Journal Title
XX(X):1–31

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A Reference-Free R-Learner for Treatment Recommendation

Junyi Zhou[1], Ying Zhang[2], and Wanzhu Tu[3]

Abstract

Assigning optimal treatments to individual patients based on their characteristics is the ultimate goal of precision medicine. Deriving evidence-based recommendations from observational data while considering the causal treatment effects and patient heterogeneity is a challenging task, especially in situations where multiple treatment options exist. Herein, we propose a reference-free R-learner based on a simplex algorithm for treatment recommendation. We showed through extensive simulation that the proposed method produced consistent and accurate recommendations that corresponded to optimal treatment outcomes. We used the method to analyze data from the Systolic Blood Pressure Intervention Trial (SPRINT) and achieved recommendations consistent with the current clinical guidelines.

Keywords

Heterogeneous treatment effect, R-learner, simplex, treatment recommendation

[1]Amgen Inc., U.S.
[2]University of Nebraska Medical Center, U.S.
[3]Indiana University School of Medicine and Fairbanks School of Public Health, U.S.

Corresponding author:
Ying Zhang, Department of biostatistics, University of Nebraska Medical Center, Nebraska, U.S.
Email: ying.zhang@unmc.edu

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

Tailoring pharmacological treatment in individual patients for maximized therapeutic effect is an essential feature of precision medicine. However, identifying individuals that are most likely to causally respond to specific therapies can be challenging, especially in situations with multiple treatment options. Among other things, recommendations must be made on a sound understanding of the *causal* effects of the treatments. At the same time, patients heterogeneity must be taken into account in determining treatment optimality.

There is a sizable literature on causal estimation of heterogeneous treatment effects (HTE) given the observed patient characteristics. Most of the methods were developed for head-to-head comparisons of treatments, and they are generally within the Rubin-Neyman's potential outcome framework (1; 2). One commonly used strategy is to estimate the response surface under a given treatment, and then derive the comparative effect by taking the difference between two estimated surfaces (3; 4; 5; 6). Under the general umbrella of the Q-learning (7; 8), various learning algorithms have been developed for HTE *estimation* (9). While intuitive in concept, these algorithms provide no performance guarantee in the presence of strong group size variation, especially when the response surfaces are trained separately (9). An alternative approach is to directly target the contrast of treatment effects and use the combined sample to alleviate the estimation bias; the approach is commonly referred to as the Advantage- or A-learning (10; 11; 12; 13). The recently proposed R-learner is one such method (14). By adopting Robinson's decomposition (15), from which it has derived its name, the R-learner expresses the HTE as a function of the observed responses and the treatment propensities. Because the overall response surface and treatment propensity function are unknown and need be estimated prior to ascertaining the least-square estimates for the HTE, the R-learner is usually implemented in a two-step procedure.

For treatment effect estimation, the original R-learner compares two treatments. While the procedure can be extended to comparisons of multiple treatments against a *prespecified* reference group, HTE estimates tend to vary with reference group selection. With multiple treatment groups, as we shall demonstrate in the simulation study, different selection can lead to inconsistent recommendations, thus causing difficulties in therapeutic optimization.

In this research, we modify the R-learner for the purpose of treatment recommendation, where the goal is to identify an optimal treatment from multiple therapeutic options, instead of focusing on pairwise treatment effect comparisons. To that end, we present a reference-free version of the R-learner to address the problem of recommendation inconsistency. Importantly, the new learner still permits comparisons between any arbitrarily chosen pairs of treatments. Implementation of the proposed method, however, requires numerical optimization under a large number of constraints, a problem we solve by using a simplex algorithm (16). We showed that the new learner was able to remove recommendation inconsistency while maintaining the HTE estimation performance.

The rest of the article is organized as follows. In Section 2, we first extend the original R-learner for HTE estimation to situations of multiple treatments, and present the reference-free R-learner for treatment recommendation. In Section 3, we present a simplex computational algorithm for implementation of the new method. In Section 4, we evaluate the numerical performance of the proposed method through an extensive simulation study. In Section 5, we illustrate the use of the method in a real clinical study for optimizing antihypertensive treatments. We end the article in Section 6 with a few remarks on the proposed method. Technical details and additional information are provided in the Appendix.

Methods

The R-Learner for Two Treatments

We follow the notation used in the Rubin-Neyman's potential outcome framework (2; 17). Let $\mathbf{X}$, Y and $T \in \{0, 1\}$ be a p-dimensional vector of patient characteristics, the observed outcome, and treatment assignment for a given patient, respectively. Here $Y^{(1)}$ and $Y^{(0)}$ are the *potential* outcomes from two available treatments, respectively labeled as 1 and 0. Since only one of the potential outcomes is observed, we write the observed outcome as $Y = 1[T = 1]Y^{(1)} + 1[T = 0]Y^{(0)}$. The HTE is defined as the expected difference between the two potential outcomes conditional on the covariates, denoted as $\tau(\mathbf{x}) = E[Y^{(1)} - Y^{(0)} \mid \mathbf{X} = \mathbf{x}]$.

In this research, we use the classical assumptions in the causal inference literature: (1) Ignorability – treatment assignment T is independent of potential outcomes $Y^{(1)}, Y^{(0)}$ given covariates $\mathbf{X}$; (2) Positivity – the propensity score $\pi(\mathbf{X}) = \Pr(T = 1 \mid \mathbf{X}) \in (0, 1)$; (3) Stable unit

treatment value assumption (SUTVA), which stipulates that the outcome in one subject depends only on the treatment that subject receives, not the treatments that others receive.

Under the ignorability assumption $\{Y^{(1)}, Y^{(0)}\} \perp T \mid \mathbf{X}$ and the Robinson's decomposition (15), the R-learner connects the HTE to the observed outcome through (14)

$$E[Y \mid \mathbf{X}, T] = m(\mathbf{X}) + (1[T = 1] - \pi(\mathbf{X}))\tau(\mathbf{X}), \tag{1}$$

where $m(\mathbf{X}) = E[Y \mid \mathbf{X}]$ is the conditional mean of Y given covariates $\mathbf{X}$, often referred to as the overall response surface, and $\pi(\mathbf{X}) = \Pr(T = 1 \mid \mathbf{X})$ is the covariate-specific propensity function for treatment 1.

Given $m(\mathbf{X})$ and $\pi(\mathbf{X})$, the HTE $\tau(\cdot)$ minimizes the mean-squares loss function

$$\ell(\tau) = E\left[(Y - m(\mathbf{X}) - (1[T = 1] - \pi(\mathbf{X}))\tau(\mathbf{X}))^2\right]. \tag{2}$$

Since $m(\cdot)$ and $\pi(\cdot)$ are rarely known in practice, a two-step procedure is often needed for the optimization of (2), by first estimating $m(\cdot)$ and $\pi(\cdot)$, and then plugging in the estimated $\hat{m}(\cdot)$ and $\hat{\pi}(\cdot)$ into (2) to ascertain $\tau(\cdot)$ that minimizes the empirical version of (2)

$$\hat{\tau}(\mathbf{X}) = \arg\min_{\tau} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}(\mathbf{X}_i) - (1[T_i = k] - \hat{\pi}(\mathbf{X}_i))\tau(\mathbf{X}_i))^2,$$

with observed dataset $\{(\mathbf{X}_i, Y_i, T_i)_{i=1}^n\}$.

Therapeutic recommendation is based on the estimated treatment effects of the drugs, given the observed patient characteristics. In a situation of two treatments, without loss of generality, we assume that the drug with a larger expected value of the conditional potential outcome is preferred. The optimal treatment k^* associated with covariates $\mathbf{X}$ is therefore determined by the sign of the estimated treatment effect

$$k^* = \begin{cases} 1 & \text{if } \hat{\tau}(\mathbf{X}) > 0, \\ 0 & \text{if } \hat{\tau}(\mathbf{X}) < 0. \end{cases}$$

The R-Learner for Multiple Treatments

The original R-learner is proposed for the estimation of treatment effect between two therapies. The method can be readily extended to comparisons of multiple therapies against a preselected reference therapy.

Suppose there exist K treatments, the potential outcomes can be written as $Y^{(k)}$ for $T \in \{1, 2, ..., K\}, K > 2$, and $k = 1, 2, ..., K$. We write the HTE of Treatment k in comparison with the reference treatment j as $\tau_j^{(k)}(x) = E[Y^{(k)} - Y^{(j)} \mid \mathbf{X}], k \neq j$, where subscript j represents the reference treatment group. The propensity function of treatment k is defined as $\pi^{(k)}(\mathbf{X}) = \Pr(T = k \mid \mathbf{X})$.

As in the two-treatment scenario, the R-learner with multiple treatment groups connects the HTE to the observed outcome through

$$E[Y \mid \mathbf{X}, T] = m(\mathbf{X}) + \sum_{k \neq j} (1[T = k] - \pi^{(k)}(\mathbf{X}))\tau_j^{(k)}(\mathbf{X}).$$

Given $m(\mathbf{X})$ and $\boldsymbol{\pi}(\mathbf{X}) = (\pi^{(1)}(\mathbf{X}), \pi^{(2)}(\mathbf{X}), ..., \pi^{(K)}(\mathbf{X}))$, the HTE

$$\boldsymbol{\tau}_j(\cdot) = \left(\tau_j^{(1)}(\cdot), ..., \tau_j^{(j-1)}(\cdot), \tau_j^{(j+1)}(\cdot) ..., \tau_j^{(K)}(\cdot)\right)$$

minimizes the mean squares loss function,

$$\arg\min_{\tau_j} E\left[\left(Y - m(\mathbf{X}) - \sum_{k \neq j} (1[T = k] - \pi^{(k)}(\mathbf{X}))\tau_j^{(k)}(\mathbf{X})\right)^2\right]. \tag{3}$$

Details of the derivation are provided in Appendix A.

