Balanced Policy Evaluation and Learning for Right Censored Data

Owen E. Leete*
Department of Biostatistics,
University of North Carolina at Chapel Hill

and

Nathan Kallus
School of Operations Research and Information Engineering,
Cornell University

and

Michael G. Hudgens
Department of Biostatistics,
University of North Carolina at Chapel Hill

and

Sonia Napravnik
Division of Infectious Diseases, Department of Medicine,
University of North Carolina at Chapel Hill

and

Michael R. Kosorok
Department of Biostatistics,
University of North Carolina at Chapel Hill

November 14, 2019

*The authors gratefully acknowledge the following funding sources: the first author was funded in part
Abstract

Individualized treatment rules can lead to better health outcomes when patients have heterogeneous responses to treatment. Very few individualized treatment rule estimation methods are compatible with a multi-treatment observational study with right censored survival outcomes. In this paper we extend policy evaluation methods to the right censored data setting. Existing approaches either make restrictive assumptions about the structure of the data, or use inverse weighting methods that increase the variance of the estimator resulting in decreased performance. We propose a method which uses balanced policy evaluation combined with an imputation approach to remove right censoring. We show that the proposed imputation approach is compatible with a large number of existing survival models and can be used to extend any individualized treatment rule estimation method to the right censored data setting. We establish the rate at which the imputed values converge to the conditional expected survival times, as well as consistency guarantees and regret bounds for the combined balanced policy with imputation approach. In simulation studies, we demonstrate the improved performance of our approach compared to existing methods. We also apply our method to data from the University of North Carolina Center for AIDS Research HIV Clinical Cohort.

Keywords: Survival data, Individualized treatment rule, Nonparametric estimation.
1 Introduction

Differences in individual patient characteristics can result in significant heterogeneity in response to treatment. An individualized treatment rule (ITR) is a function which takes patient specific characteristics and recommends a treatment. The optimal ITR recommends the treatment that maximizes the benefit with respect to some clinical outcome. Treatment decisions are made based on published treatment guidelines which often list several available treatments from which the physician needs to choose based on expert opinion and personal experience. Using data driven approaches to help inform decision making can formalize the process and improve patient outcomes. There has been a considerable amount of effort devoted to the estimation of the optimal ITR (Qian & Murphy 2011, Zhang et al. 2012, Zhao et al. 2012, 2014, Cui et al. 2017, Kallus 2018, Athey & Wager 2017).

To motivate our approach to estimating ITRs, we consider data from an observational cohort study of HIV+ patients who take different types of antiretroviral therapy (ART). The nature of this data presents several challenges to estimating ITRs. First, there are many treatment options. HIV infection is treated with combination therapies consisting of several drugs from multiple drug classes. Second, the data are frequently subject to right censoring. One measure of the effectiveness of treatment is the durability of an ART regimen, which measures how long a patient remains on the same ART. The durability is predictive of long-term patient morbidity and mortality, but it is prone to loss to follow-up. Third, the observational nature of the data means that the probability of a patient receiving a particular treatment may be related to measured and unmeasured factors including comorbid conditions and other therapies. This confounding by indication can make it difficult to establish causal relationships.

Some early methods for estimating ITRs involved a two step approach that first uses regression methodology to estimate a conditional mean for the response under each treatment and then recommends the treatment with the best estimated conditional mean value (Qian & Murphy 2011). Since these methods are based on regression modeling, they can be easily extended to meet the specific requirements of a wide array of data sources. However, there are a number of concerns about the performance of these methods (Beygelzimer & Langford 2009, Zhao et al. 2012). Much of the recent work in ITR estimation is based on outcome weighted learning (OWL), which reformulates the estimation procedure as weighted classification (Zhao et al. 2012). The original paper assumed fully observed, binary treatment data from a clinical trial. Several extensions have relaxed each of these assumptions individually (Zhao et al. 2014, Wang et al. 2016, Cui et al. 2017, Liang et al. 2018), but none of them relax all three assumptions simultaneously. Policy evaluation methods are designed to work with multiple treatments and observational data (Dudík et al. 2011, Kallus 2018, Athey & Wager 2017). However, current extensions of these methods to right censored data use multiple weighting approaches, which results in high variance estimators, or rely on strong assumptions.

In this article, we propose a new method which addresses the shortcomings in existing approaches with respect to our motivating example. We focus on balanced policy evaluation and learning, an approach proposed by Kallus (2018). Weighting approaches to policy evaluation can suffer from high variance when the weights are large. Existing vari-
ance reduction methods result in a biased estimator, but the amount of bias introduced is unknown, making it difficult to optimize the bias-variance tradeoff. Balanced policy evaluation seeks to quantify the bias-variance tradeoff nonparametrically by finding weights which minimize a measure of the conditional mean squared error (CMSE). We extend the balanced policy approach to right censored data by developing an imputation approach which replaces the censored observations with the conditional expected mean survival time. This allows us to define weighted and doubly robust estimators for policy evaluation for two or more treatments under right censoring. By more explicitly quantifying the bias-variance tradeoff and reducing the impact of censoring via imputation, the proposed approach is able to reduce the variance, leading to more efficient policy evaluation and learning.

While we focus on a single policy evaluation method, the imputation approach we propose could be used to extend any ITR estimation method to work with right censored data because it results in an estimated fully observed outcome vector. We demonstrate that the proposed approach results in improved performance when compared to existing methods. Nonetheless, the performance of balanced policy evaluation and learning can suffer in the presence of high dimensional, potentially noisy covariate information. We discuss a number of ways to mitigate this issue in the presence of right censored data, and demonstrate their use in simulation studies.

The remainder of this article is organized as follows. In Section 2, we discuss the policy evaluation approach to estimating ITRs, and introduce an imputation approach for right censored data. In Section 3, we develop the theory for the proposed method. Simulation studies are presented in Section 4. We also illustrate our method using data from the University of North Carolina Center for AIDS Research HIV Clinical Cohort in Section 5. The article concludes with a discussion of future work in Section 6. Some technical results are provided in the Appendix.

## 2 Methods

### 2.1 Problem Setup and Notation

Before discussing policy evaluation in the presence of right censoring, we first introduce some notation and describe policy evaluation and learning adapted to the fully observed survival time setting. Let $X \in \mathcal{X}$ be the observed subject-level covariate vector, where $\mathcal{X}$ is a $d$-dimensional vector space, and $A$ is the treatment, where $A \in \mathcal{A} = [m] = \{1, \ldots, m\}$ comes from a finite action space. For each $a \in \mathcal{A}$ define $\tilde{T}(a) \in [0, \infty)$ as the unrestricted failure time that would be observed under treatment $a$, and let $\hat{T} = (\tilde{T}(1), \ldots, \tilde{T}(m))$. In many studies, the unrestricted failure time $\tilde{T}$ is rarely observable due to the existence of a maximum follow-up time $\tau < \infty$. We instead consider $T = (T(1), \ldots, T(m))$, a truncated version of $\hat{T}$ at $\tau$, i.e., $T(a) = \min(\tilde{T}(a), \tau)$ for $a \in [m]$. For a covariate vector $x$, treatment $a$, and failure time $T(a)$ we define the mean-outcome function as $\mu_a(x) = E[T(a) \mid X = x]$.

In most survival studies there are some subjects for which observation of the failure is precluded by the occurrence of censoring events. For any treatment $a$, consider a censoring time $C(a)$ which is independent of $T(a)$ given $X$ and $A$. Define the observed time
\[ Y(a) = \min(T(a), C(a)) \] and the failure indicator \( \Delta = I(T(a) \leq C(a)) \). Let the failure time \( T \) and censoring time \( C \) be generated from distributions with conditional survival functions \( S(t \mid x, a) \) and \( G(c \mid x, a) \), respectively. The censored data consist of \( n \) observations of covariate, treatment, observed time, and failure indicator: \( \mathcal{D}_n = \{(X_i, A_i, Y_i(A_i), \Delta_i)\}_{i=1}^n \). We assume that the data are drawn iid from a fixed distribution: \((X, A, T, C) \sim P\), and that we observe \((X, A, Y(A))\). This definition of \((X, A, Y(A))\) implies the causal consistency assumption of the potential outcomes framework and the stable unit treatment value assumption (SUTVA), which ensures that the observed outcome for one subject is unaffected by the treatment assignments of the other subjects (Rubin 1986). We also assume that the data satisfy conditional exchangeability:

**Assumption 1.** \( T(a) \perp A \mid X \) for all \( a \in \mathcal{A} \).

This is commonly referred to as the no unmeasured confounders assumption, where the treatment decision is not influenced by the outcomes except through their mutual relationship with the observed covariates \( X \). In our setting, conditional exchangeability is equivalent to there being an unknown propensity function \( \varphi \) which, given covariate vector \( x \), assigns treatment \( a \) with probability \( \varphi_a(x) = P(A = a \mid X = x) \).

A policy is a map \( \pi : \mathcal{X} \rightarrow \Theta^m \) from the covariate space to a probability vector in the \( m \)-simplex, \( \Theta^m = \{p \in [0, 1]^m : \sum_{a=1}^m p_a = 1\} \). For an observation with covariate vector \( x \), the policy \( \pi \) specifies that treatment \( a \) should be administered with probability \( \pi_a(x) \).

