HHS Public Access

Author manuscript

Biometrika. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

Biometrika. 2015 March 1; 102(1): 151–168. doi:10.1093/biomet/asu050.

Doubly Robust Learning for Estimating Individualized Treatment with Censored Data

Y. Q. Zhao,
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, 53792, U.S.A

D. Zeng,
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, U.S.A

E. B. Laber,
Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695, U.S.A

R. Song,
Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695, U.S.A

M. Yuan, and
Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, 53792, U.S.A

M. R. Kosorok
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, U.S.A

Y. Q. Zhao: yqzhao@biostat.wisc.edu; D. Zeng: dzeng@bios.unc.edu; E. B. Laber: eblaber@ncsu.edu; R. Song: rsong@ncsu.edu; M. Yuan: myuan@stat.wisc.edu; M. R. Kosorok: kosorok@unc.edu

Summary

Individualized treatment rules recommend treatments based on individual patient characteristics in order to maximize clinical benefit. When the clinical outcome of interest is survival time, estimation is often complicated by censoring. We develop nonparametric methods for estimating an optimal individualized treatment rule in the presence of censored data. To adjust for censoring, we propose a doubly robust estimator which requires correct specification of either the censoring model or survival model, but not both; the method is shown to be Fisher consistent when either model is correct. Furthermore, we establish the convergence rate of the expected survival under the estimated optimal individualized treatment rule to the expected survival under the optimal individualized treatment rule. We illustrate the proposed methods using simulation study and data from a Phase III clinical trial on non-small cell lung cancer.

SUPPLEMENTARY MATERIALS

Supplementary material available at *Biometrika* online includes extended proofs of technical results, calculation of pseudo-outcome using Cox proportional hazards models, and additional simulation results.

Keywords

Censored data; Doubly robust estimator; Individualized treatment rule; Risk bound; Support vector machine

1. Introduction

Clinicians routinely tailor treatment to the individual characteristics of each patient. Individualized treatment rules formalize this practice by mapping patient characteristics to a recommended treatment. There is a large body work on estimation of optimal individualized treatment rules, using data from clinical trials or observational studies (Murphy, 2003; Robins, 2004; Zhao et al., 2009, 2011; Qian & Murphy, 2011; Zhang et al., 2012b). Regression-based approaches model the regression of outcome on patient covariates and treatment and infer the optimal individualized treatment rule from the modeled regression. The performance of regression-based methods depends critically on the predictive performance of the estimated regression model. In addition, because regression-based approaches require the modeling of treatment-covariate interactions, the number of terms can be large with high-dimensional covariates. An alternative class of procedures, known as classification-based methods, maximize an estimator of the marginal mean outcome over a pre-specified class of individualized treatment rules. These methods typically rely on fewer modeling assumptions about the conditional distribution of the outcome given covariates and treatment and so are potentially more robust to model misspecification; furthermore, they avoid inversion of a predictive model, which can be computationally expensive in some settings. Zhao et al. (2012) and Zhang et al. (2012a) showed that maximization of the estimated marginal mean outcome is equivalent to minimizing a weighted misclassification error with weights that are proportional to the observed clinical outcomes. Classification-based approaches have been shown to work well in settings without censoring (Zhao et al., 2012; Zhang et al., 2012a; Kang et al., 2014; Zhao et al., 2014). However, heretofore both regression-based and classification-based methods were restricted to use with non-censored data.

When the primary outcome of interest is survival time, the observations are commonly subject to right censoring because of subject dropout or administrative censoring. One approach is to fit a parametric or semiparametric survival model, including patient covariates and treatments to infer the optimal decision rule from the fitted survival model. Goldberg & Kosorok (2012) model the completely observed survival time and adjust for censoring by inverse probability of censoring weighting. These methods are intended to form high-quality predictions but may not be consistent for the optimal treatment rule (Qian & Murphy, 2011). Furthermore, parametric or semiparametric models can be sensitive to model misspecification and inverse-weighting may suffer from numerical instability when the censoring rate is high.

We extend the outcome weighted learning approach of Zhao et al. (2012) to accommodate censored data. The extension involves maximizing an estimator of the mean survival time under right censoring. The method avoids inversion of a predictive model by recasting the estimated mean survival time as a weighted misclassification rate where the weights involve

both the observed outcome and inverse probability of censoring weights. We also introduce a doubly robust version of outcome weighted learning to account for potential bias introduced by a misspecified censoring model. The method is doubly robust in the sense that the obtained individualized treatment rule is consistent for the optimal rule when the model for either survival or censoring times is correct, but not necessarily both. We use a convex relaxation idea from support vector machines (Cortes & Vapnik, 1995) to construct a computationally efficient algorithm.

2. Methodology

2·1. Value function and optimal treatment rule

Let T denote survival time, $X = (X_1,\ldots, X_d)^T \in \mathcal{X}$ denote subject covariates, and $A \in \{-1, 1\}$ denote the binary treatment assigned. Define τ to be the end of the study; because there is no information about survival beyond τ we use $T = \min(\tau, \tilde{T})$ as the outcome of interest. When we are interested in survival time on the log scale, we can use $\log(T)$ as the outcome. We assume that data are collected in a randomized trial so that treatment A is randomly assigned with a randomization probability that is completely determined by X. Thus, there are no unmeasured confounders (Rubin, 1974, 1978; Splawa-Neyman et al., 1990). Furthermore, we assume that $\pi(a; X) = \text{pr}(A = a \mid X)$ is strictly bounded away from zero with probability 1 for each a. A treatment rule, say $\mathcal{D}$, is a function from $\mathcal{X}$ into the space of treatments $\{-1, 1\}$; under $\mathcal{D}$, a patient with covariates $X = x$ is assigned treatment $\mathcal{D}(x)$. The value of a regime $\mathcal{D}$, denoted $V(\mathcal{D})$, is the expected outcome under $\mathcal{D}$. Let E denote expectation with respect to the distribution of (T, A, X) in the observed data, and $E^{\mathcal{D}}$ denote expectation under the restriction that $A = \mathcal{D}(X)$, then it can be shown (Qian and Murphy, 2011) that

$$V(\mathscr{D}) = E^{\mathscr{D}}(T) = E\{T \mid X, A = \mathscr{D}(X)\} = E\left[\frac{TI\{A = \mathscr{D}(X)\}}{\pi(A; X)}\right], \quad (1)$$

where $I(\cdot)$ is an indicator function. A treatment rule, say $\mathcal{D}^*$, is said to be optimal if $V(\mathcal{D}^*) \geq V(\mathcal{D})$ for all rules $\mathcal{D}$. To characterize $\mathcal{D}^*$, write the last term in (1) as $E\{[I\{\mathcal{D}(X) = 1\}E(T \mid A = 1, X) + I\{\mathcal{D}(X) = -1\}E(T \mid A = -1, X)]\}$ which implies

$$\mathscr{D}^*(x) = \text{sign}\{E(T \mid A = 1, X = x) - E(T \mid A = -1, X = x)\}. \quad (2)$$

Thus, $\mathcal{D}^*(x)$ is the maximizer of $E(T \mid X = x, A = a)$ with respect to a.

2·2. Outcome weighted learning with censored data

Censoring due to patient dropout is commonly seen in studies of survival time. Let C denote the potential censoring time, which could exceed τ, and assume that C and T are independent given (A, X). We observe data comprising n independent identically distributed subjects, $\{Y_i = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), X_i, A_i\}$, $i = 1, \ldots, n$, where $\Delta = I(T \leq C)$ denotes the censoring indicator. Our goal is to estimate the optimal treatment rule $\mathcal{D}^*$ using the censored data.

