

# *University of North Carolina at Chapel Hill*

The University of North Carolina at Chapel Hill Department of  
Biostatistics Technical Report Series

---

*Year 2009*

*Paper 13*

---

## Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer

Yufan Zhao\*

Michael R. Kosorok<sup>†</sup>

Donglin Zeng<sup>‡</sup>

Mark A. Socinski\*\*

\*University of North Carolina at Chapel Hill

<sup>†</sup>University of North Carolina at Chapel Hill

<sup>‡</sup>University of North Carolina at Chapel Hill

\*\*University of North Carolina at Chapel Hill

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/uncbiostat/art13>

Copyright ©2009 by the authors.

# Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer

Yufan Zhao, Michael R. Kosorok, Donglin Zeng, and Mark A. Socinski

## Abstract

Typical regimens for advanced metastatic stage IIIB/IV non-small cell lung cancer (NSCLC) consist of multiple lines of treatment. We present an adaptive reinforcement learning approach to discover optimal individualized treatment regimens from a specially designed clinical trial (a “clinical reinforcement trial”) of an experimental treatment for patients with advanced NSCLC who have not been treated previously with systemic therapy. In addition to the complexity of the problem of selecting optimal compounds for first and second-line treatments based on prognostic factors, another primary scientific goal is to determine the optimal time to initiate second-line therapy, either immediately or delayed after induction therapy, yielding the longest overall survival time. A reinforcement learning method called Q-learning is utilized which involves learning an optimal policy from patient data generated from the clinical reinforcement trial. Approximating the Q-function with time-indexed parameters can be achieved by using a modification of support vector regression which can utilize censored data. Within this framework, a simulation study shows that the procedure can extract optimal strategies for two lines of treatment directly from clinical data without relying on the identification of any accurate mathematical models. In addition, we demonstrate that the design reliably selects the best initial time for second-line therapy while taking into account the heterogeneity of NSCLC across patients.

# Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer

Yufan Zhao<sup>1</sup>, Michael R. Kosorok<sup>1,3</sup>, Donglin Zeng<sup>1</sup>,  
and Mark A. Socinski<sup>2</sup>

July 2009

## Abstract

Typical regimens for advanced metastatic stage IIIB/IV non-small cell lung cancer (NSCLC) consist of multiple lines of treatment. We present an adaptive reinforcement learning approach to discover optimal individualized treatment regimens from a specially designed clinical trial (a “clinical reinforcement trial”) of an experimental treatment for patients with advanced NSCLC who have not been treated previously with systemic therapy. In addition to the complexity of the problem of selecting optimal compounds for first and second-line treatments based on prognostic factors, another primary scientific goal is to determine the optimal time to initiate second-line therapy, either immediately or delayed after induction therapy, yielding the longest overall survival time. A reinforcement learning method called Q-learning is utilized which involves learning an optimal policy from patient data generated from the clinical reinforcement trial. Approximating the Q-function with time-indexed parameters can be achieved by using a modification of support vector regression which can utilize censored data. Within this framework, a simulation study shows that the procedure can extract optimal strategies for two lines of treatment directly from clinical data without relying on the identification of any accurate mathematical models. In addition, we demonstrate that the design reliably selects the best initial time for second-line therapy while taking into account the heterogeneity of NSCLC across patients.

**KEYWORDS:** Adaptive design; Individualized therapy; Multi-stage decision problems; Non-small cell lung cancer; Optimal policy; Q-learning; Reinforcement learning; Support vector regression.

---

<sup>1</sup>Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

<sup>2</sup>Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

<sup>3</sup>Email: kosorok@unc.edu

# 1 Introduction

There has been significant recent research activity in developing therapies that are tailored to each individual. Finding such therapies in treatment settings involving multiple decision times is a major challenge. For example, in treating advanced non-small cell lung cancer (NSCLC), patients typically experience two or more lines of treatment, and many studies demonstrate that three lines of treatment can improve survival for patients. Discovering tailored therapies for these patients is a very complex issue since effects of covariates (such as established prognostic factors or biomarkers) must be modelled within the multi-stage structure. In this article, we present a new kind of NSCLC clinical trial, based on reinforcement learning methods from computer science, that statistically finds an optimal individualized treatment plan at each decision time which is a function of available patient prognostic information. This new kind of trial extends and refines the “clinical reinforcement trial” concept developed in Zhao, et al. (2009) to enable application to NSCLC treatment and to utilize right-censored survival data.

For NSCLC patients who present with a good performance status and stage IIB/IV disease, platinum-based chemotherapy is the primary treatment which offers a modest survival advantage over best supportive care (BSC) alone. First-line treatment primarily consists of doublet combinations of platinum compounds (cisplatin or carboplatin) with gemcitabine, pemetrexed, paclitaxel, or vinorelbine (Scagliotti et al., 2008; Sandler et al., 2006; Pirker et al., 2008). These drugs modestly improve the therapeutic index of therapy, but no combination appears to be clearly superior. More recently, the addition of bevacizumab, a monoclonal antibody against vascular endothelial growth factor (VEGF), to carboplatin and paclitaxel has been shown to produce a higher response rate and longer progression-free survival and overall survival times (Sandler et al., 2006). However, this phase III study was only designed to investigate patients with histologic evidence of non-squamous cell lung cancer. Therefore, in first-line treatment of NSCLC, a very important clinical question is what tailored treatment to administer based on each individual’s prognostic factors (including the patient’s histology type, toxicity profile, smoking history, and VEGF level, etc.), among many approved first-line treatments.

All patients with advanced NSCLC who initially receive a platinum-based first-line chemotherapy inevitably experience disease progression. Approximately 50–60% of patients on recent phase III first-line trials received second-line treatment (Sandler et al., 2006). Similar to the first-line regimen, three FDA approved second-line agents (docetaxel, pemetrexed, and erlotinib) appear to have similar response and overall survival efficacy but very different toxicity profiles (Shepherd et al., 2000; Ciuleanu et al., 2008; Shepherd et al., 2005). The choice of agent should also mainly depend on a number of factors, including the patient’s comorbidities, toxicity from previous treatments, and the risk for neutropenia. A better understanding of prognostic factors in the second-line setting may allow clinicians to better select patients for second-line therapy, and lead to better designed second-line trials.

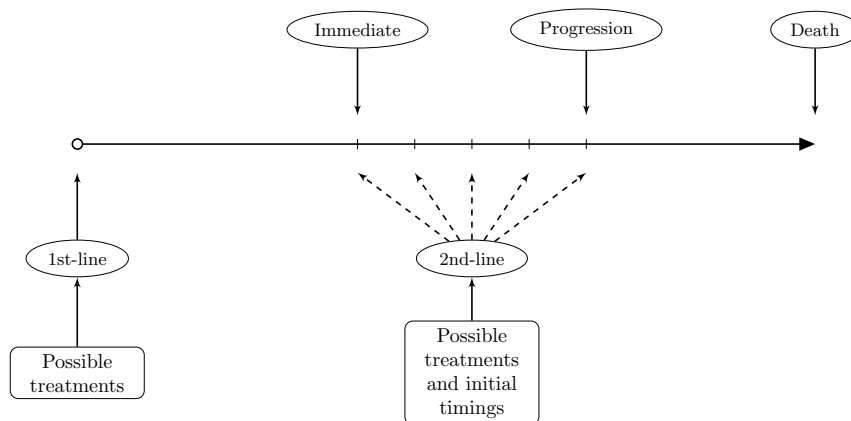


Figure 1: Treatment plan and therapy options for an advanced NSCLC trial.

The current standard treatment paradigm is to initiate second-line therapy at the time of disease progression. Recently there have been two phase III trials that have investigated other possible timings of initiating second-line therapy (Fidias et al., 2007; Ciuleanu et al., 2008). Both of these trials have revealed a statistically significant improvement in the progression-free survival, and a trend towards improved survival for the earlier use of second-line therapy. However, in terms of considering overall survival as the primary endpoint, the treatment effect revealed by these two trials is not significant. Stinchcombe and Socinski (2009) claimed that even under the best of circumstances not all patients will benefit from early initiation of second-line therapy. Hence the proper selection of patients is also critical to determining the proper time for initiation. Hence, in addition to the difficulty of discovering individualized superior therapies in second-line treatment, another primary challenge is to determine the optimal time to initiate second-line therapy, either to receive treatment immediately after completion of platinum-based therapy, or to delay to another time prior to disease progression, whichever results in the largest overall survival probability. A key goal is to provide patients with non-cross-resistant therapies capable of obtaining better response rates and longer survival time.

Some patients who maintain a good performance status and tolerate therapy without significant toxicities will receive third-line therapy (Stinchcombe and Socinski, 2008). Since there is only one FDA approved agent (Erlotinib) available for third-line treatment, we restrict our attention hereafter to finding optimal therapies for first-line and second-line only.