There are two main goals in this framework. The first goal, known as policy evaluation, is to estimate the average outcome that would be observed if treatments were assigned according to the policy \( \pi \). Because the implementation of a bad policy can be costly or dangerous, new policies are generally evaluated using historical observations (Thomas et al. 2015). The second goal, known as policy learning, attempts to identify a policy \( \pi^\ast \) which maximizes the expected outcome with respect to some reward. In this article we will assume that larger outcomes are preferable, and thus the reward would be \( T \) or \( \log(T) \).

To formally define policy evaluation, consider a policy \( \pi \) for which we wish to estimate the sample-average policy effect (SAPE),

\[
\text{SAPE}(\pi) = \frac{1}{n} \sum_{i=1}^n \sum_{a=1}^m \pi_a(X_i) \mu_a(X_i).
\]

The SAPE quantifies the average outcome that would have been observed had the treatments been assigned according to a given policy \( \pi \). The SAPE is strongly consistent for the population-average policy effect (PAPE), \( \text{PAPE}(\pi) = \mathbb{E}[\text{SAPE}(\pi)] = \mathbb{E}[\sum_{a=1}^m \pi_a(X_i) \mu_a(X_i)] \). If \( \pi^\ast \) is such that \( \pi^\ast_a(x) > 0 \iff a = \arg \max_{s \in [m]} \mu_s(x) \), then \( \pi^\ast \) is an optimal policy. That is, \( \pi^\ast \) maximizes \( \text{SAPE}(\pi) \) over all functions \( \mathcal{X} \rightarrow \Theta^m \). For an optimal policy \( \pi^\ast \), \( R(\pi) = \text{SAPE}(\pi^\ast) - \text{SAPE}(\pi) \) is the regret of \( \pi \). In policy learning, we wish to find a policy \( \hat{\pi} \) that minimizes the expected regret.

### 2.2 Existing Approaches to Policy Evaluation

Current approaches to policy evaluation generally fall into a few categories which are based on mean outcome modeling, weighting methods, or a combination thereof. The regression-
based estimator first fits regression models \( \hat{\mu}_a \) of \( \mu_a \) for all \( a \in [m] \). These models are then used to estimate the SAPE in a plug-in fashion

\[
\psi_{\text{Reg}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{m} \pi_a(X_i) \hat{\mu}_a(X_i).
\]  

(1)

This approach assumes that the regression models for each treatment are correctly specified, which can be a strong assumption when the number of treatments is large (Zhao et al. 2012).

Weighting-based estimators use covariate and treatment data to find weights, \( W(\pi, X_{1:n}, T_{1:n}) \), that reweight the outcome data to make it look as though it were generated by the policy being evaluated. These estimators are generally of the form

\[
\hat{\psi}_W(\pi) = \frac{1}{n} \sum_{i=1}^{n} W_i T_i,
\]  

(2)

where the weights are often normalized to sum to \( n \). With weights \( W \) and regression models \( \hat{\mu}_a \) we can define a doubly robust (DR) estimator

\[
\hat{\psi}_{W,\hat{\mu}}^{\text{DR}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{m} \pi_a(X_i) \hat{\mu}_a(X_i) + \frac{1}{n} \sum_{i=1}^{n} W_i (T_i - \hat{\mu}_{A_i}(X_i)),
\]  

(3)

which can be thought of as denoising the weighted estimator by subtracting the conditional mean from \( T \), or debiasing the regression-based estimator using the reweighted residuals.

A common weighting-based approach is inverse propensity weighting (IPW) (Imbens 2000). By noting that \( \text{SAPE}(\pi) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \pi_{A_i} X_i | A_i = 1 \right] \), for \( \hat{\varphi} \), an estimate of the unknown propensity function \( \varphi \), one can construct weights

\[
W_{i}^{\text{IPW}}(\pi) = \pi_{A_i} / \hat{\varphi}_{A_i}(X_i),
\]  

(4)

which gives rise to the estimator \( \hat{\psi}^{\text{IPW}} = \hat{\psi}_{W,\hat{\mu}}^{\text{IPW}}(\pi) \). Since the propensities appear in the denominator, small values of \( \hat{\varphi} \) can lead \( \hat{\psi}^{\text{IPW}} \) to have a large variance. Some ad-hoc solutions for reducing the variance exist, such as clipping, where \( \hat{\varphi}_{A_i}(X_i) \) is replaced by \( \max\{M, \hat{\varphi}_{A_i}(X_i)\} \) for some \( M \), but these methods generally have the effect of increasing the bias (Li et al. 2015).

Kallus (2018) proposed a weighting-based estimator which simultaneously controls both the bias and variance. For a generic weighted estimator \( \hat{\psi}_W(\pi) \), consider the conditional mean squared error (CMSE),

\[
\text{CMSE}(\hat{\psi}, \pi) = \mathbb{E} \left[ (\hat{\psi} - \text{SAPE}(\pi))^2 \right] | X_{1:n}, A_{1:n}.
\]  

(5)

Balanced policy evaluation takes the CMSE and decomposes it into squared bias and variance components. The bias term relies on the weights \( W \) and the unknown mean outcome function \( \mu \). Rather than directly estimating \( \mu \), the balanced policy approach defines a worst case bias by choosing the mean outcome function from a bounded class of functions that maximizes the squared bias. Once the worst case bias term has been identified, the variance term is estimated, and balanced policy weights are found by minimizing the worst case squared bias and variance with respect to the weights \( W \). Additional details for balanced policy evaluation will be provided in Section 2.4.
2.3 Extensions to Right Censored Data

Analyzing right censored data without accounting for the missing data results in bias due to censoring (Fleming & Harrington 2011), thus consistent policy evaluation requires methods designed to correct for the bias due to censoring.

Regression-based methods can be extended to right censored data by using regression models designed for censored data. Common model choices include parametric and semiparametric models such as the accelerated failure time and Cox proportional hazards models (Kalbfleisch & Prentice 2011). However, these models make restrictive assumptions about the structure of the failure time distributions which are often difficult to verify. Non-parametric models of the survival function have been proposed, such as the kernel conditional Kaplan-Meier estimate (Dabrowska 1987) or random forest based methods (Hothorn et al. 2004, Ishwaran et al. 2008, Zhu & Kosorok 2012, Steingrimsson et al. 2016, Cui et al. 2019) which define localized versions of simple survival models, such as the Kaplan-Meier estimator. Non-parametric models make less restrictive assumptions, but they are less efficient and unlikely to produce good results in small sample sizes.

Inverse propensity weighted methods can be extended to right censored data by looking only at the observed failure times, and weighting based on the probability of observation. Inverse probability of censoring weighting (IPCW), is motivated by an idea similar to IPW. Since $E[\Delta | X, A, T] = G(T | X, A)$ we have that $E[Y \Delta / G(Y | X, A) | X, A] = E[T | X, A]$.

IPCW can be combined with IPW by defining weights

$$W_{i}^{IPW,IPCW}(\pi) = \frac{\pi_{A_i}(X_i)\Delta_i}{\hat{\varphi}_{A_i}(X_i)\hat{G}(Y_i | X_i, A_i)}.$$ (6)

As long as the models for $\hat{\varphi}$ and $\hat{G}$ are correctly specified, the resulting estimator, $\psi_{IPW,IPCW}^{IPW,IPCW}(\pi)$, will be consistent for the SAPE (Anstrom & Tsiatis 2001). However, with two weights in the denominator, the issues with high variance will be compounded.

Balanced policy weights are incompatible with IPCW. Depending on when the IPC weights are applied, balanced policy weights are either no longer identifiable, or they lose the property of minimizing the worst case bias and variance. To extend balanced policy to right censored data, we use an imputation approach similar to one proposed by Cui et al. (2017). Consider an imputed failure time vector $\tilde{Y}_i \equiv \Delta_i Y_i + (1 - \Delta_i) E[T_i | X_i, A_i, Y_i, T_i > Y_i]$, where the censoring times are replaced with the conditional expected failure time given the covariates, treatment, and the knowledge that the subject survived at least until time $C_i$.

Estimating the expected failure time for the censored subjects uses a conditional estimate of the survival function. In this article we focus on estimation of the conditional expected failure time using random forest methods for right censored data because they provide nonparametric estimates with good performance for higher dimensional data (Cui et al. 2019). Random survival forests (RSF) (Ishwaran et al. 2008) and recursively imputed survival trees (RIST) (Zhu & Kosorok 2012) are based on extensions of random forests (RF) (Breiman 2001) and extremely randomized trees (ERT) (Geurts et al. 2006), respectively, where the splitting rule for each node is chosen to maximize the log-rank test statistic. Cui et al. (2019) note that a splitting rule based on the standard log-rank test statistic can pro-
duce biased results because it does not appropriately control for the censoring distribution, and instead propose a splitting rule which corrects for the bias due to censoring.

Regardless of the method used to estimate the survival function, given $\hat{S}(\cdot | X, A)$, the conditional expected survival time can be estimated as

$$
\hat{E}[T_i | X_i, A_i, Y_i, T_i > Y_i] = \int_{Y_i}^{\tau} t \, d\hat{F}(t | X, A) + \tau \hat{S}(\tau | X, A) \hat{S}(Y | X, A),
$$

and the estimated imputed failure time vector can be defined as

$$
\hat{Y}_i = \Delta_i Y_i + (1 - \Delta_i) \hat{E}[T_i | X_i, A_i, Y_i, T_i > Y_i].
$$

By using $\hat{Y}$ in place of $T$ in equation (2) we can now define a balanced policy approach to right censored data.

