*Submitted to the Annals of Applied Statistics*

# DOUBLY ROBUST ESTIMATION OF OPTIMAL TREATMENT REGIMES FOR SURVIVAL DATA - WITH APPLICATION TO AN HIV/AIDS STUDY

By Runchao Jiang[1], Wenbin Lu[1], Rui Song[1], Michael G. Hudgens[2] and Sonia Naprvavnik[2]

*North Carolina State University*[1] *and University of North Carolina at Chapel Hill*[2]

In many biomedical settings, assigning every patient the same treatment may not be optimal due to patient heterogeneity. Individualized treatment regimes have the potential to dramatically improve clinical outcomes. When the primary outcome is censored survival time, a main interest is to find optimal treatment regimes that maximize the survival probability of patients. Since the survival curve is a function of time, it is important to balance short-term and long-term benefit when assigning treatments. In this paper, we propose a doubly robust approach to estimate optimal treatment regimes that optimize a user specified function of the survival curve, including the restricted mean survival time and the median survival time. The empirical and asymptotic properties of the proposed method are investigated. The proposed method is applied to a data set from an ongoing HIV/AIDS clinical observational study conducted by the University of North Carolina (UNC) Center of AIDS Research (CFAR), and shows the proposed methods significantly improve the restricted mean time of the initial treatment duration. Finally the proposed methods are extended to multi-stage studies.

**1. Introduction.** The primary outcome of interest in many clinical studies is survival time, which can be, for example, how long patients will be alive after treatment initiation, or how long patients will stay on the current treatment before switching to other treatment. One big challenge in this setting is that the event of interest may not be observed for all patients by the end of the study, i.e., the survival times are subject to right censoring.

Traditionally, interest focused on estimating the survival functions under two treatment options and then evaluating which treatment is better. One commonly used estimator for the survival function is the Kaplan-Meier estimator (Kaplan and Meier, 1958), while the Cox proportional hazard model (Cox, 1972) is another popular tool for studying effects of covariates on survival. Sometimes, we may find the estimated survival curves intersect at one

*Keywords and phrases:* Doubly robust estimation, median survival time, optimal treatment regimen, restricted mean survival time.

or more time points; however, this does not necessarily indicate that the two treatments are equally efficacious for every patient. Moreover, it is not uncommon that two different treatments favor different sub-groups of patients, yet yield similar survival time distributions in the entire population. For example, Jiang et al. (2016) showed that the zidovudine plus didanosine treatment and zidovudine plus zalcitabine treatment led to similar survival curve estimates for HIV infected individuals with CD4 counts between 200 and 500 per milliliter. However, for the subgroup of older HIV infected patients with age greater than or equal to 34 years, the zidovudine plus didanosine treatment recipients showed slower disease progression compared to the zidovudine plus zalcitabine treatment. In contrast, for the subgroup of HIV infected patients with age less than 34 years, the zidovudine plus zalcitabine treatment was associated with slower disease progression compared to the zidovudine plus didanosine treatment.

To formalize this idea, we consider individualized treatment regimes. A treatment regime is a deterministic function that maps patient specific data to candidate treatments. An optimal treatment regimen assigns treatment individually to each patient in order to maximize some clinical outcome or utility (e.g., maximize the median survival time). Even if treatments have similar effects at the population level, we still have the potential to further improve clinical benefit by appropriate personalization.

Two popular modeling approaches to estimate the individualized treatment regime are Q-learning and A-learning (Watkins and Dayan, 1992; Murphy, 2005; Zhao et al., 2009; Murphy, 2003; Robins, 2004). When survival time is the primary endpoint of interest, Chen and Tsiatis (2001) proposed using a Cox model with treatment-covariate interaction terms to estimate the optimal individualized treatment regime. Tian et al. (2014) proposed a similar approach by fitting a Cox model with modified covariates. A concern with these approaches is that the Cox model relies on the proportional hazard assumption. Under this assumption, the regime that maximizes a short-term outcome would be the same as the regime that maximizes a long-term outcome. That one regime may be optimal for both short-term and long-term outcomes may be implausible in many cases. For example, coronary bypass surgery is not as favorable as medical therapy in the short term due to its perioperative mortality, but the advantage of surgery is evident in the long term (Zucker, 1998). In this case, the proportional hazard model is no longer suitable and thus, the associated optimal regime is questionable.

Another approach might entail finding the optimal regime which maximizes the survival probability at a particular time point, say three years after treatment. The $t$-year survival probability is a commonly used crite-

rion to compare different treatments. Bai et al. (2013) proposed a locally efficient estimator to compare treatment specific $t$-year survival probabilities. Jiang et al. (2016) proposed a doubly robust method to estimate the optimal regime for maximal $t$-year survival probability. However, one potential problem of using the $t$-year survival probability criterion is that the choice of time $t$ can be subjective. Additionally, it is difficult to balance the short-term benefit and long-term benefit by using a single value of $t$.

Alternatively, some function of the entire survival curve may be a better criterion to measure the treatment effect, due to its more composite nature. One example is the restricted mean survival time (RMST) (Irwin, 1949), which accumulates information up to a pre-determined time point and balances the short term effect and long term effect to some extent. Goldberg and Kosorok (2012) proposed a Q-learning method for censored data to estimate the dynamic optimal regime that maximizes the RMST. However, the proposed Q-learning method relies on the assumed model for the survival time. Zhao et al. (2015) developed a doubly robust method using outcome weighted learning to maximize the RMST. Another informative measure in survival analysis is the median survival time. To date, there are no methods for determining the optimal individualized treatment regime that maximizes the median survival time.

In this article, we propose a doubly robust approach to estimate the optimal treatment regime, which is an extension of the inverse propensity score weighted (IPSW) and augmented inverse propensity score weighted (AIPSW) Kaplan-Meier estimators of the $t$-year survival probability proposed in Jiang et al. (2016). The proposed methods demonstrate how to estimate the optimal treatment regime which maximizes a user-specified function of the survival curve, such as the RMST, median survival time, or $t$-year survival probability. When the user-specified function is the RMST, the proposed approach differs from Zhao et al. (2015) in two respects. First, the proposed method directly maximizes the estimated RMST while the inverse probability of censoring weighted (IPCW) outcome weighted (OW) learning approach of Zhao et al. (2015) does not because the optimization problem is transformed into a classification problem. Therefore, for the proposed method it is straightforward to derive the asymptotic distribution of the estimated RMST under the estimated optimal treatment regimen; it is not clear how to do this using the IPCW-OW learning approach. Second, the IPSW Kaplan-Meier estimator proposed in Jiang et al. (2016) is not equivalent to the IPCW estimator of the regime-specific survival function. Therefore, our proposed estimator for the regime-specific restricted mean survival time is different from the IPCW-OW-learning estimator of Zhao

et al. (2015).

The rest of the article is organized as follows. Section 2 describes an HIV/AIDS treatment study which motivates the developed methodology. Section 3 presents the proposed method. Sections 4 and 5 show simulations and the analysis results for the HIV/AIDS study, respectively. Section 6 extends the proposed method to multistage studies, followed by a discussion section. The Supplementary Appendix includes technical conditions and additional simulation results.

**2. Data.** This research is motivated by a data set from an ongoing HIV/AIDS clinical observational study conducted by the University of North Carolina (UNC) Center of AIDS Research (CFAR). The UNC CFAR Clinical Cohort was created in 2000 and includes data from over 4800 HIV infected patients (Howe et al., 2010). Antiretroviral therapy (ART) suppresses circulating levels of HIV RNA, with most patients treated with modern ART achieving and maintaining undetectable HIV RNA levels for years (Dombrowski et al., 2013). Long-term HIV RNA suppression improves immune function and lowers the risk of adverse clinical complications. A large number of antiretroviral agents are available which can be categorized into a number of classes based on the type of compound and mode of action. Until recently the three most commonly used agents included drugs from three specific classes: nucleoside reverse transcriptase inhibitors (NRTI), non-nucleoside reverse transcriptase inhibitors (NNRTI) and protease inhibitors (PI) which may or may not have been pharmacokinetically enhanced. Modern ART includes a combination of HIV antiretroviral agents; in general this combination (or regimen) includes at least three agents from at least two different classes (Gunthard et al., 2014). For an individual patient, the component agents of ART are changed as needed based on treatment failure, emergence of drug resistance, and/or issues with tolerability. Maximizing the initial treatment duration, the time between ART initiation and discontinuation or modification, is critical to optimal clinical outcomes since shorter initial treatment duration is associated with greater morbidity and mortality (Willig et al., 2008). Therefore choosing between different possible ART regimens for initial treatment is essential to long term outcomes for HIV infected individuals. In this paper we consider choosing the initial ART regimen based on individual patient characteristics in order to maximize expected initial treatment duration.

In the UNC CFAR Clinical Cohort, ART-naive patients were followed from the later of January 2000 or ART initiation until ART modification or discontinuation, loss to follow-up or administrative censoring. The study

data included 990 HIV-infected patients who were 72% male, 57% black, 28% white, 9% Hispanic and 6% of other races/ethnicities. The median age at ART initiation was 38 years, 44% were men who have sex with men (MSM) and 7% had a history of injection drug use (IDU). At ART initiation (baseline) the median CD4 cell count was 209 cells/mm$^3$ (range 1 to 1422) and the median HIV RNA level was 4.9 $\log_{10}$ copies/mL (range 1.6 to 7.2). The initial ART treatment was chosen by providers and patients based on clinical indication and included in all cases two NRTI agents with either an NNRTI or PI.