Figure 1 illustrates the treatment plan and clinically relevant patient outcomes. Therapy begins with first-line platinum-based doublets aimed at improv-

ing survival and palliating disease-related symptoms without undue toxicity. The patient is then delivered to no more than 8 cycles of treatment as recommended by the American Society of Clinical Oncology. Socinski and Stinchcombe (2007) suggest the standard initial duration of platinum-based therapy should be 3 to 4 cycles since four of the five trials investigating the duration of platinum therapy in the first-line setting have revealed equivalent survival with the shorter duration of therapy. Due to the effects of the initial treatment, patients generally experience disease progression within a median of 3–6 months, and the median survival time observed is 8 to 10 months (Schiller et al., 2002; Sandler et al., 2006). Approximately 30–40% of patients survive 1 year, and less than 15% survive 2 years (Bunn and Kelly, 1998). If the first line of treatment is successfully completed without progression or death, then a second line of therapy is administered sometime between the completion of first-line treatment and the time of first evidence of disease progression. Patients with a good performance status in second-line trials have a median survival duration of approximately 9 months (Stinchcombe and Socinski, 2008). Given the non-curative nature of chemotherapy in advanced NSCLC, we will use the overall survival time as the primary endpoint.

The primary scientific goal of the trial is to select optimal compounds for first and second-line treatments as well as the optimal time to initiate second-line therapy based on prognostic factors yielding the longest averaged survival time. Our design is based on a reinforcement learning method, called Q-learning, for maximizing the average survival time of patients as a function of prognostic factors, treatment decisions, and optimal timing. Zhao, et al. (2009) introduced the clinical reinforcement trial concept based on Q-learning for discovering effective therapeutic regimens in potentially irreversible diseases such as cancer. The concept is essentially an extension and melding of dynamic treatment regimes and sequential multiple assignment randomized trials (Murphy, 2005a) to accommodate both the presence of an irreversible disease state and a possible continuum of treatment options. The generic cancer application developed in Zhao, et al. takes into account a drug’s efficacy and toxicity simultaneously. The authors demonstrate that reinforcement learning methodology not only captures the optimal individualized therapies successfully, but is also able to improve longer-term outcomes by considering delayed effects of treatment. Their approach utilizes a simplistic reward function structure with integer values to assess the tradeoff between efficacy and toxicity. In the targeted NSCLC setting, however, this simplistic approach will not work due to the necessity of using overall survival time as the net reward to reflect the desired primary endpoint, and new methods are required.

Our proposed clinical reinforcement trial for NSCLC involves a fair randomization of patients among the different therapies in first and second-line treatments, as well as randomization of the time of initiating second-line therapy. Reinforcement learning is used to analyze the resulting data and estimate optimal individualized treatment regimens. In order to successfully handle the complex fact of heterogeneity in treatment across individuals as well as right-censored survival data, we modify the support vector regression (SVR) approach

(Vapnik, Golowich, and Smola, 1997) within a Q-learning framework to fit potentially nonlinear Q-functions for each of the two decision times (before first line and before second line). In addition, a second trial with a phase III structure is proposed to be conducted after this first trial to validate the improvement of the optimal individualized therapy against the standard of care and/or other valid competing therapies.

The remainder of this article is organized as follows. In Section 2, we provide a detailed description of the patient outcomes and Q-learning framework, followed by an introduction to SVR for estimating Q-functions and the development of a new form of SVR,  $\epsilon$ -SVR-C, for right-censored outcomes. The NSCLC trial conduct and related computational issues are presented in Section 3. In Section 4, we present a simulation study of the design to discover individualized optimal treatment strategies. We close with a discussion in Section 5.

## 2 Reinforcement Learning Framework

### 2.1 Patient Outcomes

Let  $t_1$  and  $t_2$  denote the decision times for the first and second treatment lines, respectively. After initiation of first-line chemotherapy, the time to disease progression is denoted by  $T_P$ .  $t_2$  is also the time at the completion first-line treatment, which is a fixed value usually less than  $T_P$  and determined by the number of cycles delivered in the first line of chemotherapy. We will assume for simplicity that  $T_P \geq t_2$  with probability 1. Denote the targeted time after  $t_2$  of initiating second-line therapy by  $T_M$ . Thus, according to the description of the treatment plan in Section 1, the actual time to initiate the second line is the minimum of  $t_2 + T_M$  and  $T_P$ , and the gap between the end of the first line and the beginning of the second line is  $T_M \wedge (T_P - t_2)$ , where  $\wedge$  denotes minimum. At the end of first-line therapy,  $t_2$ , clinicians make a decision about the target start time  $T_M$ . We let  $T_D$  denote the time of death from the start of therapy ( $t_1$ ), i.e., the overall survival time.

Because of the possibility of right censoring, we define the patient's censored time by  $C$  and the indicator of censoring by  $\delta = I(T_D \leq C)$ . Right censoring may be due to several reasons, including an adverse event so severe that therapy cannot be continued or the patient choosing not to receive further therapy. We will assume for now, however, that censoring is completely for administrative reasons and is thus independent of both the death time and the patient covariates. For convenience, we let  $T_1 = T_D \wedge t_2$ ,  $Y_D = I(T_D \wedge C \geq t_2)$  and  $\nu = Pr(Y_D = 1)$ , and denote  $T_2 = (T_D - t_2)I(T_D \geq t_2) = (T_D - t_2)I(T_1 = t_2)$  and  $C_2 = (C - t_2)I(C \geq t_2)$ . Note that  $T_D = T_1 + T_2$ . We can now define the total follow-up time  $T^0 = T_D \wedge C = T_1 \wedge C + Y_D(T_2 \wedge C_2)$ . The settings for determining  $T_1$ ,  $C$ ,  $T_2$  and  $T^0$  are summarized in Figure 2, including the possibilities of death or right censoring either before or after second-line therapy.

Denote patient covariate values at the  $i$ th decision time by  $\mathbf{O}_i = (O_{i1}, \dots, O_{iq})$  for  $i = 1, 2$ . Such covariates can include prognostic variables or biomarkers

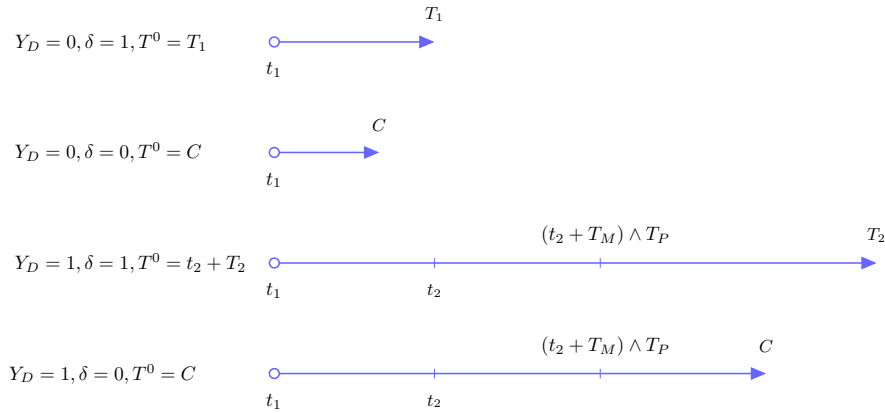


Figure 2: The four cases that determine  $T^0$ . In each case, the time of last follow-up is indicated by a right triangle. Note that  $T_P$ ,  $T_1$  and  $C$  originate at  $t_1$ , while  $T_M$  and  $T_2$  originate at  $t_2$ .

thought to be related to outcome. In first-line therapy, we assume that the death time  $T_1$  depends on the covariates  $\mathbf{O}_1$  and possible treatment  $D_1$  according to a distribution

$$[T_1 \mid \mathbf{O}_1, D_1] \sim f_1(\mathbf{O}_1, D_1; \boldsymbol{\alpha}_1),$$

where decision  $D_1$  only consists of a finite set of agents  $d_1$ . If the patient survives long enough to be treated by second-line therapy, we assume that the disease progression time  $T_P$  is  $\geq t_2$  and follows another distribution

$$[T_P \mid \mathbf{O}_1, D_1] \sim f_2(\mathbf{O}_1, D_1; \boldsymbol{\alpha}_2).$$

In addition, to account for the effects of initial timing of second-line therapy on survival,  $T_2$  given  $T_D \geq t_2$  is then given by

$$[T_2 \mid \mathbf{O}_2, D_1, D_2, T_M, T_P] \sim f_3(\mathbf{O}_2, D_1, D_2, T_M; \boldsymbol{\alpha}_3),$$

where  $D_2$  consists of a finite set of agents  $d_2$  and  $T_M$  is a continuum of initiation times for second-line therapy as described above. We assume also that  $P(T_D = t_2) = 0$ . Note the because of the independence of censoring, conditioning  $T_2$  on  $Y_D = 1$  is the same as conditioning on  $T_D \geq t_2$ . Note that this study is designed to identify the initiation time,  $T_M$ , which is associated with the best combination of treatments  $d_1$  and  $d_2$ , while maintaining longest survival  $T_D$ . Due to heterogeneities among patients, biomarker-treatment interactions, and the large number of possible shapes of  $T_2$  as functions of  $T_M$ , the distributions  $f_1, f_2$ , and  $f_3$  can be complicated and may vary between different groups of



patients. Thus, incorporating  $\mathbf{O}_i$  into models for  $f_i$  ( $i = 1, 2, 3$ ) is quite challenging, and such model-based approaches can easily become intractable (Thall et al., 2007). Another important issue is accounting for delayed effects of first-line therapy. Thall, et al., (2007) claimed that the conventional model-based approaches are not capable of handling this kind of situation very well. Based on clinical data, reinforcement learning is not only a model-free method which carries out treatment selection sequentially with time-dependent outcomes to determine optimal individualized therapy, but it can also improve longer-term outcomes by taking into account delayed effects of treatments.