Maximizing $V(\mathcal{D})$ is equivalent to minimizing $E[TI\{A \neq \mathcal{D}(X)\}/\pi(A; X)]$ according to (1). This is a weighted classification problem, where misclassification corresponds to $A \neq \mathcal{D}(X)$, and the weights are $T/\pi(A; X)$. This point of view motivated the development of outcome weighted learning for noncensored outcomes (Zhao et al., 2012). We generalize this approach to censored outcomes. Hereafter, we assume that event times and censoring times are continuous. Let $S_C^*(t|A,X)=\mathrm{pr}(C>t|A=a,X=x)$ be the conditional treatment specific survival function for the censoring time given covariates x. Recall that $T = \min(T, \tilde{\tau})$. Then,

$$E\left\{\frac{\Delta Y}{S_C^*(Y|A,X)}|A,X\right\}=E\left\{\frac{\tilde{T}I(\tilde{T}<\tau)I(C>\tilde{T})}{S_C^*(\tilde{T}|A,X)}+\frac{\tau I(\tilde{T}\geq\tau)I(C>\tau)}{S_C^*(\tau|A,X)}|A,X\right\}$$
$$=E\{\tilde{T}I(\tilde{T}<\tau)+\tau I(\tilde{T}\geq\tau)|A,X\}=E(T|A,X),$$

where we have used the conditional independence of T and C given X, A. Therefore,

$$V(\mathscr{D})=E\left[\frac{E(T|A,X)}{\pi(A;X)}I\{A=\mathscr{D}(X)\}\right]=E\left[\frac{\Delta Y}{S_C^*(Y|A,X)\pi(A;X)}I\{A=\mathscr{D}(X)\}\right]. \quad (3)$$

To obtain an estimator of $\mathcal{D}^*$ one could attempt to maximize an empirical estimate of the right-hand-side of (3). This is equivalent to minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{\Delta_i Y_i}{\hat{S}_C(Y_i|A_i,X_i)}\frac{I\{A_i\neq\mathscr{D}(X_i)\}}{\pi(A_i;X_i)}, \quad (4)$$

where $\hat{S}_C$ is a consistent estimator for S_C^*. However, direct optimization is intractable because of the discontinuous indicator functions; instead, we minimize a convex relaxation of (4). Because the objective function can be viewed as a weighted misclassification rate, we base our relaxation on support vector machines (Cortes & Vapnik, 1995). We replace $I\{A \neq \mathcal{D}(x)\}$ with a convex surrogate $\varphi\{Af(x)\}$, where $\mathcal{D}(x) = \mathrm{sign}\{f(x)\}$ and $\varphi(t) = \max(1 - t, 0)$ denotes the hinge loss. Details for estimating $\mathcal{D}^*$ are at the end of this section.

In the above formulation, a misspecified model for C given (A, X) can lead to biased estimation. Thus, we also propose a doubly robust estimator which protects against such misspecification. Let $E_{\tilde{T}}^m(T|T>t,A,X)$ denote a working model for the conditional mean residual life-time given (A, X) derived from a working survival model for $S_T(t \mid A, X)$, and let $S_C^m(t|A,X)$ denote a working model for $S_C(t \mid A, X)$. Then, using the construction given in Section 2.3.2 of van der Laan & Robins (2003), we define the augmented value function,

$$V^m(\mathscr{D})=E\left(\left[\frac{\Delta Y}{S_C^m(Y|A,X)}-\int E_{\tilde{T}}^m(T|T>t,A,X)\left\{\frac{dN_C(t)}{S_C^m(t|A,X)}+I(Y\geq t)\frac{dS_C^m(t|A,X)}{S_C^m(t|A,X)^2}\right\}\right]\frac{I\{A=\mathscr{D}(X)\}}{\pi(A;X)}\right),$$

where $N_C(t) = (1 - \Delta)I(Y \leq t)$. In addition to the inverse probability of censoring weighting, there is an augmentation term in the weights for $I\{A = \mathcal{D}(X)\}$. The following lemma shows

that $V^m(\mathcal{D})$ is equivalent to $V(\mathcal{D})$ when either working model is correct; the proof is deferred to the Supplementary Material.

Lemma 1—If either $E^m_{\tilde{T}}(T|T>t, A, X)=E(T|T>t, A, X)$ or $S^m_C(t|A, X)=S^*_C(t|A, X)$, then $V^m(\mathcal{D}) = V(\mathcal{D})$.

Define

$$R(Y,\Delta,S_C,E_{\tilde{T}})=\frac{\Delta Y}{S_C(Y|A,X)}-\int E_{\tilde{T}}(T|T>t,A,X)\left\{\frac{dN_C(t)}{S_C(t|A,X)}+I(Y\geq t)\frac{dS_C(t|A,X)}{S_C(t|A,X)^2}\right\}, \quad (5)$$

Lemma 1 shows that if either working model is correct then $V(\mathcal{D})=E[R(Y,\Delta,S^m_C,E^m_{\tilde{T}})I\{A=\mathcal{D}(X)\}/\pi(A;X)]$. Thus, we can apply the weighted classification approach to estimate the optimal treatment rule using weights $R(Y,\Delta,S^m_C,E^m_{\tilde{T}})/\pi(A;X)$.

To distinguish the two learning approaches, we call the first approach inverse censoring weighted outcome weighted learning and the second approach doubly robust outcome weighted learning. Estimation is implemented as follows:

Step 1. Fit a model for T given (A, X) to construct estimate $\hat{S}_T(T \mid A, X)$ of $S_T(T \mid, A, X)$. Estimate $E_T(T \mid T > t, A, X)$ for $t \in [0, \tau)$ by

$$\hat{E}_{\tilde{T}}(T|T>t,A,X)=\frac{\tau\hat{S}_{\tilde{T}}(\tau|A,X)}{\hat{S}_{\tilde{T}}(T|A,X)}-\int_t^{\tau-}ud\hat{S}_{\tilde{T}}(u|A,X).$$

Step 2. Fit a model for C given (A, X) to form estimate $\hat{S}_C(t \mid A, X)$ of $S_C(t \mid A, X)$.

Step 3. Calculate $W_i = \Delta_i Y_i / \hat{S}_C(Y_i \mid A_i, X_i)$ for the first approach and $W_i = R(Y_i, \Delta_i, \hat{S}_C, \hat{E}_T)$ for the second approach. If negative weights occur with the doubly robust methods we can subtract $\min_i W_i$ from all the weights.

Step 4. Use the algorithm outlined below to obtain $\hat{f}(x)$ by minimizing

$$\sum_{i=1}^{n} W_i\frac{\phi\{A_i f(X_i)\}}{\pi(A_i;X_i)}+\lambda_n\|f\|^2. \quad (6)$$

Step 5. The decision rule is $\hat{\mathcal{D}}(x) = \text{sign}\{\hat{f}(x)\}$.

We have added a regularization term $\lambda_n\|f\|^2$ to avoid overfitting in Step 4. Here, $\|f\|$ is a norm defined on the space that f belongs to, and λ_n is a tuning parameter that controls the severity of penalization. We use a linear decision function $f(x) = \theta_0 + \theta^T x$, to illustrate the algorithm utilized in this step. In this case, $\|f\|$ is the Euclidean norm of θ. Let W denote a generic weight constructed in Step 3 using one of the proposed methods. The optimization problem in Step 4 can be written as $\min \gamma\sum_{i=1}^{n} W_i\xi_i+\|\theta\|^2$ subject to $\xi \geq 0$ and $A_i(\theta^T X_i + \theta_0) \geq (1-\xi_i)$. By introducing Lagrange multipliers, we obtain the Lagrangian

$$\frac{1}{2}\|\theta\|^2+\gamma\sum_{i=1}^{n}W_i\xi_i-\sum_{i=1}^{n}\alpha_i\{A_i(\theta^TX_i+\theta_0)-(1-\xi_i)\}-\sum_{i=1}^{n}\mu_i\xi_i,\quad \alpha_i,\mu_i\geq 0.$$

By taking derivatives with respect to θ, θ_0 and ξ_i, we have $0=\sum_{i=1}^{n}\alpha_iA_i, \theta=\sum_{i=1}^{n}\alpha_iA_iX_i$ and $\alpha_i = \gamma W_i - \mu_i$. It follows that the dual problem is

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jA_iA_jX_i^TX_j \quad (7)$$

subject to $0 \leq \alpha_i \leq \gamma W_i$ and $\sum_{i=1}^{n}\alpha_iA_i=0$. The dual problem can be solved using quadratic programming. Estimates $\hat{\theta}=\sum_{\hat{\alpha}_i\neq 0}\hat{\alpha}_i\hat{A}_iX_i$ and $\hat{\theta}_0$ follow from the Karush–Kuhn–Tucker conditions. When a linear decision rule is not sufficient, the procedure can be generalized using nonlinear kernel functions. For every positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is a unique reproducing kernel Hilbert space $\mathcal{H}_k$, which is the completion of the linear span of all functions $\{k(\cdot, x), x \in \mathcal{X}\}$. The norm in $\mathcal{H}_k$, denoted by $\|\cdot\|_k$, is induced by the inner product, $\langle f,g\rangle_k=\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_jk(x_i,x_j)$, for $f(\cdot)=\sum_{i=1}^{n}\alpha_ik(\cdot,x_i)$ and $g(\cdot)=\sum_{j=1}^{m}\beta_jk(\cdot,x_j)$. A general nonlinear function $f(x)$ can be used instead of a linear function. By the representer theorem (Kimeldorf & Wahba, 1971), the minimizer must admit a representation of the form $f(\cdot)=\sum_{i=1}^{n}\alpha_ik(\cdot,X_i)$. In addition, to solve the optimization problem, we only need to compute the kernel matrix, where the inner product $X_i^TX_j$ in the dual objective function (7) is replaced by $k(X_i, X_j)$. Quadratic programming can still be applied to obtain $\hat{\alpha}_i$ ($i = 1,\ldots, n$). The resulting decision rule is $\hat{\mathcal{D}}(X) = \text{sign}\{\sum_{\hat{\alpha}_i\neq 0}\hat{\alpha}_i\hat{k}(X_i, X) + \hat{\theta}_0\}$.

