Advancing Fairness in Precision Medicine: A Universal Framework for Optimal Treatment Estimation in Censored Data

Hongni Wang*
Shandong University
of Finance and Economics
wanghongnisd@126.com

Junxi Zhang*
Concordia University
junxi.zhang@concordia.ca

Na Li
Shandong University
of Finance and Economics
naibor@163.com

Linglong Kong
University of Alberta
lkong@ualberta.ca

Bei Jiang[†]
University of Alberta
bei1@ualberta.ca

Xiaodong Yan[†]
Xi'an Jiaotong University
yanxiaodong@xjtu.edu.cn

Abstract

In healthcare and precision medicine, estimating optimal treatment strategies for right-censored data while ensuring fairness across ethnic subgroups is crucial but remains underexplored. The problem presents two key challenges: measuring heterogeneous treatment effects (HTE) under fairness constraints and dealing with censoring mechanisms. We propose a general framework for estimating HTE using nonparametric methods and integrating user-controllable fairness constraints to address these problems. Under mild regularization assumptions, our method is theoretically grounded, demonstrating the double robustness property of the HTE estimator. Using this framework, we demonstrate that optimal treatment strategies balance fairness and utility. Using extensive simulations and real-world data analysis, we uncovered the potential of this method to guide the selection of treatment methods that are equitable and effective.

1 Introduction

1.1 Background

The development of optimal treatment strategies for right-censored data is essential to advancing precision medicine and improving healthcare outcomes. These treatment strategies are generally determined by estimating survival times, with optimal treatments maximizing survival rates. Consequently, such policies are highly relevant to real-world decision-making and have significant implications at both the individual and societal levels. Nevertheless, increasing evidence indicates that the algorithms used in policy-making can inherit biases and injustices ingrained in historical data or previous algorithms. Therefore, specific populations may be unfairly disadvantaged by discriminatory treatment strategies. The process of establishing policy should ensure that it is fair to all subgroups based on factors such as race, gender and other demographic characteristics. It is therefore essential to address algorithmic bias in algorithmic fairness, as extensively reviewed in del Barrio et al. (2020); Le Quy et al. (2022); Tang et al. (2023). Although there has been growing attention to fairness, more research is needed to estimate optimal treatment strategies for right-censored data that account for fairness across subgroups simultaneously. To address this issue, we propose a universal framework for estimating HTE with user-controllable fairness constraints tailored specifically for right-censored survival data. By taking this approach, we can ensure equitable and effective treatment outcomes and optimize policies to take subgroup differences into account, ultimately advancing the practice of precision medicine equitably.

1.2 Related Works

Understanding treatment-effect heterogeneity and identifying subgroups that respond similarly to a given treatment are critical tasks across various scientific domains. A primary measure for analyzing this heterogeneity is the Conditional Average Treatment Ef-

---

fect (CATE). Recent literature has explored numerous methods for estimating HTE (e.g., (Hahn et al., 2020; Kennedy, 2020; Künzel et al., 2019; Nie and Wager, 2021; Wager and Athey, 2018)). In particular, the doubly robust estimator of HTE based on the uncentered efficient influence function is widely used in prior works. For example, this estimator is utilized in Mishler et al. (2021), which presents a loss-optimal post-processed predictor to ensure counterfactual equalized odds. In the context of unobserved confounding in covariates, Byun et al. (2024) uses this estimator to audit the fairness of certain human policies, which can be generalized to ML-based predictors. The estimation of HTE is closely connected to the development of optimal treatment strategies, also known as policy learning (e.g., (Athey and Wager, 2021; Luedtke and Van Der Laan, 2016; Murphy, 2003)). The primary goal in policy learning is to identify subgroups with CATE values that exceed a certain threshold, allowing researchers to target the most promising subgroups based on treatment efficacy (see (Ballarini et al., 2018; Schnell et al., 2016; Zhao et al., 2013)).

When estimating optimal treatment strategies in the presence of censoring, Zhao et al. (2012, 2015) introduced techniques such as inverse censoring weighted outcome-weighted learning and doubly robust outcome-weighted learning. To further address censoring mechanisms, Cui et al. (2017) proposed a random forest imputation approach specifically designed for right-censored outcomes.

Incorporating fairness into treatment effect estimation or policy design typically involves minimizing the loss function between the target and its estimate while imposing fairness constraints. This approach has led to several pioneering studies on fair policy learning (Nabi et al., 2019; Kim and Zubizarreta, 2023; Viviano and Bradic, 2024). However, these studies do not address the challenge of optimal policy learning for HTE in censored survival data. Our framework seeks to fill this gap by integrating fairness constraints into the estimation of optimal treatment strategies for right-censored data, ensuring that the resulting policies are free from discrimination based on characteristics such as gender or race.

1.3 Contributions

The paper presents an innovative approach to addressing critical gaps in precision medicine using censored survival data. This work's contributions are novel and impactful, laying the groundwork for future advances in algorithmic fairness. Here are the key contributions:

1. This is the first framework for right-censored survival data that approximates the true heteroge-

neous treatment effects (HTE) under fairness constraints using a linear projection approach.

2. The framework allows practitioners to optimize treatment strategies while carefully balancing fairness and survival value function by providing a flexible, user-adjustable fairness mechanism.

3. With robust theoretical foundations, such as double robustness and asymptotic normality, the approach ensures its statistical reliability and instills confidence in the rigorous inferences that can be made.

4. With this framework, fairness would be embedded in treatment decisions, preventing biases in treatment allocation and improving health outcomes for underrepresented groups.

The proposed framework [1] is designed to apply to a wide range of domains where fairness in decision-making is essential, such as finance, education, and public policy. Consequently, its versatility enhances its long-term significance.

2 Methodology

2.1 Motivating Example

The AIDS Clinical Trials Group Study 175 Dataset (Hammer et al., 1996) evaluated treatment with either a single nucleoside (treat = 0) or a combination of two nucleosides (treat = 1) in adults infected with human immunodeficiency virus type 1 (HIV-1). It is concluded that treatment 1 is more effective when intervention is needed for a patient, suggesting that the optimal treatment policy during the intervention phase is $g^*(X, S) \in \{0, 1\}$ (see (5) for more details), where X represents the covariate vector and S represents the sensitive attributes. In Figure 1, it shows that the proportion of males ($S = 1$) receiving treatment 1 (left bar plot) is significantly lower than that of females ($S = 0$), which suggests that the male group would be unfairly treated under the optimal treatment policy.

There is a growing awareness that biases in data can lead to unfair outcomes in predictive models (Mhasawade et al., 2021; Gervasi et al., 2022), particularly in contexts involving life-saving decisions. For example, when model-based predictions determine the allocation of limited resources, such as organ transplants, specialist referrals, or ICU services, disparities in predictions can systematically disadvantage underrepresented groups (Paulus and Kent, 2020). This is

---

[1]The code can be found in https://github.com/wangholly/FairHTEforCENSORED

Hongni Wang*, Junxi Zhang*, Na Li, Linglong Kong, Bei Jiang[†], Xiaodong Yan[†]

an urgent issue that must be addressed. Treatment strategies are often based on HTE estimations, and any bias in these estimates can greatly amplify social unfairness, leading to unfair treatment. This unfairness also exists in other policy-driven areas; for another example, there is concern that automated decision-making processes may result in unequal treatment and discriminatory outcomes for certain groups when assessing loan applicants' credit worthiness (Moldovan, 2023).
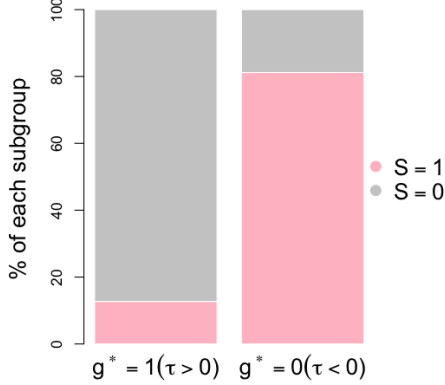

Figure 1: Unfairness proportions with respect to gender (S) for each treatment strategy in ACTG Dataset.

2.2 Problem Setup

Consider an i.i.d. sample $(Z_1, \ldots, Z_n)$ of n tuples $Z = (T, C, A, S, X) \sim \mathbb{P}$ for some distribution $\mathbb{P}$, where $T \in \mathbb{R}^+$ is the survival time, $C \in \mathbb{R}^+$ is the censoring time, $A \in \{0, 1\}$ is the assigned treatment, $S \in \{0, 1\}$ is the sensitive attribute, and $X \in \mathcal{X}$ represents the additional covariate feature vector for some compact subset $\mathcal{X}$. As commonly assumed in survival analysis, we assume that the censoring time C is independent of T given (X, S, A).

Goal Our goal is to estimate the treatment effect of treatment A on survival time T, given variables X and S, while adhering to fairness constraints related to the sensitive attribute S. Such fair estimation of treatment effect is then applied to learn an optimal treatment strategy. This involves two challenges: (i) estimating treatment effect heterogeneity under fairness constraints and (ii) handling censoring mechanisms in survival data.

Solution of Challenge (i) Assume that the survival time, $T(0), T(1)$, under treatment 0 and 1 are fully observed. Using the standard potential outcomes framework as outlined by Imbens and Rubin (2015):

- Consistency: $T = T(a)$ if $A = a$
- No unmeasured confounding: $A \perp\!\!\!\perp T(a) \mid X, S$
- Positivity: $\mathbb{P}(A = a \mid X, S) > 0$ a.s.

the conditional average treatment effect (CATE) on survival time is

$$\tau(X, S) = \mathbb{E}\left(T(1) - T(0) \mid X, S\right)$$
$$= \mu_1(X, S) - \mu_0(X, S), \quad (1)$$

where $\mu_a(X, S) = \mathbb{E}[T \mid X, S, A = a], \forall a \in \{0, 1\}$. In this case, the challenge lies in estimating CATE under fairness constraints. This challenge can be solved using the CATE estimation introduced in Section 2.3.

Solution of Challenge (ii) In contrast to the typical scenarios of estimating heterogeneous treatment effects as explored in studies such as Künzel et al. (2019); Nie and Wager (2021), the survival time T is not fully observed due to censoring mechanism. Usually, we observe $U = \min\{T, C, e\}$, accompanied by a censoring indicator $\Delta = \mathbb{1}(T \leq C)$, where e is the end time of study. To avoid the non-identifiable problem when $T > e$, we employee the finite horizon assumption in Cui et al. (2023), i.e., the survival time T admits a maximal time of study e such that $T = e$ for all $T > e$.

A preliminary approach involves substituting survival times with their conditional expected values, i.e., $\mathbb{E}(T \mid X, S, A)$, estimated through regression or nonparametric techniques (Van Buuren, 2018; Little and Rubin, 2019). We consider the following general approach,

$$T^{\text{imp}} = \Delta U + (1 - \Delta)\widehat{\mathbb{E}}(T \mid X, S, A, T > U, U), \quad (2)$$

where $\widehat{\mathbb{E}}(T \mid X, S, A, T > U, U)$ is an estimator of $\mathbb{E}(T \mid X, S, A, T > U, U)$. That is, we set $\widetilde{T} = T$ if the event occurs, and set $\widetilde{T} = T^{\text{imp}} = \widehat{\mathbb{E}}(T \mid X, S, A, T > U, U)$ if this subject is censored.

The CATE (1) can then be expressed as:

$$\widetilde{\tau}(X, S) = \mathbb{E}\left(\widetilde{T}(1) - \widetilde{T}(0) \mid X, S\right)$$
$$= \widetilde{\mu}_1(X, S) - \widetilde{\mu}_0(X, S), \quad (3)$$

where $\widetilde{\mu}_a(X, S) = \mathbb{E}[\widetilde{T} \mid X, S, A = a], \forall a \in \{0, 1\}$. In what follows, we generally denote the conditionally expected error of the imputation as $\epsilon(X, S)$, e.g., $\epsilon(X, S) = \tau(X, S) - \widetilde{\tau}(X, S)$.

