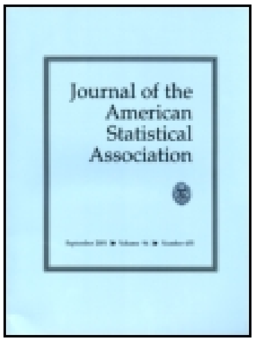

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Bayesian Nonparametric Estimation for Dynamic Treatment Regimes With Sequential Transition Times

Yanxun Xu, Peter Müller, Abdus S. Wahed & Peter F. Thall

To cite this article: Yanxun Xu, Peter Müller, Abdus S. Wahed & Peter F. Thall (2016) Bayesian Nonparametric Estimation for Dynamic Treatment Regimes With Sequential Transition Times, Journal of the American Statistical Association, 111:515, 921-950, DOI: 10.1080/01621459.2015.1086353

To link to this article: http://dx.doi.org/10.1080/01621459.2015.1086353

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, 921–950, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2015.1086353

Taylor & Francis
Taylor & Francis Group

Bayesian Nonparametric Estimation for Dynamic Treatment Regimes With Sequential Transition Times

Yanxun Xu[a], Peter Müller[b], Abdus S. Wahed[c], and Peter F. Thall[d]

[a]Division of Statistics and Scientific Computing, The University of Texas at Austin, Austin, TX, USA; [b]Department of Mathematics, The University of Texas at Austin, Austin, TX, USA; [c]Epidemiology Data Center, University of Pittsburgh, Pittsburgh, PA, USA; [d]M. D. Anderson Cancer Center, Houston, TX, USA

ABSTRACT

We analyze a dataset arising from a clinical trial involving multi-stage chemotherapy regimes for acute leukemia. The trial design was a 2 × 2 factorial for frontline therapies only. Motivated by the idea that subsequent salvage treatments affect survival time, we model therapy as a dynamic treatment regime (DTR), that is, an alternating sequence of adaptive treatments or other actions and transition times between disease states. These sequences may vary substantially between patients, depending on how the regime plays out. To evaluate the regimes, mean overall survival time is expressed as a weighted average of the means of all possible sums of successive transitions times. We assume a Bayesian nonparametric survival regression model for each transition time, with a dependent Dirichlet process prior and Gaussian process base measure (DDP-GP). Posterior simulation is implemented by Markov chain Monte Carlo (MCMC) sampling. We provide general guidelines for constructing a prior using empirical Bayes methods. The proposed approach is compared with inverse probability of treatment weighting, including a doubly robust augmented version of this approach, for both single-stage and multi-stage regimes with treatment assignment depending on baseline covariates. The simulations show that the proposed nonparametric Bayesian approach can substantially improve inference compared to existing methods. An R program for implementing the DDP-GP-based Bayesian nonparametric analysis is freely available at *www.ams.jhu.edu/ yxu70*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2014
Revised June 2015

KEYWORDS
Dependent Dirichlet process; Gaussian process; G-Computation; Inverse probability of treatment weighting; Markov chain Monte Carlo

1. Introduction

We analyze a dataset arising from a clinical trial involving multi-stage chemotherapy regimes for acute leukemia. The trial design was a 2×2 factorial for frontline therapies only. However, motivated by the idea that subsequent salvage therapies affect survival time, Wahed and Thall (2013) modeled and analyzed treatments in the trial as a dynamic treatment regime (DTR), that is, an alternating sequence of treatments or other actions and transition times between disease states. We propose a Bayesian nonparametric (BNP) approach for evaluating such DTRs in which the outcome at each stage is a random transition time between two disease states. The final overall survival (OS) time outcome of primary interest is the sum, T, of a sequence of transition times. The actually observed sequence is determined by the way that a patient's treatment regime plays out, and the mean of T may be expressed as an appropriately weighted average over all possible sequences of event times. Our proposed BNP methodology for estimating the mean of T is based on the idea of Robins' G-computation (Robins 1986, 1987).

An algorithm commonly used by oncologists in chemotherapy of solid tumors is to choose the patient's initial (frontline) treatment based on his/her baseline covariates, continue as long as the patient's disease is stable, switch to a different chemotherapy (salvage) if progressive disease (P) occurs, stop chemotherapy if the tumor is brought into complete or partial remission (C), and begin salvage if P occurs at some time after C. There are many elaborations of this in oncology, including multiple attempts at salvage therapy, use of consolidation therapy for patients in remission, suspension of therapy if severe toxicity is observed, or inclusion of radiation therapy or surgery in the regime. Another important application of this general adaptive structure occurs in treatment regimes for psychological disorders or drug addiction. For example, in treatment of schizophrenia one may replace P by a psychotic episode or other worsening of the subject's psychological status, C by a specified improvement in mental status, and death by a psychological breakdown severe enough to require hospitalization.

Denote the action at stage ℓ of the DTR by Z^{ℓ}, which may be a treatment or a decision to delay or terminate therapy. Here, stage refers to the decision point in the DTR—that is, the choice of frontline and possible salvage therapies. At each stage, one observes a disease state s_{ℓ}, such as P, C, or death (D). Let $T^{(j,r)}$ denote the transition time from disease state j to state r, with $j = 0$ the patient's initial disease status. See Figure 1 for an example (details of which will be provided later) with up to $n_{\text{stage}} = 3$ stages, $n_{\text{state}} = 4$ disease states, and a total of $n_T = 7$ different transition times. Because the actions are adaptive, the actual number of stages and observed transition times vary between patients depending on how the specific treatment-outcome sequence plays out.

CONTACT Peter Müller pmueller@math.utexas.edu Department of Mathematics, The University of Texas at Austin, Austin, TX 78712.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*

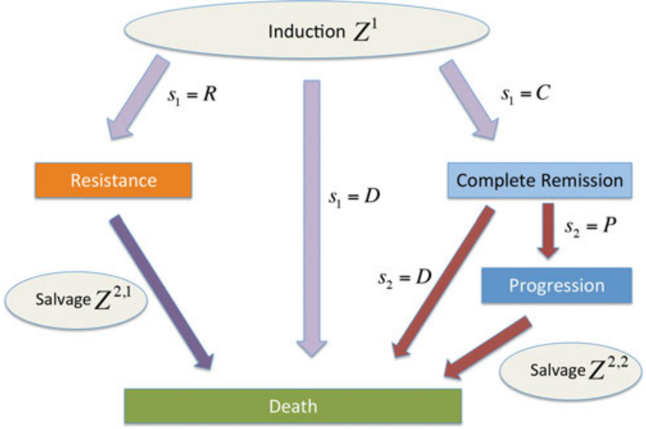


Figure 1. The scheme.

Formally, a DTR is the sequence $\boldsymbol{Z} = (Z^1, Z^2, \ldots)$, where each Z^ℓ is an adaptive action based on the patient's history $\mathcal{H}^{\ell-1}$ of previous treatments and transition times, and $\mathcal{H}^0$ is the patient's baseline covariate vector. One possible treatment-outcome sequence is $(\mathcal{H}^0, Z^1, T^{(0,C)}, Z^2, T^{(C,D)})$, in which the initial chemotherapy Z^1 was chosen based on $\mathcal{H}^0$, complete remission (C) was achieved, Z^2 was chosen based on $\mathcal{H}^1 = (\mathcal{H}^0, Z^1, T^{(0,C)})$. In this case, Z^2 would be consolidation therapy given to keep the patients in remission, that is, prevent relapse, although consolidation treatments were not included in the dataset that we will analyze. OS time is $T = T^{(0,C)} + T^{(C,D)}$. In this case, $s_1 = C$ and $s_2 = D$. Similarly, a patient brought into remission who later suffers progressive disease has sequence $(\mathcal{H}^0, Z^1, T^{(0,C)}, T^{(C,P)}, Z^2, T^{(P,D)})$ and $T = T^{(0,C)} + T^{(C,P)} + T^{(P,D)}$. We will apply BNP methods to estimate the conditional distributions of the transition times given the most recent histories, with the goal to estimate the mean of T for each possible DTR. This also will include estimates given specific baseline covariates, for so-called "individualized" therapy. Key elements of our proposed approach are quantification of all sources of uncertainty and prediction of T under a reasonable set of viable counterfactual DTRs (Wang et al. 2012). BNP methods have been used in estimating regime effects by Hill (2011) and Karabatsos and Walker (2012). Hill (2011) focused on modeling outcomes flexibly using Bayesian additive regression trees (BART), which required less assumptions in model fitting. However, the uncertainty of BART increases dramatically when there is complete treatment-subgroup confounding, and hence limited empirical counterfactuals, which often occurs in causal inference. Karabatsos and Walker (2012) proposed a nonparametric mixture model with a stick-breaking prior for the probability of treatment assignment to provide a more accurately estimated propensity score in the inverse probability of treatment weighting (IPTW) method.

Since all elements of a DTR may affect T, the clinically relevant problem is optimizing the entire regime, rather than the treatment at one particular stage. Most clinical trials or data analyses attempt to reduce variability by focusing on one stage of the actual DTR, usually frontline or first salvage treatment, or by combining stages in some manner. This often misrepresents actual clinical practice, and consequently conclusions may be very misleading. For example, an aggressive frontline cancer chemotherapy may maximize the probability of C, but it may cause so much immunologic damage that any salvage treatment given after rapid relapse, that is, short $T^{(C,P)}$, may be unlikely to achieve a second remission. In contrast, a milder induction treatment may be suboptimal to eradicate the tumor, but it may debulk the tumor sufficiently to facilitate surgical resection. Such synergies may have profound implications for clinical practice, especially because effects of multi-stage treatment regimes often are not obvious and may seem counter-intuitive. Physicians who have not been provided with an evaluation of the composite effects of entire regimes on the final outcome may unknowingly set patients on pathways that include only inferior regimes.

A major practical advantage of BNP models is that they often provide better fits to complicated data structures than can be obtained using parametric model-based methods. In the case study that we analyze here, leukemia patients were randomized among initial chemotherapy treatments but not among later salvage therapies, and the BNP model provides a good fit for each transition time distribution conditional on previous history. Failure to randomize patients in treatment stages after the first is typical in clinical trials, most of which ignore all but the first stage of therapy. In contrast, sequential multi-arm randomized treatment (SMART) designs, wherein patients are rerandomized at stages after the first, have been used in oncology trials (Thall, Millikan, and Sung 2000; Thall et al. 2007a, 2007b; Wang et al. 2012), and are being used increasingly in trials to study multi-stage adaptive regimes for behavioral or psychological disorders (Dawson and Lavori 2004; Murphy, Collins, and Rush 2007; Murphy et al. 2007; Connolly and Bernstein 2007).

While rerandomization is desirable, it is not commonly done and inference has to adjust for this lack of randomization. A wide array of methods have been proposed for evaluating DTRs from observational data and longitudinal studies, beginning with the seminal articles by Robins (1986, 1987, 1989, 1997) on G-estimation of structural nested models. Additional references include applications to longitudinal data in AIDS (Hernán, Brumback, and Robins 2000), inverse probability of treatment weighted (IPTW) estimation of marginal structural models (Murphy, Van Der Laan, and Robins 2001; van der Laan and Petersen 2007; Robins, Orellana, and Rotnitzky 2008), augmented IPTW (AIPTW) (Tsiatis 2007; Zhao et al. 2015), G-estimation for optimal DTRs (Murphy 2003; Robins 2004), and a review by Moodie, Richardson, and Stephens (2007). A variety of methods have been developed to evaluate DTRs from clinical trials (Lavori and Dawson 2000; Thall, Sung, and Estey 2002; Murphy 2005; Goldberg and Kosorok 2012; Zajonc 2012). For survival analysis, Lunceford, Davidian, and Tsiatis (2002) introduced ad hoc estimators for the survival distribution and mean restricted survival time under different treatment policies. These estimators, although consistent, were inefficient and did not exploit information from auxiliary covariates. Wahed and Tsiatis (2006) derived more efficient, easy-to-compute estimators that included auxiliary covariates for the survival distribution and related quantities of DTRs. Their estimators compared DTRs using data from a two-stage randomized trial, in which two options were available for both stages and the second-stage treatment assignments were determined by randomization. However, these estimators must be adapted for more general or more complicated designs that permit various numbers of treatment options at each stage and involve the scenarios where second-stage treatment is not randomized, but rather is determined by the attending physicians.

For settings where the DTR's final overall time, such as survival time, is the sum of a sequence of transition times, our proposed BNP approach employs a nonparametric survival regression model for each transition time conditional on the most recent history of actions and outcomes. We assume a dependent Dirichlet process prior with Gaussian process base measure (DDP-GP), and summarize a joint posterior by Markov chain Monte Carlo (MCMC) simulation. To address the important issue that Bayesian analyses depend on prior assumptions, we provide guidelines for using empirical Bayes methods to establish prior hyperparameters. Posterior analyses include estimation of posterior mean overall outcome times and credible intervals for each DTR.

The rest of the article is organized as follows. In Section 2, we review the motivating study, and give a brief review of DTRs in settings with successive transition times in Section 3. We present the DDP-GP model in Section 4. A simulation study of the BNP approach in single-stage and multi-stage regimes, with comparison to frequentist IPTW and AIPTW, is summarized in Section 5. We reanalyze the leukemia trial data in Section 6, and close with brief discussion in Section 7.

2. A Study of Multi-Stage Chemotherapy Regimes for Acute Leukemia

Our case study was a clinical trial conducted at The University of Texas M.D. Anderson Cancer Center to evaluate chemotherapies for acute myelogenous leukemia (AML) or myelo-dysplastic syndrome (MDS). Patients were randomized fairly among four frontline combination chemotherapies for remission induction: fludarabine + cytosine arabinoside (ara-C) plus idarubicin (FAI), FAI + all-trans-retinoic acid (ATRA), FAI + granulocyte colony stimulating factor (GCSF), and FAI + ATRA + GCSF. The goal of induction therapy for AML/MDS was to achieve complete remission (C), a necessary but not sufficient condition for long-term survival. Patients who did not achieve C, or who achieved C but later relapsed, were given salvage treatments as another attempt to achieve C. Following conventional clinical practice, patients were not randomized among salvage therapies, which instead were chosen by the attending physicians based on clinical judgment. Since there were many types of salvage, these are broadly classified into two categories as either containing high dose ara-C (HDAC) or other. This dataset was analyzed initially using conventional methods (Estey et al. 1999), including logistic regression, Kaplan–Meier estimates, and Cox model regression, including comparisons of the induction therapies in terms of OS that ignored possible effects of salvage therapies.

Figure 1 illustrates the actual possible therapeutic pathways and outcomes of the patients during the trial, which is typical of chemotherapy for AML/MDS. Death might occur (1) during induction therapy, (2) following salvage therapy if the disease was resistant to induction, (3) during C, or (4) following disease progression after C. Wahed and Thall (2013) reanalyzed the data from this trial by accounting for the structure in Figure 1, and identified 16 DTRs including both frontline and salvage therapies. To correct for bias due to the lack of randomization in estimating the mean OS times, they used both IPTW (Robins and Rotnitzky 1992) and G-computation based on a frequentist likelihood. In the G-computation, for each transition time they first fit accelerated failure time (AFT) regression models using Weibull, exponential, log-logistic, or lognormal distributions, and chose the distribution having smallest Bayes information criterion (BIC). They then performed likelihood-based G-computation by first fitting each conditional transition time distribution regressed on patient baseline covariates and previous transition times, and then averaging over the empirical covariate distribution.

Like Wahed and Thall, the primary goal of our analyses of the AML/MDS dataset is to estimate mean OS and determine the optimal regime. We build on their approach by replacing the parametric AFT models for transition times with the DDP-GP model. We also demonstrate the usefulness of the BNP regression model for G-computation in simulation studies of single-stage and multi-stage regimes in which treatment assignments depend on patient covariates.

3. Dynamic Regimes with Stochastic Transition Times

The case study involves more complicated structure than a stylized linear sequential study, as often is assumed in articles on DTRs that focus on basic methodology. We introduce the following notation to accommodate this more complex structure. Denote the set of possible disease states by $\{0, 1, \dots, n_{\text{state}}\}$, with 0 denoting the patient's initial state before receiving the first treatment. The pairs of states $(s_{\ell-1}, s_\ell)$ for which a transition $s_{\ell-1} \to s_\ell$ is possible at stage ℓ of the patient's therapy depend on the particular regime. Here, $s_0 = 0$ refers to the patient's initial state, before start of therapy. We will identify specific states using letters such as P, C, etc., as in the earlier examples, to replace the generic integers. For example, in cancer therapy, $s_{\ell-1} \to C$ means that a patient's disease has responded to treatment, $P \to D$ means a patient with progressive disease has died, and of course $D \to s_\ell$ is impossible. We denote the transition time from state $s_{\ell-1}$ to state s_ℓ in stage ℓ of treatment by $T^{(s_{\ell-1},s_\ell)}$, for $\ell = 1, \dots, n_{\text{stage}}$, the maximum number of stages in the DTR. In general, it might be necessary to add a third index to indicate the stage ℓ when the same transitions are possible in multiple stages. However, in our case study no ambiguity arises by simply writing $T^{(r,s)}$. To simplify notation for the transition time distributions, we denote the history of all covariates, treatments, and previous transition times through ℓ stages, before observation of $T^{(s_{\ell-1},s_\ell)}$ but including the stage ℓ action Z^ℓ by $\boldsymbol{x}^\ell = (\mathcal{H}^{\ell-1}, Z^\ell) = (\boldsymbol{x}^0, Z^1, T^{(s_0,s_1)}, \dots, T^{(s_{\ell-1},s_\ell)}, Z^\ell)$, with $\boldsymbol{x}^0 = \mathcal{H}^0$. Thus, a DTR is $\boldsymbol{Z} = (Z^1, Z^2, \dots)$, a sequence of actions for all possible stages. For example, in the leukemia trial (Figure 1), Z^1 might be FAI+ATRA given as frontline therapy, followed by salvage therapies $Z^2 =$ salvage with high dose ara-C if the disease is resistant to induction, and $Z^3 =$ other salvage if the patient first achieves a complete remission (C) but he later suffers progressive disease (P).

In the leukemia trial, the three possible outcomes following induction chemotherapy, C, R, and D, are competing risks. Thus, only one of the transition times, $T^{(0,C)}$, $T^{(0,R)}$, or $T^{(0,D)}$, is observed for each patient. The distribution of s_1 is determined by these three transition times. For example, the probability of C is

$$\begin{aligned}\Pr(s_1 = C \mid \boldsymbol{x}^0, Z^1) \\ = \Pr\left[T^{(0,C)} < \min\{T^{(0,R)}, T^{(0,D)}\} \mid \boldsymbol{x}^0, Z^1\right].\end{aligned}$$

This could be made explicit by including the states in the notation for $\boldsymbol{x}^l$. We chose not to do this for notational parsimony.

When no meaning is lost, we will further simplify notation and use a single running index on the transition times, and write $T^{(s_{\ell-1}, s_\ell)}$ as T^k, where $k = 1, \ldots, n_T$ is a running index of all possible state transitions. For example, in Figure 1 we have up to $n_{\text{stage}} = 3$ stages and $n_T = 7$ possible transitions. Similarly, we will write $\boldsymbol{x}^k$ for the corresponding covariate vector. Our use of a single index to identify stage is a slight abuse of notation since, for example, the actual second stage of therapy might differ depending on the sequence of outcomes. For example, stage 2 treatment Z^2 of a patient with sequence $(\boldsymbol{x}^0, Z^1, T^{(0,R)}, Z^2)$ is first salvage for resistant disease during induction with Z^1, while stage 3 treatment Z^3 of a patient with sequence $(\boldsymbol{x}^0, Z^1, T^{(0,C)}, T^{(C,P)}, Z^3)$ is first salvage for progressive disease after achieving response initially with Z^1. This latter example could be elaborated if, under a different regime, consolidation therapy, Z^2, was given for patients who enter C, in which case the sequence would be $(\boldsymbol{x}^0, Z^1, T^{(0,C)}, Z^2, T^{(C,P)}, Z^3)$.

Below, we will develop a general BNP model for all possible conditional distributions $p(T^k \mid \boldsymbol{x}^k)$. For any transition index k, let $\mathcal{R}^k$ denote the risk set, f^k the probability density function, and $\bar{F}^k$ the survival function of the transition time, δ_i^k is a censoring indicator with $\delta_i^k = 1$ if patient i is not censored and $\delta_i^k = 0$ if censored, and V_i^k is the observed time to the next state or censoring for patient i in risk set $\mathcal{R}^k$. For example, in the leukemia trial consider the transition $(0, R)$, corresponding to the single running index $k = 1$. The risk set is $\mathcal{R}^1 = \mathcal{R}^{(0,R)} = \{1, \ldots, n\}$. Let U_i denote the time from the start of induction to last followup for patient i. Then $\delta_i^1 = 1$ if $T_i^1 = \min(U_i, T_i^1)$ and the observed time for patient i is $V_i^1 = \min(T_i^{(0,D)}, T_i^{(0,R)}, T_i^{(0,C)}, U_i)$ since C, R, and D are competing risks. The likelihood for all possible sequences of treatments and transition times through n_T transitions is the product

$$\mathcal{L} = \prod_{k=1}^{n_T} \prod_{i \in \mathcal{R}^k} f^k(V_i^k \mid \boldsymbol{x}_i^k)^{\delta_i^k} \bar{F}^k(V_i^k \mid \boldsymbol{x}_i^k)^{1-\delta_i^k}. \qquad (1)$$

The overall time for any counterfactual sequence of transition times is the sum $T = \sum_{k=1}^{n_T} T^k$. Our goal is to estimate the mean of T for each possible $\boldsymbol{Z}$. Specific details of the likelihood are given in the Appendix.

4. A Nonparametric Bayesian Model for DTR

4.1. DDP and Gaussian Process Prior

Our motivation for using the BNP model described in this section is that it is highly robust and has full support. To specify the BNP model, we denote $Y^k = \log(T^k)$ and write the distribution of $[Y^k \mid \boldsymbol{x}^k]$ as $F^k(\cdot \mid \boldsymbol{x}^k)$. For convenience, we will refer to $\boldsymbol{x}^k$ as " covariates."We construct a BNP survival regression model for $F^k(\cdot \mid \boldsymbol{x}^k)$ by successive elaborations, starting with a model for a discrete random distribution $G^k(\cdot)$. We then use a Gaussian kernel to extend this to a prior for a continuous random distribution $F^k(\cdot)$, and finally endow the kernel means with a regression structure by expressing them as functions of $\boldsymbol{x}^k$. The latter construction extends F^k to a family $\{F^k(\cdot \mid \boldsymbol{x}^k)\}$, indexed by $\boldsymbol{x}^k$. The construction of $G^k(\cdot)$ and $F^k(\cdot)$ is outlined briefly below, by way of a brief review of BNP models. In the end, we will only use the last model $\{F^k(\cdot \mid \boldsymbol{x}^k)\}$, which we use as a sampling model for Y^k. See, for example, Müller and Mitra (2013) and Müller and Rodriguez (2013) for more extensive reviews of BNP inference. In the following discussion we temporarily drop the superindex ^{k}.

