimodal setting. Under the parametric setting, the joint distribution is unimodal - leading to a simple joint cost-survival distribution. Under the bimodal setting, we simulate data from a mixture of two cost-effectiveness distributions, each having different covariate effects in the cost and survival time models. In each of these eight settings, we simulate 200 datasets with 1500 subjects each. Details about the data generation are given in Appendix B.3.

We include the doubly-robust estimator (DR-SL) of Li et al., 2018 as a comparator. This approach involves estimating separate models for conditional mean cost and conditional mean survival time via super learner. Predictions from these models are weighted by the product of the inverse probability of treatment and inverse probability of censoring. We estimate the former using a correctly specified logistic regression - which suggests the DR estimate will be consistent but may still have substantial bias in finite samples if the models are inadequate. For the latter, we note that Li et al. did not consider the covariate-dependent censoring in their analysis. Instead, they estimate the probability of censoring in both treatment groups separately via Kaplan-Meier. Li et al. suggest using a discrete-time failure model in situations with covariate-dependent censoring. Here, we contribute to the literature by implementing this suggestion using a logistic regression. In the super learner libraries, we include regression trees, generalized additive, linear models, as well as elastic net generalized linear model (GLMnet). As recommended by Li et al., we using the bootstrap BCa interval for inference.

For the EDP-GP, we run using independent Gaussian base distributions for G_0 that are null centered

Table 3.1: Simulation Results. Average bias of posterior mean NMB (as discussed in Section 3.4) along with coverage and average width of 95% credible/confidence interval (CI) is reported for EDP-GP model. Point estimate is reported for DR-SL along with coverage and width of 95% bootstrap BCa interval. Bias is reported as a proportion of the truth. Censoring rate was 5% in the low setting and 20% in the high setting. Willingness-to-pay is set to $\kappa = 1$. Results are across 200 simulated datasets with $N = 1500$ subjects each.

Simulation Setting			EDP-GP			DR-SL		
Y Dist.	Joint Dist.	Cens.	Bias	Coverage	Width	Bias	Coverage	Width
Gaussian	Parametric	Low	-0.002	0.94	0.11	-0.001	0.95	0.18
		High	-0.002	0.97	0.12	0.003	0.95	0.30
	Bimodal	Low	-0.01	0.94	0.13	0.11	0.60	0.64
		High	-0.01	0.94	0.14	0.16	0.40	0.77
Log-Normal	Parametric	Low	-0.02	0.92	0.13	-0.001	0.96	0.12
		High	0.004	0.96	0.14	-0.01	0.96	0.13
	Bimodal	Low	-0.004	0.98	0.11	0.02	0.94	0.18
		High	0.03	0.92	0.12	0.06	0.90	0.20

with flat priors, relative to the data variance. Importantly, we use a local conditional Gaussian model for Y. We set λ_0^* to an exponential (constant) hazard. Additional details on DR-SL and EDP-GP settings are provided in Appendix B.3. To summarize, the unimodal setting with Normally distribution Y is the most favorable setting for our method since the Gaussian data generating model matches the local Gaussian model we specify. In principle, all of these settings are quite favorable to the DR-SL method since we correctly specify the propensity score model. The log-Normal setting is the least favorable to our method since our local Gaussian model is misspecified. Notice that in all censoring and Y distribution settings, the parametric data generating process yields low bias and close to nominal coverage for both methods. This is as expected since both are highly flexible models, they should perform well in simple settings. Note however, that the models diverge in the more complicated, bimodal setting. In the bimodal log-Normal setting, the DR-SL exhibits higher bias with a larger interval width relative to EDP-GP. Similarly, in the bimodal Gaussian setting, the DR-SL model exhibits particularly high bias - 11% and 16% in the low and high settings, respectively. The main challenge with DR-SL is that the underlying super learner fails to capture biomodality in the cost-effectiveness joint distribution. In contrast, the EDP partitioning picks up the bimodality - modeling each mode with separate parameters to attain a better overall fit. Finally, note that EDP-GP intervals tend to be narrower across settings.

3.7. Cost-efficacy of Endometrial Cancer Treatment

We apply our BNP method to assess the cost-effectiveness of adjuvant chemotherapy (CT) versus radiation therapy (RT) for the treatment of endometrial cancer and compare our results to the DR-SL estimate. The target population of interest are women over the age of 65 who were diagnosed with endometrial cancer before undergoing hysterectomy. Within three months after hysterectomy, patients are assigned to either adjuvant RT or CT. We select a cohort of women over the age of 65 who were diagnosed with endometrial cancer between 2000 and 2014 in the SEER-Medicare database. The first treatment after three months of diagnosis was recorded. A maximum of $\tau = 24$ months of follow-up after hysterectomy was available in this data cut. Total costs accrued by Medicare (including inpatient, outpatient, hospice, and pharmaceutical costs) were recorded along with their survival/censoring status. Covariates which are known drivers of treatment assignment (age, comorbidities, cancer stage) were extracted. Table 3.2 displays summary statistics for the sample. Notably, the 2-year survival is slightly lower in the CT arm (93% vs. 94.5%), and average total costs higher in the CT arm (51.3 vs. 42.6). This suggests worse cost-effectiveness for CT relative to RT. However, there is significant uncertainty associated with these numbers that should be quantified. Moreover, the cohorts differ substantially in terms of observed characteristics at treatment assignment. For instance, the radiation arm has a greater proportion of patients with baseline International Federation of Gynecology and Obstetrics (FIGO) stage of IB - which is more severe than IA and I-NOS. Similarly, RT harm has fewer comorbidities - with 57% (vs. 54%) having Charlson Comorbidity Index of zero. These differences could differentially affect adjuvant therapy assignment and cost-efficacy.

We use our EDP-GP approach to compute posterior point and interval estimates for NMB while adjusting for differences in observed covariates. We specify the local cost distribution, $p(Y_i \mid T_i, \delta_i, X_i, \omega_i)$, to be a log-normal distribution with parameters $\omega_i = (\beta_i, \phi_i)$. The local regression is

$$E[Y_i \mid T_i, \delta_i, A_i, L_i, \omega_i] = \exp\left\{(1, L_i, A_i, T_i, \delta_i)'\beta_i + \phi_i/2\right\}$$

This local log-normal distribution respects the non-negative nature of costs, while allowing us to capture skewness. In the model, L_i includes household income, Charlson index, and FIGO. FIGO is included as a categorical covariates, while the others are treated as continuous. We let $A = 1$

Table 3.2: Sample Characteristics: Mean and sample standard deviations reported for continuous covariates. Counts and proportions reported for categorical covariates. Standardized mean differences (SMD) are provided. Typically $SMD > .1$ indicate large differences. Monetary amounts are in thousands of 2018 U.S. Dollars.

	Radiation (N= 3,827)	Chemotherapy (N= 245)	SMD
Total Accrued Costs ($)	42.6 (36.8)	51.3 (39.7)	.23
2-yr Survival Prob.	94.5	93.0	
Age (years)	73.6 (6.2)	73.2 (6.3)	.06
Household Income ($)	60.3 (28.8)	65.6 (34.0)	.17
Charlson Index			.12
0	2176 (56.9)	131 (53.5)	
1	1056 (27.6)	65 (26.5)	
2	342 (8.9)	30 (12.2)	
≥ 3	253 (6.6)	19 (7.8)	
FIGO Stage			.5
I-NOS	353 (9.2)	23 (9.4)	
IA	1162 (30.4)	128 (52.2)	
IB	1780 (46.5)	64 (26.1)	
II/II-NOS	532 (13.8)	30 (12.2)	

indicate assignment to chemotherapy with radiation being reference.

We set prior G_0 as discussed in Section 3.3.3: $G_{0\omega}(\beta_i, \phi_i) = N(\hat{\beta}, \hat{\Sigma})IG(a_0, \hat{\phi}(a_0 - 1))$. Here, $\hat{\beta}$ are OLS estimates using $\log(Y)$ as the outcome and $\hat{\Sigma} = diag(1, .01^2, \ldots, .01^2)$. Note the latter appears overly informative, but is actually fairly wide on the exponentiated scale. That is, a prior variance of 1 implies that mean costs as large as $exp(1 \cdot 1.96) \approx 7$ times the empirical mean cost are plausible. Similarly, a prior variance of $.01$ implies covariate effects of as large as $2\% = 1 - exp(1.96 * .01)$ are *a priori* plausible in the absence of data. For ϕ, note that the variance of the log-Normal random variable, Z, is $Var[Z] = (e^{\phi} - 1)E[Z]^2$, which implies $\phi = \log[Var[Z]/E[Z]^2 + 1]$. This motivates setting $\hat{\phi} = \log[\hat{s}^2/\bar{y}^2 + 1]$, where $\hat{s}^2$ and $\bar{y}$ are the marginal variance and mean of the observed cost values. We set $a_0 = 1000$, which anchors the prior around the empirical estimate. For the effectiveness model we again follow Section 3.3.3 and set $G_{0\theta}(\theta_i) = N(\hat{\theta}_{PH}, I)$. We center the GP priors around an empirical estimate $\lambda_0^*(t) = \hat{\lambda} \approx .001$ with $b = 2000$ and $\xi = 4000$. Here, ξ is on the order of the sample size - signifying strong AR(1) smoothing. The value b is about half of ξ - putting equal *a priori* weight on the prior hazard $\lambda_0^*(t)$ and the previous hazard at time $t - 1$.

We run three MCMC chains in parallel for 5,000 iterations and discard the first 3,000 draws of each chain as burn-in. We initialize each chain with different numbers of initial cost and effectiveness

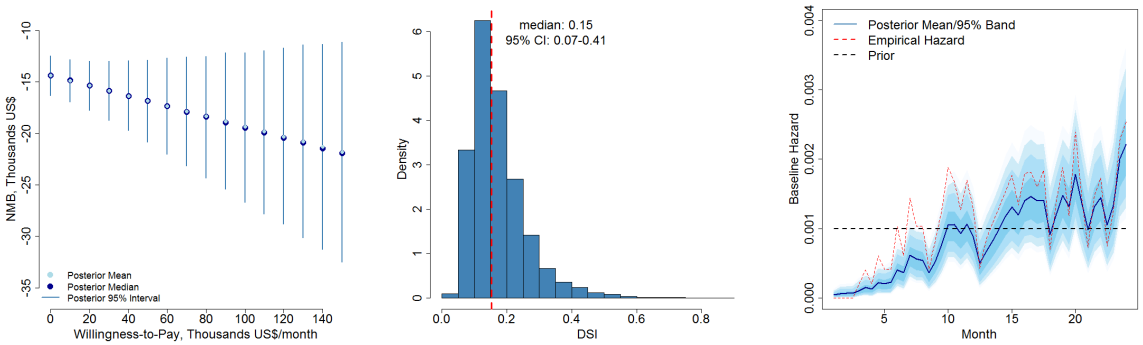


Figure 3.3: Posterior estimates of (left panel) NMB for various willingness-to-pay for each additional *month* of survival, κ. The posterior distribution of DSI in (middle panel) shows that about 15% of the variation in the individual-level NMBs is explained by the EDP induced clustering. This suggests the treatment effect may be relatively homogeneous and the NMB is a good overall average effect measure. The right panel plots the posterior baseline hazard curve along with 95%, 90%, and 80% credible bands in successively darker shades. Notice that posterior estimate is smoother version of the empirical estimate hazard in red. It is a posterior compromise between the empirical hazard and the prior constant hazard.

clusters and check that the chains converge to each other regardless of this initialization. This yields a total of 6,000 draws which we use for posterior inference. Other details and assessments of convergence are provided in Appendix B.4.

We estimate a 2-year NMB of chemotherapy over radiation to be $-\$14.5$ thousand, with 95% CI $[-\$16.6, -\$12.7]$. This assumes a willingness-to-pay of about $\kappa = \$4167$/month, or $\$50,000$/year of life gained - which is standard in cost-effectiveness analyses. This is roughly consistent with the unadjusted comparison in Table 3.2, where average total costs among chemotherapy patients was higher by about $\$9,000$. The left panel of Figure 3.3 shows average NMB as a function of κ for various κ values. Recall that by definition NMB is a linear function of κ. The intercept at $\kappa = 0$ shows an NMB that captures differences in cost only (efficacy has zero value). The negative y-intercept here reflects that even if we do not value efficacy, chemotherapy is more expensive than radiation after covariate adjustment. The negative slope of the curve reflects that adjusted efficacy (i.e. survival benefit) of chemotherapy is lower. However, the slope here is quite small, suggesting a very small difference in efficacy. This is consistent with unadjusted results - recall from Table 3.2 that 2-year survival is slightly lower among chemotherapy patients.

In terms of clustering, we compute $c^*_{1:n}$ as given in Section 3.5 and find that about 86% of the observations are grouped into two posterior mode clusters. However, in the middle panel of Figure

3.3 we see that only about 15% of the variation in the individual-level NMBs is explained by the EDP-induced partition - which suggests these clusters are not very meaningful for cost-effectiveness. This indicates low posterior evidence of treatment effect heterogeneity, suggesting average NMB may fairly characterize the cost-effectiveness profile. Finally, the right panel of Figure 3.3 shows the posterior estimate of the baseline hazard. Since continuous covariates were normalized, this represents the hazard among patients with average household income and age with Charlson index of zero and FIGO II/II-NOS. This has no explicit causal interpretation but is illustrative of the Gamma process. Notice our posterior has moved away from the constant hazard prior and towards the empirical (Nelson-Aalen) estimate shown in red. The informative AR(1) shrinkage results in a smoother posterior curve that penalizes large swings in the empirical hazard.

For comparison, we also ran the DR-SL approach where propensity score model, mean survival time model, and mean cost model were all estimated using super learner. Regression trees, GLM-net and GLM were included as candidate learners and 95% BCa intervals were estimated using 5,000 bootstrap iterations. For willingness-to-pay $\kappa = \$4167$/month, DR-SL estimates a 2-year average NMB of $-\$11.8$ with 95% CI $[-\$19.1, -\$6.0]$ in thousands. This is similar to our estimate of $-\$14.5$ $[-\$16.6, -\$12.7]$, but the DR interval is wider. More details on the DR-SL implementation are given in Appendix B.4, including a full plot of average NMB from DR-SL as in Figure 3.3. For even large willingness-to-pay values of up to 300 thousand USD per year, both approaches find a negative NMB with intervals excluding zero. This supports the relative cost-effectiveness of radiation over chemotherapy adjuvant therapy over two years.

3.8. Discussion

Cost-effectiveness is statistically challenging due to the complexities of the joint distribution of cost and survival time, such as skewness, censoring, and multi-modalities. Moreover, estimation of policy-relevant estimands with causal interpretation is complicated by confounding in observational studies. Robust causal inference for cost-effectiveness requires flexible modeling that accounts for these complexities while adjusting for confounders. In this paper, we outlined a nonparametric Bayesian solution that leverages the Gamma and enriched Dirichlet process priors to model the joint distribution of cost and survival time. We proposed cost-effectiveness estimands with causal meaning and identified them under suitable causal assumptions. We showed how our model can

be used in a Bayesian g-computation procedure that draws from the posterior of the causal effect. Finally, we show that the partition induced by the EDP can be used to explore cost-effectiveness heterogeneity and introduced the DSI diagnostic statistic for assessing how well this partition captures heterogeneity.

In simulations, we demonstrated that our procedure has adequate frequentist properties (bias, coverage, etc.) in a variety of scenarios. In complex settings, it can be comparable and, at times, outperform existing doubly-robust methods. Across almost all settings, the EDP-GP produces NMB estimates with narrower interval widths relative to the DR-SL estimates. In the data analysis, the DR-SL approach also yields wider intervals. One driver of this is the relative inefficiency of the DR-SL approach. This method only uses data from patients who are not censored and weights their contributions by the inverse probability of being uncensored. In contrast, our method uses censored patients, since they still inform the total cost distribution at their observed time. Another feature with the DR-SL is that it is a weight-based estimator (weighted both by inverse probability of treatment and censoring), which are known to be quite variable if the probability of treatment are near the bounds within subgroups. Since the EDP-GP approach is model-based, it provides more smoothing under these conditions. Finally, the bootstrap inference procedure used in the DR-SL approach can be difficult to implement in practice, where sparsity among categorical covariates leads to the occasional pathological bootstrap resample (e.g. with rank deficient matrix). This is in contrast to full posterior inference via the Bayesian bootstrap which can be more stable.

Finally, we see at least two avenues of future work and extensions. First, in our paper, we consider a setting with a single baseline treatment. This allows us to estimate the cost-effectiveness of baseline treatments, which are highly relevant in many settings. However, we may also wish to estimate the cost-effectiveness of time-varying treatment *regimes*, in addition to the effect of the initial baseline treatment. Flexible causal estimation in these settings is more complex and should be explored. Second, there has been much work on improving the computational scalability of posterior inference on Dirichlet process models, including both approximate inference via Variational Bayes and parallel MCMC procedures. Future work developing scalable inferential procedures for joint-modeling with EDPs can be useful.

CHAPTER 4

THE HIERARCHICAL BAYESIAN BOOTSTRAP FOR HETEROGENOUS TREATMENT EFFECT ESTIMATION

4.1. Introduction

Estimation of heterogeneous treatment effects - i.e., effects within strata of some other relevant variable - is popular in causal inference. These estimands are especially relevant in scenarios where treatment effects vary substantially in the population. In these scenarios, an overall estimate averaged across strata may suggest a negligible treatment effect even if there is substantial benefit within a particular stratum. Such effects can be identified under rather standard causal assumptions (no unmeasured confounding, positivity, etc.) and computed using standardization in the point-treatment setting. Within each stratum, standardization involves averaging a stratum-specific regression model adjusting for confounders and treatment over the distribution of confounders within that stratum. Fully Bayesian approaches to standardization, and causal estimation broadly, have been growing in popularity. For instance, BART regression models were used in early work by Hill (2011) to compute marginal effects and subsequently by Zeldow, Lo Re III, and Roy (2019), Hahn, Murray, and Carvalho (2020), and Henderson et al., 2018 to compute individual treatment effects. Other Bayesian nonparametric (BNP) priors such as Dirichlet process (DP) and variations such as the enriched DP and dependent DP regressions have also been used to do full posterior inference on marginal treatment effects. For instance, such methods have been developed for computing effects under zero-inflation (Oganisian, Mitra, and Roy, 2020), in the presence of missingness, (Roy et al., 2018), in mediation scenarios (Kim et al., 2017), for censored survival outcomes under competing risks (Xu et al., 2020), and to compute causal quantile effects (Xu, Daniels, and Winterstein, 2018).

To perform standardization, regression models must be averaged over the confounder distribution of the target population. For instance, Hill averages the BART over the empirical distribution when computing marginal effects. This is a flexible approach as it makes no modeling assumption about the distribution. However, it is unsatisfying from a Bayesian point of view since it uses a fixed plug-in estimate and variability of this estimate does not flow through to the posterior of the causal

effects. To address this, Wang et al. (2015) and Nethery, Mealli, and Dominici (2019) used the Rubin's Bayesian bootstrap (BB) (Rubin, 1981). Broadly, this approach models the confounder distribution as a point-mass distribution with unknown mass/weight at each observed confounder value. Posterior inference is done on these unknown weights and variability propagates through to the causal effects of interest.

The popularity of the BB for marginal estimation has lead to its adoption for heterogenous average treatment effect (HTE) estimation. These are average treatment effects within strata of some other variable. For instance, Roy, Lum, and Daniels (2017) evaluate the effect of antiretroviral therapy on various outcomes among HIV-positive patients with and without recent alcohol use. They use independent BBs to estimate the confounder distributions of recent alcohol users and non-users separately. Taddy et al. (2016) are concerned with estimating effects of a large A/B testing experiment among "new" and "old" platform users. Again, these use separate BB estimates within these two strata. More recent work by Boatman, Vock, and Koopmeiners (2020) attempts to do causal estimation in a setting where data are collected from several "supplemental sources" in addition to a "primary" data source. They then estimate a causal effect within the "primary" stratum by averaging a BART regression over a BB estimate of the confounder distribution in the primary source, separate from the other sources.

Though common, using separate BBs for HTE estimation is not ideal - especially when some strata are sparse. For instance, in our motivating data analysis we target the marginal effect of proton versus photon chemoradiotherapy on adverse event risks. The question of interest is how the effect varies across different cancer types for which chemoradiotherapy is the standard-of-care. This is complicated as some cancer types (e.g. lung) may be rare in the sample, giving us little data on the confounder distribution within these strata. By construction, the BB places zero mass on confounder values unseen within this stratum - even if this is due to small samples and not due to an *a priori* belief that unseen values are impossible. While plausible covariate values for lung cancer patients may have been observed for, say, brain cancer patients, stratum-specific BBs have no way of borrowing this information. Our main contribution is the construction of a hierarchical Bayesian bootstrap (HBB) prior for estimating stratum-specific confounder distributions in precisely such a setting. Based on the Hierarchical Dirichlet Process (HDP), our approach allows for a principled borrowing of confounder information across strata. For large strata, the HBB posterior shrinks to the

stratum-specific BB. For small strata, it is shrunk more heavily towards values seen in other strata. This approach (1) maintains the flexibility of the BB (we make no parametric assumptions about the confounder distributions), (2) provides room for efficiency gains via the induced shrinkage, and (3) is fully conjugate and agnostic to the choice of outcome model. This last property makes it compatible for several of the popular outcome modeling approached mentioned earlier.