### 2.4 Balanced Policy Evaluation with Right Censored Data

Recall from equation (5) that we defined the CMSE as the squared difference between a weighted estimator $\hat{\psi}_W(\pi)$ and the SAPE($\pi$). We define

$$
B_a(W, \pi_a; \mu_a) = \frac{1}{n} \sum_{i=1}^{n} (W_i \delta_{A_i a} - \pi_a(X_i)) \mu_a(X_i)
$$

and

$$
B(W, \pi; \mu) = \sum_{a=1}^{m} B_a(W, \pi_a; \mu_a),
$$

where $\delta_{ij} = I[i = j]$ is the Kronecker delta. Letting $\varepsilon_i = T_i - \mu_a(X_i)$, from Kallus (2018) Theorem 1, we have that

$$
\hat{\psi}_W - \text{SAPE}(\pi) = B(W, \pi; \mu) + \frac{1}{n} \sum_{i=1}^{n} W_i \varepsilon_i,
$$

and, under assumption 1,

$$
\text{CMSE}(\hat{\psi}_W, \pi) = B^2(W, \pi; \mu) + \frac{1}{n^2} W^T \Sigma W,
$$

where $\Sigma = \text{diag}(\mathbb{E}[\varepsilon_i^2 | X_i, A_i]_{i=1}^n)$. This decomposes the CMSE into squared bias and variance components. The measure of the variance will be given by the norm of weights $W^T \Lambda W$ for a positive semidefinite (PSD) matrix $\Lambda$. Instead of directly estimating the bias term, we find the worst case bias by maximizing $B^2(W, \pi; f)$ over all functions $f \in \mathcal{F}$ for a bounded function class $\mathcal{F}$.

To define the worst case bias, we look at functions, $f$, in the unit ball of a direct product of reproducing kernel Hilbert spaces (RKHS):

$$
\|f\|_{p, K_1:m, \gamma_1:m} = \left( \sum_{a=1}^{m} \|f_a\|_{K_a}^p / \gamma_a^p \right)^{1/p},
$$

8
where \( \| \cdot \|_{k_a} \) is the norm of the RKHS given by the PSD kernel \( K(\cdot, \cdot) : \mathcal{X}^2 \rightarrow \mathbb{R} \). Any choice of \( \| \cdot \| \) gives rise to the worst case CMSE objective:

\[
\mathcal{E}^2(W, \pi; \| \cdot \|, \Lambda) = \sup_{\| f \| \leq 1} B^2(W, \pi; f) + \frac{1}{n^2} W^T \Lambda W.
\]

Balanced policy evaluation is given by the estimator \( \hat{\psi}_{W^*}^{\pi} = n \sum_{i=1}^{n} W^* \hat{Y}_i \) where \( W^* = W^*(\pi; \| \cdot \|, \Lambda) \) is the minimizer of \( \mathcal{E}^2(W, \pi; \| \cdot \|, \Lambda) \) over the space of all weights \( W \) that sum to \( n \), \( W = \{ W \in \mathbb{R}^n_+ : \sum_{i=1}^{n} W_i = n \} = n \Theta^n \). Specifically,

\[
W^*(\pi; \| \cdot \|, \Lambda) \in \arg \min_{W \in W} \mathcal{E}^2(W, \pi; \| \cdot \|, \Lambda).
\] (9)

With the imputed survival time vector, \( \hat{Y} \), and the balanced policy weights, we can define the weighted and doubly robust balanced policy estimators for right censored data:

\[
\hat{\psi}_{W^*, \hat{Y}} = \frac{1}{n} \sum_{i=1}^{n} W^* \hat{Y}_i,
\]

\[
\hat{\psi}_{DR, \hat{Y}} = \frac{1}{n} \sum_{i=1}^{m} \pi_a(X_i) \hat{\mu}_a(X_i) + \frac{1}{n} \sum_{i=1}^{n} W^*_i (\hat{Y}_i - \hat{\mu}_A(X_i)).
\] (11)

Next we consider policy learning. For a given policy class \( \Pi \subset [\mathcal{X} \mapsto \Theta^m] \), let \( \hat{\pi} \in \Pi \) be the policy which maximizes the balanced policy evaluation estimator, \( \hat{\phi}_W \), using the weights defined in equation (9). We formulate this as a bilevel optimization problem:

\[
\hat{\pi}^{\text{balanced}} \in \arg \max_{\pi \in \Pi} \left\{ \hat{\psi}_W : \pi \in \Pi, W \in \arg \min_{W \in W} \mathcal{E}^2(W, \pi; \| \cdot \|, \Lambda) \right\},
\] (12)

\[
\hat{\pi}^{\text{balanced-DR}} \in \arg \max_{\pi \in \Pi} \left\{ \hat{\psi}_{W, \hat{Y}} : \pi \in \Pi, W \in \arg \min_{W \in W} \mathcal{E}^2(W, \pi; \| \cdot \|, \Lambda) \right\}.
\] (13)

It is difficult to maximize the policy \( \pi \) with respect to the full policy class \( \Pi = [\mathcal{X} \mapsto \Theta^m] \), so it is common to use a reduced class such as the parameterized policy class

\[
\Pi_{\text{logit}} = \{ \pi_a(x; \beta_a) \propto \exp(\beta_{a0} + \beta_a^T x) \}.
\] (14)

Using a reduced policy class limits the flexibility of balanced policy learning, but for moderate sample sizes the reduced complexity can allow for more accurate estimation of \( \pi^* \).

A high-level description of the proposed method is given in Algorithm 1 below.

**Algorithm 1.** Pseudo algorithm for the Balanced Policy Evaluation and Learning with Imputation

**Step 1:** Use \( \{(X_i, A_i, Y_i, \Delta_i)\}_{i=1}^{n} \) to fit a model for the conditional survival function \( S(t | X, A) \), and create the imputed survival time vector, \( \hat{Y} \), by using \( \hat{S}(t | X, A) \) to estimate the conditional expected survival times \( \hat{E}[T_i | X_i, A_i, Y_i, T_i > Y_i] \) for the censored observations.
Step 2: For a norm, $\| \cdot \|_{p, K_{1:m}, \gamma_{1:m}}$, and diagonal matrix, $\Lambda$, define a policy class $\Pi$, and select $\pi_0 \in \Pi$.

Step 3: Find the balanced policy weights associated with $\| \cdot \|_{p, K_{1:m}, \gamma_{1:m}}$, $\Lambda$, and $\pi_i$ and use the weights $W^*$ to calculate $\hat{\psi}_{W^*}(\pi_i)$ according to equation (10).

Step 4: Calculate the gradient of the policy, $\pi_i$, with respect to $\hat{\psi}_{W^*}(\pi_i)$ and use gradient ascent to find an improved policy $\pi_{i+1} \in \Pi$.

Step 5: Repeat steps 3-4 until convergence to an estimated optimal policy $\hat{\pi}$.

2.5 High Dimensional Considerations

Because the balanced policy weights are defined using a kernel norm $\| \cdot \|_K$, which depends only on the sample covariate space $X$, balanced policy evaluation and learning can be sensitive to the presence of covariates which contain redundant or noisy information. The negative effects of high-dimensional noisy data can be reduced in several ways including the choice of the kernel norm, and feature elimination, both of which we detail here.

Recall from equation (8) that we defined the norm of a function with respect to the direct product of RKHSs. One commonly used kernel is the Mahalanobis radial basis function (RBF) kernel

$$K_s(x, x') = \exp(-(x - x')^T \tilde{\Sigma}^{-1} (x - x') / s^2).$$

(15)

In Lemma 1 point (3) of Kallus (2018), it was shown that, for the norm defined in (8), if $p = 2$ and $f_a$ has a Gaussian process prior, then kernel hyperparameters, such as $s$ and $\tilde{\Sigma}$, can be selected using the marginal likelihood principle. Choosing the hyperparameters in this manner reduces the influence of the unimportant predictors. In simulations this has greatly improved the accuracy of the balanced policy estimator with respect to evaluation. However, since the redundant or noisy predictors are not eliminated, the dimension of the search space remains high, which can still adversely impact policy learning.

Reducing the dimension of the search space using variable selection (VS) methods can improve the performance of policy learning. Nearly any VS method could be used, but to keep our method as general as possible we examined non-parametric VS methods such as the risk-recursive feature elimination algorithm (risk-RFE) proposed by Dasgupta et al. (2019). Risk-RFE uses kernel machines to rank the importance of the features based on the regularized risk. Once the features are ranked in order of estimated importance, change point estimation methods can be used to identify the important predictors based on the increase in regularized risk at each step. For additional details see Dasgupta et al. (2019).