2·3. Working models for estimating $S_C(t \mid A, X)$ and $E_{\tilde{T}}(T \mid T > t, A, X)$

In our simulated experiments we used the Cox proportional hazards model for the requisite survival functions (Cox, 1972). Let Z_T and Z_C denote regressors constructed from X and A used in the Cox proportional hazards models for T and C respectively. Let $\lambda_{Ci}(t)$ and $\lambda_{Ti}(t)$ denote the hazard functions of censoring and failure times for subject i respectively. Under the Cox model, $\lambda_{Ci}(t)=\lambda_{C0}(t)\exp(\beta_C^TZ_{C_i})$ and $\lambda_{Ti}(t)=\lambda_{T0}(t)\exp(\beta_T^TZ_{Ti})$, where $\lambda_{C0}(t)$ and $\lambda_{T0}(t)$ are baseline hazard functions for censoring and failure times, respectively. The estimator for β_C, say $\hat{\beta}_C$, maximizes the partial likelihood

$$\prod_{i=1}^{n}\left\{\frac{\exp(\beta_C^TZ_{Ci})}{\sum_{Y_j\leq Y_i}\exp(\beta_C^TZ_{C_j})}\right\}^{1-\Delta_i}.$$

We use the Breslow estimator

$$\hat{\Lambda}_{C0}(t)=\int_0^t \frac{\sum_{k=1}^n dN_{Ck}(u)}{\sum_{j=1}^n I(Y_j \geq u)e^{\hat{\beta}_C^T Z_{Cj}}}$$

for the cumulative baseline hazard function $\Lambda_{C0}(t)$. An estimator of $\Lambda_C(t|A_i, X_i)$, the cumulative hazard function of censoring time for subject i, is $\exp(\hat{\beta}_C^T Z_{Ci})\hat{\Lambda}_{C0}(t)$. An estimator for $S_C(t|A_i, X_i)$ is $\hat{S}_C(t|A_i, X_i)=\exp\{-\hat{\Lambda}_{C0}(t)\}^{\exp(\hat{\beta}_C^T Z_{Ci})}$. Estimates $\hat{\beta_T}$ and $\hat{\Lambda_{T0}}(t)$ are obtained similarly. Details for estimating $E_T(T|T > t, A, X)$ are in the Supplementary Material.

3. Theoretical Results

Let $f(x)$ be the decision function with the decision rule given by $\mathcal{D}(x) = \text{sign}\{f(x)\}$, and write $V(f)$ to denote the value function $V(\mathcal{D})$. We define the pseudo value as

$$V_R(f, S_C, E_{\tilde{T}})=E\left(\frac{R(Y, \Delta, S_C, E_{\tilde{T}})I[A=\text{sign}\{f(X)\}]}{\pi(A;X)}\right).$$

Therefore, Lemma 1 can be restated as $V(f)=V_R(f, S_C^*, E_{\tilde{T}})=V_R(f, S_C, E_{\tilde{T}}^*)$, where $S_C^*(t|A, X)$ is the true conditional survival function of C given (A, X), and $E_{\tilde{T}}^*(T|T>t, A, X)$ is the true conditional mean residual lifetime given (A, X). Define convex surrogate loss function

$$L_\phi(f, S_C, E_{\tilde{T}})=\frac{R(Y, \Delta, S_C, E_{\tilde{T}})\phi\{Af(X)\}}{\pi(A;X)}, \quad (8)$$

where $\phi(t) = \max(1 - t, 0)$. Define $\mathcal{F}$ to be the set of all measurable functions from $\mathcal{X}$ into $\mathbb{R}$. Our first result states that the decision function obtained by minimizing the expectation of this surrogate loss over $\mathcal{F}$ maximizes V_R for any S_C and E_T. Furthermore, we quantify the differences using the hinge loss versus zero-one loss. The proofs are essentially the same as Theorems 3.1 and 3.2 in Zhao et al. (2012) and are thus omitted.

Lemma 2—If $\tilde{f}$ minimizes $E\{L_\phi(f, S_C, E_T)\}$ over $\mathcal{F}$ with any models for S_C and E_T, then

- **a.** $V_R(\tilde{f}, S_C, E_T) = \max_{f\in\mathcal{F}} V_R(f, S_C, E_T)$, that is, $\tilde{f}$ yields the maximum value of V_R;
- **b.** for any $f \in \mathcal{F}$,

$$V_R(\tilde{f}, S_C, E_{\tilde{T}})-V_R(f, S_C, E_{\tilde{T}}) \leq E\{L_\phi(f, S_C, E_{\tilde{T}})\}-E\{L_\phi(\tilde{f}, S_C, E_{\tilde{T}})\},$$

that is, the value lost due to using a suboptimal decision function f is bounded by the expected surrogate loss.

Our main result establishes the convergence rates for value of the estimated decision rule $\hat{f}$. As described in Section 2, we use weights that depend on the estimated survival functions, and hinge loss as a surrogate loss function. To bound the difference between the true and the empirical expectation of the surrogate, which involves random quantities, that is, estimates for survival and censoring times, we use the following assumptions:

Assumption 1: Both $\hat{E}_T(T|A = a, X = x)$ and $\hat{S}_C(t|A = a, X = x)$ converge in probability to $E^m_{\tilde{T}}(t|A, X)$ and $S^m_C(t|A, X)$ uniformly in $t \in (0, \tau]$ for every (x, a). Moreover, for some constant $\gamma > 0$,

$$E\{|R(Y, \Delta, \hat{S}_C, \hat{E}_{\tilde{T}}) - R(Y, \Delta, S^m_C, E^m_{\tilde{T}})|\} = O_p(n^{-\gamma}). \quad (9)$$

Assumption 2: For some $\eta > 0$, $S^m_C(\tau|A, X) > \eta$ with probability 1.

Assumption 1 implies that $\hat{E}_T(T|A, X)$ and $\hat{S}_C(t|A, X)$ converge to fixed functions, even if the imposed working models are wrong. Moreover, it imposes an assumption on the variance of the survival function estimators. The constant γ depends on the working models used for estimating S_C and E_T. If we assume parametric or semiparametric models, including the Cox proportional hazards model and transformation models, then $\gamma = 1/2$ in (9). Assumption 2 ensures that some subjects do not fail at the end of the study and thus have observation time τ.

In addition, we restrict the choice of reproducing kernel Hilbert space to the space associated with Gaussian radial basis function kernels, $k(x, x') = \exp(-\sigma_n^2 \|x - x'\|^2)$, $x, x' \in \mathcal{X}$, where $\sigma_n > 0$ is the kernel bandwidth parameter varying with n controlling the spread of the kernel. We can determine the complexity of $\mathcal{H}_k$ in terms of capacity bounds with respect to the empirical L_2-norm, defined as $\|f - g\|_{L_2(P_n)} = (n^{-1} \sum_{i=1}^n |f(X_i) - g(X_i)|^2)^{1/2}$, $f, g \in \mathcal{F}$ for functional class $\mathcal{F}$. For any $\varepsilon > 0$, the covering number of $\mathcal{F}$ with respect to $L_2(P_n)$, $N\{\mathcal{F}, \varepsilon, L_2(P_n)\}$, is the smallest number of $L_2(P_n)$ ε-balls needed to cover $\mathcal{F}$, where an $L_2(P_n)$ ε-ball around a function $g \in \mathcal{F}$ is the set $\{f \in \mathcal{F} : \|f - g\|_{L_2(P_n)} < \varepsilon\}$. It has been shown in Theorem 2.1 of Steinwart & Scovel (2007) that for any $\varepsilon > 0$,

$$\sup_{P_n} \log N\{B_{\mathcal{H}_k}, \varepsilon, L_2(P_n)\} \leq c_n \varepsilon^{-p}, \quad (10)$$

where $c_n = c_{p,\delta,d} \sigma_n^{(1-p/2)(1+\delta)d}$, $B_{\mathcal{H}_k}$ is the closed unit ball of $\mathcal{H}_k$, p and δ are any numbers satisfying $0 < p \leq 2$ and $\delta > 0$, $c_{p,\delta,d}$ is a constant depending only on p, δ and d, and the supremum is taken over finitely discrete probability measures P_n.

