Comment

Jingxiang Chen, Yufeng Liu, Donglin Zeng, Rui Song, Yingqi Zhao, and Michael R. Kosorok

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

KEYWORDS

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

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.

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 non-parametric 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.
(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 discussion 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 assumption 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:

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

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

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 $\hat{V}_{2} = I_{RD} \cdot \hat{Q}_{2R}(\mathcal{H}_{1R}, \hat{\pi}_{2,1}) + I_{PD} \hat{Q}_{2P}(\mathcal{H}_{1P}, \hat{\pi}_{2,2})$, 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 $\hat{V}_{2}$ 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

$$
\tilde{T} = I_{D} T^{(0,D)} + I_{RD} \left( T^{(0,R)} + \hat{V}_{2} \right) + I_{PD} \left( T^{(0,P)} + \hat{V}_{2} \right)
$$

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 $Z^{2,1}$ 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

$$
\hat{\pi}_{1} = \arg \max_{d_{1}} \hat{Q}_{1}(\mathcal{H}^{0}, d_{1}).
$$

We aim to find $d_{1}$ to maximize (2) and the maximal objective value is denoted as $\hat{V}_{1}$. The base learners with the highest $\hat{V}_{1}$ value would be desirable. For demonstration, we use linear regression

![Redefinition of the Scheme under the proposed Q-learning Framework.](image)
and exponential survival regression as the two base learners of $Q_{2R}$, $Q_{2P}$ and $Q_i$ 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 choose 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

$$ f_{2R} = \arg \min_{f} \mathbb{E}_n \left[ \frac{T^{(R,D)}(Z, f(H_{1R}))}{\pi_2(Z, H_{1R})} + \lambda_2 \| f \|^2 \right], $$

for the resistance group,

$$ f_{2P} = \arg \min_{f} \mathbb{E}_n \left[ \frac{T^{(P,D)}(Z, f(H_{1P}))}{\pi_2(Z, H_{1P})} + \lambda_2 \| f \|^2 \right], $$

for the progression group,   

where the surrogate loss function $\phi(t) = \max(1 - t, 0)$, $\mathbb{E}_n$ denotes the empirical mean over the sample, $\pi_{2}(z, H^{1}) = \Pr(Z = z | H^{1})$, $\pi_{2}(z, H^{1}) = \Pr(Z = z | H^{1})$, $\| \cdot \|^2$ denotes the square of L2 norm, and $\lambda_2$ is the tuning parameter controlling model complexity of (4). We obtain the estimate of the optimal treatment rule $d_1$ for Stage 1 via $d_1(H_{0}) = \mathbb{I}(f_1(H_{0}) > 0)$. 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 HDCAC. 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 $V_t$ 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 $S_i \sim Bernoulli(p = 0.5)$ 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 $Z^i \in \{0, 1\}$. The transition times of the competing risks R and C are generated by $T_R^{i,(R)} \sim LN(\beta^{(R)} x^{(R)}_i, \sigma^2)$ and $T_C^{i,(C)} \sim LN(\beta^{(C)} x^{(C)}_i, \sigma^2)$ where $\beta^{(R)} = (2, 0.02, 0)$, $\beta^{(C)} = (1.5, 0.03, -0.8)$, $\sigma = 0.3$ and $x^{(R)}_i = x^{(C)}_i = (1, L_i, Z_i)$. Similarly, for the next three possible transitions for which $k \in \{(R, D), (C, P), (P, D)\}$, we generate the transition time $T_k^i \sim LN(\beta^k x^k_i, \sigma^2)$ 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 $x^{(R,D)}_i = (1, L_i, Z_i, \log T^{(R)}_i, Z^1_i, S_i, Z_i, Z^{2,1})$, $x^{(C,P)}_i = (1, L_i, Z_i, \log T^{(C)}_i, \log T^{(C,P)}_i, Z^2_i, S_i, Z_i, Z^{2,2})$ and $x^{(P,D)}_i = (1, L_i, Z^1_i, \log T^{(C)}_i)$. 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 in the log survival times of the second stage are changed from being Gaussian distributed to being t distributed with degrees of freedom 10. For under this setting, $\log T^{(R)}_i = \beta^{(R)} x^{(R)}_i + \varepsilon_i$ 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 $\beta^{(R)} = (2, 0.02, 0)$, $\beta^{(C)} = (1.5, 0.03, -0.8)$, $\beta^{(R,D)} = (-0.5, 0.03, 0.2, 0.5, 0.3, 0, -0.5)$, $\beta^{(C,P)} = (1, 0.05, 1, -0.6)$ 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 $\beta^{(C,D)}$, 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, $V_t$, 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.

### 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 columns represent 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.