In our implementation for high dimensional data, we used methods where tuning hyperparameters and variable selection required a complete outcome vector. Therefore, these methods were implemented prior to Step 3 of Algorithm 1, but methods which are compatible with right censoring could be implemented prior to Step 1.
3 Theory

The properties of the proposed estimator will depend on the method used to estimate $\hat{Y}$, so to facilitate the discussion we make the following assumption:

**Assumption 2** (Convergence rate). There exists some sequence $r_n \to 0$ such that

$$\sup_{t < \tau} E_{X,A} \left| \hat{S}_n(t \mid X, A) - S(t \mid X, A) \right| = O_p(r_n).$$

Assumption 2 may seem strong, but it is met for a large number of methods for estimating $S(t \mid X, A)$. For example, the assumption holds for AFT and Cox PH models with $r_n = 1/\sqrt{n}$. Assumption 2 is also met for non-parametric methods, such as kernel conditional Kaplan-Meier and random forest based methods. For the kernel conditional Kaplan-Meier, $r_n = \log(n^{d/(d+4)}/n^{2/(d+4)})^{1/2}/n^{2/(d+4)}$, where $d$ is the covariate dimension (Dabrowska 1989). The convergence rate for the bias corrected random survival forest method proposed by Cui et al. (2019) depends on the splitting criteria, but the rate for the theoretically optimal splitting criteria is $r_n = n^{-1/(d+2)}$. Convergence rates for other random forest methods (Ishwaran et al. 2008, Zhu & Kosorok 2012) have not yet been established, but we believe that the rates will be similar to those found in Cui et al. (2019).

We show that a convergence rate for $\hat{S}_n(t \mid X, A)$ is enough to guarantee the convergence rate of $\hat{Y}$; however, Theorems 12.4 and 12.5 of Kosorok (2008) imply that it is also sufficient to have a convergence rate for an estimate of the cumulative hazard function, $\hat{\Lambda}(t \mid X, A)$.

**Lemma 1.** If there exists some sequence $r_n \to 0$ such that,

$$\sup_{t < \tau} E_{X,A} \left| \hat{S}(t \mid X, A) - S(t \mid X, A) \right| = O_p(r_n)$$

then

$$\left| \frac{1}{n} \sum_{i=1}^n (1 - \Delta_i) W_i^*(\hat{Y}_i - \tilde{Y}_i) \right| = O_p(r_n).$$

where $W^* = W^*(\pi; \| \cdot \|, \Lambda)$ are the balanced policy weights.

To prove Lemma 1 we need an additional assumption.

**Assumption 3.** Let $\hat{Y}_i^{(i)} = \hat{E}_{n-1}^{(i)}(T \mid X_i, A_i, Y_i, \Delta_i)$, where $\hat{E}_{n-1}^{(i)}$ is based on all observations except for the $i^{th}$ one. For all $n \geq 1$ and any $1 \leq i \leq n$,

$$E \left[ (1 - \Delta_i)(\hat{Y}_i - \tilde{Y}_i)^2 \right] \leq E \left[ (1 - \Delta_i)(\hat{Y}_i^{(i)} - \hat{Y}_i)^2 \right].$$

Since $\hat{Y}_i$ is based on more information than $\hat{Y}_i^{(i)}$, this assumption should be met provided $\hat{E}_{n}$ is consistent. The proof of Lemma 1 can be found in the appendix.

Consistent evaluation requires a weak form of overlap between the unknown propensity function and the target policy being evaluated:
Assumption 4 (Weak overlap). \( P(\varphi_a(X) > 0 \lor \pi_a(X) = 0) = 1 \) for all \( a \in [m] \), and \( E[\pi^2_A(X)/\varphi^2_A(X)] < \infty \).

Another requirement is well-specification of the mean-outcome function. For balanced policy evaluation the mean-outcome function is well-specified if it is in the RKHS product used to compute \( W^*(\pi) \). Otherwise, consistency is also guaranteed if the RKHS product consists of \( C_0 \)-universal kernels, defined below, such as the RBF kernel.

Definition 1. A PSD kernel \( K \) on a Hausdorff \( X \) (e.g., \( \mathbb{R}^d \)) is \( C_0 \)-universal if, for any continuous function \( g : X \rightarrow \mathbb{R} \) with compact support (i.e., for some compact \( C, \{ x : g(x) = 0 \subseteq C \}) \) and \( \eta > 0 \), there exists \( n', \alpha_1, x_1, \ldots, \alpha_n', x_n' \) such that \( \sup_{x \in X} \left| \sum_{j=1}^{n'} \alpha_j K(x_j, x) - g(x) \right| \leq \eta \).

Theorem 1. Fix \( \pi \) and let \( W_n^*(\pi) = W_n^*(\pi; f\|_p, K_{\alpha_1, \ldots, \alpha_n}, \Lambda_n) \) with \( 0 < \kappa I \leq \Lambda_n \leq \bar{\kappa} I \), \( 0 < \gamma \leq \gamma_{n,a} \leq \bar{\gamma} \) \( \forall a \in [m] \) for each \( n \). Suppose Assumptions 1-4 hold, \( \text{Var}(T|X) \) is a.s. bounded, \( \mathbb{E}[\sqrt{K_a(X, X)}] < \infty \), and \( \mathbb{E}[K_a(X, X)\pi^2_A(X)/\varphi^2_A(X)] < \infty \). Then the following two results hold:

- If \( \|\mu_a\|_{K_a} < \infty \) for all \( a \in [m] \), then \( \hat{\psi}_{W_n^*(\pi), \hat{Y}} - \text{SAPE}(\pi) = O_p(r_n + 1/\sqrt{n}) \).
- If \( K_a \) is \( C_0 \)-universal for all \( a \in [m] \), then \( \hat{\psi}_{W_n^*(\pi), \hat{Y}} - \text{SAPE}(\pi) = o_p(1) \).

Corollary 2. Suppose the assumptions of Theorem 1 hold, then

- If \( \|\mu_a\|_{K_a} < \infty \) and \( \|\bar{\mu}_{na}\|_{K_a} = O_p(1) \) for all \( a \in [m] \), then \( \hat{\psi}_{W_n^*(\pi), \bar{\mu}, \hat{Y}} - \text{SAPE}(\pi) = O_p(r_n + 1/\sqrt{n}) \).
- If \( K_a \) is \( C_0 \)-universal for all \( a \in [m] \), then \( \hat{\psi}_{W_n^*(\pi), \bar{\mu}, \hat{Y}} - \text{SAPE}(\pi) = o_p(1) \).

Proof of Theorem 1. Define

\[ \hat{\psi}_{W_n^*(\pi), T} = \frac{1}{n} \sum_{i=1}^{n} \frac{W_n^*(\pi)T_i}{\alpha_i}, \quad \text{and} \quad \hat{\psi}_{W_n^*(\pi), \hat{Y}} = \frac{1}{n} \sum_{i=1}^{n} W_n^*(\pi)\hat{Y}_i, \]

as the estimators which use the true and estimated imputed failure time vectors respectively. We can then decompose \( \hat{\psi}_{W_n^*(\pi), \hat{Y}} - \text{SAPE}(\pi) \) as follows:

\[ \hat{\psi}_{W_n^*(\pi), \hat{Y}} - \text{SAPE}(\pi) = \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i)W_i(\hat{Y}_i - T_i) + \hat{\psi}_{W_n^*(\pi), T} - \text{SAPE}(\pi). \]

By Lemma 1 we have that

\[ \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i)W_i(\hat{Y}_i - T_i) = O_p(r_n). \]
If \( \|\mu_a\|_{\mathcal{K}_a} < \infty \) for all \( a \in [m] \), then by Kallus (2018) Theorem 3, we have that 
\[
\hat{\psi}_{W_n^*(\pi), T} - \text{SAPE}(\pi) = O_p(1/\sqrt{n}),
\]
and thus 
\[
\hat{\psi}_{W_n^*(\pi), \hat{F}} - \text{SAPE}(\pi) = O_p(r_n + 1/\sqrt{n}).
\]

If \( \mathcal{K}_a \) is \( C_0 \)-universal for all \( a \in [m] \), then by Kallus (2018) Theorem 3, we have that 
\[
\hat{\psi}_{W_n^*(\pi), T} - \text{SAPE}(\pi) = o_p(1),
\]
and thus 
\[
\hat{\psi}_{W_n^*(\pi), \hat{F}} - \text{SAPE}(\pi) = o_p(1).
\]

The proof of Corollary 2 follows immediately from Lemma 1, as well as Corollary 4 of Kallus (2018) in a manner similar to the proof of Theorem 1.

Next, we establish the consistency and learning rates of the balanced policy learner uniformly over policy classes. For a given policy class \( \Pi \), we define the sample regret as
\[
\hat{R}_n(\hat{\pi}) = \max_{\pi \in \Pi} \text{SAPE}(\pi) - \text{SAPE}(\hat{\pi}),
\]
and the population regret as 
\[
R_n(\hat{\pi}) = \max_{\pi \in \Pi} \text{PAPE}(\pi) - \text{PAPE}(\hat{\pi}).
\]
Convergence of these regret quantities requires that the best-in-class policy is learnable. We quantify learnability using Rademacher complexity, but the results could be extended to VC dimension. Let us define 
\[
\hat{\mathcal{R}}_n(\mathcal{F}) = \frac{1}{2^n} \sum_{\rho_i \in \{-1, 1\}^n} \rho_i f(X_i), \quad \mathcal{R}_n(\mathcal{F}) = \mathbb{E}[\hat{\mathcal{R}}_n(\mathcal{F})].
\]
If \( \mathcal{F} \subseteq [\mathcal{X} \to \mathbb{R}^m] \), let \( \mathcal{F}_a = \{(f(\cdot))_a : f \in \mathcal{F}\} \) and set \( \mathcal{R}_n(\mathcal{F}) = \sum_{a=1}^m \mathcal{R}_n(\mathcal{F}_a) \) and \( \hat{\mathcal{R}}_n(\mathcal{F}) = \sum_{a=1}^m \hat{\mathcal{R}}_n(\mathcal{F}_a) \). We also must use a stronger version of the overlap assumption.