## 3. Methodology.

3.1. *The General Strategy.* Assume that a study consists of $n$ independently and identically distributed observations. The $i$th observation contains the $p$-dimensional covariates $X_i \in \mathcal{X}$ and the observed treatment assignment $A_i \in \mathcal{A}$. The assignment of $A_i$ may depend on $X_i$. The $i$th observation also contains $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I\{T_i \leq C_i\}$, where $T_i$ is the survival time and $C_i$ is the censoring time. An individualized treatment regime $g$ is a function that maps covariate space $\mathcal{X}$ to treatment space $\mathcal{A}$. The objective is to estimate the optimal treatment regime $g^{\text{opt}}$ which maximizes $f(S(\cdot))$, where $f$ is some pre-specified function of the survival function $S(t) = P(T > t)$. For example, $f(S(\cdot)) = S(t_0)$ is the $t_0$-year survival probability; $f(S(\cdot)) = \int_0^L S(t)dt$ is the restricted mean survival time up to time $L$; and $f(S(\cdot)) = \inf\{t : S(t) \leq 0.5\}$ is the median survival time.

For simplicity, we consider two treatment options $\mathcal{A} = \{0, 1\}$, though the proposed methods can be easily extended to cases with multiple treatment options. We are interested in estimating the optimal regime $g^{\text{opt}}$ within a class of feasible regimes $\mathcal{G}$, which is parametrized by a finite-dimensional parameter $\eta$. As an example, we may take $\mathcal{G} = \{g : g(x; \eta) = I(\eta^T \tilde{x} \geq 0)\}$, where $\eta \in \mathbb{R}^{p+1}$ and $\tilde{x} = (1, x^T)^T$. Regimes of this form recommend treatment 1 if the linear combination of the covariates $\eta^T \tilde{x}$ is greater than or equal to zero, and recommend treatment 0 otherwise, given a patient's covariate $x$. We denote the survival curve under regime $g(x; \eta)$ by $S(t; \eta)$. Estimation of the optimal regime $g^{\text{opt}}$ is equivalent to the estimation of the optimal $\eta^{\text{opt}}$. The general idea underlying the proposed method is to approximate $f(S(t; \eta))$ by $f(\widehat{S}(t; \eta))$, where $\widehat{S}(t; \eta)$ is a non-parametric estimator (defined below) of $S(t; \eta)$, and then estimate $\eta^{\text{opt}}$ by maximizing $f(\widehat{S}(t; \eta))$. The optimal regime $g^{\text{opt}}$ is then estimated by $g(x; \hat{\eta}^{\text{opt}})$, where $\hat{\eta}^{\text{opt}}$ is the maximizer of $f(\widehat{S}(t; \eta))$.

Jiang et al. (2016) proposed two propensity score based Kaplan-Meier estimators of the survival curve under any given regime. The inverse propensity

score weighted estimator for the survival curve under regime $g(x;\eta)$ is

$$\widehat{S}_I(v;\eta) = \prod_{s\le v}\left\{1 - \frac{\sum_{i=1}^n \hat{w}_{\eta,i} dN_i(s)}{\sum_{i=1}^n \hat{w}_{\eta,i} Y_i(s)}\right\}. \tag{1}$$

where $v > 0$ is the time point of interest, $N_i(s) = I\{\tilde{T}_i \le s, \delta_i = 1\}$ is the counting process, $Y_i(s) = I\{\tilde{T}_i \ge s\}$ is the at risk process, and $\hat{w}_{\eta,i}$ is an estimate of the weight $w_{\eta,i}$ for the $i$th observation. The weight $w_{\eta,i}$ is

$$w_{\eta,i} = \frac{A_i I\{\eta^T \tilde{X}_i \ge 0\} + (1-A_i)(1 - I\{\eta^T \tilde{X}_i \ge 0\})}{A_i \pi(X_i) + (1-A_i)(1-\pi(X_i))}, \tag{2}$$

where $\pi(X_i) = P(A_i = 1|X_i)$ is the propensity score for treatment assignment. To estimate the propensity score, we posit a logistic regression model with respect to $A_i$ and covariates $X_i$

$$\text{logit}(P(A_i = 1|X_i;\theta)) = \theta^T \tilde{X}_i, \tag{3}$$

where $\tilde{X}_i = (1, X_i^T)^T$. Then $\pi(X_i)$ is estimated by $\exp(\hat{\theta}^T \tilde{X}_i)/\{1+\exp(\hat{\theta}^T \tilde{X}_i)\}$ where $\hat{\theta}$ is the maximum likelihood estimator. Estimates of $\pi(X_i)$ are plugged-in to (2) to compute the estimated weights $\hat{w}_{\eta,i}$. As long as model (3) is correctly specified, $\hat{S}_I(v;\eta)$ is consistent and asymptotically normal.

Another estimator for the survival curve under regime $g(x;\eta)$ is the augmented inverse propensity score weighted estimator,

$$\widehat{S}_A(v;\eta) = \prod_{s\le v}\left(1 - \frac{\sum_{i=1}^n [\hat{w}_{\eta,i} dN_i(s) + (1-\hat{w}_{\eta,i})\hat{S}_T\{s|g_\eta(X_i), X_i\}\hat{S}_C(s) d\hat{\Lambda}_T\{s|g_\eta(X_i), X_i\}]}{\sum_{i=1}^n [\hat{w}_{\eta,i} Y_i(s) + (1-\hat{w}_{\eta,i})\hat{S}_T\{s|g_\eta(X_i), X_i\}\hat{S}_C(s)]}\right), \tag{4}$$

where $\hat{S}_T(s|a,x)$ is an estimator of $P(T_i \le s|A_i = a, X_i = x)$, the survival probability conditional on covariates and received treatment; $\hat{\Lambda}_T(s|a,x) = -\log \hat{S}_T(s|a,x)$ is the estimated cumulative hazard function for $T$; and $\hat{S}_C(s)$ is the estimated survival probability for the censoring time. The additional terms in (4), when compared to (1), contain information from regression models of the survival time $T$ and the censoring time $C$. The estimates of the survival function $\hat{S}_T\{s|g_\eta(X_i), X_i\}$ and the hazard function $\hat{\Lambda}_T\{s|g_\eta(X_i), X_i\}$ can be obtained by fitting the Cox proportional hazard model (Cox, 1972)

$$\Lambda_T(u|A, X;\beta) = \Lambda_0(u)\exp(\beta^T(X^T, A, AX^T)^T), \tag{5}$$

where $\Lambda_0(u)$ is the baseline cumulative hazard function and $\beta$ is a $(2p+1)$-dimensional parameter. Under the assumption of independent censoring, $S_C(v)$ can be consistently estimated by Kaplan-Meier estimator (Kaplan and Meier, 1958). The benefit of including the augmented terms is double robustness. As long as either model (3) or model (5) is correctly specified, $\widehat{S}_A(v;\eta)$ is consistent and asymptotically normal.

Both $\widehat{S}_I(v;\eta)$ and $\widehat{S}_A(v;\eta)$ have corresponding smoothed versions, which are denoted as $\widehat{S}_{SI}(v;\eta)$ and $\widehat{S}_{SA}(v;\eta)$ respectively. $\widehat{S}_{SI}$ has the same form as $\widehat{S}_I(v;\eta)$, except that the indicator function $I(\eta^T\tilde{X}_i > 0)$ is replaced by $\Phi(\eta^T\tilde{X}_i/h)$, where $\Phi(s)$ is the cumulative distribution function of the standard normal distribution and $h$ is the bandwidth. The same modification is applied to $\widehat{S}_A(v;\eta)$ in order to obtain $\widehat{S}_{SA}(v;\eta)$. In practice, $h$ is set to $n^{-1/3}sd(\eta^T X)$, where $sd(z)$ is the standard deviation of $z$. The smoothed versions $\widehat{S}_{SI}(v;\eta)$ and $\widehat{S}_{SA}(v;\eta)$ have the same asymptotic properties as the original versions $\widehat{S}_I(v;\eta)$ and $\widehat{S}_A(v;\eta)$ respectively, but they tend to have better finite sample performance as demonstrated in Jiang et al. (2016).

With consistent estimators for the survival curve, we can easily estimate $f(S(t;\eta))$ by $f(\widehat{S}_K(t;\eta))$ under any given regime $g(x;\eta)$, where $K = I, A, SI,$ or $SA$. The optimizer of $f(\widehat{S}_K(t;\eta))$ with respect to $\eta$, denoted by $\hat{\eta}_K^{\text{opt}}$, is a natural estimator for the optimal $\eta^{\text{opt}}$. We estimate the optimal regime $g^{\text{opt}}$ by $g(x;\hat{\eta}_K^{opt})$. In subsequent subsections, we shall discuss two important examples.

3.2. *Restricted Mean Survival.* One popular scalar summary of the survival curve is the RMST, in which case the function $f$ is defined as $f(S(t;\eta)) = \int_0^L S(v;\eta)dv$ for some pre-determined time point $L$. Let $\eta^{\text{opt}}$ be the maximizer of $f(S(t;\eta))$, such that $g(x;\eta^{\text{opt}})$ is the optimal regime. Clearly, the value of $\eta^{\text{opt}}$ may depend on the value of $L$. For simplicity, we suppress such dependence in the notation. The RMST summarizes information up to time $L$. The time $L$ may be chosen based on study specific considerations in order to balance both short-term and long-term outcomes.