Estimates of $K - 1$ HTEs against reference Treatment j, $\hat{\tau}_j^{(k)}(\cdot)$ can therefore be obtained by minimizing the empirical version of the loss function (3) via a similar two-step optimization process

$$\hat{\boldsymbol{\tau}}_j(\cdot) = \arg\min_{\boldsymbol{\tau}_j} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}(\mathbf{X}_i) - \sum_{k \neq j} (1[T_i = k]\hat{\pi}^{(k)}(\mathbf{X}_i))\tau_j^{(k)}(\mathbf{X}_i)\right)^2$$

Parallel to the two-treatment situation, the optimal treatment for an individual with characteristics $\mathbf{X}$ is recommended as

$$k^* = \begin{cases} j & \text{if } \hat{\tau}_j^{(k)}(\mathbf{X}) < 0 \text{ for } k \neq j, \\ \arg\max_k \{\hat{\tau}_j^{(k)}(\mathbf{X}), k = 1, 2, ..., K\} & \text{else.} \end{cases} \tag{4}$$

As we shall see in the examples in the simulation study, the lack of recommendation consistency presents a practical challenge that could render the R-learning method unreliable for treatment recommendation.

A Reference-Free R-Learner

To alleviate the R-learner's dependency on reference treatment selection, we consider an alternative approach in quantifying HTE. Here, a logical consideration is to redefine the HTE $\tau^{(k)}$ as a contrast between $Y^{(k)}$ and Y, thus freeing us from need of specifying one particular treatment as the reference group,

$$\tau^{(k)}(\mathbf{X}) = E(Y^{(k)} - Y \mid \mathbf{X}) = E(Y \mid T = k, \mathbf{X}) - E(Y \mid \mathbf{X}) \tag{5}$$

for $k = 1, ..., K$.

In other words, HTEs are assessed by using the difference between the expected outcome under each specific drug and the expected outcome under any drug given the personal characteristics. Under this definition, pairwise comparisons among different drugs can still be made by using

$$\tau_j^{(k)}(\mathbf{X}) = E(Y^{(k)} - Y^{(j)} \mid \mathbf{X}) \equiv \tau^{(k)}(\mathbf{X}) - \tau^{(j)}(\mathbf{X}).$$

Then, following Robinson's idea of decomposition (15), we write

$$\begin{aligned} E[Y \mid T, \mathbf{X}] &= E\left[\sum_{k=1}^{K} 1[T = k] Y^{(k)} \,\middle|\, T, \mathbf{X}\right] \\ &= \sum_{k=1}^{K} 1[T = k] \left(E(Y \mid \mathbf{X}) + \tau^{(k)}(\mathbf{X})\right) \\ &= m(\mathbf{X}) + \sum_{k=1}^{K} 1[T = k] \tau^{(k)}(\mathbf{X}), k = 1, ..., K. \end{aligned} \tag{6}$$

Under such a definition, $\boldsymbol{\tau}(\cdot) = (\tau^{(1)}(\cdot), \tau^{(2)}(\cdot), ..., \tau^{(K)}(\cdot))$ is the minimizer of the mean squares loss function

$$E\left(Y - m(\mathbf{X}) - \sum_{k=1}^{K} 1[T = k] \tau^{(k)}(\mathbf{X})\right), \tag{7}$$

given $m(\mathbf{X})$.

However, since

$$m(\mathbf{X}) = E(Y \mid \mathbf{X}) = \sum_{k=1}^{K} E(Y \mid T = k, \mathbf{X}) P(T = k \mid \mathbf{X})$$
$$= \sum_{k=1}^{K} E(Y^{(k)} \mid \mathbf{X}) \pi^{(k)}(\mathbf{X}) \tag{8}$$

the new HTE $\boldsymbol{\tau}(\cdot)$ must satisfy the constraint

$$\sum_{k=1}^{K} \tau^{(k)}(\mathbf{X}) \pi^{(k)}(\mathbf{X}) \equiv 0 \text{ for any } \mathbf{X}, \tag{9}$$

which renders the numerical optimization an intractable problem.

For implementation, we consider finding $\hat{\boldsymbol{\tau}}(\cdot)$ such that

$$\hat{\boldsymbol{\tau}}(\cdot) = \arg\min_{\tau} \quad \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}(\mathbf{X}_i) - \sum_{k=1}^{K} 1[T_i = k] \tau^{(k)}(\mathbf{X}_i) \right)^2$$
$$\text{s.t.} \quad \sum_{k=1}^{K} \tau^{(k)}(\mathbf{X}_i) \hat{\pi}^{(k)}(\mathbf{X}_i) = 0, \quad i = 1, 2, ..., n.$$

The optimization, subject to the conditions in (9) for the observed data, still poses a significant numerical challenge, as the number of constraints grows linearly with the sample size. To overcome this technical difficulty, we propose an optimization procedure for the reference-free R-learner based on the concept of simplex projection (16), which essentially converts the problem into an unconstrained optimization problem.

A Simplex Implementation of the Reference-Free R-Learner

We consider a K-vertex regular polyhedron defined in a $(K-1)$-dimensional Euclidean space. Following Zhang and Liu's notation (16), we write a K vertices simplex W in $\mathbb{R}^{K-1}$ as

$$W_k = \begin{cases} (K-1)^{-1/2} \zeta_{K-1}, & k = 1 \\ -\frac{1+K^{1/2}}{(K-1)^{3/2}} \zeta_{K-1} + \left(\frac{K}{K-1}\right)^{1/2} e_{K-1}^{(k-1)}, & 2 \le k \le K \end{cases}$$

where $\zeta_{K-1} \in \mathbb{R}^{K-1}$ is a vector of all 1's, and $e_{K-1}^{(k-1)} \in \mathbb{R}^{K-1}$ is a vector with all 0 except for the $(k-1)^{th}$ element being 1. W_k is a unit vector

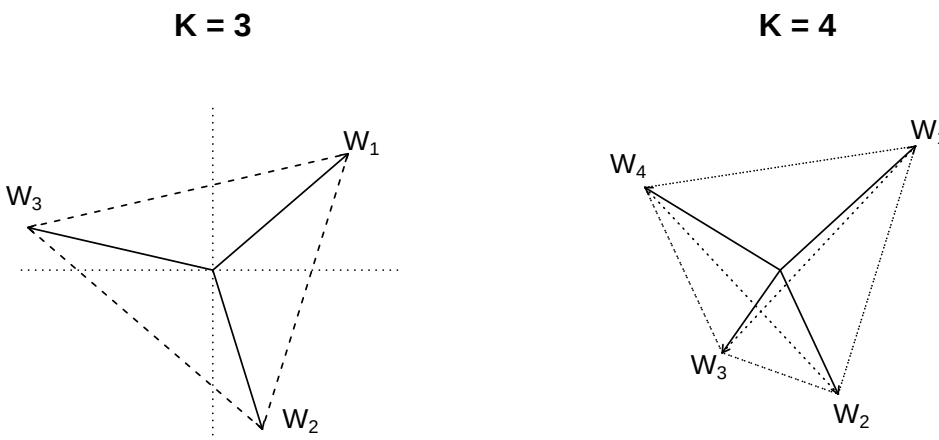

Figure 1. Illustration of simplex W when K = 3,4

pointing to the k^{th} vertex from the origin. Figure 1 illustrates the simplices W for $K = 3, 4$.

Under this definition, $\sum_{k=1}^{K} \langle \boldsymbol{v}, W_k \rangle = 0$ holds true for any arbitrary vector $\boldsymbol{v}$ in $\mathbb{R}^{K-1}$, where $\langle \cdot, \cdot \rangle$ indicates the inner product of two vectors of the same length. It implies that if we could find a mapping $f : \mathbb{R}^p \to \mathbb{R}^{K-1}$ such that

$$\langle f(\mathbf{X}), W_k \rangle = \tau^{(k)}(\mathbf{X})\pi^{(k)}(\mathbf{X}), \text{ for } k = 1, ..., K$$

the constraint (9) will be automatically satisfied. In other words, the constraints are no longer needed if we work on the function $f(\cdot)$ instead of $\boldsymbol{\tau}(\cdot)$. Replacing the $\tau^{(\cdot)}$ in (5) by $f(\cdot)$, we have

$$E(Y \mid T, \mathbf{X}) = m(\mathbf{X}) + \sum_{k=1}^{K} 1[T = k] \langle f(\mathbf{X}), W_k / \pi^{(k)}(\mathbf{X}) \rangle.$$

We estimate $f(\cdot)$ by simply minimizing the empirical mean squares loss function

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}(\mathbf{X}_i) - \sum_{k=1}^{K} 1[T_i = k] \langle f(\mathbf{X}_i), W_k / \hat{\pi}^{(k)}(\mathbf{X}_i) \rangle \right)^2 \quad (10)$$

in a two-step procedure.

To control for the complexity of the estimated function $\hat{f}(\cdot)$, we additionally propose a penalized version of the reference-free R-learner,

following Zhao and colleagues' the approach (18)

$$\hat{f}(\cdot) = \arg\min_{f} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(\mathbf{X}_i) - \sum_{k=1}^{K} 1[T_i = k]\langle f(\mathbf{X}_i), W_k/\hat{\pi}^{(k)}(\mathbf{X}_i)\rangle\right)^2 + \lambda J(f), \tag{11}$$

where $J(f)$ is the penalty for complexity in function $f(\cdot)$, and λ is a tuning parameter.