The Dirichlet process (DP) prior was first proposed by Ferguson (1973) as a probability distribution on a measurable space of probability measures. The DP is indexed by two hyperparameters, a base measure, G_0, and a precision parameter, $\alpha > 0$. If a random distribution G follows a DP prior, we denote this by $G \sim \text{DP}(\alpha, G_0)$. Denoting a beta distribution by $\text{Be}(a, b)$, if $G \sim \text{DP}(\alpha, G_0)$ then $G(A) \sim \text{Be}\{\alpha G_0(A), \alpha[1 - G_0(A)]\}$ for any measurable set A, and in particular $E\{G(A)\} = G_0(A)$. Let $\delta(\theta)$ denote a point mass at θ. Sethuraman (1994) provided a useful representation of the DP as $G = \sum_{h=0}^{\infty} w_h \delta(\theta_h)$, where $\theta_h \overset{\text{iid}}{\sim} G_0$, and the weights w_h are generated sequentially from rescaled beta distributions as $w_h/(1 - \sum_{r=1}^{h-1} w_r) \sim \text{Be}(1, \alpha)$, the so-called "stick-breaking" construction. The discrete nature of G is awkward in many applications. A DP mixture model extends the DP model by replacing each point mass $\delta(\theta_h)$ with a continuous kernel centered at θ_h. Without loss of generality, we will use a normal kernel. Let $N(\cdot;\ \mu, \sigma)$ denote a normal kernel with mean μ and standard deviation σ. The DP mixture model assumes

$$G = \sum_{h=0}^{\infty} w_h N(\cdot\,; \theta_h, \sigma). \qquad (2)$$

The use and interpretation of (2) is very similar to that of a finite mixture of normal models. In practical applications, the sum in (2) is often truncated at a reasonable finite value. This model is useful for density estimation under iid sampling from an unknown distribution, and it provides good fits to a wide variety of datasets because a mixture of normals can closely approximate virtually any distribution (Ishwaran and James 2001).

To include the regression on covariates that we will need for the survival model of each conditional transition time distribution, $F^k(\cdot \mid \boldsymbol{x}^k)$, we extend the DP mixture to a dependent DP (DDP), which was first proposed by MacEachern (1999). The basic idea of a DDP is to endow each θ_h^k with additional structure that specifies how it varies as a function of covariates $\boldsymbol{x}^k$. Writing this regression function as $\theta_h^k(\boldsymbol{x}^k)$ for the argument in each summand in (2), and returning to the conditional transition time distributions, we assume that

$$F^k(y \mid \boldsymbol{x}^k) = \sum_{h=0}^{\infty} w_h^k N(y;\ \theta_h^k(\boldsymbol{x}^k),\ \sigma^k). \qquad (3)$$

This form of the DDP, which includes both the convolution with a normal kernel and functional dependence on covariates, provides a very flexible regression model.

To complete our specification of the DDP, we will assume that the $\theta_h^k(\cdot)$'s are independent realizations from a Gaussian process (GP) prior. The GP was first popularized by O'Hagan and Kingman (1978) in Bayesian inference for a random function (unrelated to the use in a DDP prior). For more recent discussions see, for example, Rasmussen and Williams (2006), Neal (1995), and Shi et al. (2007). Temporarily suppressing the transition superindex ^{k} and running index $_h$ in (3), a GP is a stochastic process $\theta(\cdot)$ in which $(\theta(\boldsymbol{x}_1), \dots, \theta(\boldsymbol{x}_n))$ has a multivariate normal distribution with mean vector $(\mu(\boldsymbol{x}_1), \dots, \mu(\boldsymbol{x}_n))$ and $(n \times n)$ covariance matrix with (i, j) element $C(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for any set of $n \geq 1$ covariate vectors $\boldsymbol{x}_i$. We denote this by $\theta(\boldsymbol{x}) \sim \text{GP}(\mu, C)$.

We use the GP prior to define the dependence of $\theta_h^k(\boldsymbol{x}^k)$ as a function of $\boldsymbol{x}^k$ by assuming $\{\theta_h^k(\boldsymbol{x}_k)\} \sim \text{GP}(\mu_h^k, C^k)$, as a function of $\boldsymbol{x}_k$, for fixed h. That is, there is a separate GP for each term indexed by h in (3). We will refer to the DDP with a convolution using a normal kernel and a GP prior on the normal kernel means as a DDP-GP model. While the mean and covariance processes of the GP can be quite general, in practice, $C^k(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k)$ is often parameterized as a function $C(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k;\ \xi^k)$, where ξ^k is a vector of hyperparameters, and the mean function is indexed similarly by hyperparameters $\boldsymbol{\beta}_h^k$ and written as $\mu_h^k(\boldsymbol{x}^k;\ \boldsymbol{\beta}_h^k)$. In the DTR setting, since each covariate vector $\boldsymbol{x}^k$ is a history, its entries can include baseline covariates, transition times, and indicators of previous treatments or actions. To obtain numerically reasonable parameterizations of the GP functions C^k and μ_h^k, we standardize numerical-valued covariates such as age. We now have

$$\{\theta_h^k(\boldsymbol{x}^k)\} \sim \text{GP}\left(\mu_h^k(\cdot), C^k(\cdot, \cdot)\right) \qquad h = 1, 2, \dots$$

To specify the form of μ_h^k and C^k, let $i = 1, 2, \dots,$ index patients, so that $\boldsymbol{x}_i^k$ is the history of patient i at transition k, and define the indicator $\delta_{ij} = I(i = j) = 1$ if $i = j$ and 0 otherwise. We model the mean function $\mu_h^k(\cdot)$ as a linear regression, by assuming that

$$\mu_h^k\left(\boldsymbol{x}_i^k;\ \boldsymbol{\beta}_h^k\right) = \boldsymbol{x}_i^k \boldsymbol{\beta}_h^k. \tag{4}$$

For patients i and j, we assume that the covariance process takes the form

$$C^k(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k) = \exp\left\{-\sum_{m=1}^{M^k} (x_{im}^k - x_{jm}^k)^2\right\} + \delta_{ij} J^2, \quad i, j = 1, \dots, n, \tag{5}$$

where M^k is the number of covariates at transition k and J is the variance on the diagonal reflecting the amount of jitter (Bernardo, Berger, and Smith 1999), which usually takes a small value (e.g., $J = 0.1$). There are no further hyperparameters ξ^k to index the covariance function. For binary covariates, the quadratic form in (5) reduces to counting the number of binary covariates in which two patients differ. If desired, additional hyperparameters could be introduced in (5) to obtain more flexible covariance functions. However, in practice this form of the covariance matrix yields a strong correlation for observations on patients with very similar $\boldsymbol{x}^k$, and has been adopted widely (Williams 1998).

Combining all of these structures, we denote the model for the conditional distribution of the kth transition time as $F^k \sim \text{DDP-GP}\left\{\{\mu_h^k\}, C^k; \alpha^k, \{\boldsymbol{\beta}_h^k\}, \sigma^k\right\}$, recalling that the weights of the DDP are generated sequentially as $w_h^k/(1 - \sum_{r=1}^{h-1} w_r^k) \sim \text{Be}(1, \alpha^k)$. For later reference we state the full model. For $k = 1, \dots, n_T$

$$p\left(y_i^k \mid \boldsymbol{x}_i^k, F^k\right) = F^k\left(y_i^k \mid \boldsymbol{x}_i^k\right)$$
$$F^k \sim \text{DDP-GP}\left\{\{\mu_h^k\}, C^k; \alpha^k, \{\boldsymbol{\beta}_h^k\}, \sigma^k\right\}. \tag{6}$$

4.2. Determining Prior Hyperparameters

As priors for $\boldsymbol{\beta}_h^k$ in (6) we assume $\boldsymbol{\beta}_h^k \sim N(\boldsymbol{\beta}_0^k, \Sigma_0^k)$ for each transition time k, with $(\sigma^k)^{-2} \overset{\text{iid}}{\sim} \text{Ga}(\lambda_1, \lambda_2)$ and $\alpha^k \overset{\text{iid}}{\sim} \text{Ga}(\lambda_3, \lambda_4)$.

To apply the DDP-GP model, one must first determine numerical values for the fixed hyperparameters $\{\boldsymbol{\beta}_0^k, \Sigma_0^k, k = 1, 2, \dots\}$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. This is a critical step. These numerical hyperparameter values must facilitate posterior computation, and they should not introduce inappropriately strong information into the prior that would invalidate posterior inferences. With this in mind, the hyperparameters $(\boldsymbol{\beta}_0^k, \Sigma_0^k)$ for the kth transition time covariate effect distribution may be obtained via empirical Bayes by doing preliminary fits of a lognormal distribution $Y^k = \log(T^k) \sim N(\boldsymbol{x}^k \boldsymbol{\beta}_0^k, \sigma_0^k)$ for each transition k. Similarly, we assume a diagonal matrix for Σ_0^k with the diagonal values also obtained from the preliminary fit of the lognormal distribution. Once an empirical estimate of σ^k is obtained, one can tune (λ_1, λ_2) so that the prior mean of σ^k matches the empirical estimate and the variance equals 1 or a suitably large value to ensure a vague prior. Finally, information about α^k typically is not available in practice. We use $\lambda_3 = \lambda_4 = 1$.

This approach works in practice because the parameter $\boldsymbol{\beta}_0^k$ specifies the prior mean for the mean function of the GP prior, which in turn formalizes the regression of T^k on the covariates $\boldsymbol{x}^k$, including treatment selection. The imputed treatment effects hinge on the predictive distribution under that regression. Excessive prior shrinkage could smooth away the treatment effect that is the main focus. The use of an empirical Bayes type prior in the present setting is similar to empirical Bayes priors in hierarchical models. This type of empirical Bayes approach for hyperparameter selection is commonly used when a full prior elicitation is either not possible or is impractical. Inference is not sensitive to values of the hyperparameters $\boldsymbol{\lambda}$ that determine the priors of σ^k and α^k for two reasons. First, the standard deviation σ^k is the scale of the kernel that is used to smooth the discrete random probability measure generated by the DDP prior. It is important for reporting a smooth fit, that is, for display, but it is not critical for the imputed fits in our regression setting. Assuming some regularity of the posterior mean function, smoothing adds only minor corrections. Second, the total mass parameter α^k determines the number of unique clusters formed in the underlying Polya urn. However, because most clusters are small, changing the prior of α^k does not significantly change the posterior predictive values that are the basis for the proposed inference.

The conjugacy of the implied multivariate normal on $\{\theta_h^k(\boldsymbol{x}_i^k), i = 0, \ldots, n\}$ and the normal kernel in (3) greatly simplify computations, since any Markov chain Monte Carlo (MCMC) scheme for DP mixture models can be used. MacEachern and Müller (1998) and Neal (2000) described specific algorithms to implement posterior MCMC simulation in DPM models. Ishwaran and James (2001) developed alternative computational algorithms based on finite DPs, which truncated (2) after a finite number of terms. We provide details of MCMC computations in the online supplement.

4.3. Computing Mean Survival Time

We apply the Bayesian nonparametric DDP-GP model to obtain posterior means and credible intervals of mean survival time under each DTR. In the motivating leukemia trial, recall that the disease states are D (death), R (resistant disease), C (complete remission), and P (progressive disease). In stage $\ell = 1$ (induction chemotherapy), the three events D, R, and C are competing risks, so only one can be observed. We define seven counterfactual transition times T_i^k, where k indexes the transitions $(0, D)$, $(0, R)$, $(0, C)$, (R, D), (C, D), (C, P), and (P, D) (Figure 1). A dynamic treatment regime for this data may be expressed as $\boldsymbol{Z} = (Z^1, Z^{2,1}, Z^{2,2})$, where Z^1 is the induction chemo, $Z^{2,1}$ is the salvage therapy given if $s_{1i} = R$, and $Z^{2,2}$ is the salvage therapy given if $s_{1i} = C$ and $s_{2i} = P$.

Our primary goal is to estimate mean survival time for each DTR $\boldsymbol{Z}$ while accounting for baseline covariates and nonrandom treatment assignment. Under the DDP-GP model, we denote the mean survival time for a future patient under $\boldsymbol{Z}$ by

$$\eta(\boldsymbol{Z}) = E(T \mid \boldsymbol{Z}). \tag{7}$$

In terms of the seven counterfactual transition times, the survival time for a future patient $i = n + 1$ is

$$\begin{aligned} T_i = I(s_{1i} = D)T_i^{(0,D)} + I(s_{1i} = R)\left(T_i^{(0,R)} + T_i^{(R,D)}\right) \\ +I(s_{1i} = C)\left\{I(s_{2i} = D)\left(T_i^{(0,C)} + T_i^{(C,D)}\right)\right. \\ \left.+I(s_{2i} = P)\left(T_i^{(0,C)} + T_i^{(C,P)} + T_i^{(P,D)}\right)\right\}. \end{aligned} \tag{8}$$

The expectation of (8) under the DDP-GP model is evaluated by applying the law of total probability, using the same steps as in Wahed and Thall (2013). We first condition on the four possible cases, $(s_{1i} = D)$, $(s_{1i} = R)$, $(s_{1i} = C, s_{2i} = D)$, and $(s_{1i} = C, s_{2i} = P)$, compute the conditional expectation in each case, and then average across the cases. This computation requires evaluating seven expressions for the conditional mean transition times $\eta^k(\boldsymbol{Z}, \boldsymbol{x}^k) = E(T^k \mid \boldsymbol{Z}, \boldsymbol{x}^k)$ under $F^k(\cdot \mid \boldsymbol{x}^k)$, for each k. For example, $\eta^{(P,D)}(Z^1, Z^{2,2}, \boldsymbol{x}^0, T^{(0,C)}, T^{(C,P)})$ is the conditional mean remaining survival time, from P to D, given that C was achieved in stage 1 with frontline therapy Z^1, followed by P and salvage therapy $Z^{2,2}$ in stage 2. The DDP-GP models for $F^k(\cdot \mid \boldsymbol{x}^k)$, $k = 1, \ldots, n_T = 7$ define most of the marginalization for the expectation in $\eta(\boldsymbol{Z})$, leaving only conditioning on the baseline covariates $\boldsymbol{x}_i^0$. As Wahed and Thall (2013), we use the empirical covariate distribution $\widehat{p}(\boldsymbol{x}^0)$ over the observed patients to define an overall mean survival time (7). Note that the DDP-GP model does not accommodate time-varying covariates. The described evaluation of $\eta(\boldsymbol{Z})$ is an application of Robins' G-computation (Robins 1986; Robins, Hernán, and Brumback 2000). The complete expression is given as Equation (A.5) in the Appendix. In the upcoming discussion, we will use $\eta(\boldsymbol{Z})$ to evaluate the proposed approach.

5. Simulation Studies

We conducted four simulation studies to evaluate the performance of the proposed DDP-GP model for T in survival regression. The simulations focused on estimation of survival regression (simulation 1); regime effects in a study with two treatment arms and single-stage regimes (simulation 2); and regime effects in two studies with multi-stage regimes (simulations 3 and 4). For each of the latter three studies, the treatment assignment probabilities depended on patient covariates. That is, we introduced treatment selection bias. In all four simulations, we implemented inference under DDP-GP models. In simulation 1, we used a single survival regression $F(Y_i \mid \boldsymbol{x}_i)$ for a patient-specific baseline covariate vector $\boldsymbol{x}_i$. For simulation 2, we still used a single DDP-GP model $F(Y_i \mid \boldsymbol{x}_i, Z_i)$, now adding a treatment indicator Z_i to the survival regression model to estimate the causal effect. In simulations 3 and 4, we used independent DDP-GP models $F^k(Y_i^k \mid \boldsymbol{x}_i^k)$ for multiple transition times, $k = 1, \ldots, n_T$, similar to the application in our case study. For all four simulation studies, the hyperprior parameters were determined using the empirical Bayes approach described earlier. For all posterior computations, MCMC simulation was implemented with an initial burn-in of 2000 iterations and a total of 5000 iterations, thinning out in batches of 10. This worked well in all cases, with convergence diagnostics using the R package *coda* showing no evidence of practical convergence problems. Traceplots and empirical autocorrelation plots (not shown) for the imputed parameters indicated a well-mixing Markov chain.

5.1. Fitting a Survival Regression Model

In simulation 1, we considered four scenarios, with $n = 50$, 100, or 200 observations without censoring or $n = 200$ with 23% censoring. The details of simulation 1 are presented in Supplement B. Comparing the DDP-GP model with maximum likelihood estimates under the AFT model with Weibull, lognormal, and exponential distributions, the estimates under the DDP-GP model reliably recovered the shape of the true survival function and avoided the excessive bias seen with the AFT models.

5.2. Estimating a Treatment Effect in Single-Stage Regimes

Simulation 2 was designed to investigate inference under the DDP-GP model for the regime effect in a single-stage treatment setting. The simulated data represent what might be obtained in an observational setting where treatment is chosen by the attending physician based on patient covariates, rather than from a fairly randomized clinical trial. We simulated a binary treatment indicator $Z_i \in \{0 = \text{control}, 1 = \text{experimental}\}$ that depended on two continuous covariates, $\boldsymbol{x}_i = (L_i, W_i)$, for $n = 100$ patients, $i = 1, \ldots, n$. For example, L_i could be a patient's creatinine to quantify kidney function, and W_i could be body weight. We generated L_i from a mixture of normals, $L_i \sim \frac{1}{2}N(40, 10^2) + \frac{1}{2}N(20, 10^2)$, which could correspond to a subgroup of patients having worse kidney function (higher creatinine level) due to damage from prior chemotherapy. We assumed that $W_i \sim \text{Unif}(-\sqrt{12}, \sqrt{12})$, a uniform with zero mean and unit standard deviation, which could arise from standardizing a uniformly distributed raw variable. We generated the treatment indicators using the modified logistic regression model

$$p(Z_i = 1 \mid L_i, W_i) = \begin{cases} 0.05 & \text{if } \left\{1 + \exp[-2(L_i - 30)/10]\right\}^{-1} \le 0.05 \\ 0.95 & \text{if } \left\{1 + \exp[-2(L_i - 30)/10]\right\}^{-1} \ge 0.95 \\ \left\{1 + \exp[-2(L_i - 30)/10]\right\}^{-1} & \text{otherwise,} \end{cases}$$

that is, a logistic regression model with intercept 30 and slope 1/5 truncated at 0.05 and 0.95. This produces a very unbalanced treatment assignment, for example, $p(Z_i = 1 \mid L_i = 40) = 0.88$ versus $p(Z_i = 1 \mid L_i = 20) = 0.12$. This could arise in a setting where standard therapy (the "control"), $Z = 0$, is known to be nephrotoxic, while it is believed by most of the treating physicians that the experimental therapy, $Z_i = 1$, is not, so patients with high creatinine are more likely to be given the experimental therapy. In this simulation study, the goal is to estimate the comparative effect on survival of the experimental therapy versus the control. In the two treatment arms, we generated patients' responses from

$$Y(1) \sim \frac{1}{2} N\left(3 - 0.2L + \sqrt{L} - 0.1W,\ \sigma\right) + \frac{1}{2} N\left(2 - 0.2L + \sqrt{L} - 0.1W,\ \sigma\right)$$

and

$$Y(0) \sim N(-0.2L + \sqrt{L} - 0.1W,\ \sigma),$$

with $\sigma = 0.4$. We simulated 1000 trials. Note that under the simulation truth the treatment effect, $E[Y(1) - Y(0) \mid \boldsymbol{x} = (L, W)] = 2.5$, is constant across L, W.

Figure 2(a) plots the simulation truth for the mean response curve under $Z = 1$ and $Z = 0$ versus L, with $W \equiv 0$, in one randomly selected trial. The upper red solid curve represents $E[Y(1) \mid L, W = 0]$ and the lower black curve represents $E[Y(0) \mid L, W = 0]$. The red dots close to the upper curve are the observations for experimental arm patients and the black dots close to the lower curve are the observations for the control arm patients. We define an average treatment effect for the entire population under the simulation truth as $\text{ATE}^{\star} = \frac{1}{n}\sum_{i=1}^{n} E[Y_i(1) - Y_i(0)] = 2.5$.

We implemented inference for a survival regression $F(Y_i \mid \boldsymbol{x}_i, Z_i)$ using the proposed DDP-GP model (6). Figure 2(b) summarizes inference for the data from panel (a). Let $\hat{Y}_i(z) = E(Y_{n+1} \mid L_{n+1} = L_i, W_{n+1} = W_i, Z_{n+1} = z, \text{data})$ denote the posterior expected response for a future patient $n + 1$. We defined an estimated average treatment effect as $\text{ATE}_{\text{DDP}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_i(1) - \hat{Y}_i(0)]$. Figure 2(b) shows the estimated average treatment effect (horizontal red line), and credible intervals for individual effects $\hat{Y}_i(1) - \hat{Y}_i(0)$ (vertical line segments, located at L_i).

5.2.1. Inverse Probability of Treatment Weighting (IPTW)

For comparison, we also implemented inference using naive linear regression (LR), using an IPTW estimator, and an augmented IPTW (AIPTW) estimator for the average treatment effect. The LR estimator is based on a linear regression for log survival times, ignoring the lack of randomization. We use linear predictor functions $Y_i(1) = \beta_{10} + \beta_{11}L_i + \beta_{12}W_i + \epsilon_{1i}$ and $Y_i(0) = \beta_{00} + \beta_{01}L_i + \beta_{02}W_i + \epsilon_{0i}$. Denoting the least-square estimates by $\hat{\beta}_{zj}$ for $z = 0, 1$ and $j = 0, 1, 2$, the estimated means are $\hat{E}\{Y_i(z)\} = \hat{\beta}_{z0} + \hat{\beta}_{z1}L_i + \hat{\beta}_{z2}W_i$. We define an estimated average treatment effect based on the LR model as $\text{ATE}_{\text{LR}} = \frac{1}{n}\sum_i [\hat{E}\{Y_i(1)\} - \hat{E}\{Y_i(0)\}]$. Denote the propensity score $\pi_i = \text{pr}(Z_i = 1 \mid \boldsymbol{x}_i)$. The IPTW method corrects for bias due to lack of randomization by assigning each patient i a weight b_i equal to the inverse of an estimate of $p(Z_i \mid \boldsymbol{x}_i)$, the conditional probability of receiving his or her actual treatment (Robins, Hernán, and Brumback 2000). When $Z_i = 1$, $b_i = 1/\pi_i$; when $Z_i = 0$, $b_i = 1/(1 - \pi_i)$. An estimate of π_i is obtained by fitting a logistic regression model. We define the IPTW mean outcome estimator

$$\text{IPTW}(Z = z) = \frac{\sum_i I(Z_i = z)b_iY_i}{\sum_i I(Z_i = z)b_i},$$

and corresponding average treatment effect estimate $\text{ATE}_{\text{IPTW}} = \text{IPTW}(Z = 1) - \text{IPTW}(Z = 0)$.

5.2.2. Augmented IPTW (AIPTW)

The AIPTW estimate (Robins 2000) is a doubly robust generalization of the IPTW. It is consistent whenever the outcome regression model is correct and/or the propensity score model is correct. We evaluate the AIPTW estimator for average treatment effect (ATE):

$$\begin{aligned}\text{ATE}_{\text{AIPTW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\left[\frac{I(Z_i = 1)Y_i}{\hat{\pi}_i} - \frac{I(Z_i = 0)Y_i}{1 - \hat{\pi}_i}\right] \\ &- \frac{I(Z_i = 1) - \hat{\pi}_i}{\hat{\pi}_i(1 - \hat{\pi}_i)}\Big[(1 - \hat{\pi}_i)\hat{E}(Y_i \mid Z_i = 1, \boldsymbol{x}_i) \\ &+ \hat{\pi}_i\hat{E}(Y_i \mid Z_i = 0, \boldsymbol{x}_i)\Big]\Bigg\},\end{aligned} \tag{9}$$

where $\hat{\pi}_i$ is the estimated propensity score using logistic regression and $\hat{E}(Y_i \mid Z_i, \boldsymbol{x}_i)$ is estimated by a linear regression model, $i = 0, 1$.