Using the aforementioned strategy, the RMST under regime $g(x;\eta)$ can be estimated by

$$\begin{aligned}\widehat{R}_K(\eta) &= \int_0^L \widehat{S}_K(v;\eta)dv \\ &= \sum_{i=0}^{n} I(\tilde{T}_{(i)} \le L)\widehat{S}_K(\tilde{T}_{(i)},\eta)\left[\min\{\tilde{T}_{(i+1)}, L\} - \tilde{T}_{(i)}\right], \end{aligned} \tag{6}$$

where $\tilde{T}_{(0)} = 0$, $\{\tilde{T}_{(i)}\}_{i=1}^n$ are the order statistics of $\{\tilde{T}_i\}_{i=1}^n$ and $K = I, A, SI,$ or $SA$. The optimal regime that maximizes the RMST can be esti-

mated by $g(x;\hat{\eta}_K^{\text{opt}})$ where $\hat{\eta}_K^{\text{opt}}$ is the maximizer of $\widehat{R}_K(\eta)$. Note that we have four estimators of the optimal individualized regimes. Two are based on the inverse propensity score weighted approach, while the other two are based on the augmented inverse propensity score weighted approach. Theorem 1 establishes their asymptotic properties.

THEOREM 1. *Under certain regularity conditions (see Supplementary Appendix), as* $n \to \infty$,

*(i.) if model (3) is correctly specified,* $\widehat{R}_K(\hat{\eta}_K^{opt})$ *is consistent for* $R(\eta^{opt})$ *and* $\sqrt{n}(\widehat{R}_K(\hat{\eta}_K^{opt}) - R(\eta^{opt})) \to^d N(0, \sigma^2_{R,K}(\eta^{\text{opt}}))$, *for* $K = I$ *or* $SI$.

*(ii.) if either the model (3) or the model (5) is correctly specified ,* $\widehat{R}_K(\hat{\eta}_K^{opt})$ *is consistent for* $R(\eta^{opt})$ *and* $\sqrt{n}(\widehat{R}_K(\hat{\eta}_K^{opt}) - R(\eta^{opt})) \to^d N(0, \sigma^2_{R,K}(\eta^{\text{opt}}))$, *for* $K = A$ *or* $SA$.

The proof of Theorem 1 relies on $\widehat{R}_K(\eta)$ being a linear function of estimates of the $t$-year survival probability $\widehat{S}_K(t,\eta)$. Therefore, the proofs of the asymptotic properties in Theorem 1 utilize those for $\widehat{S}_K(t,\eta)$ given in Jiang et al. (2016). Consistent estimates of the asymptotic variances can also be similarly obtained. Details are presented in the Appendix. Note $\sigma^2_{R,I}(\eta^{\text{opt}}) = \sigma^2_{R,SI}(\eta^{\text{opt}})$ and $\sigma^2_{R,A}(\eta^{\text{opt}}) = \sigma^2_{R,SA}(\eta^{\text{opt}})$, i.e. smoothing does not impact the asymptotic distribution. Nevertheless, smoothing tends to improve the finite sample performance of the estimators, as demonstrated below in Section 4.

3.3. *Median Survival Time.* Another commonly used measure to characterize the survival curves in clinical studies is the median survival time. Under any regime $g(x;\eta)$, the median survival is defined as

$$\xi(\eta) = \inf\{t : S(t;\eta) \le 0.5\}. \tag{7}$$

Let $\eta^{\text{opt}}$ denote the maximizer of $\xi(\eta)$. A natural estimator for $\eta^{\text{opt}}$ is $\hat{\eta}_K^{\text{opt}} = \arg\max_\eta\{\widehat{\xi}(\eta)\}$, where $\widehat{\xi}(\eta) = \inf\{t : \widehat{S}_K(t;\eta) \le 0.5\}$ and $K = I, SI, A$ or $SA$. We estimate the optimal regime $g^{\text{opt}}$ that maximizes the median survival time by $g(;\hat{\eta}_K^{\text{opt}})$ accordingly. The proposed method can be easily extended to other cases where $q$th-quantile of the survival probability, $\xi(\eta) = \inf\{t : S(t;\eta) \le 1-q\}$, is of interest, for some $q \in (0,1)$. We let $\widehat{\xi}_K(\eta) = \inf\{t : \widehat{S}_K(t;\eta) \le 1-q\}$, $\hat{\eta}_K^{\text{opt}} = \arg\max_\eta\{\widehat{\xi}(q;\eta)\}$ and estimate $g^{\text{opt}}$ by $g(x;\hat{\eta}_K^{\text{opt}})$. As before, we have four estimators for the optimal regime. Theorem 2 establishes the asymptotic properties of these estimators.

THEOREM 2. *Under certain regularity conditions (see Supplementary Appendix), as* $n \to \infty$,

*(i.) if model (3) is correctly specified,* $\widehat{\xi}_K(\hat{\eta}_K^{opt})$ *is consistent for* $\xi(\eta^{opt})$ *and* $\sqrt{n}(\widehat{\xi}_K(\hat{\eta}_K^{opt}) - \xi(\eta^{opt})) \to^d N(0, \sigma^2_{\xi,K}(\eta^{\text{opt}}))$*, for* $K = I$ *or* $SI$*.*

*(ii.) if either the model (3) or the model (5) is correctly specified ,* $\widehat{\xi}_K(\hat{\eta}_K^{opt})$ *is consistent for* $\xi(\eta^{opt})$ *and* $\sqrt{n}(\widehat{\xi}_K(\hat{\eta}_K^{opt}) - \xi(\eta^{opt})) \to^d N(0, \sigma^2_{\xi,K}(\eta^{\text{opt}}))$*, for* $K = A$ *or* $SA$*.*

The proof of Theorem 2 makes use of the functional delta method and details are given in the Appendix. Similar to the RMST cases, smoothing does not affect the asymptotic distribution, i.e. $\sigma^2_{\xi,I}(\eta^{\text{opt}}) = \sigma^2_{\xi,SI}(\eta^{\text{opt}})$ and $\sigma^2_{\xi,A}(\eta^{\text{opt}}) = \sigma^2_{\xi,SA}(\eta^{\text{opt}})$, but it tends to improve the finite sample performance. To estimate the asymptotic variances, it can be shown that $\sigma^2_{\xi,K}(\eta^{\text{opt}}) = \sigma^2_K(\xi;\eta^{\text{opt}})/q^2(\xi;\eta^{\text{opt}})$, where $\xi = \xi(\eta^{\text{opt}})$, $\sigma^2_K(\xi;\eta^{\text{opt}})$ is the asymptotic variance of the estimator $\widehat{S}_K(\xi,\eta^{\text{opt}})$ for the $\xi$-year survival probability and $q(\xi;\eta) = -dS(t,\eta)/dt|_{t=\xi}$. A consistent estimator of $\sigma^2_{\xi,K}(\eta^{\text{opt}})$ is presented in the Appendix.

**4. Simulation Studies.** The performance of the proposed methods was investigated by several simulation studies. In the first set of simulations, for each individual $p = 2$ covariates $X_1$ and $X_2$ were independently generated from uniform$(-2, 2)$ distribution. Treatment $A$, either 1 or 0, was assigned based on a Bernoulli distribution, where the probability of assigning treatment 1 was $\pi(X_1, X_2) = \text{logit}^{-1}(X_1 - 0.5X_2)$. The survival time $T$ was generated from a linear transformation model: $h(T) = -0.5X_1 + A(X_1 - X_2) + \varepsilon$, where $h(s) = \log(e^s - 1) - 2$ and $\varepsilon$ is the error term. We considered two distributions for the error term, either extreme value distribution or logistic distribution. The censoring time was generated from uniform$(0, C_0)$, where $C_0$ was chosen such that the censoring rate was controlled at either 15% or 40%. The sample size was set to either 250 or 500. We aimed to estimate the optimal individualized treatment regimes among the regime class $\{g_\eta(X_1, X_2) = I\{\eta_0 + \eta_1 X_1 + \eta_2 X_2 \geq 0\}, \eta = (\eta_0, \eta_1, \eta_2)^T\}$, which maximize the RMST up to $L = 3$ or the median survival time. We also added the constraint $||\eta|| = 1$, to ensure the uniqueness of the the optimal regime. It is straightforward to show $\eta^{\text{opt}} = (0, 0.707, -0.707)$ for all the simulation scenarios discussed above. Under the true optimal regime $g(x;\eta^{\text{opt}})$, the RMST is 2.13 and the median survival time is 2.33 when $\varepsilon$ is extreme value distributed, and the RMST is 2.28 and the median survival time is 2.72 when $\varepsilon$ is logistic distributed.

We applied the proposed methods with $\widehat{S}_K$, where $K = I, SI, A$ and $SA$. A logistic regression model was fit with either intercept only or intercept along with a linear combination of the $X_1$ and $X_2$, to estimate the treat-

ment assignment mechanism. The former model is mis-specified, while the latter model is correctly specified. Model (5) was used to model the survival time $T$, which is correctly specified when $\varepsilon$ is extreme value distributed and mis-specified when $\varepsilon$ is logistic distributed. The Kaplan-Meier estimator was used to estimate the survival function of the censoring time $C$. The optimization was implemented by a genetic algorithm in the `R` package `rgenoud` (Mebane, Jr. and Sekhon, 2011). We ran 1000 Monte Carlo replications in each simulation scenario.

Results for the RMST are summarized in Figure 1. The first row of Figure 1 shows the true RMST of each estimated optimal regime compared to the true RMST of the true optimal regime. The true RMST was approximated by stochastically generating survival times for $5 \times 10^6$ individuals from the true survival model with treatment assignment according to the estimated optimal regimen. The true RMST was then approximated by the average of the maximum of the simulated survival times and $L = 3$. We also compare the treatment recommendation between the true optimal regime $g(x; \eta^{\text{opt}})$ and the estimated optimal regime $g(x; \hat{\eta}^{\text{opt}})$ and compute the misclassification rate (MR), shown in the second row in Figure 1. Additionally, we show the estimated RMST of the estimated optimal regime (the third row in Figure 1) and the associated empirical coverage probability (CP) of the 95% confidence interval (the fourth row in Figure 1). For brevity we only present the results for 15% censoring and sample size 250. Simulation results for other settings were similar.