Finally, for a given patient with covariates $\mathbf{X}$, the estimated reference-free HTE $\tau^{(k)}(\mathbf{X})$ is obtained by

$$\hat{\tau}^{(k)}(\mathbf{X}) = \langle \hat{f}(\mathbf{X}), W_k/\hat{\pi}^{(k)}(\mathbf{X})\rangle, \quad k = 1, 2, ..., K.$$

The optimal treatment to be recommended for this individual is

$$k^* = \arg\max_{k} \left\{\hat{\tau}^{(k)}(\mathbf{X}), k = 1, 2, ...K\right\}.$$

When $\hat{\tau}^{(k)}(\mathbf{X})$ are tied, k^* will be selected from the tied treatments with probabilities proportional to the corresponding estimated propensity scores $\hat{\pi}^{(k)}(\mathbf{X})$, mirroring the common prescribing practice of drugs of equivalent effects.

Computational Algorithm

We hereby present a computational algorithm for the proposed penalized simplex R-learner. For an observed sample with n subjects, $\{(\mathbf{X}_i, Y_i, T_i)_{i=1}^n\}$, we intend to find

$$\hat{f}(\cdot) = \arg\min_{f} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(\mathbf{X}_i) - \frac{\langle f(\mathbf{X}_i), W_{T_i}\rangle}{\hat{\pi}^{(T_i)}(\mathbf{X}_i)}\right]^2 + \lambda J(f). \tag{12}$$

for a given $\hat{m}(\cdot)$ and $\hat{\boldsymbol{\pi}}(\cdot)$.

In this work, we assume that $f(\cdot)$ takes the functional form of $f(\mathbf{X}) = \boldsymbol{z}(\mathbf{X})^T\boldsymbol{\beta}$, where $\boldsymbol{z}(\mathbf{X})$ is an expanded vector of functions of $\mathbf{X}$ with dimension q, which can be much larger than p, and $\boldsymbol{\beta}$ is a $q \times (K-1)$ matrix of model parameters. For simplicity, we use $\boldsymbol{z}$ for $\boldsymbol{z}(\mathbf{X})$. Under this setup, the inner product can be written explicitly as

$$\langle f(\mathbf{X}_i), \frac{W_{T_i}}{\hat{\pi}^{(T_i)}(\mathbf{X}_i)}\rangle = \frac{1}{\hat{\pi}^{(T_i)}(\mathbf{X}_i)}\sum_{u=1}^{q}\sum_{v=1}^{K-1} z_{iu}W_{T_i,v}\beta_{uv} = (\boldsymbol{z}_i^*)^T\boldsymbol{\beta}^*$$

where $W_{T_i,v}$ represents the v^{th} element in W_{T_i},

$$\boldsymbol{\beta}^* = (\beta_{11}, ..., \beta_{1(K-1)}, \beta_{21}, ..., \beta_{2(K-1)}, ..., \beta_{a(K-1)})^T,$$

and

$$\boldsymbol{z}_i^* = \frac{(z_{i1}W_{T_i,1}, ..., z_{i1}W_{T_i,K-1}, z_{i2}W_{T_i,1}..., z_{i2}W_{T_i,K-1}, ..., z_{iq}W_{T_i,K-1})^T}{\hat{\pi}^{(T_i)}(\mathbf{X}_i)}.$$

Then the optimization problem in (12) is converted to a standard least-squares regularization problem for a given $(\hat{m}(\cdot), \hat{\boldsymbol{\pi}}(\cdot))$

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} \left\{ \sum_{i=1}^{n} \left[y_i^* - (\boldsymbol{z}_i^*)^T \boldsymbol{\beta}^* \right]^2 + \lambda J(\boldsymbol{\beta}^*) \right\} \tag{13}$$

where $y_i^* = Y_i - \hat{m}(X_i)$. We propose to estimate $\boldsymbol{\beta}^*$ under a Lasso regularization for model parsimony, which can be implemented with R package **glmnet** developed by Friedman, Hastie and Tibshirani (19).

Proper selection of the tuning parameter λ is essential for the final treatment recommendation. Here, we used a 10-fold cross-validation process to determine the optimal λ. This method of tuning parameter selection is made available by the **glmnet** package. In Appendix B, we provide the sample R code for implementing the simplex R-learner to estimate HTE for given $\{(\mathbf{X}_i, Y_i, T_i)_{i=1}^n\}$.

Simulation Study

We conducted an extensive simulation study to evaluate the operational characteristics of the proposed method. We generated the responses from the following formulation

$$Y_i(\mathbf{X}_i, T_i) = c(\mathbf{X}_i) + \sum_{k=1}^{K} 1[T_i = k] b^{(k)}(\mathbf{X}_i) + \varepsilon_i. \tag{14}$$

where $\varepsilon_i \sim N(0, 1)$.

Herein, we considered various situations involving three treatments ($K = 3$). A total of p independent variables $\mathbf{X} = (X_1, ..., X_p)$ were generated from a standardized multivariate normal distribution, $\mathbf{X} \sim N(0, I_{p\times p})$. For a given vector of covariates $\mathbf{X}^*$, the optimal treatment

T^* was k^* such that

$$E(Y(\mathbf{X}^*, k^*)) = \max\{E(Y(\mathbf{X}^*, 1)), E(Y(\mathbf{X}^*, 2)), E(Y(\mathbf{X}^*, 3))\}. \tag{15}$$

The primary objective of the simulation study was to evaluate the simplex R-learner's ability to make correct recommendations for the optimal treatment based on the observed $\mathbf{X}$. In anticipation of the potential scenarios encountered in real application, we considered two settings for the optimal treatment design, balanced and unbalanced, and three settings for treatment assignment T: (1) balanced assignment (BA), (2) concordant assignment (CA), and (3) discordant assignment (DA). Details of these setting are given below.

The method was tested in six different data settings. For all settings, we let $c(\mathbf{X}) = X_1^2 + X_2 - X_1X_2X_3$. The two optimal treatment designs are

1. Balanced Design:

$$b^{(1)}(\mathbf{X}) = X_4^2, \quad b^{(2)}(\mathbf{X}) = X_5^2, \quad b^{(3)}(\mathbf{X}) = X_6^2,$$

where Treatments 1, 2, 3 had roughly the same chance of being the optimal treatment; and

2. Unbalanced Design:

$$b^{(1)}(\mathbf{X}) = 0.5 + X_4^2, \quad b^{(2)}(\mathbf{X}) = 1[X_5 > 0.5], \quad b^{(3)}(\mathbf{X}) = 0.5X_6,$$

where Treatment 1 had the largest outcome on average, and Treatment 3 had the the least chance to stand out. The majority of patients should be assigned to Treatment 1 instead of Treatments 2 or 3.

We used multinomial logistic regression models to simulate the distribution of treatment assignment T. The propensity scores for the treatments were modeled as follows

$$Pr(T = 1 \mid \mathbf{X}) = \pi^{(1)}(\mathbf{X}) = \frac{p_1(\mathbf{X})}{p_1(\mathbf{X}) + p_2(\mathbf{X}) + p_3(\mathbf{X})};$$
$$Pr(T = 2 \mid \mathbf{X}) = \pi^{(2)}(\mathbf{X}) = \frac{p_2(\mathbf{X})}{p_1(\mathbf{X}) + p_2(\mathbf{X}) + p_3(\mathbf{X})};$$
$$Pr(T = 3 \mid \mathbf{X}) = \pi^{(3)}(\mathbf{X}) = \frac{p_3(\mathbf{X})}{p_1(\mathbf{X}) + p_2(\mathbf{X}) + p_3(\mathbf{X})};$$

where $p_k(\cdot)$ were determined by the following three mechanisms:

1. Balanced Assignment (BA)

$$p_1(\mathbf{X}) = e^{(X_1X_4+X_1+X_4)/5}$$
$$p_2(\mathbf{X}) = e^{(X_2X_5+X_2+X_5)/5}$$
$$p_3(\mathbf{X}) = e^{(X_3X_6+X_3+X_6)/5}$$

2. Concordant Assignment (CA)

$$p_1(\mathbf{X}) = e^{(X_1X_4+X_1+3X_4+5)/5}$$
$$p_2(\mathbf{X}) = e^{(X_2X_5+X_2+1[X_5>0.5])/5}$$
$$p_3(\mathbf{X}) = e^{(X_3X_6+X_3-10X_6^2)/5}$$

3. Discordant Assignment (DA)

$$p_1(\mathbf{X}) = e^{(X_1X_4+X_1-10X_4^2)/5}$$
$$p_2(\mathbf{X}) = e^{(X_2X_5+X_2+1[X_5>0.5])/5}$$
$$p_3(\mathbf{X}) = e^{(X_3X_6+X_3+3X_6+5)/5}$$

In this simulation, BA represented the situation where treatments were assigned to the study sample with roughly equal probabilities. In CA, the treatment assignment followed the same pattern as in the case of unbalanced design, i.e., Treatment 1 had a dominant presence in the study sample. Finally, in DA, the treatment assignment reversed the pattern in the unbalanced design for the optimal treatment, i.e., assigning Treatment 3 to most individuals in the study sample.

With the generated data, we used the Reference-Free R-learner for identification of the optimal treatment. Recommendation performance of the Reference-Free R-learner was compared to that of the R-learner with a reference treatment. For the R-learner with a reference treatment, we also assessed the consistency of the recommendation when different reference treatments were used. We used additive B-splines models with 7 normalized cubic B-spline basis function (20) to estimate $m(\mathbf{X})$, $\boldsymbol{\pi}(\mathbf{X})$, and $f(\mathbf{X})$.