## 2.2 Q-Learning Framework

Q-learning (Watkins, 1989; Watkins and Dayan, 1992) is one of the most widely used reinforcement learning methods. In multi-stage decision problem, if we denote each decision point by  $t$ , state  $S_t$ , action  $A_t$ , and reward  $R_t$  are three fundamental elements of Q-learning. Q-learning assigns values to action-state pairs, and it is learning, based on  $S_t$ , how best to choose  $A_t$  to maximize an expected discounted return of the form:

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^T r_{t+T} = \sum_{k=0}^T \gamma^k r_{t+k},$$

where  $\gamma$  is the discount rate ( $0 \leq \gamma \leq 1$ ).

The algorithm therefore has a so-called  $Q$  function which calculates the quality of a state-action combination as follows:

$$Q : S \times A \rightarrow \mathbb{R}.$$

The motivation of Q-learning is that once the  $Q$  functions have been estimated, we only need to know the state to determine an action, without the knowledge of a transition model that tells us what state we might go to next. Before learning has started,  $Q$  returns a fixed value which is chosen by the designer. Then, at each time point  $t$ , the learner is given a reward value which is calculated for each combination of a state  $s_t \in S_t$ , and action  $a_t \in A_t$ . The core of the algorithm is a simple value iteration update. It assumes the old value and makes a correction based on the new information as follows (Sutton and Barto, 1998):

$$Q_t(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \times \left[ r_t + \gamma \max_{a_{t+1}} Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \right], \quad (1)$$

where  $r_t$  is the current reward given at time  $t$ ,  $\alpha_t(s_t, a_t) \in (0, 1]$  is the learning rate (or learning step-size).  $\alpha_t(s_t, a_t)$  is a constant which determines to what extent the newly acquired information will override the old information, that is, how fast learning takes place. A factor of 0 will make the learner not learn anything, while a factor of 1 will make the learner consider fully the most recent

information. We can interpret  $\gamma$  as a control to balance a learners' immediate rewards and future rewards. As  $\gamma$  approaches 1, we take future rewards into account more strongly. In the context of this paper, we let  $\gamma = 1$ , which means we fully maximize rewards over the long run. For simplicity of computation, we ignore the step-size (let  $\alpha_t(s_t, a_t) = 1$ ) for the rest of the article. All results hold with minor modifications when the step-size effects are considered. Then model (1) can be simplified to a one-step simple recursive form

$$Q_t(s_t, a_t) \leftarrow r_t + \max_{a_{t+1}} Q_{t+1}(s_{t+1}, a_{t+1}). \quad (2)$$

The Q-learning algorithm attempts to find a policy  $\pi$  that maps states to actions the learner ought to take in those states.  $\pi$  is possibly deterministic, non-stationary, and non-Markovian. We denote the optimal policy by  $\pi_t^*$ , which satisfies

$$\pi_t^* = \arg \max_{a_t} Q_t(s_t, a_t).$$

Zhao et al. (2009) performed a simulation study of a simple Q-learning approach with 6 decision time points for discovering optimal dosing for treatment of a simplistic generic cancer. While the results were encouraging, much work is needed before these methods can be applied to specific, realistic cancer scenarios, such as the NSCLC setting of this paper. For example, in their study, the choice of treatments at each decision time point is taken simply among a continuum of dosing levels. However, in NSCLC treatment with two decision time points, the action variables in the second stage become two-dimensional ( $d_2$  and  $T_M$ ). The second issue is that overall survival time, the endpoint of interest in NSCLC, cannot be utilized in the usual reward function structure in standard Q-learning, and new methodology and modeling are needed. Moreover, the presence of censoring in the reward outcome means that a fundamentally new approach for estimating the Q-function is needed.

In our clinical setting we respectively denote state and action random variables by  $O_i$  and  $D_i$  for  $i = 1, 2$ . This is consistent with notations of prognostic factors or biomarkers and treatment options used in Section 2.1. As mentioned in Section 1, we consider survival time as the primary reward. Specifically, by performing a treatment  $d_1$ , where  $d_1 \in D_1$ , the patient can transit from first line to second line treatment. Such treatment associated with prognostic factors provides the patient a progression time  $T_P$  and  $T_1$ . Moreover,  $D_2$ , which consists of two dimensional action variables consisting of both a discrete action (agent)  $d_2$  mixed with a continuous action (time)  $T_M$ , provides the patient a survival time  $T_2$  given  $T_D \geq t_2$ . After adjusting for censoring, the reward function for the first stage is just

$$T_1 \sim R_1(o_1, d_1).$$

In the second stage, the reward function is defined by  $T_2$  conditional on  $Y_D = 1$ , which is equivalent to  $T_2$  conditional on  $T_D \geq t_2$ , where  $T_2$  satisfies

$$T_2 \sim R_2(o_2, d_1, d_2, T_M).$$

Functions  $R_1$  and  $R_2$  are obtainable from  $f_1$  and  $f_3$ , defined previously, and are usually not observable. Note also that both  $T_1$  and  $T_2$  are censored rather than directly observed. In Q-learning, because for every state there are a number of possible treatments that could be taken, each treatment within each state has a value according to how long the patient will survive due to completion of that treatment. The scientific goal of our study is to find an optimal policy to maximize patients' overall survival time  $T_D$ . This is accomplished by learning which treatment (including starting time for second-line therapy) is optimal for each state.

While learning a non-stationary non-Markovian optimal policy from a clinical reinforcement trial data set

$$\{O_1, D_1, T_1 \wedge C, O_2, D_2, T_2 \wedge C_2\},$$

we denote the estimation of the optimal Q-functions based on this training data by  $\widehat{Q}_t$ , where  $t = 1, 2, 3$ . The indexes 1 and 2 correspond to the decision times  $t_1$  and  $t_2$  while index 3 is included only for mathematical convenience. According to the recursive form of Q-learning in (2), we must estimate  $Q_t$  backwards through time, that is, use the estimate  $Q_3$  from the last time point back to  $Q_1$  at the beginning of the trial. For convenience we set  $Q_3$  equal to 0. In order to estimate each  $Q_t$ , we denote  $Q_t(O_t, D_t; \theta_t)$  as a function of a set of parameters *boldsymbol* $\theta_t$ , and we allow the estimator to have different parameter sets for different time points  $t$ . Once this backwards estimation process is done, we save  $\widehat{Q}_1$  and  $\widehat{Q}_2$ , and we thereafter use them to respectively estimate optimal treatment policies

$$\widehat{\pi}_1 = \arg \max_{d_1} \widehat{Q}_1(o_1, d_1; \theta_1)$$

and

$$\widehat{\pi}_2 = \arg \max_{d_2, T_M} \widehat{Q}_2(o_2, d_2, T_M; \theta_2),$$

for new patients. Since the resulting estimated optimal policies are functions of patient covariates, the resulting treatment regimens are individualized. These individualized treatment regimens should also be evaluated in a follow-up confirmatory phase III trial comparing the optimal regimens with the standard of care or other appropriate fixed (i.e., non-individualized) treatments.

### 2.3 Support Vector Regression

A strength of Q-learning is that it is able to compare the expected reward for the available treatments without requiring a model of the relationship. To achieve this, the main task is to estimate the  $Q$  functions for finding the corresponding optimal policy. However, challenges may arise due to the complexity of the structure of the true  $Q$  function, specifically, the non-smooth maximization operator in the recursive equation (2).

Nonparametric statistical methods are appealing for estimating  $Q$  functions due to their robustness and flexibility. For instance, using random forest (RF) or extremely randomized trees (ERT) techniques is very effective for extracting a

well-fitted  $Q$  functions (Ernst, Geurts, and Wehenkel, 2005; Geurts, Ernst, and Wehenkel, 2006; Guez, Vincent, Avoli, and Pineau, 2008; Zhao et al., 2009). Besides the RF and ERT methods, other methodologies for fitting  $Q$  include, but are not limited to, neural networks, kernel-based regressions (Ormoneit and Sen, 2002), and support vector machines (SVM) (Vapnik, 1995). Our experience so far indicates that both SVR and ERT work quite well and their accuracy is approximately equivalent, although ERT is more computationally intense.

