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.

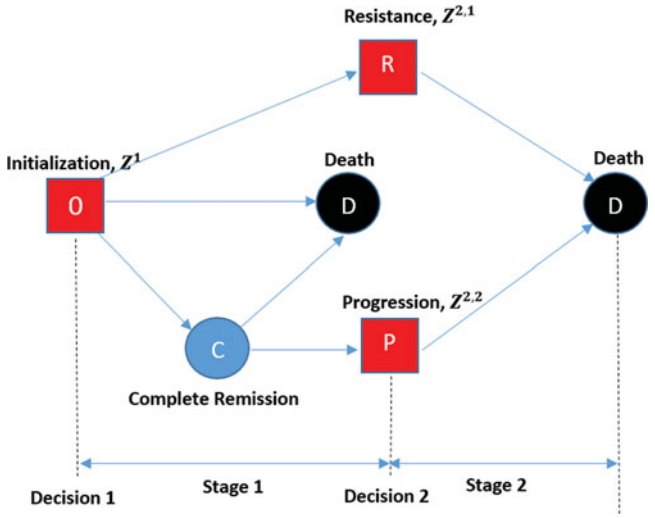

Figure 1. Redefinition of the Scheme under the proposed Q-learning Framework. The states in red square boxes (i.e., initialization, resistance and progression) are the treatment decision-making points that are used to split the two stages. Complete remission (C) is not considered as a splitting point 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:

$$\hat{\pi}_{2,1} = \arg\max_{Z^{2,1}} \left\{ \hat{Q}_{2R}\left(\mathcal{H}^{1R}, Z^{2,1}\right) \right\}, \text{ for the resistance group,}$$
$$\hat{\pi}_{2,2} = \arg\max_{Z^{2,2}} \left\{ \hat{Q}_{2P}\left(\mathcal{H}^{1P}, Z^{2,2}\right) \right\}, \text{ for the progression group,} \quad (1)$$

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

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

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

$$f_{2R} = \arg\min_f \mathbb{E}_n \left[\frac{T^{(R,D)} \phi(Z^{2,1} f(\mathcal{H}^{1R}))}{\pi_{2,1}(Z^{2,1}, \mathcal{H}^{1R})} + \lambda_2 \|f\|^2 \right], \quad \text{for the resistance group,}$$

$$f_{2P} = \arg\min_f \mathbb{E}_n \left[\frac{T^{(P,D)} \phi(Z^{2,2} f(\mathcal{H}^{1P}))}{\pi_{2,2}(Z^{2,2}, \mathcal{H}^{1P})} + \lambda_2 \|f\|^2 \right], \quad \text{for the progression group,} \tag{3}$$

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

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

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 $i = 1, \ldots, N$. 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 ($Z^{2,1}$, age) and ($Z^{2,2}$, 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]