Several notable modifications to the bootstrap have been proposed which are distinct from our work. For instance, Makela, Si, and Gelman (2018) developed a two-stage Bayesian bootstrap for a cluster-randomized study setting. Here, clusters/strata are sampled and then individuals are sampled within a cluster. A key problem here is how to account for strata that exist, but are never sampled. This is distinct from our problem where strata are known and fixed and the issue is to borrow information across them. Approaches such as "bag-of-little bootstraps" (Barrientos and Pea, 2020; Kleiner et al., 2014) have been proposed with the goal of scaling bootstrap to large datasets. The idea run separate bootstraps on sub-samples, then combine in such a way as to approximate the overall bootstrap distribution. However, we are not concerned with estimating the overall data distribution, but stratum-specific distributions. Finally, several "smoothed" bootstraps have been developed (Efron and Gong, 1983; Silverman and Young, 1987; Wang, 1995). The view here is that the Efron's bootstrap is sampling from the empirical distribution that places uniform mass on each observed data value. This point-mass distribution is convoluted with a kernel to induce smoothness. While indeed a step in the right direction, it is unsatisfactory from the perspective of HTE estimation. We could, for instance, estimate stratum-specific smoothed bootstrap distributions. This will indeed place some mass on the unseen values, but this mass is allocated via an ad-hoc kernel, rather than informed by data in the other strata. Specification of this kernel is also a hurdle, which BB does not face. As we will see, however, we can provide a probabilistic motivation for the smoothed bootstrap as an improper case of the HBB.

In the next section, we introduce some notation and motivate the causal problem more precisely before outlining the HBB and related computation. After, we will discuss simulation studies assessing the performance of the HBB relative to dominant approaches in the causal literature under a variety of settings. We end with an analysis contrasting the risk of adverse events for proton versus photon therapies across various cancer types.

4.2. Background and Motivation

Suppose we observe outcome Y for subjects assigned to treatment $A \in \{0, 1\}$ along with some confounders $L = (W, V)$ that are measured pre-treatment. These are variables which we believe to be drivers of both treatment and outcome. In the HTE setting, this set often consists of V - a discrete variable taking on values $v \in \{1, 2, \ldots, K\}$ along which we wish to make causal comparisons - and variables W which we would like to average over. The levels of V are sometimes constructed using multiple covariates - e.g. V marking all combinations of discrete V_1 and V_2. Using potential outcomes notation (Rubin, 1978), one popular causal estimand is the heterogeneous, or stratum-specific, average treatment effect (HTE) $\Psi(v) = E[Y^1 - Y^0 \mid V = v]$ - the average difference in outcomes had everyone in the stratum $V = v$ taken treatment 1 versus 0. Note this is distinct from the target of *individualized* treatment effect (ITE) estimation, which are individual-level effects.

While we could estimate $E[Y \mid A = a, V]$ with observed data, in general $E[Y \mid A = a, V] \neq E[Y^a \mid V]$. That is, the average outcome among subjects treated with $A = a$ in V may not be the same as the average outcome had *everyone* in V taken treatment $A = a$. This is due to confounding: treated subjects may be a non-representative subset of the patients in stratum V (e.g. systematically sicker and, therefore, more likely to have worse outcomes). Under well-known causal identification assumptions, we can estimate $\Psi(v)$ by integrating the difference in stratum-specific outcome regressions over the conditional distribution of W (see Web Appendix A in Supporting Information)

$$\Psi(v) = \int_{\mathcal{W}} \Big\{ E[Y \mid A = 1, V = v, W] - E[Y \mid A = 0, V = v, W] \Big\} dP_v(W) \tag{4.1}$$

where $P_v(W) = P(W \mid V = v)$. This formula is known as standardization - a special case of the g-formula (Robins, 1986) in the point-treatment setting. The same general approach can be used to compute an overall average treatment effect (ATE) $\Psi = E[Y^1 - Y^0]$ by integrated the outcome regression over the joint $P(L) = P(W, V)$. The estimand $\Psi(v)$ is more relevant than Ψ in settings where, due to variability within the population, the ATE is not a meaningful measure of the treatment effect.

Suppose we observe n independent subjects with data, $D = \{Y_i, A_i, W_i, V_i\}_{1:n}$. Let $S_v = \{i : V_i = v\}$ contain the indices of subjects in stratum $V = v$ and let n_v denote the cardinality of S_v such that

$n = \sum_v n_v$. Bayesian inference typically proceeds by obtaining a posterior over $E[Y \mid A, V, W]$ and $P_v(W)$ which together induce a posterior over the target $\Psi(v)$. As discussed in the introduction, many BNP models exist for the former. Efficient estimation of the latter via the HBB is the chief objective of this paper, but first we review some popular alternatives. One approach is to plug in the empirical distribution $\hat{P}_v(w) = \frac{1}{n_v} \sum_{i \in S_v} \delta_{W_i}(w)$ - where $\delta_x(\cdot)$ denotes the degenerate distribution at x. For compactness we sometimes denote these as simply P_v and δ_x. This places uniform mass of $1/n_v$ on each confounder vectors observed in stratum v.

To our knowledge, Wang et al. (2015) first proposed using Rubin's Bayesian bootstrap (BB) (Rubin, 1981) over this empirical approach and it has since become popular as it accounts for variability in the empirical estimate (Nethery, Mealli, and Dominici, 2019; Saarela et al., 2015; Xu, Daniels, and Winterstein, 2018). To summarize the BB, it models the covariate distribution as $P_v(w) = \sum_{i \in S_v} \pi_i^v \delta_{W_i}(w)$, but unlike the empirical approach the weights, $\pi^v = \{\pi_i^v\}_{i \in S_v}$, are considered unknown parameters that completely determine P_v. A prior over these these weights is then a prior over P_v. Noting that the weight vector lives in the simplex, $\pi^v \in \{\mathbb{R}^{n_v} : \pi_i^v > 0 \; \forall i \in S_v, \; \sum_{i \in S_v} \pi_i^v = 1\}$, BB places an improper Dirichlet prior over this space $\pi^v \sim Dir(0_{n_v})$, where 0_{n_v} is the n_v-dimensional zero vector. This is a conjugate model with posterior $\pi^v \mid \{W_i\}_{i \in S_v} \sim Dir(1_{n_v})$, where 1_{n_v} is the n_v-dimensional vector of ones. Note that this is done for each $V = v$, *separately*. This is the approach used for HTE estimation in the Bayesian causal inference literature by Boatman, Vock, and Koopmeiners (2020), Roy, Lum, and Daniels (2017), and Taddy et al. (2016). This common approach does have several advantages. First, it retains the flexibility of the empirical distribution. Note that the posterior expectation of each π_i^v is $1/n_v$. Second, unlike the empirical estimate, variability in this estimate flows through to the posterior of $\Psi(v)$ since the weights are not fixed at $1/n_v$. Third, it is computationally easy to sample due to conjugacy and, fourth, it is agnostic to the choice of outcome model. However, it becomes problematic for sparse strata where few values of W are observed. Under the BB, P_v assigns zero probability to values of W that are unseen in stratum v. This is undesirable because there are many values that we may think are *a priori* plausible. Indeed, we may observe such values in other strata. Since the BB estimates of P_v are done independently, the posterior estimate of P_v cannot borrow this information - yielding less stable estimates of $\Psi(v)$. In essence, the proposed HBB retains these desirable properties of the BB while addressing the small-strata shortcomings by "partially pooling" the estimates of P_v.

4.3. The Hierarchical Bayesian Bootstrap

Let $W^v = \{W_i\}_{i \in S_v}$ denote the observed confounders in stratum v and let $W = \{W^v\}_{v=1:K}$ denote the full set of confounders. We model W^v as following an unknown distribution $W^v \mid P_v \sim P_v$ and propose a prior for P_v that borrows information across V. The DP is a stochastic process that generates random, discrete distributions. Due to its flexibility and conjugacy, it has become a popular prior for unknown distributions in Bayesian analysis. Suppose we place a DP prior on each P_v, denoted $P_v \sim DP(\alpha P_{0v})$. The realizations of P_v are centered around a "mean" distribution of P_{0v}, with $\alpha > 0$ controlling the dispersion of these realizations around P_{0v}. This is flexible because the posterior of P_v under a DP is a compromise between the prior mean, P_{0v}, and the empirical distribution in stratum $V = v$, $\sum_{i \in S_v} \delta_{W_i}(w)$, with relative weight controlled by α. However, each P_v is centered around its own P_{0v}, preventing any borrowing of information across strata. This motivates the hierarchical DP (HDP) of Teh et al. (2006), which centers the P_v around a *common* mean distribution P_0 and adds a DP hyperprior on P_0. We note that while the following development may seem rather involved, the actual posterior computation will be fully conjugate and efficient. Under the HDP prior, the full model for the covariates is

$$\begin{aligned} W_i \mid P_v &\sim P_v \ \text{ for } i \in S_v \\ P_v \mid \alpha, P_0 &\sim DP(\alpha P_0) \ \text{ for } v = 1, \ldots, K \\ P_0 \mid \gamma, P_* &\sim DP(\gamma P_*) \end{aligned} \tag{4.2}$$

The DP hyperprior on P_0 implies that the random P_0 are discrete - allocating mass to atoms. Due to this discreteness, the distributions P_v have support on the same atoms as P_0 but allocate mass differently across these atoms in a way that is local to V. Since the DP is conjugate, the posterior of P_v conditional on P_0 is another DP: $P_v \mid P_0, \alpha, W^v \sim DP(\alpha P_0 + \sum_{i \in S_v} \delta_{W_i})$. Similarly the marginal posterior of P_0 is also a DP: $P_0 \mid W \sim DP(\gamma P_* + \sum_{i=1}^{n} \delta_{W_i})$. For the Hierarchical BB, we set $\gamma = 0$ in (4.2) and denote this prior on P_v as $P_v \mid \alpha \sim HBB(\alpha)$. This yields posterior under the $HBB(\alpha)$

$$\begin{aligned} P_v \mid P_0, \alpha, W^v &\sim DP(\alpha P_0 + \sum_{i \in S_v} \delta_{W_i}) \\ P_0 \mid W &\sim DP(\sum_{i=1}^{n} \delta_{W_i}) \end{aligned} \tag{4.3}$$

With $\gamma = 0$, P_0 are random distributions centered around the empirical distribution $P_0 \mid W \sim DP(\sum_{i=1}^{n} \delta_{W_i})$ This distribution is discrete with an atom at each of the n observed W_i. A P_0 can be drawn from this posterior by drawing a vector of weights $\pi_{1:n} \sim Dir(1_n)$, where $\pi_{1:n} = (\pi_1, \pi_2, \ldots, \pi_n)$. This draw of P_0 can then be represented as $P_0 = \sum_{i=1}^{n} \pi_i \delta_{W_i}$. Note that this is exactly the BB. However, now we have an additional layer of uncertainty as the stratum-specific distributions must be drawn around this P_0: $P_v \mid P_0, \alpha, W^v \sim DP(\alpha(\sum_{i=1}^{n} \pi_i \delta_{W_i}) + \sum_{i \in S_v} \delta_{W_i})$.

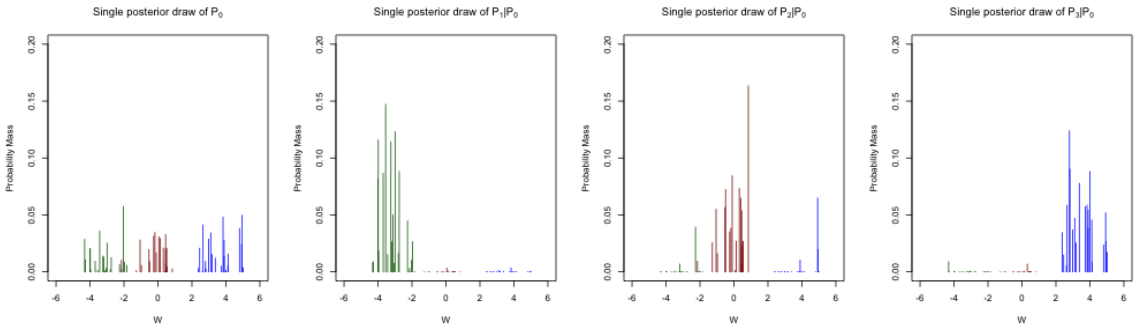


Figure 4.1: Draw from posterior of P_v under prior $P_v \sim HBB(2)$ with simulated scalar W_i for $n = 90$ subjects from $V = 1, 2, 3$. These 90 atoms are represented by vertical bars with colors indicating stratum of the atom. The height of the lines represent probability mass drawn from the HBB posterior. Left panel: a draw of P_0 - recall this is centered around the empirical distribution (i.e. line 2 in (4.3)). The next panel shows a draw from the Dirichlet Process posterior of P_v conditional on this draw of P_0 - i.e. line one of (4.3). Note that P_1, P_2, and P_3 place positive mass on *all* observed atoms. For instance, independent BB estimates of P_2 would put place 0 mass on all atoms but the red - unlike the third panel.

Again, conditional on a draw of P_0, each P_v is a discrete distribution with atoms at each of the observed n points in the *entire* sample. Combining like terms in the summations, however, we see that atoms observed in stratum $V = v$ have a weight of $\alpha\pi_i + 1$ - higher than the weight on atoms unseen in stratum $V = v$, which is $\alpha\pi_i$. To see this, note that in expectation (over many draws of P_v), the posterior distribution of W^v can be represented as a Pólya Urn (Blackwell and MacQueen, 1973):

$$\begin{aligned} P_v(W = w \mid P_0, \alpha, W^v) &= \frac{\alpha}{\alpha + n_v}(\sum_{i=1}^{n} \pi_i \delta_{W_i}) + \frac{1}{\alpha + n_v} \sum_{i \in S_v} \delta_{W_i} \\ &= \frac{1}{\alpha + n_v} \Big\{ \sum_{i \notin S_v} \alpha\pi_i \delta_{W_i} + \sum_{i \in S_v} (1 + \alpha\pi_i) \delta_{W_i} \Big\} \end{aligned} \tag{4.4}$$

Again due to the finitely many atoms, we can draw a P_v from this posterior by drawing from an n-dimensional Dirichlet distribution with the i^{th} concentration parameter being $\alpha\pi_i$ for $i \notin S_v$ and

$1 + \alpha\pi_i$ for $i \in S_v$. Intuitively, this can be seen as adding an additional α subjects from the marginal distribution into stratum V. These "pseudo-subjects" can take on any observed value in the marginal, even if they are unobserved in the stratum - thus, borrowing information. As with the posterior update for P_0, a draw from this Dirichlet distribution yields an n-dimensional set of weights $\pi^v_{1:n}$ and thus a draw of P_v is given by $P_v(w) = \sum_{i=1}^n \pi_i^v \delta_{W_i}$. Note that in the above we used a common α across strata. This is without loss of generality, as each stratum can have its own α_v without changing the results. We will turn to specification of these hyperparameters after discussing computation.

4.3.1. Posterior Computation via MCMC

Here we describe posterior HTE inference under a $HBB(\alpha)$ prior for P_v via Markov Chain Monte Carlo (MCMC). At each iterations $m = 1, \ldots, M$, we

1. Obtain a posterior draw of P_0 by drawing weights $\pi^{(m)}_{1:n} \sim Dir(1_n)$ then forming $P_0^{(m)}(w) = \sum_{i=1}^n \pi_i^{(m)} \delta_{W_i}(w)$.

2. For each $v = 1, \ldots, K$, obtain a posterior draw, $P_v^{(m)}$, conditional on $P_0^{(m)}$. We do this by drawing $\pi^{v(m)}_{1:n} \sim Dir(\eta_n^{(m)})$, where $\eta_n^{(m)}$ is the n-dimensional vector with element i being $\alpha\pi_i^{(m)}$ if $i \notin S_v$ and $(1 + \alpha\pi_i^{(m)})$ if $i \in S_v$. Note the sum of the elements in $\eta_n^{(m)}$ is $\alpha + n_v$. This now forms a draw of $P_v^{(m)}(w) = \sum_{i=1}^n \pi_i^{v(m)} \delta_{W_i}(w)$.

Now to estimate the HTEs, suppose we also have M posterior draws of the regression $E[Y \mid A, W, V]$, denoted by $\mu^{(m)}(A, W, V)$. This can be from any model. For instance, in a GLM this could be $\mu^{(m)}(A, W, V) = g^{-1}(\beta_0^{(m)} + W'\beta_w^{(m)} + V'\beta_v^{(m)} + \beta_A^{(m)} A)$ where g^{-1} is the inverse link function. This could also be a posterior draw $\mu^{(m)}(A, W, V) = f^{(m)}(A, W, V)$ where $f^{(m)}$ is the posterior draw of a sum-of-trees model under a $f \sim BART$ prior. To estimate the HTE, we include a third step

3. Integrate over HBB draw of P_v from Step 2, $P_v^{(m)}$.

$$\begin{aligned} \Psi^{(m)}(v) &= \int_{\mathcal{W}} \left\{ \mu^{(m)}(1, W_i, v) - \mu^{(m)}(0, W_i, v) \right\} dP_v^{(m)}(W) \\ &= \sum_{i=1}^n \pi_i^{v(m)} \left\{ \mu^{(m)}(1, W_i, v) - \mu^{(m)}(0, W_i, v) \right\} \end{aligned} \tag{4.5}$$

Repeating this procedure for each of the draws yields a set of M draws from the posterior of $\Psi(v)$, $\{\Psi^{(m)}(v)\}_{1:M}$, for each stratum $v = 1, \ldots, K$. Note that the W_i from all subjects contribute to $\Psi^{(m)}(v)$. However, values from the stratum and values outside the stratum are weighted differently according to $\pi_i^{v(m)}$.

4.3.2. Some Limiting Cases and Hyperparameter Choice

Here we consider the limiting behavior of the HBB by analyzing (4.4) conditional on $P_0(w) = \sum_{i=1}^{n} \pi_i \delta_{W_i}$ and the choice of hyperparameter. Note that for $\alpha = 0$, the first term in line one of (4.4) disappears and our estimate reduces to $P_v(W = w \mid P_0, \alpha, W^v) = \frac{1}{n_v} \sum_{i \in S_v} \delta_{W_i}$. This is the empirical distribution within stratum v and represents a *completely unpooled* estimate where values of W unseen in stratum v have no mass. Thus, there is no borrowing of information. This is also the posterior mean of the BB estimate of P_v. Now consider the other extreme where $\alpha >> n_v$. In this case (4.4) reduces to $P_v(w \mid P_0, \alpha, D) = \sum_{i=1}^{n} \pi_i \delta_{W_i}$ - the BB estimate of the entire marginal distribution (over V) that places expected mass $E[\pi_i] = 1/n$ on each observed value of W in the entire sample. That, is we have *completely pooled* all the stratum-specific distributions. The parameter α controls the posterior compromise between these extremes for a particular stratum. The idea of partial-pooling is to balance the bias-variance tradeoff, with fully pooled estimates favoring reduction in variance over potential increase in bias and fully unpooled estimates favoring a reduction of bias over potential increase in variance. Of course, partial-pooling by its nature may induce bias, especially if the confounder distributions in the sub-populations are very different. While it may be tempting to view the introduction of a user-specified parameter α as a limitation, we have just shown above that the dominant BB approach already makes a very strong prior choice of $\alpha = 0$ - implicitly favoring the completely unpooled scenario, even if some partial pooling to reduce variability is sensible. The explicit introduction of α is more sensible as we do not lock users into this strong prior.