In the present article we apply SVR as our main method to fit  $Q$  functions and learn optimal policies using a training data set. SVR provides a compromise between the parametric and purely nonparametric approaches. The ideas underlying SVR are similar but slightly different from SVM within the margin-based classification scheme. To illustrate, consider the case where the rewards in the training data set are not censored. At each stage, given  $(\mathbf{x}_i, y_i)_{i=1}^n$ , where attributes  $\mathbf{x}_i \in \mathbb{R}^m$  and label index  $y_i \in \mathbb{R}$ , the goal in SVR is to find a function  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  that closely matches the target  $y_i$  for the corresponding  $\mathbf{x}_i$ . Note that in our simulation study in Section 4,  $\mathbf{x}_i$  may be replaced by information of states along with actions and  $y_i$  may be replaced by survival time, respectively. Instead of the hinge loss function used in SVM, one of the popular loss functions involved in SVR is known as the  $\epsilon$ -insensitive loss function (Vapnik, 1995), which is defined as

$$L(f(\mathbf{x}_i), y_i) = (|f(\mathbf{x}_i) - y_i| - \epsilon)_+, \quad (3)$$

where  $\epsilon > 0$  and the subscript  $+$  denotes taking the positive part. That is, as long as the absolute difference between the actual and the predicted values is less than  $\epsilon$ , the empirical loss is zero, otherwise there is a cost which grows linearly. SVR is more general and flexible than least-squares regression, since it allows a predicted function that has at most  $\epsilon$  deviation from the actually obtained targets  $y_i$  for all the training data. The lack of differentiability in (3) implies a difficulty for efficient optimization, but SVR solves an alternative optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \xi'} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C_E \sum_{i=1}^n (\xi_i + \xi'_i), \\ \text{subject to} \quad & (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) - y_i \leq \epsilon + \xi_i, \\ & y_i - (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \leq \epsilon + \xi'_i, \\ & \xi_i, \xi'_i \geq 0, \quad i = 1, \dots, n. \end{aligned} \quad (4)$$

$\mathbf{w}^T \Phi(\mathbf{x}_i) + b$  is defined as a separating hyperplane, where  $\Phi$  is a nonlinear transformation which maps data into a feature space.  $\xi_i$  and  $\xi'_i$  are slack variables and  $C_E$  is the cost of error. By minimizing the regularization term  $\frac{1}{2} \|\mathbf{w}\|^2$  as well as the training error  $C_E \sum_{i=1}^n (\xi_i + \xi'_i)$ , SVR can avoid both overfitting and underfitting of the training data. A class of functions called kernels  $K : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$  such that  $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$  (for example, the Gaussian kernel is  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\zeta \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ ) are used in SVR to guarantee

that any data set becomes arbitrarily separable as the data dimension grows. Since the SVR function is derived within this reproducing kernel Hilbert space (RKHS) context, the explicit knowledge of both  $\Phi$  and  $\mathbf{w}$  are not needed if we have information regarding  $K$ . In this case, problem (4) is equivalent to solving an optimization dual problem equipped with Lagrange multipliers  $\lambda_i$ :

$$\begin{aligned} \min_{\lambda, \lambda'} \quad & \frac{1}{2}(\boldsymbol{\lambda} - \boldsymbol{\lambda}')^T K(\mathbf{x}_i, \mathbf{x}_j)(\boldsymbol{\lambda} - \boldsymbol{\lambda}') - \sum_{i=1}^n (y_i - \epsilon)\lambda'_i + \sum_{i=1}^n (y_i + \epsilon)\lambda_i, \\ \text{subject to} \quad & \sum_{i=1}^n (\lambda_i - \lambda'_i) = 0, \\ & 0 \leq \lambda_i, \lambda'_i \leq C_E, \quad i = 1, \dots, n. \end{aligned} \quad (5)$$

Both parameters  $\zeta$  and  $C_E$  in SVR are obtained by utilizing cross validation to achieve good performance. Once the above formulation is solved to get the optimal  $\lambda_i$  and  $\lambda'_i$ , the approximating function at  $\mathbf{x}$  is given by:

$$f(\mathbf{x}) = \sum_{i=1}^n (\lambda'_i - \lambda_i)K(\mathbf{x}_i, \mathbf{x}) + b. \quad (6)$$

Unfortunately, this approach as is cannot be implemented in the presence of censoring.

## 2.4 Support Vector Regression for Censored Subjects

Note that we have in the prior section assumed that all patients are followed until they die. In conducting an NSCLC trial, a common problem is the right censoring caused by patients who do not complete the study and drop out of the study without further measurements. Possible reasons for patients dropping out of the study include, adverse reactions, lack of improvement, unpleasant study procedures, and other factors related or unrelated to the trial procedure and treatments. As mentioned previously, we assume in this paper that censoring is independent of death.

In general, we denote interval censored data by  $(\mathbf{x}_i, l_i, u_i)_{i=1}^n$ . If the patient experiences the death event and  $T_D$  is observed rather than being interval censored then we include  $T_D$  and denote such an observation by  $(\mathbf{x}_i, y_i)$ . In other words, when we observe  $T_D$  exactly ( $\delta = 1$ ), we let  $l_i = u_i = y_i$ . Note that by letting  $u_i = +\infty$  we can easily construct a right censored observation  $(\mathbf{x}_i, l_i, +\infty)$ .

One naive way to handle censored data within Q-learning by using SVR is to consider only those samples for which the survival times  $T_D$  are known exactly. Such an approach which totally ignores censoring will both reduce and bias the sample for statistical analysis and inference. Thus the more patients that are censored, or the earlier they are censored, the more unreliable the results will be. An SVR procedure that targets interval censored subjects was introduced by

Shivaswamy, Chu, and Jansche (2007). The key component of their procedure is a loss function, defined as  $L(f(\mathbf{x}_i), l_i, u_i) = \max(l_i - f(\mathbf{x}_i), f(\mathbf{x}_i) - u_i)_+$ . However, this loss function does not have  $\epsilon$ -insensitive properties, that is, it does not allow  $\epsilon$  or other deviations from the predicted  $f(\mathbf{x}_i)$ , especially when  $l_i = u_i = y_i$ . In this article, we propose a modified SVR algorithm with  $\epsilon$ -insensitive loss function (called  $\epsilon$ -SVR-C) to make use of both survival times and censoring times in the data set and to reduce the potential bias which may be caused by performing a classical SVR with censored data.

Given the interval censored data set  $(\mathbf{x}_i, l_i, u_i)_{i=1}^n$ , our modified loss function is defined as

$$L(f(\mathbf{x}_i), l_i, u_i) = \max(l_i - \epsilon - f(\mathbf{x}_i), f(\mathbf{x}_i) - u_i - \epsilon)_+. \quad (7)$$

The main difference between (3) and (7) is that  $y_i$  is separated into two parts which are replaced by  $l_i$  and  $u_i$ , respectively. We remark that this loss function does not penalize values of  $f(\mathbf{x}_i)$  if it is between  $l_i - \epsilon$  and  $u_i + \epsilon$ . On the other hand, the cost grows linearly if this output is more than  $u_i + \epsilon$  or less than  $l_i - \epsilon$ . Figure 3 shows the loss function of the modified SVR. Note that when  $u_i = +\infty$ , this loss function becomes one sided, which means there is no empirical error if  $f(\mathbf{x}_i) \geq l_i - \epsilon$ . In addition, when the data is not observed as censored, our modified SVR algorithm reduces to the classical SVR.

Defining index sets  $L = \{i : l_i > -\infty\}$  and  $U = \{i : u_i < +\infty\}$ , the corresponding modified SVR optimization formulation is:

$$\min_{\mathbf{w}, b, \xi, \xi'} \frac{1}{2} \|\mathbf{w}\|^2 + C_E \left( \sum_{i \in L} \xi_i + \sum_{i \in U} \xi'_i \right),$$

$$\begin{aligned} \text{subject to} \quad & (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) - u_i \leq \epsilon + \xi_i, \quad i \in U, \\ & l_i - (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \leq \epsilon + \xi'_i, \quad i \in L, \\ & \xi_i \geq 0, \quad i \in L; \quad \xi'_i \geq 0, \quad i \in U. \end{aligned}$$

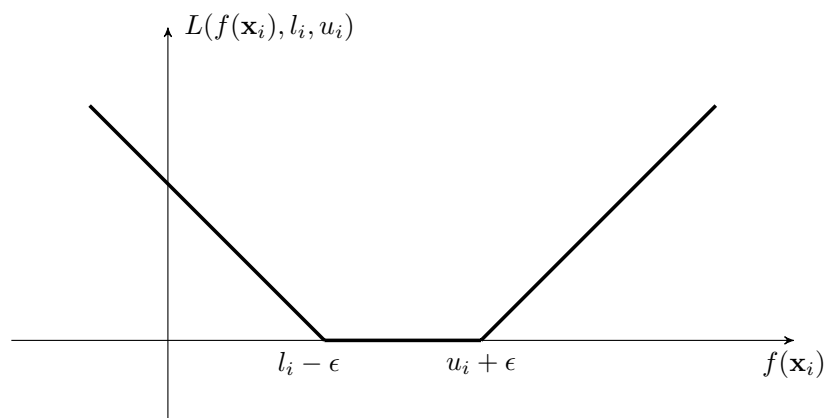
Similarly to classical SVR, the dual can be presented as follows by introducing the Lagrange multiplier  $\lambda_i$ :

$$\min_{\lambda, \lambda'} \frac{1}{2} (\boldsymbol{\lambda} - \boldsymbol{\lambda}')^T K(\mathbf{x}_i, \mathbf{x}_j) (\boldsymbol{\lambda} - \boldsymbol{\lambda}') - \sum_{i \in L} (l_i - \epsilon) \lambda'_i + \sum_{i \in U} (u_i + \epsilon) \lambda_i,$$