Figure 2(b) shows ATE_{DDP}, ATE_{LR}, ATE_{IPTW}, and $\text{ATE}_{\text{AIPTW}}$ for one simulated dataset under this simulation setup. We found $E(\text{ATE}_{\text{DDP}} \mid \text{data}) = 2.31$, with 90% posterior credible interval $(1.89, 2.96)$, compared with the simulation truth $\text{ATE}^{\star} = 2.5$. In contrast, $\text{ATE}_{\text{LR}} = 4.13$ overestimates, while the IPTW method underestimates, with $\text{ATE}_{\text{IPTW}} = 1.11$. The AIPTW method reports $\text{ATE}_{\text{AIPTW}} = 2.73$. In Figure 2(b), the vertical green and blue segments are marginal 90% posterior credible intervals for the treatment effect (under the DDP-GP model) at each observed L value. Lengths of posterior credible intervals larger than 2 are highlighted by blue segments. Note how the uncertainty bounds grow wider in the range where there is less overlap across treatment groups, that is, over a range of covariate

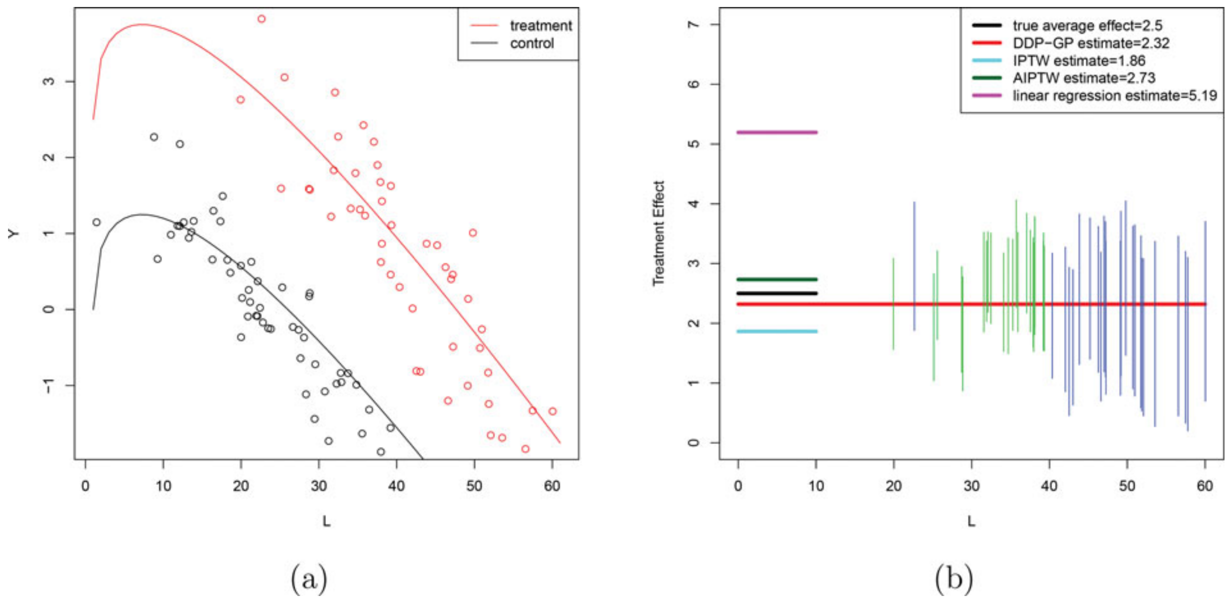


Figure 2. Simulation 2. (a) Simulated data for one (treatment, control) pair. The upper red solid curve represents $E[Y(1) \mid X]$, the lower black curve represents $E[Y(0) \mid X]$ given $W = 0$. The red dots close to the upper curve are the treated observations and the black dots close to the lower curve are the untreated. (b) Average treatment effect estimations ATE* (black solid line), ATE_{DDP} (red line), ATE_{IPTW} (turquoise blue), $\text{ATE}_{\text{AIPTW}}$ (dark green), ATE_{LR} (heliotrope). The vertical line segments are marginal 90% posterior intervals for the treatment effect at each L value from treated observations (under the DDP-GP model).

values for which we do not observe reliable empirical counterfactuals for each data point (e.g., $L > 50$). Most of the credible intervals reasonably cover the true treatment effect.

Figure 2(b) reports inference for one hypothetical dataset. For a comparison of average behavior, we carried out extensive simulations and report the distribution of estimated regime effects across these simulations. We compared the regime effect estimates obtained by DDP-GP, IPTW, AIPTW, and LR based on data from 1000 simulated trials. Figure 3 shows density plots of the distributions of estimated regime effects. Compared to the estimates obtained from DDP-GP or AIPTW, the IPTW estimates are much more variable, ranging from 1.14 to 7.13. The LR estimates are highly biased, and overestimate the true effects. The distribution of estimated regime effects under the DDP-GP model is highly concentrated around the simulation truth.

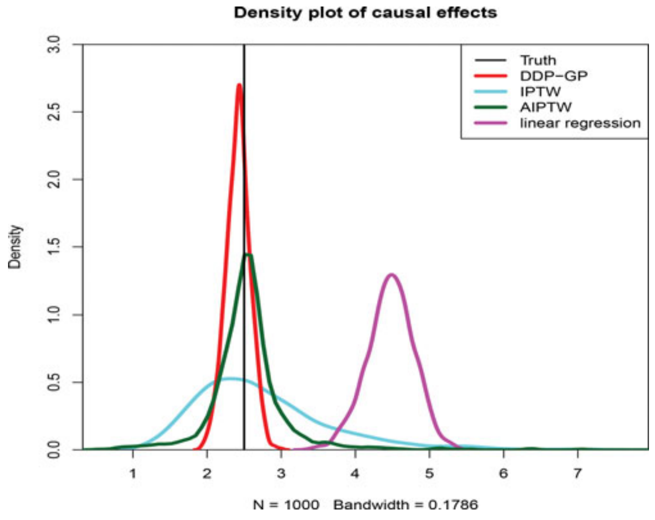


Figure 3. Simulation 2. The density plot of estimated regime effects by DDP-GP, IPTW, AIPTW and linear regression in 1000 trials. The truth is indicated by a black vertical line.

5.3. Regime Effect for Multi-Stage Regimes

Simulation 3 was designed to examine inference on strategy effects for multi-stage regimes with a general DTR setup. This simulation is similar to the scenario in Moodie, Richardson, and Stephens (2007). We simulated samples of size $n = 200$. Patients were randomized to initial induction therapy or not, coded as $Z_i^1 = a_1$ and $Z_i^1 = a_2$, with the randomization probabilities based on their baseline CD4 counts, which were simulated as $L_i \sim N(450, 10^2)$. For frontline therapy, we used the model $p(Z_i^1 = a_1 \mid L_i) = 0.8\, I(L_i < 450) + 0.2\, I(L_i \geq 450)$. To focus on covariate-dependent induction and salvage therapies, we assumed for simplicity that all patients were resistant to the induction therapy. Let $X \sim \text{LN}(m, s)$ denote a lognormal random variable with $\log(X) \sim N(m, s)$, we simulated the times $T_i^{(0,R)} \sim \text{LN}(2 + 0.005L_i, 0.3)$. The salvage treatment for each patient Z_i^2 was assigned with probability $p(Z_i^2 = 1 \mid Z_i^1, T_i^{(0,R)}) = Z_i^1 \text{expit}(1 - 0.003\, T_i^{(0,R)}) + (1 - Z_i^1)\text{expit}(-0.8 - 0.004\, T_i^{(0,R)})$, where $\text{expit}(u) = e^u/(1 + e^u)$. For the first-stage transition times, we generated transition times $T_i^{(R,D)} \sim \text{LN}(\boldsymbol{\beta}^{(R,D)}\boldsymbol{x}_i^{(R,D)}, 0.3)$, where $\boldsymbol{\beta}^{(R,D)} = (-0.5,\ 0.03,\ 0.2, 0.5,\ 0.3)$ and $\boldsymbol{x}_i^{(R,D)} = (1,\ L_i,\ Z_i^1, \log(T_i^{(0,R)}), Z_i^2)$.

The goal is to estimate mean survival time for each DTR (Z^1, Z^2). We have four possible DTRs in this simulation. We applied the Bayesian nonparametric DDP-GP model, IPTW, and AIPTW (Zhang et al. 2013) to each simulated dataset to estimate mean survival for each of the four possible DTRs. When implementing IPTW and AIPTW, we estimated the propensity score using logistic regression and the outcome model using AFT regression models with a lognormal distribution. For the nonparametric Bayesian inference, we defined independent DDP-GP models $F^k(Y_i^k \mid \boldsymbol{x}_i^k)$ as in (6) for each of the $n_T = 2$ log transition times $Y_i^k = \log T_i^k$. Figure 4(a) compares the mean survival estimates using boxplots of (Estimated mean survival −

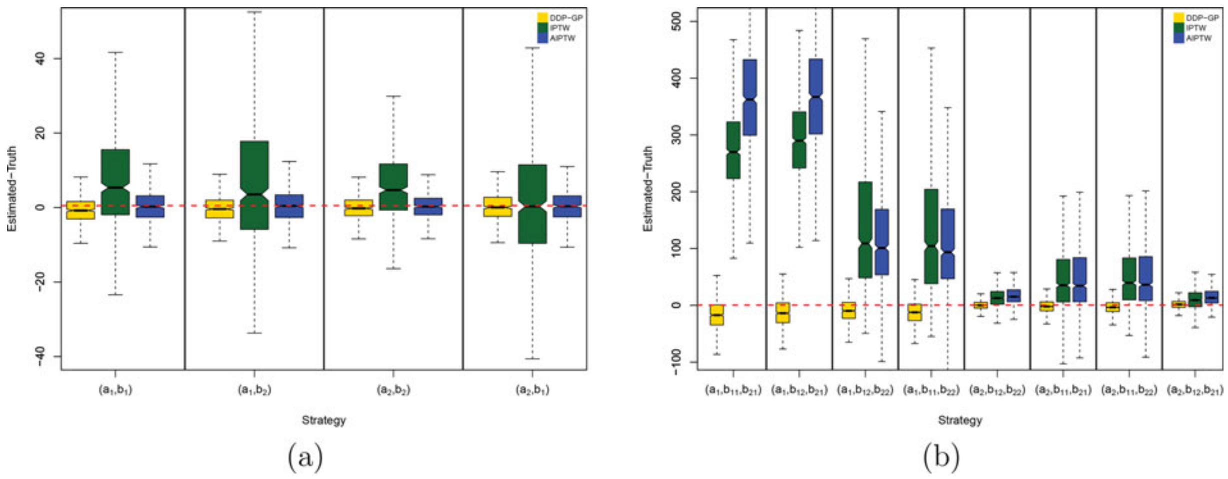


Figure 4. (a) Simulation 3 and (b) simulation 4. The yellow boxplots show posterior estimated mean OS using the DPP-GP model under each of the regimes as a difference with the simulation truth over 1000 simulations. The green and blue boxes show the corresponding inferences under the IPTW and AIPTW approaches, respectively. In each notched box-whisker plot, the box shows the interquartile range (IQR) from 1st quantile ($Q1$) to 3rd quantile ($Q3$), and the mid-line is the median. The top whisker denotes $Q3 + 1.5 * \text{IQR}$ and the bottom whisker $Q1 - 1.5 * \text{IQR}$. The notch displays a confidence interval for the median, that is, median$\pm 1.57 * \text{IQR}/\sqrt{1000}$.

Simulation truth), based on 1000 simulated datasets, arranged by inference method (DDP-GP, IPTW, and AIPTW) and by the four possible DTRs (the four subplots). Note that the DDP-GP and the AIPTW estimates are on average closer to the truth and have much smaller variability, compared to the IPTW estimates, across all four strategies. Because we use the same outcome regression models as the simulation truth when implementing the AIPTW method, it performs well in this simulation study. In summary, both the DDP-GP and the AIPTW methods show satisfactory performance in this example, although the DDP-GP estimates show slightly smaller variability than the AIPTW estimates.

Simulation 4 is a stylized version of the leukemia data that we will analyze in Section 6. We simulated samples of size $n = 200$ and patients' blood glucose values $L_i \sim N(100, 10^2)$. Patients initially were randomized equally between two induction therapies $Z^1 \in \{a_1, a_2\}$. We then generated a response (see below). Patients who were resistant (R) to the assigned induction therapies were then assigned salvage treatment $Z^{2,1} \in \{b_{11}, b_{12}\}$. Salvage treatments were randomized using the rule $p(Z^{2,1} = b_{11} \mid L_i) = 0.8\,I(L_i < 100) + 0.2\,I(L_i \geq 100)$. Patients who achieved C and subsequently suffered disease progression (P), were given salvage treatment $Z^{2,2} \in \{b_{21}, b_{22}\}$, using $p(Z^{2,2} = b_{21} \mid L_i) = 0.2\,I(L_i < 100) + 0.85\,I(L_i \geq 100)$. Finally, the survival time for each patient was evaluated as

$$T_i = \begin{cases} T_i^{(0,R)} + T_i^{(R,D)} & \text{if patient } i \text{ had} \\ \quad (L, Z^1, T^{(0,R)}, Z^{2,1}) & \\ T_i^{(0,C)} + T_i^{(C,P)} + T_i^{(P,D)} & \text{if patient } i \text{ had} \\ \quad (L, Z^1, T^{(0,C)}, T^{(C,P)}, Z^{2,2}). & \end{cases}$$

We simulated the times of the two competing risks R and C as $T_i^{(0,R)} \sim \text{LN}(\boldsymbol{\beta}^{(0,R)}\boldsymbol{x}_i^{(0,R)}, \sigma^{(0,R)})$ and $T_i^{(0,C)} \sim \text{LN}(\boldsymbol{\beta}^{(0,C)}\boldsymbol{x}_i^{(0,C)}, \sigma^{(0,C)})$, where $\boldsymbol{\beta}^{(0,R)} = (2,\ 0.02,\ 0)$, $\boldsymbol{\beta}^{(0,C)} = (1.5,\ 0.03,\ -0.8)$, with $\boldsymbol{x}_i^k = (1, L_i, Z_i^1)$ for $k \in \{(0,R), (0,C)\}$. For the three possible second-stage transitions $k \in \{(R,D), (C,P), (P,D)\}$, we generated (competing) transition times $T_i^k \sim \text{LN}(\boldsymbol{\beta}^k \boldsymbol{x}_i^k, \sigma^k)$, where $\boldsymbol{\beta}^{(R,D)} = (-0.5,\ 0.03,\ 0.2, 0.5,\ 0.3)$, $\boldsymbol{\beta}^{(C,P)} = (1,\ 0.05,\ 1,\ -0.6)$, $\boldsymbol{\beta}^{(P,D)} = (0.8,\ 0.04,\ 1.5,\ -1,\ 0.5,\ 0.5)$, with covariate vectors $\boldsymbol{x}_i^{(R,D)} = (1,\ L_i,\ Z_i^1, \log(T_i^{(0,R)}), Z_i^{2,1})$, $\boldsymbol{x}_i^{(C,P)} = (1, L_i, Z_i^1, \log(T_i^{(0,C)}))$, and $\boldsymbol{x}_i^{(P,D)} = (1, L_i, Z_i^1,\ \log(T_i^{(0,C)}), \log(T_i^{(C,P)}), Z_i^{2,2})$. We simulated $N = 1000$ trials with 15% censoring.

The goal is to estimate mean survival time for each DTR $(Z^1, Z^{2,1}, Z^{2,2})$. We performed inference under the Bayesian nonparametric DDP-GP model, IPTW, and AIPTW for each simulated dataset to estimate mean survival for each of the eight possible DTRs. When implementing IPTW and AIPTW, we estimated the propensity score using logistic regression and the outcome model using AFT regression models with a log-normal distribution. For the nonparametric Bayesian inference, we defined independent DDP-GP models $F^k(Y_i^k \mid \boldsymbol{x}_i^k)$ for each of the $n_T = 5$ log transition times $Y_i^k = \log T_i^k$. Figure 4(b) compares mean survival estimates using boxplots of (Estimated mean survival − Simulation truth), based on 1000 simulated datasets. The boxplots are arranged by inference method (DDP-GP, IPTW, AIPTW) and by all eight possible DTRs. In this simulation, both the propensity score model and the outcome model are incorrect when we implement the IPTW and AIPTW methods. In this case, the DDP-GP estimates on average are much closer to the truth and have much smaller variability, compared to the IPTW and AIPTW estimates, across all eight strategies as shown in Figure 4(b).

6. Evaluation of the Leukemia Trial Regimes

6.1. Leukemia Data—Inference for the Survival Regression

To analyze the AML-MDS trial data under the proposed DDP-GP model, we first implement posterior inference for six of the $n_T = 7$ transition times. The exception is $T^{(C,D)}$. Due to the limited sample size—only 9 patients died after C without first suffering disease progression (P)—we do not implement the DDP-GP model, and instead use an intercept-only Weibull AFT model. Table 1 summarizes the data. The table reports the

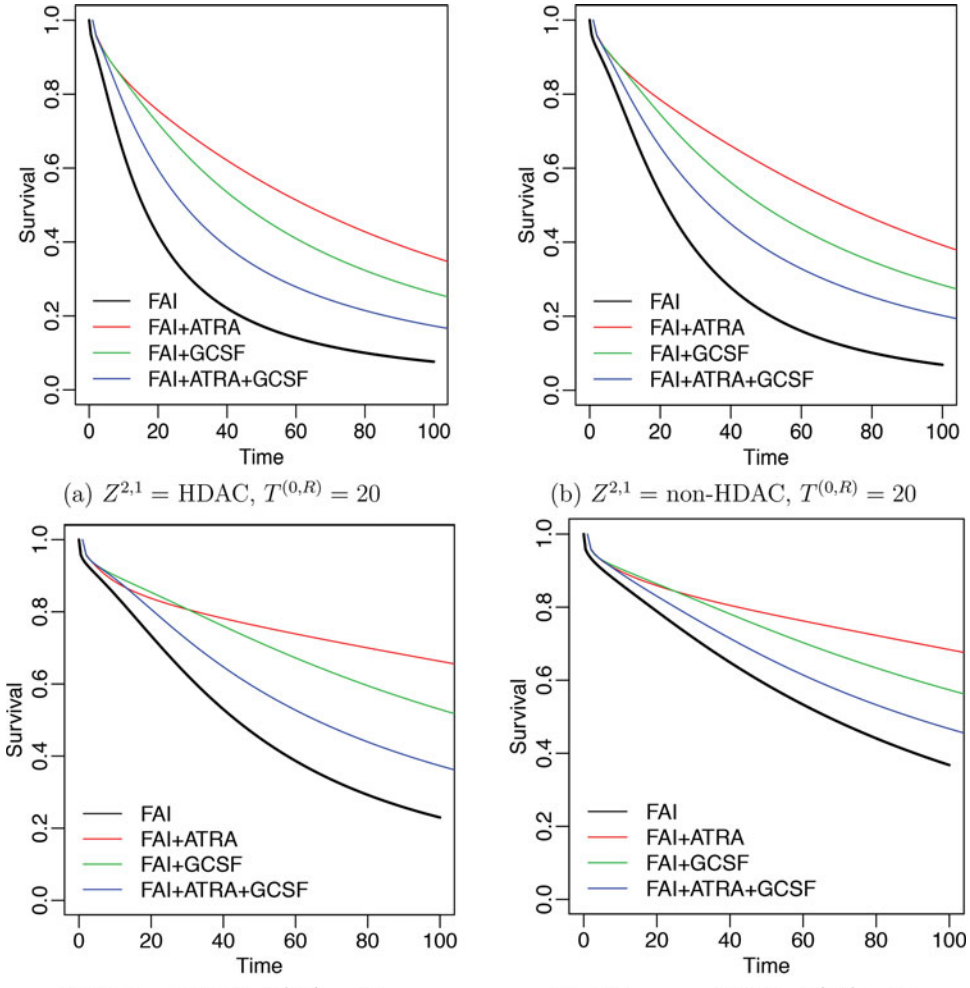
Table 1. The sample median of each transition time is given, with lower 25% quantile and upper 75% quantile in the parenthesis next to each median.

	Resistance		Die after resistance		
Induction	N	T^{R} (days)	Salvage	N	$T^{(R,D)}$ (days)
All	39	59 (47,84)	All	37	76 (27,187)
FAI	17	63 (41,97)	HDAC	25	65 (21,154)
FAI+ATRA	13	59 (55,76)			
FAI+GCSF	4	77 (43.5,106.75)	non-HDAC	12	146 (79, 376.75)
FAI+ATRA+GCSF	5	51 (48, 65)			
	CR		Die after progression		
Induction	N	T^{C} (days)	Salvage	N	$T^{(P,D)}$ (days)
All	102	32 (27,41)	All	83	120 (45,280)
FAI	20	31 (29, 44)	HDAC	47	106 (45,175.5)
FAI+ATRA	26	31 (25.25, 35)			
FAI+GCSF	28	35.5 (28,42.75)	non-HDAC	36	147.5 (42.75, 592.25)
FAI+ATRA+GCSF	28	32 (26,41)			

number of patients and median transition times for some selected transitions.

We first report results for $T^{(R,D)}$. Of 210 patients, 39 (18.57%) experienced resistance to their induction therapies. The rate of resistance varied across regimes, from 31% for patients receiving FAI, 24% for FAI plus ATRA, 7.8% for FAI plus GCSF, and 10% for FAI plus ATRA plus GCSF. The times to treatment resistance were longer, with greater variability in the FAI plus GCSF arm compared to the other three arms. Among the 39 patients who were resistant to induction therapies, 27 were given HDAC as salvage treatment, of whom 2 were censored before observing death. Figure 5 summarizes survival regression under the proposed DDP-GP model by plotting posterior predicted survival functions for a hypothetical future patient at age 61 with poor prognosis cytogenetic abnormality. The figure shows posterior predicted survival functions, arranged by different induction therapies Z^{1} (the four curves in each panel), $T^{(0,R)}$, and $Z^{2,1}$ (as indicated in the subtitle). Figure 5 shows that patients with shorter $T^{(0,R)}$ had lower predicted survival once their cancer became resistant. Also, patients with $s_1 = R$ who received $Z^{2,1}$ = HDAC as salvage had worse predicted survival than patients who received salvage treatment with non-HDAC. Similar results can be obtained for other transition times.

Next, we summarize results of the survival regression for $T^{(C,P)}$. Among the $n = 210$ patients, 102 (48.6%) achieved C, with C rates of 37%, 48%, 53%, and 56% in the FAI, FAI plus ATRA, FAI plus GCSF, and FAI plus GCSF plus ATRA arms,



(a) $Z^{2,1}$ = HDAC, $T^{(0,R)} = 20$

(b) $Z^{2,1}$ = non-HDAC, $T^{(0,R)} = 20$

(c) $Z^{2,1}$ = HDAC; $T^{(0,R)} = 55$

(d) $Z^{2,1}$ = non-HDAC; $T^{(0,R)} = 55$

Figure 5. Survival regression for $T^{(R,D)}$ in the AML-MDS trial. Panels (a)–(d) show the posterior estimated survival functions for a future patient at age 61 with poor prognosis cytogenetic abnormality, with $T^{(0,R)}$ and $Z^{2,1}$ as indicated. Survival curves are shown for four induction therapies. Black, red, green, and blue curves indicate Z^{1} = FAI, FAI+ATRA, FAI+GCSF, and FAI+ATRA+GCSF, respectively.