Hyperparameter guidance:

To guide decisions about α, recall that we can interpret $\alpha > 0$ as adding an additional α pseudo-subjects from the marginal distribution of W to the n_v subjects in stratum v. Higher α places more weight on the pseudo-subjects - who may have values unseen in stratum $V = v$ (i.e. more shrinkage towards the marginal). The relative mass on a point seen within the stratum relative

to an unseen point is approximately $\rho = \frac{1+\alpha/n}{\alpha/n} = \frac{n}{\alpha} + 1$. This is seen in (4.4) when substituting π_i with its posterior expectation of $1/n$. For example, if we add $\alpha = n$ pseudo-subjects, then on average the atoms seen in stratum v are about as likely as the atoms not seen in stratum v. This is fairly aggressive shrinkage. For some $M > 0$, one option is to set $\alpha = \frac{n \cdot M}{n_v}$ which implies a relative weight of $\rho = \frac{n_v}{M} + 1$. We should now subscript this parameter as α_v as it is stratum-specific depending on n_v - but we omit this notation where there is no ambiguity. Here, M is user-specified and can be interpreted as the minimum desired sample size in each stratum. This may partially be set depending on the number of confounders we are integrating over and the complexity of the joint distribution. For instance, with well-behaved, standard joint distribution (e.g. multivariate Gaussian), $M = 30$ subjects within a stratum may be sufficient to estimate the distribution. On the other hands, if the covariates are complex, skewed, and multimodal we may need a larger M to obtain a good nonparametric estimate such a distribution. Note that strata with size $n_v << M$ implies $\rho \approx 1$ which corresponds to heavy shrinkage. Conversely, for large strata with $n_v >> M$, ρ gets larger - placing increasingly more weight on atoms within stratum v only. This reduces shrinkage proportional to n_v. Figure 4.2 depicts draws from the posterior of P_v under a prior $P_v \sim HBB(nM/n_v)$ with synthetic data. Note that strata that are more sparse (relative to M) have distribution draws that are more heavily shrunk towards the marginal. However, we place positive mass on all points observed in the sample. While in some ways it could be more satisfying to specify a prior over α, this would

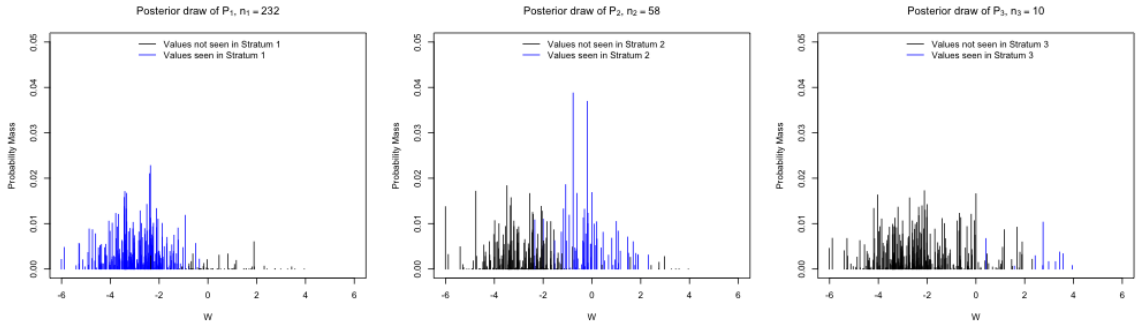


Figure 4.2: Draw from posterior of P_v under prior $P_v \sim HBB(nM/n_v)$ with $n = 300$ scalar confounders simulated for $v = 1, 2, 3$ strata. Here we set $M = 30$. Note that for stratum $V = 1$ we have far greater observations than M and so the draw of P_1 places most mass on atoms seen in this stratum. Stratum 2 has size slightly larger than M and so places $\rho = 58/30 + 1 \approx 3$ times more weight on atoms seen in the stratum. Stratum 3 only has 10 subjects, and so places $\rho = 10/30 + 1 \approx 1$ equal weight on all atoms. This last case represents heaviest shrinkage.

complicate the posterior computation with a non-conjugate update. The existing BB's popularity

due in large part to its conjugate Dirichlet updates and keeping α user-specified maintains this important property.

The smoothed bootstrap as a limiting case:

The smoothed bootstrap has been proposed as one way of placing mass on unseen values of W. In this section, we briefly show how this is a limiting case of the $HBB(\alpha)$ prior on a mixing distribution when $\alpha \to 0$. The smooth bootstrap estimate of P_v is given by $\hat{P}_v(w) = \frac{1}{n_v}\sum_{i \in S_v} K_h\Big(\frac{w-W_i}{h}\Big)$. Smoothness is induced by convoluting a user-specified symmetric kernel, K_h, with the empirical distribution and the parameter h controlling smoothness. For concreteness, suppose the kernel is chosen to be standard Normal $K_h(\frac{w-W_i}{h}) = N(\frac{w-W_i}{h}; 0, 1)$. Then this bootstrap model is a mixture of n_v kernels centered around each observed W_i with variance h^2. The mixing distribution is the empirical distribution giving weight $1/n_v$ to each mixture component. Now consider a Bayesian mixture model with *unknown* mixing distribution P_v, written as $P(w \mid P_v) = \int_{\mathcal{W}} K_h\big(\frac{w-W}{h}\big) dP_v(W)$. Here, W are random with distribution P_v and w is a particular value. With an $HBB(\alpha)$ prior on the mixing distribution, recall that the mean of P_v is given via the Pólya Urn in (4.4). Plugging this urn expression in for P_v yields

$$P(w) = \int_{\mathcal{W}} K_h(\frac{w-W}{h})\Big\{\frac{\alpha}{\alpha+n_v}P_0(W) + \frac{1}{\alpha+n_v}\sum_{i \in S_v}\delta_{W_i}(W)\Big\}$$

In the improper limit as $\alpha \to 0$, the left term in the Pólya Urn goes to 0. Distributing the kernel and noting that the integral over $\mathcal{W}$ is non-zero only at each observed W_i, we get $P(w) = \frac{1}{n_v}\sum_{i \in S_v} K_h\Big(\frac{w-W_i}{h}\Big)$. This is exactly the smoothed bootstrap estimate $\hat{P}_v$. Thus, we have a probabilistic motivation for the smoothed bootstrap via the HBB, formally linking our work with this previous result. Later in the discussion we elaborate on how this link may motivate future work on a "smoothed" HBB.

4.4. Simulation Experiments

In this section we assess the behavior of the HBB relative to other approaches under a variety of settings via simulation. In all settings, we simulate 1000 datasets with $n = 300$ observations from $K = 4$ strata of varying sparsity. On average, the strata counts are $n_1 = 120$, $n_2 = 90$, $n_3 = 60$, $n_4 = 30$. Thus, stratum 4 is the most sparse stratum and stratum 1 is the least sparse. In each

simulated data set, we simulate a vector, W, of 10 confounders for each subject conditional on stratum V. The treatment indicator A itself is simulated as a function of stratum membership and confounders. We simulate a binary outcome model conditional on V, W, and A from a logistic model. In the true outcome model, each stratum has a different (conditional) treatment effect, leading to true HTEs that vary across strata. These represent challenging scenarios with several confounders and small samples that are often encountered in applied work.

For each simulated dataset, we use a correctly specified Bayesian logistic regression with wide, null-centered Gaussian priors. This is to focus attention on the confounder distribution models. After posterior sampling for the regression, we compute a causal risk difference, $\Psi(v) = E[Y^1 \mid V = v] - E[Y^0 \mid V = v]$, by integrating the regression over the confounder distribution model under both treatment interventions and taking the difference. We integrate over four confounder distribution models: the empirical distribution, the stratum-specific BB estimate, the HBB, and the oracle. By "oracle" we mean a Monte Carlo integration over draws from the true stratum-specific confounder distribution. For the HBB, we set $P_v \sim HBB(nM/n_v)$ with $M = 100$ in all settings. We assess the bias, variance, coverage, and precision of posterior estimates for $\Psi(1)$ (the effect in the least sparse stratum) and $\Psi(4)$ (the effect in the most sparse stratum) across simulation results in Table 4.1. Additional details about the simulation study can be found in Web Appendix C in Supporting Information.

In the first setting, we consider a relatively simple multivariate Gaussian generating distribution for W, which does not vary across V. In this "homogeneous Gaussian" setting, we see little difference in performance among the 4 methods in the least sparse stratum ($V = 1$). This is desirable as we would want the HBB to perform similar to other methods in such populous stratum. In the sparse stratum ($V = 4$), the HBB has slightly lower bias with lower MSE (equal up to three decimal places). Notably, the HBB borrows information across strata to yield, on average, smaller interval lengths than the BB (.46 v .478) while maintaining nominal coverage of around 95%. Note that the BB produces a wider interval than the empirical distribution as well (.478 v .459) this is because the BB accounts for uncertainty in the confounder distribution estimate.

The second setting considers a more difficult scenario where W is marginally generated from a location mixture of Gaussians. Each W is generated from a 10-dimensional multivariate normal but with different mean for each stratum. Thus, borrowing information from different strata is expected

Table 4.1: Simulation results: MSE, absolute bias, empirical variance of the posterior mean along with the width and coverage of the 95% credible interval across 1,000 simulation runs. MSE is computed as average of the squared difference between posterior mean and truth across simulations. Empirical variance is computed as the variance of the 1000 posterior means. In general, the HBB trades off bias for gains in efficiency, leading to overall reduction in MSE for sparse strata. Performance is generally similar to BB in more populous strata. The performance is particularly good in the complicated Gamma mixture setting, where stratum 4 has too few observations from the tail of the Gamma-distributed W to estimate $P_4(W)$ reliably via BB. The HBB, however, is able to borrow tail values observed in the other strata.

	Model	MSE	Bias	Variance	Interval Width	Coverage
Setting 1: Homogeneous Gaussian						
Stratum 1	Empirical	0.005	0.007	0.005	0.256	0.930
	BB	0.005	0.007	0.005	0.260	0.930
	HBB	0.005	0.007	0.005	0.258	0.933
	Oracle	0.005	0.008	0.005	0.255	0.928
Stratum 4	Empirical	0.013	0.002	0.013	0.459	0.944
	BB	0.013	0.002	0.013	0.478	0.951
	HBB	0.013	0.000	0.013	0.460	0.948
	Oracle	0.013	0.000	0.013	0.457	0.945
Setting 2: Gaussian Mixture						
Stratum 1	Empirical	0.005	0.003	0.005	0.261	0.938
	BB	0.005	0.003	0.005	0.264	0.939
	HBB	0.005	0.007	0.005	0.253	0.941
	Oracle	0.005	0.004	0.005	0.259	0.934
Stratum 4	Empirical	0.014	0.003	0.014	0.465	0.949
	BB	0.014	0.003	0.014	0.484	0.952
	HBB	0.011	0.018	0.010	0.440	0.950
	Oracle	0.013	0.000	0.013	0.463	0.957
Setting 3: Bernoulli Mixture						
Stratum 1	Empirical	0.007	0.005	0.007	0.310	0.933
	BB	0.007	0.005	0.007	0.312	0.936
	HBB	0.007	0.012	0.006	0.300	0.930
	Oracle	0.007	0.006	0.007	0.310	0.931
Stratum 4	Empirical	0.023	0.010	0.022	0.577	0.953
	BB	0.023	0.010	0.022	0.584	0.953
	HBB	0.021	0.032	0.020	0.544	0.945
	Oracle	0.022	0.011	0.022	0.575	0.95
Setting 4: Gamma Mixture						
Stratum 1	Empirical	0.005	0.013	0.005	0.268	0.942
	BB	0.005	0.013	0.005	0.272	0.947
	HBB	0.006	0.022	0.006	0.288	0.934
	Oracle	0.005	0.009	0.005	0.260	0.950
Stratum 4	Empirical	0.032	0.092	0.023	0.587	0.904
	BB	0.032	0.092	0.023	0.592	0.907
	HBB	0.011	0.002	0.011	0.405	0.943
	Oracle	0.010	0.018	0.009	0.371	0.933

to come at the expense of more increased bias. Indeed, in stratum 4 we see that absolute bias is about six times higher for HBB relative to BB (.018 v .003), however variation is also reduced (.01 v .014) leading to about a 20% reduction of MSE (.011 v .014). The HBB interval is narrower relative to BB (.440 v .484) while maintaining close to nominal coverage. In stratum 1 we see equivalent MSE across methods.

In the third setting, we consider the case where W is comprised of independent Bernoulli realizations - with separate probability vectors for each stratum. Each vector can have 2^{10} possible values, but there are far fewer than 2^{10} observations in any of the stratum. This complicates estimation of $Pv(W)$. In the sparse stratum $V = 4$, we see the HBB has nearly three times higher absolute bias (.032 v .01) but has reduced variability (.020 v .022). The MSE is reduced by about 8% (.021 v .023) with the HBB. Notably, the HBB interval is, on average, narrower while maintaining close to nominal coverage. While, in any stratum, we observe far fewer than the 2^{10} possible values of W, the HBB is able to borrow values seen in other strata.

Lastly, in the fourth setting we consider an even more complicated scenario where W is generated from a 10-dimensional location mixture of Gamma distributions. Each stratum has a different mean and, importantly, skewness. This scenario is designed to assess the tail-behavior of the HBB. As shown in Table 4.1, the HBB performs especially well in this complicated scenario. In stratum 4, the MSE, bias, and variance are lower than the BB. Intervals are narrower and coverage is closer to the nominal rate (94.3%). The small sample size in stratum 4 leads to too few covariate observations from the tail of the skewed Gamma to have a reliable nonparametric estimate of $P_4(W)$. This leads to poor BB estimates. However, in this setting the HBB borrows realizations from the tail in other strata - leading to a better estimate of $P_4(W)$.

4.5. Adverse Event Risk of Proton versus Photon Therapy

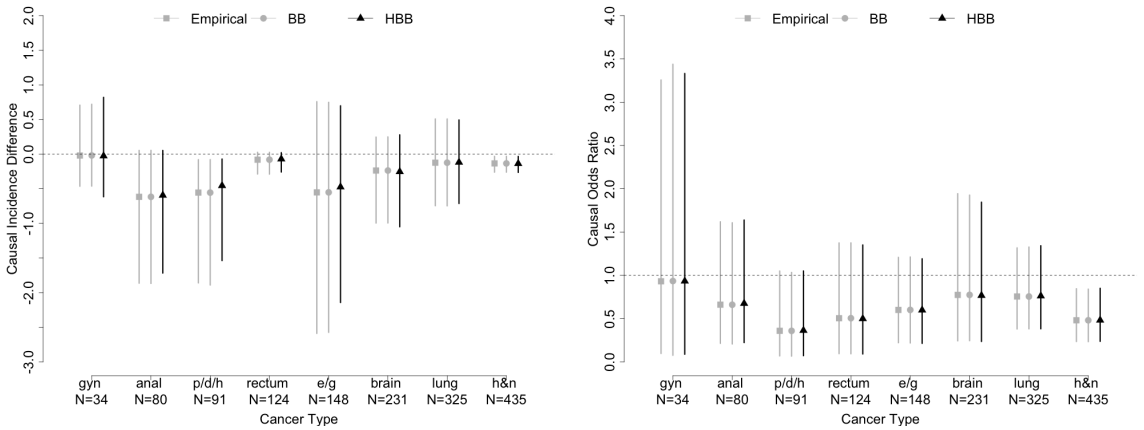
In this section we conduct posterior inference for casual contrasts of proton versus photon therapy among patients being treated for various locally-advanced cancers. For the cancers under consideration, standard-of-care therapy is a combination of chemotherapy and radiation - known as concurrent chemoradiotherapy (CRT). However, many modalities of radiation exist. The most common modality used in CRT has been photon radiation. In recent year, proton radiation therapy has become more accessible alternative to patients as barriers to access have eased and health sys-

tems have adopted the necessary technology. The idea of proton therapy is to deliver radiation in a more targeted way to the cancer site, while being less damaging to healthy tissue relative to photon. Observational data were collected from $n = 1468$ adult patients diagnosed with non-metastatic cancer and treated with CRT at the University of Pennsylvania Health Systems from 2011-2016.

Our data includes assigned treatment to CRT with either proton or photon radiation, several confounders measured at the time of treatment initiation, as well as the count of adverse events for a follow-up period of 90 days after treatment initiation. All patients in the sample had complete follow-up for at least 90 days. Previous research on this data (Baumann et al., 2020) has focused on the comparative risk of adverse events for patients on proton versus photon radiation. One hypothesis is that the more targeted nature of proton therapy will lead to fewer adverse events. Importantly, the differential risk may vary across cancer types. To address these questions, we conduct two analyses. In the first, we estimate a causal incidence difference between proton and photon patients across cancer type strata using a Poisson GLM for the adverse event count. In the second, we estimate of causal odds ratio for risk of any adverse event nonparametrically using BART. In the process we illustrate how the HBB can be combined with both parametric and nonparametric models for different outcome types. It can also be used to estimate different marginal causal contrasts (incidence differences, odds ratios, risk ratios, etc).

4.5.1. Parametric Model for Causal Incidence Difference

In this setting, our outcome is a count of adverse events over the 90-day follow-up, $Y \in \{0\} \cup \mathbb{Z}^+$. We observe data across $K = 8$ cancer types (e.g., lung, head and neck, and esophagus/gastric) indicated by V. Let $A = 1$ denote proton while $A = 0$ denote photon. Finally, let W be a vector of confounders including baseline age, race, sex, body-mass index (BMI), insurance plan, and charlson comorbidity index (a measure of baseline health status). We specify a conditional Poisson outcome model with the regression below. We adjust for race, sex, and insurance plan as categorical covariates. BMI, age, and charlson index are included as continuous covariates. More details on specification and prior choices are given in Web Appendix D in the Supporting Information. The mean of the Poisson distribution is modeled as $E[Y_i \mid A_i, W_i, V_i = v] = \exp\{\beta^v + W_i'\eta^v + A_i\theta^v\}$. Though parametric, such models are common in practice. Note we allow coefficients to vary across strata. Our target of interest here is the causal incidence difference within each stratum $\Psi(v) = E[Y^1 \mid V = v] - E[Y^0 \mid V = v]$. A negative value indicates lower incidence of adverse



Poisson Outcome Models

BART Outcome Models

Figure 4.3: Posterior mean and 95% credible interval estimates of stratum-specific causal contrasts under Poisson model (left) and BART (right). For both models, we set minimum desired sample size of $M = 100$. The abbreviations are gynecological (gyn), pancreas/duodenum/hepatobiliary (p/d/h), esophagus/gastric (e/g), and head/neck (h&n). Similar strata definitions were used in previous clinical studies (Baumann et al., 2020) and may be justified by anatomical closeness of affected organs.

events due to proton therapy relative to photon. To obtain this, we integrate the above regression over various estimators of $P_v(W)$. The left panel of Figure 4.3 displays results under three different estimates of P_v - including the HBB (with $M = 100$), BB, and the empirical distribution of W in each stratum. While the estimates for $\Psi(v)$ are largely similar across strata, note the HBB intervals are typically slightly shorter. Similarly, the point estimates are typically higher in these strata. This may partially reflect the trading off of increasing biased for reduced variability, as demonstrated in the simulations. However, these simulation results were averages across many runs. In any single data analysis, HBB need not produce narrower intervals.

Interpreting posterior estimates of $\Psi(v)$ in Figure 4.3, we see that the proton and photon therapies' effect on adverse event incidence are largely comparable across cancer type - with posterior distributions centered either near zero or very wide around 0 (as indicated by 95% credible intervals). Of course, these causal interpretations are subject to the validity of the required identification assumptions discussed earlier and in Web Appendix A. Moreover, these inferences are conditional on the very rigid parametric assumptions. For instance, it assumes linear (on log-scale) and additive covariate effects, in addition to a poisson outcome distribution. In the next section, we consider a

nonparametric estimation via BART.

4.5.2. Nonparametric Inference for Causal Odds Ratio via BART

Here we illustrate how the HBB can be used in conjunction with a nonparametric model for a binary outcome to obtain HTEs more robust to model misspecification. In this context let $Y \in \{0, 1\}$ be a binary indicator of any adverse event over the 90-day followup period. Then, we specify a conditional Bernoulli model for Y with regression $E[Y_i \mid A_i, W_i, V = v] = \Phi\big(f_v(W_i, A_i)\big)$ with prior $f_v \sim BART$ for $v = 1, \ldots, K$. This is the probit specification of BART outlined in Chipman, George, and McCulloch (2010). Above, Φ is the standard Normal distribution function and $f_v \sim BART$ is shorthand for the sum-of-trees model $f_v(W_i, A_i) = \sum_{j=1}^{J} g_j^v(W_i, A_i)$ with J trees, g_j^v. BART is characterized by a prior on the structure of each tree, g_j^v, consisting of terminal node parameters, splitting rules, and tree depth. Here we estimate stratum-specific models, with separate BART priors on each function. Thus, for each stratum v, we can get posterior draws of f_v under each treatment $A = a$. In this case our target is the stratum-specific causal odds ratio $\Psi(v) = \frac{E[Y^1|V=v]/(1-E[Y^1|V=v])}{E[Y^0|V=v]/(1-E[Y^0|V=v])}$. Values of $\Psi(v)$ less than one indicate lower risk of any adverse event due to proton therapy, relative to photon. Using standardization, we can compute each expectation by integrating $\Phi(f_v(W, a))$ over $P_v(W)$.The right panel of Figure 4.3 displays posterior results for $\Psi(v)$ under three different estimate of P_v - including the HBB (with M=100), BB, and the empirical distribution of W in each stratum. We notice that while point and interval estimates are generally similar across strata, the HBB intervals are somewhat narrower. However, according to these results, there is little posterior evidence for a reduction of adverse event risk due to proton therapy. While point estimates of the odds ratios are below one across strata, there is significant posterior uncertainty about the direction and magnitude of these effects, as indicated by the wide 95% credible intervals mostly overlapping one.