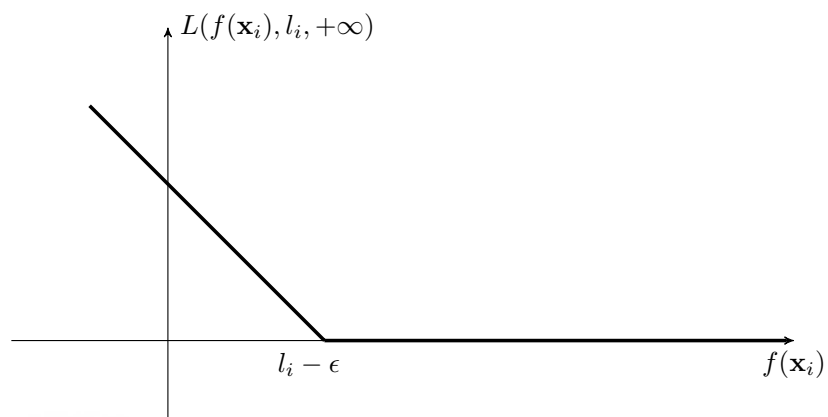
$$\text{subject to} \quad \sum_{i \in L} \lambda'_i - \sum_{i \in U} \lambda_i = 0,$$

$$0 \leq \lambda_i, \lambda'_i \leq C_E, \quad i = 1, \dots, n.$$

Once the above formulation is solved to get the optimal  $\lambda_i$  and  $\lambda'_i$ , the approximate function at  $\mathbf{x}$  can be obtained and has the same form as equation (6). Based on results for non-censored Q-learning with classical SVR, it is expected that the  $\epsilon$ -SVR-C behaves similarly, with the estimated policies  $\hat{\pi}$  being more robust to censored data and being more optimal than results where the censored patients are simply ignored. The effectiveness of  $\epsilon$ -SVR-C will be demonstrated in a small simulation study reported later in this paper.



(a)



(b)

Figure 3: Modified SVR loss functions for interval censored data (a) and right censored data (b).

### 3 Trial Conduct and Computational Strategy

Different populations of patients with NSCLC appear to have different clinical and molecular characteristics, so clinical trials that investigate the activity of different agents, and incorporate patient selection based on clinical factors, are required. We will now describe a virtual clinical reinforcement trial which provides a realistic approximation to a potentially real NSCLC trial that evaluates two-line treatment strategies for patients with NSCLC who have not been treated previously with systemic therapy. As mentioned in Section 1, while many new single agents with potential clinical efficacy currently are being produced at an increasing rate, the number of doublet combinations in the first line that can be evaluated clinically is limited. Considering the number of possible agents that may be of interest in the second line, the limitations are even greater.

Without loss of generality, suppose for simplicity that strategies are based on four FDA approved therapies (either single agents or doublets), which we denote by  $A_i, i = 1, \dots, 4$ . In our study we assume that the second line treatment must be different from the first. When designing the trial, two of the four agents  $A_1$  and  $A_2$  are selected for first-line treatment, while  $A_3$  and  $A_4$  are selected for second line. A total of  $N$  patients are recruited into the trial and fairly randomized at enrollment between  $A_1$  and  $A_2$ , and each patient is followed through to completion of first-line treatment, given such patient is not dead or lost to follow-up from the study. We fix this duration  $t_2 - t_1$  as 2.8 months, although other lengths are possible, depending on the number of cycles of treatment. At the end of first-line treatment, patients are randomized again between agents  $A_3$  and  $A_4$ . Moreover, another important decision necessary to make at this point is when to initiate the second-line treatment. Thus, the initiation for second-line treatment could be randomized to as early as  $t_2$  or as late as  $T_P$  (recall that  $T_P$  denotes the time of patient's disease progression). This will be accomplished by randomizing to a target initiation time  $T_M$  over the interval  $[0, 2]$  (in months) and then initiating second line therapy at  $T_M \wedge (T_P - t_2)$ . At the end of the trial, the patient data is collected and Q-learning is applied, in combination with SVR applied at each time point, to estimate the optimal treatment rule as a function of patient variables and biomarkers, at  $t_1$  and  $t_2$ .

The trial described above was motivated by the desire to compare several agents as well as timing in a randomized fashion, the belief that different agents combined with different timing given consecutively may have different effects for different populations of patients, and the desire to determine a sound basis for selecting individualized optimal strategies for evaluation in a future clinical trial. Putting this all together, the entire algorithm for Q-function estimation and optimal treatment discovery can be summarized as follows:

1. Inputs: If  $t = 1$ , a set of training data consists of attributes  $\mathbf{x}_i$  (states  $o_1$ , actions  $d_1$ ) and index  $y_i$  (censored rewards  $\{T_1 \wedge C, \delta\}$ ), i.e.  $\{(o_1, d_1, T_1 \wedge C, \delta)_i, i = 1, \dots, n\}$ ; if  $t = 2$  and  $Y_D = 1$ , a set of training data  $\{(o_2, d_2, T_M, T_2 \wedge C_2)_j, j = 1, \dots, n'\}$ , where  $n' \leq n$  since patients may die or be censored



before second-line therapy.

2. Initialization: Let  $\widehat{Q}_3$  be a function equal to zero.
3.  $Q_2$  is fitted with  $\epsilon$ -SVR-C through the following equation:

$$Q_2(o_2, d_2, T_M) = T_2 + \text{error},$$

where  $T_2$  is assessed through the censored observation  $\{T_2 \wedge C_2, \delta\}$ . This is possible to do since we are restricting ourselves in this step to patients for whom  $Y_D = 1$ .

4.  $Q_1$  is fitted with  $\epsilon$ -SVR-C through the following equation:

$$\begin{aligned} Q_1(o_1, d_1) &= T_1 + I(T_1 = t_2) \max_{d_2, T_M} \widehat{Q}_2(o_2, d_2, T_M) + \text{error} \\ &= T_1 + I(T_1 = t_2) \widehat{T}_2 + \text{error}, \end{aligned}$$

where  $T_1 + I(T_1 = t_2) \widehat{T}_2$  is assessed through the censored observation  $(\tilde{X}, \tilde{\delta}) = (T_1 \wedge C + Y_D \widehat{T}_2, \delta + (1 - \delta) Y_D)$ . The reason this works can be summarized in two steps: First, we can show after some algebra that  $\tilde{X} = \widehat{T}_D \wedge \tilde{C}$  and  $\tilde{\delta} = I(\widehat{T}_D \leq \tilde{C})$ , where  $\widehat{T}_D = T_1 + I(T_1 = t_2) \widehat{T}_2$  and  $\tilde{C} = CI(C < t_2) + \infty I(C \geq t_2)$ , and thus we have independent right censoring of the quantity  $\widehat{T}_D$ . Second, since  $Q_1$  needs to model the expectation of  $T_D$  given the covariates  $(O_1, D_1)$ , it is appropriate if we replace  $T_D$  with the quantity  $T_1 + I(T_1 = t_2) E(T_2 | O_1, D_1, T_D \geq t_2)$ . Since  $\widehat{T}_2$  is an estimate of the latter conditional expectation, our approach is valid.

5. For the SVR computations in steps 3 and 4, if a Gaussian kernel is applied, we use a straightforward coarse grid search with  $C_E = 2^{-5}, 2^{-3}, \dots, 2^{15}$  and  $\zeta = 2^{-15}, 2^{-13}, \dots, 2^3$ , evaluated at each candidate pair  $(C_E, \zeta)$ , and then select the one that yields the highest cross-validation rate.
6. Given  $\widehat{Q}_1$  and  $\widehat{Q}_2$ , the individualized optimal policies  $\widehat{\pi}_1$  and  $\widehat{\pi}_2$  for application to future patients are computed.

## 4 Simulation Study

To demonstrate that the tailored therapy for NSCLC found by using the proposed clinical reinforcement trial is superior, we employ an extensive simulation study to assess the proposed approach on virtual clinical reinforcement trials of patients, and then evaluate using phase III trial-like comparisons between the estimated optimal regimen and the various possible fixed treatments.

## 4.1 Data Generating Models

Based on historical research, it is well known that the rate of disease progression or death for patients with advanced NSCLC increases over time. Consequently, in order to generate simulated data, we simply consider that  $T_1$ ,  $T_P - t_2$ , and  $T_2$  conditional on  $T_D \geq t_2$  follow different exponential distributions. Many alternative models are also possible.

Let  $\exp(x)$  denote an exponential distribution with mean  $e^x$ . Also let  $W_t$  and  $M_t$  be patient prognostic factors observable at  $t = 1, 2$  (corresponding to times  $t_1$  and  $t_2$ ) which will be defined shortly. For a patient given first-line treatment  $d_1$ , we assume that  $T_1 = \tilde{T}_1 \wedge t_2$ , where

$$[\tilde{T}_1 | D_1] \sim \exp(\alpha_{D_1} + \beta_{D_1}W_1 + \kappa_{D_1}M_1 + \tau_{D_1}W_1M_1). \quad (8)$$

If  $\tilde{T}_1 \geq t_2$ , we generate  $T_M$  from a uniform  $[0, 2]$  distribution. We now absorb  $T_P$  into  $T_M$  for modeling  $T_2$  given  $T_D \geq t_2$  through an intent-to-treat structure (basically, we can ignore  $T_P$  since it depends only on  $D_1$ ,  $M_1$  and  $W_1$  and not on  $T_M$ ). In addition, for a patient given second-line treatment  $d_2$  and initiation time  $T_M$ , we assume the death time

$$[T_2 | D_1, D_2] \sim \exp(\alpha_{D_{12}} + \beta_{D_{12}}W_2 + \kappa_{D_{12}}M_2 + h(T_M; \varphi)), \quad (9)$$

where  $h(T_M; \varphi)$  is a function depending on the parameter  $\varphi$  which reflects the effect of timing  $T_M$  on death. The total time to death is then  $T_D = T_1 + I(T_1 = t_2)T_2$ . We then need to generate the right censoring time  $C$  uniformly from the interval  $[t_1, t_1 + u]$ . To find  $u$ , we estimate the unconditional survival function  $\hat{S}(t)$  for the failure time  $T_D$ , where “unconditional” refers to taking expectation over the covariates  $D_i, W_i, M_i (i = 1, 2)$ , and  $T_M$  of the conditional survival function  $T_D$ . Then,  $u$  is the solution to

$$\frac{1}{u} \int_{t_1}^{t_1+u} \hat{S}(x) dx = p,$$

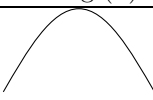
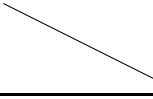
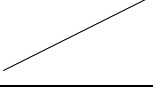
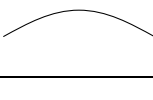
where  $p$  is the desired probability of censoring.