One sufficient way[2] to obtain the estimation $\widehat{\mathbb{E}}(T \mid X, S, A, T > U, U)$ is through the random survival forest[3] (Ishwaran et al., 2008), which can be used to estimate the survival function $S(u \mid X, S, A) = \Pr(T > u \mid X, S, A)$. We denote the estimated survival function

---

[2]One can apply other estimators or imputation methods.
[3]The corresponding R package is `randomForestSRC`.

as $\widehat{S}(u \mid X, S, A)$. Due to the relationship

$$
\begin{aligned}
&\mathbb{E}(T \mid X, S, A, T > U, U) \\
&= U + \mathbb{E}(T - U \mid X, S, A, T > U, U) \\
&= U + \frac{\int_U^e S(u \mid X, S, A) du}{S(U \mid X, S, A)},
\end{aligned}
$$

where $\mathbb{E}(T - U \mid X, S, A, T > U, U)$ is the mean residual survival time. We can obtain estimation

$$
\widehat{\mathbb{E}}(T \mid X, S, A, T > U, U) = U + \frac{\int_U^e \widehat{S}(u \mid X, S, A) du}{\widehat{S}(U \mid X, S, A)},
$$

with which, T^{imp} in (2) can be obtained.

Learning an Optimal Treatment Strategy A treatment strategy or policy, $g : (X, S) \to \{0, 1\}$, is defined as a function that maps pre-treatment variables to treatment decisions. The CATE function, τ, is utilized to derive optimal strategies or identify specific interest subgroups. Hospitals typically suggest a treatment to maximize the *survival value function*, which is defined as:

$$
\mathcal{V}(g) = \mathbb{E}\{T(1)g(X, S) + T(0)(1 - g(X, S))\}. \quad (4)
$$

The optimal treatment strategy g^* is selected to optimize the survival value function, i.e., $g^* = \arg\max_{g \in \mathcal{G}} \mathcal{V}(g)$, where $\mathcal{G} = \{g | g : (X, S) \to \{0, 1\}\}$.

It can be readily demonstrated that the optimal policy that maximizes $\mathcal{V}(g)$ is

$$
g^*(X, S) = \mathbb{1}\{\tau(X, S) > 0\}. \quad (5)
$$

The optimal strategy, g^*, targets individuals for whom the treatment is given to produce a longer survival time. In a broader context, one might also seek to develop a targeting policy that identifies specific subgroups of interest:

$$
g_{\mathcal{I}}(X, S) = \mathbb{1}\{\tau(X, S) \in \mathcal{I}\},
$$

where $\mathcal{I}$ is an interval specifying the interest subgroup.

2.3 Proposed Framework

In this section, we present a framework to estimate CATE based on $(\widetilde{T}, X, S, A)$, where $\widetilde{T} = T$ if T is observed, $\widetilde{T} = T^{\text{imp}}$ otherwise. Specifically, our goal is to find the estimation of CATE, $\widehat{\widetilde{\tau}}(X, S)$, while satisfying fairness constraints, so that the optimal treatment strategy g^* can be estimated. Note that $(\widetilde{T}, X, S, A)$ is complete after the imputation.

Consider $\tau(X, S) \in \text{span}(\boldsymbol{b}(X, S))$, where $\boldsymbol{b}(X, S) := [b_1(X, S), \ldots, b_k(X, S)]^\top$ is a sequence of finite many basis functions for some $k \in \mathbb{Z}^+$, and span $(\boldsymbol{b}(X, S))$ is the space spanned by $\boldsymbol{b}(X, S)$. To estimate $\tau(X, S)$ under fairness constraints, we minimize the mean square

error (MSE) of $\widehat{\widetilde{\tau}}(X, S)$ subject to fairness constraints. This can be formulated as the following constrained stochastic optimization problem:

$$
\min_{\beta \in \mathcal{R}^k} \mathbb{E}\left[\left\{\widetilde{T}(1) - \widetilde{T}(0) + \epsilon(X, S) - \beta^\top \boldsymbol{b}(X, S)\right\}^2\right] \text{ (P)}
$$

$$
\text{s.t. } \left|\mathbb{E}\left\{\text{UF}_j(Z)\beta^\top \boldsymbol{b}(X, S)\right\}\right| \le \delta_j, j \in |m|,
$$

where $|m| = \{1, \cdots, m\}$, $\text{UF}_j(Z) : \mathbb{R}^+ \times (X, S) \times \{0, 1\} \to \mathbb{R}^+$ represents unfairness measure described later, and δ_j represents a predetermined tolerance level for the maximum acceptable degree of unfairness in the j-th criterion. When sensitive attributes are multi-categorical, it is natural to consider using pairwise unfairness constraints. For instance, when $S \in \{s_1, \cdots, s_L\}$, fairness constraints (as proposed in (6), (7), (9) below) can be incorporated for each pair (s_i, s_l) with $i \ne l$. If fairness is achieved between each pair of variables, then overall fairness among all variables is ensured. For simplicity, this paper will focus on implementing fairness between binary variables. The solution to the program above provides the coefficients of the best-fitting function for $\tau(X, S)$ within the finite-dimensional model space spanned the basis functions $\boldsymbol{b}(X, S) = (b_1(X, S), \ldots, b_k(X, S))^\top$, subject to the m fairness constraints.

The unfairness measures $\text{UF}_j(Z)$ can be defined as follows to represent different fairness notions as described in Mishler and Kennedy (2022).

(1)Demographic parity can be applied by defining

$$
\text{UF}_j(Z) = \frac{\mathbb{1}\{S = s\}}{\mathbb{E}(\mathbb{1}\{S = s\})} - \frac{\mathbb{1}\{S = s'\}}{\mathbb{E}(\mathbb{1}\{S = s'\})}, \quad (6)
$$

which leads to

$$
\left|\mathbb{E}\left\{\beta^\top \boldsymbol{b}(X, S) \mid S = s\right\} - \mathbb{E}\left\{\beta^\top \boldsymbol{b}(X, S) \mid S = s'\right\}\right| \le \delta_j,
$$

ensuring our fitted models are marginally independent of the sensitive attribute.

(2) Conditional statistical parity can be attained by defining

$$
\text{UF}_j(Z) = \frac{(\mathbb{1}\{S = s\})\mathbb{1}\{C(X, S) \in \mathcal{I}\}}{\mathbb{E}\left[(\mathbb{1}\{S = s\})\mathbb{1}\{C(X, S) \in \mathcal{I}\}\right]} \\
- \frac{\mathbb{1}\{S = s'\}\mathbb{1}\{C(X, S) \in \mathcal{I}\}}{\mathbb{E}\left[\mathbb{1}\{S = s'\}\mathbb{1}\{C(X, S) \in \mathcal{I}\}\right]}, \quad (7)
$$

which implies

$$
\begin{aligned}
&|\mathbb{E}\left\{\beta^\top \boldsymbol{b}(X, S) \mid S = s, C(X, S) \in \mathcal{I}\right\} - \\
&\mathbb{E}\left\{\beta^\top \boldsymbol{b}(X, S) \mid S = s', C(X, S) \in \mathcal{I}\right\}| \le \delta_j. \quad (8)
\end{aligned}
$$

Here, $C : \mathcal{X} \times \{0, 1\}$ represents the legitimate factor used to specify the condition under which the estimated treatment effect should remain independent of the sensitive attribute.

(3) Equal opportunity, which ensures demographic parity concerning the treatment effect, can be applied by defining

$$\mathrm{UF}_j(Z) = \frac{(\mathbb{1}\{S=s\})\mathbb{1}\{\widetilde{\tau}(X,S)>0\}}{\mathbb{E}[(\mathbb{1}\{S=s\})\mathbb{1}\{\widetilde{\tau}(X,S)>0\}]} - \frac{\mathbb{1}\{S=s'\}\mathbb{1}\{\widetilde{\tau}(X,S)>0\}}{\mathbb{E}[\mathbb{1}\{S=s'\}\mathbb{1}\{\widetilde{\tau}(X,S)>0\}]}, \quad (9)$$

which leads to

$$\begin{aligned} | \; \mathbb{E}\left\{\beta^\top \boldsymbol{b}(X,S) \mid S=s, \widetilde{\tau}(X,S)>0\right\} - \\ \mathbb{E}\left\{\beta^\top \boldsymbol{b}(X,S) \mid S=s', \widetilde{\tau}(X,S)>0\right\} |\le \delta_j. \quad (10) \end{aligned}$$

One can trivially generalize $\widetilde{\tau}(X,S) > 0$ to $\widetilde{\tau}(X,S) \in \mathcal{I}$ for some interval $\mathcal{I}$ in a boarder context.

2.4 Estimation

To solve the problem (P), we need to estimate (P) using suitable data-driven estimations. Based on the identification assumptions (Imbens and Rubin (2015)), we can derive the following identity:

$$\begin{aligned} \mathbb{E}\left[\widetilde{T}(a)\boldsymbol{b}(X,S)\right] &= \mathbb{E}\left[\widetilde{\mu}_a(X,S)\boldsymbol{b}(X,S)\right] \\ &= \mathbb{E}\left[\frac{\mathbb{1}\{A=a\}\widetilde{T}}{\pi_a(X,S)}\boldsymbol{b}(X,S)\right], \end{aligned}$$

where $\pi_a(X,S) = \mathbb{P}[A=a \mid X,S]$ is the *propensity score*. Then one may estimate the *counterfactual parameter* $\mathbb{E}\left[\widetilde{T}(a)\boldsymbol{b}(X,S)\right]$ using either the plug-in (PI) estimator $\mathbb{P}_n\left\{\widehat{\widetilde{\mu}}_a(X,S)\boldsymbol{b}(X,S)\right\}$ or the inverse-probability-weighted estimator $\mathbb{P}_n\left\{A\widetilde{T}/\widehat{\pi}_a(X,S)\boldsymbol{b}(X,S)\right\}$ depending on the quality of information to model the observational outcome or treatment process. Here $\widehat{\widetilde{\mu}}_a$ and $\widehat{\pi}_a$ are some estimators of $\widetilde{\mu}_a$ and π_a, respectively.

To solve the optimal problem (P), we introduce a more efficient semiparametric estimation method to estimate the components in (P). Let φ_a denote the uncentered *efficient influence function (EIF)* for the parameter $\mathbb{E}\left[\widetilde{T}(a)\right] = \mathbb{E}\{\mathbb{E}[\widetilde{T} \mid X,S,A=a]\}$, which is defined by

$$\varphi_a(Z;\eta) = \frac{\mathbb{1}(A=a)}{\pi_a(X,S)}\left\{\widetilde{T} - \widetilde{\mu}_A(X,S)\right\} + \widetilde{\mu}_a(X,S),$$

with a set of the nuisance components $\eta = \{\pi_a, \widetilde{\mu}_a\}$ (Kennedy (2017, 2022)).