4.6. Discussion

The confounder distribution is a key unknown that must be estimated flexibly when making causal inferences. It is still more important in the context of HTEs where some strata are too sparse to allow reliable nonparametric estimation. In this paper we show that straightforward application of the Bayesian bootstrap can be improved upon in these scenarios with the HBB. The proposed HBB shares covariate information across strata to achieve more stable stratum-specific causal estimates.

The approach is computationally tractable, compatible with arbitrary outcome models, and makes no parametric assumptions about the distributions. As shown in the data analysis, it can be used to compute a variety of marginal causal contrasts.

We emphasize that potential applications of the HBB go beyond estimation of stratum-specific average causal effects. For instance, another popular causal estimand is the average treatment effect on the treated (ATT). This is defined as the average difference in potential outcomes among those assigned treatment. A Standardization-type procedure can be used here as well and requires integrating a regression over the distribution of confounders among the treated, $P(W \mid A = 1)$. If there are too few treated subjects to get a reliable nonparametric estimate of this distribution, it may be reasonable to borrow covariate information from untreated subjects, $P(W \mid A = 0)$, by shrinking towards the marginal via the HBB.

Lastly, our discussion of the connection between the HBB and the smoothed bootstrap motivates an extension to a "smoothed HBB". In Section 4.3.2, an $HBB(0)$ prior on the mixing distribution corresponds to a smoothed bootstrap within a stratum but prevents borrowing of information. We could in principle set $\alpha > 0$. In this case, the posterior becomes a hierarchical DP mixture of K_h - thus borrowing information across strata while modeling the distribution as a smooth mixture. If, for instance, K_h is a Gaussian kernel, we speculate the strength of the shrinkage could be informed by the $L - 2$ distance in covariate values across strata. While such distance-based shrinkage would be appealing, posterior computation for such a mixture model is much more involved - requiring updating the kernel parameters - and requires good default choices of K_h. An advantage of the HBB is that we require no specification of distance metric/kernel and maintain computational ease. However, this extension would be interested to pursue in the future.

CHAPTER 5

DISCUSSION AND CONCLUDING REMARKS

This dissertation developed (BNP) techniques for estimating causal effects in a variety of complicated observational data settings. The common thread of these methods is the joint emphasis on flexible estimation and uncertainty quantification. Nonparametric priors over the high-dimensional models presented in this work allow us to model observed data objects with only very minimal assumptions about the data generating mechanism. This greatly reduces the risk of bias due to misspecification that is common in parametric modeling procedures. Simultaneously, the fully Bayesian nature of our models allows for posterior uncertainty quantification over the models and the causal estimates.

This joint emphasis on uncertainty *and* flexibility supplies BNP methods with some of the best elements of classical statistics (typically concerned with the former) and machine learning (typically concerned with the latter). In fact, many BNP methods can be thought of as leveraging a latent probability model that underlies a given heuristic machine learning algorithm. This view has led to developments such as Bayesian additive regression trees (BART), which is a probabilistic analogue of regression tree-based algorithms. Similarly, the DP mixtures used in this dissertation can be thought of as probabilistic analogues of "mixture-of-experts" models used in machine learning. Other well-known methods in machine learning such as least absolute shrinkage and selection operator (LASSO) and Ridge regression can be motivated as maximum a posteriori inference under specific priors (double exponential in the first case and Gaussian for the latter).

Indeed, interest in BNP methods have seen a revival in computer science and machine learning under the guise of "probabilistic machine learning." While this term has many usages in the literature, Ghahramani, 2015 describes it as "the probabilistic approach to modeling uses probability theory to express all forms of uncertainty." He writes that "probability distributions are used to represent all the uncertain unobserved quantities in a model (including structural, parametric, and noise-related) and how they relate to the data. Then the basic rules of probability theory [presumably, Bayes' rule] are used to infer the unobserved quantities given the observed data." Section 2 of that piece discusses the role of nonparametric priors in detail. Similarly, Murphy, 2021 describes probabilistic

machine learning as "..machine learning, but from a probabilistic perspective. Roughly speaking, this means that we treat all unknown quantities ... as random variables, that are endowed with probability distributions...". From a statistics perspective, we can recognize this exactly as the Bayesian approach to inference, but with the added flexibility of nonparametric Bayes that go back to at least Ferguson, 1973.

Separately, the causal inference literature has increasingly focused on flexible semiparametric estimation of target quantities and moved away from parametric modeling. Before these advances, the approach was typically reversed: parametric models were fit to data and researchers tried to back out estimates of causal parameters from those models. One issue with this approach is that if the parametric model is incorrect, then estimators of the causal effects based on transformations of the parameters would have little meaning. Modern causal inference techniques - such as the ones employed in this dissertation - begin with the causal target of interest, identify them as a functional of observed data objects, then model those objects with minimal assumptions about their structure. In our case, we leverage BNP models of these observed-data objects. While these models have a high-dimensional parameter set that are difficult to interpret, the downstream causal effect estimates they form have a concrete causal interpretation under clearly specified assumptions.

The rising interest of nonparametric Bayes in machine learning coupled with the emphasis on semiparametric estimation in causal inference inspires optimism about the future of Bayesian nonparametric causal inference. These includes causal inference in dynamic treatment regime settings, sensitivity analyses for violations of causal assumptions, and and instrumental variable methods in the presence of unmeasured confounding - to list just a few.

APPENDIX A

APPENDICES FOR CHAPTER 2

A.1. Derivation of Posterior Quantities

A.1.1. Conditional Posterior

The conditional posterior of ω_i can be expressed up to a proportionality constant as

$$\begin{aligned} p(\omega_i|\omega_{1:(i-1)}, D) &\propto p(D|\omega_{1:i})p(\omega_i|\omega_{1:(i-1)}) \\ &\propto \frac{\alpha}{\alpha+i-1}p(D|\omega_{1:i})G_0(\omega_i) + \frac{1}{\alpha+i-1}\sum_{j<i} p(D_j|\omega_j)I(\omega_i=\omega_j) \\ &\propto \frac{\alpha}{\alpha+i-1}p(\omega_i|D_i, G_0)\int_{\omega_i} p(D|\omega_{1:i})dG_0(\omega_i) \\ &\quad + \frac{1}{\alpha+i-1}\sum_{j<i} p(D_j|\omega_j)I(\omega_i=\omega_j) \end{aligned} \tag{A.1}$$

The second line follows by substituting the Pólya Urn scheme for $p(\omega_i|\omega_{1:(i-1)})$. The last line follows from the fact that $p(\omega_i|D_i, G_0) = \frac{p(D|\omega_{1:i})G_0(\omega_i)}{\int_{\omega_i} p(D|\omega_{1:i})dG_0(\omega_i)}$.

A.1.2. Posterior Predictive Distribution

Causal inference using standardization is based on the posterior predictive mean of the outcome under some intervention $A = a$. Denote this as $\tilde{y}^a$. Letting tildes denote posterior predictive draws throughout,

$$p(\tilde{y}^a|D) = \int_{\omega_{1:n}}\int_{\tilde{l}}\int_{\tilde{\omega}} p(\tilde{y}^a|\tilde{l},\tilde{\omega},\omega_{1:n},D)p(\tilde{l}|\tilde{\omega},\omega_{1:n},D)p(\tilde{\omega}|\omega_{1:n})p(\omega_{1:n}|D)\; d\tilde{\omega}\; d\tilde{L}\; d\omega_{1:n} \tag{A.2}$$

Conventionally, we assume that that $\tilde{y}^a \perp \omega_{1:n}, D|\tilde{\omega}$ and $\tilde{l} \perp \omega_{1:n}, D|\tilde{\omega}$. That is, conditional on new parameter draws, the new outcome draw is independent of previous observations and their parameters.

$$p(\tilde{y}^a|D) = \int_{\omega_{1:n}}\int_{\tilde{l}}\int_{\tilde{\omega}} p(\tilde{y}^a|\tilde{l},\tilde{\omega})p(\tilde{l}|\tilde{\omega})p(\tilde{\omega}|\omega_{1:n})p(\omega_{1:n}|D)\; d\tilde{\omega}\; d\tilde{L}\; d\omega_{1:n} \tag{A.3}$$

Assuming ignorability and consistency hold,

$$p(\tilde{y}^a|D) = \int_{\omega_{1:n}} \int_{\tilde{l}} \int_{\tilde{\omega}} p(\tilde{y}|A=a,\tilde{l},\tilde{\omega}) p(\tilde{l}|\tilde{\omega}) p(\tilde{\omega}|\omega_{1:n}) p(\omega_{1:n}|D)\ d\tilde{\omega}\ d\tilde{L}\ d\omega_{1:n} \tag{A.4}$$

Recall that $\omega_i|\omega_1,\ldots,\omega_{i-1}$ from the Pólya Urn Blackwell and MacQueen, 1973

$$\omega_i|\omega_{1:(i-1)} \sim \frac{\alpha}{\alpha+i-1} G_0(\omega_i) + \frac{1}{\alpha+i-1} \sum_{j=1}^{i-1} I(\omega_i = \omega_j)$$

Substituting $i = n+1$ yields

$$\tilde{\omega}|\omega_{1:n} \sim \frac{\alpha}{\alpha+n} G_0(\tilde{\omega}) + \frac{1}{\alpha+n} \sum_{j=1}^{n} I(\tilde{\omega} = \omega_j)$$

Substituting this into Equation A.4 yields,

$$\begin{aligned}
p(\tilde{y}^a|D) &= \int_{\omega_{1:n}} \int_{\tilde{l}} \int_{\tilde{\omega}} p(\tilde{y}|A=a,\tilde{l},\tilde{\omega}) p(\tilde{l}|\tilde{\omega}) \Big[\frac{\alpha}{\alpha+n} G_0(\tilde{\omega}) + \frac{1}{\alpha+n} \sum_{j=1}^{n} I(\tilde{\omega} = \omega_j) \Big] p(\omega_{1:n}|D)\ d\tilde{\omega}\ d\tilde{L}\ d\omega_{1:n} \\
&= \int_{\omega_{1:n}} \int_{\tilde{l}} \Big[\int_{\tilde{\omega}} p(\tilde{y}|A=a,\tilde{l},\tilde{\omega}) p(\tilde{l}|\tilde{\omega}) \frac{\alpha}{\alpha+n} G_0(\tilde{\omega}) d\tilde{\omega} \\
&\quad + \frac{1}{\alpha+n} \sum_{j=1}^{n} \int_{\tilde{\omega}} p(\tilde{y}|A=a,\tilde{l},\tilde{\omega}) p(\tilde{l}|\tilde{\omega}) I(\tilde{\omega} = \omega_j) d\tilde{\omega} \Big] p(\omega_{1:n}|D)\ d\tilde{L}\ d\omega_{1:n} \\
&= \int_{\omega_{1:n}} \int_{\tilde{l}} \Big[\frac{\alpha}{\alpha+n} \int_{\tilde{\omega}} p(\tilde{y}|A=a,\tilde{l},\tilde{\omega}) p(\tilde{l}|\tilde{\omega}) G_0(\tilde{\omega}) d\tilde{\omega} \\
&\quad + \frac{1}{\alpha+n} \sum_{j=1}^{n} p(\tilde{y}|A=a,\tilde{l},\omega_j) p(\tilde{l}|\omega_j) \Big] p(\omega_{1:n}|D)\ d\tilde{L}\ d\omega_{1:n}
\end{aligned} \tag{A.5}$$

Integrating the above over y yields the posterior predictive mean

$$\begin{aligned}
E(\tilde{y}^a|D) = \int_{\omega_{1:n}} \int_{\tilde{l}} \Big[&\frac{\alpha}{\alpha+n} \int_{\tilde{\omega}} E(\tilde{y}|A=a,\tilde{l},\tilde{\omega}) p(\tilde{l}|\tilde{\omega}) G_0(\tilde{\omega}) d\tilde{\omega} \\
&+ \frac{1}{\alpha+n} \sum_{j=1}^{n} E(\tilde{y}|A=a,\tilde{l},\omega_j) p(\tilde{l}|\omega_j) \Big] p(\omega_{1:n}|D)\ d\tilde{L}\ d\omega_{1:n}
\end{aligned} \tag{A.6}$$

We can evaluate these integrals via Monte Carlo. Assume we have $t = 1,\ldots,T$ draws from the

posterior $\omega_{1:n}^{(t)} \sim p(\omega_{1:n}|D)$. How to obtain these draws is the subject of Web Appendix B. For each posterior draw, we propose a new set of parameters from the prior $\omega_0^{(t)} \sim G_0$. We also take several draws from $p(\tilde{l} \mid \omega_i^{(t)})$

A.1.3. Posterior Predictive Propensity Score

We can use the model to estimate the propensity score for each subject i using the posterior predictive probability of treatment conditional on subject i's covariates.

$$\begin{aligned} P(\tilde{A}|\tilde{L}=l,D) = \frac{p(\tilde{A},l|D)}{p(l|D)} &= \frac{\int_{\tilde{\omega}}\int_{\omega_{1:n}} p(\tilde{A},l,\tilde{\omega},\omega_{1:n}|D)\ d\tilde{\omega}\ d\omega_{1:n}}{\int_{\tilde{\omega}}\int_{\omega_{1:n}} p(l,\tilde{\omega},\omega_{1:n}|D)\ d\tilde{\omega}\ d\omega_{1:n}} \\ &= \frac{\int_{\tilde{\omega}}\int_{\omega_{1:n}} p(\tilde{A}|l,\tilde{\omega})p(l|\tilde{\omega})p(\tilde{\omega}|\omega_{1:n})p(\omega_{1:n}|D)\ d\tilde{\omega}\ d\omega_{1:n}}{\int_{\tilde{\omega}}\int_{\omega_{1:n}} p(l|\tilde{\omega})p(\tilde{\omega}|\omega_{1:n})p(\omega_{1:n}|D)\ d\tilde{\omega}\ d\omega_{1:n}} \end{aligned} \tag{A.7}$$

Again, substituting the Pólya Urn distribution,

$$\begin{aligned} P(\tilde{A}|l,D) &= \frac{\int_{\tilde{\omega}}\int_{\omega_{1:n}} p(\tilde{A}|l,\tilde{\omega})p(l|\tilde{\omega})\big[\frac{\alpha}{\alpha+n}G_0(\tilde{\omega}) + \frac{1}{\alpha+n}\sum_{j=1}^{n} p(\tilde{A}|l,\tilde{\omega})p(l|\tilde{\omega})\delta_{\omega_j}(\tilde{\omega})\big]p(\omega_{1:n}|D)\ d\tilde{\omega}\ d\omega_{1:n}}{\int_{\tilde{\omega}}\int_{\omega_{1:n}} p(l|\tilde{\omega})\big[\frac{\alpha}{\alpha+n}G_0(\tilde{\omega}) + \frac{1}{\alpha+n}\sum_{j=1}^{n}\delta_{\omega_j}(\tilde{\omega})\big]p(\tilde{\omega}_{1:n}|D)\ d\tilde{\omega}\ d\omega_{1:n}} \\ &= \frac{\int_{\omega_{1:n}}\big[\frac{\alpha}{\alpha+n}\int_{\tilde{\omega}} p(\tilde{A}|l,\tilde{\omega})p(l|\tilde{\omega})G_0(\tilde{\omega})\ d\tilde{\omega} + \frac{1}{\alpha+n}\sum_{j=1}^{n} p(\tilde{A}|l,\omega_j)p(l|\omega_j)\big]p(\omega_{1:n}|D)\ d\omega_{1:n}}{\int_{\omega_{1:n}}\big[\frac{\alpha}{\alpha+n}\int_{\tilde{\omega}} p(l|\tilde{\omega})G_0(\tilde{\omega})\ d\tilde{\omega} + \frac{1}{\alpha+n}\sum_{j=1}^{n} p(l|\omega_j)\big]p(\tilde{\omega}_{1:n}|D)\ d\omega_{1:n}} \end{aligned} \tag{A.8}$$

Again, given T posterior draws $\omega_{1:n}^{(t)}$ indexed by t, we can perform a Monte Carlo evaluation of the integral

$$P(\tilde{A}=1|l,D) \approx \frac{1}{T}\sum_{t=1}^{T} \frac{\frac{\alpha}{\alpha+n}\int_{\tilde{\omega}} p(\tilde{A}=1|l,\tilde{\omega})p(l|\tilde{\omega})G_0(\tilde{\omega})\ d\tilde{\omega} + \frac{1}{\alpha+n}\sum_{j=1}^{n} p(\tilde{A}=1|l,\omega_j^{(t)})p(l|\omega_j^{(t)})}{\frac{\alpha}{\alpha+n}\int_{\tilde{\omega}} p(l|\tilde{\omega})G_0(\tilde{\omega})\ d\tilde{\omega} + \frac{1}{\alpha+n}\sum_{j=1}^{n} p(l|\omega_j^{(t)})} \tag{A.9}$$

A.2. Metropolis-in-Gibbs Sampler and Relabeling Strategy

We use a Metropolis-in-Gibbs sampler for posterior inference. First, introduce latent cluster indicators for the n subjects at iteration t of the algorithm, $c_{1:n}^{(t)}$. Let $\mathcal{K}^{(t)}$ be the set of unique cluster labels at iteration t. In this iteration, each c_i may take on one of $K^{(t)}$ unique values, indexed by k. Associated k^{th} cluster is a set of cluster specific parameters $\omega_k^{(t)} = (\gamma_k^{(t)}, \beta_k^{(t)}, \phi_k^{(t)}, \eta_k^{(t)}, \theta_k^{(t)})$.

The MCMC procedure alternates between updating the cluster-specific parameters, ω_k, conditional on $c_{1:n}$. Then updates $c_{1:n}$ conditional on ω_k. The procedure is given in Algorithm 1.

Algorithm 1 Metropolis-in-Gibbs Posterior Sampler for Zero-inflated DP Mixture

1: Initialize $c_{1:n}^{(0)}$ to $K^{(0)}$ initial clusters with unique labels $\mathcal{K}^{(0)}$.

2: Initialize parameters $\omega_k^{(0)}$ for each of these clusters.

3: **for** t=1:T **do**

4: Update Cluster-specific Parameters

5: **for** k in $\mathcal{K}^{(t-1)}$ **do**

6: $\beta_k^{(t)} \sim p(\beta|\phi_k^{(t-1)}, D) \propto \prod_{i|y_i>0, c_i=k} N(y_i|x_i'\beta, \phi_k^{(t-1)}) \cdot G_0(\omega)$

7: $\phi_k^{(t)} \sim p(\phi|\beta_k^{(t)}, D) \propto \prod_{i|y_i>0, c_i=k} N(y_i|x_i'\beta_k^{(t)}, \phi) \cdot G_0(\omega)$

8: $\theta_k^{(t)} \sim p(\theta_k|D) \propto \prod_{i|c_i=k} p(l_i|\theta_k) \cdot G_0(\omega)$

9: $\eta_k^{(t)} \sim p(\eta_k|D) \propto \prod_{i|c_i=k} Ber(A_i|expit(m'\eta_k)) \cdot G_0(\omega)$

10: $\gamma_k^{(t)} \sim p(\gamma_k|D) \propto \prod_{i|c_i=k} Ber(z_i|expit(x'\gamma_k)) \cdot G_0(\omega)$

11: Conditional on $\omega_k^{(t)}$ for all $k \in \mathcal{K}^{(t-1)}$, update $c_{1:n}$

12: Propose a new cluster with parameters drawn from the base distribution, $\omega_{new} \sim G_0$

13: **for** i=1:n **do**

Update to existing cluster...

14: $P(c_i = k|c_{-i}, \{\omega_k^{(t)} : \forall k \in \mathcal{K}^{(t-1)}\}, D) \propto p(D_i|\omega_k^{(t)}) \cdot \frac{\sum_{j \neq i} I(c_i = c_j)}{\alpha + n}$

...or to the newly proposed cluster.

15: $P(c_i \neq k, \forall k \in \mathcal{K}^{(t-1)} \mid c_{-i}, \{\omega_k^{(t)} : \forall k \in \mathcal{K}^{(t-1)}\}, D) \propto p(D_i|\omega_{new}) \cdot \frac{\alpha}{\alpha + n}$

In practice we assume prior independence, so that $G_0 = p(\beta|\mu_\beta)p(\phi|\mu_\phi)p(\theta|\mu_\theta)p(\eta|\mu_\eta)p(\gamma|\mu_\gamma)$. Here, the μ terms represent hyperparameters. In this setting, the prior distribution factors so that in line 6, for example, $G_0(\omega) \propto p(\beta|\mu_\beta)$. Furthermore, we could choose $p(\beta|\mu_\beta)$ to be a multivariate

Gaussian with hyper mean vector and covariance matrix $\mu_\beta = (\lambda, \Sigma)$. This allows us to performing the sampling in line 6 using conjugacy. The idea is the same for the covariate model update in line 8 - where we could specify independent beta priors for binary variables, normal priors for continuous covariates, and Dirichlet priors for categorical variables. More complicated distributions can be chosen to better model correlations between the parameters at the expense of computational complexity.