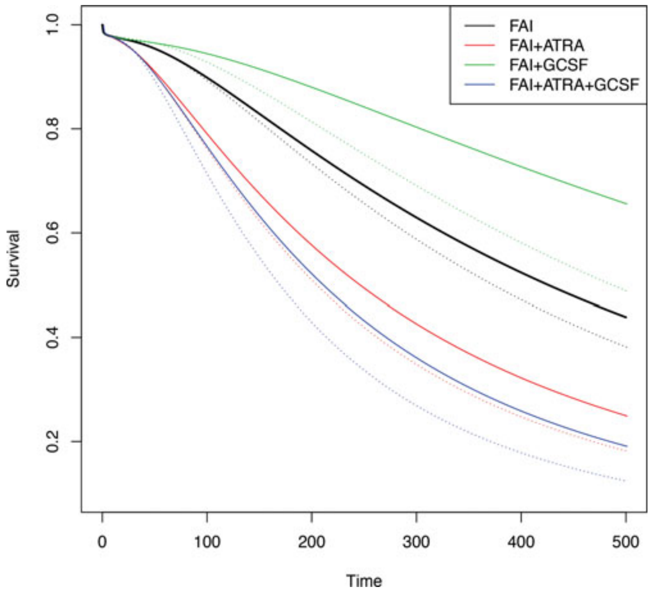


Figure 6. The effect of $T^{(0,C)}$ on $T^{(C,P)}$ at age 61 with poor cytogenetic abnormality. Black, red, green, and blue curves represent induction treatments FAI, FAI+ATRA, FAI+GCSF, and FAI+ATRA+GCSF, respectively. Solid lines and dotted lines represent $T^{(0,C)} = 20$ and $T^{(0,C)} = 30$, respectively. The longer it takes to achieve *C*, the shorter the period of time that the patient remained in *C*.

respectively. Of the 102 patients who achieved CR, 93 experienced disease progression before death or being lost to follow-up. Among these 93 relapsed patients, 53 received salvage treatment with HDAC. For a hypothetical future patient at age 61 with poor prognosis cytogenetic abnormality, Figure 6 summarizes survival regression functions for each of the four induction therapies, with solid lines representing $T^{(0,C)} = 20$ and dotted lines representing $T^{(0,C)} = 30$. The four dotted lines are below the four corresponding solid lines, indicating that $T^{(0,C)}$ was associated with $T^{(C,P)}$. This observation coincides with the well-known phenomenon in chemotherapy for AML or MDS that, regardless of induction therapy, the longer it takes to achieve *C*, the shorter the period that the patient remains in *C*.

Similarly, we summarize results for the survival regression for $T^{(P,D)}$. For a patient with poor prognosis cytogenetic abnormality, Figure 7 shows the posterior predicted survival functions under different combinations of induction therapy and age. Panels (a) and (c) show the survival functions of a patient assigned salvage treatment HDAC with age 46 or 76, while panels (b) and (d) plot the corresponding survival functions for the patient assigned non-HDAC as salvage. Four different colors

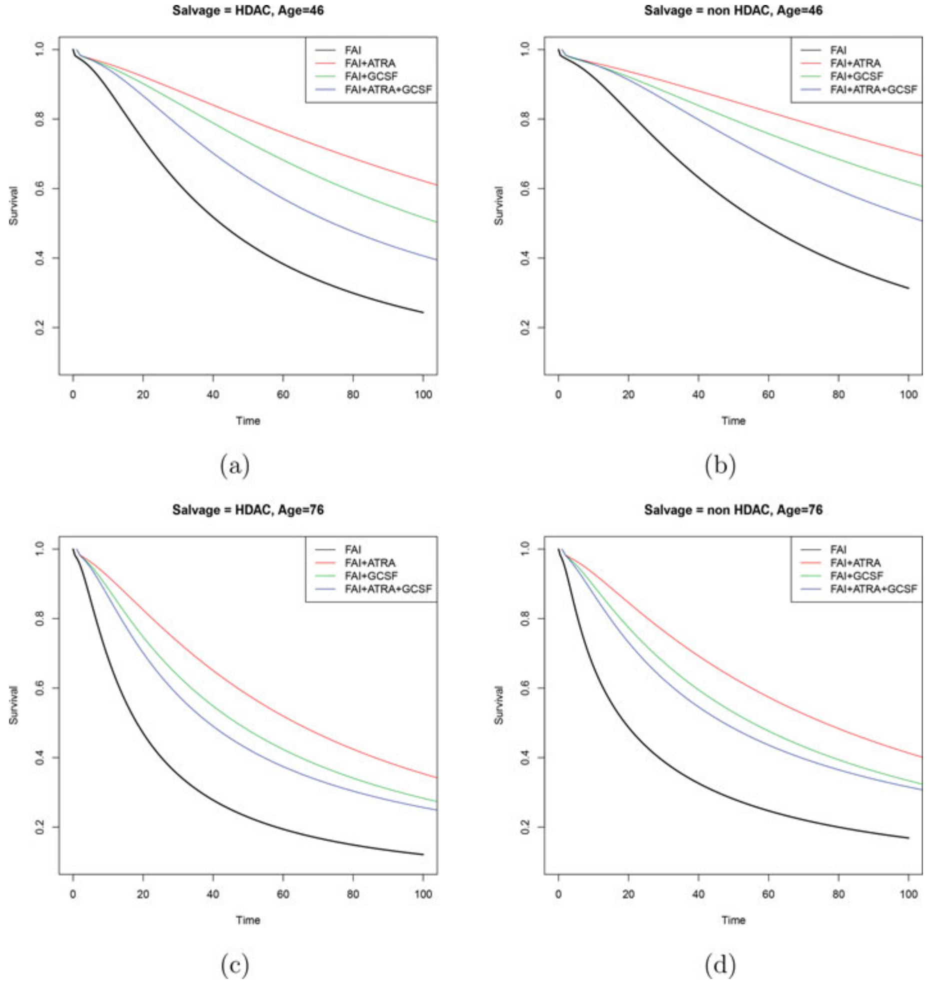


Figure 7. AML-MDS trial data in transition (P, D): Panels (a) and (c) show the posterior estimated survival functions of patient at age 46 and 76 with poor cytogenetic abnormality assigned to salvage treatment HDAC for four induction therapies, respectively. Panels (b) and (d) show the posterior estimated survival functions of patient at age 46 and 76 with poor cytogenetic abnormality assigned to salvage treatment non-HDAC for four induction therapies, respectively. Black, red, green, and blue curves represent induction treatments FAI, FAI+ATRA, FAI+GCSF, and FAI+ATRA+GCSF, respectively.

Table 2. Mean overall survival time under the IPTW method and the posterior mean and 90% credible interval (CI) under the DDP-GP model.

	Estimated mean OS times (days)		
		DDP-GP	
Regime (A, B_1, B_2)	IPTW	Posterior mean	90% CI
(FAI, HDAC, HDAC)	191.67	390.35	(286.47 545.6)
(FAI, HDAC, other)	198.18	416.34	(295.84 581.73)
(FAI, other, HDAC)	216.59	394.2	(287.15 538.63)
(FAI, other, other)	222.42	420.19	(296.51 579.05)
(FAI+ATRA, HDAC, HDAC)	527.43	572.9	(416.63 829.12)
(FAI+ATRA, HDAC, other)	458.85	617.15	(434.4 905.82)
(FAI+ATRA, other, HDAC)	532.29	573.46	(413.59 830.39)
(FAI+ATRA, other, other)	464.39	617.71	(434.49 900.32)
(FAI+GCSF, HDAC, HDAC)	326.15	542.06	(393.49 725.23)
(FAI+GCSF, HDAC, other)	281.78	578.24	(419.69 781.05)
(FAI+GCSF, other, HDAC)	327.66	542.5	(392.77 726.08)
(FAI+GCSF, other, other)	283.36	578.68	(421.46 781.26)
(FAI+ATRA+GCSF, HDAC, HDAC)	337.44	458.34	(327.91 651.21)
(FAI+ATRA+GCSF, HDAC, other)	285.64	502.48	(360.29 727.44)
(FAI+ATRA+GCSF, other, HDAC)	362.56	459.42	(328.09 651.61)
(FAI+ATRA+GCSF, other, other)	309.62	503.56	(358.84 726.88)

represent the four induction therapies. Figure 7 shows that residual survival time after disease progression following C was associated with both age and salvage therapy. Older patients were more likely to have shorter residual life once their disease progressed, and patients given HDAC as salvage died more quickly than patients given non-HDAC salvage.

6.2. Estimating the Regime Effects

In the AML-MDS trial, the four induction therapies and two salvage therapies define a total of 16 regimes. Mean survival time estimates under each of the 16 regimes were calculated using posterior inference under independent DDP-GP models $F^k(Y_i^k \mid \boldsymbol{x}_i^k)$ for each of the $n_T = 7$ transition times. For comparison, we also evaluated mean survival times using the IPTW method. See Equation (A.7) in the Appendix for details. Table 2 summarizes the results using IPTW and the DDP-GP model, including 90% credible intervals. Figure 8 shows boxplots of the marginal posterior distributions of survival times under the DDP-GP model for the 16 regimes.

The two methods give very different estimates for mean survival time, with the DDP-GP likelihood-based estimator much larger than the corresponding IPTW estimator for most regimes. The differences are expected due to the distinct properties of these two methods. The IPTW estimator uses the covariates to estimate the regime probability weights. In contrast, the DDP-GP likelihood-based method computes mean survival time, using G-computation, accounting for patients' covariates and previous transition times in addition to treatment followed by marginalizing over the empirical covariate distribution to obtain $\eta(\mathbf{Z})$. Additionally, the IPTW estimate is calculated from the overall samples, whereas the likelihood-based DDP-GP method models each transition time distribution separately, which reduces the effective sample size for each model fit and thus increases the overall variability even though they share the same prior for the $\boldsymbol{\beta}^k$'s.

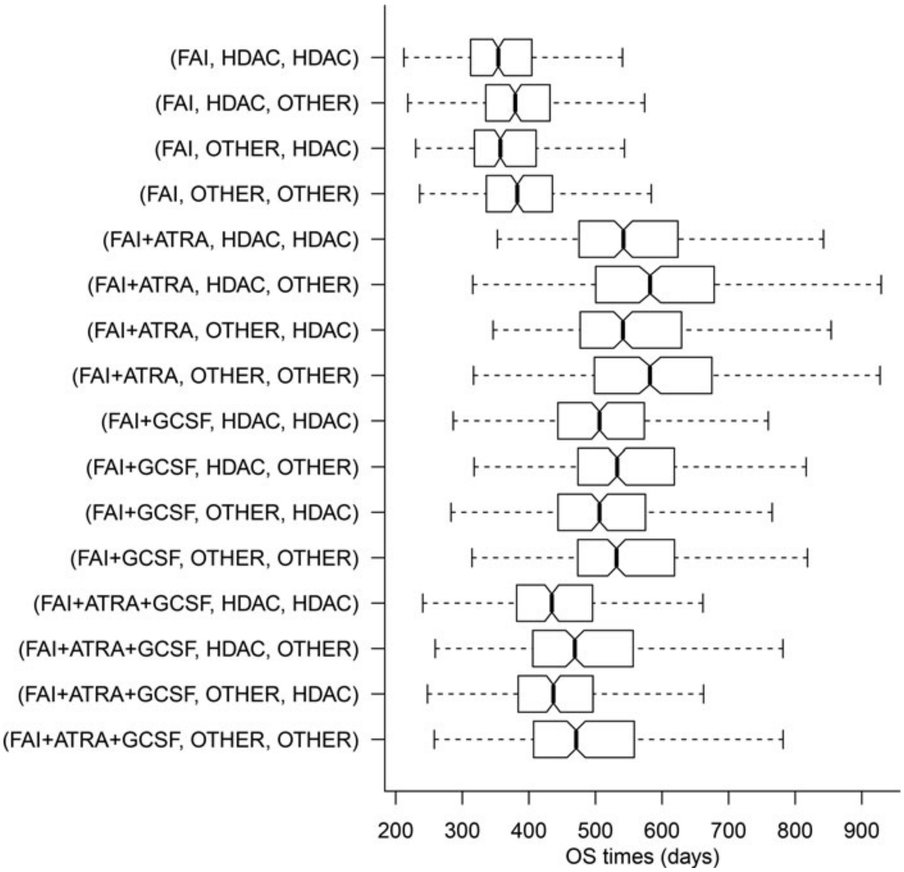


Figure 8. Marginal posterior distributions of overall survival time under the DDP-GP model for all 16 regimes.

For both methods, the estimates were smallest for the four regimes with FAI as induction therapy regardless of salvage treatment, and the 90% credible intervals were relatively small for these inferior regimes. Under the IPTW method, the estimates were largest for the four regimes with FAI plus ATRA as induction therapy, and the best regime is (FAI+ATRA, other, HDAC). With the DDP-GP likelihood-based approach, FAI plus ATRA as induction also gave the largest estimates, except for the regimes (FAI+GCSF, HDAC, other) and (FAI+GCSF, other, other), while the best regime is (FAI+ATRA, other, other). Most importantly, the DDP-GP likelihood-based approach showed that (FAI + ATRA, $Z^{2,1}$, other) was superior to (FAI + ATRA, $Z^{2,1}$, HDAC) regardless of $Z^{2,1}$. Therefore, our results suggest that (1) FAI plus ATRA was the best induction therapy, (2) if the patient's disease was resistant to FAI plus ATRA, then it was irrelevant whether the salvage therapy contained HDAC, and (3) if patients experienced progression after achieving C with FAI plus ATRA, then salvage therapy with non-HDAC was superior.

These conclusions, although not confirmatory, contradict those given by Estey et al. (1999), who concluded that none of the three adjuvant combinations FAI plus ATRA, FAI plus GCSF, or FAI plus ATRA plus GCSF were significantly different from FAI alone with respect to either survival or event-free survival time, based on consideration of only the frontline therapies by applying conventional Cox regression and hypothesis testing.

7. Conclusions

We have proposed a Bayesian nonparametric DDP-GP model for analyzing survival data and evaluating joint effects of induction-salvage therapies in clinical trials, using the posterior estimates, to predict survival for future patients. The Bayesian paradigm works very well, and the simulation studies suggest that our DDP-GP method yields more reliable estimates than IPTW and AIPTW. The DDP-GP model can be extended easily to multivariate outcomes. In Equation (2), this could be done by replacing the normal distribution with a multivariate normal distribution as the base measure. A referee has noted that, in settings where interpretability is important, our proposed BNP approach could be applied in the context of a policy search algorithm (Orellana, Rotnitzky, and Robins 2010; Zhang et al. 2012a, 2012b; Zhao et al. 2012; Zhang et al. 2013; Zhao et al. 2014, 2015).

We employed two different methods to evaluate the 16 possible two-stage regimes for choosing induction and salvage therapies in the leukemia trial data. The IPTW method estimates the regime effect by using covariates only to compute the assignment probabilities of salvage therapies to correct for bias. In contrast, likelihood-based G-computation under the DDP-GP model accounts for all possible outcome paths, the transition times between successive states, and effects of covariates and previous outcomes, on each transition time. Although the two methods gave different numerical estimates of mean survival time, they both reached the conclusion that FAI plus ATRA was the best induction therapy and FAI was the worst induction therapy. Although our current models are set up for two-stage treatment regimes, they easily can be extended to other applications with multi-stage regimes.

Appendix: The Complete Expression of Likelihood and the IPTW Method

A.1. Likelihood

The following structure is adapted from Wahed and Thall (2013), and is included here for completeness. The risk sets of the seven transition times in the leukemia trial are defined as follows. Let $\mathcal{R}^0 = \{1, \dots, n\}$ denote the initial risk set at the start of induction chemotherapy, and $\mathcal{R}^{(0,r)} = \{i : s_{1i} = r\}$ for $r = D, C, R$, so $\mathcal{R}^0 = \mathcal{R}^{(0,D)} \cup \mathcal{R}^{(0,C)} \cup \mathcal{R}^{(0,R)}$. Similarly, $\mathcal{R}^{(C,P)} = \{i : s_{1i} = C, s_{2i} = P\}$ is the later risk set for $T^{(P,D)}$.

To record right censoring, let U_i denote the time from the start of induction to last followup for patient i. We assume that U_i is conditionally independent of the transition time given prior transition times and other covariates. Censoring of event times occurs by competing risk and/or loss to followup. For patient i in the risk set for transition time T_i^k, let $\delta_i^k = 1$ if patient i is not censored and 0 if patient i is right censored. For example, $\delta_i^{(0,D)} = 1$ for $i \in \mathcal{R}^0$ if $T_i^{(0,D)} = \min(U_i, T_i^{(0,D)}, T_i^{(0,C)}, T_i^{(0,R)})$. Similarly, $\delta_i^{(R,D)} = 1$ for $i \in \mathcal{R}^{(0,R)}$ if $T_i^{(0,R)} + T_i^{(R,D)} < U_i$ and $\delta_i^{(P,D)} = 1$ for $i \in \mathcal{R}^{(C,P)}$ if $T_i^{(0,C)} + T_i^{(C,P)} + T_i^{(P,D)} < U_i$.

For $i \in \mathcal{R}^0$, let $V_i^0 = \min(T_i^{(0,D)}, T_i^{(0,R)}, T_i^{(0,C)}, U_i)$ denote the observed time for the stage 1 event or censoring. For $i \in \mathcal{R}^{(0,C)}$ let $V_i^C = \min(T_i^{(C,D)}, T_i^{(C,P)}, U_i - T_i^{(0,C)})$ denote the observed event time for the competing risks D and P and loss to followup. Similarly, for $i \in \mathcal{R}^{(0,R)}$, let $V_i^R = \min(T_i^{(R,D)}, U_i - T_i^{(0,R)})$, and for $i \in \mathcal{R}^{(C,P)}$ let $V_i^{(C,P)} = \min(T_i^{(P,D)}, U_i - T_i^{(0,C)} - T_i^{(C,P)})$.

The joint likelihood function is the product $\mathcal{L} = \mathcal{L}_1\mathcal{L}_2\mathcal{L}_3\mathcal{L}_4$. The first factor $\mathcal{L}_1$ corresponds to response to induction therapy,

$$\mathcal{L}_1 = \prod_{i\in\mathcal{R}^0} \prod_{r\in\{D,R,C\}} f^{(0,r)}\left(V_i^0 \mid \boldsymbol{x}_i^{(0,r)}\right)^{\delta_i^{(0,r)}} \bar{F}^{(0,r)}\left(V_i^0 \mid \boldsymbol{x}_i^{(0,r)}\right)^{1-\delta_i^{(0,r)}}, \tag{A.1}$$

where $\bar{F}^k = 1 - F^k$. The second factor $\mathcal{L}_2$ corresponds to patients $i \in \mathcal{R}^{(0,R)}$ who experience resistance to induction and receive salvage $Z^{2,1}$,

$$\mathcal{L}_2 = \prod_{i\in\mathcal{R}^{(0,R)}} f^{(R,D)}\left(V_i^R \mid \boldsymbol{x}_i^{(R,D)}\right)^{\delta_i^{(R,D)}} \bar{F}^{(R,D)}\left(V_i^R \mid \boldsymbol{x}_i^{(R,D)}\right)^{1-\delta_i^{(R,D)}}. \tag{A.2}$$

The third factor $\mathcal{L}_3$ is the likelihood contribution from patients achieving C,

$$\mathcal{L}_3 = \prod_{i\in\mathcal{R}^{(0,C)}} \prod_{k=(C,D),(C,P)} f^k\left(V_i^C \mid \boldsymbol{x}_i^k\right)^{\delta_i^k} \bar{F}^k\left(V_i^C \mid \boldsymbol{x}_i^k\right)^{1-\delta_i^k}. \tag{A.3}$$

The fourth factor $\mathcal{L}_4$ is the contribution from patients who experience tumor progression after C,

$$\mathcal{L}_4 = \prod_{i\in\mathcal{R}^{(C,P)}} f^{(P,D)}\left(V_i^{(C,P)} \mid \boldsymbol{x}_i^{(P,D)}\right)^{\delta_i^{(P,D)}} \bar{F}^{(P,D)}\left(V_i^{(C,P)} \mid \boldsymbol{x}_i^{(P,D)}\right)^{1-\delta_i^{(P,D)}}. \tag{A.4}$$

The mean survival time of a patient treated with regime $\mathbf{Z} = (Z^1, Z^{2,1}, Z^{2,2})$ is

$$\eta(\mathbf{Z}) = \int \left[p\left(s_1 = D \mid \boldsymbol{x}^0, Z^1\right) \eta^{(0,D)}\left(\boldsymbol{x}^0, Z^1\right) \right] d\widehat{p}(\boldsymbol{x}^0)$$

$$
\begin{aligned}
&+\int \Big\{ p\left(s_1 = R \mid \boldsymbol{x}^0, Z^1\right) \Big[\eta^R\left(\boldsymbol{x}^0, Z^1\right) \\
&+\int \eta^{(R,D)}\left(\boldsymbol{x}^0, Z^1, Z^{2,1}, T^{(0,R)}\right) d\mu(T^{(0,R)}) \Big] \Big\} d\widehat{p}(\boldsymbol{x}^0) \\
&+\int p\left(s_1 = C \mid \boldsymbol{x}^0, Z^1\right) \Big[\eta^C\left(\boldsymbol{x}^0, Z^1\right) \\
&+\int \Big[p\left(s_2 = D \mid s_1 = C, \boldsymbol{x}^0, Z^1, T^{(0,C)}\right) \eta^{(C,D)}\left(\boldsymbol{x}^0, Z^1, T^C\right) \\
&+p\left(s_2 = P \mid s_1 = C, \boldsymbol{x}^0, Z^1, T^{(0,C)}\right) \Big[\eta^{(C,P)}\left(\boldsymbol{x}^0, Z^1, T^{(0,C)}\right) \\
&+\int \eta^{(P,D)}\left(\boldsymbol{x}^0, Z^1, Z^{2,2} T^{(0,C)}, T^{(C,P)}\right) d\mu\left(T^{(C,P)}\right) \Big] \\
&\times d\mu\left(T^{(0,C)}\right) \Big] d\widehat{p}(\boldsymbol{x}^0). \qquad \text{(A.5)}
\end{aligned}
$$

A.2. IPTW

We compute the IPTW estimates for overall mean survival with regime $\mathbf{Z}$ as

$$
\text{IPTW}(\mathbf{Z}) = \sum_{i=1}^{n} w_i(\mathbf{Z}) T_i \Big/ \sum_{i=1}^{n} w_i(\mathbf{Z}), \qquad \text{(A.6)}
$$

where

$$
\begin{aligned}
w_i(\mathbf{Z}) = \frac{I(\mathbf{Z} = \mathbf{Z}_i)\delta_i}{\hat{K}(U_i)} \Big[& I(s_{1i} = D) + I(s_{1i} = R) \\
& \times I_i(Z^{2,1}) / \hat{\Pr}\left(Z^{2,1} \mid s_{1i} = R, Z^1, \boldsymbol{x}_i^0, T_i^{(0,R)}\right) \\
& + I(s_{1i} = C, s_{2i} = D) \\
& + I(s_{1i} = C, s_{2i} = P) I_i(Z^{2,2}) / \hat{\Pr}\Big(Z^{2,2} \mid s_{1i} = C, \\
s_{2i} = P, Z^1, \boldsymbol{x}_i^0, T_i^{(0,C)}, T_i^{(C,P)}\Big) \Big]. &
\end{aligned}
\qquad \text{(A.7)}
$$

In (A.7), $\hat{K}$ is the Kaplan–Meier estimator of the censoring survival distribution $K(u) = P(U \geq t)$ at time t. $I_i(Z)$ is an indicator of treatment Z and 0 otherwise, and $\hat{\Pr}(Z^{2,1} \mid s_{1i} = C, Z^1, \boldsymbol{x}_i^0, T_i^{(0,R)})$ is the probability of receiving salvage treatment $Z^{2,1}$ estimated using logistic regression, and similarly for $\hat{\Pr}(Z^{2,2} \mid s_{1i} = C, s_{2i} = P, Z^1, \boldsymbol{x}_i^0, T_i^{(0,C)}, T_i^{(C,P)})$. The above estimator has been shown to be consistent under suitable assumptions (Scharfstein, Rotnitzky, and Robins 1999; Wahed and Thall 2013).