**Assumption 5** (Strong overlap). \( \exists \alpha \geq 1 \) such that \( \mathbb{P}(\varphi_a(X) > 1/\alpha) = 1 \) \( \forall a \in [m] \).

**Theorem 3.** Fix \( \Pi \subseteq [\mathcal{X} \to \Theta^n] \) and let \( W_n^*(\pi) = W_n^*(\pi; \|f\|_{\mathcal{K}_1 \cup \mathcal{K}_2}, \gamma_n, m, \lambda_n) \) with \( 0 < \gamma_n \leq \lambda_n \leq \pi I, 0 < \gamma \leq \gamma_n, \lambda_n, \Lambda_n \subseteq \pi I, 0 < \gamma_n, \lambda_n \leq \pi I \) \( \forall a \in [m] \), \( n \), and \( \pi \in \Pi \). Suppose assumptions 1-2 and 5 hold, \( |\epsilon_i| \leq B \) a.s. bounded, and \( \sqrt{\mathcal{K}_a(X, X)} \leq \Gamma \) \( \forall a \in [m] \) for \( \Gamma \geq 1 \). If \( \hat{\pi}_n^{\text{balanced}} \) is as in (12), then the following results hold:
- If \( \|\mu_a\|_{\mathcal{K}_a} < \infty \) for all \( a \in [m] \), then 
  \[ R_n(\hat{\pi}_n^{\text{balanced}}) = O_p(\mathcal{R}_n(\Pi) + r_n + 1/\sqrt{n}), \]
- If \( \mathcal{K}_a \) is \( C_0 \)-universal for all \( a \in [m] \), then 
  \[ R_n(\hat{\pi}_n^{\text{balanced}}) = o_p(1). \]

If \( \hat{\pi}_n^{\text{balanced-DR}} \) is as in (13), then the following results hold:
- If \( \|\mu_a\|_{\mathcal{K}_a} < \infty \) and \( \|\mu_{na}\|_{\mathcal{K}_a} = O_p(1) \) for all \( a \in [m] \), then 
  \[ R_n(\hat{\pi}_n^{\text{balanced-DR}}) = O_p(\mathcal{R}_n(\Pi) + r_n + 1/\sqrt{n}), \]
- If \( \mathcal{K}_a \) is \( C_0 \)-universal for all \( a \in [m] \), then 
  \[ R_n(\hat{\pi}_n^{\text{balanced-DR}}) = o_p(1). \]

The proof of Theorem 3 follows directly from Assumption 2 and Lemma 1, as well as Corollary 7 of Kallus (2018). All the same results hold when replacing \( \mathcal{R}_n(\Pi) \) with \( \hat{\mathcal{R}}_n(\Pi) \) and/or replacing \( R_n \) with \( \hat{R}_n \).
4 Simulation Studies

We have conducted simulation studies to assess the performance of the proposed method in comparison with existing alternatives. The models we compare are the simple weighted and doubly robust versions of the balanced policy approach with the imputed failure time vector defined in equation (7), a normalized clipped version of the approach which combines IPW with IPCW as defined in equation (6), and a normalized clipped version of the IPW approach combined with the imputed vector from (7). In all settings, both the optimal and estimated policies are members of the reduced policy class \( \Pi_{\text{logit}} \) defined in (14). The kernel used to define the worst case bias in the balanced policy approach is the Mahalanobis RBF kernel in equation (15), where the hyperparameters are chosen by the marginal likelihood method using Gaussian process regression. The same kernel is used for all of the treatments, \( \| \cdot \|_{\kappa_a} = \| \cdot \|_K \) for all \( a \in [m] \). For each setting, the conditional expected survival times, \( \hat{Y}_i \), are estimated using RIST with the tuning parameter settings suggested in Zhu & Kosorok (2012). Propensity scores are estimated using a correctly specified Gaussian process classifier. Due to different censoring mechanisms, the weights used for IPCW are estimated differently for each setting. For the first setting the censoring weights are estimated using RIST with the recommended tuning parameter settings. For the second setting the censoring weights are estimated using the correctly specified accelerated failure time (AFT) model. A testing dataset with size 10000 is used to approximate the population regret \( \text{PAPE}(\pi^*) - \text{PAPE}(\hat{\pi}) \). For each setting the log survival time is used to estimate the optimal policy on simulated datasets of size \( N = 500, 1000, \) and 2000. Each simulation is repeated 100 times.

4.1 Simulation Settings

Our first setting is a slight modification of the setting found in Kallus (2018). To begin, we sample the covariate vector from a multivariate normal distribution \( X \sim \mathcal{N}(0, \Sigma_d) \), where \( d = 10 \) and \( \Sigma_d \) is compound symmetric covariance with diagonal elements of 1 and off diagonal elements of 0.2. The treatments are assigned with probabilities which depend on the first two elements of \( X \). Specifically \( A|X \sim \text{Multinoulli}(p_1(X), \ldots, p_5(X)) \) where \( p_a(X) \propto N((X_1, X_2) - \bar{X}_a, I_{2 \times 2}) \) for \( a \in [m] \), \( \bar{X}_1 = (0, 0), \bar{X}_2 = (1, 0), \bar{X}_3 = (0, 1), \bar{X}_4 = (-1, 0), \) and \( \bar{X}_5 = (0, -1) \). The unrestricted failure times are drawn from a log-normal distribution where \( \log(\tilde{T}) | A \sim N(\mu_A(X), \sigma^2 = 1) \) where \( \mu_a(x) = \exp(1) - \exp(1 - 1/\|x - \chi_a\|_2) \), and \( \chi_a = (\text{Re}, \text{Im})(e^{-i2\pi a/5}/\sqrt{2}) \) for \( a \in [5] \), \( \epsilon_i \sim N(0, 1) \), and \( (\text{Re}, \text{Im}) \) are the real and imaginary components respectively. The observed failure time \( T = \min(\tau, \tilde{T}) \), where \( \tau = 3.5 \). This mean outcome process results in an optimal treatment policy consisting of 5 equal sized wedges arranged radially about the origin. The censoring times are drawn from a log-normal distribution where \( \log(C) | A \sim N(\mu_C(X), \sigma^2 = 2) \) where \( \mu_C(x) = (5/2) - (1/2) \exp(1 - 1/\|X_1 + X_2\|_2) \). The censoring rate is approximately 45%.

In the second setting, the failure times are drawn from log-normal distributions where the treatment effect has both prognostic and prescriptive elements. The 10-dimensional covariate vector, \( X \), is drawn from a uniform\((-1, 1)\) distribution with a compound symmetric covariance structure by first drawing from a multivariate normal distribution \( \mathcal{N}(0, \Sigma_d) \)
where \( \Sigma_d \) is the same as in setting 1. The CDF transformation was applied to the normally distributed covariates which where then scaled to be within -1 and 1. The treatments are assigned with probabilities which depend on the first three elements of \( X \), where \( A \mid X \sim \text{Multinoulli}(p_1(X), \ldots, p_3(X)) \) where \( p_a(X) \propto N((X_1, X_2, X_3) - \overline{X}_a, I_{3\times3}) \) for \( a \in [3], \overline{X}_1 = (-0.5, -0.5, 0.4), \overline{X}_2 = (0, 0, -0.75), \overline{X}_3 = (0.5, 0.5, 0.4). \) The unrestricted failure times are drawn from a log-normal distribution where \( \log(T) = 0.2 - 0.6X_1 + 0.2X_2 + 0.4X_3 + \mu_0 + \epsilon \), where \( \mu_0 = 0.2X_1 - 0.3X_2, \mu_2 = 0.1 + 0.1X_3, \mu_3 = -0.1 - 0.2X_1 + 0.4X_2 - 0.2X_3, \) and \( \epsilon \sim N(0, \sigma^2 = 1). \) The failure times \( T = \min(\tau, T) \), where \( \tau = 1.5. \) The censoring times are drawn from a log-normal distribution which depends on the assigned treatment, where \( \log(C) = 0.6 - 0.4X_1 + 0.3X_2 + 0.8X_3 + \nu_a + \epsilon \), with \( \nu_1 = -0.1X_1 - 0.2X_2, \nu_2 = 0.1X_2, \nu_3 = -0.1 + 0.3X_2 - 0.4X_3 \) and \( \epsilon \sim N(0, \sigma^2 = 1). \) The censoring rate is about 45%.

### 4.2 Simulation Results

The results for setting 1 can be found in Figure 1. This setting is designed to have mismatch between the unknown propensity function and the optimal policy such that the inverse propensity weights are large, resulting in high variability for the IPW with IPCW, and IPW with Imputation approaches. As expected, the increased variance of the IPW based estimators results in reduced performance. The more stable balanced policy with imputation approach performs better than the other approaches. In this setting the feature elimination method effectively eliminates noise variables, even for small sample sizes, which increases performance over the methods using a higher dimensional feature space. We also examined doubly robust approaches for each of the methods, but the mean model was not accurate enough to improve performance over the simple weighted methods. The results for the doubly robust approaches can be found in the supplemental material.