Let $f^m = \operatorname{argmin}_{f \in \mathcal{F}} E\{L_\phi(f, S^m_C, E^m_T)\}$. According to Lemma 2, f^m also maximizes $V_R(f, S^m_C, E^m_T)$, and $|V_R(f^m, S^m_C, E^m_T - V_R(f, S^m_C, E^m_T))| \leq |E\{L_\phi(f, S^m_C, E^m_T)\} - E\{L_\phi(f^m, S^m_C, E^m_T)\}|$ for any

function *f*. Hence, the convergence rate of the value using the estimated rule under the hinge loss will dominate the rate under the 0–1 loss. Define the approximation error function

$$a(\lambda)=\inf_{f\in\mathcal{H}_k}[E\{L_\phi(f,S_C^m,E_{\tilde{T}}^m)\}+\lambda\|f\|_k^2-E\{L_\phi(f^m,S_C^m,E_{\tilde{T}}^m)\}].$$

The following theorem bounds the excess value optimal treatment rule relative to the doubly robust estimator of the optimal treatment rule. Its proof can be found in the Appendix.

Theorem 1—Assume Assumptions 1 and 2 hold and that $\lambda_n \to 0$ and $\lambda_n n^{\min(2\gamma,1)} \to \infty$ as $n \to \infty$. If we estimate $\hat{f}$ within a reproducing kernel Hilbert space $\mathcal{H}_k$ associated with Gaussian radial basis function kernels, then with probability greater than $1 - 2e^{-b}$,

$$\begin{aligned} V(f^*)-V(\hat{f}) \le 2\sup_{f\in\mathcal{F}}|V_R(f,S_C^*,E_{\tilde{T}}^*)-V_R(f,S_C^m,E_{\tilde{T}}^m)|+a(\lambda_n) \\ +M_p{c_n}^{2/(p+2)}(n\lambda_n)^{-2/(p+2)}+M_p\lambda_n^{-1/2}{c_n}^{2/(p+2)}n^{-2/(p+2)} \\ +Kb(n\lambda_n)^{-1}+2Kbn^{-1}\lambda_n^{-1/2}+O_p(n^{-\gamma}\lambda_n^{-1/2}), \end{aligned} \tag{11}$$

where M_p is a constant depending on p and K is a sufficiently large positive constant.

In this theorem, $f^*(x) = E(T|A = 1, X = x) - E(T|A = -1, X = x)$ gives the optimal treatment rule. On the right-hand-side of (11), the first term reflects the estimation bias from the working models for T and C. The second term is the approximation error due to using the $\mathcal{H}_k$ space. The last term is the stochastic variability of estimating S_C^m and $E_{\tilde{T}}^m$. The rest of the terms contain the empirical loss function for estimation of the optimal treatment rule. In particular, the convergence rate γ depends on the estimating procedure applied to the two working models.

A corollary is that when either the model for survival time or the model for censoring time is correctly specified, with probability greater than $1 - 2e^{-b}$, we have

$$\begin{aligned} V(f^*)-V(\hat{f}) \le\ & a(\lambda_n) \\ & +M_p{c_n}^{2/(p+2)}(n\lambda_n)^{-2/(p+2)} \\ & +M_p\lambda_n^{-1/2}{c_n}^{2/(p+2)}n^{-2/(p+2)} \\ & +Kb(n\lambda_n)^{-1} \\ & +2Kbn^{-1}\lambda_n^{-1/2} \\ & +O_p(n^{-\gamma}\lambda_n^{-1/2}). \end{aligned}$$

Remark 1—With the hinge loss, it has been shown that if the reproducing kernel Hilbert space is rich enough, the optimizer within the reproducing kernel Hilbert space approaches the optimal treatment decision rule as the sample size goes to ∞ for appropriately chosen tuning parameters. The Gaussian kernel is one such kernel, which can induce a reproducing kernel Hilbert space that is flexible enough to approximate the optimal decision rule. While

the approximation error term usually goes to zero as $\lambda_n \to 0$, the other term controlling the stochastic error will increase. The optimal bandwidth λ_n can be obtained by setting the orders of $a(\lambda_n)$ and $\max\{c_n^{2/(p+2)}(n\lambda_n)^{-2/(p+2)}, n^{-\gamma}\lambda_n^{-1/2}\}$ equal to each other. The approximation error function $a(\lambda_n)$ is usually related to the data-generating distribution, especially the behavior close to the decision boundary, which is the true optimal decision rule if either working model is correct. Intuitively, we should be able to learn the treatment rule more rapidly for well-separated optimal treatment classes, that is, distributions that have low density near the boundary.

This behavior can be characterized in terms of the size of the set of points that are close to boundary $f^*(x) = 0$ (Tsybakov, 2004; Steinwart & Scovel, 2007). There exist a constant c_1 such that $a(\lambda_n) \leq c_1\lambda_n^{q/q+1}$, when using a Gaussian kernel with its kernel band-width σ_n varying with λ_n as $\sigma_n = \lambda_n^{-1/\{(q+1)d\}}$ and $c_n = \lambda_n^{-(1-p/2)(1+\delta)/(q+1)}$. Here, $q > 0$ is the noise exponent that characterizes the distribution close to the boundary (Steinwart & Scovel, 2007), and a larger q indicates a better separation between two treatment classes. More details on q are provided in the Appendix. An optimal choice of λ_n that balances bias and variance is $\lambda_n = \max\{n^{-2(q+1)/\{(4+p)q+2+(2-p)(2+\delta)\}}, n^{-2(q+1)\gamma/(2q+1)}\}$. The achieved convergence rate of the value due to the estimated rule versus the optimal value is thus $\max[n^{-2q/\{(4+p)q+2+(2-p)(2+\delta)\}}, n^{-2q\gamma/(2q+1)}]$.

The rate consists of two parts: the first part reflects the rate of convergence in estimating the optimal decision rule, which is consistent with the results without censoring (Zhao et al., 2012); the second part is related to survival function estimation. When q is sufficiently large and δ and p are close to zero, the convergence rate is close to $n^{-\gamma}$, where γ is determined by the survival function estimator. A Cox model for the survival function estimates leads to $\gamma = 1/2$. Other working models can also be applied, such as transformation models (Zeng & Lin, 2007), nonparametric methods based on kernel type estimators (Dabrowska, 1989), or machine learning techniques (Zhu & Kosorok, 2012). However, the rate $n^{-\gamma}$ can be slower than $O_p(n^{-1/2})$ for certain estimators.

Remark 2—Although the theoretical results are derived only for doubly robust estimators, inverse censoring weighted estimators enjoy the property stated in Theorem 1, as it is a special case obtained by setting the augmentation term to zero. However, the first term on the right-hand-side of (11) will change unless the censoring model is correctly specified.

4. Simulation Studies

4·1. Preliminaries

We aim to maximize the value function in terms of the survival time on the log scale. We compare the inverse censoring weighted and the doubly robust methods for selecting optimal individualized treatment with Cox regression and Q-learning adjusted with censoring weights (Goldberg & Kosorok, 2012). For Cox regression, we fit a proportional hazards model with treatment-by-covariate interactions, and identify the optimal individualized treatment based on the predicted outcomes. To apply Q-learning, we fit $Q(X, A) = \Phi(X, A)\theta$, where $\Phi(X, A) = (1, X, A, XA)$, to the log of the failure time. We also apply a

regularized version of *Q*-learning, called L_2 Q-learning, where an L_2 penalty is used for regularization. The estimator is obtained as

$$\operatorname{argmin}_\theta \sum_{i=1}^{n} \{\log(Y_i) - \Phi(X_i, A_i)\theta\}^2 \frac{\Delta_i}{\hat{S}_C(Y_i|A_i, X_i)} + \lambda_n \|\theta\|^2,$$

where λ_n is a tuning parameter to be selected using cross-validation, and $\hat{S}_C(Y|A, X)$ is the estimated conditional survival function of C given (A, X) that can be obtained using Cox regression. The estimated optimal decision rule is $\hat{D}(x) = \operatorname{argmax}_{a\in\{-1,1\}} \Phi(x, a)\hat{\theta}$.