When the propensity score model was correctly specified, the $I$, $SI$, $A$ and $SA$ approaches (the left four box plots of each panel in the first two rows) all provided good estimates of the true optimal treatment regime, i.e., the simulated $R(\hat{\eta}^{\text{opt}})$'s were very close to the upper bounds $R(\eta^{\text{opt}})$ and the MRs were very close to zero. When the propensity score model was misspecified but the regression model was correctly specified (the right four box plots of each panel in the left column), the $I$ and $SI$ approaches had relatively large biases as expected, while the $A$ and $SA$ approaches still performed well, demonstrating the double robustness of the $A$ and $SA$ methods. When both the propensity score model and the regression model were misspecified (the right four box plots of each panel in the right column), the $I$, $SI$, $A$ and $SA$ approaches all had some biases, however, the $A$ and $SA$ approaches had much smaller biases than those of the $I$ and $SI$ approaches. This implies that the augmented approaches can help to reduce the biases even when the regression model is misspecified. In addition, as shown in the box plots given in the third row, the estimated RMST based on the unsmoothed approaches $I$ and $A$ all have relatively large biases, which in turn lead to empirical CP

less than the nominal level. But the smoothing technique helps to reduce the biases of the estimated RMST and thus improves the associated empirical CP. In particular, when the propensity score model was correctly specified, the $SI$ and $SA$ approaches have empirical CP close to 95%, while when the propensity score model was misspecified but the regression model was correctly specified, the $SA$ approach has correct empirical CP. The results for the median survival time were given in Figure 2. The findings for the median survival time are similar to those for the RMST.

For the simulations described above, the optimal treatment regimes obtained by maximizing the $t$-year survival probability, restricted mean survival time and median survival time are all the same. In this setting the proposed methods are expected to perform similarly to the method of Jiang et al. (2016) for maximizing the $t$-year survival probability and the method of Zhao et al. (2015) for maximizing the RMST. This is demonstrated empirically by the results in Table 1 of the Supplementary Appendix.

Additional simulations were conducted with $p = 10$ covariates. Specifically, the same simulation setting as above was considered, but eight additional "noise" covariates were generated independent of $T$, each independently generated from uniform$(-2, 2)$. Here, we only considered the model with the extreme value distribution for the error term. For each setting, we generated 500 data sets with sample size 250. Results for the RMST and median survival time are given in Tables 2 and 3 of the Supplementary Appendix, respectively. Table entries give the true RMST and the true median survival times under the estimated optimal treatment regimes, MRs of the estimated optimal treatment regimes, and the average computation time (in seconds) per run. These results indicate that the proposed methods work reasonably well for $p = 10$. The estimated optimal treatment regimes for $p = 10$ covariates give slightly smaller values of the RMST and median survival times with nearly doubled MRs compared with the results for $p = 2$ covariates. This is expected because eight noise variables are added but the estimated optimal treatment regimes are not sparse. As a result, the MRs (comparing the estimated optimal regime with the true sparse regime) increased almost one fold. On the other hand, the RMST and median survival time values of the estimated regimes only decreased slightly because the estimators of $\eta$ are still close to the true value and the RMST and median survival time values of the estimated regimes are less sensitive to the biases of the estimators of $\eta$ compared with the MRs. In addition, the computation time increased 1.5 - 5.5 times compared with simulations with $p = 2$ covariates.

Next, we conducted simulation studies for a setting where the regimes

maximizing the median and restricted mean survival times are different. Specifically, the survival time $T$ was generated by

$$T = 12+0.5\sin(\pi(X_1-X_2))+0.25(1+X_1+2X_2)^2+A(0.5+2X_1-X_2)+(1+2AX_1^2)e,$$

where $X_1$ and $X_2$ were generated independently from uniform$(-2, 2)$, $A$ was generated from Bernoulli distribution with success probability 0.5 and $e$ was an independent error generated from an exponential distribution with mean 0 and variance 1. Under this data generating mechanism, the true optimal treatment regime for maximizing the median survival time is given by $\eta^{\text{opt}} = (0.760, 0.169, -0.690)$, which is different from the regimes that maximize the $t$-year survival probability and restricted mean survival time. The median survival time under the optimal treatment regime $g(x; \eta^{\text{opt}})$ is 14.935. The censoring time $C$ was generated from uniform$(0, C_0)$, where $C_0$ was chosen to give a censoring rate of 0.25. We compare the proposed methods, the method of Jiang et al. (2016) for maximizing the $t$-year survival probability, the method of Zhao et al. (2015) for maximizing the RMST, and Cox regression with the linear baseline covariate effects and linear treatment-covariate interaction effects. Results based on 500 simulated data sets each with sample size 250 are summarized in Table 1. Table entries give the estimates of $\eta$, the true median survival times under the estimated optimal treatment regimens (denoted by $V$), and the MRs of the estimated optimal treatment regimes. Based on the results, compared with other methods, the proposed method for maximizing the estimated median survival time gives estimators of $\eta$ closer to its true value, and leads to estimated optimal treatment regimes with larger median survival times and smaller MRs. Note the treatment effect is relatively small in this simulation setting, such that the advantage of the proposed method is more pronounced when comparing MR rather than $V$.

**5. Application.** In this section, we apply these methods to the UNC CFAR HIV Clinical Cohort study data. Our objective is to identify the optimal treatment regime that results in the expected longest initial treatment duration. Particularly, we aim to find the optimal regime that maximizes the restricted mean initial treatment duration up to day 4000. Day 4000 is chosen so that approximately 99% of event times are less than the time point of interest. The covariates include age, gender (male vs female), race (black, white, Hispanic, or other), MSM (yes, no, or unknown), IDU (yes, no, or unknown), CD4 count, and viral load (VL). Categorical variables are transformed into dummy variables, resulting in an 11-dimensional covariate vector $X$. The NNRTI plus NRTI combination is coded as treatment 1,

TABLE 1

*Simulation results comparing the proposed methods for maximizing the median survival time (denoted by Med) and restricted mean survival time (denoted by RM), the method of Jiang et al. (2016) (denoted by tyear), the method of Zhao et al. (2015) (denoted by Zhao) and the Cox regression (denoted by Cox). I and A denote the inverse probability weighted and augmented inverse probability weighted estimation methods, respectively; SI and SA denote the corresponding smoothing counterparts. Table entries are averages of estimates of $\eta_j$ for $j = 0, 1, 2$, median survival times under the estimated optimal treatment regimes (V), and misclassification rates of the estimated optimal treatment regimes (MR). The numbers in parenthesis are the standard deviations of the corresponding estimates.*

| Method | | $\eta_0$ | $\eta_1$ | $\eta_2$ | $V$ | MR |
|---|---|---|---|---|---|---|
| Med | I | 0.57 (0.32) | 0.24 (0.39) | -0.54 (0.29) | 14.84 (0.08) | 0.20 (0.15) |
| Med | SI | 0.58 (0.31) | 0.27 (0.38) | -0.50 (0.32) | 14.85 (0.07) | 0.21 (0.15) |
| Med | A | 0.59 (0.32) | 0.23 (0.40) | -0.50 (0.31) | 14.85 (0.07) | 0.20 (0.16) |
| Med | SA | 0.58 (0.31) | 0.27 (0.39) | -0.49 (0.33) | 14.85 (0.07) | 0.21 (0.16) |
| RM | I | 0.24 (0.27) | 0.74 (0.16) | -0.51 (0.20) | 14.79 (0.04) | 0.39 (0.08) |
| RM | SI | 0.22 (0.24) | 0.78 (0.15) | -0.48 (0.19) | 14.80 (0.03) | 0.41 (0.07) |
| RM | A | 0.22 (0.25) | 0.77 (0.14) | -0.50 (0.18) | 14.79 (0.04) | 0.40 (0.07) |
| RM | SA | 0.21 (0.22) | 0.79 (0.14) | -0.48 (0.18) | 14.80 (0.03) | 0.41 (0.07) |
| tyear | I | 0.29 (0.23) | 0.82 (0.12) | -0.34 (0.24) | 14.77 (0.06) | 0.43 (0.06) |
| tyear | SI | 0.17 (0.26) | 0.86 (0.09) | -0.35 (0.18) | 14.79 (0.04) | 0.46 (0.04) |
| tyear | A | 0.24 (0.27) | 0.82 (0.13) | -0.33 (0.26) | 14.76 (0.08) | 0.45 (0.06) |
| tyear | SA | 0.14 (0.27) | 0.86 (0.11) | -0.34 (0.22) | 14.77 (0.05) | 0.46 (0.05) |
| Zhao | I | 0.05 (0.55) | 0.32 (0.51) | -0.26 (0.53) | 14.41 (0.51) | 0.45 (0.19) |
| Zhao | A | -0.08 (0.68) | 0.15 (0.47) | -0.18 (0.51) | 14.22 (0.58) | 0.49 (0.23) |
| Cox | | 0.66 (0.14) | 0.54 (0.15) | -0.47 (0.10) | 14.81 (0.04) | 0.28 (0.06) |

while the PI plus NRTI combination as treatment 0. The primary outcome of interest is time to the discontinuation of the initial treatment, which is defined as either a change in the anchor agent (PI or NNRTI), or discontinuing ART for more than 30 days. Among all 990 study patients, 35% were observed to have the event of interest during follow-up, and the remaining patients were censored at their last known clinical encounter.

To estimate the optimal treatment regime, we applied the $I$, $SI$, $A$ and $SA$ approaches. We first fit the logistic regression model (3) to estimate the propensity score. Table 2 shows the estimated coefficients, standard errors and p-values of the estimates. As expected patients with lower CD4 cell counts, indicating more advanced HIV disease progression, were more likely to be prescribed a PI-based regimen, because the commonly used PI had demonstrated greater CD4 cell count recovery and less drug resistance associated with virologic failure in comparison to the commonly used NNRTI during the years of this study. Women were also more likely to be prescribed a PI-based regimen during these years because of concerns the primary NNRTI used may have had teratogenic effects (Panel on Antiretro-

viral Guidelines for Adults and Adolescents, 2016).