Next, we use cross-fitting to allow for arbitrarily complex nuisance estimators $\widehat{\eta}$. Specifically, we split the data into K disjoint groups, each with size n/K approximately, by drawing variables $(B_1, \ldots, B_n)$ independent

of the data, with $B_i = b$ indicating that subject i was split into group $b \in \{1, \ldots, K\}$. Then the semiparametric estimator for $\mathbb{E}\left[\widetilde{T}(a)\boldsymbol{b}(X,S)\right]$ based on the EIF and sample splitting are given by

$$\begin{aligned} \frac{1}{K}&\sum_{b=1}^{K} \mathbb{P}_n^b \left\{\varphi_a\left(Z;\widehat{\eta}_{-b}\right)\boldsymbol{b}(X,S)\right\} \\ &\equiv \mathbb{P}_n\left\{\varphi_a\left(Z;\widehat{\eta}_{-B}\right)\boldsymbol{b}(X,S)\right\}, \end{aligned}$$

where we let $\mathbb{P}_n^b$ denote empirical averages only over the set of units in group $b \{i : B_i = b\}$ and let $\widehat{\eta}_{-b}$ denote the nuisance estimator constructed only using those units $\{i : B_i \ne b\}$. This suggests

$$\mathbb{P}_n\left\{\varphi_1\left(Z;\widehat{\eta}_{-B}\right) - \varphi_0\left(Z;\widehat{\eta}_{-B}\right)\boldsymbol{b}(X,S)\right\} \quad (11)$$

as our estimator for $\mathbb{E}\left[\left(\widetilde{T}(1) - \widetilde{T}(0)\right)\boldsymbol{b}(X,S)\right] = \mathbb{E}[\widetilde{\tau}(X,S)\boldsymbol{b}(X,S)]$. Under weak regularity conditions, this cross-fitting-based semiparametric estimator attains the efficiency bound with the double robustness property and thus allows us to employ flexible machine learning methods while achieving the $\sqrt{n}$-rate of convergence and valid inference (Kennedy (2017)).

The estimation of the fairness constraints in (P) relies on the fairness notion. In what follows, we present the estimation of $\mathrm{UF_j}\boldsymbol{b}(X,S)$ based on the fairness criterias given in (6), (7) and (9).

To estimate the unfairness measure (6) corresponding to demographic parity, one may simply use the following sample-average estimator:

$$\begin{aligned} \mathbb{P}_n&\left\{\widehat{\mathrm{UF}}_j\boldsymbol{b}(X,S)\right\} = \\ &\mathbb{P}_n\left[\left\{\frac{(1-\mathbb{1}\{S=s\})}{\mathbb{P}_n(1-\mathbb{1}\{S=s\})} - \frac{\mathbb{1}\{S=s\}}{\mathbb{P}_n(\mathbb{1}\{S=s\})}\right\}\boldsymbol{b}(X,S)\right], \end{aligned}$$

which is naturally $\sqrt{n}$-consistent without any need for nuisance estimation. Similarly, the empirical estimator is used when conditional statistical parity in (7) is ued. To estimate the unfairness measure (9) corresponding to demographic parity with respect to treatment, we will employ the margin condition to manage the non-smooth component. In other words, we restrict the closeness of $\widetilde{\mu}_1(X,S)$ and $\widetilde{\mu}_0(X,S)$ using their probability, which is imposed by the margin condition:

Definition 2.1. (Margin Condition). For any margin exponent $\alpha > 0$ and for all $t \ge 0$,

$$\mathbb{P}\left(|\widetilde{\mu}_1(X,S) - \widetilde{\mu}_0(X,S)| \le t\right) \lesssim t^\alpha. \quad (12)$$

This margin condition allows us to use the following

plug-in type estimator with cross fitting:

$$\mathbb{P}_n \left\{ \widehat{\mathrm{UF}}_j \boldsymbol{b}(X,S) \right\} =$$

$$\frac{1}{K} \sum_{b=1}^{K} \mathbb{P}_n^b \left(\boldsymbol{b}(X,S) \left[\frac{(1-S)\phi(X,S)}{\frac{1}{K}\sum_{b=1}^{K}\mathbb{P}_n^b \left[(1-S)\phi(X,S)\right]} \right.\right.$$

$$\left.\left. - \frac{S\phi(X,S)}{\frac{1}{K}\sum_{b=1}^{K}\mathbb{P}_n^b \left[S\phi(X,S)\right]} \right] \right),$$

where $\phi(X,S) = \mathbb{1}\left\{ \widehat{\widetilde{\mu}}_{1,-b}(X,S) - \widehat{\widetilde{\mu}}_{0,-b}(X,S) > 0 \right\}$.

The above estimator attains $\sqrt{n}$-consistency and asymptotic normality if

$$\mathbb{P}\left[\mathbb{1}\left\{ \widehat{\widetilde{\mu}}_1 - \widehat{\widetilde{\mu}}_0 > 0 \right\} \neq \mathbb{1}\left\{ \widetilde{\mu}_1 - \widetilde{\mu}_0 > 0 \right\} \right] = o_\mathbb{P}(1)$$

and $\max_a \left\| \widehat{\widetilde{\mu}}_a - \mu_a \right\|_{\infty,\mathbb{P}}^\alpha = o_\mathbb{P}\left(n^{-\frac{1}{2}}\right)$ as discussed in Kennedy (2020, 2022).

Because our fairness function involves the potential survival time, we can utilize estimators similar to those employed for our objective function. Specifically, if the fairness function $\mathrm{UF}_j(\widetilde{T}(0), \widetilde{T}(1), X, S)$ is a smooth function of $\widetilde{T}(0), \widetilde{T}(1)$, then we can use the following EIF-based semiparametric estimator:

$$\mathbb{P}_n \left\{ \widehat{\mathrm{UF}}_j \boldsymbol{b}(X,S) \right\} =$$

$$\mathbb{P}_n \left\{ \mathrm{UF}_j \left(\varphi_0\left(Z; \widehat{\eta}_{-B}\right), \varphi_1\left(Z; \widehat{\eta}_{-B}\right), X, S\right) \boldsymbol{b}(X,S) \right\},$$

which is asymptotically normal and efficient according to the same logic as used for (11).

Consequently, our approximating program can be found as the following convex quadratic program:

$$\min_{\beta \in \mathcal{R}^k} \quad \frac{1}{2}\beta^\top \mathbb{P}_n \left\{ \boldsymbol{b}(X,S)\boldsymbol{b}(X,S)^\top \right\} \beta$$

$$- \beta^\top \mathbb{P}_n \{ \boldsymbol{b}(X,S)\epsilon(X,S) \}$$

$$- \beta^\top \mathbb{P}_n \left[\{ \varphi_1\left(Z; \widehat{\eta}_{-B}\right) - \varphi_0\left(Z; \widehat{\eta}_{-B}\right) \} \boldsymbol{b}(X,S) \right] \quad (\widehat{P})$$

$$\text{s.t.} \left| \beta^\top \mathbb{P}_n \left\{ \widehat{\mathrm{UF}}_j \boldsymbol{b}(X,S) \right\} \right| \leq \delta_j, j \in J.$$

The above optimization can be readily solved using off-the-shelf solvers. Let $\widehat{\beta}$ be an optimal solution to $(\widehat{P})$. Our proposed estimator for $\tau(X,S)$ is then given by $\widehat{\widetilde{\tau}}(X,S) = \widehat{\beta}^\top \boldsymbol{b}(X,S)$, therefore, the optimal treatment strategy g^* can be estimated by $\widehat{g}(X,S) = \mathbb{1}\{\widehat{\widetilde{\tau}}(X,S) > 0\}$.

3 Theoretical Properties

3.1 Inferences of Estimation $\widehat{\beta}$

Let β^* and $\widehat{\beta}$ be the solutions to the optimal problems (P) and $(\widehat{P})$, respectively. One advantage of the

proposed method is the ability to conduct statistical inference on $\widehat{\beta}$. Based on the proposed estimator, we introduce the following assumptions to state the asymptotic properties of $\widehat{\beta}$.

(A1) $\mathbb{E}[\boldsymbol{b}(X,S)\boldsymbol{b}(X,S)^\top]$ is positive definite.

(A2) There exists some $\epsilon > 0$ such that $\widehat{\pi}_a \in [\epsilon, 1-\epsilon]$ with probability one.

(A3) $||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}} = o_\mathbb{P}(1)$ or $||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} = o_\mathbb{P}(1)$.

(A4) $||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} = o_\mathbb{P}(\frac{1}{\sqrt{n}})$.

(A5) $\mathbb{P}_n\left(\widehat{\mathrm{UF}}_j\boldsymbol{b}(X,S)\right) - \mathbb{E}[\mathrm{UF}_j\boldsymbol{b}(X,S)] = O_\mathbb{P}(\frac{1}{\sqrt{n}})$.

(A6) $\sqrt{n}\left(\mathbb{P}_n\left(\widehat{\mathrm{UF}}_j\boldsymbol{b}(X,S)\right) - \mathbb{E}[\mathrm{UF}_j\boldsymbol{b}(X,S)]\right)$ is asymptotically normal $N(0, \sigma^2)$ for some $\sigma^2 < \infty$.

Assumption (A1) ensures the uniqueness of optimal solution of (P) and the quadratic growth condition holds at the optimal solution in (P), such assumption can be replaced by a weaker second-order condition[4] (see e.g., Section 2.4 in Still (2018) for more details.). Assumptions (A2)-(A4) are commonly used conditions in causal inference. It is worth pointing out that assumptions (A2)-(A4) do not include the imputation error. Assumptions (A5)-(A6) are guarantees of convergence of the estimations of fairness constraints at a proper rate.

Theorem 3.1. *Let β^* and $\widehat{\beta}$ be the solutions to the optimal problems* (P) *and* $(\widehat{P})$, *respectively.*

(i) *Under Assumptions (A1)-(A3), and (A5),*

$$||\widehat{\beta} - \beta^*||_2 =$$

$$O_\mathbb{P}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right) \quad (13)$$

(ii) *Under addition Assumptions (A4) and (A6), and assuming Linear Independence Constraint Qualification* (LICQ) *and Strict Complementarity* (SC)[5] *hold at β^*,*

$$||\widehat{\beta} - \beta^*||_2 = O_\mathbb{P}\left(\frac{1}{\sqrt{n}}\right), \quad (14)$$

$$\sqrt{n}(\widehat{\beta} - \beta^*) \xrightarrow{d} N(0, \sigma_0^2), \text{ where } \sigma_0^2 < \infty. \quad (15)$$

Result (i) in Theorem 3.1 provides doubly robust property of $\widehat{\beta}$, implying accuracy of the fair estimator $\widehat{\beta}$ if either π_a or μ_a is accurately estimated. Furthermore, result (ii) in Theorem 3.1 provides asymptotically normality of $\widehat{\beta}$, which provides ability of giving statistical

---

[4]See the condition in Appendix E.

[5]See definitions of LICQ and SC assumptions in Appendix D.

inferences with respect to $\widehat{\beta}$ via bootstrap. It is worth pointing out that the imputation error does not affect the estimation error of β^*. Therefore, Theorem 3.1 provides a solid theoretical guarantee of using the proposed estimations in Section 2.4.

3.2 Properties of Estimated Survival Value Function $\widetilde{\mathcal{V}}(\widehat{g}(X,S))$

With the proposed estimation of CATE, $\widehat{\beta}^\top \mathbf{b}(X,S)$, one can estimate the optimal treatment rule g^* in (5) by

$$\widehat{g}(X,S) = \mathbb{1}\{\widehat{\beta}^\top \mathbf{b}(X,S) > 0\}. \qquad (16)$$

The performance of the estimated survival value function $\widetilde{\mathcal{V}}(\widehat{g}) = \mathbb{E}[\widetilde{T}(1)\widehat{g}(X,S) + \widetilde{T}(0)(1 - \widehat{g}(X,S))]$ can be evaluated, under the margin condition, using the L_q loss of the proposed CATE estimator and the imputation error in the following lemma.