4·2. Simulation study

Ten independent covariates, $X_1, \ldots, X_{10}$, were generated from the uniform distribution on [0, 1]. Treatments were generated from $\{-1, 1\}$ with equal probabilities 0·5. Four different scenarios are presented, corresponding to different combinations of correct or incorrect survival time and censoring time models. Specifically, we generated $\tilde{T}$ or C from the accelerated failure time or Cox models in different scenarios, while we always used a proportional hazards model as a working model for both $\tilde{T}$ and C given (A, X). Regarding the specification of the model basis, we include treatment covariate interaction terms in the survival function modeling since we are interested in whether certain characteristics moderate treatment effects. Conversely, we do not model the censoring time with interaction terms unless we have full knowledge of the data, because it is not typical to posit a complex model for the censoring mechanism in practice. Details for calculating the doubly robust weights using Cox working models are given in the Supplementary Material.

For each scenario, a test data set of size 10,000 is generated to evaluate the estimated rules. The decision rules are estimated from training data using the proposed methods as described in Section 2 and the competitors. The sample sizes for the training data sets were taken to equal to 100, 200 and 400, and the simulations were repeated 1000 runs for each sample size. A linear basis is applied for model fitting in Q-learning. Linear kernels were used for both the implementation of inverse censoring weighted and doubly robust outcome weighted learning. We also explored the use of Gaussian kernels, and found that the performances were comparable to the linear kernel. The learning procedure was implemented using a Library for Support Vector Machines developed in Chang & Lin (2011). The tuning parameter λ_n in (6) was chosen using 5-fold cross validation over a pre-specified grid, with the criterion being the empirical pseudo value function. Specifically, for each tuning parameter, we partitioned the training data into 5 parts, each of which serves as the validation set once while the other 4 parts of the data are utilized for estimation. We sum up the empirical pseudo values calculated across the validation sets from the corresponding trained decision rules, and choose the optimal tuning parameter as the one maximizing the summed value.

In the following, we consider four generative models.

Case 1: The true models are Cox models for both $\tilde{T}$ and C. The survival time T is the minimum of $\tau = 1.5$ and $\tilde{T}$, where $\tilde{T}$ is generated with hazard rate function

$$\lambda_{\tilde{T}}(t|A,X)=\lambda_{\tilde{T}0}(t)\exp\{0\cdot 6X_1-0\cdot 8X_2+(0\cdot 6-0\cdot 4X_1-0\cdot 2X_2-0\cdot 4X_3)A\},$$

and $\lambda_{\tilde{T}0}(t) = 2t$. The censoring time C is generated with hazard rate function

$$\lambda_C(t|A,X)=\lambda_{C0}(t)\exp\{0\cdot 5X_1+0\cdot 5X_2+(-1+0\cdot 4X_1-0\cdot 6X_2)A\},$$

where $\lambda_{C0}(t) = 2t$. The censoring percentage is around 56%. The optimal decision boundary is linear with $\mathcal{D}^*(X) = -\text{sign}(0{\cdot}6 - 0{\cdot}4X_1 - 0{\cdot}2X_2 - 0{\cdot}4X_3)$. We use the Cox regression model with covariates (X_1, X_2, X_3, A, X_1A, X_2A, X_3A) to model survival and censoring times respectively. Therefore, both models are correctly specified.

Case 2: The true model for T is a Cox model, and the true model for C is an accelerated failure time model. The survival time T is the minimum of $\tau = 2$ and $\tilde{T}$, where

$$\lambda_{\tilde{T}}(t|A,X)=\lambda_{\tilde{T}0}(t)\exp\{-1\cdot 5X_1+0\cdot 5X_2+(1-0\cdot 7X_1-1\cdot 2X_2)A\}.$$

We let censoring time C be generated from an accelerated failure time model with

$$\log(C)=-0\cdot 3+0\cdot 6X_1-0\cdot 1X_2+0\cdot 3X_3+(0\cdot 2-X_1-2X_2+0\cdot 5X_3)A+\varepsilon,$$

where ε is generated from $N(0, 1)$. The censoring percentage is around 34%. The optimal decision boundary is $\mathcal{D}^*(X) = -\text{sign}(0{\cdot}5 - 0{\cdot}1X_1 - 0{\cdot}6X_2)$. We use the Cox model as the working model for both T and C given (A, X). Specifically, we use (X_1, X_2, A, X_1A, X_2A) as covariates to model survival time, and (X_1, X_2, X_3) to model censoring time. Therefore, the working model is correctly specified for T but incorrect for C.

Case 3: The true model for T is an accelerated failure time model, and the true model for C is a Cox model. The survival time T is the minimum of $\tau = 2$ and $\tilde{T}$, which is generated with

$$\log(\tilde{T})=-0\cdot 5-0\cdot 8X_1+0\cdot 7X_2+0\cdot 2X_3+(0\cdot 6-0\cdot 4X_1-0\cdot 1X_2-0\cdot 4X_3)A+\varepsilon.$$

The censoring time C is generated from the Cox proportional hazards model, where

$$\lambda_C(t|A,X)=\lambda_{C0}(t)\exp\{-0{\cdot}5X_1-0{\cdot}5X_2+0{\cdot}2X_3-(1-0{\cdot}5X_1+0{\cdot}3X_2-0{\cdot}5X_3)A\},$$

and $\lambda_{C0}(t) = 2t$. The censoring percentage is around 45%. The optimal decision boundary is linear with $\mathcal{D}^*(X) = \text{sign}(0{\cdot}6 - 0{\cdot}4X_1 - 0{\cdot}1X_2 - 0{\cdot}4X_3)$. We use the Cox model for both T and C given (A, X). Specifically, we use (X, A, XA) to model survival time, and (X_1, X_2, X_3, A,

X_1A, X_2A, X_3A) to model censoring time. Therefore, the working model is correctly specified for C but incorrect for T.

Case 4: The true models are accelerated failure time models for both T and C. The survival time T is the minimum of $\tau = 2.5$ and $\tilde{T}$, where

$$\log(\tilde{T}) = -0{\cdot}2 - 0{\cdot}5X_1 + 0{\cdot}5X_2 + 0{\cdot}3X_3 + (0{\cdot}5 - 0{\cdot}1X_1 - 0{\cdot}6X_2 + 0{\cdot}1X_3)A + \varepsilon.$$

The censoring time C is generated from an accelerated failure time model with

$$\log(C^*) = 0{\cdot}5 - 0{\cdot}8X_1 + 0{\cdot}4X_2 + 0{\cdot}4X_3 + (0{\cdot}5 - 0{\cdot}1X_1 - 0{\cdot}6X_2 + 0{\cdot}3X_3)A + \varepsilon,$$

and ε is generated from $N(0, 1)$. The censoring percentage is around 31%. The optimal decision boundary is linear with $\mathcal{D}^*(X) = \text{sign}(0{\cdot}5 - 0{\cdot}1X_1 - 0{\cdot}6X_2 + 0{\cdot}1X_3)$. We use ($X$, A, XA) to model survival time, and X to model censoring time. Therefore, both working models are incorrectly specified.

Since we know the true data generating mechanism under every scenario, for each of 1000 replicates, we calculate the values based on the logarithm of the survival time using the constructed rule from different methods. Figure 1 shows these values when $n = 100$, where larger values indicate longer survival. Additional results using other sample sizes are provided in the Supplementary Material.

In general, inverse censoring weighted and doubly robust outcome weighted learning have satisfactory performances. Inverse censoring weighted outcome weighted learning performs better when the censoring model is correctly specified, see Fig. 1(a) and 1(c). Indeed, doubly robust outcome weighted learning requires estimating both censoring and survival probabilities, which yield a higher variability compared with that of the inverse censoring weighted outcome weighted learning. The strength of doubly robust approach can be seen when the censoring model is mis-specified but the survival model is correct, since it can correct the bias from using only inverse censoring weighting, see Fig. 1(b). When the survival data are truly generated from the Cox model, Cox regression with correct basis results in the best performances, see Fig. 1(a) and 1(b). However, the strength is lost when the survival time is generated from an accelerated failure time model. Although Q-learning is improved by L_2-regularization, possibly by reducing overfitting, Q-learning based methods can have suboptimal performances even when the censoring model is correctly specified but survival time is generated from a Cox model, see Fig. 1(a). This is due to model misspecification when inverse censoring weighted Q-learning models the logarithm of the survival time.

We also consider a nonlinear example of possible model misspecification in the Supplementary Material. The number of covariates is increased to 30, and $\tilde{T}$ and C are generated from Cox models with complex effects. When the censoring or survival model is correctly specified, we use the true sets of covariates for model fitting. If the model is

incorrect for survival time or censoring time, we use a linear basis. In addition to a linear kernel, we apply both methods with a Gaussian kernel. We can see that the gain from using a Gaussian kernel is pronounced, since it may better approximate the nonlinear treatment decision rules.