In the first step, we set

$$m(\mathbf{X}) = \sum_{u=1}^{p}\sum_{v=1}^{7} \alpha_{uv}^{(1)} B_{uv}(X_u),$$

$$\log\frac{\pi^{(1)}(\mathbf{X})}{\pi^{(3)}(\mathbf{X})} = \sum_{u=1}^{p}\sum_{v=1}^{7} \alpha_{uv}^{(2)} B_{uv}(X_u), \quad \log\frac{\pi^{(2)}(\mathbf{X})}{\pi^{(3)}(\mathbf{X})} = \sum_{u=1}^{p}\sum_{v=1}^{7} \alpha_{uv}^{(3)} B_{uv}(X_u)$$

and $\quad \pi^{(1)}(\mathbf{X}) + \pi^{(2)}(\mathbf{X}) + \pi^{(3)}(\mathbf{X}) = 1,$

where the interior nodes of $\{B_{uv}(X_u)\}_{v=1}^{7}$ were chosen to be the sample quantiles of the data corresponding to covariate X_u. We used a Lasso regularization (21) in the **glmnet** package with the identity and multinomial logit link functions, respectively, to train the parsimonious estimates, $\hat{m}(\mathbf{X})$ and $\hat{\boldsymbol{\pi}}(\mathbf{X})$.

In the second step, we set $f(\mathbf{X}) = \boldsymbol{z}(\mathbf{X})^T\boldsymbol{\beta}$, where

$$\boldsymbol{z}(\mathbf{X}) = (B_{11}(X_1), ..., B_{17}(X_1), \cdots, B_{p1}(X_p), ..., B_{p7}(X_p))^T$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{11}^{(1)} & \cdots & \beta_{17}^{(1)} & \beta_{21}^{(1)} & \cdots & \beta_{p7}^{(1)} \\ \beta_{11}^{(2)} & \cdots & \beta_{17}^{(2)} & \beta_{21}^{(2)} & \cdots & \beta_{p7}^{(2)} \end{bmatrix}^T$$

which gives rise to the model parameter dimension of $q = 7p$ in the computational algorithm; see Section 3. The **glmnet** package with the identity link function for Lasso regularization was applied to the converted least-squares problem (13) to obtain parsimonious estimates, $\hat{f}(\mathbf{X})$.

To alleviate the bias in the model parameter estimates caused by estimating nuisance parameters in the first step using the same observations, we adopted the idea of cross-fitting in all settings (22). In particular, we used 2-fold data split cross-fitting (23), which divided the original data set in two halves. We used one half of the sample to estimate nuisance parameters m and $\boldsymbol{\pi}$, and then we plugged in the resultant estimates into (13) to estimate $\boldsymbol{\beta}^*$ and to obtain $\hat{\tau}$ using the other half of the data. The same two half samples were reversed in their roles for estimating the nuisance parameters and τ. Finally, the two estimators of $\hat{\tau}$ were averaged as the final estimator of HTE.

We generated training samples of sizes n = 2000 or 4000. Performance was evaluated in independently generated testing samples of size 10000.

Table 1. Mean ± standard devision for the allocation of optimal treatment and treatment assignment in test data with 10000 observations based on 100 repetitions

	Optimal Treatment Allocation		**Treatment Assignment**		
	Balanced	Unbalanced	BA	CA	DA
Treatment 1	3333 ± 53	8012 ± 41	3336 ± 48	6305 ± 50	1026 ± 29
Treatment 2	3330 ± 48	1565 ± 35	3331 ± 49	2669 ± 41	2668 ± 46
Treatment 3	3338 ± 46	423 ± 17	3333 ± 44	1025 ± 31	6306 ± 50

The number of variables p was set to either 9 or 18, but only the first 9 covariates were used for generation of Y and T; the rest were noise. We also examined how the number of noise variables in the models had affected the performance of treatment recommendation by altering p. For each parameter setting, we repeated the simulation 100 times. The summary of optimal treatment distribution as well as recommended assignment across the three treatments are listed in Table 1.

We report in Table 2 the recommendation accuracy of the optimal treatments for each method, which was calculated as the proportion of the recommended treatment by an R-learner that matches the one determined by the maximum expected benefit given by (15) in the test sample. The simulation highlighted the operational characteristics of the R-learner with a reference treatment – Its recommendation accuracy varied by choices of the reference treatment. The Reference-Free R-learner, while maintaining the recommendation accuracy in all settings compared to the R-learner with a reference treatment, had no concern of inconsistent recommendations in CA and DA.

To further assess the magnitudes of recommendation inconsistency by the R-learner with a reference treatment, we compared the results when different reference treatments were chosen. We calculated the percentages of disagreement in the order of estimated treatment effects when different treatments were chosen as the reference. Table 3 shows that in the balanced design, around 20%-40% of individuals were recommended different treatments by the R-learner when different reference treatments were used. Levels of disagreement were reduced to under 20% in the unbalanced design. This is anticipated as the optimal treatment is predominately allocated to one treatment in the study sample, and thus making the recommendation easier. Magnitudes of disagreement increased with the number of treatments offered in the study, suggesting

Table 2. Comparison of recommendation accuracy for the optimal treatment among the R-learners.

	Balanced			Unbalanced		
	BA	CA	DA	BA	CA	DA
N=2000, p = 9						
S	83 ± 2.2%	67.1 ± 5.7%	68.5 ± 6.7%	83.4 ± 2.9%	80 ± 2%	79.5 ± 4.3%
R1	84.3 ± 2.1%	67 ± 5.7%	61 ± 2.3%	83.6 ± 2.8%	81.5 ± 3.2%	79.2 ± 2.8%
R2	84.4 ± 2%	66.6 ± 6.7%	66.7 ± 6.2%	81.8 ± 2.4%	80.6 ± 2.4%	78.7 ± 3.1%
R3	84.2 ± 2%	61 ± 2.5%	66.7 ± 5.8%	83.9 ± 2.7%	82.5 ± 2.7%	79.5 ± 2.8%
N=2000, p = 18						
S	81.8 ± 2%	64.2 ± 4.9%	66.4 ± 5.8%	82.5 ± 2.5%	80.2 ± 2%	79.6 ± 2.8%
R1	83 ± 1.9%	63.9 ± 4.4%	60.2 ± 1.9%	82.4 ± 2.3%	81.4 ± 2.5%	79.5 ± 1.4%
R2	82.9 ± 2%	62.9 ± 4.9%	63.4 ± 4.9%	81 ± 1.7%	80.3 ± 2%	78.6 ± 2.6%
R3	82.8 ± 1.8%	60.1 ± 1.9%	64.2 ± 4.6%	82.8 ± 2.3%	82.2 ± 2.5%	79.1 ± 2.9%
N=4000, p = 9						
S	86.1±1.4%	75.1 ± 5.3%	76.1 ± 4%	85.2 ± 2.5%	81.2 ± 2.7%	83 ± 2.6%
R1	87.3 ± 1.5%	74.1 ± 6.6%	62.9 ± 2.3%	85.7 ± 2.7%	84.8 ± 2.9%	80.2 ± 1%
R2	87.6 ± 1.7%	73.8 ± 6.6%	74.4 ± 6.5%	83.7 ± 2.9%	82.7 ± 3.1%	80.7 ± 2.8%
R3	87.5 ± 1.6%	63.1 ± 3.1%	74.6 ± 6.3%	86.4 ± 2.4%	85.2 ± 2.9%	83.4 ± 2.8%
N=4000, p = 18						
S	85 ± 1.5%	72.1 ± 4.9%	73.2 ± 3%	84.4 ± 2.4%	81.4 ± 2.2%	82.4 ± 2.7%
R1	86.4 ± 1.6%	69.7 ± 5.8%	61.9 ± 1.6%	84.6 ± 2.5%	84 ± 2.8%	80.1 ± 0.8%
R2	86.6 ± 1.7%	69 ± 5.7%	69.4 ± 5.6%	82.4 ± 2.2%	81.8 ± 2.6%	79.9 ± 2.6%
R3	86.6 ± 1.4%	61.5 ± 1.6%	70.2 ± 5.7%	85.5 ± 2.4%	84.4 ± 2.9%	82.2 ± 2.7%

Note: S: The Reference-Free R-learner; R1: The R-learner with treatment 1 as the reference; R2: The R-learner with treatment 2 as the reference; R3: The R-learner with treatment 3 as the reference.

that recommendations are less reliable when the R-learner with a reference treatment is used, in contrast to the Reference-Free R-learner.

As shown in Tables 2 and 3, recommendation accuracy increased with the size of the training sample, and the accuracy is not strongly affected by the number of noise variables included in the models.

We also compared the predicted outcomes under the recommended treatment with the actual observed outcomes. The Boxplots in Figure 2 show the distributions of the average of predicted outcomes under the recommended treatment with 100 replications when n = 2000 and p = 9. A larger average value of the outcome indicated a greater overall benefit of the recommended treatment. Here, the predicted average outcome under the original treatment assignment represents the scenario of no optimization. On the other extreme is the predicted average outcome under the ideal (truly optimal) treatment assignment. Together, they represent two essential benchmarks for evaluating the magnitudes of improvement associated with various R-learning methods. Both the Reference-Free

Table 3. Probability of disagreement of ranking treatment effects in the R-learner with a reference treatment.