The results for setting 2 can be found in Figure 2. Again the higher variability of the IPW with IPCW, and IPW with Imputation approaches reduced performance when compared to the balanced policy approach. The doubly robust approaches for the balanced policy with imputation and IPW with Imputation increased performance compared to the simple weighted versions for small sample sizes. We examined the effect of feature elimination in this setting, but the method frequently eliminated important predictors, especially in the smaller sample sizes. The results using feature elimination can be found in the supplemental material.

### 5 Application to HIV Study

We apply the proposed method to analyze data from the UNC CFAR HIV Clinical Cohort (UCHCC). In the UCHCC study, patients were followed from initiation of an integrase strand transfer inhibitor (INSTI) in combination with at least two other antiretroviral (ARV) agents from at least one other ARV therapy class representing current standard of care for initial ART in high income clinical settings. Patients were followed until the first of ART modification or discontinuation, death, loss to follow-up, or administrative censoring. The study data included 957 HIV-infected patients who were 72% male, 62%
black, 29% white, 9% Hispanic or other races/ethnicities. The median age at treatment initiation was 44 years, 50% were men who have sex with men (MSM) and 9% had a history of injection drug use (IDU). At INSTI initiation (baseline), the median CD4 cell count was 514 cells/mm$^3$ (range 9 to 2970.0) and the median HIV RNA level was 1.8 log10 copies/mL (range 1.3 to 7.6). The median number of prior antiretroviral (ARV) compounds was 3 (range 0 to 17) and 29% had no prior ARV experience. The INSTI regimen was chosen by providers and patients based on clinical indication and included in all cases one integrase strand transfer inhibitor (INSTI) and two or more nucleoside reverse-transcriptase inhibitors (NRTIs).

The covariates include age, male (yes, no), race (black, white, or other), MSM (yes, no), IDU (yes, no), baseline CD4 count, baseline viral load (VL), an indicator if baseline RNA is undetectable, the number of prior ARV agents, and an indicator if the patient is treatment naïve. Categorical variables are transformed into dummy variables, resulting in an 11-dimensional covariate vector $X$.

The treatment regimens of interest are Raltegravir (RAL), Elvitegravir (EVG), and Dolutegravir (DTG), each with 2+ NRTIs. The primary outcome of interest is time to the discontinuation of the initial INSTI regimen, which is defined as either a change in the INSTI agent or discontinuing ART for more than two weeks. The compounds under consideration received FDA approval at different times, so the maximum follow-up time is different for each regimen. The left plot in Figure 3 shows the Kaplan-Meier curves for the UCHCC data by treatment over the unrestricted follow-up window. This plot indicates that there may be large treatment differences for the longer follow-up times, but we chose
to limit our analysis to 2.5 years (913 days) of follow-up time. This time period was chosen because it is near the 98th percentile of observed treatment discontinuations for the regimen with the least follow-up time, which ensures sufficient data coverage to support the analysis.

Among the 957 study subjects, 416 (43%) were observed to discontinue treatment, and 319 (33%) were censored due to loss to follow-up during the first 2.5 years. The 222 (23%) patients still being followed after 2.5 years were administratively censored.

We applied balanced policy evaluation and learning to the UCHCC data using all available variables with the kernel defined in (15) and the policy class in (14). The estimated optimal treatment strategy identified by balanced policy evaluation and learning was to treat every patient with DTG and 2+ NRTIs. Each of the three INSTIs evaluated in these analyses have different barriers to resistance evolution, tolerability profiles and dosing frequency, likely affecting the durability of the regimen differentially across different patient groups. Additional analyses stratified by whether patients were ART naive at INSTI initiation or by age at INSTI initiation, supported the findings of the primary analyses, the details of which may be found in the supplemental material.

In order to see how our method performs in comparison to existing methods on realistic data when meaningful subgroup effects are present, we modified the data to equalize baseline differences and highlight heterogeneity. Specifically, a constant scaling factor was applied to the observed times for each of the treatments so that the marginal treatment effects between any two treatments was negligible. Due to the effects of differential censoring rates between treatments, the average policy effect was estimated using the regression-based
For this analysis we used a reduced set of predictors which were identified by riskRFE as being most likely to have meaningful subgroup effects. The predictors were age, treatment naive status, baseline CD4, baseline RNA, and undetectable baseline RNA indicator. The conditional expected survival times, $\hat{Y}$, and the censoring probabilities used for IPCW were estimated using RIST with the tuning parameter settings suggested in Zhu & Kosorok (2012). Propensity scores were estimated using a Gaussian process classifier. The balanced policy evaluation again used the kernel defined in (15) and the policy class in (14).

We compare the balanced policy with imputation approach to several other methods. The regression-based method as defined in equation (1) with $\hat{\mu}$ estimated using RIST, and two weighted methods as defined in equation (2). The weighted methods are: IPW with IPCW using the weights defined in equation (6), and IPW with imputation using $\hat{Y}$ and the weights defined in equation (4).

For comparison between methods, we estimate the optimal policy with a cross-validation type analysis. Specifically, the data are partitioned into ten roughly equal-sized parts. For each method under consideration, we estimate the optimal policy on nine parts of the data, and then compute several estimates of the SAPE using the remaining tenth part. By applying the above procedure, holding out a different part of the data each time, we were able to get out-of-sample predictions which should better represent expected optimal SAPE for future subjects. This approach was applied to 100 different partitions of the data and the results are presented in Table 1 and Figure 4.

From Table 1, we observe that the proposed approach produces policies with larger mean estimated SAPE with smaller standard errors across all three evaluation criteria. The policies found by the regression-based estimator had high estimated SAPE with respect to its corresponding criteria, $\hat{\psi}_{\text{Reg}}(\hat{\pi})$, but low estimated SAPE with respect to the weight based criteria, $\hat{\psi}_{\text{IPW}}(\hat{\pi})$ and $\hat{\psi}_{\text{IPW,IPCW}}(\hat{\pi})$. The proposed method was able to find policies with high estimated SAPE for all criteria, indicating that the policies found were truly the best in terms of SAPE, as their improvement was insensitive to the choice of validation estimation method.

Table 2 shows the mean and standard deviations of percentage of subjects that were
Table 1: Analysis of modified HIV data: Mean (sd) of estimated SAPE.

| Method                      | $\psi_{\text{Reg}}(\hat{\pi})$ | $\psi_{\text{IPW}}(\hat{\pi})$ | $\psi_{\text{IPW,IPCW}}(\hat{\pi})$ |
|-----------------------------|----------------------------------|----------------------------------|-------------------------------------|
| Reg                         | 667 (0.63)                       | 608 (12.1)                       | 601 (12.7)                          |
| IPW+IPCW                    | 658 (2.37)                       | 630 (12.0)                       | 617 (14.1)                          |
| IPW+$\hat{Y}$               | 660 (2.26)                       | 626 (12.5)                       | 620 (15.7)                          |
| BP+$\hat{Y}$                | 667 (0.43)                       | 649 (6.1)                        | 647 (6.9)                           |

Reg: Regression-based approach; IPW+IPCW: IPW combined with IPCW; IPCW+$\hat{Y}$: IPW with imputed outcome vector; BP+$\hat{Y}$: Balanced policy with imputed outcome vector.

Figure 4: Boxplots showing the estimated SAPE for the modified HIV data. Larger is better. Reg: Regression-based approach; IPW+IPCW: IPW combined with IPCW; IPCW+$\hat{Y}$: IPW with imputed outcome vector; BP+$\hat{Y}$: Balanced policy with imputed outcome vector.

Table 2: Analysis of modified HIV data: Mean (sd) of the percentage of subjects assigned to each treatment.

| Method                      | RAL (sd)          | EVG (sd)          | DTG (sd)         |
|-----------------------------|-------------------|-------------------|------------------|
| Reg                         | 42.1% (1.62)      | 51.0% (1.75)      | 6.9% (0.77)      |
| IPW+IPCW                    | 32.2% (3.98)      | 57.4% (3.25)      | 10.4% (1.83)     |
| IPW+$\hat{Y}$               | 36.6% (3.38)      | 54.9% (3.02)      | 8.5% (2.10)      |
| BP+$\hat{Y}$                | 48.1% (1.17)      | 49.4% (0.94)      | 2.5% (0.73)      |

Reg: Regression-based approach; IPW+IPCW: IPW combined with IPCW; IPCW+$\hat{Y}$: IPW with imputed outcome vector; BP+$\hat{Y}$: Balanced policy with imputed outcome vector.
assigned to each of the three treatment options for the different approaches over the 100 partitions of the data. The standard deviation of the treatment assignment percentages is smallest for the proposed method, indicating that it is less sensitive to small perturbations of the data when compared to the other approaches. The variability was largest for the inverse weighted methods which is likely a result of the higher variance of these estimators. The censoring rates were modest for RAL and EVG, but high for DTG. Estimating the average treatment effect in the presence of high censoring likely led to an unstable estimate and we believe that adjusting based on this estimate led to the unusually low numbers of subjects assigned to DTG.

6 Discussion

The proposed method appears to be more effective at finding optimal treatment policies, when compared to adapting previous methods to the right-censored setting. The use of unbiased weighted estimators is likely to result in high variance estimates of the SAPE which can complicate the process of finding the policy that is optimal for the population of interest. The proposed method uses a balanced policy evaluation and learning approach combined with imputation of the censoring times to reduce the variance of the estimator. Both of these approaches are likely to increase the bias, especially for small sample sizes. However, our results show that the reduction in variance offsets the increased bias and allows for more effective training of an optimal treatment policy.