Lemma 3.2. *Assuming the margin condition holds with margin exponent $\alpha \in (0,\infty)$. Then*

$$|\mathcal{V}(g^*(X,S)) - \widetilde{\mathcal{V}}(\widehat{g}(X,S))|$$
$$\lesssim \parallel \widehat{\beta}^\top \mathbf{b}(X,S) - \widetilde{\tau}(X,S) \parallel_{q,\mathbb{P}}^\gamma + \parallel \epsilon(X,S) \parallel_{1,\mathbb{P}},$$

where $\gamma = \alpha + 1$ if $q = \infty$, $\gamma = \frac{q(\alpha+1)}{q+\alpha}$ if $q \in [1,\infty)$.

3.3 Trade-off between Fairness and Survival Value Function

We first introduce the following notations to characterize the trade-off between fairness and the survival value function.

- **Estimation error of $\widetilde{\tau}$:** $\widetilde{\beta}$ is

$$\underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathbb{E}\Big[\Big\{\widetilde{T}(1) - \widetilde{T}(0) + \epsilon(X,S)$$
$$- \beta^\top \boldsymbol{b}(X,S)\Big\}^2\Big]$$

  is the unconstrained optimal parameter.

- **Estimation error of $\widehat{\beta}$:**
  $T_{1,n} =$
  $O_\mathbb{P}\Big(\max_a \|\widehat{\pi}_a - \pi_a\|_{2,\mathbb{P}}\|\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a\|_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\Big).$

- **Cost due to fairness constraints:**
  $T_2 = O\Big(\parallel \sum_j \sqrt{\lambda_j}\mathrm{UF}_j(Z) \parallel_{2,\mathbb{P}} \parallel \mathbf{b}(X,S) \parallel_{2,\mathbb{P}}\Big)$, where λ_j is the Lagrange multiplier associated with the j-th fairness constraint in (P).

Theorem 3.3. *Under Assumptions (A1)-(A3), (A5) and the margin condition with margin exponent $\alpha \in (0,\infty)$, we have*

(i) $|\mathcal{V}(g^*(X,S)) - \widetilde{\mathcal{V}}(\widehat{g}(X,S))| \lesssim \parallel \epsilon(X,S) \parallel_{1,\mathbb{P}} + T_{1,n}^\gamma + T_2^\gamma + \parallel \widetilde{\beta}^\top \mathbf{b}(X,S) - \widetilde{\tau}(X,S) \parallel_{q,\mathbb{P}}^\gamma$, *where $\gamma = \alpha + 1$ if $q = \infty$, and $\gamma = \frac{q(\alpha+1)}{q+\alpha}$ if $q \in [1,\infty)$.*

(ii) $\mathbb{P}(g^*(X,S) \neq \widehat{g}(X,S)) \lesssim \parallel \epsilon(X,S) \parallel_{1,\mathbb{P}} + T_{1,n}^\alpha + T_2^\alpha + \parallel \widetilde{\beta}^\top \mathbf{b}(X,S) - \widetilde{\tau}(X,S) \parallel_{\infty,\mathbb{P}}^\alpha$.

Theorem 3.3 presents upper bounds for the estimation error of the optimal survival value function (result (i)) and the probability of $g^*(X,S) \neq \widehat{g}(X,S)$ (result (ii)). The upper bounds provide a clear trade-off between fairness and estimation errors. For instance, the upper bounds involve four parts: the estimation error of $\tau(X,S)$, the imputation error, the estimation error of nuisance estimations $T_{1,n}$, the unfairness constraints T_2 in terms of the Lagrange multipliers. Therefore, the upper bounds will increase when fairness is increasing (i.e., λ_j is increasing. Therefore T_2 will increase), or/and when the estimations of $\tau(X,S)$, the imputation method and nuisance parameters have decreasing accuracy. Although the estimation error of $\tau(X,S)$ is unavoidable (due to the modeling error and optimization algorithm error), it may close to 0 when μ_a lies in the function space spanned by basis functions $\mathbf{b}$ and when the optimization algorithm is accurate. The second term evaluates the imputation error, which can be small with advanced imputation methods such as the proposed random survival forest. The third term, $T_{1,n}$, in the upper bounds converges to 0 with rate $\sqrt{n}$ when either π_a or $\widetilde{\mu}_a$ is estimated with an error with order $\frac{1}{\sqrt{n}}$, or when both π_a and $\widetilde{\mu}_a$ are estimated with errors with order $n^{-\frac{1}{4}}$. Importantly, the term T_2 characterizes how much the cost of accuracy is due to the fairness constraints involved in the optimal problem (P). The smaller the fairness tolerance level δ_j is, the larger Lagrange multiplier λ_j should be selected, therefore, the larger T_2 becomes. On the other hand, when the fairness tolerance level δ_j is very large, λ_j is close to 0. Therefore, T_2 is close to 0.

In short, there is a cost to accuracy associated with fairness constraints when estimating the optimal treatment rule at a specific level of fairness. Therefore, Theorem 3.3 characterizes the trade-off between fairness and the accuracy of the estimation of the optimal survival value function, $\mathcal{V}(g^*(X,S))$. Specifically, when less fairness is required (large δ_j and small λ_j), the estimation error of $\mathcal{V}(g^*(X,S))$ is smaller. On the other hand, when more fairness is required (small δ_j and large λ_j), the estimation error of $\mathcal{V}(g^*(X,S))$ is larger according to T_2. This scenario aligns with the trade-off between fairness and faithfulness discussed in the literature on fairness learning. With the proposed framework, this trade-off is quantified in analytic form in Theorem 3.3, shedding light on fairness studies.

4   Experiments

4.1   Real Data Analysis

We apply the proposed framework to the ACTG Dataset mentioned in Section 2.1. To enhance the illustration of the proposed framework, we run experiment on another real data set, the HCC dataset, in Appendix C. We further consider a simulation study in Appendix A to explain our motivation and the proposed method.

The ACTG dataset include 22 features and 1 sensitive attribute, *gender*. The outcome of interest is the survival time of the patients after receiving certain treatments. The median duration of the experimental follow-up was 143 weeks, and the censoring rate was 75.64%. As a preparation, we removed the multivariate treatment indicator 'trt' and retained the binary variable 'treat'. Specifically, $A = 0$ refers the treatment ZDV, and $A = 1$ refers other treatments (ZDV + ddI; ZDV + Zal; ddI only).

Firstly[6], we impute the censoring data as described in Section 2.2 using the random survival forest. We use roughly 75% of the data to estimate g^* (and therefore $\mathcal{V}(g^*)$) using the proposed framework and the remaining 25% as the testing set.

With the estimated optimal treatment strategy $\hat{g}$ from our method, Figure 2 shows that the proportions of males and females in each treatment group are nearly equal, demonstrating the fairness of the proposed estimation process. To further highlight the fairness achieved by our framework, we compare the density functions of the estimated treatment effect $\widehat{\tilde{\tau}}(X, S)$ in Figure 3 with fairness constraint when $\delta = 0$ (bottom plot) and without fairness constraint when $\delta = \infty$ (top plot). In this case, $\delta = 90$ is equivalent to having no fairness constraint ($\delta = \infty$) since the constraint becomes inactive. Without fairness constraints, we can see from the top plot, the center of the estimated treatment effect for the male group ($S = 1$) is significantly lower than that for the female group ($S = 0$), suggesting that male patients are less likely to receive treatment 1 compared to female patients. This imbalance reflects an unfair treatment effect estimation, which could lead to biased treatment decisions. Conversely, in the bottom plot, the centers of the estimated treatment effects for both groups coincide under the fairness constraint ($\delta = 0$), indicating an equitable distribution of treatment effects between male and female groups. Additionally, the centers of the estimated treatment effects align with the center of the overall density of $\widehat{\tilde{\tau}}(X, S)$, further demonstrating the effectiveness of our

---

[6]Experimental details are presented in Appendix B.

proposed fair estimation method for treatment effects $\tau(X, S)$.

To illustrate the theoretical findings in Section 3, which demonstrate that fairness in optimal policies comes at a cost, we estimate $g^*(X, S)$ across different values of δ ranging from 0 to 90. Simultaneously, we estimate the survival value function $\mathcal{V}(g^*(X, S))$ and the level of unfairness for each δ. In Figure 4, we show how the estimated survival value function changes as the level of unfairness varies. The results clearly indicate that the survival value function increases as we allow more lenient fairness tolerance (δ), confirming that there is a cost to achieving fairness in terms of reduced survival time. As demonstrated in Theorem 3.3, a trade-off exists between the estimation error of $\tilde{\mathcal{V}}(\hat{g}(X, S))$ and fairness. Figure 5 further verifies our theoretical findings by showing the decreasing trend of the estimation error of $\tilde{\mathcal{V}}(\hat{g}(X, S))$ as unfairness increases. As the fairness constraint loosens, the estimation error diminishes, eventually stabilizing at a low error level when δ is sufficiently large.

The selection of δ in the fairness constraints depends heavily on the specific requirements of users' task. For example, if fairness is the primary priority in the task, δ should be close to 0. On the other hand, if accuracy is the primary priority in the task, one should let δ increase. Interestingly, from Figure 4 and 5 on real datasets and the results in simulation study in Appendix A, we observe that the estimation error decreases rapidly as δ increases around 0, but then slows down significantly as δ continues to increase. Therefore, if fairness is not the primary priority but still required, one may select δ at the point where the decay in estimation error starts to slow down, optimizing accuracy while ensuring a certain level of fairness. As discussed in many prior works (Chzhen and Schreuder, 2022; Feldman et al., 2015), a sustainable selection rule involves accepting some degree of compromise on fairness to preserve accuracy, allowing $\delta > 0$ and preserving significant accuracy. In specific contexts, when certain unfairness rules-such as the 80% rule in Feldman et al. (2015)-are applied, one may set $\delta = 0.2$.

5   Conclusion

In this paper, we introduce a novel framework that incorporates fairness constraints into the estimation of optimal treatment strategies for right-censored data. Our approach begins with the imputation of right-censored data, followed by the estimation of heterogeneous treatment effects using nonparametric methods under user-controllable fairness constraints. This framework is supported by strong theoretical guarantees, including double robustness and the asymptotic normality of the estimator, which enable valid statistical inference. Ad-

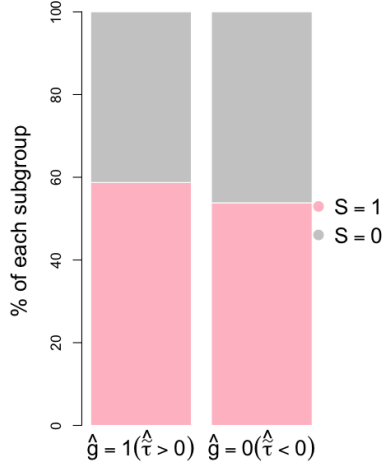**Hongni Wang*, Junxi Zhang*, Na Li, Linglong Kong, Bei Jiang[†], Xiaodong Yan[†]**

Figure 2: Fairness proportions with respect to gender (S) for each treatment strategy in ACTG Dataset.
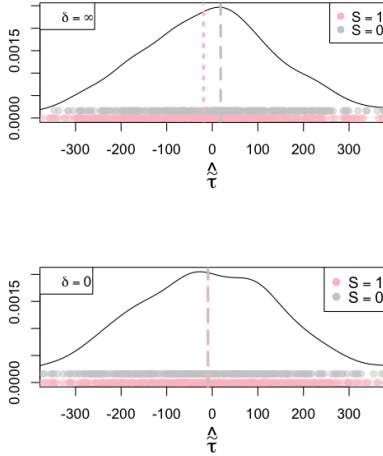

Figure 3: Densities of $\widehat{\widetilde{\tau}}(X, S)$ without a fairness constraint (when $\delta = \infty$) and with a fairness constraint (when $\delta = 0$) in the ACTG dataset. The two vertical dashed lines correspond to $\mathbb{P}_n(\widehat{\widetilde{\tau}}(X, S) \mid S = 0)$ and $\mathbb{P}_n(\widehat{\widetilde{\tau}}(X, S) \mid S = 1)$.

ditionally, we examine the trade-off between fairness and the estimation error of the survival value function.