	Balanced			Unbalanced		
	BA	CA	DA	BA	CA	DA
N=2000, p = 12						
R1 vs. R2	0.306	0.406	0.374	0.187	0.211	0.173
R1 vs. R3	0.297	0.362	0.341	0.187	0.207	0.163
R2 vs. R3	0.303	0.376	0.382	0.175	0.167	0.144
Average	0.302	0.381	0.365	0.183	0.195	0.160
N=2000, p = 24						
R1 vs. R2	0.364	0.419	0.352	0.206	0.202	0.161
R1 vs. R3	0.351	0.327	0.329	0.210	0.187	0.143
R2 vs. R3	0.360	0.347	0.407	0.172	0.146	0.137
Average	0.358	0.364	0.363	0.196	0.179	0.147
N=4000, p = 12						
R1 vs. R2	0.250	0.338	0.407	0.146	0.199	0.184
R1 vs. R3	0.256	0.399	0.400	0.135	0.206	0.153
R2 vs. R3	0.254	0.412	0.335	0.157	0.195	0.162
Average	0.253	0.383	0.380	0.146	0.200	0.167
N=4000, p = 24						
R1 vs. R2	0.298	0.365	0.374	0.159	0.205	0.192
R1 vs. R3	0.304	0.365	0.368	0.156	0.198	0.155
R2 vs. R3	0.306	0.375	0.366	0.168	0.170	0.172
Average	0.303	0.369	0.370	0.161	0.191	0.173

and the R-learners with a reference treatment significantly improved the treatment benefit in comparison to the observed data where the treatment assignment was not related to the benefit. The figure shows that in most of the tested scenarios, the predicted overall benefits based on the R-learners were close to the ideal case in which each individual is assigned to the recommended treatment defined by (15). This provides another perspective in addition to the accuracy in assessing the performance of R-learners in term of treatment benefits. In all of the simulated cases, the overall benefit was quite similar among all R-learners, but

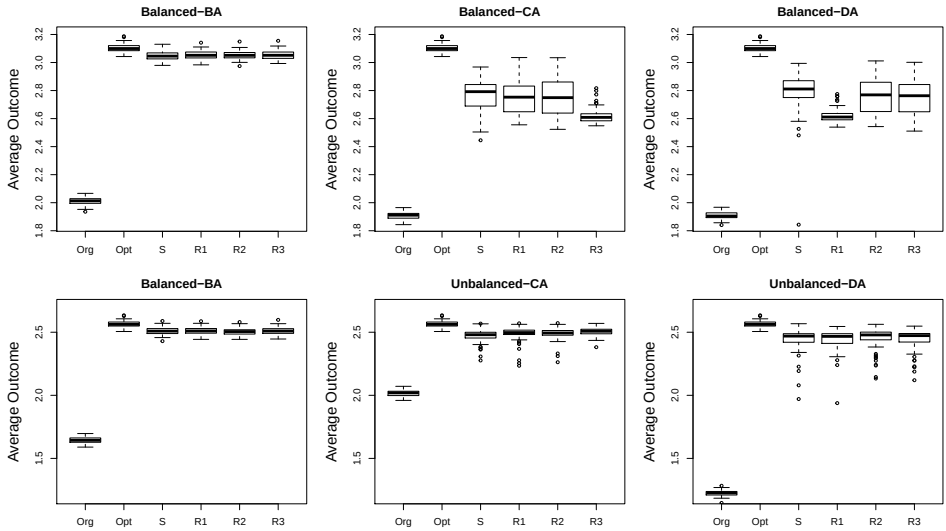

Figure 2. Comparison of the predicted average outcome of all individuals based on the recommended treatment from the R-learners with n = 2000 and p = 12. "Org": Observed outcomes under the original assignment; "Opt": Expected outcomes under the optimal treatment by the true model; "S": Expected outcomes under the recommended treatment by the Reference-Free R-learner; "R1-R3": Expected outcomes under the recommended treatment by the R-learners with treatment 1-3 as the reference; "Balanced": Balanced design for optimal treatment; "Unbalanced": Unbalanced design for optimal treatment.

in some cases (balanced design with CA and DA), the Reference-Free R-learner apparently outperformed the R-learners with a reference treatment. This result was consistent with what has been shown regarding the recommendation accuracy. Appendix C provided additional boxplots for the cases $(n, p) \in \{(2000, 18), (4000, 9), (4000, 18)\}$ which showed similar patterns in Figure 2.

Application to SPRINT Study

To illustrate the proposed method, we analyzed the data from the Systolic Blood Pressure Intervention Trial (SPRINT) a randomized clinical trial aimed at reducing cardiovascular complications in individuals with hypertension, by aggressively lowering systolic blood pressure (SBP). Importantly, the SPRINT study did *not* randomize patients on drug assignments. Instead, the SPRINT intervention only set a lower SBP goal, and it left the therapeutic decisions to the treating physicians on how to bring down SBP. For this reason, the study provided a perfect platform

for us to examine the drugs used in individual patients and treatment outcomes. In this analysis, we used the SPRINT data to evaluate the SBP benefits of different classes of antihypertensive agents, by using the proposed Reference-Free R-learner. The SPRINT data are publicly available by the National Heart, Lung, and Blood Institute, through its Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) (`https://biolincc.nhlbi.nih.gov/home/`) under signed Research Materials Distribution Agreements (RMDA).

The SPRINT Group (24) showed that by setting the SBP goal to 120 mm Hg, they were able to reduce the combined risk of myocardial infarction, acute coronary syndrome, stroke, acute decompensated heart failure, and cardiovascular-related death by approximately 25%. As previously stated, physicians made treatment decisions on who should receive what medications. We applied the Reference-Free R-learner to the SPRINT data to identify optimal treatment for each individual to achieve the maximal SBP reduction. Towards this end, we demonstrate how to accomplish such a goal.

For the purpose of illustration, we considered three classes of antihypertensive agents: Thiazide (or thiazide-type) diuretics, calcium channel blockers (CCB), and angiotensin-converting enzyme (ACE) inhibitors. These are the three main classes of drugs used by the study. Specifically, the study formulary offered chlorthalidone, a thiazide-type diuretic, amlodipine, a calcium channel blocker, and lisinopril, an ACE inhibitor to study participants.

With these three medications, we had $K = 3$ and $T \in \{\text{amlodipine}, \text{chlorthalidone}, \text{lisinopril}\}$. The analysis included a total of 2011 study participants who had been on one of the above medications during the study period, often with other concurrent medications. We considered the first therapeutic episode in which a patient received one of the three medications. Patient response to the medications was measured by the change of SBP before and after the initiation of the therapy. Fourteen patient characteristics were included as covariates in determining the optimal treatment; see Table 4.

We applied the proposed method to the SPRINT data to seek therapeutic recommendations that maximize SBP reduction in individual patients with given characteristics. A summary of the recommendations made by the Reference-Free R-learner are reported in Table 5. The original drug

Table 4. Summary Statistics for Characteristics of the Selected Subjects

Continuous Variables	
Systolic blood pressure: mm Hg	130.9±13.0
Potassium: mmol/L	4.2±0.5
Sodium: mmol/L	139.9±2.4
Estimated GFR: ml/min/1.73 m^2	77.1±20.4
Serum Creatinine: mg/dL	1.0±0.3
Total Cholesterol: mg/dL	192.0±40.6
HDL: mg/dL	52.2±17.2
Framingham Risk Score	17.2±2.4
Age: yr	67.0±9.2
Body-mass index (BMI)	29.2±5.5
Categorical Variables	
Gender: no.(%)	
Female	626(31.1)
Male	1385(68.9)
Smoke Status: no.(%)	
Not Current Smoker	1700(84.5)
Current Smoker	311(15.5)
No. of Antihypertensive Agents: no.(%)	
None	428(21.3)
One	1034(51.4)
Two	449(22.3)
Three	94(4.7)
Four	6(0.3)
Race: no.(%)	
Spanish	167(8.3)
White	1217(60.5)
Black	598(29.7)
Other	29(1.4)

Note: Plus-minus values are means ± standard deviation. Systolic blood pressures, potassium (K), sodium (NA), and estimated GFR (eGFR) were measured at last visit. The other characteristics were collected before trial embarked. GFR denotes glomerular filtration rate. HDL is high-density lipoprotein. Body-mass index is the weight devided by the square of height (kg/m^2).

assignments were relatively balanced among the three drugs. The simplex R-learner, however, suggested that 78.8% (1584 of 2011) of total subjects should be put on chlorthalidone, the thiazide-type diuretic, for improved SBP benefit. Importantly, the algorithm's recommendation is highly consistent with the current clinical guidelines for hypertension treatment. The eighth Joint National Committee guidelines (JNC-8) considered thiazide diuretics as the first line treatment for essential hypertension, and recommended its use as the initial treatment for most patients, either as a monotherapy or in combinations with other antihypertensive agents (25). The JNC-8 panel further concluded that although ACE inhibitors, angiotensin receptor blockers, and calcium channel blockers are acceptable alternatives, thiazide-type diuretics still have the best evidence of efficacy. Compared with other thiazide diuretics, such as hydrochlorothiazide, chlorthalidone is more potent, and has a much long duration of action (26).

As shown in Table 5, the mean SBP reductions based on the recommended assignments were clearly greater than those for the observed treatment assignments. For example, SBP was reduced when patients were put on chlorthalidone. For patients that our algorithm identified for chlorthalidone, the estimated benefit was 1.85 mm Hg. The magnitude of SBP reduction appeared modest, in part because the comparisons were not between an active drug and a placebo, but were among drugs with established efficacy. We note that previous studies showed that ACE inhibitors and calcium channel blocks were able to reduce SBP by 3-5 mm Hg, *when compared against placebo controls* (27; 28; 29). Viewed in this context, the blood pressure benefit of the recommended therapy is certainly not trivial. The trend of the differences between the recommended and actual treatments clearly indicated the added SBP-lowering effects in the recommended therapy.