The imputation approach that we propose can be used to extend any policy evaluation method to right censored data. In the simulation studies, we demonstrated that combining inverse propensity weighting with the imputation approach leads to better performance when compared with the method that uses both inverse propensity and inverse probability censoring weights. We showed that the imputation approach is compatible with any consistent estimate of the conditional survival or cumulative hazard function. We proved that the convergence rates for the conditional expected failure times is of the same order as the convergence rates for the conditional survival or cumulative hazard function estimator.

There are a number of ways in which the proposed method may be improved or extended. As currently formulated, the proposed method is limited to comparatively small datasets, both in terms of number of observations and the dimension of the covariate space. Like the original balanced policy evaluation method, the proposed method is more computationally intensive than existing approaches. There may be more efficient optimization procedures, such as alternating descent or boosting with a faster approximate method that could allow the proposed method to be applied to larger datasets. Additionally, when no parametric or semiparametric relationship can be assumed for failure time distribution, the convergence rate of the conditional expected survival times can be quite slow, especially when the number of covariates are large. We have discussed some potential solutions, but a more principled approach may be able to further increase the performance. It would also be of interest to extend our approach to dynamic treatment regimes where we wish to identify an optimal treatment sequence in a censored survival setting.
7 Appendix

Proof of Lemma 1. Noting that
\[
\frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i) W_i (\hat{Y}_i - T_i) = \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i) W_i (\hat{Y}_i - \tilde{Y}_i) + \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i) W_i (\hat{Y}_i - T_i),
\]
we can bound each term separately. Focusing on the rate at which \( \hat{Y} \to \tilde{Y} \), consider the form of \( \hat{Y} \)
\[
\int_{t_{min}}^{t_{max}} t \, d\hat{F}(t \mid X, A) + \tau \hat{S}(\tau \mid X, A) \frac{\int_{t_{min}}^{t_{max}} t \, d\hat{S}(t \mid X, A)}{\hat{S}(Y \mid X, A)} = - \int_{t_{min}}^{t_{max}} t \, d\hat{S}(t \mid X, A) + \tau \hat{S}(\tau \mid X, A) \frac{\int_{t_{min}}^{t_{max}} t \, d\hat{S}(t \mid X, A) + \tau \hat{S}(\tau \mid X, A)}{\hat{S}(Y \mid X, A)}.
\]
The first quantity depends on the pair \((\hat{S}(t \mid X, A), \hat{S}(Y \mid X, A))\) through the two maps \((A, B) \mapsto (A, 1/B) \mapsto \int_{t_{min}}^{t_{max}} t \, d\hat{S}(t \mid X, A)\). The derivative of the composition map is given by
\[
(\alpha, \beta) \mapsto \left( \alpha, \frac{-\beta}{\hat{S}(Y \mid X, A)^2} \right) \mapsto \int_{t_{min}}^{t_{max}} t \, d\alpha(t) + \frac{\int_{t_{min}}^{t_{max}} t \, d\hat{S}(t \mid X, A) \beta(y)}{\hat{S}(Y \mid X, A)}.
\]
Similarly, the second quantity depends on the pair \((\hat{S}(t \mid X, A), \hat{S}(Y \mid X, A))\) through the two maps \((A, B) \mapsto (A, 1/B) \mapsto \tau A/B\). The derivative of the composition map is given by
\[
(\alpha, \beta) \mapsto \left( \alpha, \frac{-\beta}{\hat{S}(Y \mid X, A)^2} \right) \mapsto \int_{t_{min}}^{t_{max}} t \, d\alpha(t) + \frac{\tau \alpha(\tau)}{\hat{S}(Y \mid X, A)} + \frac{\tau \hat{S}(\tau \mid X, A) \beta(y)}{\hat{S}(Y \mid X, A)}. \]
By the Hadamard differentiability of the maps, Theorems 2.8 and 12.1 of Kosorok (2008), and Assumption 2 we have that \( \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i - \bar{Y}_i = O_p(r_n) \). Returning to \( \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i) W_i (\hat{Y}_i - \bar{Y}_i) \), by Cauchy-Schwarz we have
\[
\frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i) W_i (\hat{Y}_i - \bar{Y}_i) \leq \left( \frac{1}{n} \|W\|_2^2 \right)^{\frac{1}{2}} \left( \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2 \right)^{\frac{1}{2}}.
\]
The proof of Kallus (2018) Theorem 3 implies that \( \frac{1}{n} \|W\|_2^2 = O_p(1) \), and thus
\[
\left| \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i) W_i (\hat{Y}_i - \bar{Y}_i) \right| \leq \left( \frac{1}{n} \|W\|_2^2 \right)^{\frac{1}{2}} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i)(\hat{Y}_i - \bar{Y}_i)^2 \right)^{\frac{1}{2}}.
\]
By assumption 3 and the exchangeability of the data, we have
\[
E \left[ \frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i)(\hat{Y}_i - \bar{Y}_i)^2 \right] \leq E \left[ (1 - \Delta_n)(\hat{Y}_n - \bar{Y}_n)^2 \right].
\]
Now,
\[
E \left[ (1 - \Delta_i)(\hat{Y}_n^{(n)} - \bar{Y}_n)^2 \right] \leq \tau E \left[ \left| \hat{E}_{n-1}(T \mid X_n, A_n, T > Y_n) - E(T \mid X_n, A_n, T > Y_n) \right| \wedge \tau \right]
\]
(since both $\hat{E}_{n-1}[T]$ and $E[T]$ are constrained to be $\in (0,\tau]$)

$$
= \tau E \left[ \left( \hat{E}_{n-1}(T | X, A, T > Y) - E(T | X, A, T > Y) \right) \wedge \tau \right]
= \tau E \left[ \left( -2\tau \hat{S}_{n-1}(\tau | X, A) + Y \hat{S}_{n-1}(Y | X, A) + \int_Y \hat{S}_{n-1}(t | X, A) dt \right) \hat{S}_{n-1}(Y | X, A) \right.
+ \left. \frac{2\tau S(\tau | X, A) - YS(Y | X, A) - \int_Y S(t | X, A) dt}{S(Y | X, A)} \right] \wedge \tau .
$$

(*1)

By adding and subtracting

$$
\frac{2\tau S(\tau | X, A) - YS(Y | X, A) - \int_Y S(t | X, A) dt}{\hat{S}_{n-1}(Y | X, A)}
$$

combining terms and simplifying, we have

$$
(*1) \leq 8\tau^2 E \left[ \sup_{t \leq \tau} \left( \frac{\hat{S}_{n-1}(t | X, A) - S(t | X, A)}{\hat{S}_{n-1}(\tau | X, A)} \right) \wedge \tau \right].
$$

(*2)

If we assume that $\inf_{X,A} S(\tau | X, A) \geq \delta$ for some $\delta > 0$, then,

$$
(*2) \leq \frac{16\tau^3}{\delta} \left[ \sup_{t \leq \tau} \left( \frac{\hat{S}_{n-1}(t | X, A) - S(t | X, A)}{\hat{S}_{n-1}(\tau | X, A)} \right) + \Pr \left( \hat{S}_{n-1}(t | X, A) \leq \frac{\delta}{2} \right) \right].
$$

(*3)

But

$$
\Pr \left( \hat{S}_{n-1}(t | X, A) \leq \frac{\delta}{2} \right) \leq \Pr \left( \hat{S}_{n-1}(t | X, A) - S(t | X, A) \leq -\frac{\delta}{2} \right)
= \frac{E \left[ \sup_{t \leq \tau} \left( \hat{S}_{n-1}(t | X, A) - S(t | X, A) \right) \right]}{\delta/2},
$$

thus

$$
(*3) \leq \frac{8\tau^3}{\delta} \left( 1 + \frac{2}{\delta} \right) E \left[ \sup_{t \leq \tau} \left( \hat{S}_{n-1}(t | X, A) - S(t | X, A) \right) \right].
$$