Since there are no conjugate priors for the logistic models, the updates in lines 9 and 10 are done using a Metropolis step. We use a multivariate normal jumping distribution centered around $\eta_k^{(t-1)}$ and $\gamma_k^{(t-1)}$, respectively. The cluster assignment update beginning in line 13 is the most time-consuming part of the algorithm as it requires updating each subject one at a time. Nevertheless, it is simple to implement with existing statistical software.

At each iteration of the sampler, clusters can die (become unoccupied). New clusters can also appear. If the proposed cluster's parameters in line 12 happens to fit a subject's data better than the existing clusters, then the probability in line 15 will dominate those in line 14.

We compute $n \times n$ adjacency matrix, $\boldsymbol{M}^{(t)}$, for each posterior draw, $t = 1, \ldots, T$. The $(i,j)^{th}$ entry, $M_{ij}^{(t)} = I(c_i^{(t)} = c_j^{(t)})$, of this matrix is an indicator for whether subject i was clustered with subject j. Taking the element-wise mean of this matrix over Gibbs iterations, t, yields an $n \times n$ posterior mode matrix, $\boldsymbol{M}^*$ displaying the frequency with which subject i was clustered with subject j.

To obtain a hard classification status for each patient, we search for the $\boldsymbol{M}^{(t)}$ that is closest to the posterior mode matrix, $\boldsymbol{M}^*$, in the L_2 sense. That is, we search for the posterior $\boldsymbol{M}^{(t)}$ that yield the lowest $\sum_{i,j}(M_{ij}^{(t)} - M_{ij}^*)^2$. This approach provides an unambiguous way of classifying subjects according to the posterior mode in the presence of label switching. The clusters can then be summarized in terms of their outcome and observed confounder distributions.

Moreover, we can view $\boldsymbol{M}^*$ as a posterior adjacency matrix, represented as a network diagram where each subject is represented by a node and the edge between subject i and j has length given by M_{ij}^* - the posterior probability of subject i and j being clustered together.

We alluded to the computational complexity of the algorithm earlier. In our data analysis, running 40,000 iterations (after 20,000 burn-in) with about 1,000 subjects and six covariates took about 3.3

hours on a standard Windows 64-bit machine with 16GB of RAM, which we found to be reasonable. In settings with larger sample sizes, several MCMC chains can be run in parallel. For example, four chains that each take 10,000 draws after a 20,000 burn-in will yield a total of 40,000 draws in much less time. If gains from parallelization is still insufficient, other algorithms for sampling from DP posteriors exist that may potentially scale better with sample size. For example, the split-merge algorithm (Jain and Neal, 2004) explores the posterior space of cluster assignments by proposing splits and merges of existing clusters, instead of clustering one subject at a time.

Finally, we end with some guidance on choosing the initial number of clusters in the model. It is advisable to conduct a few MCMC runs with different amounts of initial clusters and monitor some posterior quantity of interest in each run. For example, we could choose to monitor marginal causal effect chains for different numbers of starting clusters. The chains can be visually inspected to make sure they mix well. A Gelman-Rubin $\hat{R}$ statistic could be computed - with $\hat{R} < 1.1$ indicating adequate mixing.

A.3. Causal Identification Assumptions

We can identify $E\left[Y_i^{A_i=1} - Y_i^{A_i=0}\right]$ under these assumptions:

- Ignorability: $Y_i^{A_i=a} \perp A_i = a|L_i$. Conditional on observed confounders, potential cost is independent of treatment assignment. Unmeasured confounding, for example, would be a violation of this assumption.

- Consistency: $Y_i^{A_i=a} = Y_i|A_i = a$. That is, Y_i observed under the actual treatment $A_i = a$ is equal to $Y_i^{A_i=a}$. Non-adherence to treatment assignment is an example of a violation of consistency.

- No interference: $Y_i^{A_i=a} \perp A_j, \ \ \forall i \neq j$. one subject's treatment assignment does not impact another's potential outcome. This assumption may not hold in vaccine studies, for example, where one subject's vaccination status may impact another subject's infection status.

- Positivity: $0 < P(A_i = 1|L_i) < 1$. If $P(A_i = 1|L_i) = 1$, then there would be some subpopulation, in terms of L, for which we would not have any control subjects to compare against.

A.4. Simulation Details

We simulate 1000 data sets of 3000 subjects each under two data generating processes: a clustered setting and parametric setting. For the clustered setting, we simulate $i \in 1, \ldots, 3000$ subjects in the following way:

- Draw cluster indicators, c_i with uniform probability $P(c_i = 1) = P(c_i = 2) = P(c_i = 3) = 1/3$.
- Draw confounder vector $L_i \sim P(L_i|\theta_{c_i}) = N(L_1|\mu_{c_i}, \phi_{c_i}) \prod_{j=2}^{5} Ber(L_j|p_{c_i})$. Where $\theta_{c_i} = (\mu_{c_i}, \phi_{c_i}, p_{c_i})$ and the cluster-specific parameters are $\theta_1 = (.9, .1^2, .25)$, $\theta_2 = (0, .1^2, .5)$, and $\theta_3 = (.375, .1^2, .75)$.
- Draw treatment $A_i|L_i \sim Ber(p_{c_i})$, where $p_1 = expit(.5 - 1L_1 + \sum_{j=2}^{4} L_j)$, $p_2 = expit(.2 + 2L_1 - 2\sum_{j=2}^{4} L_j)$, and $p_3 = expit(.8 + 1L_1 - \sum_{j=2}^{4} L_j)$.
- Draw structural zero indicator, $Z_i|A_i, L_i \sim Ber(pz_{c_i})$, where $pz_1 = expit(-2 - 2A_i + L_1 + \sum_{j=2}^{4} L_j)$, $pz_2 = expit(-2 + .5A_i - 2L_1 + 2\sum_{j=2}^{4} L_j)$, and $pz_3 = expit(2.5 + 1A_i - L_1 - \sum_{j=2}^{4} L_j)$.
- If $Z_i = 1$, then set $Y_i = 0$. Otherwise, draw $Y_i \sim Gamma(shape_{c_i}, scale_{c_i})$ where $scale_1 = 10000$, $scale_2 = 20000$, and $scale_3 = 30000$. The shape parameters are $shape_1 = 4 + A + 2L_1 + \sum_{j=2}^{4} L_j$, $shape_2 = 5 + A - 5L_1 + \sum_{j=2}^{4} L_j$, and $shape_3 = 7 + A + 1L_1 + \sum_{j=2}^{4} L_j$

For the parametric setting,

- Draw confounder vector $L_i \sim P(L_i|\theta) = N(L_1|\mu, \phi) \prod_{j=2}^{5} Ber(L_j|p)$. Where $\theta = (\mu, \phi, p) = (0, .1^2, .5)$.
- Draw treatment $A_i|L_i \sim Ber(p)$, where $p = expit(2 + 3L_1 - \sum_{j=2}^{4} L_j)$.
- Draw structural zero indicator, $Z_i|A_i, L_i \sim Ber(pz)$, where $pz = expit(-2 + .5A_i - 2L_1 + \sum_{j=2}^{4} L_j)$.
- If $Z_i = 1$, then set $Y_i = 0$. Otherwise, draw $Y_i \sim Gamma(shape, scale)$ where $scale = \mu/1e9$, $shape = \mu^2/1e9$, and $\mu = 3e5 + 50000A - 100000L1 + \sum_{j=2}^{4} L_j$.

We implement BART using the **BayesTree** package, the doubly robust estimator using the **twang** package, and code the Gamma hurdle and Gamma +.01 models in **Stan** with improper, uniform

priors for all parameters. All Bayesian results are based on 5000 posterior draws after 5000 burn-in iteration.

In both DGPs, for the DP mixture model, we use the base distribution, G_0, is the product of the following priors:

- A Gaussian prior on L_1 with mean equal to the empirical mean and variance equal to the empirical variance scaled by 10. Independent $Beta(1,1)$ priors on $L_{2:4}$.
- A Gaussian prior on the coefficient vector of the non-zero outcome model (β). The mean of this prior is centered around coefficient estimates from a linear model estimated using non-zero outcome values. The prior covariance matrix was set to be diagonal with variances equal to variances from a linear model estimated using non-zero outcomes.
- An $InvGamma(5, 10000)$ parameter is specified for the variance of the non-zero outcomes.
- The coefficient vectors of the logistic regression for both structural zeros and treatment were assigned a Gaussian priors with a zero mean vector and variances set to 2. This is relatively flat on the odds ratio scale.
- Both of the logistic regression parameter vectors were sampling using a Metropolis step with a Gaussian jumping distributions with diagonal covariance matrices. The diagonal elements were set to $.05$ for the treatment model and .005 for the structural zero model.

A.5. Data Analysis Details

We take 40,000 posterior draws after allowing for 20,000 burn-in. We initialize the sampler with 5 clusters. Normal priors on continuous covariate distributions. A Normal hyperprior is placed on the mean of this prior with empirical means and standard deviations (scaled by 10 to be slightly wider). Informative $InvGam(10, 10)$ hyperprior placed on the variance. $Beta(1,1)$ priors are placed on binary covariates distributions.

A Multivariate normal prior is placed on the outcome regression coefficients. The mean is set to linear regression coefficient estimates using only subjects with positive costs. The prior covariance was set to be diagonal with variance set to the diagonal of the previously mentioned regression's

covariance matrix. We scale these variances by 100 to make the prior a little wider. Treatment model and zero-inflation model regresion parameters are also Gaussian, centered around zero with variance of 2. Note that on an odds ratio scale this places most prior density on regression odds ratios ranging from .014 to 70.

Multivariate normal jumping distributions with diagonal variances are used for the Metropolis steps in both the treatment and zero-inflation models. The variance for the jumping distribution was set to .025. Lastly, we choose to estimate the concentration parameter rather than setting it. We place an $InvGam(1,1)$ prior on the parameter and implement a metropolis step using a Gaussian jumping distribution with variance 1.

For standardization, we evaluated the necessary integrals with 30,000 Monte Carlo iterations per posterior draw. The resulting MCMC chains are given

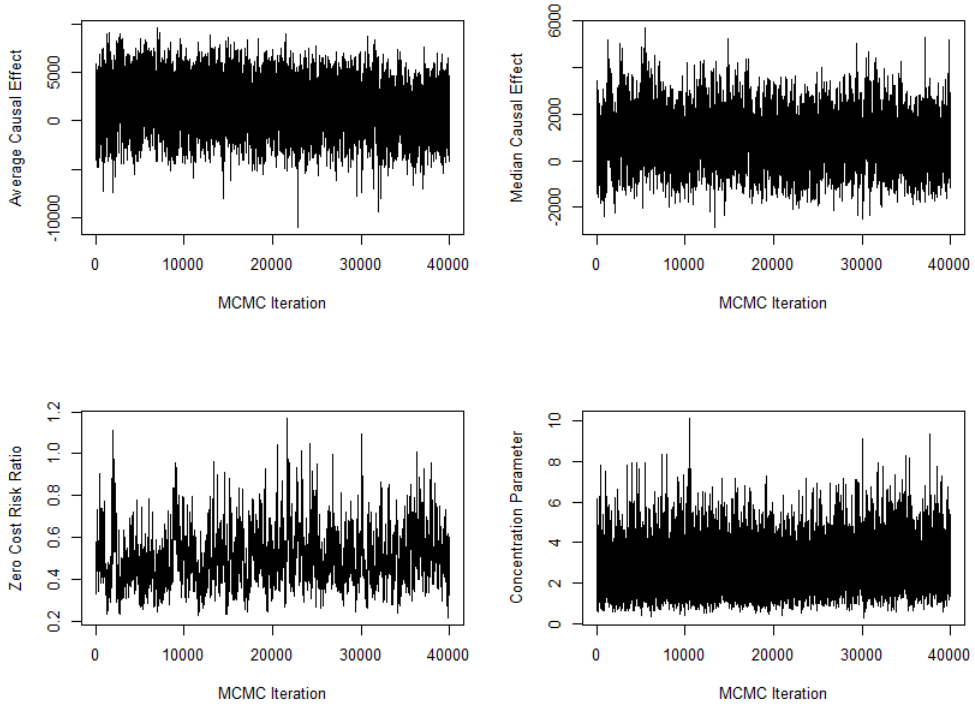


Figure A.1: Chains of 40,000 post-burn-in posterior draws of relevant quantities presented in the data analysis section.

Table A.1: Summary Statistics by Posterior Mode Cluster Assignment: Means are reported for continuous variables. Percentages are reported for categorical variables. All monetary amounts are in thousands of 2018 U.S. Dollars. Columns are ordered from lowest average cost to highest average cost. Small clusters were omitted for compactness.

	Red (n=452)	Green (n=160)	Blue (n=288)	Orange (n=127)
Total Inpatient Costs ($)	9.4	10.6	24	71.1
Radiation	91.8%	93.1%	89.9%	89.8%
Age (years)	72.4	73.2	73.9	74.9
Household Income ($)	74.6	69.9	68.4	64.5
White	90.5%	88.8%	82.6%	85.8%
Diabetic	0.0%	40%	38.2%	25%
CCI				
0	100%	15.0%	23.6%	22.0%
1	0.0%	85.0%	35.1%	31.5%
2	0.0%	0.0%	29.5%	15.7%
≥ 3	0.0%	0.0%	11.7%	30.7%
Grade = 1	21.7%	25%	21.2%	26.8%
FIGO Stage I-N0 or I-A	40.3%	39.4%	38.9%	42.5%

A.6. Zero-Inflated DP Mixture with Log-Transformed Outcome

For applications with non-negative outcomes close to zero, a local Gaussian conditional outcome distribution will likely yield negative predicted values. In these settings it may be desirable to have a predictive distribution with non-negative support. In this section, we outline a simple modification to our model to accommodate such scenarios. This involves log-transforming non-zero outcome values, then applying our model to the transformed data. This is equivalent to assuming a log-Normal conditional distribution for non-zero outcomes, as opposed to a Normal distribution.

At the prediction step, we can exponentiate log-outcomes back to the original scale. Since we use a fully Bayesian model, all inference is conducted using the posterior so that inference on the original

scale is still probabilistically valid. Specifically, define a transformed outcome

$$\dot{Y}_i = \begin{cases} 0 & Y_i = 0 \\ \log(Y_i) & Y_i > 0 \end{cases}$$

We can now use the proposed zero-inflated conditional outcome distribution with this transformed outcome

$$\dot{Y}_i \mid A_i, L_i, \beta_i, \gamma_i, \phi_i \sim \pi\left(x_i'\gamma_i\right)\delta_0\left(\dot{y}_i\right) + \left(1 - \pi\left(x_i'\gamma_i\right)\right) \cdot N\left(\dot{y}_i \mid x_i'\beta_i, \phi_i\right)$$

We will illustrate this with a toy simulation example. We simulate a zero-inflated outcome, Y_i, with a single covariate, X_i, and no treatment for $n = 600$ subjects. Subjects are generated from three unobserved, latent clusters $c_i \in \{1, 2, 3\}$ in the following way

$$\begin{aligned} p(c_i = k) &= \frac{1}{3}, \text{ for } k \in \{1, 2, 3\} \\ X_i &\sim N(\mu^{(c_i)}, \sigma^{(c_i)}) \\ Z_i | X_i &\sim Ber(expit(\gamma_0^{(c_i)} + \gamma_1^{(c_i)} X_i)) \\ Y_i | X_i, Z_i &\sim \begin{cases} 0 & Z_i = 1 \\ \log N(\beta_0^{(c_i)} + \beta_0^{(c_i)} X_i, \tau^{(c_i)}) & Z_i = 0 \end{cases} \end{aligned}$$

The true values are

- For $c_i = 1$: $\beta_0 = 4, \beta_1 = 1/10, \gamma_0 = -.5, \gamma_1 = -.5, \mu = 10, \tau = 1, \sigma = 3$.
- For $c_i = 2$: $\beta_0 = 5, \beta_1 = 1/5, \gamma_0 = -1, \gamma_1 = .05, \mu = 0, \tau = 1, \sigma = 2$.
- For $c_i = 3$: $\beta_0 = 3, \beta_1 = 1/10, \gamma_0 = -2, \gamma_1 = -.2, \mu = -10, \tau = 1, \sigma = 4$.

We compute the transformed outcome $\dot{Y}_i$ as defined above. Below is a figure of the data on both the original and log scale. Notice the outcome values near zero ($Y < 500$), the zero-inflation, skew-ness, and multimodality - the pathological features our model aims to capture. We will show that running our proposed model with the transformed outcome, $\dot{Y}$, will capture these complexities on the original scale.

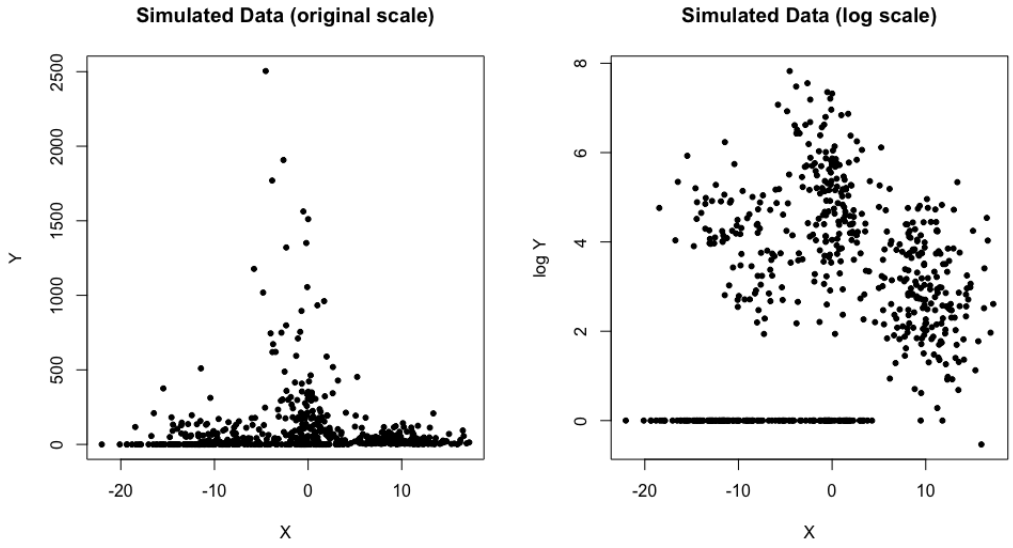


Figure A.2: Simulated data on both original and log-scale.

We run our model with the transformed outcome, omitting the treatment model since in this toy example we have no treatment.

$$\dot{Y}_i \mid X_i \sim \pi(\gamma_{0i} + \gamma_{1i} X_i)\delta_0(\dot{y}_i) + (1 - \pi(\gamma_{0i} + \gamma_{0i} X_i)N(\dot{y} \mid \beta_{0i} + \beta_{1i} X_i, \tau_i)$$

$$X_i \sim N(x_i | \mu_i, \sigma_i)$$

$$\omega_i | G \sim G$$

$$G | G_0, \alpha \sim DP(\alpha G_0)$$

Above, $\omega_i = (\gamma_{0i}, \gamma_{1i}, \beta_{0i}, \beta_{1i}, \mu_i, \sigma_i, \tau_i)$. We set G_0 to be the product of the following prior distributions (subscript i omitted for clarity):

- 2-dimensional Gaussian distribution for (β_0, β_1) centered around OLS parameter estimates using only positive outcomes. The prior covariance is the estimated covariance matrix from this OLS regression.
- 2-dimensional Gaussian distribution for (γ_0, γ_1) centered around the zero-vector with a diag-

onal covariance $diag(1, 1)$.

- Gaussian distribution for μ centered around the sample mean $\bar{X}$ with variance equal to sample variance $S_x = (n-1)^{-1} \sum_i (X_i - \bar{X})^2$
- $InvGam(shape = 2, rate = S_x)$ prior for σ.
- $InvGam(2, 1)$ prior for τ.
- $Gamma(1, 1)$ prior on α.

We initialize the model with 5 clusters and retain 1000 posterior draws after a 2000 draw burn-in. Below is a figure depicting clustering and prediction results. In column 1, shapes indicated posterior mode cluster assignment on both the original and log scale. In column 2, the observed data distribution is shown with the 2D-density contours. A single predictive outcome draw for each subject is depicted with the points. Note that the predictive draws are distributed similarly to the contours - indicating a good fit on both the log and original scales.