5. Application

We illustrate the proposed methods using advanced non-small-cell lung cancer data (Socinski et al., 2002), which is collected in a two-arm randomized trial with survival time as the primary endpoint. Non-small-cell lung cancer is the leading cause of cancer-related mortality, and approximately 30% to 40% of all new cases present with stage IV or stage IIIB disease. To investigate the optimal duration of therapy that maximizes survival, a prospective randomized phase III trial was initiated in 1998. Patients with advanced non-small-cell lung cancer were recruited and randomized to either four cycles of carboplatin/paclitaxel or continuous therapy with carboplatin/pactaxel until disease progression. The study enrolled 230 subjects; however, we restrict our analysis to the 224 subjects with complete information. In the analysis sample, 112 subjects were assigned to each treatment. The censoring rate was 32%. The baseline covariates include age ranging from 32 to 82 with median 63, sex with 138 male and 86 female, race with 162 white, 54 black and 8 other, performance status with 117 Karnofsky performance status 90% to 100% and 97 70% to 80%, and stage with 30 Stage IIIB and 194 Stage IV.

In addition to the proposed methods, we apply Cox regression, inverse censoring weighted Q-learning and L_2 Q-learning with a linear basis. We consider two sets of working models: we first use Cox regression with basis (X, A, XA) for both survival time and censoring time, and then use the Kaplan–Meier estimator for censoring time as an alternative. A treatment decision rule $\mathcal{D}$ is evaluated based on its empirical value adjusted for censoring. The empirical value is calculated by $\sum_{i=1}^{n} \tilde{R}_i I\{A_i = \mathcal{D}(X_i)\} / \sum_{i=1}^{n} I\{A_i = \mathcal{D}(X_i)\}$, with $\tilde{R}$ equal to

$$\frac{\Delta \log(Y)}{\hat{S}_C(Y|A,X)} - \int \hat{E}_{\tilde{T}}\{\log(T)|T>t,A,X\} \left\{ \frac{dN_C(t)}{\hat{S}_C(t|A,X)} + I(Y_i \geq t) \frac{d\hat{S}_C(t|A,X)}{\hat{S}_C(t|A,X)^2} \right\},$$

where $\hat{S}_C(t \mid A, X)$ and $\hat{E}_T(T \mid T > t, A, X)$ are the estimated censoring probability and residual life conditional on patients characteristics. To avoid overfitting, we employ a cross-validated analysis. At each run, we partition the whole data set into 5 pieces, where 4 parts of the data are used as training data to estimate the individualized treatment rules, and the remaining part is the validation set for implementing the estimated rules, with empirical values stored for each method respectively. The cross-validated values are obtained by averaging the empirical values on all 5 validation subsets. The procedure is repeated 100 times. The averages and standard errors of these values are reported in Table 1, where larger values correspond to longer survival time.

Both inverse censoring weighted and doubly robust outcome weighted learning methods lead to higher values more frequently. We see a comparable performance between the two

proposed approaches, which is reasonable if the censoring distribution is correctly specified, although doubly robust outcome weighted learning may have a larger variance. Since the number of covariates is not large, the performances of inverse censoring weighted Q-learning and L_2 Q-learning are similar. They could have difficulties in identifying the optimal treatment rules if the model for survival time does not fit well, even if the censoring weight is correctly specified. Also, when we apply Cox regression to model the censoring time, none of the covariates has a significant effect. Thus, working models with either Kaplan–Meier estimators or Cox regression estimators yield similar results. We then apply the proposed methods to the whole data set using Cox regression working models for both T and C. The treatment recommendations from inverse censoring weighted outcome weighted learning recommends that 119 patients be assigned to continuous therapy with carboplatin/pactaxel, while by using doubly robust outcome weighted learning, 122 out of 224 patients should be given the continuous therapy. By checking the empirical value, we find that both strategies yield close values: 5·301 for inverse censoring weighted outcome weighted learning and 5·567 for doubly robust outcome weighted learning. In fact, sometimes we may have equivalent treatment strategies if there are no differential treatment effects on a subset of the patients. The treatment decision rule produced by inverse censoring weighted Q-learning and L_2 Q-learning however, only leads to an empirical value of 4·756, and an empirical value of 4·744 by using Cox regression.

6. Discussion

As one reviewer pointed out, the proposed method is only one possible reduction of optimizing treatment rules to a weighted classification problem. Alternative choices have been proposed for continuous outcomes (Zhang et al., 2012b; Rubin & van der Lann, 2012), which can be generalized to the censoring data setup.

There may exist multiple treatments for comparison. One approach for extending the proposed framework to handle this case is to incorporate techniques developed in multicategory classification (Lee et al., 2004; Liu & Yuan, 2011). Another important extension is to settings in which there are a large number of variables. In this setting, penalized methods in classification by using sparse penalties could be adapted (Zhu et al., 2004).

Effective management of a chronic illness requires individualized treatment recommendations that are responsive to a patient's changing health status. Dynamic treatment regimens formalize a dynamic treatment plan as a sequence of treatment rules, one per stage of clinical intervention, that map current patient information to a recommended treatment. Longer life expectancy and an aging population have created a surge in the rate of chronic illness related death. Thus, there is increasing interest in dynamic treatment regimens (Thall et al., 2002; Murphy, 2003; Robins, 2004; Moodie et al., 2007; Zhao et al., 2011; Goldberg & Kosorok, 2012; Huang et al., 2014). Extension of the proposed approach for survival data to the dynamic setting is of great interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the U.S. National Institutes of Health. We are grateful to the editors and the reviewers for their insightful comments which have led to important improvements in this paper.

References

Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. J Am Statist Assoc. 2006; 101:138–156.

Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995; 20:273–297.

Cox DR. Regression models and life-tables (with discussion). J R Statist Soc B. 1972; 34:187–220.

Dabrowska DM. Uniform consistency of the kernel conditional Kaplan–Meier estimate. Ann Statist. 1989; 17:1157–1167.

Goldberg Y, Kosorok MR. Q-learning with censored data. Ann Statist. 2012; 40:529–560.

Huang X, Ning J, Wahed AS. Optimization of individualized dynamic treatment regimes for recurrent diseases. Statist Med. 2014 In press.

Kang C, Janes H, Huang Y. Combining biomarkers to optimize patient treatment recommendations. Biometrics. 2014 In Press.

Kimeldorf G, Wahba G. Some results on Tchebycheffian spline functions. J Math Anal Appl. 1971; 33:82–95.

Lee Y, Lin Y, Wahba G. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. J Am Statist Assoc. 2004; 99:67–81.

Liu Y, Yuan M. Reinforced multicategory support vector machines. J Comput Graph Statist. 2011; 20:909–919.

Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007; 63:447–455. [PubMed: 17688497]

Murphy SA. Optimal dynamic treatment regimes (with discussion). J R Statist Soc B. 2003; 65:331–366.

Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Statist. 2011; 39:1180–1210.

Robins, JM. Optimal structural nested models for optimal sequential decisions. In: Lin, DY.; Heagerty, PJ., editors. Proc 2nd Seattle Symp Biostatist. New York: Springer; 2004.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974; 66:688–701.

Rubin DB. Bayesian inference for causal effects: The role of randomization. Ann Statist. 1978; 6:34–58.

Rubin DB, van der Lann MJ. Statistical issues and limitations in personalized medicine research with clinical trials. J Educ Psychol. 2012; 8 Article 18.

Socinski MA, Schell MJ, Peterman A, Bakri K, Yates S, Gitten R, Unger P, Lee J, Lee JH, Tynan M, Moore M, Kies MS. Phase III trial comparing a defined duration of therapy versus continuous therapy followed by second-line therapy in advanced-stage IIIB/IV nonsmall-cell lung cancer. J Clin Oncol. 2002; 20:1335–1343. [PubMed: 11870177]

Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments (engl. transl by D M Dabrowska and T P Speed). Statist Sci. 1990; 5:465–472.

Steinwart I, Scovel C. Fast rates for support vector machines using Gaussian kernels. Ann Statist. 2007; 35:575–607.

Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. J Am Statist Assoc. 2002; 97:29–39.

Tsybakov AB. Optimal aggregation of classifiers in statistical learning. Ann Statist. 2004; 32:135–166.

van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. New York: Springer-Verlag; 2003.

Zeng D, Lin D. Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). J R Statist Soc B. 2007; 69:507–564.

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012a; 1:103–114. [PubMed: 23645940]

Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012b; 68:1010–1018. [PubMed: 22550953]

Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Statist Med. 2009; 28:3294–3315.