Note that in our simulation study we straightforwardly use exponential pdfs (8) and (9) to replace  $f_1$  and  $f_3$  and we drop  $f_2$ , where  $(f_1, f_2, f_3)$  were described in Section 2.1. For the sake of simplicity, in these density functions only two state variables, quality of life (QOL)  $W_t$  and tumor size  $M_t$ , are considered as patient prognostic factors or biomarkers to be related to outcome. We consider these two factors because they are patient based, realistically easy to measure, can predict therapeutic benefit after treatment of chemotherapy, and, more importantly, they are significant prognostic factors for survival (Socinski et al., 2007). In addition, state variables for the next decision are generated by the simple dynamic models  $W_2 = W_1 + T_M \dot{W}_1$  and  $M_2 = M_1 + T_M \dot{M}_1$ , where  $\dot{W}_1$  and  $\dot{M}_1$  are constants.

The parameter vector for patients who only experience first-line treatment is

$$\boldsymbol{\theta}_1 = (\alpha_{D_1}, \beta_{D_1}, \kappa_{D_1}, \tau_{D_1}),$$

Table 1: The scenarios studied in the simulation. Sample size = 100/group.

Group	State Variables	Status	Timing ( $h$ )	Optimal Regimen
1	$W_1 \sim N(0.25, \sigma^2)$ $M_1 \sim N(0.75, \sigma^2)$	$W_1 \downarrow M_1 \uparrow$		$A_1 A_3 2$
2	$W_1 \sim N(0.75, \sigma^2)$ $M_1 \sim N(0.75, \sigma^2)$	$W_1 \uparrow M_1 \uparrow$		$A_1 A_4 1$
3	$W_1 \sim N(0.25, \sigma^2)$ $M_1 \sim N(0.25, \sigma^2)$	$W_1 \downarrow M_1 \downarrow$		$A_2 A_3 3$
4	$W_1 \sim N(0.75, \sigma^2)$ $M_1 \sim N(0.25, \sigma^2)$	$W_1 \uparrow M_1 \downarrow$		$A_2 A_4 2$

otherwise, it is

$$\theta_2 = (\alpha_{D_1}^P, \beta_{D_1}^P, \kappa_{D_1}^P, \tau_{D_1}^P, \alpha_{D_{12}}, \beta_{D_{12}}, \kappa_{D_{12}}, \varphi).$$

Parameters vectors  $\theta_1$  and  $\theta_2$  as well as the shape of the time-related function  $h(T_M; \varphi)$  vary among different patients. Note that two patients who receive different decisions with the same first-line treatment, say  $(A_1, A_3)$  and  $(A_1, A_4)$ , both contribute data for estimating  $Q_1$ .

## 4.2 Clinical Scenarios

To construct a set of scenarios reflecting the interaction between two lines of treatment, we temporarily assume that a large portion of patients survive long enough to be treated by second-line therapy, that is, we adjust the parameters so that  $\nu = 0.8$  averaged across all patients. Other than the constraint on  $\nu$ , each clinical scenario under which we will evaluate the design in the simulation study is built by a unique set of fixed values of  $(\alpha_{D_1}^P, \beta_{D_1}^P, \kappa_{D_1}^P, \tau_{D_1}^P, \alpha_{D_{12}}, \beta_{D_{12}}, \kappa_{D_{12}})$ . The remaining fixed parameter values needed for the simulations are those that determine how  $T_2$  varies as a function of  $T_M$ . To implement this, we specified four corresponding model-based cases of each function  $h(T_M; \varphi)$  in terms of their numerical values at each  $T_M$ . All of these underscore the importance of specifying the optimal regimen to target a subpopulation of patients with distinct characteristics.

Hence, to facilitate interpretation of reinforcement learning strategies for capturing individualized therapies, four scenarios are specified and summarized in Table 1. In group 1 and 4, initial timing of second-line therapy for survival time ( $T_2$ ) are functions that form an inverse-U (quadratic) shape with  $T_M$ ,

while initial timing in group 2 and 3 for  $T_2$  are functions that linearly decrease and increase with  $T_M$ , respectively. Each group thus consists of a combination  $(A_i, A_j)$  as well as  $T_M$  from Table 1 (where  $i = 1, 2$  and  $j = 3, 4$ ), with the fixed values of  $\alpha_{D_1}^P, \beta_{D_1}^P, \kappa_{D_1}^P, \tau_{D_1}^P, \alpha_{D_{12}}, \beta_{D_{12}}, \kappa_{D_{12}}$ , and  $\varphi$  as described above.

Note that whatever combination of two-line treatment  $(A_i, A_j)$  is evaluated, all patients within one group share the same trend of  $T_2$  versus  $T_M$ . However, we assume there is only one strategy that will yield the longest survival in each group. For convenience, we denote “1, 2, 3” as the location of optimal initiation of second-line therapy, say “immediate, intermediate, delayed”, respectively. For example, as claimed in the last column in Table 1,  $A_1A_32$  indicates that the two-line treatments  $(A_1, A_3)$  along with an intermediate initiation time point is the optimal regimen for group 1. The inverse-U-shaped function  $T_2$  for  $T_M$  corresponds to the case where patients have relatively low QOL at enrollment but relatively large tumor size, hence, this optimal intermediate initiation of second-line therapy is recommended to delay treatment a short time for patients who may have severe symptoms and low tolerance of chemotherapy, but not to fully delay due to the possibility of death. In scenario 2, due to the good QOL and large tumor size at enrollment, it is optimal for the second-line therapy to begin immediately after first-line therapy, hence,  $A_1A_41$  is the optimal regimen for these patients. Similarly, in scenario 3, treatment  $A_2A_33$  is considered the superior treatment since we believe fully that delaying the initiation of second-line therapy at the time of disease progression will improve survival and palliate symptoms. Although scenario 4 has optimal regimen  $A_2A_42$ , due to the flat shape of  $T_2$  versus  $T_M$ , there is no significant improvement between delaying and not delaying the initiation of second-line therapy. In this manner, many plausible effects of treatment are captured, at least to some degree, including both reversible and irreversible toxicities resulting from chemotherapy.

### 4.3 Simulation Methods and Results

First, according to various  $(W_1, M_1)$  as described in Table 1, a non-censored sample of  $N = 100$  virtual patients for each of the four disease profile groups (with total sample size  $n = 400$ ) is generated.  $\hat{Q}_1$  and  $\hat{Q}_2$  are computed via the algorithm given in Section 3. The predicted optimal strategies are then computed, and an independent testing sample of size 100 per disease profile group (hence also totalling 400) is also generated. For evaluation purposes, we then assign all virtual test patients to all possible combinations of  $(A_i, A_j) \times \{\text{immediate, intermediate, delayed}\}$  as well as the estimated optimal strategy, resulting in 13 possible treatments. Patients’ outcomes (overall survival) conducted by our estimated optimal regimens and different 12 fixed regimens are all evaluated. This is similar in spirit to a virtual phase III trial with  $5200 = 13 \times 400$  patients, except that the estimated effects will be more precise. Moreover, we repeated the simulations 10 times for the training sample trial (with total sample size  $n = 400$ ). Then, 10 estimated optimal strategies learned from these 10 training trials were applied to the same testing patients described above. All of the results for each of the 13 treatments are averaged over the

Table 2: Comparisons between true optimal regimens and estimated optimal regimens for overall survival (month). Each training dataset is of size 100/group with 10 simulation runs. The testing dataset is of size 100/group.