Looking ahead, we plan to extend this framework to handle multiple treatment options beyond binary treatments, while preserving fairness constraints within the optimal treatment strategy. Furthermore, we aim to expand the data structure to address both right-censored data and other forms of missing data. These enhancements will be central to our future research endeavors.
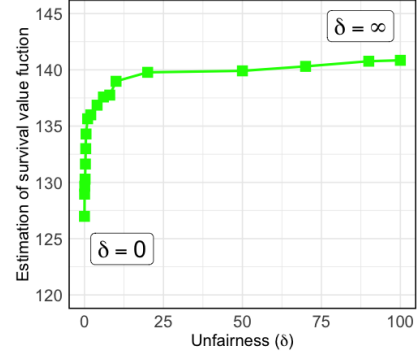
Figure 4: The cost of $\tilde{\mathcal{V}}(\widehat{g}(X, S))$ with respect to the unfairness in the ACTG dataset.
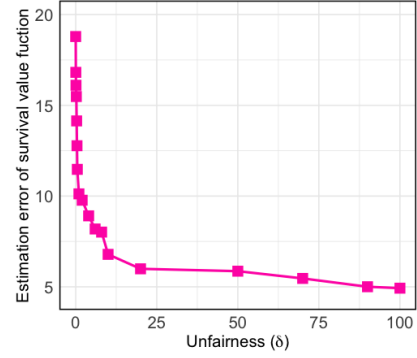

Figure 5: The estimation error of $\mathcal{V}(g^*(X, S))$ with respect to the unfairness in the ACTG dataset.

Acknowledgements

References

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633.

Ballarini, N. M., Rosenkranz, G. K., Jaki, T., König, F., and Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PloS one*, 13(10):e0205971.

Byun, Y., Sam, D., Oberst, M., Lipton, Z., and Wilder, B. (2024). Auditing fairness under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 4339–4347. PMLR.

Chzhen, E. and Schreuder, N. (2022). A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211.

Cui, Y., Zhu, R., and Kosorok, M. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic journal of statistics*, 11(2):3927.

del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2020). Review of mathematical frameworks for fairness in machine learning.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Gervasi, S. S., Chen, I. Y., Smith-McLallen, A., Sontag, D., Obermeyer, Z., Vennera, M., and Chawla, R. (2022). The potential for bias in machine learning and opportunities for health insurers to address it: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs*, 41(2):212–218.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge university press.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, pages 841–860.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research*, pages 141–167.

Kennedy, E. H. (2017). Semiparametric theory. *arXiv preprint arXiv:1709.06418*.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 5.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

Kim, K. and Zubizarreta, J. R. (2023). Fair and robust estimation of heterogeneous treatment effects for policy learning. In *International Conference on Machine Learning*, pages 16997–17014. PMLR.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452.

Li, X., Zhou, Q., Wu, Y., and Yan, Y. (2023). Multicategory matched learning for estimating optimal individualized treatment rules in observational studies with application to a hepatocellular carcinoma study. *arXiv preprint arXiv:2302.05287*.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Luedtke, A. R. and Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713.

Mhasawade, V., Zhao, Y., and Chunara, R. (2021). Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666.

Mishler, A. and Kennedy, E. H. (2022). Fade: Fair double ensemble learning for observable and counterfactual outcomes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1053–1053.

Mishler, A., Kennedy, E. H., and Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400.

Moldovan, D. (2023). Algorithmic decision making methods for fair credit scoring. *IEEE Access*, 11:59729–59743.

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

Nabi, R., Malinsky, D., and Shpitser, I. (2019). Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. PMLR.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Paulus, J. K. and Kent, D. M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, 3(1):99.

Schnell, P. M., Tang, Q., Offen, W. W., and Carlin, B. P. (2016). A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*, 72(4):1026–1036.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2021). *Lectures on stochastic programming: modeling and theory*. SIAM.

Still, G. (2018). Lectures on parametric optimization: An introduction. *Optimization Online*, page 2.

Tang, Z., Zhang, J., and Zhang, K. (2023). What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Viviano, D. and Bradic, J. (2024). Fair policy targeting. *Journal of the American Statistical Association*, 119(545):730–743.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Zhang, N., Jiang, W., Zhang, Y., Song, T.-Q., Lv, J., Gu, J., Wu, Y., Qian, J., Tian, D., Guo, Q., et al. (2020). Individualised tailored assessment of therapeutic alternatives for hcc patients within the milan criteria. *Gut*, 69(10):1893–1895.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L.-J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539.

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

Zhao, Y.-Q., 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.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes] All statements of the assumptions are included in Section 3 and Appendix D

   (b) Complete proofs of all theoretical results. [Yes] All proofs are included in Appendix F.

   (c) Clear explanations of any assumptions. [Yes] All explanations of our assumptions are included in Section 3 and Appendix E.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Code and data are included in supplementary materials.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Training details are included in Appendix B.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Appendix B.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Computational details are included in Appendix B.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] Code and data are included in https://github.com/wangholly/FairHTEforCENSORED.

   (d) Information about consent from data providers/curators. [Yes]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes] The sensitive attributes in the dataset is explained in Section 4.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Simulation Study

In this section, we conduct a simulation study to demonstrate the proposed estimators' theoretical properties and finite-sample performance when dealing with right-censored survival outcomes. We generate synthetic data according to the following process:

$$S \sim \text{ Bernoulli } (0.5), [X_1, X_2]^\top \mid S \sim \mathcal{N}\left([0, 3S-1]^\top, I_2\right),$$
$$\mathbb{P}(A = 1 \mid X, S) = \text{ expit }\left(W^\top[1, 0, 0] + SX_1\right), A \in \{0, 1\},$$
$$\mu_A(X, S) = AX_2^3/2 + \log\left(SX_1^2 + 10\right) + \exp\left(-SX_2/5\right) + SX_1,$$
$$T^A = \mu_A(X, S) + \epsilon, \epsilon \sim N(0, 1), \quad T = T(1)A + T(0)(1 - A)$$

where expit and I_2 denote the inverse logit function and the 2×2 identity matrix, T^A is potential survival time, and the true survival time T using a Cox model with treatment-specific baseline hazard functions:

$$\lambda(u \mid A, X, S) = \lambda_0(u \mid A) \exp\left(T^A\right),$$

The end of the study is fixed to be 2.7, and the censoring time C is generated from an exponential distribution with parameter 0.2 to induce a 17.7% censoring rate. We impute the censoring time using the random survival forest as described in Section 2.2. Then, $\tau(X, S) = X_2^3/2$ and $g^*(X, S) = \mathbb{1}(X_2 > 0)$. Our results are presented in Figure 6.
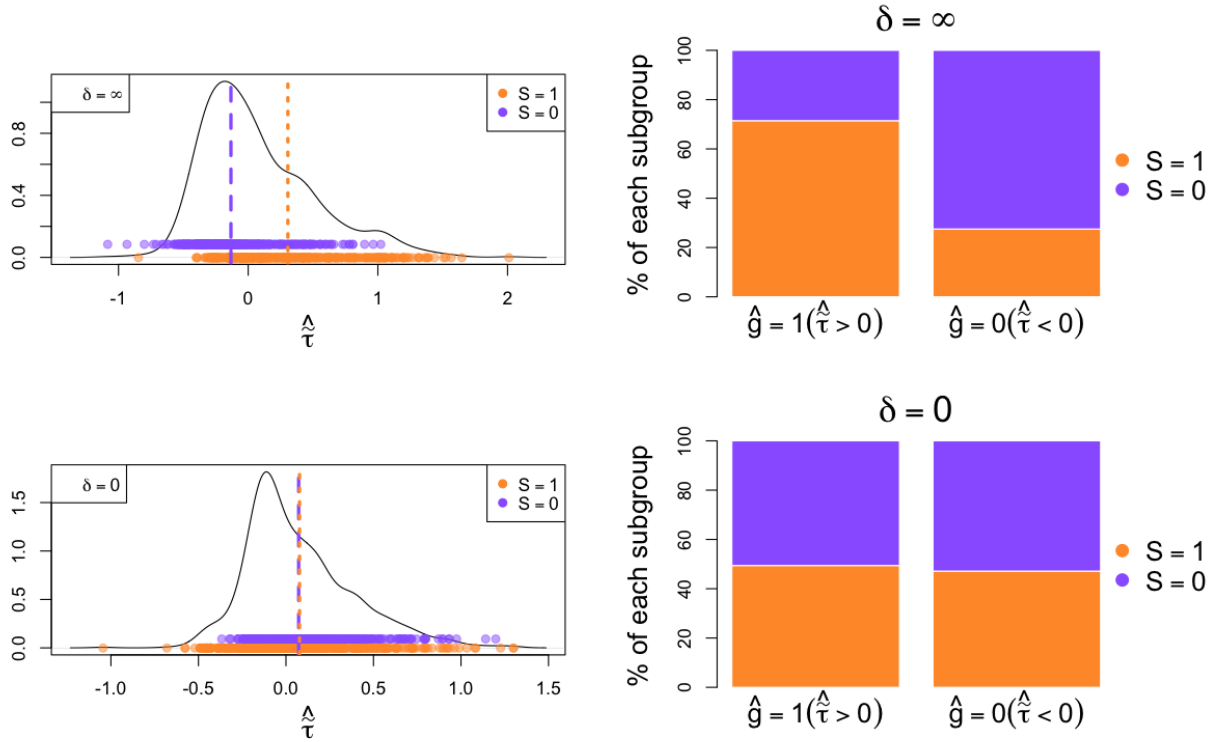

Figure 6: Densities (Left) and proportions for sensitive attribute S (Right) of the unfair data (when $\delta = \infty$) fair data (when $\delta = 0$).