Zhao Y, Zeng D, Socinski MA, Kosorok MR. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics. 2011; 67:1422–1433. [PubMed: 21385164]

Zhao YQ, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Statist Assoc. 2014 In Press.

Zhao YQ, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Statist Assoc. 2012; 107:1106–1118.

Zhu J, Rosset S, Hastie TJ, Tibshirani RJ. 1-norm support vector machines. Adv Neural Inf Process Syst. 2004; 16:49–56.

Zhu R, Kosorok MR. Recursively imputed survival trees. J Am Statist Assoc. 2012; 107:331–340.

Appendix

Proof of Theorem 1

First, it can be established that

$$V(f^*)-V(\hat{f}) \leq \sup_{f\in\mathscr{F}} V_R(f,S_C^m,E_{\tilde{T}}^m)-V_R(\hat{f},S_C^m,E_{\tilde{T}}^m)+2\sup_{f\in\mathscr{F}}|V_R(f,S_C^*,E_{\tilde{T}}^*)-V_R(f,S_C^m,E_{\tilde{T}}^m)|.$$

According to Lemma 2(a), $V_R(f^m,S_C^m,E_{\tilde{T}}^m)=\sup_{f\in\mathscr{F}} V_R(f,S_C^m,E_{\tilde{T}}^m)$, where $f^m=\operatorname{argmin}_{f\in\mathscr{F}} E\{L_\phi(f,S_C^m,E_T^m)\}$. Hence, it suffices to derive the convergence rate of $V_R(f^m,S_C^m,E_{\tilde{T}}^m)-V_R(\hat{f},S_C^m,E_{\tilde{T}}^m)$.

Let $f_{\lambda_n}^m=\operatorname{argmin}_{f\in\mathscr{H}_k}[E\{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)\phi\{Af(X)\}/\pi(A;X)\}+\lambda_n\|f\|_k^2]$. Then,

$$
\begin{aligned}
&V_R(f^m, S_C^m, E_{\tilde{T}}^m) - V_R(\hat{f}, S_C^m, E_{\tilde{T}}^m) \\
&\quad \le a(\lambda_n) + \Bigg(n^{-1} \sum_{i=1}^{n} \Big[\lambda_n \|\hat{f}\|_k^2 \\
&+ \frac{R(Y_i,\Delta_i,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A_i \hat{f}(X_i)\}}{\pi(A_i;X_i)} - \lambda_n \|f_{\lambda_n}^m\|_k^2 - \frac{R(Y_i,\Delta_i,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A_i f_{\lambda_n}^m(X_i)\}}{\pi(A_i;X_i)} \Big] \\
&- E\left[\lambda_n \|\hat{f}\|_k^2 + \frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A\hat{f}(X)\}}{\pi(A;X)} - \lambda_n \|f_{\lambda_n}^m\|_k^2 - \frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A f_{\lambda_n}^m(X)\}}{\pi(A;X)} \right] \Bigg) \\
&\quad + E\left[\frac{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)\phi\{A\hat{f}(X)\}}{\pi(A;X)} - \frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A\hat{f}(X)\}}{\pi(A;X)} \right] \\
&\quad + E\left[\frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A f_{\lambda_n}^m(X)\}}{\pi(A;X)} - \frac{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)\phi\{A f_{\lambda_n}^m(X)\}}{\pi(A;X)} \right] \\
&\quad = a(\lambda_n) + (I) + (II) + (III).
\end{aligned}
$$

To bound (*II*) and (*III*), we consider the class of functions

$$
\begin{aligned}
\mathscr{B} = \{ & R\{Y,\Delta,S_C(\beta_C,\Lambda_{C0}),E_{\tilde{T}}(\beta_T,\Lambda_{T0})\} : \beta_C \in \mathbb{R}^d, \|\beta_C - \beta_C^m\| < \delta_0,\ \beta_T \in \mathbb{R}^d, \\
& \|\beta_T - \beta_T^m\| < \delta_0, \Lambda_{C0}, \Lambda_{T0} \text{ are bounded monotone functions in } [0,\tau], \\
& \sup_t |\Lambda_{C0}(t) - \Lambda_{C0}^m(t)| < \delta_0,\ \sup_t |\Lambda_{T0}(t) - \Lambda_{T0}^m(t)| < \delta_0 \},
\end{aligned}
$$

where δ_0 is a small constant, and $\beta_T^m, \beta_C^m, \Lambda_{C0}^m(t), \Lambda_{T0}^m(t)$ are the limits of $\hat{\beta_T}$, $\hat{\beta_C}$, $\hat{\Lambda_{C0}}$ and $\hat{\Lambda_{T0}}$ based on the Cox models. Then $|R\{Y, \Delta, S_C(\beta_C, \Lambda_{C0}), E_T(\beta_C, \Lambda_{C0})\}|/\pi(A, X)$ can be bounded from above by a constant, say M.

Trivial bounds for $\|\hat{f}\|_k$ and $\|f_{\lambda_n}^m\|_k$ are obtained as $\|\hat{f}\|_k \le M\lambda_n^{-1/2}$ and $\|f_{\lambda_n}^m\|_k \le M\lambda_n^{-1/2}$. For every $f \in M\lambda_n^{-1/2} B_{\mathscr{H}_k}$, $|(1-Af)^+| \le 1 + M\lambda_n^{-1/2} = B$. Thus,

$$
\left| E\left[\frac{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)\phi\{Af(X)\}}{\pi(A;X)} - \frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{Af(X)\}}{\pi(A;X)} \right] \right| = O_p(n^{-\gamma}\lambda_n^{-1/2}).
$$

We use empirical process theory to bound (*I*). Define the functional class

$$
\begin{aligned}
\mathscr{L} = \Bigg\{ & \lambda_n \|f\|_k^2 + \frac{R(Y,\Delta,S_C,E_{\tilde{T}})\phi\{Af(X)\}}{\pi(A;X)} - \frac{R(Y,\Delta,S_C,E_{\tilde{T}})\phi\{Af_{\lambda_n}^m(X)\}}{\pi(A;X)} - \lambda_n \|f_{\lambda_n}^m\|_k^2, \\
& f \in M\lambda_n^{-1/2} B_{\mathscr{H}_k}, R(Y,\Delta,S_C,E_{\tilde{T}}) \in \mathscr{B} \Bigg\},
\end{aligned}
$$

and $\mathscr{G} = \{E(l) - l : E(l) = \varepsilon, l \in \mathscr{L}\}$. Let $Z = \sup_{g\in\mathscr{G}} n^{-1} \sum_{i=1}^{n} g(X_i)$. Since $E(g) = 0$, $g \in \mathscr{G}$, it follows from Lemma S.1 in the Supplementary Material, by setting $\rho = 1$, that $\mathrm{pr}\{Z \ge 2E(Z) + \sigma(Kb)^{1/2} n^{-1/2} + 2KBbn^{-1}\} \le e^{-b}$, where $B = O(\lambda_n^{-1/2})$. Furthermore, $\sigma^2 \le c_n' \varepsilon$ following the arguments for proving Theorem 3.4 in Zhao et al. (2012), given that $E(l^2) \le c_n' E(l)$, where $c_n' = O(\lambda_n^{-1})$. In addition, for $f \in M\lambda_n^{-1/2} B_{\mathscr{H}_k}$,

$$E(Z)=E\left\{\sup_{g\in\mathcal{G}} n^{-1}\sum_{i=1}^{n} g(X_i)\right\}=E\left[\sup_{E(l^2)\le c_n'\varepsilon}\left|E\{l(X)\}-n^{-1}\sum_{i=1}^{n} l(X_i)\right|\right].$$