Group	Optimal regimen	Optimal timing	True survival	Selected timing	Predicted survival		
					Min	Mean	Max
1	$A_1A_32$	3.80	16.00	3.92	15.83	15.93	16.00
2	$A_1A_41$	2.80	15.33	2.94	14.96	15.13	15.28
3	$A_2A_33$	4.80	18.37	4.62	17.75	17.99	18.27
4	$A_2A_42$	3.80	20.75	4.11	20.60	20.86	20.97
Average			17.61		17.28	17.48	17.63

400 test patients. As shown in Figure 4, among regular regimens, assigning all testing patients to  $A_2A_32$  will yield the averaged longest survival among the 12 fixed treatments at 11.29 months. It thus appears that, in terms of adaptively selecting best strategies for each group, the optimal regimen obtained by Q-learning with SVR is superior due to its average (over 10 simulations) survival of 17.48 months. The survival curves for the groups (based on the Kaplan-Meier estimates) are shown in Figure 5, which demonstrates the effectiveness of the proposed approach for prolonging survival. Because of this encouraging result, it is worthwhile to deeply investigate whether our approximations are close to the exact solution. To carefully examine this comparison, we assign test patients from each disease profile group to the corresponding true optimal regimen described in Table 1 to obtain the “True survival” column of Table 2. The minimum, maximum, and mean values of averaged predicted survival for each group are computed based on these 10 trials, respectively. The results are summarized in Table 2. The averaged predicted survival over all groups is shown as 17.48 (which is consistent with the number shown below the “optimal” bar in Figure 4), this number along with minimum 17.28 and maximum 17.63 are all pretty close to true optimal survival 17.61. In addition to claim that the frequencies of selecting optimal regimens ( $A_i, A_j$ ) as true regimen is 100%, the averaged selected optimal timings are shown in the fifth column of Table 2. Note that they are close to true optimal timings for each group. In terms of estimation, under each of the scenarios 1-3 our methods perform very similarly and slightly underestimates the true optimal survival. In contrast, our method slightly overestimates the true optimal survival in scenario 4.

Second, although our Q-learning method with  $N = 100$  per group using SVR leads to an apparently small bias for estimating individualized optimal regimens, an examination of performance influenced by the sample size is worthwhile. We repeated the simulations 10 times for each specified sample size while varying  $N$  from 2 to 600 per group. The results are illustrated in Figure 6, which shows that the method’s reliability is very sensitive to  $N$  when  $N \leq 80$ , with the averaged survival for the estimated optimal strategy increasing from 14.192

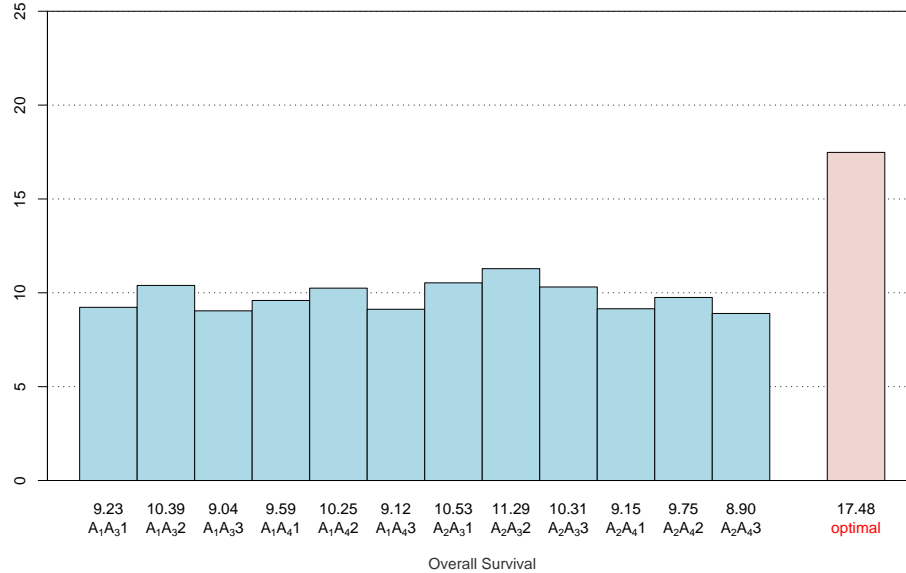


Figure 4: Performance of optimal individualized regimen versus other 12 combinations.

when  $N = 2$  to 17.479 when  $N = 80$ . The boxplots also show that both the variance and estimation bias of predicted survival are getting smaller when the sample size becomes larger. When  $N \geq 100$ , our methods appear to do a very reliable job of selecting the best strategy. Hence, in the setting we study here, the sample sizes required to reach an excellent approximation are similar to and not larger than the sizes required for typical phase III trials.

Third, in order to compare performance of  $\epsilon$ -SVR-C for censored subjects to ignoring the censored cases and using SVR, from 400 training samples over 10 simulations run, we randomly censor as described in Section 4.1 to achieve a targeted proportion of censoring  $p$ , estimate the optimal treatment policy using  $\epsilon$ -SVR-C, throw out the censored observations and use SVR to estimate the optimal policy, and then apply 400 testing patients to the estimated regimens to estimate the average survival. This is done for 25%, 50%, and 75% censoring proportions  $p$ , respectively. The boxplots are presented in Figure 7. For instance, in panel (a) we generate 10 training trials with 25% censoring. The left boxplot in Figure 7(a) indicates the performance for the optimal policy estimated under our proposed method without any censoring. The middle boxplot indicates the performance based on  $\epsilon$ -SVR-C applied to the 25% censored data, while, in the right boxplot, we simply delete the 25% of patients which are right-censored and apply SVR to the remaining data to estimate the optimal policy. This basic process is repeated across the three different censoring levels.

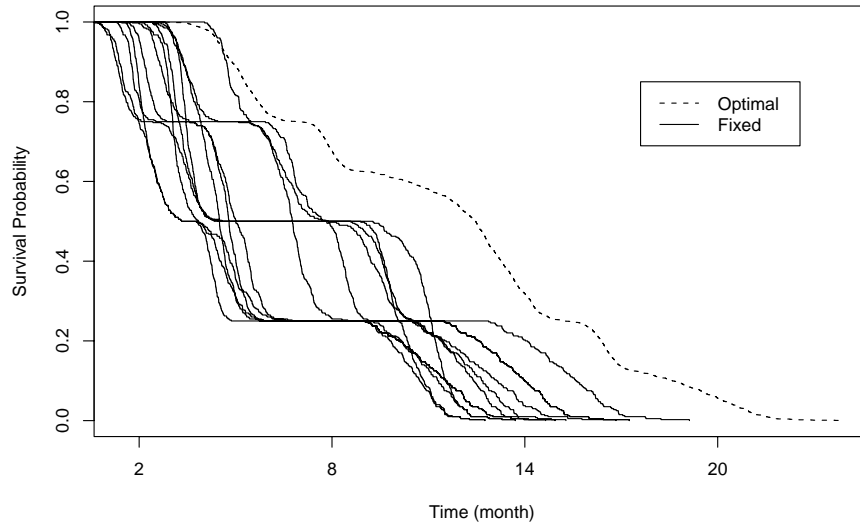


Figure 5: Survival functions for testing sample treated by 13 different regimens (12 fixed treatments plus optimal regimen).

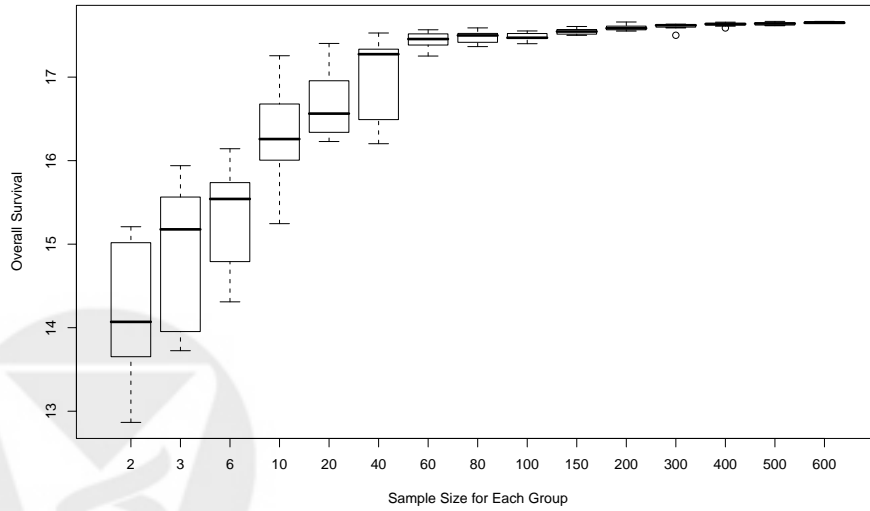


Figure 6: Sensitivity of the predicted survival to the sample size.

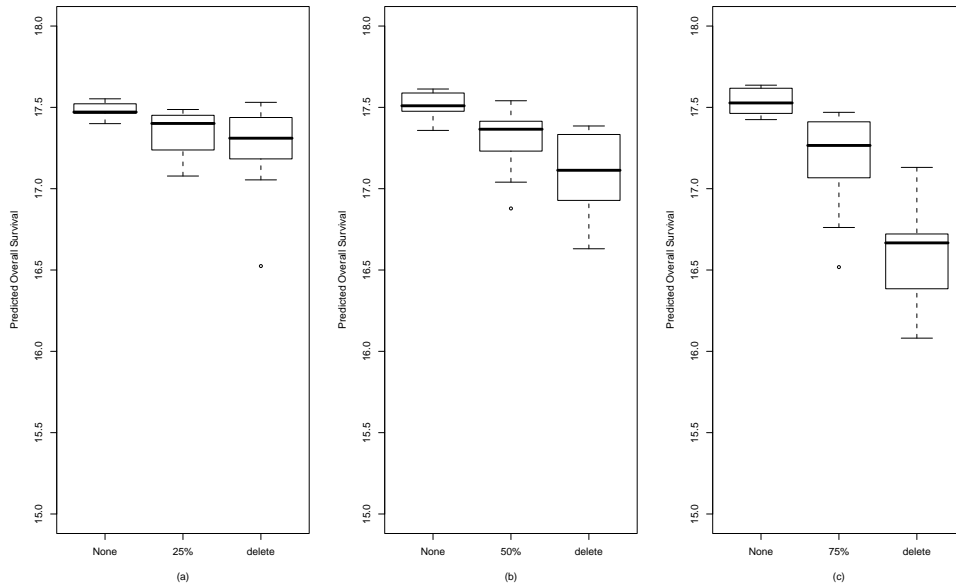


Figure 7: Boxplots of the predicted survival for the optimal policy estimated from  $\epsilon$ -SVR-C by using training data with 25% (a), 50% (b), and 75% (c) fraction of right censored subjects. In each panel, from left to right, the three boxplots indicate performance from no censoring and using SVR, from right-censoring and using  $\epsilon$ -SVR-C, and from throwing out censoring and using SVR, respectively.