Supplementary Materials

The supplementary material includes the details of MCMC posterior sampling for the proposed DDP-GP model and more simulation studies.

Funding

This research was supported by NCI/NIH grant R01 CA157458-01A1 (Yanxun Xu and Peter Müller) and R01 CA83932 (Peter F. Thall). This research was also supported in part by NIH through resources provided by the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory, under grant 1S10OD018495-01.

References

Bernardo, J., Berger, J., and Smith, A. D. F. (1999), "Regression and Classification Using Gaussian Process Priors," in *Bayesian Statistics 6: Proceedings of the Sixth Valencia Inter-National Meeting* (Vol. 6), eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 475. [925]

Connolly, S., and Bernstein, G. (2007), "Practice Parameter for the Assessment and Treatment of Children and Adolescents With Anxiety Disorders," *Journal of the American Academy of Child and Adolescent Psychiatry*, 46, 267–283. [922]

Dawson, R., and Lavori, P. W. (2004), "Placebo-Free Designs for Evaluating New Mental Health Treatments: The use of Adaptive Treatment Strategies," *Statistics in Medicine*, 23, 3249–3262. [922]

Estey, E. H., Thall, P. F., Pierce, S., Cortes, J., Beran, M., Kantarjian, H., Keating, M. J., Andreeff, M., and Freireich, E. (1999), "Randomized Phase II Study of Fludarabine+ Cytosine Arabinoside+ Idarubicin±All-trans Retinoic Acid±Granulocyte Colony-Stimulating Factor in Poor Prognosis Newly Diagnosed Acute Myeloid Leukemia and Myelodysplastic Syndrome," *Blood*, 93, 2478–2484. [923,933]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [924]

Goldberg, Y., and Kosorok, M. R. (2012), "Q-Learning With Censored Data," *Annals of Statistics*, 40, 529–560. [922]

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000), "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men," *Epidemiology*, 11, 561–570. [922]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [922]

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173. [924,926]

Karabatsos, G., and Walker, S. G. (2012), "A Bayesian Nonparametric Causal Model," *Journal of Statistical Planning and Inference*, 142, 925–934. [922]

Lavori, P. W., and Dawson, R. (2000), "A Design for Testing Clinical Strategies: Biased Adaptive Within-Subject Randomization," *Journal of the Royal Statistical Society*, Series A, 163, 29–38. [922]

Lunceford, J. K., Davidian, M., and Tsiatis, A. A. (2002), "Estimation of Survival Distributions of Treatment Policies in Two-Stage Randomization Designs in Clinical Trials," *Biometrics*, 58, 48–57. [922]

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science, American Statistical Association*, pp. 50–55. [924]

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238. [926]

Moodie, E. E., Richardson, T. S., and Stephens, D. A. (2007), "Demystifying Optimal Dynamic Treatment Regimes," *Biometrics*, 63, 447–455. [922,928]

Müller, P., and Mitra, R. (2013), "Bayesian Nonparametric Inference—Why and How," *Bayesian Analysis*, 8, 269–302. [924]

Müller, P., and Rodriguez, A. (2013), "Nonparametric Bayesian Inference," IMS-CBMS Lecture Notes. IMS, 270. Shaker Heights, OH: Institute of Mathematical Statistics. [924]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [922]

——— (2005), "An Experimental Design for the Development of Adaptive Treatment Strategies," *Statistics in Medicine*, 24, 1455–1481. [922]

Murphy, S. A., Collins, L. M., and Rush, A. J. (2007), "Customizing Treatment to the Patient: Adaptive Treatment Strategies," *Drug and Alcohol Dependence*, 88, S1–S3. [922]

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., and TenHave, T. (2007), "Developing Adaptive Treatment Strategies in Substance Abuse Research," *Drug and Alcohol Dependence*, 88, S24–S30. [922]

Murphy, S., Van Der Laan, M., and Robins, J. (2001), "Marginal Mean Models for Dynamic Regimes," *Journal of the American Statistical Association*, 96, 1410–1423. [922]

Neal, R. (1995), "Bayesian Learning for Neural Networks," Ph.D. dissertation, Graduate Department of Computer Science, University of Toronto. [925]

——— (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265. [926]

O'Hagan, A., and Kingman, J. (1978), "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society*, Series B, 40, 1–42. [925]

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010), "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content," *The International Journal of Biostatistics*, 6. [933]

Rasmussen, C., and Williams, C. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press. [925]

Robins, J. (1986), "A New Approach to Causal Inference in Mortality Studies With a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 7, 1393–1512. [921,922,926]

——— (1987), "Addendum to "A New Approach to Causal Inference in Mortality Studies With a Sustained Exposure Period – Application to Control of the Healthy Worker Survivor Effect," *Computers & Mathematics With Applications*, 14, 923–945. [921,922]

——— (1989), "The Analysis of Randomized and Non-Randomized Aids Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," *Health Service Research Methodology: A Focus on AIDS*, 113–159. [922]

——— (1997), "Causal Inference From Complex Longitudinal Data," in *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane, New York: Springer, pp. 69–117. [922]

——— (2000), "Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models," in *Proceedings of the American Statistical Association* (Vol. 1999), pp. 6–10. [927]

——— (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, eds. D. Y. Lin and P. J. Heagerty, New York: Springer, pp. 189–326. [922]

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560. [926,927]

Robins, J. M., Orellana, L., and Rotnitzky, A. (2008), "Estimation and Extrapolation of Optimal Treatment and Testing Strategies," *Statistics in Medicine*, 27, 4678–4721. [922]

Robins, J. M., and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology*, eds. N. P. Jewell, K. Dietz, and V. T. Farewell, New York: Springer, pp. 297–331. [923]

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120. [934]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [924]

Shi, J. Q., Wang, B., Murray-Smith, R., and Titterington, D. M. (2007), "Gaussian Process Functional Regression Modeling for Batch Data," *Biometrics*, 63, 714–723. [925]

Thall, P. F., Logothetis, C., Pagliaro, L. C., Wen, S., Brown, M. A., Williams, D., and Millikan, R. E. (2007a), "Adaptive Therapy for Androgen-Independent Prostate Cancer: A Randomized Selection Trial of Four Regimens," *Journal of the National Cancer Institute*, 99, 1613–1622. [922]

Thall, P. F., Millikan, R. E., and Sung, H.-G. (2000), "Evaluating Multiple Treatment Courses in Clinical Trials," *Statistics in Medicine*, 19, 1011–1028. [922]

Thall, P. F., Sung, H.-G., and Estey, E. H. (2002), "Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials," *Journal of the American Statistical Association*, 97, 29–39. [922]

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007b), "Bayesian and Frequentist Two-Stage Treatment Strategies Based on Sequential Failure Times Subject to Interval Censoring," *Statistics in Medicine*, 26, 4687–4702. [922]

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, New York: Springer. [922]

van der Laan, M. J., and Petersen, M. L. (2007), "Causal Effect Models for Realistic Individualized Treatment and Intention to Treat Rules," *International Journal of Biostatistics*, 3, 3. [922]

Wahed, A. S., and Thall, P. F. (2013), "Evaluating Joint Effects of Induction–Salvage Treatment Regimes on Overall Survival in Acute Leukaemia," *Journal of the Royal Statistical Society*, Series C, 62, 67–83. [921,923,926,933,934]

Wahed, A. S., and Tsiatis, A. A. (2006), "Semiparametric Efficient Estimation of Survival Distributions in Two-Stage Randomisation Designs in Clinical Trials With Censored Data," *Biometrika*, 93, 163–177. [922]

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2012), "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer," *Journal of the American Statistical Association*, 107, 493–508. [922]

Williams, C. K. (1998), "Prediction With Gaussian Processes: From Linear Regression to Linear Prediction and Beyond," in *Learning in Graphical Models*, ed. M. I. Jordan, Dordrecht, The Netherlands: Springer, pp. 599–621. [925]

Zajonc, T. (2012), "Bayesian Inference for Dynamic Treatment Regimes: Mobility, Equity, and Efficiency in Student Tracking," *Journal of the American Statistical Association*, 107, 80–92. [922]

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a), "Estimating Optimal Treatment Regimes From a Classification Perspective," *Stat*, 1, 103–114. [933]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [933]

——— (2013), "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100, 681–694. [928,933]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [933]

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2014), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [933]

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015), "Doubly Robust Learning for Estimating Individualized Treatment With Censored Data," *Biometrika*, 102, 151–168. [922,933]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1200911

Comment

Qian Guan, Eric B. Laber, and Brian J. Reich

Department of Statistics, North Carolina State University, Raleigh, NC, USA

1. Introduction

We congratulate Xu, Muller, Wahed, and Thall (hereafter, XMWT) on an interesting and novel article and we thank the Editors for organizing these discussions. The integration of nonparametric Bayesian (NPB) methods and optimal treatment regimes was long overdue; we expect that the work of XMWT will be the first in a long and fruitful line of research. There is currently a great amount of interest in non-parametric methods for estimation of optimal treatment regimes; for example, recently proposed methodology includes kernel-based methods (Zhao et al. 2009, 2012, 2015, 2014; Zhou et al. 2015), nearest-neighbor methods (Zhou and Kosorok 2016), generalized additive models (Moodie, Dean, and Sun 2013), boosting (Kang, Janes, and Huang 2014), and trees (Laber and Zhao 2015; Doove et al. 2015; Zhang et al. 2015). This interest stems in part from a desire to mitigate the risk of model misspecification as it is well-known that the optimal treatment regime can be a highly nonlinear function of patient covariates even under simple generative models (Laber, Linn, and Stefanski 2014; Schulte et al. 2014). NPB methods are in line with this trend and, as we discuss below, may possess a number of advantages over existing methods in the context of policy-search algorithms.

However, nonparametric estimation of an optimal treatment regime often comes at the price of a loss of interpretability within a domain context which can be a major detriment to scientific progress, especially if estimation is done a secondary, hypothesis-generating analysis. In this discussion, we argue that NPB methods can have tremendous value as an engine for policy-search algorithms used to estimate an optimal treatment regime within a prespecified class (Robins, Orellana, and Rotnitzky 2008; Orellana, Rotnitzky, and Robins 2010; Zhao et al. 2012; Zhang et al. 2012, 2013; Zhao et al. 2014, 2015). An advantage of policy-search methods is that the prespecified class can be chosen to ensure parsimony and interpretability, etc. However, applying NPB methods for this purpose is challenging because of the curse of dimensionality, we expect that the authors will have some insight into this issue.

In Section 2, we discuss whether NPB methods are necessary if the primary goal is to evaluate the marginal mean outcome under a small number of fixed regimes and compare the method of XMWT with a flexible accelerated failure time model (Wahed and Thall 2013). In Section 3, we describe a schematic for NPB methods for policy-search with sequential transition times. We make concluding remarks and discuss directions for future research in Section 4.

2. Nonparametric Bayesian Methods for Evaluating Fixed Treatment Regimes

XMWT use nonparametric Bayesian (NPB) methods to evaluate a finite set of fixed treatment regimes. This requires estimating a series of densities conditional on a patient's covariate and outcome trajectory. Thus, to estimate the marginal mean outcome under each fixed regime, they must take the intermediate step of performing conditional density estimation; this density can be high-dimensional, especially if one desires to incorporate accumulating longitudinal patient information to individualize multi-stage treatment decisions (see Section 3). Furthermore, as XMWT illustrate, implementation of NBP methods for evaluating fixed treatment regimes is highly non-trivial. In contrast, existing regression-based or inverse-probability weighting estimators of the marginal mean outcome under a fixed treatment regime involve minimal modeling and computational burden (Zhang et al. 2012, 2013). However, inverse-weighting methods can be highly variable especially if the number of treatment combinations is large and the sample size is small; and parametric regression-based estimators can be highly biased if the regression model is misspecified. NPB methods seem to provide a nice compromise in that modeling the conditional densities introduces enough structure to reduce variability yet the class of densities is sufficiently rich to avoid severe misspecification. A similar, but much simpler compromise, is to use a flexible regression-based estimator to estimate the marginal mean outcome under each fixed regime (this approach was advocated by Taylor, Cheng, and Foster 2015).

2.1. A Simple Semiparametric Model

To illustrate the use of a flexible regression-based estimator, we apply a semiparametric accelerated failure time model based on Wahed and Thall (2013); using the notation of XMWT, this model postulates that the transition T (the model is applied separately for each arm of the study) follows the additive model

$$\log(T) = \alpha + \sum_{l=1}^{p} b_l(x_l) + \epsilon, \qquad (1)$$

CONTACT Eric Laber eblaber@ncsu.edu Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

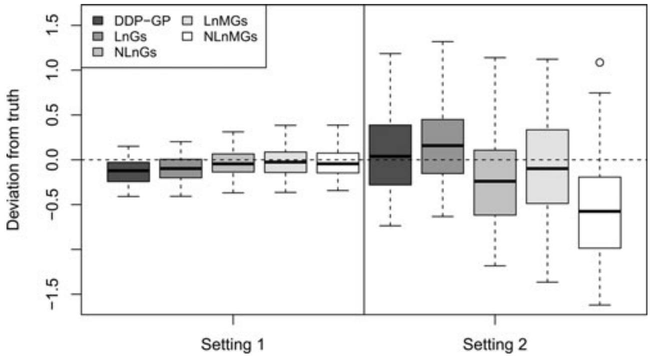


Figure 1. Simulation results for single-stage trials under Settings 1 (left) and 2 (right). The boxplots show the differences between estimated and true causal effects over the 100 simulated datasets using the dependent Dirichlet process model ("DDP-GP") and the semiparametric models with linear ("Ln") and nonlinear ("NLn") means and Gaussian ("Gs") and mixture of Gaussians ("MGs") residuals.

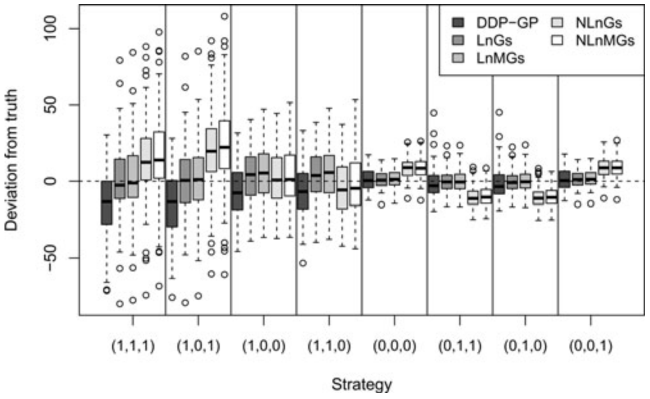


Figure 2. Simulation results for three-stage trials under Setting 3. The boxplots show the differences between estimated and true causal effects over the 100 simulated datasets using the Dependent Dirichlet Process model ("DDP-GP") and the semiparametric models with linear ("Ln") and nonlinear ("NLn") means and Gaussian ("Gs") and mixture of Gaussians ("MGs") residuals.

where $x = (x_1, \dots, x_p)$ are the covariates, b_l are smooth functions; and ϵ is an independent error term with smooth density. For binary covariates (e.g., treatment indicators) we take $b_l(x_l) = x_l\beta_l$; for continuous covariates we model b_l using M b-spline basis functions, $b_l(u) = \sum_{m=1}^{M} \psi_m(u)\gamma_{lj}$, where $\psi_1, \dots, \psi_M$ are a fixed b-splines basis functions and γ_{lj} are unknown coefficients. The density of ϵ is assumed to be a finite mixture of J normals, $f(u) = \sum_{j=1}^{J} p_j\phi(u; \mu_j, \sigma_1^2)$, where ϕ is the Gaussian density function. For illustration, we fix $J = M = 5$ and use the following priors: $\alpha, \beta_j \overset{\text{iid}}{\sim} \text{Normal}(0, 100^2)$; $\gamma_{lj} \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_2^2)$; $\mu_j \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_3^2)$; $(p_1, \dots, p_J) \sim \text{Dirichlet}(1, \dots, 1)$; and $\sigma_k^2 \overset{\text{iid}}{\sim} \text{InvGamma}(0.1, 0.1)$. Censoring is handled using standard data augmentation methods (Tanner and Wong 1987), and given the complete imputed dataset all parameters have conjugate full conditional distributions leading to straightforward Gibbs sampling. We generate 5000 MCMC samples and discard the first 2000 as burn-in. For the three-stage clinical trial data described below, this requires around 20 sec on a standard PC (as opposed to approximately 50 min for the fully nonparametric model of XMWT).

We compare this model to the DDP-GP method of XMWT in a suite of simulation experiments. To explore the effects of different types of model misspecification we fit the above model with linear ("Ln") mean, that is, $M = 1$ and $b_l(x_l) = x_l$, and nonlinear ("NLn") mean, $M = 5$ as above, and with Gaussian ("Gs") residuals, that is, $J = 1$ and $\mu_1 = 0$, and a mixture of J Gaussian ("MGs") densities as above. This gives four models for comparison in Figures 1, 2, and 3. For all scenarios we generate and fit 100 simulated datasets and report boxplots of the estimated sampling distribution of the difference between causal effect estimation errors as in Figure 4 of XMWT.

2.2. Setting 1: Single-Stage Additive Model

We simulated data as in Simulation 2 in XMWT. The model in (1) is applied with $x = (Z, W, L)$. As in our simplified model (1), the mean under this scenario is a nonlinear function of the covariates and the residual distribution is a mixture of normals that does not depend on the covariates. As expected, the "NLnMGs" model with nonlinear mean and non-Gaussian errors yields the most accurate estimates (Figure 1, left). Surprisingly, even for the complex data-generation scheme the linear Gaussian model ("LnGs") outperforms the DDP-GP.

2.3. Setting 2: Single-Stage Nonadditive Model

To explore the extent to which violation of model assumptions degrades performance, we simulated data from an even more pathological case than Setting 1. In this setting, the covariates L and W and the treatment group indicator Z are generated as in Setting 1, but the responses are generated as

$$Y(Z) = P_Z(x)N(4, 1) + [1 - P_Z(x)]N(1, 1),$$

where $\eta = -0.2L + \sqrt{L} - 0.1W$, $\text{logit}[P_0(x)] = \eta - 1$, and $\text{logit}[P_1(x)] = \eta + 2$. This model cannot be written in the form of (1) because the mean is not an additive function of the covariates and the residual distribution depends on the covariates. In this case, the DDP-GP model outperforms all four variations of (1).

2.4. Setting 3: Multistage Linear Model

We also compared the semiparametric model in (1) with the DDP-GP using data simulated from the three-stage clinical trial specified by Simulation 4 in XMWT. The semiparametric model

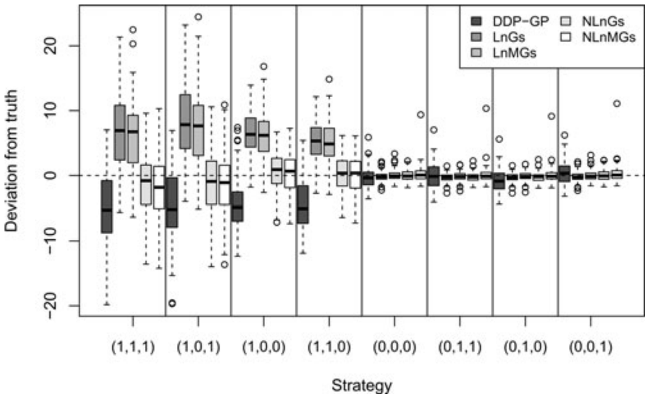


Figure 3. Simulation results for three-stage trials under Setting 4. The boxplots show the differences between estimated and true causal effects over the 100 simulated datasets using the Dependent Dirichlet Process model ("DDP-GP") and the semiparametric models with linear ("Ln") and nonlinear ("NLn") means and Gaussian ("Gs") and mixture of Gaussians ("MGs") residuals.

(1) was applied separately of the each arm of the trial. In this setting all log survival time are generated as Gaussian with mean set to a linear combination of the covariates. The best-performing model in Figure 2 is indeed the multivariate regression model “LnGs,” followed closely by the linear mean model with mixture of Gaussian residuals. The comparison of the DDP-GP to other nonlinear mean models is mixed; the DDP-GP model performs well in some cases and poorly in others.

2.5. Setting 4: Multi-Stage Nonlinear Model

In the final simulation, we considered a three-stage trial with nonlinear mean function by modifying Setting 3’s assumptions about the log survival distributions as follows. We add a quadratic term in the first stage, $x_i^{(0,R)} = x_i^{(0,C)} = (1, L_i^2, Z_i^1)$, with $\beta^{(0,R)} = (2, 0.01, 0)$ and $\beta^{(0,C)} = (1.5, 0.02, -0.8)$. We also added a square root transformation (which is valid because the log survival times are positive almost surely) at the second stage

$$x_i^{(R,D)} = \left(1, L_i^2, Z_i^1, \sqrt{\log\left(T_i^{(0,R)}\right)}, Z_i^{2,1}\right)$$

$$x_i^{(C,P)} = \left(1, L_i^2, Z_i^1, \sqrt{\log\left(T_i^{(0,C)}\right)}\right)$$

$$x_i^{(P,D)} = \left(1, L_i^2, Z_i^1, \sqrt{\log\left(T_i^{(0,C)}\right)}, \log\left(T_i^{(C,P)}\right), Z_i^{2,2}\right)$$

with $\beta^{(R,D)} = (-0.5, 0.015, 0.2, 1.0, 0.3)$, $\beta^{(C,P)} = (1.0, 0.03, 1.0, -1.0)$, and $\beta^{(P,D)} = (0.8, 0.01, 1.5, -1.0, 0.7, 0.5)$. As shown in Figure 3, the full “NlnMGs” model consistently outperforms the DDP-GP model, and the multiple linear model “LnGs” also often produces smaller errors than the DDP-GP.

2.6. Summary

While the full BNP model is theoretically appealing, estimating a conditional density that varies over several covariates is a daunting task. These additional simulations illustrate that for the small datasets considered here, quite often simpler models yield faster and more stable estimates than the full BNP model. In some nonlinear and/or non-Gaussian cases, we find that even multiple linear regression outperforms the DDP-GP approach. Therefore, despite the elegance of the full BNP approach, we conclude that there is still room for standard model fitting combined with careful regression diagnostics.

3. Nonparametric Bayesian Methods for Policy-Search

Policy-search methods use a flexible model for the mean under any regime and estimate the optimal regime with the maximizer of this fitted model over a prespecified class of regimes. NPB methods are well-suited to policy-search as they permit flexible modeling of the mean outcome (or other suitable functional) and also allow Bayesian machinery to be leveraged to conduct inference.

To simplify our development, we assume a two-stage decision problem with no censoring; though the transition times are random, possibly outcome-dependent, and not all patients progress through both stages. The extension to censored data is straightforward. We assume that the observed data are $\mathcal{D} = \{(\boldsymbol{X}_{1,i}, A_{1,i}, T_{1,i}, \boldsymbol{X}_{2,i}, A_{2,i}, T_{2,i})\}_{i=1}^n$ which comprise n independent, identically distributed trajectories of the form $(\boldsymbol{X}_1, A_1, Y_1, \boldsymbol{X}_2, A_2, Y_2)$, where: $\boldsymbol{X}_1 \in \mathbb{R}^{p_1}$ denotes baseline patient covariates; $A_1 \in \mathcal{A}_1$ denote the first treatment actually received; $T_1 \in \mathbb{R}_+$ denotes the time until death or transition to the second treatment stage; $\boldsymbol{X}_2 \in \mathbb{R}^{p_2}$ denotes interim covariate information collected during the first treatment period, for example, disease progression, remission, etc.; $A_2 \in \mathcal{A}_2$ denotes the second treatment actually received; and $T_2 \in \mathbb{R}_+$ denotes time until death in the second stage. Define $\boldsymbol{H}_1 = \boldsymbol{X}_1$ and $\boldsymbol{H}_2 = (\boldsymbol{X}_1^\intercal, A_1, T_1, \boldsymbol{X}_2^\intercal)^\intercal$ so that $\boldsymbol{H}_t$ is the information available to the decision maker at time $t = 1, 2$.