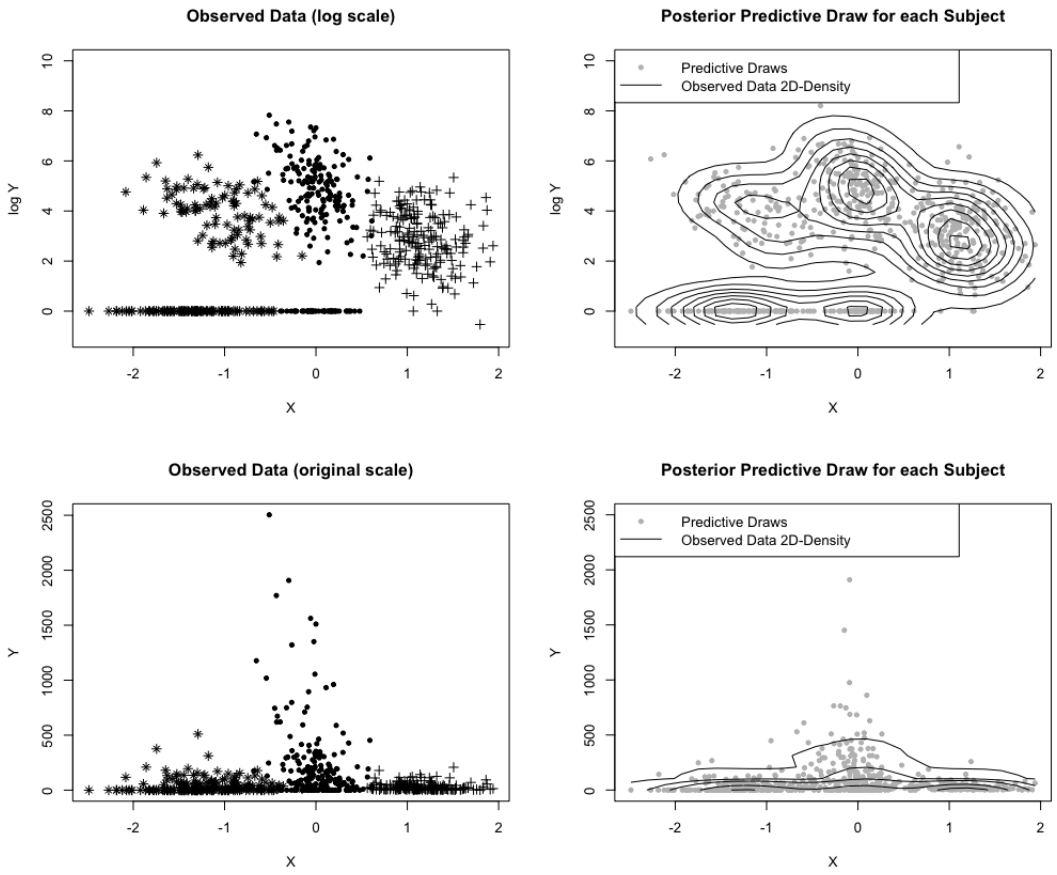


Figure A.3: Prediction and clustering results on original and log scales.

Below is a QQ plot similar to the one presented in the manuscript.

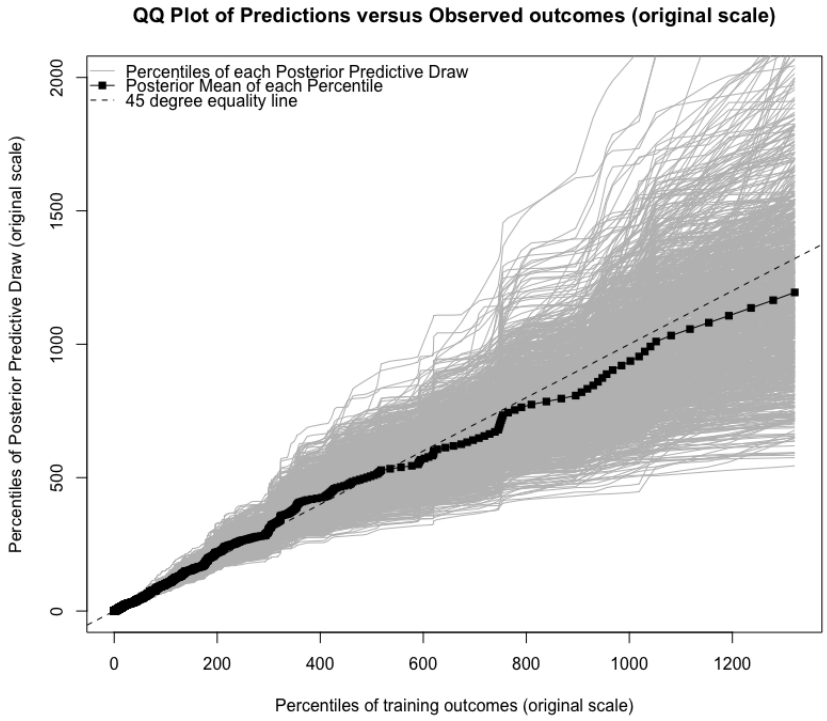


Figure A.4: Percentiles of predictive draws versus observed percentiles, on original scale.

This shows that the posterior predictions on the original scale of the data faithfully represent the data used for training - indicating a good fit despite being trained on the log-scale.

APPENDIX B

APPENDICES FOR CHAPTER 3

B.1. Identification of Causal Net Monetary Benefit

Recall that we are interested in estimating $\Psi = E[MV^{1,0}] - E[MV^{0,0}]$, where the expectation implicitly conditional on the parameters governing the joint cost-survival distribution. We can identify each term of Ψ. Starting with an iterated expectation over L,

$$\begin{aligned}
E[MV^{a,0}] &= E_{\mathcal{L}}[E_{\mathcal{Y},\mathcal{D}}[MV^{a,0} \mid L, \omega_{1:n}, \theta_{1:n}, \lambda_0]] \\
&= E_{\mathcal{L}}[E_{\mathcal{Y},\mathcal{D}}[MV^{a,0} \mid, A = a, \delta = 0, L, \omega_{1:n}, \theta_{1:n}, \lambda_0]] \\
&= E_{\mathcal{L}}[E_{\mathcal{Y},\mathcal{D}}[MV \mid A = a, \delta = 0, L, \omega_{1:n}, \theta_{1:n}, \lambda_0]] \\
&= \int_{\mathcal{L}} E_{\mathcal{Y},\mathcal{D}}[MV \mid A = a, \delta = 0, L, \omega_{1:n}, \theta_{1:n}, \lambda_0] dP(L) \\
&= \int_{\mathcal{L}} \int_{\mathcal{Y},\mathcal{D}} (D\kappa - Y) p(Y, T \mid A = a, L, \delta = 0, \omega_{1:n}, \theta_{1:n}, \lambda_0) dP(L)
\end{aligned}$$

Note above, $\mathcal{Y}$ and $\mathcal{D}$ are the spaces we integrate over. This last line is each term of Equation (3.3). The second line follows from joint ignorability (IA.1), allowing us to condition on $A = a, \delta = 0$ after first conditioning on L. The third line follows from joint consistency, IA.2, allowing us to drop the superscripts on monetary value. These are extensions of the usual conditional ignorability and consistency assumptions under censoring (Robins, Hernán, and Brumback, 2000) extended to handle a bivariate cost-survival time outcome. The interference assumption, IA.4, allows us to write $MV^{a_{1:n},0_{1:n}} = MV^{a,0}$. That is, each subject's potential monetary value is independent of others' treatments or censoring status. Said another we, we learn nothing about someone else's potential monetary value by learning another's treatment assignment. Joint positivity, IA.3, is requires so that we do not condition on a zero-probability event in the second equality. The expression above identifies a causal estimand that is purely a function of unknown parameters. Thus a posterior distribution over the parameters induces a posterior distribution over monetary value.

B.2. Posterior Computation

Gamma Process Prior Specification

This appendix provides additional details for updating the baseline hazard model with a dependent Gamma process prior (Nieto-Barajas and Walker, 2002). Much of this is a detailed overview of the results established by Nieto-Barajas and others in their 2002 paper and outlined in documentation of the **BGPhazard** R package. We provide an abbreviated presentation adapted to the context of our joint model for the reader's convenience.

Consider observing right-censored survival time data for $i = 1, \ldots, n$ subjects with survival time T_i and death indicator δ_i. Consider a partition, $\{\tau_v\}_{v=1:V}$, of the time interval such that $0 < \tau_1 < \tau_2 < \cdots < \tau_V$ where $\tau_V > max_i(T_i)$. In a setting with fixed study end, τ, we could set $\tau_V = \tau$. In this case we consider equally-spaced interval such that $\Delta_v = \tau_v - \tau_{v-1}$ for all v. A piecewise constant hazard model can be defined as

$$\lambda_0(t) = \sum_{v=1}^{V} \lambda_{0v} I(\tau_{v-1} < t \leq \tau_v)$$

If *a priori* the baseline hazard $\lambda_0(t) \sim GP(b\lambda_0^*, b, \xi = 0)$, then the hazard rate in each interval follows $\lambda_{0v} \sim Gam(b\lambda_{0v}^*, b)$, where the first argument in the shape and the second argument is the rate. In the shape, we've defined $\lambda_{0v}^* = \{\Lambda_0^*(\tau_v) - \Lambda_0^*(\tau_{v-1})\}/\Delta_v$, where Λ_0^* denotes the prior cumulative hazard. Thus the prior mean hazard at each interval is $E[\lambda_{0v}] = \lambda_{0v}^*$. This is known as the independent Gamma process prior because the hazard at each increment is independent *a priori*. The *dependent* Gamma process of Nieto-Barajas extends this process to introduce dependence between hazards in nearby increments - providing a smoother estimate that is less dependent on choice of time partition. They do this by introducing latent processes $\{c_v\}_{1:V}$ and $\{u_v\}_{1:V}$ and is denoted with GP, as above, but with $\xi > 0$. The process is initialized with $\lambda_1 \sim Gam(b\lambda_{01}^*, b)$. Now for $v \in \{1, 2, \ldots \tau - 1\}$, we have $u_v \mid \lambda_v, c_v \sim Pois(c_v \lambda_{0v})$ and $\lambda_{0v+1} \mid u_v, c_v \sim Gam(b\lambda_{0v+1}^* + u_v, b + c_v)$. The conditional prior mean of this process is

$$E[\lambda_{0v} \mid \lambda_{0v-1}] = \frac{b\lambda_{0v}^* + c_{v-1}\lambda_{0v-1}^*}{b + c_{v-1}}$$

So the prior mean baseline hazard rate in current interval v is a weighted average of the prior

baseline hazard rate, λ^*_{0v}, in the current time interval and the prior baseline hazard rate in the previous time interval, λ^*_{0v-1}. This is the induced AR(1) smoothness of the dependent Gamma Process. Following, Nieto-Barajas we place a hyperprior on $\{c_v\}_{1:V}$, assuming $c_v \mid \xi \overset{iid}{\sim} Exp(\xi)$. Where the prior mean is $E[c_v] = \xi$. The magnitude of ξ (relative to b) controls the aggressiveness of the prior AR(1) shrinkage. if $\xi >> b$, then on average $c_{v-1} >> b$ at all intervals v, meaning that $E[\lambda_{0v} \mid \lambda_{0v-1}] \approx \lambda^*_{0v-1}$. Similarly, if $\xi << b$, then $E[\lambda_{0v} \mid \lambda_{0v-1}] \approx \lambda_{0v}$ - i.e. almost no shrinkage to the previous hazard. It can be shown above that setting $\xi = 0$ above reduces this to the independent Gamma process.

Thus, the notation $\lambda_0 \sim GP(b\lambda^*_0, b, \xi)$ denotes this prior for the piecewise constant model $\lambda_0(t)$. Specifically, the joint prior is

$$p(\lambda_{01:V}, c_{1:V}, u_{1:V} \mid b, \xi) = p(\lambda_1)p(u_1 \mid \lambda_{01}, c_1)\prod_{v=2}^{V} p(u_v \mid \lambda_v, c_v)p(\lambda_{0v} \mid u_{v-1}, c_{v-1})\prod_{v=1}^{V} p(c_v \mid \xi) \qquad \text{(B.1)}$$

With hyperparameters b, ξ, and λ^*_0. Notational dependence on λ^*_0 has been suppressed for compactness. This can be combined with the likelihood for the observed data to obtain conditional posteriors for each of the three parameter blocks, $\lambda_{01:V}, c_{1:V}$, and $u_{1:V}$. We discuss likelihood construction in the next section. b

Gamma Process Likelihood Construction

Now we consider the $GP(b\lambda^*_0, b, \xi)$ prior for the baseline hazard in a proportional hazard model $\lambda(t \mid X_i, \theta_i) = \lambda_0(t)\exp\left(X_i'\theta_i\right)$, where $\lambda_0(t) = \sum_{v=1}^{V} \lambda_{0v} I(\tau_{v-1} < t \leq \tau_v)$. Specifically, our goal is to find the posterior $p(\{\lambda_{0v}\}_{1:V}, \{c_v\}_{1:V}, \{u_v\}_{1:V} \mid \mathcal{D})$, where $\mathcal{D}$ indicates the observed data.

For convenience in presentation, define $\eta_i = X_i'\theta_i$. Also note that under the piece-wise constant model, the cumulative hazard is $\Lambda_i(t) = \int_0^t \lambda_0(s)e^{\eta_i}ds = \sum_{v=1}^{V} \lambda_{0v}e^{\eta_i}\Delta_v(t)$. Here, $\Delta_v(t) = (t - \tau_{v-1})I(t \in (\tau_{v-1}, \tau_v]) + \Delta_v I(t > \tau_v)$.

Conditional on θ_i, standard survival likelihood construction with right-censored data yields

$$p(T_i \mid X_i, \theta_i, \delta_i, \lambda_{01:V}) = \prod_{i|\delta_i=1} f(T_i \mid X_i, \theta_i) \prod_{i|\delta_i=1} S(T_i \mid X_i, \theta_i)$$

Subjects with an event contribute to the likelihood via the density, f, and censored subjects contributed via the survival function S, both of which can be expressed in terms of the hazard. Denote

λ_{0v_i} as the hazard rate of the increment in which subject i died. The density evaluated at subject i's death time is,

$$f(T_i \mid X_i, \eta_i) = \lambda_0(T_i)e^{-\Lambda_i(T_i)} = \lambda_{v_i}e^{\eta_i}\exp\Big\{-\sum_{v=1}^{V}\lambda_{0v}e^{\eta_i}\Delta_v(T_i)\Big\} \tag{B.2}$$

The survival function in terms fo the hazard is,

$$S(T_i \mid X_i, \theta_i) = \exp\Big\{-\Lambda_i(T_i)\Big\} = \exp\Big\{-\sum_{v=1}^{V}\lambda_{0v}e^{\eta_i}\Delta_v(T_i)\Big\}$$

So the full likelihood is

$$p(T_i \mid X_i, \theta_i, \delta_i, \lambda_{01:V}) = \Big(\prod_{i|\delta_i=1}\lambda_{0v_i}\Big)\exp\Big\{\sum_{i|\delta_i=1}\eta_i\Big\}\exp\Big\{-\sum_{v=1}^{V}\lambda_{0v}\Big(\sum_{i=1}^{n}e^{\eta_i}\Delta_v(T_i)\Big)\Big\} \tag{B.3}$$

Gamma Process Posterior Updates

The likelihood (B.3) can be combined with the joint prior (B.1) to obtain the following conditional posteriors distributions for $u_{1:V}$, $c_{1:V}$, and $\lambda_{01:V}$. Note all of these distributions are also conditional on data, $\mathcal{D}$. First, the conditional posterior distribution of $\{c_v\}_{1:V}$ is

$$p(c_v \mid u_v, \lambda_{ov+1}, \lambda_{0v}) \propto \begin{cases} c_v^{u_v}\exp\Big\{-(\lambda_{0v}+\lambda_{0v+1}+\frac{1}{\xi})c_v\Big\}(b+c_v)^{\lambda^*_{0v+1}+u_v} & v = 1,\ldots,V-1 \\ Gam(u_v+1, \lambda_{0v}+\frac{1}{\xi}) & v = V \end{cases} \tag{B.4}$$

For $v = 1,\ldots,V-1$ this update is not conjugate. We sample each c_v separately using Adaptive Metropolis-Hastings with separate proposal variances for each c_v. The proposal variances are tuned every few iterations in the burn-in period to target a 23.4% acceptance rate, which has been shown to be optimal in around 10-dimensional sampling problems (**roberts2001**). The latent process $\{u_v\}_{1:V}$ can be updated from the following conditional posterior,

$$p(u_v \mid c_v, \lambda_{0v+1}, \lambda_{0v}) \propto \begin{cases} \frac{\left[c_v\lambda_{0v}\lambda_{0v+1}(b+c_v)\right]^{u_v}}{\Gamma(u_v+1)\Gamma(\lambda^*_{0v+1}+u_v)} & v = 1,\ldots,V-1 \\ Pois(c_v\lambda_{0v}) & v = V \end{cases} \tag{B.5}$$

Note here u_v is integer-valued and non-conjugate for $v = 1, \ldots, V-1$. To sample from these conditional posteriors, we use grid sampling with a large grid of points $\{0, \ldots, 10000\}$. Finally, the conditional posteriors of the hazard rate in each interval is given by

$$p(\lambda_{0v} \mid -, D) = \begin{cases} Gam\big(d_1 + u_1 + \lambda^*_{01}, c_1 + b + \sum_{i=1}^n e^{\eta_i} \Delta_1(T_i)\big) & v = 1 \\ Gam\big(d_v + u_v + u_{v-1} + \lambda^*_{0v}, b + c_v + c_{v-1} + \sum_{i=1}^n e^{\eta_i} \Delta_v(T_i)\big) & v = 2, \ldots, V \end{cases} \tag{B.6}$$

Above, d_v is the number of deaths in interval v. Note that the conditional distribution is fully conjugate for all v and can be sampled directly. Note also that this update is the only Gamma Process update that involves data. The processes $u_{1:V}$ and $c_{1:V}$ are latent and the updates do not involve data - but they do induce a dependence between the $\lambda_0 v$, which now must be updated sequentially and in order.

Concentration Parameters

The two concentration parameters of the EDP, α_θ and α_ω, are given $Gam(1, 1)$ priors. We follow the implementation in Roy et al., 2018. Details can be found in the supplement to their 2018 paper.

Monte Carlo Integration for Monetary Value

The expectation can be expressed as

$$\mu(a, 0) = \kappa E[D \mid -] - \int_0^\tau \int_0^\infty E[Y \mid D, -] p(D \mid -) dY dD$$

Note we use "$-$" to denote the conditioning set, which was made explicit in the main body of the paper.

- The first term, $E[D \mid -]$, (average death time within 2-years under treatment a) can be computed in closed form. Since we partition time interval (see Appendix B.2) into K intervals, the probability of dying in interval k is $p(t \in [\tau_k, \tau_k + 1] \mid -)$. Within each interval, death time is

uniform - so mean is $\frac{\tau_{k+1}+\tau_k}{2}$.

$$E[D \mid -] \approx \sum_{k=1}^{K} \frac{\tau_{k+1}+\tau_k}{2} \cdot p(t \in [\tau_k, \tau_k + 1] \mid -)$$

At every iteration, $p(t \in [\tau_k, \tau_k + 1] \mid -)$ is given by substituting the parameter draws in this iteration into Equation (B.2).

- Second term: For each subject, draw death interval proportional to $p(t \in [\tau_k, \tau_k + 1] \mid -)$. Then, within each interval draw a death time t^* uniformly within that interval. Compute $E[Y \mid T = t^*, -]$ using this drawn value and the parameter draws in the current iteration.

B.3. Simulation Details

Data Generation

In the log-normal setting, we simulate data as follows. For subject $i = 1, \ldots, N$,

- Simulate latent cluster membership: $c_i \sim Ber(p_c)$, a 5-dimensional confounder L_i. This vector contains one continuous confounder drawn from a standard Normal distribution in the first entry and four binary confounders draw from Bernoulli distribution with probability .5.

- Simulate treatment:

$$A_i \sim Ber(expit(0 + (.1, .5, -.5, .5, -5)' L_i))$$

- Simulate survival time, T_i: from a Weibull distribution (using the proportional hazard parameterization) with shape 10 and scale $exp(\eta_i)$. Where

$$\eta_i = c_i \cdot [(0, .1, -.1, .1, -.1)' L_i] + (1 - c_i) \cdot [(1, -.1, .1, -.1, .1)]) + (-3 + 2c_i) A_i$$

  Notice that the treatment effect on survival is bimodal, along with the covariate effects.

- Simulate a covariate-dependent censoring time: C_i, from the same Weibull as above.