Note that $r_n \to 0$ iff $r_{n-1} \to 0$, so the rates should be similar. Turning our attention to $\frac{1}{n} \sum_{i=1}^n (1 - \Delta_i) W_i (\tilde{Y}_i - T_i)$. Let $F_i = \sigma(X_i, A_i, Y_i, \Delta_i)$ be the sigma field for a single observation and $\overline{F}_n = \sigma( \bigcup_{1 \leq i \leq n} F_i )$. For any $i$, note that

$$
\mathbb{E}[(1 - \Delta_i) W_i (\tilde{Y}_i - T_i)] = \mathbb{E} \left[ \mathbb{E}[(1 - \Delta_i) W_i (\tilde{Y}_i - T_i) | \overline{F}_n] \right]
= \mathbb{E} \left[ (1 - \Delta_i) W_i \left( \mathbb{E}[T_i | X_i, A_i, Y_i] - \mathbb{E}[T_i | X_i, A_i, Y_i] \right) \right] = 0.
$$
For $1 \leq i \neq j \leq n$, also note that
\[
\mathbb{E}[(1 - \Delta_i)W_i(\tilde{Y}_i - T_i)(1 - \Delta_j)W_j(\tilde{Y}_j - T_j)]
\]
\[
= \mathbb{E}
\left[
(1 - \Delta_i)(1 - \Delta_j)W_iW_j\mathbb{E}[(\tilde{Y}_i - T_i)(\tilde{Y}_j - T_j) | \sigma(F_i \cup F_j)]
\right],
\]
but,
\[
\mathbb{E}[(\tilde{Y}_i - T_i)(\tilde{Y}_j - T_j) | \sigma(F_i \cup F_j)]
\]
\[
= \mathbb{E}
\left[
\mathbb{E}[(\tilde{Y}_i - T_i) | X_i, A_i, Y_i, \Delta_i]\mathbb{E}[(\tilde{Y}_j - T_j) | X_j, A_j, Y_j, \Delta_j]
\right] = 0.
\]
This implies that $\frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i)W_i(\tilde{Y}_i - T_i)$ is approximately zero with variance
\[
\frac{1}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(1 - \Delta_i)W_i^2(\tilde{Y} - T)^2] \right] = \frac{1}{n} \mathbb{E}[(1 - \Delta_i)W_i^2(\tilde{Y} - T)^2] \to 0,
\]
by exchangeability and by the fact that $\limsup_{n \to \infty} \mathbb{E}[(1-\Delta_i)W_i^2(\tilde{Y} - T)^2] < \infty$. Combining these results, we obtain
\[
\frac{1}{n} \sum_{i=1}^{n} (1 - \Delta_i)W_i(\tilde{Y}_i - T_i) = O_p(r_n).
\]
References

Anstrom, K. J. & Tsiatis, A. A. (2001), ‘Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data’, *Biometrics* **57**(4), 1207–1218.

Athey, S. & Wager, S. (2017), ‘Efficient policy learning’, *arXiv preprint arXiv:1702.02896*.

Beygelzimer, A. & Langford, J. (2009), The offset tree for learning with partial labels, in ‘Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining’, ACM, pp. 129–138.

Breiman, L. (2001), ‘Random forests’, *Machine Learning* **45**(1), 5–32.

Cui, Y., Zhu, R., Kosorok, M. et al. (2017), ‘Tree based weighted learning for estimating individualized treatment rules with censored data’, *Electronic Journal of Statistics* **11**(2), 3927–3953.

Cui, Y., Zhu, R., Zhou, M. & Kosorok, M. (2019), ‘Consistency of survival tree and forest models: splitting bias and correction’, *arXiv preprint arXiv:1707.09631*.

Dabrowska, D. M. (1987), ‘Non-parametric regression with censored survival time data’, *Scandinavian Journal of Statistics* pp. 181–197.

Dabrowska, D. M. (1989), ‘Uniform consistency of the kernel conditional kaplan-meier estimate’, *The Annals of Statistics* pp. 1157–1167.

Dasgupta, S., Goldberg, Y., Kosorok, M. R. et al. (2019), ‘Feature elimination in kernel machines in moderately high dimensions’, *The Annals of Statistics* **47**(1), 497–526.

Dudık, M., Langford, J. & Li, L. (2011), ‘Doubly robust policy evaluation and learning’, *arXiv preprint arXiv:1103.4601*.

Fleming, T. R. & Harrington, D. P. (2011), *Counting processes and survival analysis*, Vol. 169, John Wiley & Sons.

Geurts, P., Ernst, D. & Wehenkel, L. (2006), ‘Extremely randomized trees’, *Machine learning* **63**(1), 3–42.

Hothorn, T., Lausen, B., Benner, A. & Radespiel-Tröger, M. (2004), ‘Bagging survival trees’, *Statistics in medicine* **23**(1), 77–91.

Imbens, G. W. (2000), ‘The role of the propensity score in estimating dose-response functions’, *Biometrika* **87**(3), 706–710.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. et al. (2008), ‘Random survival forests’, *The annals of applied statistics* **2**(3), 841–860.
Kalbfleisch, J. D. & Prentice, R. L. (2011), *The statistical analysis of failure time data*, Vol. 360, John Wiley & Sons.

Kallus, N. (2018), Balanced policy evaluation and learning, *in ‘Advances in Neural Information Processing Systems 31’, pp. 8895–8906.

Kosorok, M. R. (2008), *Introduction to empirical processes and semiparametric inference.,* Springer.

Li, L., Chen, S., Kleban, J. & Gupta, A. (2015), Counterfactual estimation and optimization of click metrics in search engines: A case study, *in ‘Proceedings of the 24th International Conference on World Wide Web’, ACM, pp. 929–934.

Liang, M., Ye, T. & Fu, H. (2018), ‘Estimating individualized optimal combination therapies through outcome weighted deep learning algorithms’, *Statistics in medicine 37*(27), 3869–3886.

Qian, M. & Murphy, S. A. (2011), ‘Performance guarantees for individualized treatment rules’, *Annals of statistics 39*(2), 1180–1210.

Rubin, D. B. (1986), ‘Comment: Which ifs have causal answers’, *Journal of the American Statistical Association 81*(396), 961–962.

Steingrimsson, J. A., Diao, L., Molinaro, A. M. & Strawderman, R. L. (2016), ‘Doubly robust survival trees’, *Statistics in medicine 35*(20), 3595–3612.

Thomas, P. S., Theocharous, G. & Ghavamzadeh, M. (2015), High-confidence off-policy evaluation, *in ‘Twenty-Ninth AAAI Conference on Artificial Intelligence’.

Wang, Y., Wu, P., Liu, Y., Weng, C. & Zeng, D. (2016), Learning optimal individualized treatment rules from electronic health record data, *in ‘Healthcare Informatics (ICHI), 2016 IEEE International Conference on’, IEEE, pp. 65–71.

Zhang, B., Tsiatis, A. A., Laber, E. B. & Davidian, M. (2012), ‘A robust method for estimating optimal treatment regimes’, *Biometrics 68*(4), 1010–1018.

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M. & Kosorok, M. R. (2014), ‘Doubly robust learning for estimating individualized treatment with censored data’, *Biometrika 102*(1), 151–168.

Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M. R. (2012), ‘Estimating individualized treatment rules using outcome weighted learning’, *Journal of the American Statistical Association 107*(499), 1106–1118.

Zhu, R. & Kosorok, M. R. (2012), ‘Recursively imputed survival trees’, *Journal of the American Statistical Association 107*(497), 331–340.

25
Supplemental Material for “Balanced Policy Evaluation and Learning for Right Censored Data”

Owen E. Leete
Department of Biostatistics,
University of North Carolina at Chapel Hill
and
Nathan Kallus
School of Operations Research and Information Engineering,
Cornell University
and
Michael G. Hudgens
Department of Biostatistics,
University of North Carolina at Chapel Hill
and
Sonia Napravnik
Division of Infectious Diseases, Department of Medicine,
University of North Carolina at Chapel Hill
and
Michael R. Kosorok
Department of Biostatistics,
University of North Carolina at Chapel Hill

November 14, 2019
Additional Simulation Study Results

Figure 5 shows the results for simulation setting 1 for the simple weighted and doubly robust estimators. For this setting, the mean was a complicated function of the covariates. The nonparametric estimate of the mean was not accurate enough for the doubly robust methods over the corresponding simple weighted versions.

![Boxplots showing estimated population regret for setting 1. Smaller is better.](image)

IPW+IPCW: IPW combined with IPCW; IPCW+Ŷ: IPW with imputed outcome vector; BP+Ŷ: Balanced policy with imputed outcome vector. Doubly robust versions used the correctly specified AFT model for Ŷ.

Figure 6 shows the results for simulation setting 2 with and without variable selection. For the smaller sample sizes, the variable selection method used was unable to reliably identify the meaningful variables. For the largest sample size, variable selection was able to improve the performance of the policy learners, which is most clearly seen in the IPW+Ŷ and IPW+IPCW estimators.

Additional Application Results

The differences between the different INSTIs during the first 2.5 years of follow-up were quite small. A larger difference between RAL and EVG does not start to appear in the Kaplan-Meier plots until about 3 years of follow-up, indicating that there may be some effects that do not manifest until later in the treatment course.

Based on clinical experience, we believe that treatment naïve patients should respond differently than treatment experienced patients and younger patients should respond differently than older patients.

Figure 7 shows survival curves broken down by treatment and treatment experience, and by treatment and age. For both treatment naïve and treatment experienced patients,
Dolutegravir is associated with longer survival times. Similarly, for patients aged 30 years and younger as well as patients over 30, Dolutegravir appears to be superior to the other treatment options.

Figure 8 shows the average projected survival curves under several ITR estimation methods, as well as the observed survival curve for the modified data. The regression based method resulted in a projected survival rate which is less than the observed survival rate in the data. The other methods examined resulted in improved projected survival rates when compared to the observed rate, but the best projected survival rates were associated with balanced policy evaluation with imputation.
Figure 7: The left plot shows the survival curves for each treatment broken down by treatment naïve vs. experienced. The right plot shows the survival curves for each treatment broken down by age $\leq 30$ vs age $> 30$. In both cases Dolutegravir was associated with longer survival times in each subpopulation.

Figure 8: Average projected survival curves under several ITR estimation approaches compared to the observed survival curve. Reg: Regression-based approach; IPW+IPCW: IPW combined with IPCW; IPCW+$\hat{Y}$: IPW with imputed outcome vector; BP+$\hat{Y}$: Balanced policy with imputed outcome vector.