A treatment regime is a pair of functions, $\boldsymbol{\pi} = (\pi_1, \pi_2)$, one per stage of intervention, so that under $\boldsymbol{\pi}$ a patient presenting with $\boldsymbol{H}_t = \boldsymbol{h}_t$ at time t is recommended treatment $\pi_t(\boldsymbol{h}_t)$. We restrict attention to regimes in a prespecified class Π which might be constrained to ensure parsimony or interpretability; we assume that all regimes are feasible (Schulte et al. 2014) so that any $\boldsymbol{\pi} \in \Pi$ satisfies $\pi_t(\boldsymbol{H}_t) \in \psi_t(\boldsymbol{H}_t)$ with probability one, where $\psi_t(\boldsymbol{h}_t) \subseteq \mathcal{A}_t$ is the set of treatments that can be feasibly (ethically) assigned to a subject with $\boldsymbol{H}_t = \boldsymbol{h}_t$ for $t = 1, 2$. To define an optimal treatment regime we use the language of potential outcomes (Rubin 1978; Splawa-Neyman, Dabrowska, and Speed 1990). Define $\boldsymbol{H}_2^*(a_1)$ to be the potential second stage history under first-stage treatment a_1, $T_2^*(a_1, a_2)$ the potential time in the second stage under treatment sequence (a_1, a_2). Let $D^*(a_1) = D\{\boldsymbol{H}_2^*(a_2)\}$ denote the indicator that a patient transitions to the second stage; $T_2^*(a_1, a_2)$ is only defined if $D^*(a_1) = 0$. In concordance with XMWT, the potential outcome of interest is $T_1^*(a_1) + T_2^*(a_1, a_2)\{1 - D^*(a_1)\}$. The expected outcome under a regime $\boldsymbol{\pi}$ is

$$V(\boldsymbol{\pi}) = \mathbb{E}\left(\sum_{(a_1,a_2)} \left[T_1^*(a_1) + T_2^*(a_1, a_2)\left\{1 - D^*(a_1)\right\}\right] \right.$$
$$\left. \times\ 1_{a_1=\pi_1(\boldsymbol{H}_1)} 1_{a_2=\pi_2\left\{\boldsymbol{H}_2^*(a_1)\right\}} \right).$$

The optimal regime, say $\boldsymbol{\pi}^{\text{opt}} \in \Pi$, satisfies $V(\boldsymbol{\pi}^{\text{opt}}) \geq V(\boldsymbol{\pi})$ for all $\pi \in \Pi$.

To express $\boldsymbol{\pi}^{\text{opt}}$ in terms of the data generating model we make the following assumptions which are standard in the context of estimating optimal treatment regimes (Robins 2004; Schulte et al. 2014): (A1) consistency, $\boldsymbol{H}_2 = \boldsymbol{H}_2^*(A_1)$, $T_2 = T_2^*(A_1, A_2)$; (A2) sequential ignorability, $\{\boldsymbol{H}_2^*(a_1), T_2^*(a_1, a_2)\}_{(a_1,a_2)\in\mathcal{A}_1\times\mathcal{A}_2} \perp A_t | \boldsymbol{H}_t$; and (A3) positivity, $P(A_t = a_t | \boldsymbol{H}_t = \boldsymbol{h}_t) \geq \epsilon$ for some $\epsilon > 0$, all $a_t \in \psi_t(\boldsymbol{h}_t)$, and almost all $\boldsymbol{h}_t$ for $t = 1, 2$. In addition, we make the stable unit treatment value assumption (SUTVA). Under the foregoing assumptions, it can be shown that for any $\pi \in \Pi$

$$
\begin{aligned}
V(\boldsymbol{\pi}) = \int [t_1 + t_2(1-d)]\, & f_{T_2|_{H_2},A_2}(t_2|\boldsymbol{h}_2, a_2)\delta_{\pi_2(\boldsymbol{h}_2)}(a_2) \\
& f_{H_2|A_1,H_1}(\boldsymbol{h}_2|a_1, \boldsymbol{h}_1)\delta_{\pi_1(\boldsymbol{h}_1)}(a_1) f_{H_1}(\boldsymbol{h}_1) \\
& \times d\lambda(t_2, \boldsymbol{h}_2, \boldsymbol{h}_1, a_1, a_2), \qquad (2)
\end{aligned}
$$

where λ is a dominating measure, $\delta_u(\cdot)$ denotes a point mass at u, and $f_{W|Z}(w|z)$ denotes the conditional density of W given Z, and we have used the fact that t_1, d are functions of $\boldsymbol{h}_2$. Given the expression for $V(\boldsymbol{\pi})$ in (2) a natural approach to estimating $\boldsymbol{\pi}^{\text{opt}}$ is to first construct estimators $\widehat{f}_{T_2|\mathbf{H}_2,A_2}$, $\widehat{f}_{\mathbf{H}_2|\mathbf{H}_1,A_1}$, and $\widehat{f}_{\mathbf{H}_1}$ of $f_{T_2|\mathbf{H}_2,A_2}$, $f_{\mathbf{H}_2|\mathbf{H}_1,A_1}$, and $f_{\mathbf{H}_1}$ and subsequently

$$
\begin{aligned}
& \widehat{\boldsymbol{\pi}}_n^{\text{plug}-\text{in}} \\
& = \arg\max_{\pi\in\Pi} \int [t_1 + t_2(1-d)]\, \widehat{f}_{T_2|H_2,A_2}(t_2|\boldsymbol{h}_2, a_2)\delta_{\pi_2(\boldsymbol{h}_2)}(a_2) \\
& \widehat{f}_{H_2|A_1,H_1}(\boldsymbol{h}_2|a_1, \boldsymbol{h}_1)\delta_{\pi_1(\boldsymbol{h}_1)}(a_1)\widehat{f}_{H_1}(\boldsymbol{h}_1) \\
& \times d\lambda(t_2, \boldsymbol{h}_2, \boldsymbol{h}_1, a_1, a_2). \qquad (3)
\end{aligned}
$$

Of course one of the primary difficulties with this approach is estimating the requisite conditional densities which may be high-dimensional (the problem becomes even worse as the number of treatment stages increases and the dimension of the history grows). An appeal of nonparametric Bayesian (NPB) methods is the ability to impose structure on these densities by anchoring them to parsimonious parametric densities through the prior; thus, at least in principle, NPB methods permit as much flexibility as the data can shoulder. One approach would be to use a use a mixture of Gaussian processes prior anchored to normal (log) linear models, that is, one might postulate the following models

$$
\begin{aligned}
\log(T_2) &= \boldsymbol{H}_{2,0}^{\intercal}\beta_{2,0} + A_2\boldsymbol{H}_{2,1} + \delta_2(\boldsymbol{H}_2, A_2) \\
\boldsymbol{H}_2 &= \Omega_{1,0}\boldsymbol{H}_1 + A_1\Omega_{1,1}\boldsymbol{H}_1 + \delta_1(\boldsymbol{H}_1, A_1) \\
\log(T_1) &= \boldsymbol{H}_{1,0}^{\intercal}\beta_{1,0} + A_1\boldsymbol{H}_{1,1}^{\intercal}\beta_{1,1} + \delta_0(\boldsymbol{H}_1, A_1),
\end{aligned}
$$

where $\boldsymbol{H}_{j,0}, \boldsymbol{H}_{j,1}$ are known features constructed from $\boldsymbol{H}_j$, $\beta_{2,0}, \beta_{2,1}, \Omega_{1,0}, \Omega_{1,1}, \beta_{1,0}, \beta_{1,1}$ are unknown parameters, and $\delta_2, \delta_1, \delta_0$ are errors with unspecified distributions. The distribution of $\boldsymbol{H}_1$ can be modeled using standard methods. Using the DDP-GP prior of XMWT for δ_j with a strong prior on the covariance kernel of each Gaussian process being identically zero corresponds to strong weighting on the (log)linear models with normal errors.

A schematic for NPB policy-search is as follows. Postulate a prior $\mathcal{G}$ for $\boldsymbol{f} = (f_{T_2|\mathbf{H}_2,A_2}, f_{\mathbf{H}_2|\mathbf{H}_1,A_1}, f_{\mathbf{H}_1})$, and let $P(\boldsymbol{f}|\mathcal{D})$ denote the posterior distribution of f. The an NPB estimator could be taken to be the posterior mode

$$
\begin{aligned}
\widehat{\boldsymbol{\pi}}_n^{\text{NPB}-\text{mode}} &= \arg\max_{\pi\in\Pi} \mathbb{E}\left\{V_{\boldsymbol{f}}(\pi)\middle|\mathcal{D}\right\} \\
&= \arg\max_{\pi\in\Pi} \int \left(\int [t_1 + t_2(1-d)]\, f_{T_2|H_2,A_2} \right. \\
&\quad \times (t_2|\boldsymbol{h}_2, a_2)\delta_{\pi_2(\boldsymbol{h}_2)}(a_2) \\
&\quad f_{H_2|A_1,H_1}(\boldsymbol{h}_2|a_1, \boldsymbol{h}_1)\delta_{\pi_1(\boldsymbol{h}_1)} \\
&\quad \left. \times (a_1) f_{H_1}(\boldsymbol{h}_1) d\lambda(t_2, \boldsymbol{h}_2, \boldsymbol{h}_1, a_1, a_2) \right) dP(\boldsymbol{f}|\mathcal{D}),
\end{aligned}
$$

where $V_{\boldsymbol{f}}(\pi)$ is the marginal mean outcome computed using (2) with density $\boldsymbol{f}$. An advantage of this form is that it permits: (i) estimation within a prespecified class which may be restricted to be interpretable, respect cost constraints, etc.; (ii) inferences that are notoriously difficult under a frequentist paradigm; and (iii) estimation of regimes that optimize a functional other than a mean, for example, a quantile or measure of variability. For example, to illustrate (ii) suppose Π_1 is a class of regimes composed of decision rules representable as trees and Π_2 is the space of all feasible regimes, then one can compute a posterior credible set for the difference $\max_{\pi\in\Pi_2} V_{\boldsymbol{f}}(\pi) - \max_{\pi\in\Pi_1} V_{\boldsymbol{f}}(\pi)$. If this credible interval lies above zero then one may conclude that the class of regimes Π_1 is not sufficiently expressive and opt to enrich this class. In contrast, standard asymptotic approaches for constructing a confidence interval, for example, the bootstrap or series approximations, for this difference cannot be applied without modification because the marginal mean outcome $V(\pi)$ is a nonsmooth functional of the underlying generative distribution (e.g., Van Der Vaart 1991; Hirano and Porter 2009; Laber et al. 2014; Chakraborty, Laber, and Zhao 2014). It would be interesting to investigate the operating characteristics of an NPB approach to these inferential problems.

4. Discussion

NPB methods have great potential for estimation of and inference for optimal treatment regimes. Under suitable identifiability conditions, they can be used to estimate regimes that optimize essentially any smooth functional of the potential outcome distribution. Furthermore, the underlying Bayesian machinery makes it straightforward to construct credible sets for functionals for which standard frequentist inference procedures cannot be applied without modification. However, it is not apparent whether these credible sets will have good frequentist operating characteristics. Investigation of frequentist properties of NPB methods in the context of treatment regimes would be of great interest.

5. Simulation of Comparing Several Models

5.1. *Single-Stage Regimes*

We simulated the data under the same setting as indicated in simulation 2 in the article. There are two continuous covariates, $x_i = (L_i, W_i)$, for n = 100 patients, $i = 1, \ldots, n$. We generated L_i from a mixture of normals, $L_i \sim \frac{1}{2}N(40, 10^2) + \frac{1}{2}N(20, 10^2)$. We generated W_i from a uniform with zero mean and unit standard deviation, $W_i \sim \text{Unif}(-\sqrt{12}, \sqrt{12})$. We generated the treatment indicators $Z_i \in \{0 = \text{control}, 1 = \text{experimental}\}$

using the modified logistic regression model

$$p(Z_i = 1|L_i, W_i) = \begin{cases} 0.05 & \text{if}\{1 + \exp[-2(L_i - 30)/10]\}^{-1} \leq 0.05 \\ 0.95 & \text{if}\{1 + \exp[-2(L_i - 30)/10]\}^{-1} \geq 0.95 \\ \{1 + exp[-2(L_i - 30)/10]\}^{-1} & \text{otherwise} \end{cases}$$

We generated patients' responses differently for different treatments from

$$Y(1) \sim \frac{1}{2}N(3 - 0.2L + \sqrt{L} - 0.1W, \sigma)$$
$$+\frac{1}{2}N(2 - 0.2L + \sqrt{L} - 0.1W, \sigma)$$

and

$$Y(0) \sim N(-0.2L + \sqrt{L} - 0.1W, \sigma)$$

with $\sigma = 0.4$. Under the simulation truth the treatment effect, $E[Y(1) - Y(0)|x = (L, W)] = 2.5$, is constant across L, W.

We implemented inference for $F(Y_i|x_i, Z_i)$ using the DDP-GP model. Let $\hat{Y}_i(z) = E(Y_{n+1}|L_{n+1} = L_i, W_{n+1} = W_i, Z_{n+1} = z, \text{data})$ denote the posterior expected response for a future patient. The estimated average treatment effect(ATE) is defined as $\frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_i(1) - \hat{Y}_i(0)]$.

We then applied linear regression method to the simulated data to estimate the average treatment effect. We used data of the patients who received treatment 1 to fit the linear model of treatment 1: $Y_i(1) = \beta_{10} + \beta_{11}L_i + \beta_{12}W_i + \epsilon_i$. We used data of patients with treatment 0 to fit: $Y_i(0) = \beta_{00} + \beta_{01}L_i + \beta_{02}W_i + \epsilon_i$. Denoting the least-square estimates by $\hat{\beta}_{zj}$ for $z = 0, 1$ and $j = 0, 1, 2$. We applied two fitted regression model to estimate mean potential repsonse of each patient given treatment 1 and treatment 0: $\hat{E}\{Y_i(z)\} = \hat{\beta}_{z0}L_i + \hat{\beta}_{z2}W_i$, for $z = 0, 1$, and $i = 1, ..., n$. Then we can estimate average treatment effect as $\frac{1}{n}\sum_i[\hat{E}\{Y_i(1)\} - \hat{E}\{Y_i(0)\}]$.

Then we used IPTW to estimate ATE. The IPTW method assigns each patients i a weight b_i equal to the inverse of an estimate of $p(Z_i|x_i)$. Let $\pi_i = p(Z_i = 1|x_i)$. An estimate of π_i is obtained by fitting a logistic regression model. When $Z_i = 1$, $b_i = 1/\pi_i$; when $Z_i = 0$, $b_i = 1/(1 - \pi_i)$. We define the IPTW mean outcome estimator

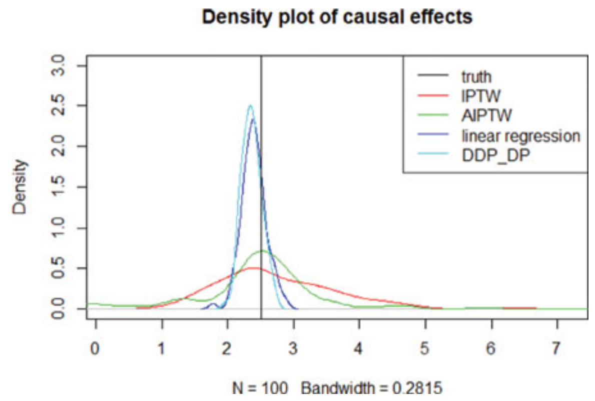
$$\text{IPTW}(Z = z) = \frac{\sum_i I(Z_i = z)b_iY_i}{\sum_i I(Z_i = z)b_i}$$

and corresponding average treatment effect estimate $\text{ATE}_{\text{IPTW}} = \text{IPTW}(Z = 1) - \text{IPTW}(Z = 0)$.

Finally, We used AIPTW, which is a doubly robust generalization of the IPTW, to estimate ATE.

$$\text{ATE}_{\text{AIPTW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{I(Z_i = 1)Y_i}{\hat{\pi}_i} - \frac{I(Z_i = 0)Y_i}{1 - \hat{\pi}_i}\right]\right.$$
$$- \frac{I(Z_i = 1) - \hat{\pi}_i}{\hat{\pi}_i(1 - \hat{\pi}_i)}\left[(1 - \hat{\pi}_i)\hat{E}(Y_i|Z_i = 1, x_i)\right.$$
$$\left.\left.+\hat{\pi}_i\hat{E}(Y_i|Z_i = 0, x_i)\right]\right\}.$$

We simulated 100 trials and draw the density plot of estimated regime effects by DDP-GP linear regression, IPTW and AIPTW in 100 trials. The truth is indicated by a black vertical line.



Under this simulation setting, it seems linear regression worked as well as DDP-GP model, which are less variable and close to true effects. However, DDP-GP model took much longer time to fit by MCMC sampling.

5.2. Multistage Regimes

Also, we simulated the data under the same setting as indicated in simulation 4 in the article. We simulated samples of size $n = 200$ and patients' blood glucose values $L_i \sim N(100, 10^2)$. Patients initially were randomized equally between two induction therapies $Z^1 \in \{a_1, a_2\}$. Patients who were resistant (R) to the assigned induction therapies were then assigned salvage treatment $Z^{2,1} \in \{b_{11}, b_{12}\}$. Salvage treatments were randomized using the rule $p(Z^{2,1} = b_{11}|L_i) = 0.8(L_i < 100) + 0.2I(L_i \geq 100)$. Patients who achieved C and subsequently suffered disease progression (P), were given salvage treatment $Z^{2,2} \in \{b_{21}, b_{22}\}$, using $p(Z^{2,2} = b_{21}|L_i) = 0.2(L_i < 100) + 0.85I(L_i \geq 100)$. The survival time for each patient was evaluated as

$$T_i = \begin{cases} T_i^{(0,R)} + T_i^{(R,D)} & \text{if patient } i \text{ had sequence} \\ & \qquad (L, Z^1, T^{(0,R)}, Z^{2,1}) \\ T_i^{(0,C)} + T_i^{(C,P)} & \text{if patient } i \text{ had sequence} \\ \quad +T_i^{(P,D)} & \qquad (L, Z^1, T^{(0,C)}, T^{(C,P)}, Z^{2,2}). \end{cases}$$

We simulated t$T^{(0,R)} \sim LN(\beta^{(0,R)}x_i^{(0,R)}, \sigma^{(0,R)})$ and $T^{(0,C)} \sim LN(\beta^{(0,C)}x_i^{(0,C)}, \sigma^{(0,C)})$, where $\beta^{(0,R)} = (2, 0.02, 0)$, $\beta^{(0,C)} = (1.5, 0.03, -0.8)$, with $x_i^{(0,R)} = (1, L_i, Z_i^1)$ for $k \in \{(0, R), (0, C)\}$. For the second stage transitions $k \in \{(R, D), (C, P), (P, D)\}$, we generated $T_i^k \sim LN(\beta^k x_i^k, \sigma^k)$, where $\beta^{(R,D)} = (-0.5, 0.03, 0.2, 0.5, 0.3)$, $x_i^{(R,D)} = (1, L_i, Z_i^1, \log(T_i^{(0,R)}), Z_i^{2,1})$, $\beta^{(C,P)} = (1, 0.05, 1, -0.6)$, $x_i^{(C,P)} = (1, L_i, Z_i^1, \log(T_i^{(0,C)}))$, $\beta^{(C,D)} = (0.8, 0.04, 1.5, -1, 0.5, 0.5)$, $x_i^{(P,D)} =$

$(1, L_i, Z_i^1, \log(T_i^{(0,C)}), \log(T_i^{(C,P)}), Z_i^{2,2})$. We simulated 100 trials.

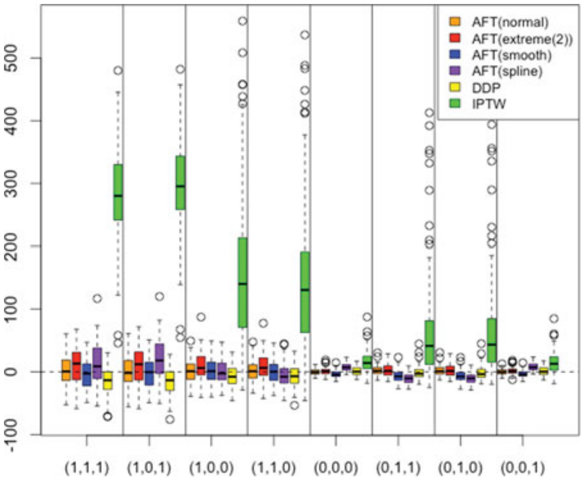
We used several different models to fit the simulated data and estimated mean survival time for each DTR $(Z^1, Z^{2,1}, Z^{2,2})$ using fitted models. The models we have implemented including DDP-GP model proposed in the article, IPTW and AFT regression models with different specification of error distribution. The AFT model is in the form

$$\log(T_i) = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + \sigma\epsilon_i.$$

We assume two parametric error(ϵ) distributions for AFT models: normal distribution and extreme value distribution with two parameters. Accordingly, the distribution of T is lognormal distribution and weibull distribution, respectively. We also implemented a semiparametric AFT model with smooth error distribution, which is expressed as a mixture of G-spines. We compared the estimated treatment regime effect with the truth. The differences are summarized in the boxplot, which is arranged by different methods and by eight possible DTRs.

In our simulation scenario, the true distribution for the error term is normal distribution. We can see from the boxplot, when the error distribution is correctly specified (normal) in the AFT model, the estimated mean survival time is very close to the truth; when the error distribution is not correctly specified (extreme value) in the AFT model, the estimated mean survival time is a little bit biased. AFT model with error distribution expressed as a mixture of G-splines is more flexible. We don't need to specify any parametric error distribution for the AFT model so as to reduce bias caused by misspecification.

Compared with DDP-GP model proposed in the article, semiparametric AFT model takes much shorter time to fit and performs even a little bit better in this simulation scenario. Maybe it is because survival time between states is linearly generated in the simulation setting. DDP-GP model might be able to accommodate to more complicated data structure better.



6. Personalized Treatment Regime

The proposed approach in the article gave one general choice of treatment regimes for all the patients since they evaluated mean overall survival time by averaging over the empirical covariate distribution. We think we can apply the DDP-GP approach in the context of policy search algorithm by evaluating individual-specific treatment regime and gave treatment recommendations for patients based on their baseline covariates and disease progression.