- Simulate Observed time observed time: Draw $Z_i \sim Unif(0, 1)$ and simulate censoring indicator $\bar{\delta}_i = I(C_i < D_i) \cdot I(Z_i < p_\delta)$. If $\bar{\delta}_i = 1$, then $T_i = min(C_i, D_i)$.

- Simulate accumulated cost up to T_i:

$$Y_i \sim \log N\Big(mean = \mu_i,\ sd = .05\Big)$$

where

$$\mu_i = 2c_i + (.1, .2, .2, .2, .2)' L_i - 2T_i + .3A_i$$

Here we have a bimodal cost distribution (different means depending on c_i) but homogeneous treatment effect on costs.

- Output observed data $D_i = (Y_i, T_i, \delta_i = 1 - \bar{\delta}_i, L_i, A_i)$.

In the Normal setting, we simulate data as above with the following modifications:

- Simulate survival and censoring times time with $\log$ scale parameter

$$\eta_i = c_i \cdot [(-1, .1, -.1, .1, -.1)' L_i] + (1 - c_i) \cdot [(1, -.1, .1, -.1, .1)]) + 2c_i \cdot A_i$$

Note again that treatment and covaraite effects are bimodal (dependent on c_i).

- Simulate outcome data from a *Normal* distribution with standard deviation .5 and mean

$$\mu_i = 5 + 5c_i + (.1, .5, .5, .5, .5)' L_i - 3A_i + T_i$$

- Here the treatment and covariate effects on Y are homogeneous.

We simulate each dataset with $N = 1500$. In the bimodal setting, $p_c = .5$. In the parametric setting, the $p_c = 0$ - so all subjects are from the same cluster. We set $p_\delta = .4$ in the high setting to target 20% censoring and $p_\delta = .1$ in the low setting to target 5% censoring. For each setting Normal/log-Normal -p_δ-p_c combination, we simulate 200 such datasets.

EDP-GP Prior Settings

First we discuss the settings for the log-Normal data generating mechanism. For the Gamma Process prior, we partition the interval from $[0, max(T_i)]$ into equal size increments of $.1$. We set $\xi = 1e - 6$ to be quite small (very flat) to allow the likelihood to drive the posterior estimate. We set

$b = \xi$ thus inducing an AR1 dependence between increments that is as informative as the shrinkage towards λ_0^*, which we set to an exponential hazard with rate 400 - close to the average empirical hazard rate across time points. Notice the actual baseline hazard is generated from a Weibull, so our prior is deliberately misspecified as it likely would be in practice.

The prior on θ_i, $G_{0\theta}$ is set to a multivariate Gaussian with zero mean vector and diagonal covariance $3^2 I_6$. Where I_6 is the 6×6 identity matrix, where 6 is the number of covariates (5 confounders and one treatment indicator). This is flat on the hazard ratio scale.

Since we fit a Gaussian conditional model for Y, the prior $G_{0\omega}$ is a product of a prior on the covariate effects and prior on the variance. Regarding the former, we again use a multivariate Gaussian with zero mean vector and covariance $3^2 I_7$, where the identity matrix has a diagonal entry for the five confounders, treatment indicator, and observed time. This is fairly flat relative to the true conditional outcome variance (on log scale) of $.05^2$. The prior for the variance is set to an inverse gamma distribution. In the bi-modal setting we set this distribution to have shape and scale equal to 20. This centers the prior variance around 1. In the parametric/unimodal setting we use a slightly tighter prior around 1 - with shape and rate equal to 100. These tighter settings like 20 and 100 help regularize the Gaussian model we fit to the skewed Y data.

For the Normal data generating mechanism much of the settings above is the same. We only change the shape parameter of the inverse gamma distribution on the conditional cost variance to be 5 with a rate of 20. This is a fairly flat prior.

For each data set, we run the MCMC sampler for 7000 iterations and discard the first 2000 as burn-in. This yields 5,000 posterior draws which we use for inference about NMB. In all settings, we initialize the model with three ω clusters, each having three θ sub-clusters. This initialization is very different from the true data generating mechanism that either generates data from a single $\omega - \theta$ cluster and two ω (top-level) clusters.

Since we fit a Gaussian model, each cluster's conditional ω posterior is conjugate with our Normal-Inverse-Gamma prior. This is a simple update. For the θ cluster parameters we use a Metropolis update with Gaussian jumping distribution. The jumping covariance is identity with .1 along the diagonals. Similarly, we use a Metropolis step to update $\{c_{1:v}\}_{1:V}$ (see Appendix B.2) at each step. Each c_v is updated from an independent Gaussian jumping distribution with variance .5. We adapt

both of these jumping distribution variances every 25 iterations starting from iteration 50 and ending at iteration 200 to target an acceptance rate of 23.4% per **roberts2001**

Doubly-Robust Implementation

Here we describe the doubly-robust NMB estimator of Li et al., 2018 implemented in our simulations. The cost and survival time models are estimated using super learner with regression trees, generalized additive models, generalized linear models, and GLM-Net included in the ensemble. We use a correctly specified logistic regression for the treatment model. This is quite generous since doubly-robust estimators are guaranteed to be consistent with a correctly specified treatment model (though the convergence rate can be quite slow if the outcome model is very misspecified.).

Since we have covariate dependent censoring, we estimate the inverse censoring weights using a discrete-time failure model as described in Section 3.1.1 of their paper. To summarize, these weights are computed using estimates of the probability of censoring at each time point, conditional on not having been censored before that time point. This is estimated using a logistic regression of a censoring indicator at each time point on simulated confounders, treatment and time-level fixed effects. Intervals are computed using a 95% BCa interval after 1502 bootstrap iterations (BCa intervals require more bootstrap iterations than observations in the sample).

B.4. Data Analysis Details

We partition the interval from $[0, 24]$ into increments of $.5$. To sample from conditional posterior of $\{c_v\}_{1:V}$ (as mentioned Appendix B.2) we use a Metropolis-Hastings update from jumping variance of $.5$. To sample from the posterior of θ (the covariate effects of the hazard model) we use a joint Metropolis-Hastings update with an initial identity covariance matrix multiplied by $.1$ along the diagonal. For both samplers, we adapt these jumping variances every 25 iterations starting from iteration 50 to iteration 200. Every 25^{th} iteration we use the previous 25 draws to target an acceptance rate of 23.4%, as per **roberts2001** Since we assume a log-normal cost distribution, posterior updates are conjugate using log-transformed cost. Figure B.1 contains some diagnostic plots with a discussion in the caption. These plots show the MCMC chains to be well-mixed and model fit to be adequate. The total run-time was approximately 50 hours when parallelizing the three chains.

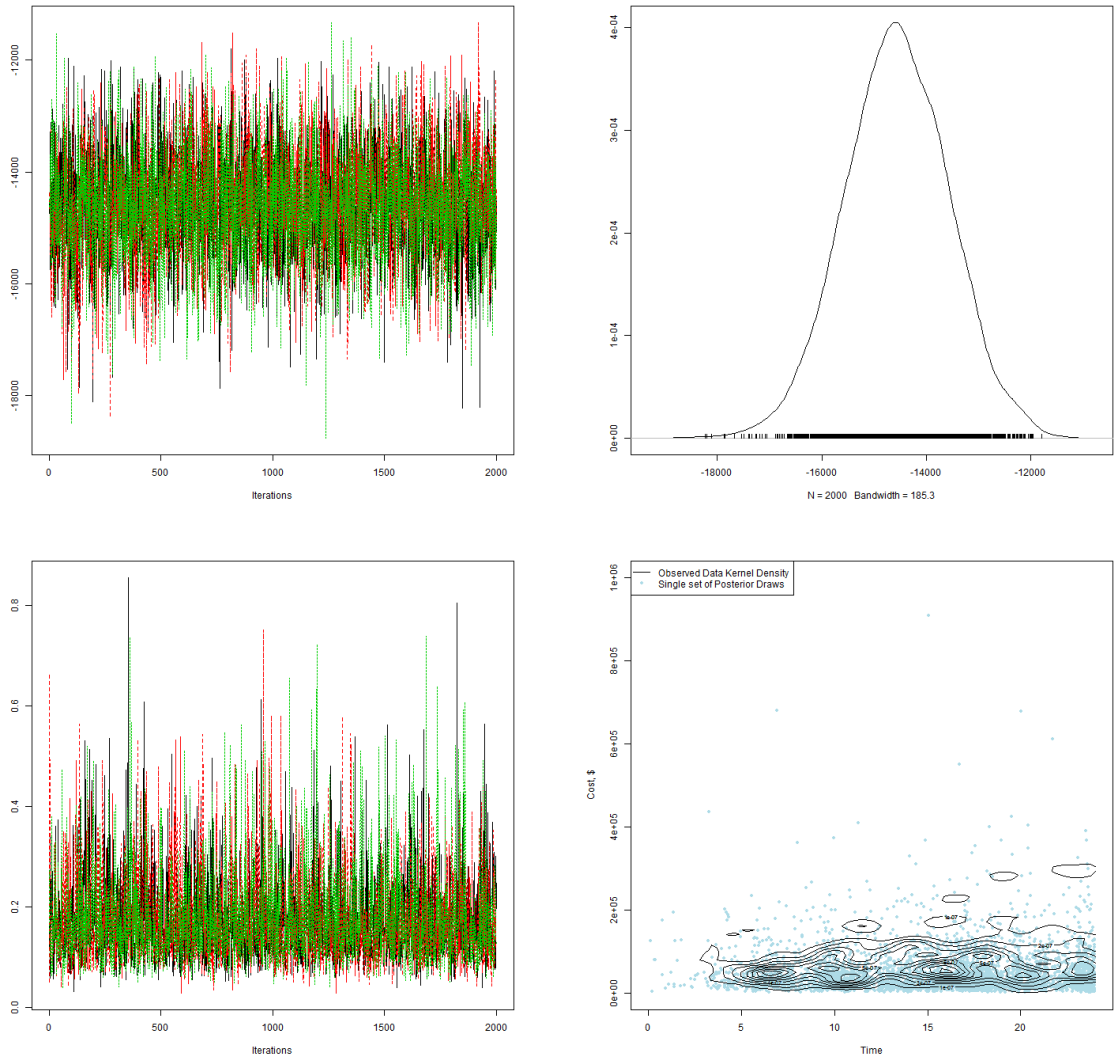


Figure B.1: Diagnostic plots supporting data analysis results. Top row: traceplots of three MCMC chains of posterior NMB draws (left) and distribution of the combined posterior NMB draws of all chains (right). These NMB draws are based on $\kappa = \$50,000/12$. All three chains mix after starting with different initial clusters and seeds. Corresponding posterior is unimodal and peaked around $\$14,500$. Panel C shows the traceplots of three MCMC chains for DSI, which mix well. Finally, panel D shows a kernel density estimate of the joint observed time and cost distribution. In blue we show a single set of posterior predictive draws of joint cost and observed time. This shows adequate model fit: the posterior predictive is placing mass around the observed data. Moreover, the posterior predictive allows for occasional large cost draws. This indicates the local log-Normal cost distribution is able to capture skewness. If, for instance, the posterior predictive draws did not overlap with the observed data, we would be suspicious of the model fit.

For the doubly-robust (DR-SL) implementation of Li et al., 2018, we estimate the propensity score model, cost model, and survival model using super learner with regression trees, GLM, and GLMnet as candidates. Inverse censoring probability weights were estimated using a discrete-time failure model described in Section 3.1.1 Li et al., 2018. This is a logistic model that predicts the probability of censoring at each time point, conditional on not having been censored before that time point. The discretization is at the monthly level, thus there are 24 intervals in which one can be censored over $\tau = 24$ months. The resulting model is used to predict the probability being censored at the observed time, for each subject. The inverse of this probability is the weight used in the DR approach. We include all Age, Household income, Charlson Index, and FIGO stage as covariates in each model. Due to small cell counts, we combined FIGO stage II and II-NOS into a single category. In the discrete-time failure model, we include a fixed effect for each month, 1-24. Due to sparsity, we included month as a continuous covariate rather than categorical in this model. In Figure B.2, displays NMB estimates from this DR-SL model in gray, along with the EDP-GP estimates for reference. Note the larger uncertainty in the DR-SL model.

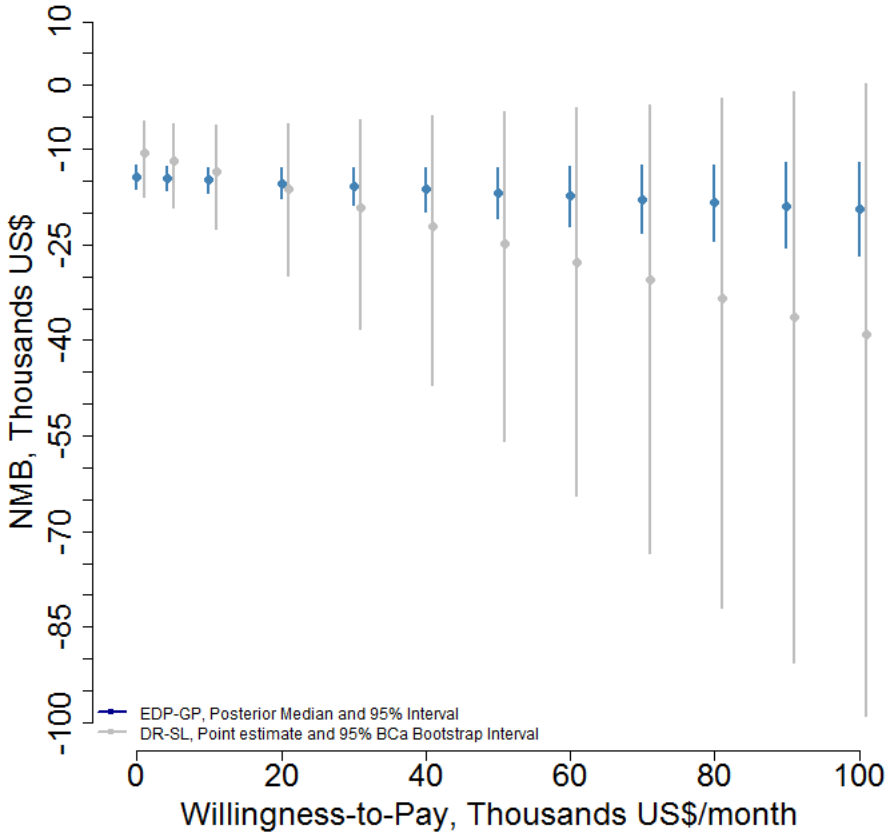


Figure B.2: NMB mean and 95% bootstrap intervals for various willingness to pay from the DR-SL model in gray. The EDP-GP estimates from Figure 3.3 are shown in blue for reference.

APPENDIX C

APPENDICES FOR CHAPTER 4

C.1. Identification of HTE

Here we identify the HTE in the point-treatment setting discussed in the paper. Recall the HTE is the average treatment effect within stratum v, $\Psi(v) = E[Y^1 \mid V = v] - E[Y^0 \mid V = v]$. Consider the term $E[Y^a \mid V = v]$ and now iterate expectation over W:

$$E[Y^a \mid V = v] = \int_{\mathcal{W}} E[Y^a \mid W, V = v] dP_v(W)$$

Now we assume conditional ignorability. Specifically that within stratum v, once we condition on confounders W, treatment assignment is independent of potential outcome, $Y^a \perp A \mid W, V = v$. This implies that $E[Y^a \mid W, V = v] = E[Y^a \mid A = a, W, V = v]$,

$$E[Y^a \mid V = v] = \int_{\mathcal{W}} E[Y^a \mid A = a, W, V = v] dP_v(W)$$

Now, we assume consistency. That is, the outcome actually observed under treatment assignment $A = a$ actually equals the outcome that would occur under treatment $A = a$, i.e. $Y^a = Y$. This would be violated if, for instance, there is non-adherence to treatment assignment. This yields,

$$E[Y^a \mid V = v] = \int_{\mathcal{W}} E[Y \mid A = a, W, V = v] dP_v(W)$$

So we have identified each term of $\Psi(v)$ as a regression averaged over $P_v(W) = P(W \mid V = v)$. Note that we implicitly make a positivity and non-adherence assumption. By conditioning on $A = a$ within W and V, we are assuming that treatment probability is bounded $0 < P(A = 1 \mid W, V = v) < 1$ or else we would be conditioning on a zero-probability even. This is also known as "overlap". Causally, it would suggest that there is some level and W within stratum V where we only observed patients assigned to one of the two treatments. We cannot estimate a causal effect in this region of the data without (likely incorrect) extrapolation. Moreover, for a particular sample we have assumed that each subjects potential outcome $Y_i^{a_i}$ is unaffected by others' treatment assignment. If subject j's treatment assignment impacts subject i's potential outcome, then we would have had to index

the potential outcome with this treatment as well, $Y_i^{a_i,a_j}$.

C.1.1. Posterior Derivations

Here we provide a derivation of the posterior distribution of each P_v using Dirichlet Distributions - the finite-dimensional analogue of the Dirichlet Process. This is to supplement the conjugacy results used in the main text. Suppose our model for the conditional covariate distribution, $P_v(W) = P(W \mid V = v)$, is

$$P_v(W \mid \pi^v) = \sum_{i=1}^{n} \pi_i^v \cdot \delta_{W_i}(W)$$

We have K such distributions for each of the K levels of V. Consider the Dirichlet prior on each $\pi^v = (\pi_1^v, \pi_2^v, \ldots, \pi_n^v)$ conditional on the $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ and α mentioned in the text,

$$\pi^v \mid \pi, \alpha \sim Dir(\alpha\pi)$$

Note, we could do everything in terms of v-specific concentration parameters, but use a common α for compactness. Now place Dirichlet hyperprior on π:

$$\pi \mid \gamma \sim Dir(\gamma 1_n)$$

Note that the HBB corresponds to setting $\gamma = 0$ and that α is user-specified but we will leave γ as it is for now. So the joint posterior is

$$p(\pi^1, \pi^2, \ldots \pi^K, \pi \mid \alpha, \gamma, W, V) \propto \Big\{ \prod_{v=1}^{K} \frac{\Gamma(\sum_{i=1}^{n} \alpha\pi_i)}{\prod_{i=1}^{n} \Gamma(\alpha\pi_i)} \prod_{i=1}^{n} (\pi_i^v)^{\alpha\pi_i + \delta_v(V_i) - 1} \Big\} p(\pi \mid \gamma) \tag{C.1}$$

The objective is to sample the π^v. To do this, we sample from the joint and simply ignore draws of π. Note that the joint can be expressed as a marginal posterior for π and independent conditional posteriors for π^v

$$p(\pi^1, \pi^2, \ldots \pi^K, \pi \mid \alpha, \gamma, W, V) = \{\prod_{v=1}^{K} p(\pi^v \mid \pi, \alpha, \gamma, W, V)\} p(\pi \mid \alpha, \gamma, W, V)$$

Thus to sample from the joint, we first sample π from the marginal posterior. Then conditional on π, we can sample the π^v independently. These are exactly Step 1 and 2, respectively, in the algorithm

of Section 4.3.1. We now derive this marginal posterior and then turn to the conditional posteriors of π^v. To get the marginal, integrate out each of the π^v in (C.1)

$$p(\pi \mid \alpha, \gamma, W, V) \propto \Big\{ \prod_{v=1}^{K} \int_{\Pi_v} \frac{\Gamma(\sum_{i=1}^{n} \alpha\pi_i)}{\prod_{i=1}^{n} \Gamma(\alpha\pi_i)} \prod_{i=1}^{n} (\pi_i^v)^{\alpha\pi_i + \delta_v(V_i) - 1} d\pi^v \Big\} p(\pi \mid \gamma)$$
$$\propto \Big\{ \prod_{v=1}^{K} \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha\pi_i)} \frac{\prod_{i \in S_v}^{n} \Gamma(\alpha\pi_i + 1) \prod_{i \notin S_v}^{n} \Gamma(\alpha\pi_i)}{\Gamma(\alpha + n_v)} \Big\} p(\pi \mid \gamma)$$

Above, Π_v is the n-dimensional simplex we integrate over. This result follows because the integral is over the kernel of a Dirichlet distribution, with concentration parameter vector $\alpha\pi_i + \delta_v(V_i)$ and recognizing that $\sum_{i=1}^{n} \alpha\pi_i = \alpha$ since π_i sum to 1. Continuing the derivation, we cancel like terms from the numerator and denominators and note that $\Gamma(\alpha\pi_i + 1) = \alpha\pi_i \Gamma(\alpha\pi_i)$. Therefore, $\frac{\Gamma(\alpha\pi_i + 1)}{\Gamma(\alpha\pi_i)} = \alpha\pi_i$ and we have

$$p(\pi \mid \alpha, \gamma, W, V) \propto \Big\{ \prod_{v=1}^{K} \frac{\Gamma(\alpha)\alpha^{n_v}}{\Gamma(\alpha + n_v)} \Big\} (\prod_{i=1}^{n} \pi_i) p(\pi \mid \gamma)$$

Now, note that in the last line the term in brackets is constant with respect to π, so we can eliminate it and maintain proportionality. Then, substituting the prior $p(\pi \mid \gamma = 0) = Dir(0_n) \propto \prod_{i=1}^{n} \pi_i^{-1}$,

$$p(\pi \mid \alpha, \gamma, W, V) \propto (\prod_{i=1}^{n} \pi_i) \prod_{i=1}^{n} \pi_i^{-1} \propto \prod_{i=1}^{n} \pi_i^{1-1}$$

This is the kernel of $Dir(1_n)$ - the posterior of Rubin's bootstrap. Thus, to draw from this marginal posterior, we can draw $\pi \sim Dir(1_n)$. This is the distribution we sample from in Step 1 of the algorithm in Section 4.3.1.