A closer examination of the distributions of the covariates was also revealing: Patients recommended for chlorthalidone tended to have higher baseline levels of serum potassium and sodium; see Figure 3. Chlorthalidone, like other thiazide diuretics, lowers blood pressure by disposing excessive sodium through urine. Reduced sodium load helps suppress extracellular fluid volume, and thus causing blood pressure to drop. The drug is especially effective for individuals with a higher sodium load (30). But chlorthalidone can also cause excessive urinary disposal

Table 5. Allocation of Amlodipine, Chlorthalidone, and Lisinopril from original assignment and algorithm recommendation

		Assigned	Recommended	Assigned as Recommended
Amlodipine	Size	507	145	18
	Benefit (mm Hg)	0.04(13.61)	1.57(5.74)	3.61(14.69)
Chlorthalidone	Size	608	1584	356
	Benefit (mm Hg)	0.78(14.4)	1.85(8.58)	2.06(15.59)
Lisinopril	Size	896	282	42
	Benefit (mm Hg)	-0.66(14.53)	-0.34(4.95)	1.62(14.72)
Overall	Size	2011	2011	416
	Benefit (mm Hg)	-0.05(14.27)	1.52(8.02)	2.08(15.43)

Note: Benefit is the average of SBP reduction, either observed (Assigned column), or estimated (Recommended column). The standard deviations are given in the parentheses.

of potassium. Because human body is highly sensitive to the level of potassium, which varies within a narrow range, loss of potassium could lead to dangerous situations of hypokalemia. In clinical care, thiazide diuretics are often prescribed together with drugs with potassium-sparing property, such as triamterene (31). For this reason, chlorthalidone is rarely prescribed to patients with low baseline level of potassium, as shown by Figure 3. In summary, the Reference-Free R-learner recommendations were in general supported by the known biological functions of these drugs, and were consistent with current clinical practice.

Concluding Remarks

Treatment optimization and recommendation are essential tasks in the practice of precision medicine. Evidence-based recommendations must reflect the causal effects estimated from real clinical data (32; 12; 13). Yet the problem is not entirely the same as causal effect estimation, where the focus is on direct comparisons of therapeutic effects of different treatments. When multiple treatment options exist, recommendations based on head-to-head pairwise comparisons between each treatment and a preselected reference treatment are not suggested due to the fact that recommendations may vary with a different reference treatment preselected. In this research, we repurposed the R-learning method (14) for treatment recommendation. Specifically, we proposed a reference-free

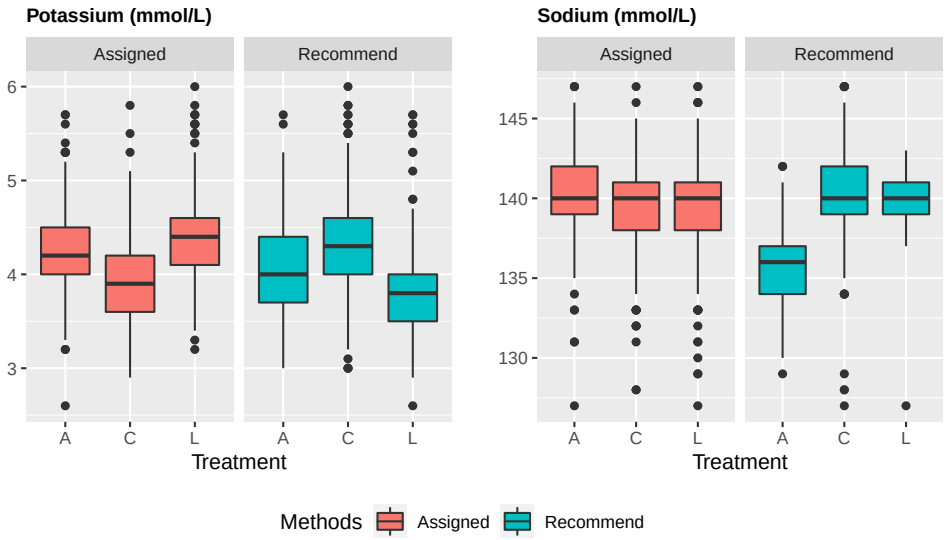

Figure 3. Comparison of distribution of selected variables among three treatments. Treatment "A": Amlodipine; Treatment "C": Chlorthalidone; Treatment "L": Lisinopril. Assigned: Observed assignments; Recommend: Recommendations by Simplex R-learner.

method that uses the overall mean response as the reference group. This modification, however, introduces a large number of constraints to the optimization, and thus increasing the numerical difficulty of optimization. To overcome the challenges in optimization, we introduced a simplex algorithm, which helped us achieve consistent recommendations while sustaining a good numerical performance under multiple treatments. The method broadens the scope of R-learner's application and allows it be used for both causal effect estimation and treatment recommendation. An important feature of the proposed method, one that sets it apart from the R-learner with a reference treatment, is its ability to generate therapeutic recommendations that are not influenced by the selection of the reference treatment group. Further, our use of additive B-splines estimation in the simplex framework has also simplified the computation while relaxing the parametric assumptions on $f(\cdot)$. Extensive simulations demonstrated a satisfactory performance in various simulation settings. Not only did it produce a high level of recommendation accuracy in comparison to the R-learner with a reference treatment, it also retained a robust performance in the presence of noise in the covariates. The latter feature makes it particularly appealing in high dimensional data applications.

While the additive B-splines model worked well with the LASSO regularization in this research, the nonparametric tree-based method such as XGboost (33) could be a viable alternative to account for the non-linearity and model sparsity. In order to apply the tree-based methods, however, the loss function calculation needs to be updated at each node to account for the simplex structure, which inevitably increases the computation complexity.

It should be recognized that seeking individual treatment rule (ITR) by subgroup identification (34; 35; 36; 37; 38; 39; 40) is also a viable approach for treatment optimization. It can be shown that the optimal HTE problem is equivalent to the ITR problem; see Appendix D. In comparison to the methods requiring discontinuous loss functions, the proposed method is generally easier to implement computationally, unless alternative smooth functions can be found to approximate the discontinuous loss functions. In addition, the proposed method produces model-free estimates for the casual effects associated with all the treatments as byproducts.

The method presented in this paper is for continuous outcomes. Extending the Robinson decomposition to other types of casual effects, such as odds ratios for binary outcomes or the hazard ratios for survival outcomes, remains to be investigated. This limitation notwithstanding, we contend that the Reference-Free R-learner can be used for treatment optimization in a broad class of practical applications.

Appendix

A. Derivation of the R-learner with multiple treatments

When there are $K > 2$ treatments, the observed outcome Y is the one of the potential outcomes from the data and hence can be denoted by $Y = \sum_{k=1}^{K} 1[T = k]Y^{(k)}$. If treatment j is selected as the reference, then HTE for the k^{th} treatment can be defined as $\tau_j^{(k)}(\mathbf{X}) = E[Y^{(k)} - Y^{(j)} \mid \mathbf{X}], k \neq j$, and the probability of receiving treatment k is $\pi^{(k)}(\mathbf{X}) = \Pr(T = k \mid \mathbf{X})$. The conditional expectation of observed outcome Y can be written as

$$
\begin{aligned}
E[Y \mid T, \mathbf{X}] &= E\left[\sum_{k=1}^{K} 1[T=k]Y^{(k)} \;\middle|\; T, \mathbf{X}\right] = \sum_{k=1}^{K} 1[T=k]E\left[Y^{(k)} \mid T, \mathbf{X}\right] \\
&= \sum_{k\neq j} 1[T=k]\left(E[Y^{(j)} \mid \mathbf{X}] + \tau_j^{(k)}(\mathbf{X})\right) + 1[T=j]E[Y^{(j)} \mid \mathbf{X}] \\
&= E[Y^{(j)} \mid \mathbf{X}] + \sum_{k\neq j} 1[T=k]\tau_j^{(k)}(\mathbf{X})
\end{aligned}
$$

and we also note that overall conditional mean outcome

$$
\begin{aligned}
m(\mathbf{X}) = E[Y \mid \mathbf{X}] &= \sum_{k=1}^{K} E\left[Y \mid T=k, \mathbf{X}\right] \Pr(T=k \mid \mathbf{X}) \\
&= \sum_{k=1}^{K} E\left[Y^{(k)} \mid T, \mathbf{X}\right] \pi^{(k)}(\mathbf{X}) \quad \text{(Ignorability Assumption)} \\
&= \sum_{k\neq j} \left(E[Y^{(j)} \mid \mathbf{X}] + \tau_j^{(k)}(\mathbf{X})\right) \pi^{(k)}(\mathbf{X}) + E[Y^{(j)} \mid \mathbf{X}]\pi^{(j)}(\mathbf{X}) \\
&= E[Y^{(j)} \mid \mathbf{X}] + \sum_{k\neq j} \tau_j^{(k)}(\mathbf{X})\pi^{(k)}(\mathbf{X})
\end{aligned}
$$

which results in

$$
E[Y \mid T, \mathbf{X}] = m(\mathbf{X}) + \sum_{k=j} \left(1[T=k] - \pi^{(k)}(\mathbf{X})\right) \tau_j^{(k)}(\mathbf{X}).
$$

This extends the Robinson decomposition in a multiple treatment scenario and hence for given $m(\mathbf{X})$ and $\boldsymbol{\pi}(\mathbf{X})$ the HTE, $\boldsymbol{\tau}(\cdot)$ is the one that minimizes the mean-squares loss function

$$
\arg\min_{\tau} \left\{ E\left[\left(Y - m(\mathbf{X}) - \sum_{k\neq j} \left(1[T=k] - \pi^{(k)}(\mathbf{X})\right) \tau_j^{(k)}(\mathbf{X}) \right)^2 \right] \right\}.
$$