Since $|\beta_C-\beta_C^m|$ and $|\beta_T-\beta_T^m|$ are bounded by δ_0, they lie in a hypercube of $\mathbb{R}^{2d}$. Moreover, $\{\Lambda_{C0}:\sup_t|\Lambda_{C0}(t)-\Lambda_{C0}^m(t)|<\delta_0\}$ is a class of monotone functions, so is $\{\Lambda_{T0}:\sup_t|\Lambda_{T0}(t)-\Lambda_{T0}^m(t)|<\delta_0\}$. The function in $\mathcal{B}$ is Lipschitz continuous with respect to all these parameters and the Lipschitz constant is less than a constant W. There exists a constant K, depending on d, such that the bracketing number for $\mathcal{B}$ satisfies $N_{[\cdot]}\{\mathcal{B}, \varepsilon W, L_2(P)\}\le K(\delta_0/\varepsilon)^{2d+2}$. According to (10), $\sup_{P_n}\log N\{\mathcal{G}, \varepsilon, L_2(P_n)\}\le c_n\varepsilon^{-p}$, and therefore

$$E(Z)\le c_p M\lambda_n^{-\frac{1}{2}}\max\left\{(M^2\lambda_n c_n'\varepsilon)^{(2-p)/4}{c_n}^{1/2}n^{-1/2}, {c_n}^{2/(2+p)}n^{-2/(2+p)}\right\},$$

where c_p is a constant depending on p. See Proposition 5.5 in Steinwart & Scovel (2007) and references therein. Consequently,

$$\begin{aligned}\mathrm{pr}\Bigg(\left|n^{1/2}\left[n^{-1}\sum_{i=1}^{n} l(X_i)-E\{l(X)\}\right]\right|>\left(c_n'\varepsilon Kb\right)^{1/2}n^{-1/2}+2KBbn^{-1}\\ +2c_pM\lambda_n^{-1/2}\max\left\{(M^2\lambda_n c_n'\varepsilon)^{(2-p)/4}{c_n}^{1/2}n^{-1/2}, {c_n}^{2/(2+p)}n^{-2/(2+p)}\right\}\Bigg)\le e^{-b}.\end{aligned}$$

Given that $\mathcal{L}$ is convex, if $l\in\mathcal{L}$ satisfies $n^{-1}\sum_{i=1}^{n} l(X_i)\le\alpha\varepsilon$ and $E\{l(X)\}\ge\varepsilon$, there exists $l'\in\mathcal{L}$ such that $n^{-1}\sum_{i=1}^{n} l'(X_i)\le\alpha\varepsilon$ and $E\{l'(X)\}=\varepsilon$. Thus, with probability at least $1-e^{-b}$, every $l\in\mathcal{L}$ with $n^{-1}\sum_{i=1}^{n} l(X_i)\le\alpha\varepsilon$ satisfies $El\in\varepsilon$ (Bartlett et al., 2006; Steinwart & Scovel, 2007). Since

$$n^{-1}\sum_{i=1}^{n}\left[\lambda_n\|\hat{f}\|_k^2+\frac{R(Y_i,\Delta_i,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A_i\hat{f}(X_i)\}}{\pi(A_i;X_i)}-\lambda_n\|f_{\lambda_n}^m\|_k^2-\frac{R(Y_i,\Delta_i,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A_i f_{\lambda_n}^m(X_i)\}}{\pi(A_i;X_i)}\right]\le 0<\alpha\varepsilon,$$

with probability at least $1-e^{-b}$,

$$E\left[\lambda_n\|\hat{f}\|_k^2+\frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{A\hat{f}(X)\}}{\pi(A;X)}-\lambda_n\|f_{\lambda_n}^m\|_k^2-\frac{R(Y,\Delta,\hat{S}_C,\hat{E}_{\tilde{T}})\phi\{Af_{\lambda_n}^m(X)\}}{\pi(A;X)}\right]\le\varepsilon.$$

It follows that,

$$\mathrm{pr}\left[|(I)|>\left\{4c_pM\lambda_n^{-1/2}(M^2\lambda_n c_n')^{(2-p)/4}c_n^{1/2}n^{-1/2}\right\}^{4/(p+2)}+c_pM\lambda_n^{-1/2}c_n^{2/(2+p)}n^{-2/2+p}+c_n'Kbn^{-1}+2KBbn^{-1}\right]\leq 2e^{-b},$$

with $c_n'=O(\lambda_n^{-1})$ and $B=O(\lambda_n^{-1/2})$. Using M_p as a new constant depending on p, we subsequently obtain the desired results.

Geometric noise exponent

The approximation error depends on the noise component q, called the geometric noise exponent (Steinwart & Scovel, 2007). Let

$$\eta(x)=\frac{E\{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)|X=x,A=1\}-E\{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)|X=x,A=-1\}}{E\{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)|X=x,A=1\}+E\{R(Y,\Delta,S_C^m,E_{\tilde{T}}^m)|X=x,A=-1\}}+1/2.$$

Hence, $2\eta(x)-1$ is the decision boundary for the optimal treatment decision rules when we use the pseudo-outcomes. We further define $\mathcal{X}^+=\{x\in\mathcal{X}:2\eta(x)-1>0\}$, and $\mathcal{X}^-=\{x\in\mathcal{X}:2\eta(x)-1<0\}$. A distance function to the boundary between $\mathcal{X}^+$ and $\mathcal{X}^-$ is $\Delta(x)=d(\tilde{x},\mathcal{X}^+)$ if $x\in\mathcal{X}^-$, $\Delta(x)=d(\tilde{x},\mathcal{X}^-)$ if $x\in\mathcal{X}^+$ and $\Delta(x)=0$ otherwise, where $d(\tilde{x},\mathcal{O})$ denotes the distance of x to a set $\mathcal{O}$ with respect to the Euclidean norm. Then the distribution P is said to have geometric noise exponent $0<q<\infty$, if there exists a constant $C>0$ such that

$$E\left[\exp\left\{-\frac{\Delta(X)^2}{t}\right\}|2\eta(X)-1|\right]\leq Ct^{qd/2},t>0.$$

$\Delta(X)$ actually measures the size of the set of points that are close to the opposite class. Indeed, if the data are distinctly separable, that is, when $|2\eta(x)-1|>\delta>0$, for some constant δ, and η is continuous, q can be very large. If either model for the survival time or the censoring time is correctly specified, $2\eta(x)-1$ is the optimal treatment decision rule, where $\mathrm{sign}\{2\eta(x)-1\}=\mathrm{sign}\{f^*(X)\}$.

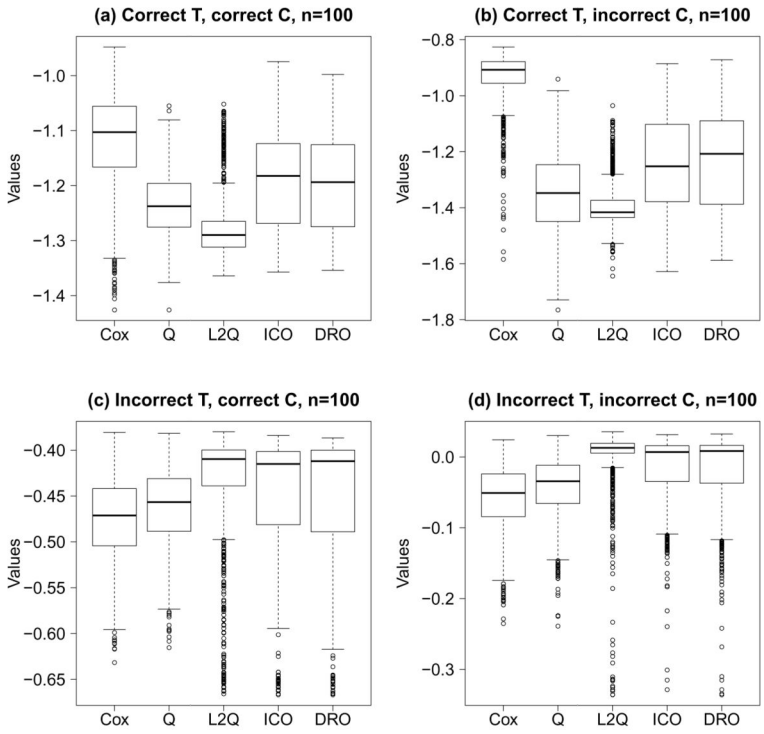

Fig. 1.

Boxplots of values of estimated rules using different methods, representing the logarithm of the survival time with higher values being more preferable. Cox, Cox model; Q, inverse censoring weighted Q-learning; L2Q, inverse censoring weighted L_2 Q-learning; ICO, inverse censoring weighted outcome weighted learning with linear kernel; DRO, doubly robust outcome weighted learning with linear kernel.

Table 1

Mean (s.e.) cross-validated days on log sale using different working models for C, with working model for T being a Cox regression model with basis (X, A, XA)

	Mean (s.e.) Cross-validated Values				
Working model for *C*	**Cox**	**Q**	**L_2Q**	**ICO**	**DRO**
Cox model, basis (X, A, XA)	5·061 (0·381)	5·234 (0·578)	5·258 (0·575)	5·339 (0·404)	5·257 (0·526)
Kaplan–Meier	5·061 (0·381)	5·002 (0·621)	5·000 (0·632)	5·473 (0·166)	5·446 (0·273)

s.e., standard errors; Cox, Cox model; Q: inverse censoring weighted Q-learning; L_2Q, inverse censoring weighted L_2 Q-learning; ICO, inverse censoring weighted outcome weighted learning with linear kernel; DRO, doubly robust outcome weighted learning with linear kernel.