TABLE 2
*Estimated coefficients (Est.), standard errors (s.e.), and Wald test p-values (p-val) from the fitted logistic regression model*

| | Int. | age | gender male | $\text{race}_1$ black | $\text{race}_2$ white | $\text{race}_3$ Hispanic | $\text{msm}_1$ yes | $\text{msm}_0$ no | $\text{idu}_1$ yes | $\text{idu}_0$ no | CD4 | VL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Est. | -1.36 | -0.00 | 0.53 | -0.16 | -0.59 | 0.15 | 0.11 | -0.10 | -0.12 | 0.11 | 0.23 | 0.04 |
| s.e. | 0.78 | 0.01 | 0.20 | 0.29 | 0.30 | 0.36 | 0.25 | 0.27 | 0.32 | 0.21 | 0.06 | 0.04 |
| p-val | 0.08 | 0.67 | 0.01 | 0.57 | 0.05 | 0.67 | 0.66 | 0.71 | 0.72 | 0.61 | 0.00 | 0.35 |

For the augmented estimation methods, we fit the proportional hazards model (5). Table 3 presents the estimated coefficients in the optimal regimes obtained by the $I$, $SI$, $A$ and $SA$ approaches. Overall, the four estimated optimal treatment regimes give relatively similar treatment allocation rules. Here we examine the results using the regime $g(x; \hat{\eta}_{SA}^{\text{opt}})$ obtained by the SA approach. We estimate the restricted mean survival times under the estimated optimal regimes and compare these restricted mean survival times with those under the fixed treatment regimes by assigning all patients to one treatment. The restricted mean survival time is 2776 days under the regime $g(\hat{\eta}_{SA}^{\text{opt}})$, 2637 days if all the patients were given treatment 1, and 2339 days if all the patients were given treatment 0. Figure 3 shows the estimated survival curves under the regime $g(\hat{\eta}_{SA}^{\text{opt}})$, the fixed regimes, and the observed treatment assignment (i.e., the empirical regime). The estimated survival curve under the estimated optimal treatment regime $g(\hat{\eta}_{SA}^{\text{opt}})$ is uniformly better than under the empirical and fixed regimes, indicating that the estimated optimal individualized treatment regime may lead to improved clinical outcomes if used in routine medical care. Additionally, the estimated survival curve if all patients were assigned to treatment 1 led to better patient outcomes than if all patients were assigned to treatment 0. Given the antiretroviral agents used in these calendar years these findings are not surprising. The NNRTI used as an anchor agent for treatment 1 continues to be recommended for initial HIV treatment; however, with one exception the PIs included in treatment 0 are no longer recommended as initial treatment (Gunthard et al., 2014). Table 4 shows the 95% confidence intervals of the difference between the restricted mean survival times under the estimated optimal regimes obtained by the I, A, SI and SA approaches and the fixed regimes. The estimated optimal treatment regimes significantly increase the restricted mean survival times of the initial treatment duration compared with the fixed treatment regimes.

Next, we compare treatment allocation of the observed treatment assign-

TABLE 3
*Estimated coefficients of optimal treatment regimes by the I, SI, A and SA methods.*

| | int. | age | gender | $\text{race}_1$ | $\text{race}_2$ | $\text{race}_3$ | $\text{msm}_1$ | $\text{msm}_0$ | $\text{idu}_1$ | $\text{idu}_0$ | CD4 | VL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $I$ | 0.38 | -0.02 | -0.26 | -0.15 | -0.53 | -0.58 | -0.16 | -0.13 | -0.23 | -0.18 | 0.11 | 0.09 |
| $SI$ | 0.48 | -0.02 | -0.12 | 0.13 | -0.35 | -0.46 | -0.09 | -0.10 | -0.43 | -0.44 | 0.03 | 0.09 |
| $A$ | 0.25 | -0.01 | 0.54 | -0.36 | -0.39 | -0.17 | -0.26 | 0.10 | 0.38 | 0.08 | -0.26 | 0.21 |
| $SA$ | 0.48 | -0.02 | -0.09 | 0.07 | -0.36 | -0.47 | -0.10 | -0.08 | -0.43 | -0.44 | 0.03 | 0.09 |

TABLE 4
*Confidence intervals for the difference of estimated restricted mean survival times.*

| | compared to trt. 1 | compared to trt. 0 |
|---|---|---|
| $I$ | $(63, 286)$ | $(232, 707)$ |
| $SI$ | $(54, 249)$ | $(199, 694)$ |
| $A$ | $(20, 139)$ | $(123, 631)$ |
| $SA$ | $(47, 231)$ | $(189, 684)$ |

ment and the estimated optimal regime $g(x; \hat{\eta}_{SA}^{\text{opt}})$ in Table 5. Overall only 55% of the patients received the ART estimated to be the optimal ART by the SA approach. Moreover, the SA approach estimated that 85% patients who received a PI should have received an NNRTI, but only 14% of patients who received an NNRTI would have fared better if they had received a PI-based ART. These findings are supported by the estimated survival curves given in Figure 3 since the survival function if the whole population received treatment 1 is uniformly better than if all patients received treatment 0.

TABLE 5
*Comparison between the observed treatment assignment and recommended treatment by the regime* $g(\hat{\eta}_{SA}^{opt})$.

| | | A | |
|---|---|---|---|
| | | 0 | 1 |
| $g(\hat{\eta}_{SA}^{\text{opt}})$ | 0 | 63 | 76 |
| | 1 | 367 | 484 |

We also compare treatment allocation of the empirical regime and the estimated optimal regime $g(x; \hat{\eta}_{SA}^{\text{opt}})$ across strata of each demographic and clinical patient characteristic of interest. Figure 4 and 5 present the results for the categorical and continuous covariates, respectively. For continuous covariates age, CD4 count and VL, we discretized them into four ranges based on quartiles. Consistent with observations for the entire study population (Figure 3), the estimated optimal regime overwhelmingly favored

initiating an NNRTI-based ART versus a PI-based ART across all patient characteristics. In nearly all cases a greater proportion of patients were allocated to treatment 1 (an NNRTI) by the estimated optimal regime than were observed to receive treatment 1. During the years of this study the PI used most frequently, in comparison to the predominantly used NNRTI, had slightly lower efficacy in reducing circulating HIV RNA levels, but was associated with slightly greater CD4 cell count recovery and lower antiretroviral drug resistance evolution with virologic failure (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2016). These known properties of the primary anchor agents available at the time, in addition to slightly different tolerability profiles of the antiretrovirals under consideration, likely influenced the channeling bias observed in clinical care and shaped the estimated optimal regime results. For example, this effect can be observed for CD4 cell count (Figure 5) where it is clear that patients with lower CD4 were more likely to be prescribed a PI-based ART than those at higher CD4 cell counts. A further example is age, in general patients at older ages enter HIV care and start ART at lower CD4 where a PI-based ART may have been more effective. In general men entered HIV care, and hence started ART, at lower CD4 cell counts in this clinical cohort, therefore as expected the estimated optimal regime was PI-based in a greater proportion of men than women (Figure 4). On the other hand a PI-based regime was prescribed to women at a higher proportion than men. In part this may be related to the efficacy and tolerability differences in the agents used, as described above. Additionally there were clinical concerns that the primary NNRTI available at the time had teratogenic effects, and therefore women of reproductive age may have been steered away from NNRTI use.

## 6. Extension.

6.1. *Framework.* In this section, we extend the proposed methods from Section 3 to multi-stage studies, where treatment assignment is made at multiple time points based on patients' covariate information available at each time point. For simplicity, we consider a two-stage study, with two treatment options at each stage. Assume treatment $A_1$ is assigned $s$ days after the initial treatment $A_0$. The objective is still to maximize either the RMST up to time $L$ or the median survival time.

For each patient, baseline covariates $X_0$ are collected at the first visit and the initial treatment $A_0 \in \mathcal{A}_0 = \{0, 1\}$ is assigned based on $X_0$. The follow-up visit is scheduled at $s$ days after the initial visit. If the patient is still at risk at the second visit, additional covariates $X_1$ are collected and the follow-up treatment $A_1 \in \mathcal{A}_1 = \{0, 1\}$ is given based on the accumulated

information $X_0$, $A_0$ and $X_1$. Thus, the observed data is $\{(X_{0i}, A_{0i}, X_{1i}I\{\tilde{T}_i > s\}), A_{1i}I\{\tilde{T}_i > s\}, \tilde{T}_i, \delta_i), i = 1, \ldots, n\}$.