B. R Code of the Simplex Algorithm for the Reference-Free R-learner

```
require(glmnet)
require(doMC)
registerDoMC(cores = 4)
# 4 is the number of cores for parallel computing
```

```
SimplexR <- function(x, # data projected on functional spaces z(X)
                     w, # observed treatment assignment T
                     y, # observed outcome Y
                     x.test,
                     # test data projected on functional spaces
                     pi.train,
                     # estimated pi in first stage \hat{\pi}
                     pi.test,
                     # predicted pi for test data based on estimated
                     # model in first stage
                     m.train   # estimated m in first stage, \hat{m}
 ) {
  nobs = dim(x)[1]
  pobs = dim(x)[2]

  ################## Simplex Method ####################
  k = length(unique(w))
  z = rep(1, k-1)
  e = diag(x = 1, k-1)
  W1 = (k-1)^(-0.5) * z
  W = cbind(W1,  (k/(k-1))^(0.5)*e - z*(1+sqrt(k))/(k-1)^1.5 )

  # transform data x (z in section 3) into desired x.new (z*) ##
  x.whole = cbind(1, x)
  x.new = sapply(seq(nobs), function(i){
    as.vector(outer(W[,w[i]] ,x.whole[i,]))/pi.train[i,w[i]]
  })
  x.new = t(x.new)
  penalty_f = c(rep(0,k-1), rep(1, pobs*(k-1)))
  fit.tau = cv.glmnet(x.new, y-m.train, family = "gaussian",
     parallel = TRUE, penalty.factor = penalty_f, intercept=FALSE)

  ## estimate tau for testing data ##
  x.test.whole = cbind(1, x.test)
  best.beta = coef(fit.tau, s="lambda.min")
  best.beta = matrix(best.beta[-1], nrow = pobs+1, byrow = T)
  est.tau = x.test.whole %*% best.beta %*% W / pi.test
  return(est.tau)
}
```

C. Additional simulation results

We include additional boxplots mentioned in Section 4 for the simulation scenarios with $(n, p) \in \{(2000, 24), (4000, 12), (4000, 24)\}$. Figure 4, 5, and 6 demonstrate that for these cases, the pattern in Figure 2 still remains, and the Reference-Free R-learner outperformed the R-learner with a reference treatment in the balanced design with CA and DA.

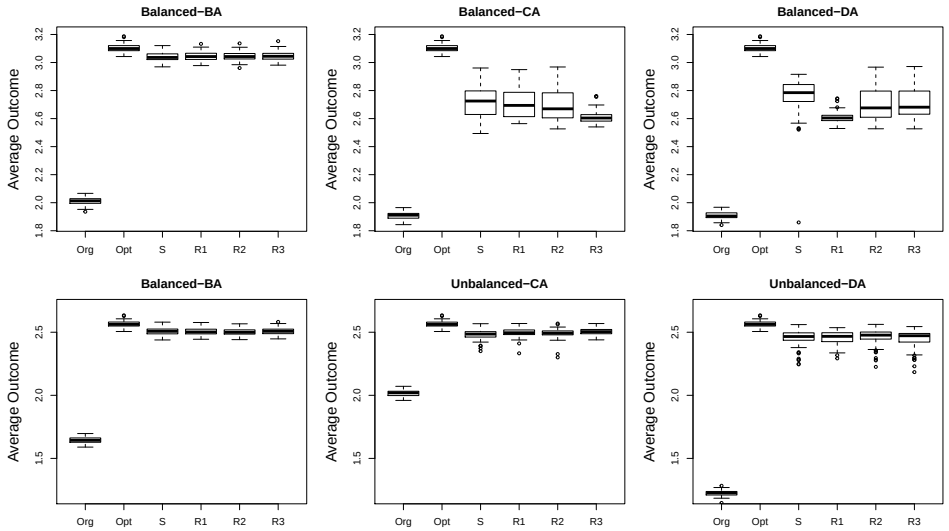

Figure 4. Comparison of the predicted average outcome of all individuals based on the recommended treatment from the R-learners with n = 2000 and p = 24. “Org”: Observed outcomes under the original assignment; “Opt”: Expected outcomes under the optimal treatment by the true model; “S”: Expected outcomes under the recommended treatment by the Simplex R-learner; “R1-R3”: Expected outcomes under the recommended treatment by the standard R-learner with treatment 1-3 as the reference; “Balanced”: Balanced design for optimal treatment; “Unbalanced”: Unbalanced design for optimal treatment.

D. Connection between the Reference-Free R-learner and ITR

Following the definition given by Qian and Murphy,(4) the value function in multiple treatments scenario with $K > 2$ is

$$V_{d(X)} = E\left(\sum_{k=1}^{K} 1[d(\mathbf{X}) = k]Y^{(k)} \mid \mathbf{X}\right)$$

given civariate $\mathbf{X}$, where $d(\cdot) \in \{1, 2, ..., K\}$ is the ITR.

Here we show that ITR problem defined as

$$d^*(\cdot) = \arg\max_{d(\cdot)} V_{d(\mathbf{X})}$$

is equivalent to the optimal HTE problem, that is

$$k^* = \arg\max_{k}\{\tau_*^{(k)}, k = 1, 2, ...K\},$$

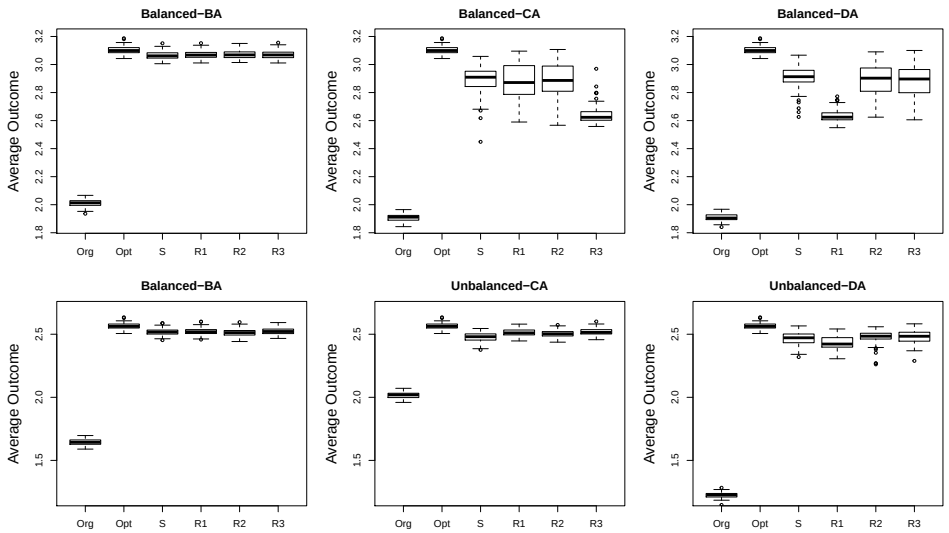

Figure 5. Comparison of the predicted average outcome of all individuals based on the recommended treatment from the R-learners with n = 4000 and p = 12. "Org": Observed outcomes under the original assignment; "Opt": Expected outcomes under the optimal treatment by the true model; "S": Expected outcomes under the recommended treatment by the Reference-Free R-learner; "R1-R3": Expected outcomes under the recommended treatment by the R-learners with treatment 1-3 as the reference; "Balanced": Balanced design for optimal treatment; "Unbalanced": Unbalanced design for optimal treatment.

where $\boldsymbol{\tau}_*(\cdot) = \left(\tau_*^{(1)}(\cdot), \tau_*^{(2)}(\cdot), ..., \tau_*^{(K)}(\cdot)\right)$ satisfies

$$\boldsymbol{\tau}_*(\cdot) = \arg\min_{\boldsymbol{\tau}} E\left(Y - m(\mathbf{X}) - \sum_{k=1}^{K} 1[T=k]\tau^{(k)}(\mathbf{X})\right)^2$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \tau_*^{(k)}(\mathbf{X})\pi^{(k)}(\mathbf{X}) = 0 \quad \text{for any } \mathbf{X}$$

Note that the mean-squares loss function can be further written as

$$\int E\left\{\sum_{k=1}^{K}\left(Y^{(k)} - m(\mathbf{X}) - \tau^{(k)}(\mathbf{X})\right)^2 \pi^{(k)}(\mathbf{X}) \mid \mathbf{X}\right\} dF(\mathbf{X}),$$

where F is the distribution function of $\mathbf{X}$. We define a functional $\mathcal{L}$ of $\boldsymbol{\tau}(\cdot)$ and a Lagrange multiplier λ as

$$\mathcal{L}(\boldsymbol{\tau}, \lambda) = \int \bar{\mathcal{L}}(\boldsymbol{\tau}(\mathbf{x}), \lambda) dF(\mathbf{x}),$$