References

Chakraborty, B., Laber, E. B., and Zhao, Y.-Q. (2014), "Inference About the Expected Performance of a Data-Driven Dynamic Treatment Regime," *Clinical Trials*, 11, 408–417. [939]

Doove, L., Dusseldorp, E., Van Deun, K., and Van Mechelen, I. (2015), "A Novel Method for Estimating Optimal Tree-Based Treatment Regimes in Randomized Clinical Trials," Abstract, available at *https://lirias.kuleuven.be/handle/123456789/493191*. [936]

Hirano, K., and Porter, J. R. (2009), "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701. [939]

Kang, C., Janes, H., and Huang, Y. (2014), "Combining Biomarkers to Optimize Patient Treatment Recommendations," *Biometrics*, 70, 695–707. [936]

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014), "Interactive Model Building for Q-Learning," *Biometrika*, 101, 831–847. [936]

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014), "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, 8, 1225. [939]

Laber, E. B., and Zhao, Y. Q. (2015), "Tree-Based Methods for Individualized Treatment Regimes," *Biometrika*, 102, 501–514. [936]

Moodie, E. E. M., Dean, N., and Sun, Y. R. (2013), "Q-Learning: Flexible Learning About Useful Utilities," *Statistics in Biosciences*, 6, 1–21. [936]

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010), "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main content," *The International Journal of Biostatistics*, 6. [936]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics* (Vol. 179 of *Lecture Notes in Statistics*), eds. D. Y. Lin, and P. J. Heagerty, New York: Springer, pp. 189–326. [938]

Robins, J. M., Orellana, L., and Rotnitzky, A. (2008), "Estimation and Extrapolation of Optimal Treatment and Testing Strategies," *Statistics in Medicine*, 27, 4678–4721. [936]

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58. [938]

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014), "Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Statistical Science*, 29, 640–661. [936,938]

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5, 465–472. [938]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [937]

Taylor, J. M. G., Cheng, W., and Foster, J. C. (2015), "Reader Reaction to "A Robust Method for Estimating Optimal Treatment Regimes" by Zhang et al. (2012)," *Biometrics*, 71, 267–273. [936]

Van Der Vaart, A. (1991), "On Differentiable Functionals," *The Annals of Statistics*, 178–204. [939]

Wahed, A. S., and Thall, P. F. (2013), "Evaluating Joint Effects of Induction–Salvage Treatment Regimes on Overall Survival in Acute Leukaemia," *Journal of the Royal Statistical Society*, Series C, 62, 67–83. [936]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [936]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013), "Robust Estimation Of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100, 681–694. [936]

Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015), "Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes," *Biometrics*, 71, 895–904. [936]

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009), "Reinforcement Learning Design for Cancer Clinical Trials," *Statistics in Medicine*, 28, 3294–3315. [936]

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [936]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [936]

Zhao, Y. Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2014), "Doubly Robust Learning for Estimating Individualized Treatment With Censored Data," *Biometrika*, 102, 151–168. [936]

Zhou, X., and Kosorok, M. R. (2016), "Nearest Neighbor Rules for Optimal Treatment Regimes," *Under Review*, 1, 1–15. [936]

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015), "Residual Weighted Learning for Estimating Individualized Treatment Rules," *Journal of the American Statistical Association*, forthcoming. DOI:10.1080/01621459.2015.1093947. [936]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1200914

Comment

Jingxiang Chen[a], Yufeng Liu[b], Donglin Zeng[a], Rui Song[c], Yingqi Zhao[d], and Michael R. Kosorok[e]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [b]Department of Statistics and Operations Research, Department of Biostatistics, Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [c]Department of Statistics, North Carolina State University, Raleigh, NC, USA; [d]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; [e]Department of Biostatistics, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

ABSTRACT
Xu, Müller, Wahed, and Thall proposed a Bayesian model to analyze an acute leukemia study involving multi-stage chemotherapy regimes. We discuss two alternative methods, Q-learning and O-learning, to solve the same problem from the machine learning point of view. The numerical studies show that these methods can be flexible and have advantages in some situations to handle treatment heterogeneity while being robust to model misspecification.

KEYWORDS
Dynamic treatment regimes; Multi-stage chemotherapy regimes; O-learning; Q-learning

1. Introduction

There is increasing recognition that optimal therapies should account for individual heterogeneity and be adaptive over time. Thus, in recent clinical trials and observational studies, dynamic treatment regimes (DTR) have drawn significant attention. We congratulate Xu, Müller, Wahed, and Thall on their contribution in proposing a novel applicable and competitive method for analyzing the clinical trial for acute leukemia involving multi-stage chemotherapy regimes. Specifically, there is a sequence of treatments beginning at induction and followed by subsequent salvage therapies which depend on disease stage. The combination of these therapies affect patient overall survival time, which consists of the sum of the transition times between each involved disease stage. To evaluate joint effects of induction-salvage therapies on patient survival, Xu et al. (2016) build a Bayesian nonparametric survival regression model, assuming a Dependent Dirichlet Process prior with Gaussian Process (DDP-GP) base for each transition time. The numerical results show that such a Bayesian paradigm can produce an accurate estimate for the joint effects of induction-salvage therapies when compared with IPTW and AIPTW (Zhang et al. 2013). Moreover, the authors indicate that such a model could be extended to the situation where the therapy effect is heterogeneous in the population.

In addition to the Bayesian methods, there are some recently developed machine learning tools that have achieved success in estimating individualized DTRs which are somewhat more frequentist in perspective. In this article, we would like to introduce two representatives, Q-learning and O-learning, and illustrate how they can be used to solve the same problem addressed in Xu et al. (2016). A major advantage of these two alternative approaches is their relaxed assumptions on the joint distribution of feature variables and clinical outcomes such as survival time. Specifically, one does not need to model the entire process to construct the optimal treatment regimes. For Q-learning, conditional expectations are modeled but not the entire process. For O-learning, only the treatment decision boundary and propensity score (when needed) are modeled. These reductions in modeling requirements can be significant relative to approaches which require modeling of the entire process. In this article, we investigate the performances of Bayesian DDP-GP proposed in Xu et al. (2016), Q-learning and O-learning when certain assumptions fail, including (1) when the treatment effect is heterogeneous in the population and (2) when the log transition times are not Gaussian.

The article is organized as follows. In Sections 2 and 3, we briefly introduce the general ideas of Q-learning and O-learning and focus on how to modify them for the DTR setup in Xu et al.

CONTACT Michael R. Kosorok kosorok@unc.edu Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599.

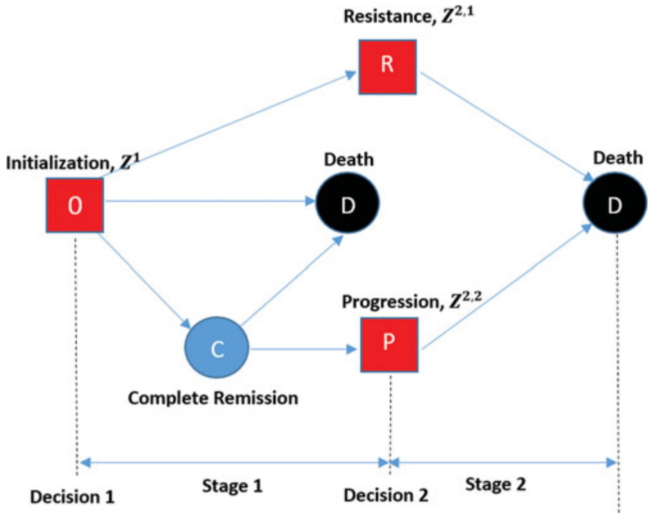


Figure 1. Redefinition of the Scheme under the proposed Q-learning Framework. The states in red square boxes (i.e., initialization, resistance and progression) are the treatment decision-making points that are used to split the two stages. Complete remission (C) is not considered as a splitting point since no decision action can be taken. Censoring time could happen at the end of each stage.

(2016). In Section 4, we present simulation studies comparing the Bayesian DDP-GP model and the proposed methods when the assumptions hold or fail. In Section 5, we apply the proposed Q-learning to the multi-stage acute leukemia trial data. We conclude with a brief discussions in Section 6.

2. Q-Learning in Finding the Dynamic Treatment Regimes

Q-learning is a reinforcement learning method that can be used to estimate the optimal personalized treatment strategy in a sequence of clinical decisions over time (Murphy 2003). It aims to estimate a sequence of time-varying Q functions by taking the patients' state and the clinical decision at each stage as inputs. In the end, Q-learning returns the estimated Q function and the corresponding optimal treatments for each stage. Next, we present how to adapt the Q-learning to solve the DTR problems discussed in Xu et al. (2016) and use their acute Leukemia example for illustration.

Using similar notations as in Xu et al. (2016), we let T^k represent the transition times from the n_T possible state transitions and let k be one of the following transitions in the acute Leukemia example: $(0, R)$, $(0, D)$, $(0, C)$, (C, D), (R, D) , (C, P), and (P, D). In addition, we use Z^1, $Z^{2,1}$ and $Z^{2,2}$ to represent the indicator of the frontline therapy, the salvage with High Dose Ara-C (HDAC) for those having resistance and the other salvage for those patients who first achieve a complete remission but suffer progressive disease later. To explain the proposed Q-learning, we need to clarify the definition of stages and states under our framework. We define the end of the stage as either the decision making time point or the failure time point. To be specific, the first stage starts at the beginning of the study when patients are randomly assigned to frontline therapy groups and ends when one of the following events occur: resistance (R), progression (P), death (D) and missing to follow-up. The reason that we do not mark complete remission (C) as the start of the second stage is that no decision action can be taken at this point. Figure 1 illustrates our definition above, where the states in red square boxes are all the decision making points so that the second stage begins when either resistance or progression occurs. Furthermore, we allow data censoring to happen at the end of each stage, that is, before resistance, before progression or before death.

Based on the defined stages in Figure 1, we introduce the steps of a backward Q-learning strategy in finding the optimal therapy at each stage. For simplicity, we only consider the two stage setting and similarly one can extend the method for multiple-stage situations. Starting at the second stage, we assume that the two state transitions (i.e., (R, D) and (P, D)) are independent of each other. In this way, treating the two transition times $T^{(R,D)}$ and $T^{(P,D)}$ as the response, we can formulate the optimal therapy estimation problem in Stage 2 as follows:

$$\begin{aligned} \hat{\pi}_{2,1} &= \underset{Z^{2,1}}{\arg\max} \left\{ \hat{Q}_{2R}\left(\mathcal{H}^{1R}, Z^{2,1}\right) \right\}, \text{ for the resistance group,} \\ \hat{\pi}_{2,2} &= \underset{Z^{2,2}}{\arg\max} \left\{ \hat{Q}_{2P}\left(\mathcal{H}^{1P}, Z^{2,2}\right) \right\}, \text{ for the progression group,} \end{aligned} \tag{1}$$

where the two Q functions (i.e., $\hat{Q}_{2R}$ and $\hat{Q}_{2P}$) are respectively for the resistance group and progression group at the beginning of Stage 2, and allow both of them to have either a parametric or nonparametric form. The $\mathcal{H}^{1R}$ and $\mathcal{H}^{1P}$ in (1) denote all the information at the end of the first stage for the two corresponding groups. They may contain the baseline covariates, initial treatment, observed time to event during the first stage and all the measurements at the end of the first stage. Based on $\hat{\pi}_{2,1}$ and $\hat{\pi}_{2,2}$, we define the estimated value function at Stage 2 as $\hat{\mathcal{V}}_2 = I_{RD} \cdot \hat{Q}_{2R}(\mathcal{H}^{1R}, \hat{\pi}_{2,1}) + I_{PD}\hat{Q}_{2P}(\mathcal{H}^{1P}, \hat{\pi}_{2,2})$, where the indicator functions, I_{RD} and I_{PD}, indicate whether the patient is in the resistance or progression state at the beginning of Stage 2. The quantity $\hat{\mathcal{V}}_2$ indicates the expected survival time at Stage 2 for a given individual assuming that the optimal treatment is given at Stage 2.

Once the optimal therapy is estimated for the second stage, we consider adding the transition time at Stage 1 into the response and then estimate the corresponding optimal treatment as follows. First, we compute the pseudovalue

$$\tilde{T} = I_D T^{(0,D)} + I_{RD}\left(T^{(0,R)} + \hat{\mathcal{V}}_2\right) + I_{PD}\left(T^{(0,P)} + \hat{\mathcal{V}}_2\right)$$

for each individual, where I_D indicates whether the patient has either failure or no follow-up in the first stage. Let $\mathcal{H}^0$ represents all the information in baseline covariate measurements and let d_1 is the possible decision action for the first clinical stage, which has the same parameter space as Z^1 in this case. The pseudovalue $\tilde{T}$ replaces the observed values of both $T^{(R,D)}$ and $T^{(P,D)}$ with the corresponding expected times if the optimal treatment were applied at the second stage. We then regress $\tilde{T}$ on $\mathcal{H}^0$ and d_1 to obtain the estimated Stage 1 Q-function $\hat{Q}_1(\mathcal{H}^0, d_1)$. Stage 1 optimal treatment is then estimated as

$$\hat{\pi}_1 = \underset{d_1}{\arg\max} \hat{Q}_1(\mathcal{H}^0, d_1). \tag{2}$$

We aim to find d_1 to maximize (2) and the maximal objective value is denoted as $\hat{\mathcal{V}}_1$. The base learner with the highest $\hat{\mathcal{V}}_1$ value would be desirable. For demonstration, we use linear regression and exponential survival regression as the two base learners of Q_{2R}, Q_{2P} and Q_1 in the numeric studies below.

The proposed Q-learning method can be quite flexible in certain situations. Specifically, since Q-learning does not fit the entire process of the transitions, we do not necessarily need any distributional assumption to build the model. Furthermore, the base learners in either (1) or (2) do not have to be linear or parametric. Thus, we can chose the model which fits the data the best. For example, some nonparametric learning tools could end up with high-prediction accuracy when the variable relationship is complex. Such tools include random forest, boosting and kernel methods (Hastie Tibshirani, and Friedman 2011). In addition, the heterogeneity of the treatment effect can also be detected by simply including the treatment-covariate interaction terms into the Q functions at each stage.

3. O-Learning in Finding the Dynamic Treatment Regimes

Estimating the overall treatment effect in a population is not always necessary when detecting the optimal treatment at each stage. Accordingly, another possible approach is one of the O-learning extensions to dynamic treatment regimes, that is, backward outcome weighted learning (BOWL, Zhao et al. 2015a). BOWL provides a new paradigm for framing the optimal DTR identification and formulates it into a weighted classification problem with the clinical outcome as weights. The estimation of BOWL proceeds backward to find the optimal treatment rule at each stage to maximize the cumulative rewards over the subsequent time. To apply BOWL to solve the problem discussed in Figure 1, we need to modify the steps by introducing the indicator functions used in the Q-learning approach above. Specifically, we first write the BOWL algorithm for the second stage as

$$
\begin{aligned}
f_{2R} &= \arg\min_{f} \mathbb{E}_n \left[\frac{T^{(R,D)}\phi(Z^{2,1} f(\mathcal{H}^{1R}))}{\pi_{2,1}(Z^{2,1}, \mathcal{H}^{1R})} + \lambda_2 \| f \|^2 \right], \\
&\qquad\qquad\qquad\qquad \text{for the resistance group}, \\
f_{2P} &= \arg\min_{f} \mathbb{E}_n \left[\frac{T^{(P,D)}\phi(Z^{2,2} f(\mathcal{H}^{1P}))}{\pi_{2,2}(Z^{2,2}, \mathcal{H}^{1P})} + \lambda_2 \| f \|^2 \right], \\
&\qquad\qquad\qquad\qquad \text{for the progression group}, \qquad (3)
\end{aligned}
$$

where the surrogate loss function $\phi(t) = \max(1 - t, 0)$, $\mathbb{E}_n$ denotes the empirical mean over the sample, $\pi_{2,1}(z, \mathcal{H}^1) = \Pr(Z^{2,1} = z|\mathcal{H}^1)$, $\pi_{2,2}(z, \mathcal{H}^1) = \Pr(Z^{2,2} = z|\mathcal{H}^1)$, $\| \cdot \|^2$ denotes the square of L_2 norm, and λ_2 is the tuning parameter that controls model complexity. After the classifiers f_{2R} and f_{2P} are obtained, the corresponding estimate of the optimal treatment rule for the second stage, d_2, can be calculated through $\hat{d}_2(\mathcal{H}^1) = I_{RD} \cdot I(f_{2R}(\mathcal{H}^{1R}) > 0) + I_{PD} \cdot I(f_{2P}(\mathcal{H}^{1P}) > 0)$. Based on $\hat{d}_2$, we have the classifier for the first stage as

$$
\begin{aligned}
f_1 = \arg\min_{f_1} \mathbb{E}_n \Bigg[\Bigg(& I_D \frac{T^{(0,D)}}{\pi_1(Z^1, \mathcal{H}^0)} \\
& + I_{RD} \frac{I(Z^{2,1} = \hat{d}_2(\mathcal{H}^1)) \cdot (T^{(0,R)} + T^{(R,D)})}{\pi_1(Z^1, \mathcal{H}^0)\pi_{2,1}(Z^{2,1}, \mathcal{H}^1)} \\
& + I_{PD} \frac{I(Z^{2,2} = \hat{d}_2(\mathcal{H}^1)) \cdot (T^{(0,P)} + T^{(P,D)})}{\pi_1(Z^1, \mathcal{H}^0)\pi_{2,2}(Z^{2,2}, \mathcal{H}^1)} \Bigg) \\
& \cdot \phi(Z^1 f_1(\mathcal{H}^0)) + \lambda_1 \| f_1 \|^2 \Bigg], \qquad (4)
\end{aligned}
$$

where $\pi_1(z, \mathcal{H}^0) = \Pr(Z^1 = z|\mathcal{H}^0)$, and λ_1 is the tuning parameter controlling model complexity of (4). We obtain the estimate of the optimal treatment rule d_1 for Stage 1 via $\hat{d}_1(\mathcal{H}^0) = I(f_1(\mathcal{H}^0) > 0)$. Essentially, BOWL aims to assign the patients having good clinical outcome to the same treatment they received and to assign the opposite treatment otherwise. The advantage of the adjusted BOWL is that its estimate is obtained under a nonparametric framework, so that BOWL can effectively handle the potentially complex relationships between sequential treatments and prognostic variables at each stage.

So far, the adjusted BOWL cannot be used directly for data with censoring. However, one can develop such an extension by estimating the distribution of censoring times in each stage as Zhao et al. (2015b) has done in the single stage scenario. In this article, we do not cover such an extension but only apply BOWL to simulated datasets which do not have censoring.

4. Simulation Studies

In this section, we compare the DDP-GP Bayesian model in Xu et al. (2016) with the adjusted Q-learning and O-learning introduced in Sections 2 and 3. Specifically, for Q Learning, we choose two popular base learners for the Q function: linear regression (Q-learn-1 in Table 1) and exponential survival regression (Q-learn-2) with the transition times as the response, as described previously. We include all the interaction terms between treatments and baseline covariates at each stage. To make a fair comparison, in addition to the original DDP-GP Bayesian model (DDP-GP-1), we also implement a modified version which has these interaction terms in the mean structure (DDP-GP-2). In the O-learning implementation, we treat both f_1, f_{2R}, and f_{2P} as linear classifiers for simplicity. Also for simplicity, we do not include censoring in the simulations.

We consider four simulation scenarios arising from Simulation 4 of Xu et al. (2016). First, we add a new variable S and consider both situations where the true model either includes or exclude interactions between S and the salvage treatment with HDAC. Second, we discuss the scenarios when the underlying Gaussian distribution assumption fails for the log survival time to examine model robustness against distribution misspecification. In addition, since the proposed O-learning is not yet capable of handling censoring, we always let the transition events happen before censoring for all the patients. In each simulation setting, we generate a single, fixed population set of size $N = 2000$ and then sample $n = 200$ training observations from this population 50 times. For each such sample, the selected methods are applied to the generated sample and then used to predict the optimal treatment for both the sample and the population. The model performance is then evaluated by the estimated value function $\hat{\mathcal{V}}_1$ for the combined sample and population groups. We now introduce the setting details for the four simulation cases as follows.

Simulation 1a: Gaussian distribution with no interaction term. Similar to Simulation 4 in Xu et al. (2016), we first generate the patients' baseline blood glucose L and the new baseline subgroup indicator S as $L_i \sim N(100, 10^2)$ and $S_i \sim \text{Bernoulli}(p = 0.5)$ for $i = 1, \dots, N$. It is clear that neither of these two variables is time dependent. In the first stage, we randomly assign patients into one of the induction therapy groups $Z^1 \in \{0, 1\}$. The transition times of the competing risks R and C are generated by $T_i^{(0,R)} \sim \text{LN}(\beta^{(0,R)} x_i^{(0,R)}, \sigma^2)$ and $T_i^{(0,C)} \sim \text{LN}(\beta^{(0,C)} x_i^{(0,C)}, \sigma^2)$ where $\beta^{(0,R)} = (2, 0.02, 0)$, $\beta^{(0,C)} = (1.5, 0.03, -0.8)$, $\sigma = 0.3$ and $x_i^{(0,R)} = x_i^{(0,C)} = (1, L_i, Z_i^1)$. Similarly, for the next three possible transitions for which $k \in \{(R, D), (C, P), (P, D)\}$, we generate the transition time $T_i^k \sim \text{LN}(\beta^k x_i^k, \sigma^2)$ with coefficients to be $\beta^{(R,D)} = (-0.5, 0.03, 0.2, 0.5, 0.3, 0, 0)$, $\beta^{(C,P)} = (1, 0.05, 1, -0.6)$ and $\beta^{(P,D)} = (0.8, 0.04, 1.5, -1, -1, -0.5, 0)$. The corresponding covariate vectors are $x_i^{(R,D)} = (1, L_i, Z_i^1, \log T_i^{(0,R)}, Z_i^{2,1}, S_i, S_i \cdot Z^{2,1})$, $x_i^{(P,D)} = (1, L_i, Z_i^1, \log T_i^{(0,C)}, \log T_i^{(C,P)}, Z_i^{2,2}, S_i \cdot Z_i^{2,2})$ and $x_i^{(C,P)} = (1, L_i, Z_i^1, \log T_i^{(0,C)})$. One can tell that in this case, the new factor S_i is not influential on the treatment selection at all.

Simulation 1b: T distribution with no interaction term. The only difference between Simulation 1b and Simulation 1a is that all the error terms of the log survival time in each stage are changed from being Gaussian distributed to being t distributed with degrees of freedom 10. For example, under this setting, $\log T_i^{(0,R)} = \beta^{(0,R)} x_i^{(0,R)} + \varepsilon_i$ where $\varepsilon_i \sim t(df = 10)$. All the underlying coefficients remain the same.

Simulation 2a: Gaussian distribution with interaction terms. Compared with Simulation 1a, the only change made in this case is to include the nonzero underlying interaction coefficients. Specifically, we have the underlying coefficients as $\beta^{(0,R)} = (2, 0.02, 0)$, $\beta^{(0,C)} = (1.5, 0.03, -0.8)$, $\beta^{(R,D)} = (-0.5, 0.03, 0.2, 0.5, 0.3, 0, -0.5)$, $\beta^{(C,P)} = (1, 0.05, 1, -0.6)$ and $\beta^{(P,D)} = (0.8, 0.04, 1.5, -1, -1, -0.5, 1)$. Such a setting introduces a heterogeneous treatment effect caused by the different values of S_i in the second stage for both resistance and progression groups. For example, according to the new $\beta^{(P,D)}$, one can tell that the HDAC therapy will only help increase the survival time of those patients undergoing progression who have $S_i = 1$ at baseline.

Simulation 2b: T distribution with interaction terms. The difference between Simulation 2b and Simulation 2a is similar to that between the first two simulations, that is, all the error terms for the log survival times are now changed from being Gaussian distributed to being t distributed with degrees of freedom 10.

The predicted optimal value function, that is, $\hat{\mathcal{V}}_1$, for all the selected models is presented in Table 1 for both the samples and populations. Higher values indicate better outcomes from the treatment regimes being estimated. For simulation 1a, the five selected methods perform similarly in terms of the expected value function while the modified DDP-GP model has a larger variance compared to the predicted optimal Q functions. When the Gaussian distribution assumption no longer holds in Simulation 1b, both of the DDP-GP models and Q-learning with exponential survival regression come up with lower expected value function. This decrease in performance could originate from the improper assumptions on the transition time distribution. When the true model contains the treatment–covariate interaction terms—and thus the optimal treatment varies from patient to patient—neither the original DDP-GP nor the modified DDP-GP models perform as well as the remaining three models. The Q-learning with exponential survival model achieves the highest average value function in this case. This performance may indicate that minor parametric model misspecification may not be a severe problem for Q-learning. In the last setting, where the Gaussian assumptions no longer hold but treatment–covariate interaction terms are present, the O-learning performs best. Generally speaking, Q-learning and O-learning appear to perform better under model misspecification, while O-learning appears to be the most robust to model misspecification but perhaps more variable than Q-learning.