We first assess the effectiveness of the proposed estimators in mitigating unfairness in the estimated CATE. To this end, we generate a sample of 2000 observations and estimate $\widehat{\widetilde{\tau}}$ with $K = 2$ splits under two conditions: $\delta = \infty$ (i.e., without fairness constraints) and $\delta = 0$ (i.e., with exact fairness constraints). Then, we generate a sample of 1000 observations as test set. In our simulation settings, setting $\delta = 90$ effectively means having no constraints, as it is equivalent to $\delta = \infty$, rendering the fairness constraint inactive. We then compare the
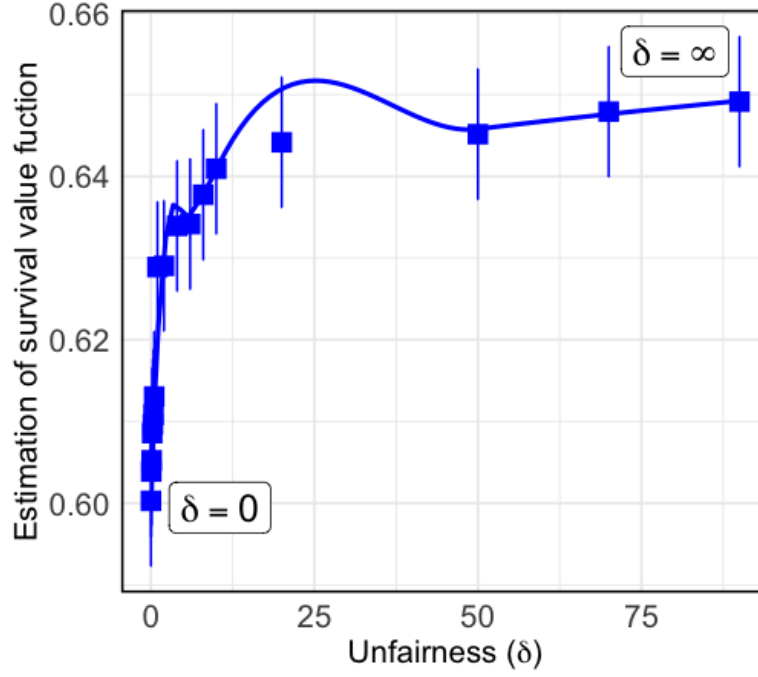
Figure 7: The cost of $\tilde{\mathcal{V}}(\widehat{q}(X,S))$ with respect to the unfairness in the simulation data.
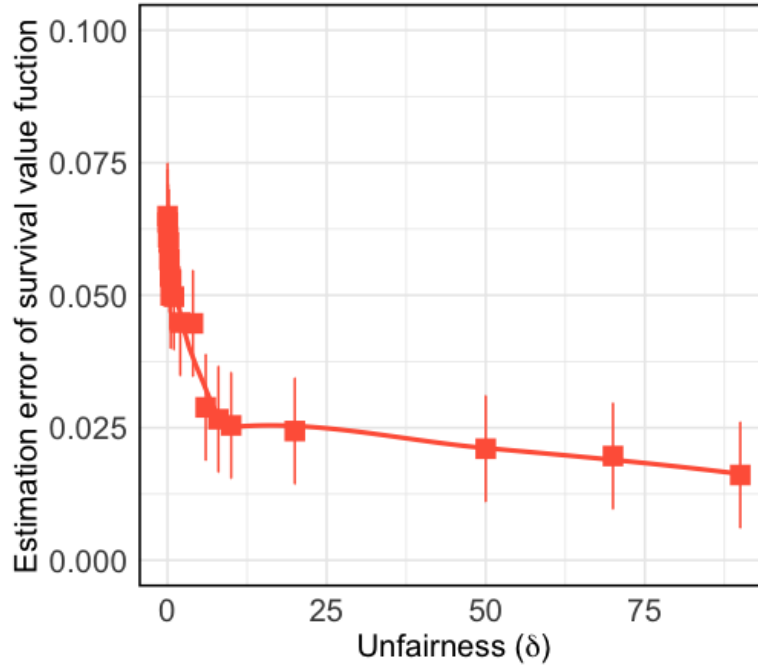

Figure 8: The estimation error of $\mathcal{V}(g^*(X,S))$ with respect to the unfairness in the simulation data.

densities of $\widehat{\tilde{\tau}}$ for each value of δ and illustrate how individuals from different groups are distributed in terms of S on the left of Figure 6. The right bar figures in Figure 6, we compute the proportions of $S=1$ to $S=0$ for each decision of $\widehat{g}(X,S) = \mathbb{1}\{\widehat{\tilde{\tau}}(X,S) > 0\}$ under both $\delta = \infty$ and $\delta = 0$ conditions.

The density plots in Figure 6, without fairness constraints, show a significant violation of the independence criterion in $\widehat{\tilde{\tau}}(X,S)$; individuals belonging to group $S=1$ (or $S=0$) are predominantly found in the treatment

strategy where $\widehat{\widetilde{\tau}}(X, S) > 0$ (or $\widehat{\widetilde{\tau}}(X, S) < 0$). This results in disproportionate (unfair) policies, as shown at the top of the bar plots in Figure 6, where individuals with $S = 1$ are more likely to be treated. However, this issue is largely resolved when the fairness constraint is applied with $\delta = 0$. In the lower-left corner of Figure 6, we observe that the conditional sample means of $\widehat{\widetilde{\tau}}(X, S)$ for $S = 0$ and $S = 1$ are nearly identical, and many individuals in the $S = 1$ (or $S = 0$) group are shifted to the left (or right) compared to the $\delta = \infty$ scenario. The results in policies where the treated and untreated groups are more balanced along the sensitive attribute, giving individuals with $S = 0$ a greater chance to be treated, as shown in the lower right corner of Figure 6.

To demonstrate the theoretical findings in Section 3, which state that the fairness in optimal policies comes at a price, we estimate $g^*(X, S)$ across different values of δ ranging between 0 and 90. This time, we also estimate the survival value function $\mathcal{V}(g^*(X, S))$ and the unfairness for different δ.

In Figures 7 and 8, we present the estimated survival time function and its associated errors across varying degrees of fairness constraints. The results show that the survival value function improves when we adopt a more lenient fairness tolerance (δ), substantiating the cost associated with achieving fairness through diminished survival time. This phenomenon aligns with Theorem 3.3, which articulates a trade-off between the accuracy of the estimated survival value function $\tilde{\mathcal{V}}(\hat{g}(X, S))$ and the degree of fairness enforced. This trade-off highlights the intrinsic costs of incorporating fairness into treatment decisions.

B   Experiment Details

In this section, we will present all the experimental details and some additional results to support the main text.

B.1   Data and algorithm details

We consider data from The AIDS Clinical Trials Group Study 175 Dataset (ACTG) which evaluated treatment with either a single nucleoside or a combination of two nucleosides in adults infected with human immunodeficiency virus type 1 (HIV-1), whose CD4 cell counts ranged from 200 to 500 cells per cubic millimeter. The dataset contains 2,139 instances and 23 features. The covariates include age, weight, hemophilia, homosexual activity, history of IV drug use, Karnofsky score, Non-ZDV antiretroviral therapy pre-175, ZDV in the 30 days prior to 175, ZDV prior to 175, days pre-175 anti-retroviral therapy, race, gender, antiretroviral history, antiretroviral history stratification, symptomatic indicator, treatment indicator, censoring indicator, indicator of off-trt before 96+/-5 weeks, CD4, CD4 at 20+/-5 weeks, CD8, CD8 at 20+/-5 weeks. We use roughly 75% of the data to estimate g^* (and therefore $\mathcal{V}(g^*)$), reserving the remaining 25% as a testing set.

We use the `randomForestSRC` package in R to impute censored data. For computing the convex quadratic program, we employ the `quadprog` solver in R.

B.2   Additional results

To illustrate the normality convergence of $\hat{\beta}$ in Theorem 3.1, we plot the density of some arbitrary linear combination of $\hat{\beta}$ in Figure 9.

B.3   Computational details

All experiments were performed using R version 2024.09.0+375 on a Mac with an Apple M2 Pro chip and 16 GB of RAM. The average time spent for the training process to obtain an estimation of CATE is around 14 seconds on the ACTG dataset.

C   Experiment on HCC Data

Hepatocellular carcinoma (HCC), the most common form of liver cancer, accounts for approximately 75% of all liver cancer cases. We analyzed data from an observational study of HCC that enrolled 537 patients who met the Milan criteria between 2009 and 2019 in Tianjin, China (Zhang et al., 2020). The study investigated three treatment modalities: liver transplantation (LT), liver resection (LR), and locoregional ablation (LA). Among the patients, 172 underwent LT, 191 underwent LR, and 174 underwent LA. According to Li et al. (2023), LT
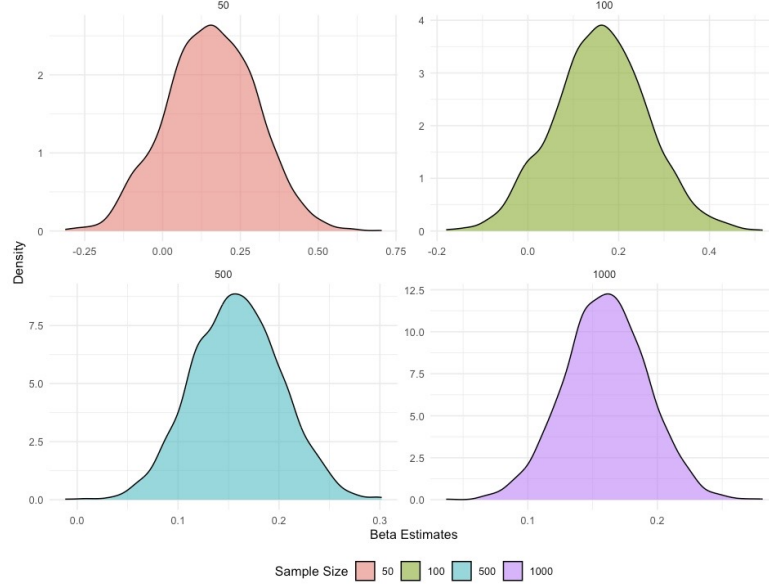
Figure 9: Density functions of some arbitrary linear combinations of $\hat{\beta}$.

is associated with the most favorable treatment effect, whereas LR is associated with the least favorable effect. Consequently, we excluded data from patients treated with LA, resulting in a censoring rate of 73.46%. We then recoded the treatment types, assigning LT as 1 and LR as 0. Next, we used the `randomForestSRC` package in R to impute the censored data. We allocate approximately 75% of the data to estimate g^*, and use the remaining portion as a testing set.

Figure 10 reveals that the proportion of males ($S = 1$) receiving treatment 1 (left bar plot) is significantly lower than that of females ($S = 0$), indicating potential unfair treatment towards the male group under the optimal treatment policy. In this scenario, setting $\delta = 90$ is equivalent to having no fairness constraint ($\delta = \infty$), as the constraint becomes inactive. Using the estimated optimal treatment strategy $\hat{g}$ from our method, Figure 11 shows that the proportions of males and females in each treatment group are nearly equal, demonstrating the fairness achieved through our estimation process.

To underscore this fairness, we compare the density functions of the estimated treatment effect $\widehat{\widehat{\tau}}(X, S)$ in Figure 12. The bottom plot applies a fairness constraint ($\delta = 0$), while the top plot does not ($\delta = \infty$). In the top plot, where no fairness constraint is applied, the center of the estimated treatment effect for males ($S = 1$) is significantly lower than for females ($S = 0$), suggesting that male patients are less likely to receive treatment 1 compared to female patients. This disparity indicates an unfair estimation of treatment effects, which could lead to biased treatment decisions. In contrast, in the bottom plot, where the fairness constraint ($\delta = 0$) is enforced, the centers of the estimated treatment effects for both genders align, demonstrating an equitable distribution of treatment effects. Additionally, these centers align with the overall density center of $\widehat{\widehat{\tau}}(X, S)$, further validating the effectiveness of our proposed method in achieving fair treatment effect estimations for $\tau(X, S)$.

D Regularity Assumptions and Definitions

In this section, we will provide some common used regularities and definitions in constraint nonlinear optimization, we refer Still (2018) for more details of these regularities and definitions.

Let $\mathcal{C}_{fair} := \{\beta | \mathbb{E}[\mathrm{UF}_j(Z)\beta^\top \mathbf{b}(X, S)]| \le \delta_j, j \in J\}$. Define a set of inequality constraints $\mathcal{C} = \{\beta | g_j(\beta) \le 0, 1 \le j \le M\}$. For some specific points $\bar{\beta} \in \mathcal{C}$, we define the active index set as follows.

Definition D.1 (Active set). J_0 is called an active index set if $J_0(\bar{\beta}) = \{1 \le j \le M | g_j(\bar{\beta}) = 0\}$.

Definition D.2 (LICQ). The condition, linear independent constraint qualification (LICQ) is satisfied with respect to the optimal problem (P) at $\bar{\beta} \in \mathcal{S}$ if the vectors $\Delta_\beta g_j(\bar{\beta})$, $j \in J_0$ are linear independent.