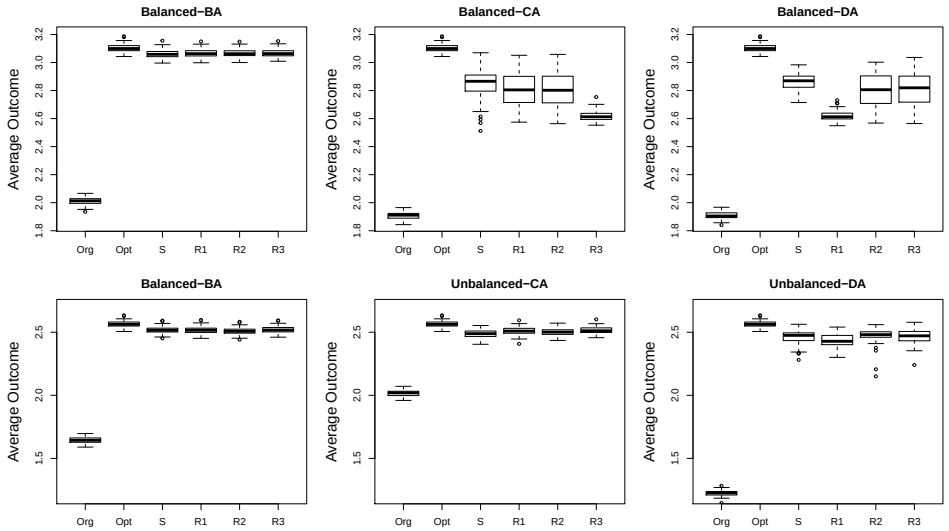

Figure 6. Comparison of the predicted average outcome of all individuals based on the recommended treatment from the R-learners with n = 4000 and p = 24. "Org": Observed outcomes under the original assignment; "Opt": Expected outcomes under the optimal treatment by the true model; "S": Expected outcomes under the recommended treatment by the Reference-Free R-learner; "R1-R3": Expected outcomes under the recommended treatment by the R-learners with treatment 1-3 as the reference; "Balanced": Balanced design for optimal treatment; "Unbalanced": Unbalanced design for optimal treatment.

where

$$\bar{\mathcal{L}}(\boldsymbol{\tau}(x),\lambda) = E\Big\{ \sum_{k=1}^{K} \Big(Y^{(k)} - m(\mathbf{X}) - \tau^{(k)}(\mathbf{X})\Big)^2 \pi^{(k)}(\mathbf{X}) - \lambda \sum_{k=1}^{K} \tau^{(k)}(\mathbf{X})\pi^{(k)}(\mathbf{X}) \mid \mathbf{X}=\mathbf{x}\Big\}.$$

We calculate the partial functional derivative with respect to the k^{th} component

$$\begin{aligned}
&\frac{\delta \mathcal{L}(\boldsymbol{\tau},\lambda)}{\delta \tau^{(k)}} \\
&= \int \frac{\bar{\mathcal{L}}(\tau^{(1)},...,\tau^{(k)}+\varepsilon\phi^{(k)},...\tau^{(K)})}{d\varepsilon}\Big|_{\varepsilon=0} dF(\mathbf{x}) \\
&= \int \Big(-E\Big[2\{Y^{(k)} - m(\mathbf{X}) - \tau^{(k)}(\mathbf{X})\} \mid \mathbf{X}=\mathbf{x}\Big] - \lambda\Big)\, \pi^{(k)}(\mathbf{x})\phi^{(k)}(\mathbf{x}) dF(\mathbf{x}).
\end{aligned}$$

To satisfy the first order condition for optimization, it follows that

$$E\Big[Y^{(k)} - m(\mathbf{X}) - \tau^{(k)}(\mathbf{X}) \mid \mathbf{X}=\mathbf{x}\Big] = -\lambda/2, \quad \text{for } k = 1,2,...,K.$$

Thus, for arbitrary treatment i and j,

$$E\left[Y^{(i)} - m(\mathbf{X}) - \boldsymbol{\tau}_*^{(i)}(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\right] = E\left[Y^{(j)} - m(\mathbf{X}) - \boldsymbol{\tau}_*^{(j)}(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\right].$$

If $\tau_*^{(i)}(\mathbf{x}) > \tau_*^{(j)}(\mathbf{x})$ for a given $\mathbf{x}$, then any value $v_{ij} \in [\tau_*^{(j)}(\mathbf{x}), \tau_*^{(i)}(\mathbf{x})]$ satisfies

$$E\left[Y^{(i)} - m(\mathbf{X}) - v_{ij} \mid \mathbf{X} = \mathbf{x}\right] > E\left[Y^{(j)} - m(\mathbf{X}) - v_{ij} \mid \mathbf{X} = \mathbf{x}\right].$$

Hence $E\left[Y^{(i)} \mid \mathbf{X} = \mathbf{x}\right] > E\left[Y^{(j)} \mid \mathbf{X} = \mathbf{x}\right]$. It is also noted that for any $i \neq j$, $E[Y^{(i)} \mid \mathbf{X} = \mathbf{x}] > E[Y^{(j)} \mid \mathbf{X} = \mathbf{x}]$ immediately implies $\tau_*^{(i)}(\mathbf{x}) > \tau_*^{(j)}(\mathbf{x})$. This completes the proof.

Acknowledgments

The work was partially supported by NIH grants HL095086, HL228494, AA025208, AA026969, GM115458. The manuscript was prepared using SPRINT research data obtained from the NHLBI.

References

[1] Neyman JS. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences* 1923; 10: 1–51.

[2] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; 66(5): 688.

[3] Moodie EE, Platt RW and Kramer MS. Estimating response-maximized decision rules with applications to breastfeeding. *Journal of the American Statistical Association* 2009; 104(485): 155–165.

[4] Qian M and Murphy SA. Performance guarantees for individualized treatment rules. *Annals of Statistics* 2011; 39(2): 1180.

[5] Foster JC. *Subgroup Identification and Variable Selection from Randomized Clinical Trial Data.* PhD Thesis, The University of Michigan, 2013.

[6] Athey S and Imbens GW. Machine learning methods for estimating heterogeneous causal effects. *stat* 2015; 1050(5): 1–26.

[7] Watkins CJ and Dayan P. Q-learning. *Machine learning* 1992; 8(3): 279–292.

[8] Chakraborty B and Moodie E. Statistical methods for dynamic treatment regimes. *Springer-Verlag doi* 2013; 10: 978–1.

[9] Künzel SR, Sekhon JS, Bickel PJ et al. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 2019; 116(10): 4156–4165.

[10] Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; 65(2): 331–355.

[11] Robins JM. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*. Springer, pp. 189–326.

[12] Schulte PJ, Tsiatis AA, Laber EB et al. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics* 2014; 29(4): 640.

[13] Shi C, Fan A, Song R et al. High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics* 2018; 46(3): 925.

[14] Nie X and Wager S. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika* 2020; DOI:10.1093/biomet/asaa076. URL https://doi.org/10.1093/biomet/asaa076. Asaa076, https://academic.oup.com/biomet/advance-article-pdf/doi/10.1093/biomet/asaa076/33788449/asaa076.pdf.

[15] Robinson PM. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 1988; : 931–954.

[16] Zhang C and Liu Y. Multicategory angle-based large-margin classification. *Biometrika* 2014; 101(3): 625–640.

[17] Imbens GW and Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[18] Zhao Q, Small DS and Estefaie A. Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society: Series B* 2022; 84. DOI:10.1111/rssb.12483.

[19] Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2010; 33(1): 1.

[20] Schumaker L. *Spline functions: basic theory*. Cambridge University Press, 2007.

[21] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 1996; 58(1): 267–288.

[22] Chernozhukov V, Chetverikov D, Demirer M et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 2018; 21(1): C1–C68. DOI:10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097. https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf.

[23] Knaus MC, Lechner M and Strittmatter A. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal* 2021; 24(1): 134–161.

[24] Group SR. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine* 2015; 373(22): 2103–2116.

[25] James PA, Oparil S, Carter BL et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the eighth joint national committee (jnc 8). *JAMA* 2014; 311(5): 507–520.

[26] Carter BL, Ernst ME and Cohen JD. Hydrochlorothiazide versus chlorthalidone: evidence supporting their interchangeability. *Hypertension* 2004; 43(1): 4–9.

[27] Yusuf S, Sleight P, Pogue Jf et al. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *The New England Journal of Medicine* 2000; 342(3): 145.

[28] MacMahon S, Sharpe N, Gamble G et al. Randomized, placebo-controlled trial of the angiotensin-converting enzyme inhibitor, ramipril, in patients with coronary or other occlusive arterial disease. *Journal of the American College of Cardiology* 2000; 36(2): 438–443.

[29] Pitt B, Byington RP, Furberg CD et al. Effect of amlodipine on the progression of atherosclerosis and the occurrence of clinical events. *Circulation* 2000; 102(13): 1503–1510.

[30] Moser M and Feig PU. Fifty years of thiazide diuretic therapy for hypertension. *Archives of Internal Medicine* 2009; 169(20): 1851–1856.

[31] Tu W, Decker BS, He Z et al. Triamterene enhances the blood pressure lowering effect of hydrochlorothiazide in patients with hypertension. *Journal of General Internal Medicine* 2016; 31(1): 30–36.

[32] Murphy SA, van der Laan MJ, Robins JM et al. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* 2001; 96(456): 1410–1423.

[33] Chen T and Guestrin C. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016; DOI:10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

[34] Zhao Y, Zeng D, Rush AJ et al. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 2012; 107(499): 1106–1118.

[35] Zhang B, Tsiatis AA, Davidian M et al. Estimating optimal treatment regimes from a classification perspective. *Stat* 2012; 1(1): 103–114.

[36] Xu Y, Yu M, Zhao YQ et al. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* 2015; 71(3): 645–653.

[37] Zhang B and Zhang M. C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* 2018; 74(3): 891–899.

[38] Zhu R, Zhao YQ, Chen G et al. Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* 2017; 73(2): 391–400.

[39] Zhang C, Chenm J, Fu H et al. Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. *Statistica Sinica* 2018; .

[40] Qi Z, Liu D, Fu H et al. Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *Journal of the American Statistical Association* 2020; 115(530): 678–691.