5. Application to the Leukemia Trial Regimes

Due to the censoring issue as mentioned early, we only apply Q-learning illustrated in Section 2 to the Leukemia clinical trial regimes dataset in Xu et al. (2016). Although BOWL can be extended to censored data, this is beyond the scope of the current article. We choose the exponential survival regression as the base learner. In contrast to Xu et al. (2016), we let both $\mathcal{H}^1$ and $\mathcal{H}^0$ further contain the interaction term between the baseline age and therapy. As a consequence, we find that both interactions of $(Z^{2,1}, \text{age})$ and $(Z^{2,2}, \text{age})$ are statistically significant under an $\alpha = 0.1$ significance level when implementing the second stage Q-learning. According to the estimated coefficients, we find that for patients suffering resistance, the

Table 1. Simulation Studies: The estimated value function for sample and population (Pop.) including means and the corresponding standard deviations (in parentheses) over the 50 replicates. The true model column represents the situation where we plug in the true coefficients and true optimal treatment to calculate the value function; DDP-GP-1 and DDP-GP-2 stand for the situations where the Bayesian DDP-GP model excludes and includes the interaction terms; Q-learn-1 and Q-learn-2 denote the cases when we use Q-learning with linear regression and exponential survival regression as the base learner.

Cs	Stat.	True Model	DDP-GP-1	DDP-GP-2	Q-learn-1	Q-learn-2	O-learn
1a	Sample	7.95 (0.39)	6.78 (0.03)	6.84 (0.93)	7.16 (0.05)	**7.19 (0.04)**	6.63 (0.15)
	Pop.	7.65 (0)	6.77 (0.01)	6.82 (0.93)	7.15 (0.03)	**7.17 (0.02)**	6.52 (0.15)
1b	Sample	7.51 (0.34)	6.47 (0.06)	6.36 (1.92)	**6.98 (0.12)**	6.48 (0.18)	6.89 (0.17)
	Pop.	7.22 (0)	6.47 (0.07)	6.67 (1.86)	**6.99 (0.11)**	6.48 (0.18)	6.87 (0.13)
2a	Sample	7.89 (0.15)	6.55 (0.09)	6.68 (1.38)	7.20 (0.12)	**7.57 (0.07)**	7.29 (0.15)
	Pop.	8.02 (0)	6.43 (0.08)	6.56 (2.06)	7.19 (0.12)	**7.55 (0.06)**	7.28 (0.12)
2b	Sample	7.35 (0.25)	5.78 (0.10)	5.92 (2.61)	6.32 (0.17)	6.02 (0.16)	**6.87 (0.20)**
	Pop.	7.64 (0)	5.58 (0.11)	5.82 (2.30)	6.31 (0.16)	6.01 (0.17)	**6.73 (0.21)**

Table 2. Application of Q-learning with exponential survival regression to the Leukemia Trial Regimes: selected coefficient estimates in Stage 2. Z^2 represents $Z^{2,1}$ for the resistance group and $Z^{2,2}$ for the progression group.

Group	Resistance		Progression	
Terms	Estimate	Std	Estimate	Std
Z^2	2.84	1.63	0.65	1.03
$Z^2 \cdot$ age	−0.05	0.03	−0.03	0.02
age	−0.01	0.02	−0.002	0.01

Table 3. Application of Q-learning with exponential survival regression to the Leukemia Trial Regimes: selected coefficient estimates in Stage 1. For the treatment Z^1, the level 3,4,5,6 indicate FAI, FAI+ATRA, FAI+GCSF, FAI+ATRA+GCSF respectively and we choose level FAI as the reference.

Terms	Z^1			$Z^1 \cdot$ age			Age
Treatment Level	4	5	6	4	5	6	–
Estimate	− 0.11	− 1.73	− 1.08	0.01	0.04	0.02	− 0.04
Std	0.99	0.98	1.03	0.02	0.02	0.02	0.01

HDAC group always has a longer survival time than the non-HDAC group, which is consistent with the discoveries of Xu et al. (2016). For the patients suffering progression in the second stage, however, Q-learning finds that the HDAC would only be effective for the young age group (those patients younger than 22 years old approximately). In the first-stage implementation, Q-learning draws a similar conclusion as in the second stage in that the interaction between the therapy Z^1 and age is statistically significant when controlling for the cytogenetic abnormality level. The estimated coefficients show that FAI+ATRA would be the best therapy in the younger age group (<54) while FAI+GCSF would be the optimal therapy for the older age group. This conclusion is slightly different from the one drawn by only considering treatments main effect (Figure 8 in Xu et al. (2016)) but seems to be implied by Figure 6 in Xu et al. (2016). The Q-learning value function estimate indicates that it is possible to increase the average survival time by 81 days by assigning the estimated optimal treatment. We display the coefficient estimates for the treatment factor, age and their interactions in Table 3 (Stage 1) and Table 2 (Stage 2).

6. Discussion

In summary, the Bayesian DDP-GP model can perform very well when the distribution assumptions hold and the model specification is correct according to the numeric examples. In practice, one might also need to pay attention to the cases where some exceptions to the model assumptions happen, and in these settings the proposed Q-learning and O-learning methods are good alternatives. In particular, O-learning focuses on finding a decision treatment rule to maximize an objective function which reflects the benefit of using such a rule. According to its algorithm, O-learning does not calculate the overall treatment effect directly as is done in the Bayesian DDP-GP model. Q-learning concentrates on maximizing the cumulative reward by specifying the relationship between the Q-function and treatment at each stage. On the one hand, both O- and Q- learning methods can have more flexible model specifications and do not depend on assumptions regarding the response distribution. On the other hand, since the Bayesian DDP-GP model aims to make inference based on the posterior distribution of the estimate, it can additionally conduct tests of the null hypotheses of treatment effects and thus control type-I error as long as the distribution assumptions hold. This makes power analysis and sample size calculation more straightforward. In contrast, sample size computations for Q-learning and O-learning are more complicated and the increased model flexibility may necessitate larger sample sizes to achieve the same power. Subgroup analysis, which aims to identify subgroups of patients with enhanced treatment effects, may be viewed as an intermediate method for assessing treatment effects and facilitating power analysis and sample size calculations (Yusuf et al., 1991; Brookes et al., 2004; Rothwell, 2005; Shen and He, 2015; Fan, Song, and Lu 2016).

We would also like to point out some recent literature for dynamic treatment regimes for survival outcomes. Goldberg and Kosorok (2012) developed Q-learning for right-censored data when the censoring is completely independent of both the failure time and patient covariates. Jiang et al. (2015) developed optimal dynamic treatment regimes for maximizing t-year survival probability. Bai et al. (2015) considered optimal dynamic treatment regimes for survival endpoints using locally-efficient, doubly-robust estimators from a classification perspective. While extremely promising, some barriers to general use of these methods in practice remain, warranting the need for ongoing research.

References

Bai, X., Tsiatis, A., Lu, W., and Song, R. (2015), "Optimal Treatment Regimes for Survival Endpoints Using Locally-Efficient Doubly-Robust Estimator from a Classification Perspective" *Statistics in Medicine*. Under revision, doi:10.1007/s10985-016-9376-x. [946]

Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A., and Peters, T. J. (2004), "Subgroup Analyses in Randomized Trials: Risks of Subgroup-Specific Analyses; Power and Sample Size for the Interaction Test," *Journal of clinical epidemiology*, 57, 229–236. [946]

Fan, A., Song, R., and Lu, W. (2016), "Change-Plane Analysis for Subgroup Detection and Sample Size Calculation," *Journal of the American Statistical Association*. To appear, doi:10.1080/01621459.2016.1166115. [946]

Goldberg, Y., and Kosorok, M. R. (2012), "Q-Learning with Censored Data," *Annals of Statistics*, 40, 529–560. [946]

Hastie, T., Tibshirani, R., and Friedman, J. (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [944]

Jiang, R., Lu, W., Song, R., and Davidian, M. (2015), "On Estimation of Optimal Treatment Regimes for Maximizing t-Year Survival Probability," *Journal of the Royal Statistical Society*, Series B. To appear. [946]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [943]

Rothwell, P. M. (2005), "Subgroup Analysis in Randomised Controlled Trials: Importance, Indications, and Interpretation," *The Lancet*, 365, 176–186. [946]

Shen, J., and He, X. (2015), "Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model," *Journal of the American Statistical Association*, 110, 303–312. [946]

Xu, Y., Müller, P., Wahed, A. S., and Thall, P. F. (2016), "Bayesian Nonparametric Estimation for Dynamic Treatment Regimes with Sequential Transition Times," *Journal of the American Statistical Association*, 111, this issue. [942,943,944,945]

Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H. A. (1991), "Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials," *Journal of the American Medical Association*, 266, 93–98. [946]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013), "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100, 681–694. [942]

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015a), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [944]

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015b), "Doubly Robust Learning for Estimating Individualized Treatment With Censored Data," *Biometrika*, 102, 151–168. [944]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1200915

Comment

Lorenzo Trippa and Giovanni Parmigiani

Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Harvard Cancer Center, Boston, MA, USA

We congratulate the authors for the development of an excellent Bayesian nonparametric methodology to evaluate dynamic treatment regimes (DTR) . This challenging task is motivated in their article by the need of comparing, in cancer research, competing treatment strategies that involve selection of front-line therapies followed by salvage treatments at tumor recurrence. Beyond this important application, this new methodology can be directly used in other medical areas, as well as different disciplines such as economics and pedagogy.

A DTR is a plan, or more formally a function, which selects a treatment, based on an individual's time-varying profile including patients' characteristics, disease history, and previous interventions. Different treatments can be selected by the DTR during the course of an individual's disease history. The authors provide an elegant framework to identify optimal or nearly optimal DTRs at the completion of a partially randomized study, where randomization only involves the assignment of the front-line treatment but not the subsequent second-line treatments. Partial randomization and the need for integrating information on nonrandomized second-line interventions, which are not controlled by the investigators, are common in trials of advanced cancer. In this context, flexible modeling of disease histories is particularly valuable and, in the absence of reliable short-term surrogate outcomes, can support superiority/noninferiority evaluation of new experimental treatments.

The authors' framework can be summarized into two main components, (i) a flexible probability model for the entire disease history, that accounts for baseline individual characteristics and time-varying interventions, and (ii) a posteriori evaluation of candidate DTRs based on integration of a key metric, for example, life expectancy, over the posterior distribution. These two components are then followed, as should be the case in such a complex analysis, by a critical review of the approach considering possible weaknesses of the prior and modeling assumptions. Nonparametric models, in this setting, cannot avoid assumptions that require careful considerations, such as the absence of unmeasured confounders that drive the assignment of second-line treatments. These are potential sources of bias and erroneous conclusions in the comparison of DTRs with which any modeling approach needs to grapple. It is important to give them proper weight in the communication of the results especially in view of planning future studies.

We found the proposed Bayesian nonparametric approach attractive for several reasons. We mention three next.

First, a pragmatic consideration: selection of appropriate parametric assumptions in DTR can be extremely time consuming and involve a large number of goodness of fit summaries and modeling decisions, even with a modest number of treatments and relatively simple criteria to classify patients during their disease history. A flexible nonparametric model can allow the user, to a large extent, to circumvent this tedious process. In our view, this is an important advantage of nonparametric approaches in this context.

Second, as in several other estimation problems, Bayesian modeling has the advantage of allowing straightforward and solid integration of the uncertainty propagating from each component of the disease history model, ranging from the estimation of the response probabilities under competing front-line treatments to the estimation of the survival curves after progression under various salvage treatments. A natural alternative to Bayesian approaches could be resampling methods tailored to DTR; see, for example, Chakraborty, Laber, and Zhao (2013).

Third, the authors' framework allows one to easily move beyond the key metric of expected survival under competing DTRs. Author Peter Thall has been influential in the development of joint models of multiple outcomes in medicine, through an important series of articles that explain the advantages of joint modeling of toxicities and efficacy outcomes, for example survival, to optimize the treatment dose in oncology (Thall and Cook 2004). Building on the methodology introduced by Xu and colleagues, we envision more complex Bayesian models that involve a broader spectrum of conditions and clinically relevant events in the individual disease history. Potentially,

CONTACT Lorenzo Trippa ltrippa@jimmy.harvard.edu Dana-Farber Cancer Institute, Harvard Cancer Center, Boston, MA 02215, USA.

posterior samples of key parameters can be stored and interrogated at a later time to estimate *personalized* optimal DTRs that account for both individual conditions and the individual preferences. DTRs ranking can, for example, depend strongly on the risk of severe toxicities for some patients, and on the probability of tumor response for other patients.

Post-processing of the posterior distributions generated by Xu and colleagues could further enrich the set of results deliverable to the medical community. For example, it can be useful to identify the key components that drive differences in expected survival (or other metrics) across DTRs. This requires translating a complex posterior distribution into clearly interpretable summaries such as "Treatment A is recommended as first-line treatment. Although the response rates and times to tumor recurrence are comparable across treatments A, B, and C, there is evidence that patients who respond to treatment A have longer survival after progression." In this example, the disease history component that drives the recommendation is clearly identified: survival after tumor progression for patients that responded to treatment A. This type of summaries could facilitate the interpretation of DTRs comparisons, and have the potential of pointing attention in the right direction for follow-up studies that consider the same or related malignancies.

We next provide a similar example based on the dataset analyzed by Xu and colleagues.

Response rates across first-line treatments are comparable and, when we focus on patients that respond before two months from enrollment, we observe similar risks of recurrenceduring the first year from enrollment. For the set of 50 patients who respond within two months but recurred within 1 year, expected survival after recurrence appears to be markedly inferior for the FAI first-line treatment group compared to the others. We identified this subset of patients with rudimentary exploratory analyses. Post-processing of the posterior distribution can provide a more principled way to identify and emphasize disease history components that drive variation in overall survival, or other key metrics, across DTRs. This example illustrates such variation, and suggests that an even more productive use of the approach of Xu and colleagues could be achieved by using post-processing to produce highly representative of the nonparametric posterior distribution. Adapting existing post-processing principles (Hahn and Carvalho 2015) and subsets selection techniques (as those developed by Peter Müller and coauthors; Müller, Sivaganesan, and Laud 2010) to DTR, as well as developing new post-processing methodologies to summarize complex DTRs posterior distributions, could further strengthen the applicability and impact of the authors' approach.

In conclusion, we greet with enthusiasm the highly innovative contribution of Xu and colleagues.

References

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013), "Inference for Optimal Dynamic Treatment Regimes Using an Adaptive *m*-Out-of-*n* Bootstrap Scheme," *Biometrics*, 69, 714–723. [947]

Hahn, P. R., and Carvalho, C. M. (2015), "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective," *Journal of the American Statistical Association*, 110, 435–448. [948]

Müller, P., Sivaganesan, S., and Laud, P. (2010), "A Bayes Rule for Subgroup Reporting," in *Frontiers of Statistical Decision Making and Bayesian Analysis*, eds. M.-H. Chen, P. Müller, D. Sun, K. Ye, and D. Dey, New York: Springer, pp. 277–284. [948]

Thall, P. F., and Cook, J. D. (2004), "Dose-Finding Based on Efficacy–Toxicity Trade-Offs," *Biometrics*, 60, 684–693. [947]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1200917

Rejoinder

Yanxun Xu[a], Peter Müller[b], Abdus S. Wahed[c], and Peter Thall[d]

[a]Division of Statistics and Scientific Computing, The University of Texas at Austin, Austin, TX, USA; [b]Department of Mathematics, The University of Texas at Austin, Austin, TX, USA; [c]Epidemiology Data Center, University of Pittsburgh, Pittsburgh, PA, USA; [d]M. D. Anderson Cancer Center, Houston, TX, USA

1. We thank Chen, Liu, and Zeng et al. (CLZ) for their discussion of our Bayesian nonparametric (BNP) model-based methodology. They propose Q-learning methods as alternative approaches, and provide simulations to evaluate several methods, dubbed Q-learn-1, Q-learn-2, and O-learning, and two methods, based on versions of our DDP-GP model, that they call DDP-GP-1 and DDP-GP-2. The simulation results in their Table 1 show superior performance for Q-learn-1, Q-learn-2, and O-learn, in terms of closeness of the estimate to the simulation truth. However, the DDP-GP-1 and DDP-GP-2 are not the proposed BNP model. Rather, like the other methods, these two methods aim to optimize a value function, which is based on expected survival time following either resistance to induction chemo or progression. Each value function is a function of Q-functions and estimated optimal Stage 2 decisions, although the specific forms of their Q-functions are unclear.

CONTACT Peter Müller pmueller@math.utexas.edu Department of Mathematics, University of Texas at Austin, 1 University Station C1200, Austin, TX 78712.

Simulation Scenario 2a of CLZ

Regime	True Mean OS (days)	DDP-GP Estimate of Mean OS
(1,1,1)	199.14	201.64
(1,0,1)	195.39	201.63
(1,0,0)	195.45	201.76
(1,1,1)	199.20	201.77
(0,0,0)	158.22	147.24
(0,1,1)	165.69	166,45
(0,1,0)	165.69	166.46
(0,0,1)	158.21	147.23

The BNP approach has a different goal. To be specific, denote T = OS time, x = patient covariates, θ = model parameter, $\mathbf{Z}$ = DTR, $\mu_T(x, \mathbf{Z}, \theta) = \mathrm{E}(T|x, \mathbf{Z}, \theta)$, and $\mu_T(\mathbf{Z}, \theta) = \int_x \mu_T(x, \mathbf{Z}, \theta) dp(x)$. The goal is to estimate the posterior mean of $\mu_T(\mathbf{Z}, \theta)$ for each $\mathbf{Z}$, by computing $\hat{\mu}_T(\mathbf{Z}) = \mathrm{E}\{\mu_T(\mathbf{Z}, \theta)|\text{data}\}$ under the DDP-GP model. These posterior estimates can be used to choose a nominally optimal $\mathbf{Z}$, although the substantial variability in the estimates for the leukemia dataset renders any claim of optimality somewhat questionable. As illustrated in our Figure 7, one also may compute the covariate-specific posterior estimate $\hat{\mu}_T(x, \mathbf{Z}) = \mathrm{E}\{\mu_T(x, \mathbf{Z}, \theta)|\text{data}\}$ for given x, to evaluate individualized therapies.

The numerical values in the column "True Model" of Table 1 of CLZ are not mean OS times. To clarify this point, since simulation 1a of CLZ is the same as our simulation 4, below we tabulate $\hat{\mu}_T(x, \mathbf{Z})$ obtained from the actual DDP-GP method, based on 50 replications per case, under simulation scenario 2a of CLZ.

CLZ conclude "In summary, the Bayesian DDP-GP model can perform very well when the distribution assumptions hold and the model specification is correct according to the numeric examples." We disagree with the latter qualification in this conclusion. The DDP-GP model has full support and, essentially, is a mixture model that can fit virtually any distribution with high accuracy. Consequently, in any case, the posterior mean OS will be close to the truth, subject to the usual limitations of overall sample size and number of subjects per regime. Our additional simulation results under Scenario 2a studied by CLZ appear to confirm this.

Two additional points are worth mentioning. In our motivating example, finding the optimal sequential decision by direct comparison of all possible policies is not prohibitively difficult, since one can list all possible DTRs, write down a likelihood, and compute posteriors, with the optimal policy the $\mathbf{Z}$ maximizing $\hat{\mu}_T(\mathbf{Z})$. For more complicated problems, finding the optimal sequential policy can be much more difficult. We agree that, in such settings, Q-learning, O-learning, and similar methods may be preferable, since writing down and fitting a full likelihood may not be practical.

2. We thank Trippa and Parmigiani (TP) for their kind words, and their useful discussion of many key issues. We agree that potential effects of unmeasured confounders are very important, and that any inferences about DTRs should be qualified by noting the possibility of such effects. We agree with TP that an important extension of our methodology will be to evaluate and optimize DTRs based on multiple outcomes at each stage of the regime. This has been described by Lee et al. (2015, 2016), in a phase I–II dose-finding setting with a utility function $U(Y_E, Y_T)$ quantifying the trade-off between discrete efficacy Y_E and toxicity Y_T, in each of two cycles of therapy. However, the development of Lee et al. assumes a Bayesian hierarchical model, rather than a BNP model. To use trade-off utilities with continuous variables, aside from censoring, at calendar time t let $T(t)$ denote a subject's survival time and $B(t)$ the subject's total toxicity burden, as defined in Hobbs, Thall, and Lin (2016). A utility $U(t) = U\{T(t), B(t)\}$ can be established using the elicitation process given in Thall et al. (2013), to explicitly quantify the trade-off between $T(t)$ and $B(t)$. One may use $U(t)$ as the objective function for evaluating multi-stage regimes, assuming an appropriate Bayesian nonparametric model for the joint distribution of $[T(t), B(t)|\mathbf{Z}, x]$. This could be the basis for optimizing either overall or personalized regimes. Similar utility based analyses also have been carried out in a semi-competing risks setting by Murray et al. (2016). It is also worth noting that, if one does not wish to assume a BNP model in multi-stage treatment settings, trade-off utilities still may be used as the basis for defining a Q-function approach.

   We also agree with TP that principled post-processing of the posterior, beyond simply estimating mean OS, is a very important undertaking that potentially can identify important disease history components relevant to therapeutic decision making. This is an important area for future research in DTR settings.

3. We thank Guan, Laber, and Reich (GLR) for their detailed discussion of our methodology, their argument that Bayesian nonparametric methods can serve as "an engine for policy-search algorithms," and their presentation of interesting alternative methods.

   GLR propose and study a set of semi-parametric AFT models for transition times with covariate effects modeled using b-splines and additive residuals assumed to follow a mixture of normals. This greatly reduces computing time for this type of model, compared to the DDP-GP. To place this computational advantage into perspective, however, it should be kept in mind that, in practice, given the four models specified by GLR or some similar set of candidate models, one would need to perform goodness-of-fit analysis as a basis for either choosing one best model, or possibly do model averaging. The time needed to carry out this additional model criticism would be an additional consideration. Still, it is very interesting to see cases where the DDP-GP does not perform as well as some of the simpler proposals of GLR. It seems clear that this is a rich area for future research.

References

Hobbs, B., Thall, P. F., and Lin, S. (2016), "Bayesian Group Sequential Clinical Trial Design Using Total Toxicity Burden and Progression-Free Survival," *Journal of Royal Statistical Society*, Series C, 65, 273–297. [949]

Lee, J., Thall, P. F., Ji, Y., and Muller, P. (2015), "Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity," *Journal of the American Statistical Association*, 110, 711–722. [949]

——— (2016), "A Practical Decision-Theoretic Phase I–II Design for Ordinal Outcomes in Two Cycles," *Biostatistics*, 17, 304–319. [949]

Murray, T. A., Thall, P. F., Yuan, Y., McAvoy, S., and Gomez, D. R. (2016), "Robust Treatment Comparison Based on Utilities of Semi-Competing Risks in Non-Small-Cell Lung Cancer," *Journal of the American Statistical Association* (in press). [949]

Thall, P. F., Nguyen, H. Q., Braun, T. M., and Qazilbash, M. (2013), "Using Joint Utilities of the Times to Response and Toxicity to Adaptively Optimize Schedule-Dose Regimes," *Biometrics*, 69, 673–682. [949]