Definition D.3 (SC). Let $L(\beta, \gamma)$ be the Lagrangian. The condition, strict complementarity (SC) is satisfied
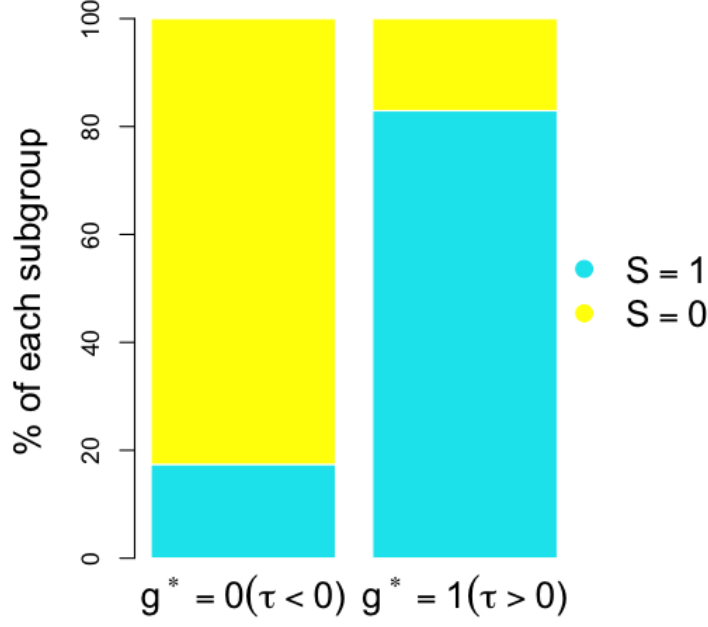
Figure 10: Unfairness proportions with respect to gender (S) for each treatment strategy in HCC Dataset.

with respect to the optimal problem (P) at $\bar{\beta} \in \mathcal{S}$ if, with multipliers $\bar{\gamma}_j \geq 0$, $j \in J_0(\bar{\beta})$, the Karush-Kuhn-Tucker (KKT) condition

$$\Delta_\beta L(\bar{\beta}, \bar{\gamma}) = \Delta_\beta \mathcal{L}(\bar{\beta}) + \sum_{j \in J_0(\bar{\beta})} \bar{\gamma}_j \Delta_\beta g_j(\bar{\beta}) = 0\,,$$

is satisfied.

LICQ is arguably one of the most widely-used constraint qualifications that admit the first-order necessary conditions. SC means that if the j-th inequality constraint is active then the corresponding dual variable is strictly positive, so exactly one of them is zero for each $1 \leq j \leq m$.

E Discussion of assumption (A1)

Assumption (A1) can be replaced by the following weaker assumption with the second-order condition (Shapiro et al., 2021):

(A1') For each optimal solution β^* of (P), $\eta^\top \mathbb{E}[\mathbf{b}(X, S)^\top \mathbf{b}(X, S)]\eta > 0$, for

$$\forall \eta \in \{\eta \in \mathbb{R}^k | \mathbf{b}(X, S)^\top \eta \leq 0, \mathrm{UF}_j(Z)\mathbf{b}(X, S)^\top \eta \leq 0, \quad j \in J_0(\beta^*)\}\backslash\{0\}.$$

Assumption (A') implies β^* is locally isolated and therefore the quadratic growth condition holds at β^* (Shapiro et al., 2021).

F Proofs

This section includes all the proofs of the main results in the main text.

Lemma F.1. *Under assumption (A1) or (A1'), we have*

$$||\widehat{\beta} - \beta^*|| = O_\mathbb{P}\left(\max\left\{||\widehat{C} - C||_2, ||\widehat{L} - L||_F\right\}\right)\,, \tag{17}$$
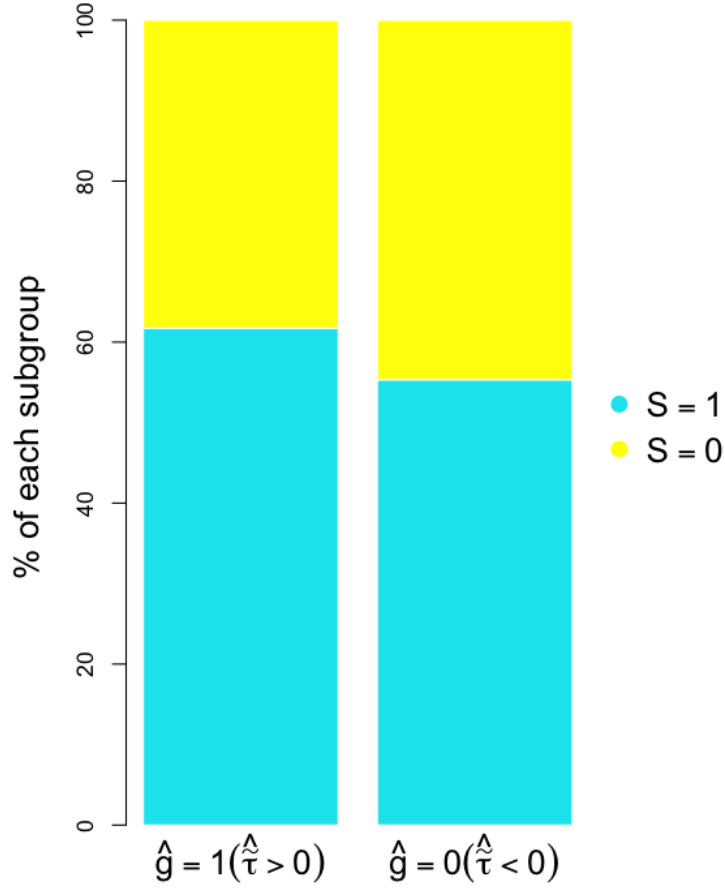
Figure 11: Fairness proportions with respect to gender (S) for each treatment strategy in HCC Dataset.

where

(1). $C = \left\{ \mathbb{E}[\mathbf{b}(X,S)\mathbf{b}(X,S)^\top]_{i,l}, \mathbb{E}[(Y(1) - Y(0))\mathbf{b}(X,S)]_i, \mathbb{E}[\epsilon(X,S)\mathbf{b}(X,S)]_i \right\}_{i,j=1}^k$, *and its estimator*
$\widehat{C} = \left\{ \mathbb{P}_n[\mathbf{b}(X,S)\mathbf{b}(X,S)^\top]_{i,l}, \mathbb{P}_n \left\{ (\varphi_1(Z;\widehat{\eta}_{-B}) - \varphi_0(Z;\widehat{\eta}_{-B})) \, \boldsymbol{b}(X,S) \right\}, \mathbb{P}[\epsilon(X,S)\mathbf{b}[W]] \right\}_{i,j=1}^k$.

(2). $L = [\mathbb{E}[\mathrm{UF}_1\mathbf{b}(X,S)], \cdots, \mathbb{E}[\mathrm{UF}_m\mathbf{b}(X,S)]]^\top$, *and its estimator*
$\widehat{L} = \left[\mathbb{P}_n\left[\widehat{\mathrm{UF}}_1\mathbf{b}(X,S) \right], \cdots, \mathbb{P}_n\left[\widehat{\mathrm{UF}}_m\mathbf{b}(X,S) \right] \right]^\top$.

Proof. First, we rewrite the original optimal problem (P) as

$$\min_{\beta \in \mathbb{R}^k} f(\beta, C) \text{ s.t. } L\beta \leq \delta, \tag{18}$$

where $\delta = [\delta_1, \cdots, \delta_m]^\top$. Similarly, the estimated optimal problem ($\widehat{P}$) can be rewritten as

$$\min_{\beta \in \mathbb{R}^k} f(\beta, \widehat{C}) \text{ s.t. } \widehat{L}\beta \leq \delta. \tag{19}$$

Under assumption (A1) or (A1'), the second order condition holds for the optimal β^* in (18), therefore, β^* is Lipschitz stable by Theorem 6.4 in Still (2018). This implies the result immediately. □

Hongni Wang*, Junxi Zhang*, Na Li, Linglong Kong, Bei Jiang[†], Xiaodong Yan[†]
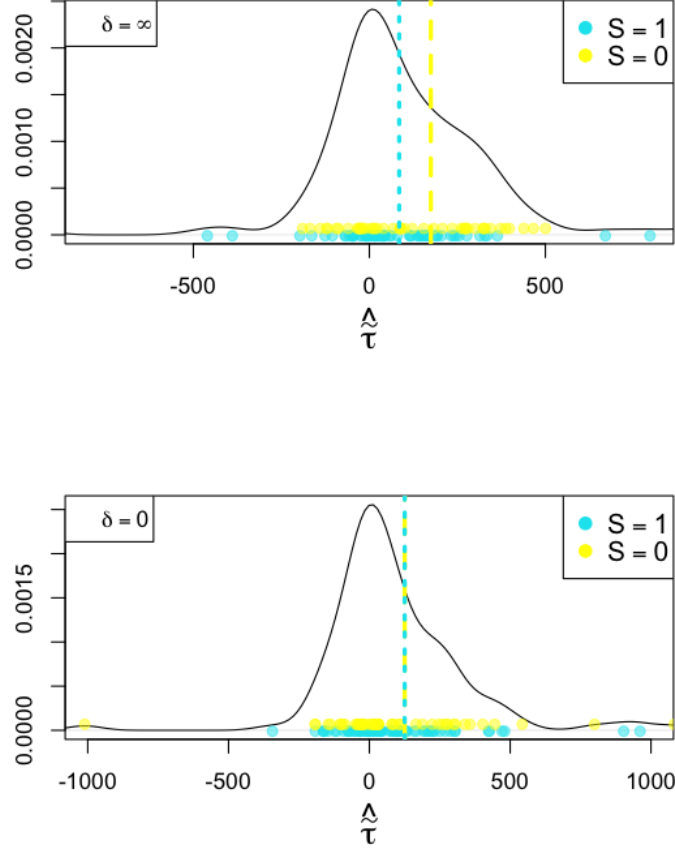
Figure 12: Densities of $\widehat{\widetilde{\tau}}(X, S)$ without a fairness constraint (when $\delta = \infty$) and with a fairness constraint (when $\delta = 0$) in the HCC dataset. The two vertical dashed lines correspond to $\mathbb{P}_n(\widehat{\widetilde{\tau}}(X, S) \mid S = 0)$ and $\mathbb{P}_n(\widehat{\widetilde{\tau}}(X, S) \mid S = 1)$.

To proof Theorem 3.1, we need to characterize the order of the right hand side in (17). This can be done by the following lemma.