Now, the conditional posterior of each π^v conditional on π is much simpler. Just absorb all terms not involving π_i^vin (C.1) into the proportionality constant and we have

$$p(\pi^v \mid \pi, \alpha, \gamma, W, V) \propto \prod_{i=1}^{n} (\pi_i^v)^{\alpha\pi_i + \delta_v(V_i) - 1}$$

Which is proportional to a $\pi^v \sim Dir\Big(\alpha\pi_1 + \delta_v(V_1), \alpha\pi_2 + \delta_v(V_2), \ldots, \alpha\pi_n + \delta_v(V_n)\Big)$. This is the distribution we sample from in Step 2 of the algorithm in Section 4.3.1.

C.1.2. Simulation Details

Here we provide details for the simulation study in Section 4.4. In each setting, we simulate 1000 data sets with $n = 300$ subjects as follows. For $i = 1, \ldots, 300$

1. Simulate stratum allocation:

$$V_i \sim Multinom(1; \frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10})$$

   The parameter vectors gives the probability of assignment to stratum 1, 2, 3, and 4, respectively.

2. Simulate 10-dimensional confounder vector $W_i = (W_i^p)_{p=1:10}$,

$$W_i \mid V_i = v \sim p(W \mid V = v)$$

   The form of $p(W \mid V = v)$ varies with simulation setting and is specified below.

3. Simulate treatment assignment, A_i, from Bernoulli with probability

$$P(A = 1 \mid W_i, V_i = v) = expit(\eta_v + W_i'\beta)$$

4. Simulate binary outcome, Y_i, from a Bernoulli with probability

$$P(Y = 1 \mid W_i, V_i = v) = expit(-1 + \gamma_v + W_i'\theta + \alpha_v A_i)$$

Note in the above that W_i impacts both treatment assignment (via β) and outcome (via θ) - so it is a confounder. Similarly, V_i impacts both treatment assignment (via η_v) and outcome (via γ_v). Note that the conditional treatment effect, α_v, varies across stratum - so this is a complex scenario with treatment effect heterogeneity across strata. This yields a simulated data set $\{Y_i, A_i, W_i, V_i\}_{i=1:n}$. We simulate 1000 such data sets across four settings.

The covariate distribution $p(W \mid V)$ has a different family governed by different parameters in each of the four settings, $1 - 4$:

1. $W_i^p \sim N(0,1)$ for all $V = v$.

2. $W_i^p \mid V = v \sim N(\mu_v, 1)$ where $\mu_v \in \{-2, 0, 2, 4\}$ for $v = 1, \ldots 4$, respecting order. Marginal of V, the distribution of W is a location mixture of normals.

3. $W_i^p \mid V = v \sim Ber(p_v)$ where $p_v \in \{.8, .6, .4, .2\}$ for $v = 1, \ldots 4$, respecting order.

4. $W_i^p \mid V = v \sim Gam(shape = \frac{1}{2}\tau_v, rate = \frac{1}{2})$. Here $\tau_v \in \{8, 6, 4, 1\}$ for $v = 1, \ldots 4$, respecting order.

All settings share these simulation parameters:

- Set $\beta = \theta = (1, -1, 1, -1, 1, -1, 1, -1, 1, -1)$.
- Set $\eta_v \in (0, -.5, .5, .5)$ for $v = 1, \ldots, 4$ in order.
- $\gamma_v \in (-.1, -.5, .1, .5)$ for $v = 1, \ldots, 4$ in order.
- $\alpha_v \in (1, -1.5, 1, 1.5)$ for $v = 1, \ldots, 4$ in order.

Using each simulated dataset, we specify the following logistic regression

$$P(Y \mid A, W, V = v) = expit\Big(\omega_0 + \omega_v + W'\omega_W + \omega_v^* A\Big)$$

Normal priors with mean zero and standard deviation 3 were placed on each parameter. We obtain $M = 5000$ posterior samples $\{\omega_0, \omega_1^{(m)}, \ldots, \omega_4^{(m)}, \omega_W^{(m)}, \omega_1^{*(m)}, \ldots, \omega_4^{*(m)}\}_{m=1:M}$ after discarding the first 5000 draws as burn-in. Sampling was done via Hamiltonian Monte Carlo as implemented in Stan. These samples were combined with HBB as described in Section 4.3.1.

C.1.3. Data Analysis Details

Here we provide additional details about the data analysis in the main text. In the parametric Poisson model, we include the following covariates for each stratum except gynecological cancer.

- treatment: binary with one indicating proton.
- race: categorical with levels white, black, and other.

- sex: binary with one indicating male.
- insurance: categorical with levels medicare, private, and other.
- body-mass index: normalized.
- age: normalized
- charlson index: logged.

For gynecological cancer, there is no need to adjust for sex. We specify$N(0,1)$ priors on all covariates except in the following instances: in the models for E/G, brain, anal, and rectum, we use tighter $N(0,.1)$ priors on the other race coefficient. Similarly, for the P/D/H model we use a $N(0,.1)$ prior on other insurance. The tight priors are to regularize coefficients that explode due too little variation in insurance status or race in a particular stratum. Non-bayesian analyses typically omit such variables (equivalent to a prior that the coefficient is exactly 0), but we choose to include them with a tight prior around 0 as a compromise. Note that the $N(0,1)$ prior may seem overly informative, but on the log scale it is quite flat. It puts sufficient volume at incident rate ratios within $\exp(\pm 1.96)$ or within $(.14, 7.1)$.

For posterior sampling, we use hamiltonian monte carlo as implemented in Stan. We call Stan in R using the rstan package. For inference, we retain 10000 posterior draws after a 10000 burn-in. After obtaining these draws, we use HBB as described in Section 4.3.1.

For the BART model, we adjust for all of the same covariates. Draws of f_v under particular treatments were obtained using the BayesTree R package. We retain 1000 posterior draws for inference after discarding the first 1000 as burn-in. For the BART hyperpriors, we increase the power parameter from the default of 2 to 3. This is to favors more shallow trees which provides more regularization. After draws of f_v are obtained, we combine with HBB draws as described in Section 4.3.1.

Finally, we note that the effects in the gynecological cancer model, in particular, is highly variable. As there were only 4 subjects treated with proton therapy in this stratum and none of the four had events, this coefficient is not identifiable with data. This is manifest in the large interval in both the Poisson and BART models.

BIBLIOGRAPHY

Athey, S and Wager, S (2019). *Estimating Treatment Effects with Causal Forests: An Application*. arXiv: 1902.07409 [stat.ME].

Baio, G (2014). Bayesian models for cost-effectiveness analysis in the presence of structural zero costs. *Statistics in Medicine* 33.11, 1900–1913. ISSN: 10970258. DOI: 10.1002/sim.6074. arXiv: 1307.5243. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285321/pdf/sim0033-1900.pdf.

Bang, H and Tsiatis, AA (2000). Estimating medical costs with censored data. *Biometrika* 87.2, 329–343. ISSN: 00063444. DOI: 10.1093/biomet/87.2.329. URL: http://www.jstor.org/stable/2673467.

Barcella, W, De Iorio, M, Baio, G, and Malone-Lee, J (2016). A Bayesian nonparametric model for white blood cells in patients with lower urinary tract symptoms. *Electron. J. Statist.* 10.2, 3287–3309. DOI: 10.1214/16-EJS1177. URL: https://doi.org/10.1214/16-EJS1177.

Barrientos, AF and Pea, V (2020). Bayesian Bootstraps for Massive Data. *Bayesian Anal.* 15.2, 363–388. DOI: 10.1214/19-BA1155. URL: https://doi.org/10.1214/19-BA1155.

Baumann, BC, Mitra, N, Harton, JG, Xiao, Y, Wojcieszynski, AP, Gabriel, PE, Zhong, H, Geng, H, Doucette, A, Wei, J, O?Dwyer, PJ, Bekelman, JE, and Metz, JM (2020). Comparative Effectiveness of Proton vs Photon Therapy as Part of Concurrent Chemoradiotherapy for Locally Advanced Cancer. *JAMA Oncology* 6.2, 237–246. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2019.4889. eprint: https://jamanetwork.com/journals/jamaoncology/articlepdf/2757520/jamaoncology_baumann_2019_oi_190095.pdf. URL: https://doi.org/10.1001/jamaoncol.2019.4889.

Binder, DA (1978). Bayesian cluster analysis. *Biometrika* 65.1, 31–38. ISSN: 0006-3444. DOI: 10.1093/biomet/65.1.31. eprint: https://academic.oup.com/biomet/article-pdf/65/1/31/685711/65-1-31.pdf. URL: https://doi.org/10.1093/biomet/65.1.31.

Blackwell, D and MacQueen, JB (1973). Ferguson Distributions Via Polya Urn Schemes. *Ann. Statist.* 1.2, 353–355. DOI: 10.1214/aos/1176342372. URL: https://doi.org/10.1214/aos/1176342372.

Boatman, JA, Vock, DM, and Koopmeiners, JS (2020). Borrowing from Supplemental Sources to Estimate Causal Effects from a Primary Data Source. *arXiv preprint arXiv:2003.09680*.

Chipman, HA, George, EI, and McCulloch, RE (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4.1, 266–298. DOI: 10.1214/09-AOAS285. URL: https://doi.org/10.1214/09-AOAS285.

Dahl, DB (2006). "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model". In: *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, 201–218. DOI: 10.1017/CBO9780511584589.011.

Ding, P and Li, F (2018). Causal Inference: A Missing Data Perspective. *Statist. Sci.* 33.2, 214–237. DOI: 10.1214/18-STS645. URL: https://doi.org/10.1214/18-STS645.

Efron, B and Gong, G (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician* 37.1, 36–48. ISSN: 00031305. URL: http://www.jstor.org/stable/2685844.

Ferguson, TS (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* 1.2, 209–230. DOI: 10.1214/aos/1176342360. URL: https://doi.org/10.1214/aos/1176342360.

George, E, Laud, P, Logan, B, McCulloch, R, and Sparapani, R (2018). Fully Nonparametric Bayesian Additive Regression Trees. *arXiv preprint arXiv:1807.00068*.

Ghahramani, Z (2015). Probabilistic machine learning and artificial intelligence. *Nature* 521.7553, 452–459.

Ghosh, SK, Mukhopadhyay, P, and Lu, J-C (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference* 136.4, 1360 –1375. ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2004.10.008. URL: http://www.sciencedirect.com/science/article/pii/S0378375804004008.

Hahn, PR, Murray, JS, and Carvalho, CM (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *Bayesian Analysis*. Advance publication. DOI: 10.1214/19-BA1195. URL: https://doi.org/10.1214/19-BA1195.

Handorf, EA, Heitjan, DF, Bekelman, JE, and Mitra, N (2019). Estimating cost-effectiveness from claims and registry data with measured and unmeasured confounders. *Statistical Methods in Medical Research* 28.7. PMID: 29468944, 2227–2242. DOI: 10.1177/0962280218759137. eprint: https://doi.org/10.1177/0962280218759137. URL: https://doi.org/10.1177/0962280218759137.

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

Henderson, NC, Louis, TA, Rosner, GL, and Varadhan, R (2018). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics* 21.1, 50–68. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxy028. eprint: https://academic.oup.com/biostatistics/article-pdf/21/1/50/31564932/kxy028.pdf. URL: https://doi.org/10.1093/biostatistics/kxy028.

Hill, JL (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics* 20.1, 217–240. DOI: 10.1198/jcgs.2010.08162. eprint: https://doi.org/10.1198/jcgs.2010.08162. URL: https://doi.org/10.1198/jcgs.2010.08162.

Huang, Y (2002). Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association* 97.457, 318–327. ISSN: 01621459. DOI: 10.1198/016214502753479446. URL: https://www.tandfonline.com/action/journalInformation?journalCode=uasa20.

Jain, S and Neal, RM (2004). A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics* 13.1, 158–182. DOI: 10.1198/1061860043001. eprint: https://doi.org/10.1198/1061860043001. URL: https://doi.org/10.1198/1061860043001.

Keil, AP, Daza, EJ, Engel, SM, Buckley, JP, and Edwards, JK (2017). A Bayesian approach to the g-formula. *Statistical Methods in Medical Research* 0.0. PMID: 29298607. DOI: 10.1177/0962280217694665. eprint: https://doi.org/10.1177/0962280217694665. URL: https://doi.org/10.1177/0962280217694665.

Kim, C, Daniels, MJ, Marcus, BH, and Roy, JA (2017). A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* 73.2, 401–409. DOI: 10.1111/biom.12575. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12575. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12575.

Kleiner, A, Talwalkar, A, Sarkar, P, and Jordan, MI (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76.4, 795–816. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/24774569.

Li, J, Vachani, A, Epstein, A, and Mitra, N (2018). A doubly robust approach for costeffectiveness estimation from observational data. *Statistical Methods in Medical Research* 27.10, 3126–3138. ISSN: 14770334. DOI: 10.1177/0962280217693262.

Lin, DY (2003). Regression analysis of incomplete medical cost data. *Statistics in Medicine* 22.7, 1181–1200. ISSN: 02776715. DOI: 10.1002/sim.1377.

Lin, D (2000). Linear regression analysis of censored medical costs. *Biostatistics* 1.1, 35–47. ISSN: 1465-4644. DOI: 10.1093/biostatistics/1.1.35.

Lin, D, Feuer, E, Etzioni, R, and Wax, Y (1997). Estimating medical cost from incomplete data. *Biometrics* 53.2, 419–434. URL: http://dlin.web.unc.edu/files/2013/04/LinEA97.pdf.

Linero, AR, Sinha, D, and Lipsitz, SR (2018). Semiparametric Mixed-Scale Models Using Shared Bayesian Forests. *ArXiv e-prints*. arXiv: 1809.08521 [stat.ME].

Makela, S, Si, Y, and Gelman, A (2018). Bayesian inference under cluster sampling with probability proportional to size. *Statistics in Medicine* 37.26, 3849–3868. DOI: 10.1002/sim.7892. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7892. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7892.

Muller, P, Quintana, F, Jara, A, and Hanson, T (2015). *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics. Springer International Publishing. ISBN: 9783319189673. URL: https://books.google.com/books?id=eH5ErgEACAAJ.

Murphy, KP (2021). *Probabilistic Machine Learning: An introduction*. MIT Press. URL: probml.ai.

Neal, RM (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* 9.2, 249–265. DOI: 10.1080/10618600.2000.10474879. eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/10618600.2000.10474879. URL: https://amstat.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879.

Nethery, RC, Mealli, F, and Dominici, F (2019). Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *Ann. Appl. Stat.* 13.2, 1242–1267. DOI: 10.1214/18-AOAS1231. URL: https://doi.org/10.1214/18-AOAS1231.

Nieto-Barajas, LE and Walker, SG (2002). Markov Beta and Gamma Processes for Modelling Hazard Rates. *Scandinavian Journal of Statistics* 29.3, 413–424. DOI: 10.1111/1467-9469.00298. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9469.00298. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9469.00298.

Oganisian, A, Mitra, N, and Roy, JA (2020). A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *Biometrics*. DOI: 10.1111/biom.13244. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13244.

Oganisian, A and Roy, JA (2021). A practical introduction to Bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine* 40.2, 518–551. DOI: https://doi.org/10.1002/sim.8761.

Petersen, ML, Porter, KE, Gruber, S, Wang, Y, and Laan, MJ van der (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21.1. PMID: 21030422, 31–54. DOI: 10.1177/0962280210386207. eprint: https://doi.org/10.1177/0962280210386207. URL: https://doi.org/10.1177/0962280210386207.

Rasmussen, CE and Williams, CKI (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press. ISBN: 026218253X.

Robins, J (1986). A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling* 7.9, 1393 –1512. ISSN: 0270-0255. DOI: https://doi.org/10.1016/0270-0255(86)90088-6. URL: http://www.sciencedirect.com/science/article/pii/0270025586900886.

Robins, JM, Hernán, MA, and Brumback, B (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11.5, 551.

Rodrguez, CE and Walker, SG (2014). Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. *Journal of Computational and Graphical Statistics* 23.1, 25–45. DOI: 10.1080/10618600.2012.735624. eprint: https://doi.org/10.1080/10618600.2012.735624. URL: https://doi.org/10.1080/10618600.2012.735624.

Roy, J, Lum, KJ, and Daniels, MJ (2017). A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics* 18.1, 32–47. DOI: 10.1093/biostatistics/kxw029. eprint: /oup/backfile/content_public/journal/biostatistics/18/1/10.1093_biostatistics_kxw029/3/kxw029.pdf. URL: http://dx.doi.org/10.1093/biostatistics/kxw029.

Roy, J, Lum, KJ, Zeldow, B, Dworkin, JD, Re III, VL, and Daniels, MJ (2018). Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics* 74.4, 1193–1202. DOI: 10.1111/biom.12875. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12875. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12875.

Rubin, DB (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Ann. Statist.* 6.1, 34–58. DOI: 10.1214/aos/1176344064. URL: https://doi.org/10.1214/aos/1176344064.

Rubin, DB (1981). The Bayesian Bootstrap. *Ann. Statist.* 9.1, 130–134. DOI: 10.1214/aos/1176345338. URL: https://doi.org/10.1214/aos/1176345338.

Saarela, O, Stephens, DA, Moodie, EEM, and Klein, MB (2015). On Bayesian estimation of marginal structural models. *Biometrics* 71.2, 279–288. DOI: 10.1111/biom.12269. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12269. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12269.

Silverman, BW and Young, GA (1987). The Bootstrap: To Smooth or Not to Smooth? *Biometrika* 74.3, 469–479. ISSN: 00063444. URL: http://www.jstor.org/stable/2336686.

Stephens, M (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, 795–809.

Taddy, M, Gardner, M, Chen, L, and Draper, D (2016). A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business and Economic Statistics* 34.4, 661–672. DOI: 10.1080/07350015.2016.1172013. eprint: https://doi.org/10.1080/07350015.2016.1172013. URL: https://doi.org/10.1080/07350015.2016.1172013.

Teh, YW, Jordan, MI, Beal, MJ, and Blei, DM (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101.476, 1566–1581. DOI: 10.1198/016214506000000302. eprint: https://doi.org/10.1198/016214506000000302. URL: https://doi.org/10.1198/016214506000000302.

Wade, S, Dunson, DB, Petrone, S, and Trippa, L (2014). Improving Prediction from Dirichlet Process Mixtures via Enrichment. *J. Mach. Learn. Res.* 15.1, 1041–1071. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2627435.2638569.

Wang, C, Dominici, F, Parmigiani, G, and Zigler, CM (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* 71.3, 654–665. DOI: 10.1111/biom.12315. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12315. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12315.

Wang, S (1995). Optimizing the smoothed bootstrap. *Annals of the Institute of Statistical Mathematics* 47.1, 65–80.

Xu, D, Daniels, MJ, and Winterstein, AG (2018). A Bayesian nonparametric approach to causal inference on quantiles. *Biometrics* 74.3, 986–996. DOI: 10.1111/biom.12863. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12863. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12863.

Xu, Y, Müller, P, Wahed, AS, and Thall, PF (2016). Bayesian Nonparametric Estimation for Dynamic Treatment Regimes With Sequential Transition Times. *Journal of the American Statistical Association* 111.515. PMID: 28018015, 921–950. DOI: 10.1080/01621459.2015.1086353. eprint: https://doi.org/10.1080/01621459.2015.1086353. URL: https://doi.org/10.1080/01621459.2015.1086353.

Xu, Y, Scharfstein, D, Mller, P, and Daniels, M (2020). A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics*. kxaa008. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxaa008. eprint: https://academic.oup.com/biostatistics/advance-article-pdf/doi/10.1093/biostatistics/kxaa008/33012489/kxaa008.pdf. URL: https://doi.org/10.1093/biostatistics/kxaa008.

Zeldow, B, Lo Re III, V, and Roy, J (2019). A semiparametric modeling approach using Bayesian Additive Regression Trees with an application to evaluate heterogeneous treatment effects. *Ann. Appl. Stat.* 13.3, 1989–2010. DOI: 10.1214/19-AOAS1266. URL: https://doi.org/10.1214/19-AOAS1266.