6.2. *Methods.* We want to find the optimal dynamic regime $g = (g_0, g_1)$ which maximizes the restricted mean survival time or median survival time, respectively. As before, we consider regimes of the form

$$g_0(x_0; \eta_0) = I\{\eta_0^T(1, x_0^T) \geq 0\},$$
$$g_1(x_0, x_1; \eta_1) = I\{\eta_1^T(1, x_0^T, g_0(x_0; \eta_0), x_1^T)) \geq 0\}.$$

Equation (1) is still applicable in the multi-stage studies, if the weight function is replaced with

$$w^{(2)} = \frac{I(\tilde{T}_i \leq s)\delta_i}{\hat{S}_C(\tilde{T}_i)} \times \frac{I\{A_{0i} = g_0(X_{0i}; \eta_0)\}}{\hat{\pi}_{A_0}(X_{0i})} + \frac{I(\tilde{T}_i > s)}{\hat{S}_C(s)} \times \frac{I\{A_{0i} = g_0(X_{0i}; \eta_0), A_{1i} = g_1(X_{0i}, X_{1i}; \eta_1)\}}{\hat{\pi}_{A_0}(X_{0i})\hat{\pi}_{A_1}(X_{0i}, A_{0i}, X_{1i})},$$

where $\hat{\pi}_{A_0}(X_{0i}) = \hat{\pi}_0(X_{0i})A_{0i} + \{1 - \hat{\pi}_0(X_{0i})\}(1 - A_{0i})$, $\hat{\pi}_{A_1}(X_{0i}, A_{0i}, X_{1i}) = \hat{\pi}_1(X_{0i}, A_{0i}, X_{1i})A_{1i} + \{1 - \hat{\pi}_1(X_{0i}, A_{0i}, X_{1i})\}(1 - A_{1i})$, and $\hat{\pi}_0(X_{0i})$ and $\hat{\pi}_1(X_{0i}, A_{0i}, X_{1i})$ are the maximum likelihood estimates of the propensity scores $P(A_{0i} = 1|X_{0i})$ and $P(A_{1i} = 1|X_{0i}, A_{0i}, X_{1i}, \tilde{T}_i > s)$, respectively. See Jiang et al. (2016) for details. Let $\widehat{S}_I^{(2)}(u, \eta)$ denote the resulting estimator of the survival function $S^{(2)}(u, \eta)$ under the regime $g(x_0, x_1; \eta)$.

We can also apply the kernel smoothing technique to improve finite sample performance in the multistage setting. Specifically, we replace the indicator functions $g_0(X_{0i}; \eta_0)$ and $g_1(X_{0i}, X_{1i}; \eta_1)$ in $\widehat{S}_I^{(2)}(u; \eta)$ by $\Phi\left(\eta_0^T(1, X_{0i}^T)/h_0\right)$ and $\Phi\left(\eta_1^T(1, X_{0i}^T, g_0(X_{0i}; \eta_0), X_{1i}^T)/h_1\right)$, respectively, where $h_0$ and $h_1$ are bandwidths. The resulting smoothed estimator is denoted by $\widehat{S}_{SI}^{(2)}(u; \eta)$. A natural estimator of the optimal dynamic treatment regime is given by $\hat{g}_\eta^{\text{opt},(2)} = \{g_0(X_0; \hat{\eta}_{K,0}^{\text{opt},(2)}), g_1(X_0, X_1; \hat{\eta}_{K,1}^{\text{opt},(2)})\}$, where $\hat{\eta}_K^{\text{opt},(2)} = (\hat{\eta}_{K,0}^{\text{opt},(2)}, \hat{\eta}_{K,1}^{\text{opt},(2)})$ maximizes $f(\widehat{S}_K^{(2)}(t; \eta))$, $k = I$ or $SI$, and $f$ is a user-specified function, such as the RMST or median survival time.

Let $\widehat{R}_K^{(2)}(\hat{\eta}_K^{\text{opt},(2)})$ and $\widehat{\xi}_K^{(2)}(\hat{\eta}_K^{\text{opt},(2)})$ denote the estimated RMST and median survival time under the estimated optimal dynamic treatment regime $\hat{g}_\eta^{\text{opt},(2)}$, respectively. Jiang et al. (2016) showed that $\hat{S}_K^{(2)}(u, \eta)$ is a consistent estimator for $S^{(2)}(u, \eta)$ no matter whether $u \geq s$ or $u < s$. Following the proof in Jiang et al. (2016), it can be shown that the estimators $\widehat{R}_K^{(2)}(\hat{\eta}_K^{\text{opt},(2)})$ and $\widehat{\xi}_K^{(2)}(\hat{\eta}_K^{\text{opt},(2)})$ are consistent and asymptotically normal. Theorem 3 establishes the asymptotic properties of these estimators.

THEOREM 3. *Under certain regularity conditions (see Supplementary Appendix), when model $\pi_{A_0}$ and $\pi_{A_1}$ are correctly specified, as $n \to \infty$,*

*(i.)* $\widehat{R}_K^{(2)}(\hat{\eta}_K^{opt,(2)})$ *is consistent for* $R^{(2)}(\eta^{opt,(2)})$ *and* $\sqrt{n}\{\widehat{R}_K^{(2)}(\hat{\eta}_K^{opt,(2)}) - R^{(2)}(\eta^{opt,(2)})\} \to^d N\{0, \sigma^2_{R,K}(\eta^{\text{opt},(2)})\}$, *for* $K = I$ *or* $SI$.

*(ii.)* $\widehat{\xi}_K^{(2)}(\hat{\eta}_K^{opt,(2)})$ *is consistent for* $\xi^{(2)}(\eta^{opt,(2)})$ *and* $\sqrt{n}\{\widehat{\xi}_K^{(2)}(\hat{\eta}_K^{opt,(2)}) - \xi^{(2)}(\eta^{opt,(2)})\} \to^d N\{0, \sigma^2_{\xi,K}(\eta^{\text{opt},(2)})\}$, *for* $K = I$ *or* $SI$.

As before, smoothing does not impact the asymptotic distribution. In addition, the asymptotic variances of the estimators can be consistently estimated in a similar fashion as for the one-stage estimators. We conducted simulation studies to investigate the finite sample properties of the proposed two-stage estimators. The simulation settings and results are given in the Supplementary Appendix. Both $I$ and $SI$ based methods performed well and again smoothing helped improve the finite sample performance.

**7. Discussion.** In this paper, we proposed a doubly robust estimation method for obtaining the optimal treatment regime which maximizes a prespecified function of the survival function, including the RMST and median survival time as special cases. The proposed method can be employed to determine optimal individualized treatment regimes that balance short-term and long-term treatment effects on survival, thus providing optimal regimens that target clinically meaningful quantities of interest. Extensions to multistage studies were also developed, broadening the scope of settings where this method can be applied.

There are several possible avenues of future related research. For instance, in survival analysis it is common for competing risks to be present. In the HIV context, the initial treatment may be discontinued due to several competing reasons. Thus it would be of interest to extend the proposed method to incorporate competing risks. One approach could entail deriving nonparametric estimators of the cumulative incidence function associated with a given treatment regime and then determining the optimal treatment regime which maximizes a prespecified function of the cumulative incidence function. Another common occurrence in survival analysis, especially in HIV studies, is interval censoring wherein the failure time is known only to occur within some interval. Extensions of the proposed methods to allow for interval censoring is another possible area of future research. Finally, similar to the value search method of Jiang et al. (2016), the proposed methods directly maximize the estimated RMST or median survival time using a genetic algorithm. A computational limitation of such algorithms is the inability to handle high dimensional covariates. Thus extensions of the proposed method

to allow for high dimensional covariates could also be considered.

## APPENDIX

**Proof of Theorem 1.** As shown in Jiang et al. (2016), for any time point $t < \tau$, $\widehat{S}_K(t; \hat{\eta}_K^{\text{opt}})$ is consistent for $S(t; \eta^{\text{opt}})$ and

$$\sqrt{n}\{\widehat{S}_K(t; \hat{\eta}_K^{\text{opt}}) - S(t; \eta^{\text{opt}})\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{i,K}(t; \eta^{\text{opt}}, \alpha^*) + o_p(1),$$

where $\zeta_{i,K}(t; \eta, \alpha^*)$ is the $i$th influence function for $\widehat{S}_K(t; \eta, \hat{\alpha})$ and $\alpha^*$ includes all parameters in the treatment assignment model and/or the regression model. For any pre-determined time point $L$, the RMST up to $L$ is a continuous function of $S(t; \eta)$. By applying the continuous mapping theorem, $\widehat{R}_K(\hat{\eta}^{\text{opt}})$ is consistent for $R(\eta^{\text{opt}})$.

By applying the delta method, we have

$$\sqrt{n}\left\{\widehat{R}_K(\hat{\eta}_K^{\text{opt}}, \hat{\alpha}) - R(\eta^{\text{opt}})\right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \rho_{i,K}(\eta^{\text{opt}}, \alpha^*) + o_p(1)$$

where

$$\rho_{i,K}(\eta^{\text{opt}}, \alpha^*) = -\int_0^L S(t; \eta^{\text{opt}}) \zeta_{i,K}(t; \eta^{\text{opt}}, \alpha^*)\, dt.$$

Thus, $\widehat{R}_K(\hat{\eta}_K^{\text{opt}}, \hat{\alpha})$ is asymptotic normal with variance $\sigma^2_{R,K}(\eta^{\text{opt}}) = E[\{\rho_{i,K}(\eta^{\text{opt}}, \alpha^*)\}^2]$, which can be consistently estimated by

$$\hat{\sigma}^2_{R,K}(\hat{\eta}^{\text{opt}}) = n^{-1} \sum_{i=1}^{n} \left[\int_0^L \widehat{S}_K(t; \hat{\eta}_K^{\text{opt}})\} \zeta_{i,K}(t; \hat{\eta}_K^{\text{opt}}, \hat{\alpha})\, dt\right]^2.$$

**Proof of Theorem 2.** Recall that median survival time is also a continuous function of survival time. Define $\phi(S(t; \eta)) = S^{-1}(0.5; \eta) = \inf\{t : S(t; \eta) \geq 0.5\}$. We have $\xi(\eta) = \phi(S(t; \eta))$ and $\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}}) = \phi(\widehat{S}_K(t; \hat{\eta}_K^{\text{opt}}))$. Applying the continuous mapping theorem, $\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}})$ can be shown to be consistent for $\xi(\eta^{\text{opt}})$ .