<table>
<thead>
<tr>
<th>Cs</th>
<th>Stat.</th>
<th>True Model</th>
<th>DDP-GP-1</th>
<th>DDP-GP-2</th>
<th>Q-learn-1</th>
<th>Q-learn-2</th>
<th>O-learn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>Sample</td>
<td>7.95 (0.39)</td>
<td>6.78 (0.03)</td>
<td>6.84 (0.93)</td>
<td>7.16 (0.05)</td>
<td>7.19 (0.04)</td>
<td>6.63 (0.15)</td>
</tr>
<tr>
<td></td>
<td>Pop.</td>
<td>7.65 (0.0)</td>
<td>6.77 (0.01)</td>
<td>6.82 (0.93)</td>
<td>7.15 (0.03)</td>
<td>7.17 (0.02)</td>
<td>6.52 (0.15)</td>
</tr>
<tr>
<td>1b</td>
<td>Sample</td>
<td>7.51 (0.34)</td>
<td>6.47 (0.06)</td>
<td>6.36 (1.92)</td>
<td>6.98 (0.12)</td>
<td>6.48 (0.18)</td>
<td>6.89 (0.17)</td>
</tr>
<tr>
<td></td>
<td>Pop.</td>
<td>7.22 (0.0)</td>
<td>6.47 (0.07)</td>
<td>6.76 (1.86)</td>
<td>6.99 (0.11)</td>
<td>7.19 (0.17)</td>
<td>6.87 (0.13)</td>
</tr>
<tr>
<td>2a</td>
<td>Sample</td>
<td>7.89 (0.15)</td>
<td>6.55 (0.09)</td>
<td>6.68 (1.38)</td>
<td>7.20 (0.12)</td>
<td>7.57 (0.07)</td>
<td>7.29 (0.15)</td>
</tr>
<tr>
<td></td>
<td>Pop.</td>
<td>8.02 (0.0)</td>
<td>6.43 (0.08)</td>
<td>6.56 (2.06)</td>
<td>7.19 (0.12)</td>
<td>7.55 (0.06)</td>
<td>7.28 (0.12)</td>
</tr>
<tr>
<td>2b</td>
<td>Sample</td>
<td>7.35 (0.25)</td>
<td>5.78 (0.10)</td>
<td>5.92 (2.61)</td>
<td>6.32 (0.17)</td>
<td>6.02 (0.16)</td>
<td>6.87 (0.20)</td>
</tr>
<tr>
<td></td>
<td>Pop.</td>
<td>7.64 (0.04)</td>
<td>5.58 (0.11)</td>
<td>5.82 (2.30)</td>
<td>6.31 (0.16)</td>
<td>6.01 (0.17)</td>
<td>6.73 (0.21)</td>
</tr>
</tbody>
</table>

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 $H^1$ and $H^0$ 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
Leukemia Trial Regimes: selected coefficient estimates in Stage 1. For the treatment
Leukemia Trial Regimes: selected coefficient estimates in Stage 2.

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, sam-
ple 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 analy-
sis, 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 perspec-
tive. While extremely promising, some barriers to general use of
these methods in practice remain, warranting the need for ongo-
ing research.

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 treat-
ment 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 distri-
bution. 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

<table>
<thead>
<tr>
<th>Terms</th>
<th>$Z^1$</th>
<th>$Z^1 \cdot \text{age}$</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>4.01</td>
<td>1.73</td>
<td>-1.08</td>
</tr>
<tr>
<td>Std</td>
<td>0.99</td>
<td>0.98</td>
<td>1.03</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Terms</th>
<th>$Z^2$</th>
<th>$Z^2 \cdot \text{age}$</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.05</td>
<td>0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>Std</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.002</td>
</tr>
</tbody>
</table>

Table 3. Application of Q-learning with exponential survival regression to the
Leukemia Trial Regimes: selected coefficient estimates in Stage 1. For the treatment
$Z^2$, the level $3, 4, 5, 6$ indicate FAI, FAI+ATRA, FAI+GCSF, FAI+ATRA+GCSF respectively
and we choose level FAI as the reference.

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 sec-
ond 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 implement-
ation, Q-learning draws a similar conclusion as in the sec-
ond stage in that the interaction between the therapy $Z^1$ and
age is statistically significant when controlling for the cytoge-
netic 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 indi-
cates that it is possible to increase the average survival time by
81 days by assigning the estimated optimal treatment. We dis-
play the coefficient estimates for the treatment factor, age and
their interactions in Table 3 (Stage 1) and Table 2 (Stage 2).

References

tor from a Classification Perspective” Statistics in Medicine. Under revi-
sion, doi:10.1007/s10985-016-9376-x. [946]
Subgroup-Specific Analyses; Power and Sample Size for the Interaction Test,” Journal of clinical epidemiology, 57, 229–236. [946]
[946]
Annals of Statistics, 40, 529–560. [946]
[944]
mal Treatment Regimes for Maximizing t-Year Survival Probability,”
Royal Statistical Society, Series B, 65, 331–355. [943]
Rothwell, P. M. (2005), “Subgroup Analysis in Randomised Controlled Tri-
[946]
Shen, J., and He, X. (2015), “Inference for Subgroup Analysis With a Struct-
tured Logistic-Normal Mixture Model,” Journal of the American Statistical
Association, 110, 303–312. [946]
metric Estimation for Dynamic Treatment Regimes with Sequential Transition Times,” Journal of the American Statistical Association, 111, this
issue. [942, 943, 944, 945]
domized Clinical Trials,” Journal of the American Medical Association, 266, 93–98. [946]
mation of Optimal Dynamic Treatment Regimes for Sequential Treatment
Decisions,” Biometrika, 100, 681–694. [942]