Lemma F.2. *Under conditions (A1)-(A3) and (A5), we have*

$$||\widehat{C} - C||_2 = O_{\mathbb{P}}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right)$$

$$||\widehat{L} - L||_F = O_{\mathbb{P}}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right).$$

Proof. This proof follows directly from the doubly robust estimation of the expectation of potential outcome using efficient influence function. For instance, by Section 4.4 in Kennedy (2016) and the fact that $|\mathbb{P}_n(h(X, S)) -$

$\mathbb{E}[h(X,S)]| = O_{\mathbb{P}}(\frac{1}{\sqrt{n}})$ (the central limit theorem) for any $h(X,S)$,

$$||\widehat{C} - C||_2 = (\mathbb{P}_n - \mathbb{P})(\varphi(Z)) + O_{\mathbb{P}}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right)$$

$$= O_{\mathbb{P}}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right).$$

$$||\widehat{L} - L||_F = (\mathbb{P}_n - \mathbb{P})(\varphi(Z)) + O_{\mathbb{P}}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right)$$

$$= O_{\mathbb{P}}\left(\max_a ||\widehat{\pi}_a - \pi_a||_{2,\mathbb{P}}||\widehat{\widetilde{\mu}}_a - \widetilde{\mu}_a||_{2,\mathbb{P}} + \frac{1}{\sqrt{n}}\right).$$

The results in this lemma follows directly by the central limit theorem. $\qquad\square$

Proof of Theorem 3.1:

Proof. Result (i) in Theorem 3.1 follows from Lemma F.1 and F.2. Under additional assumption (A4), one has $\widehat{\beta}$ is a $\sqrt{n}-$consistent estimator of β^*. With assumption (A6), LICQ and SC assumption, we have asymptotic normality for both elements $||\widehat{C} - C||_2$ and $||\widehat{L} - L||_F$, which implies the asymptotic normality of $||\widehat{\beta} - \beta^*||_2$. The variance is $\sigma_0 = \begin{bmatrix} \nabla_\beta^2 L(\beta^*, \lambda^*) & A^\top \\ A & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}^\top \gamma$. In this expression, $\gamma \sim N(0, \sigma^2)$ is the limiting distribution of assumption (A6) above Theorem 3.1. $L(\beta, \lambda)$ is the Lagrangian associate with problem (P), β^* and λ^* are the true solutions of $L(\beta, \lambda)$. $A = [\nabla_\beta g_j(\beta^*) : j \in J_0(\beta^*)]$, where $g_j(\beta) = \mathbb{E}\{\mathrm{UF}_j(Z)\beta^\top \mathbf{b}(X,S)\}$ and $J_0(\beta)$ is the active index set. $\qquad\square$

Proof of Proposition 3.2:

Proof. Let $\widetilde{g}(X,S) = \mathbb{1}\{\widetilde{\tau}(X,S) > 0\}$ to be the imputed optimal solution of the imputed value function $\widetilde{\mathcal{V}}(g(X,S)) = \mathbb{E}[\widetilde{T}(1)g(X,S) + \widetilde{T}(0)(1 - g(X,S))]$. The optimal value function is therefore denoted as $\mathcal{V}(\widetilde{g}(X,S))$. Therefore,

$$|\mathcal{V}(g^*(X,S)) - \widetilde{\mathcal{V}}(\widehat{g}(X,S))| \le |\mathcal{V}(g^*(X,S)) - \widetilde{\mathcal{V}}(g^*(X,S))| + |\widetilde{\mathcal{V}}(g^*(X,S)) - \widetilde{\mathcal{V}}(\widehat{g}(X,S))|. \tag{20}$$

For the first part in (20), we have

$$|\mathcal{V}(g^*(X,S)) - \widetilde{\mathcal{V}}(g^*(X,S))| = |\mathbb{E}[(Y(1) - \widetilde{Y}(1))g^*(X,S) + (Y(0) - \widetilde{Y}(0))(1 - g^*(X,S))]|$$
$$\le |\mathbb{E}[\epsilon(X,S)g^*(X,S)]| + |\mathbb{E}[\epsilon(X,S)g^*(X,S)]| \le 2||\epsilon(X,S)||_{1,\mathbb{P}}.$$

For the second therm in (20), recall that

$$|\widetilde{\mathcal{V}}(g^*(X,S)) - \mathcal{V}(\widehat{g}(X,S))| = |\mathbb{E}[\widetilde{\tau}(X,S)(\mathbb{1}\{\widehat{\widetilde{\tau}}(X,S) > 0\} - \mathbb{1}\{\widetilde{\tau}(X,S) > 0\})]|$$
$$\le \mathbb{E}[|\widetilde{\tau}(X,S)|\mathbb{1}\{|\widetilde{\tau}(X,S)| \le |\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\}],$$

where the last inequality holds since $|\mathbb{1}\{\widehat{\widetilde{\tau}}(X,S) > 0\} - \mathbb{1}\{\widetilde{\tau}(X,S) > 0\}| \le \mathbb{1}\{|\widetilde{\tau}(X,S)| \le |\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\}$. Therefore, by Lemma 5.1 in Audibert and Tsybakov (2007)

$$|\widetilde{\mathcal{V}}(g^*(X,S)) - \mathcal{V}(\widehat{g}(X,S))| \le ||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{\infty,\mathbb{P}}\mathbb{P}(\{|\widetilde{\tau}(X,S)| \le |\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\})$$
$$\lesssim ||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{\infty,\mathbb{P}}^{\alpha+1}.$$

This proof case when $q = \infty$ and $\gamma = \alpha + 1$.

On the other hand, by Lemma 5.2 in Audibert and Tsybakov (2007), we have for any $d > 0$,

$$|\widetilde{\mathcal{V}}(g^*(X,S)) - \mathcal{V}(\widehat{g}(X,S))| \le \mathbb{E}[|\widetilde{\tau}(X,S)|\mathbb{1}\{|\widetilde{\tau}(X,S)| \le |\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\}\mathbb{1}\{|\widetilde{\tau}(X,S)| \le d\}]$$
$$+ \mathbb{E}[|\widetilde{\tau}(X,S)|\mathbb{1}\{|\widetilde{\tau}(X,S)| \le |\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\}\mathbb{1}\{|\widetilde{\tau}(X,S)| > d\}]$$
$$\le \mathbb{E}[|\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\mathbb{1}\{|\widetilde{\tau}(X,S)| \le d\} + \mathbb{E}[|\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|\mathbb{1}\{|\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)| > d\}.$$

Applying the Hölder inequality and the Markov inequality will lead to

$$|\widetilde{\mathcal{V}}(g^*(X,S)) - \mathcal{V}(\widehat{g}(X,S))|$$

$$\leq ||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{q,\mathbb{P}} \mathbb{P}(|\widetilde{\tau}(X,S)| \leq d)^{\frac{q-1}{q}} + ||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{q,\mathbb{P}} \left(\frac{\mathbb{P}(|\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)|)^q}{d^q} \right)^{\frac{q-1}{q}}$$

$$\lesssim ||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{q,\mathbb{P}} d^{\frac{q-1}{q}} + ||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{q,\mathbb{P}}^q d^{1-q},$$

which follows from the margin condition. The last expression attains its maximal at $||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{q,\mathbb{P}}^{\frac{q(1+\alpha)}{q+\alpha}}$ when $d = O(||\widetilde{\tau}(X,S) - \widehat{\widetilde{\tau}}(X,S)||_{q,\mathbb{P}}^{\frac{q}{q+\alpha}})$. This proves the case when $q \in [1,\infty)$ and $\gamma = \frac{q(\alpha+1)}{q+\alpha}$. $\qquad\square$

Proof of Theorem 3.3:

Proof. From the result of Proposition 3.2, we have

$$|\mathcal{V}(g^*(X,S)) - \widetilde{\mathcal{V}}(\widehat{g}(X,S))| \lesssim \| \widehat{\beta}^\top \mathbf{b}(X,S) - \widetilde{\tau}(X,S) \|_{q,\mathbb{P}}^\gamma + \| \epsilon(X,S) \|_{1,\mathbb{P}}$$

$$\lesssim \| \widetilde{\beta}^\top \mathbf{b}(X,S) - \widetilde{\tau}(X,S) \|_{q,\mathbb{P}}^\gamma + \| \widehat{\beta}^\top \mathbf{b}(X,S) - \beta^* \mathbf{b}(X,S) \|_{q,\mathbb{P}}^\gamma + \| \widehat{\beta}^\top \mathbf{b}(X,S) - \beta^* \mathbf{b}(X,S) \|_{q,\mathbb{P}}^\gamma + \| \epsilon(X,S) \|_{1,\mathbb{P}},$$

where the third term is less or equal to $T_{1,n}^\alpha$ under the assumptions by Theorem 3.1 and the Minkowski's inequality. Therefore, we will focus on the second term $\| \widehat{\beta}^\top \mathbf{b}(X,S) - \beta^* \mathbf{b}(X,S) \|_{q,\mathbb{P}}^\gamma$.

Recall that

$$\widetilde{\beta} = \arg\min_{\beta \in \mathbb{R}^k} \mathbb{E} \left[\left\{ \widetilde{T}(1) - \widetilde{T}(0) + \epsilon(X,S) - \beta^\top \boldsymbol{b}(X,S) \right\}^2 \right] \tag{21}$$

. On the other hand, since the constraint set in (P) includes only linear constraints, thus strong duality holds, and there is a dual solution $\lambda = (\lambda_j)_{j=1}^m$ to (P) such that any solution of (P) is also a solution of

$$\arg\min_{\beta \in \mathbb{R}^k} \mathbb{E} \left[\left\{ \widetilde{T}(1) - \widetilde{T}(0) + \epsilon(X,S) - \beta^\top \boldsymbol{b}(X,S) \right\}^2 \right] + \sum_{j=1}^m \lambda_j [\beta^\top \mathbb{E}[\mathrm{UF_j}(\mathrm{Z}) \mathbf{b}(X,S)]]^2.$$

Let $F(\beta,\theta) := \mathbb{E} \left[\left\{ \widetilde{T}(1) - \widetilde{T}(0) + \epsilon(X,S) - \beta^\top \boldsymbol{b}(X,S) \right\}^2 \right] + \sum_{j=1}^m (\beta^\top \theta_j)^2$, and denote the solution of $\arg\min_\beta F(\beta,\theta)$ as $\beta(\theta)$. Then, $\widetilde{\beta} = \beta(0)$ and $\beta^* = \beta(\theta^*)$, where $\theta_j^* = \sqrt{\lambda_j} \mathbb{E}[\mathrm{UF_j}(\mathrm{Z}) \mathbf{b}(X,S)]$ for $j = 1, \cdots, m$.

By assumption (A1), the optimization (21) is strongly convex, and thus the quadratic growth condition holds at $\widetilde{\beta}$. Therefore, for all β we have $c||\widetilde{\beta} - \beta||_2^2 \leq F(\beta,0) - F(\widetilde{\beta},0)$, where $c > 0$ is some constant. We can take $\beta = \beta^*$, which gives

$$c||\widetilde{\beta} - \beta^*||_2^2 \leq F(\beta^*,0) - F(\widetilde{\beta},0)$$
$$= F(\beta^*,0) - F(\beta^*,\theta^*) + F(\beta^*,\theta^*) - F(\widetilde{\beta},\theta^*) + F(\widetilde{\beta},\theta^*) - F(\widetilde{\beta},0)$$
$$\leq \left(||\widetilde{\beta}||_2^2 + ||\beta||_2^2 \right) ||\theta^*||_2^2,$$

where the last inequality follows from the definition of F and the fact that $F(\beta^*,\theta^*) \leq F(\widetilde{\beta},\theta^*)$. Then by the Cauchy-Schwarz and the Jensen's inequality, we have

$$||\widetilde{\beta} - \beta^*||_2^2 \lesssim \left\| \mathbb{E} \left[\left(\sum_{j=1}^m \sqrt{\lambda_j} \mathrm{UF_j}(\mathrm{Z}) \right) \mathbf{b}(X,S) \right] \right\|_2^2$$

$$\leq \left\| \sum_{j=1}^m \sqrt{\lambda_j} \mathrm{UF_j}(\mathrm{Z}) \right\|_{2,\mathbb{P}}^2 ||\mathbb{E}[\mathbf{b}(X,S)]||_2^2$$

$$\leq \left\| \sum_{j=1}^m \sqrt{\lambda_j} \mathrm{UF_j}(\mathrm{Z}) \right\|_{2,\mathbb{P}}^2 ||\mathbf{b}(X,S)||_{2,\mathbb{P}}^2.$$

This proves the form of T_2, therefore proves result (i). The result (ii) can be proved by the same logic and Lemma 5.1 in Audibert and Tsybakov (2007), i.e.,

$$\mathbb{P}\left(g^*(X,S) \neq \widehat{g}(X,S)\right) \leq \parallel \epsilon(X,S) \parallel_{1,\mathbb{P}} + \parallel \widehat{\beta}^{\top}\mathbf{b}(X,S) - \widetilde{\tau}(X,S) \parallel_{\infty,\mathbb{P}}^{\alpha},$$

where the second term on the right hand side can be handled similarly as the proof of result (i). $\qquad\square$