To derive the limiting distribution of $\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}})$, we follow the steps in Gill et al. (1997, Section IV.3.4). When regularity condition A10 holds, $\phi$ is compactly differentiable at $S$. We have

$$\sqrt{n}\{\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}}) - \xi(\eta^{\text{opt}})\} = \frac{1}{q(\xi; \eta^{\text{opt}})} \sqrt{n}\{\widehat{S}_K(\xi, \hat{\eta}_K^{\text{opt}}) - S(\xi, \eta^{\text{opt}})\} + o_p(1).$$

Thus,

$$\sqrt{n}\{\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}}) - \xi(\eta^{\text{opt}})\} \to^d N\left(0, \frac{\sigma_K^2(\xi;\eta^{\text{opt}})}{q^2(\xi;\eta^{\text{opt}})}\right),$$

where $\sigma_K^2(\xi;\eta^{\text{opt}})$ is the asymptotic variance of the estimator $\widehat{S}_K(\xi, \hat{\eta}_K^{\text{opt}})$ as derived in Jiang et al. (2016). A consistent estimator of $\sigma_{\xi,K}^2(\eta^{\text{opt}}) = \sigma_K^2(\xi;\eta^{\text{opt}})/q^2(\xi;\eta^{\text{opt}})$ can be obtained as $\hat{\sigma}_K^2(\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}}); \hat{\eta}_K^{\text{opt}})/\hat{q}(\widehat{\xi}_K(\hat{\eta}_K^{\text{opt}}); \hat{\eta}_K^{\text{opt}})$, where

$$\hat{q}(v;\eta) = -\frac{1}{h}\int_0^\infty \varphi\left(\frac{v-u}{h}\right) d\widehat{S}_K(u;\eta),$$

$\varphi(x)$ is the density function for the standard normal distribution and $h = sd(\tilde{T}_i) * n^{-1/5}$ is the bandwidth.

**Proof of Theorem 3.** The proof of Theorem 3 is similar to those of Theorem 1 and 2, and is omitted here.

## ACKNOWLEDGEMENT

The authors thank the Editor, the Associate Editor and three referees for their comments that substantially improved the article. This research was supported by the University of North Carolina at Chapel Hill Center for AIDS Research (CFAR), an NIH funded program P30 AI50410. WL and RS were partially supported by NIH grant P01 CA142538, and MH was partially supported by NIH grant R01 AI029168.

## REFERENCES

Bai, X., Tsiatis, A. A., and O'Brien, S. M. (2013). Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics*, 69(4):830–839.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–220.

Dombrowski, J. C., Kitahata, M. M., Van Rompaey, S. E., Crane, H. M., Mugavero, M. J., Eron, J. J., Boswell, S. L., Rodriguez, B., Mathews, W. C., Martin, J. N., Moore, R. D., and Golden, M. R. (2013). High levels of antiretroviral use and viral suppression among persons in HIV care in the United States, 2010. *Journal of Acquired Immune Deficiency Syndromes*, 63(3):299–306.

Gill, R. D., Keiding, N., and Andersen, P. K. (1997). *Statistical Models Based on Counting Processes*. Springer, New York.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *The Annals of Statistics*, 40(1):529–560.

Gunthard, H. F., Aberg, J. A., Eron, J. J., Hoy, J. F., Telenti, A., Benson, C. A., Burger, D. M., Cahn, P., Gallant, J. E., Glesby, M. J., Reiss, P., Saag, M. S., L, T. D., Jacobsen, D. M., and Volberding, P. A. (2014). Antiretroviral treatment of adult HIV infection:

2014 recommendations of the International Antiviral Society-USA panel. *The Journal of American Medical Association*, 312(4):410–425.

Howe, C. J., Cole, S. R., Napravnik, S., and Eron Jr, J. J. (2010). Enrollment, retention, and visit attendance in the University of North Carolina Center for AIDS Research Clinical Cohort, 2001–2007. *AIDS Research and Human Retroviruses*, 26(8):875–881.

Irwin, J. O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene*, 47:188–189.

Jiang, R., Lu, W., Song, R., and Davidian, M. (2016). On estimation of optimal treatment regimes for maximizing $t$-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, in press.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):pp. 457–481.

Mebane, Jr., W. R. and Sekhon, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*, 42(11):1–26.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005). A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097.

Panel on Antiretroviral Guidelines for Adults and Adolescents (2016). *Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents*. Department of Health and Human Services, Available at https://aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3-4):279–292.

Willig, J. H., Abroms, S., Westfall, A. O., Routman, J., Adusumilli, S., Varshney, M., Allison, J., Chatham, A., Raper, J. L., Kaslow, R. A., Saag, M. S., and Mugavero, M. J. (2008). Increased regimen durability in the era of once daily fixed-dose combination antiretroviral therapy. *AIDS*, 22(15):1951–1960.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315.

Zhao, Y., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, in press.

Zucker, D. M. (1998). Restricted mean life with covariates: Modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442):702–709.

Department of Statistics,
North Carolina State University,
Raleigh, NC, USA
E-mail: lu@stat.ncsu.edu

Department of Biostatistics,
University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

School of Medicine,
University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

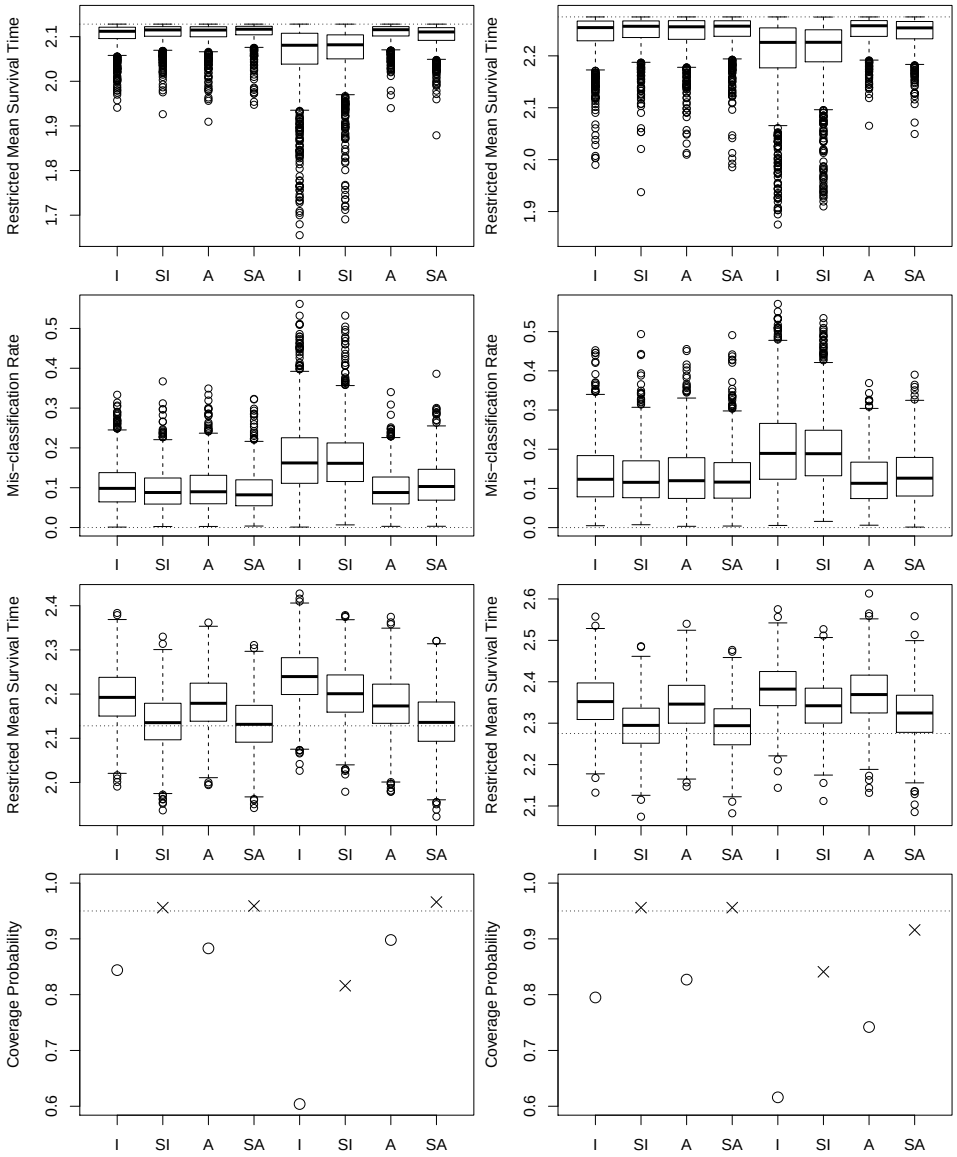


FIG 1. *Simulation results for maximizing the RMST. The left column is for extreme value distributed error, while the right column is for logistic distributed error. The first row displays box plots for* $R(\hat{\eta}^{opt})$*, with the horizontal lines indicating the upper bound* $R(\eta^{opt})$*. The second row displays box plots for MR, with the horizontal lines indicating zero. The third row displays box plots for* $\widehat{R}(\hat{\eta}^{opt})$*, with the horizontal lines indicating the true value* $R(\eta^{opt})$*. The fourth row presents the empirical coverage probability of the confidence interval of* $\widehat{R}_K(\hat{\eta}^{opt})$*, with the horizontal lines indicating the nominal level of* 95%*. Within each panel, the left half of the plot is for correctly specified logistic regression, while right half of the plot is for misspecified logistic regression.*