This means that the data presented in the left columns of all three panels are random replications of the same data scenario. As can be observed in Figure 7, as the fraction of right-censoring increases, there is an increasing decline in performance resulting from throwing out censored observations. In contrast, our proposed approach (the middle boxplot of each panel) can robustly estimate the optimal policy under censoring, with only a minor increase in bias as censoring increases. Clearly, in terms of averaged predicted survival in all cases, the  $\epsilon$ -SVR-C algorithm outperforms the method which totally ignores the censored data, particularly when the censoring proportion is large.

## 5 Discussion

We have proposed a clinical reinforcement trial design for discovering individualized therapy for multiple lines of treatment in a group of patients with advanced NSCLC. The incorporation of Q-learning with the proposed  $\epsilon$ -SVR-C appears to successfully identify optimal treatment strategies tailored to appropriate sub-



populations of patients. While our method has been utilized for the two decision points at hand, the general concepts and algorithms of this approach could be applied, with suitable modification, to design future trials having similar goals but for possibly different diseases. Although overall survival time is considered among many clinicians to be the appropriate primary endpoint in late stage NSCLC, a potentially important alternative outcome to consider in future cancer clinical reinforcement trial research is quality-of-life-adjusted survival (Gelber, et al., 1995). This may require some modification of the proposed  $\epsilon$ -SVR-C methodology.

In this article, we studied the prediction accuracy of our method with varying sample sizes. The simulation studies show that with sample size  $N \geq 100$  our method can yield a small estimation bias. However, an important and challenging question is: how do we determine an appropriate sample size for a clinical reinforcement trial to reliably obtain a treatment policy that is very close to the true optimal policy? This sample size calculation is related to the statistical learning error problem. Recently, there has been considerable interest in studying the generalization error for Q-learning. Murphy (2005b) derived finite sample upper bounds in a closely related setting which depends on the number of observations in the training set, the number of decision points, the performance of the approximation on the training set, and the complexity of the approximation space. We believe further development of this theory is needed to better understand how the performance of Q-learning with SVR is related to the sample size of the training data in clinical reinforcement trials. We hope that this article will serve to stimulate interest in these issues.

## Acknowledgements

The authors would like to thank the Reinforcement Learning Group at the University of North Carolina for many stimulating exchanges. The research was funded in part by grant CA075142 from the U.S. National Cancer Institute and from pilot funding provided by the Center for Innovative Clinical Trials at the UNC Gillings School of Global Public Health.

## References

- [1] Bunn, P. and Kelly, K. (1998). New chemotherapeutic agents prolong survival and improve quality of life in non-small cell lung cancer: A review of the literature and future directions. *Clinical Cancer Research* **4**, 1087–1100.
- [2] Ciuleanu, T. E., Brodowicz, T., Belani, C. P., Kim, J., Krzakowski, M., Laack, E., Wu, Y., Peterson, P., Adachi, S., and Zielinski, C. C. (2008). Maintenance pemetrexed plus best supportive care (BSC) versus placebo plus BSC: A phase III study. *Journal of Clinical Oncology* **26**, May 20 suppl, abstract 8011.

- [3] Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch model reinforcement learning. *Journal of Machine Learning Research* **6**, 503–556.
- [4] Fidias, P., Dakhil, S., Lyss, A., Loesch, D., Waterhouse, D., Cunneen, J., Chen, R., Treat, J., Obasaju, C., and Schiller, J. (2007). Phase III study of immediate versus delayed docetaxel after induction therapy with gemcitabine plus carboplatin in advanced non-small-cell lung cancer: Updated report with survival. *Journal of Clinical Oncology* **25**, June 20 suppl, LBA7516.
- [5] Gelber, R. D., Cole, B. F., Gelber, S., and Goldhirsch, A. (1995). Comparing treatments using quality-adjusted survival: The Q-TWiST method. *American Statistician* **49**, 161–169.
- [6] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning* **11**, 3–42.
- [7] Guez, A., Vincent, R., Avoli, M., and Pineau, J. (2008). Adaptive treatment of Epilepsy via batch-mode reinforcement learning. *Innovative Applications of Artificial Intelligence*.
- [8] Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24**, 1455–1481.
- [9] Murphy, S. A. (2005b). A generalization error for Q-learning. *Journal of Machine Learning Research* **6**, 1073-1097.
- [10] Ormoneit, D. and Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning* **49**, 161–178.
- [11] Pirker, R., Szczesna, A., Von Pawel, J., Krzakowski, M., Ramlau, R., Park, K., Gatzemeier, U., Bajeta, E., Emig, M., and Pereira, J. R. (2008). FLEX: A randomized, multicenter, phase III study of cetuximab in combination with cisplatin/vinorelbine (CV) versus CV alone in the first-line treatment of patients with advanced non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology* **26**, May 20 suppl, abstract 3.
- [12] Sandler, A., Gray, R., Perry, M. C., Brahmer, J., Schiller, J. H., Dowlati, A., Lilenbaum, R., and Johnson, D.H. (2006). Paclitaxel-Carboplatin alone or with bevacizumab for non-small-cell lung cancer. *The New England Journal of Medicine* **355**, 2542–2550.
- [13] Scagliotti, G.V., Parikh, P., Von Pawel, J., Biesma, B., Vansteenkiste, J., Manegold, C., Serwatowski, P., Gatzemeier, U., Digumarti, R., Zukin, M., Lee, J. S., Mellemegaard, A., Park, K., Patil, S., Rolski, J., Goksel, T., de Marinis, F., Simms, L., Sugarman, K. P., and Gandara, D. (2008). Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naïve patients with advanced-stage non-small-cell lung cancer. *Journal of Clinical Oncology* **26**, 3543-3551.

- [14] Schiller, J. H., Harrington, D., Belani, C. P., Langer, C., Sandler, A., Krook, J., Zhu, J., and Johnson, D. H. (2002). Comparison of four chemotherapy regimens for advanced non-small cell lung cancer. *New England Journal of Medicine* **346**, 92–98.
- [15] Shepherd, F. A., Dancey, J., Ramlau, R., Mattson, K., Gralla, R., O'Rourke, M., Levitan, N., Gressot, L., Vincent, M., Burkes, R., Coughlin, S., Kim, Y., and Berille, J. (2000). Prospective randomized trial of docetaxel versus best supportive care in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal of Clinical Oncology* **18**, 2095–2103.
- [16] Shepherd, F. A., Pereira, J. R., Ciuleanu, T., Tan, E. H., Hirsh, V., Thongprasert, S., Campos, D., Maoleekoonpiroj, S., Smylie, M., Martins, R., van Kooten, M., Dediu, M., Findlay, B., Tu, D., Johnston, D., Bezjak, A., Clark, G., Santabarbara, P., and Seymour, L. (2005). Erlotinib in previously treated non-small-cell lung cancer. *The New England Journal of Medicine* **353**, 123–132.
- [17] Shivaswamy, P., Chu, W., and Jansche, M. (2007). A Support Vector Approach to Censored Targets. *Proceedings of the International Conference on Data Mining*. Omaha, NE.
- [18] Socinski, M. A. and Stinchcombe, T. E. (2007). Duration of first-line chemotherapy in advanced non small-cell lung cancer: less is more in the era of effective subsequent therapies. *Journal of Clinical Oncology* **25**, 5155–5157.
- [19] Socinski, M. A., Crowell, R., Hensing, T. E., Langer, C. J., Lilenbaum, R., Sandler, A. B., and Morris D. (2007). Treatment of non-small cell lung cancer, stage IV. ACCP evidence-based clinical practice guidelines. *Chest* **132**, 3, supplement.
- [20] Stinchcombe, T. E. and Socinski, M. A. (2008). Considerations for second-line therapy of non-small cell lung cancer. *The Oncologist* **13**, 28–36.
- [21] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA.
- [22] Thall, P. F., Wooten, L. H., Logothetis, C., J., Millikan, R. E., and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine* **26**, 4687–4702.
- [23] Watkins, C. J. C. H. (1989). *Learning From Delayed Rewards*. Ph.D. Thesis, King's College, Cambridge, UK.
- [24] Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning* **8**, 279–292.

- [25] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer: New York.
- [26] Vapnik, V., Golowich, S., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems* **9**, 281–287.
- [27] Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine (Accepted)*.