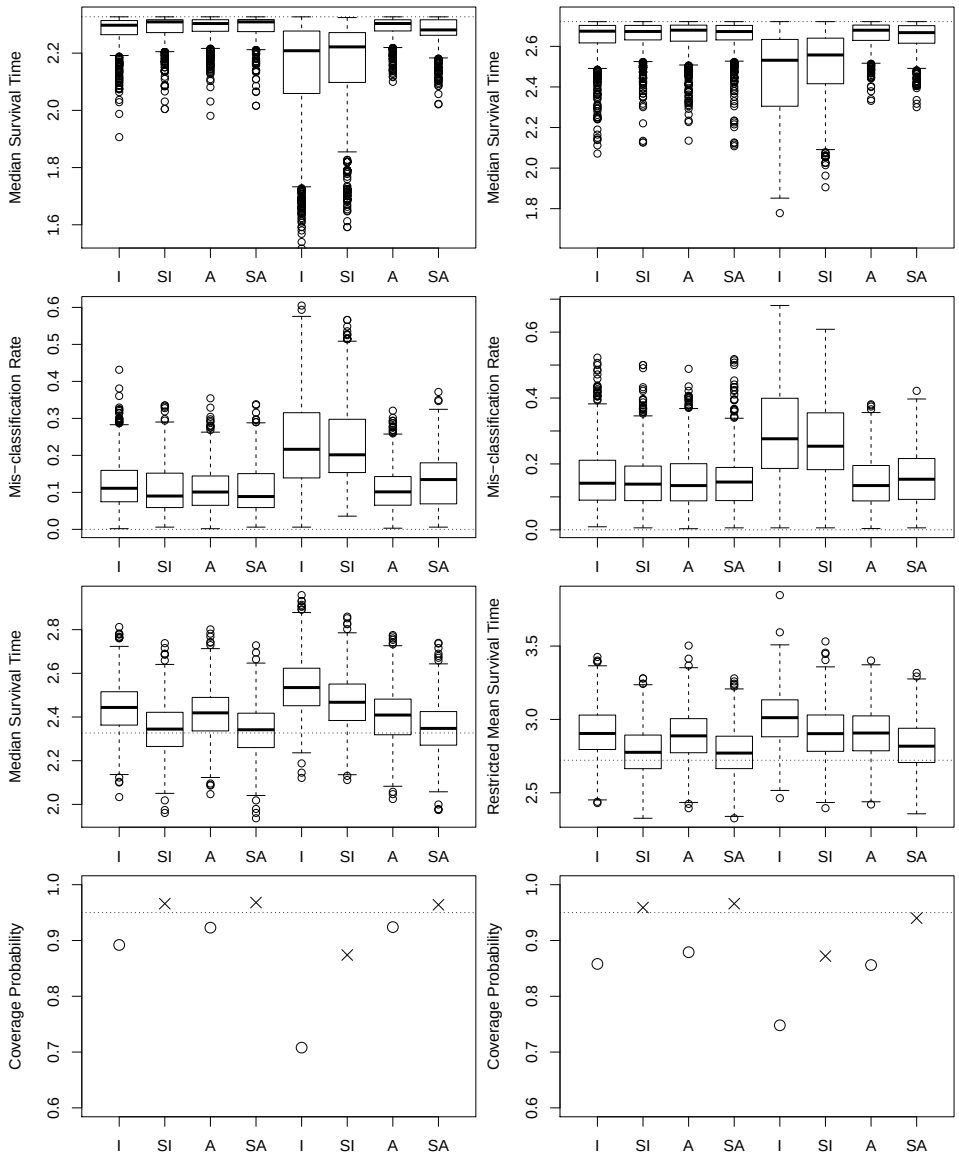


FIG 2. *Simulation results for maximizing the median survival time. See Figure 1 for detailed descriptions of the plots.*

FIG 3. *Survival function estimates if all the patients followed* $g(x;\hat{\eta}_{SA}^{opt})$ *(solid line), the observed treatment (dashed line), received treatment 1 (dotted line) or received treatment 0 (dotted dashed line).*

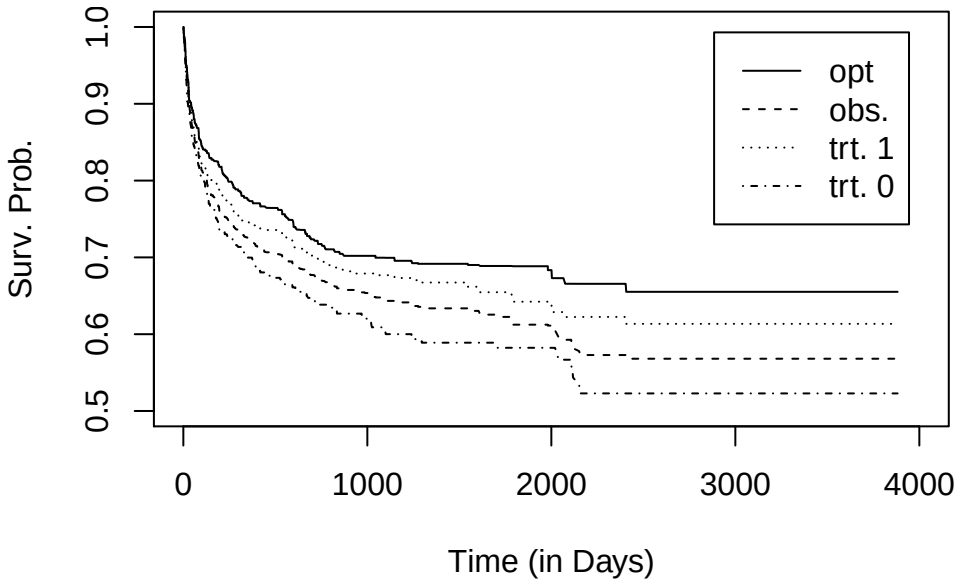

FIG 4. *Comparison of treatment allocation percentages stratified on each categorical covariates. The left panel is for the observed assignment while the right panel is for the estimated optimal treatment regime (denoted by gSA).*

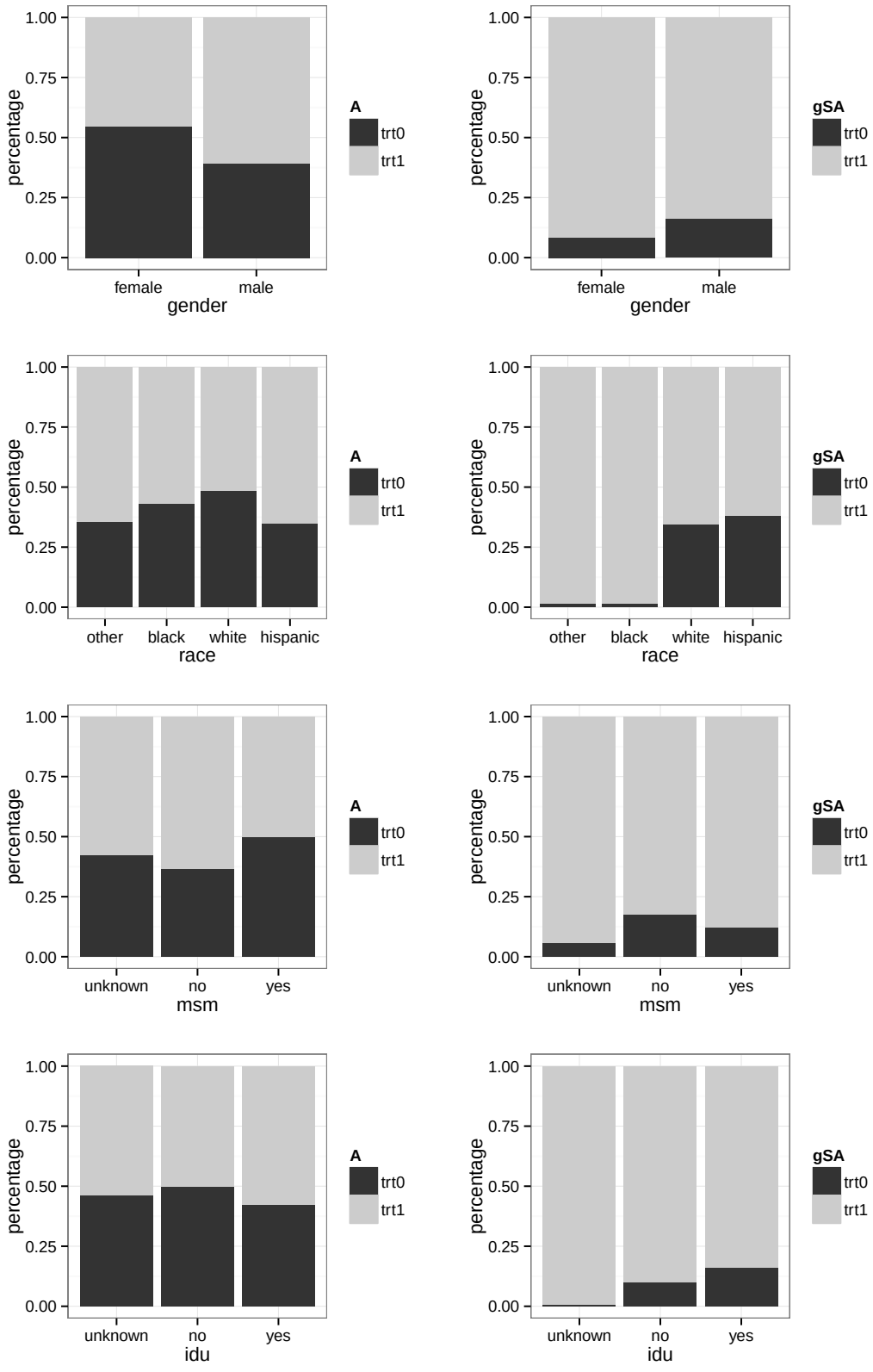

FIG 5. *Comparison of treatment allocation percentages stratified on each continuous covariates (based on quartiles). The left panel is for the observed assignment while the right panel is for the estimated optimal treatment regime (denoted by gSA).*