HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

J Am Stat Assoc. 2021 ; 116(533): 283–294. doi:10.1080/01621459.2020.1725522.

Improved doubly robust estimation in learning optimal individualized treatment rules

Yinghao Pan, **Ying-Qi Zhao***

Department of Mathematics and Statistics, University of North Carolina at Charlotte

Public Health Sciences Division, Fred Hutchinson Cancer Research Center

Abstract

Individualized treatment rules (ITRs) recommend treatment according to patient characteristics. There is a growing interest in developing novel and efficient statistical methods in constructing ITRs. We propose an improved doubly robust estimator of the optimal ITRs. The proposed estimator is based on a direct optimization of an augmented inverse-probability weighted estimator (AIPWE) of the expected clinical outcome over a class of ITRs. The method enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent when either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among the class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. Simulation studies show that the estimated ITRs obtained from our method yield better results than those obtained from current popular methods. Data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study is analyzed as an illustrative example.

Keywords

Double robustness; Individualized treatment rule; Personalized medicine; Propensity score

1 Introduction

In recent years, personalized medicine, or precision medicine, has received tremendous attention in clinical practice and medical research (Hamburg and Collins, 2010; Chan and Ginsburg, 2011; Collins and Varmus, 2015). Its development originates from the fact that patients often exhibit heterogenous responses to treatments. A drug that works for the majority of individuals may not work for a subgroup of patients with certain characteristics. For example, trastuzumab is shown to be effective for treating HER2-overexpressing metastatic breast cancer as it is specifically designed to target HER2 amplification (Vogel et al., 2002). Individualized treatment rules (ITRs) formalize personalized treatment decisions, which recommend treatments using patients' own information, with the optimal ITR

*Contact: Ying-Qi Zhao, yqzhao@fredhutch.org, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109.

SUPPLEMENTARY MATERIAL

The online supplementary materials contain the appendices for the article.

maximizing the mean of a pre-specified clinical outcome if followed by the patient population.

Using data collected from clinical trials or observational studies, numerous methods have been developed on estimation of optimal ITRs. One approach is to fully or partly specify a model of the clinical outcome given treatment and covariates, and then use the fitted model to infer the optimal ITR. This includes Q-learning (Qian and Murphy, 2011) and A-learning (Murphy, 2003; Robins, 2004; Blatt et al., 2004). Q-learning models the conditional mean of the outcome given treatment and covariates while A-learning directly models the differential treatment effects between treatments. However, one drawback of Q- and A-learning is that the optimal treatment rule is indirectly estimated through posited regression models, and thus sensitive to model misspecification. Value-search or direct-maximization methods offer an alternative to regression-based methods by directly maximizing an estimator of the marginal mean outcome over a pre-specified class of ITRs (Zhao et al., 2012; Zhang et al., 2012; Zhou et al., 2017; Zhao et al., 2019), thereby separating the class of decision rules from the posited regression models.

In particular, Zhang et al. (2012) estimated the optimal ITR by maximizing an augmented inverse-probability weighted estimator (AIPWE) for the population mean outcome over a class of ITRs. The aforementioned estimator is doubly robust (DR) in the sense that it consistently estimates the optimal ITR if either the propensity score or the outcome regression model is correctly specified. Doubly robust estimation has enjoyed great popularity in missing data and causal inference models (Scharfstein et al., 1999; Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005). The DR estimators require specification of two nuisance working models, one for the missingness or treatment assignment mechanism, and another one for the distribution of complete data or potential outcomes. Historically, estimation of the nuisance parameters indexing the working models in DR estimators had received little attention, partly because the asymptotic properties of the DR estimators do not depend on the choice of nuisance parameter estimates when both working models are correctly specified (Tsiatis, 2007). As a result, standard maximum likelihood estimators are used, i.e., logistic regression for the propensity score model, and linear regression for the outcome model. This standard practice starts to change after Kang and Schafer (2007) cautioned against the use of DR estimators when both working models are misspecified. Several discussion articles (Robins et al., 2007; Tsiatis and Davidian, 2007; Tan, 2007) further pointed out that the choice of nuisance parameter estimates can have a dramatic impact on the properties of the DR estimators when at least one working model is misspecified. Indeed, in the context of estimating optimal ITRs using DR methods, there is still room for improved performance. For example, as illustrated in simulation studies of Zhao et al. (2019), the usual DR estimator (Zhang et al., 2012) can be inefficient, i.e, exhibits a large variation when the outcome regression model is misspecified. The poor performance may be partly a consequence of the default use of maximum likelihood estimators for the coefficients in the misspecified outcome regression model (Cao et al., 2009). This motivates us to develop improved DR approaches for learning optimal ITRs.

Several improved DR estimators have been proposed in missing data and causal inference models for the purpose of variance reduction. The nuisance parameters indexing the outcome model are estimated so as to minimize the variance of the DR estimator under a correctly specified propensity score model (Rubin and van der Laan, 2008; Cao et al., 2009; Tan, 2010; Tsiatis et al., 2011). In this article, we propose to estimate the optimal ITR by maximizing an improved DR estimator of the population mean outcome among a set of ITRs. Our proposed estimator is doubly robust. In addition, it achieves the smallest variance among its class of DR estimators when the propensity score models are correctly specified, regardless of the specification of the outcome models. As we demonstrate, this approach leads to estimated optimal regimes achieving comparable or better performance than those from Zhang et al. (2012).

The heterogeneity in response to treatments exists not only between patients but also within each patient. A patient's response to treatment can change over time because individual characteristics, and the nature of disease itself, evolve. This motivates the development of dynamic treatment regimes (DTRs) (Murphy, 2003), which are sequential decision rules that adapt over time to the clinical status of each patient. At each decision point, the available patient history data are used as input for the decision rule, and an individualized treatment is recommended for the next stage. Construction of optimal DTRs has been of great interest, where several methods are developed to handle multi-stage problems (Zhang et al., 2013; Laber et al., 2014; Schulte et al., 2014; Zhao et al., 2015; Wallace and Moodie, 2015; Liu et al., 2018). In this paper, we also discuss extending the proposed method to estimate optimal DTRs with added efficiency and robustness.

This article proposes 2 major contributions to the literature. (1) We propose improved DR approaches for estimating optimal ITRs, which has not been investigated in the field of personalized medicine. (2) Current literature such as Cao et al. (2009) and Tsiatis et al. (2011) employed inverse-probability weighted estimating equations to estimate the nuisance parameters. Instead, we propose *augmented* inverse-probability weighted estimating equations for this purpose, which brings further stability.

The remainder of the article is organized as follows. In Section 2, we introduce background information and review existing doubly robust estimators in learning optimal ITRs. We then formally describe the proposed improved doubly robust estimator in single-stage optimal treatment problems. Theoretical results are presented in Section 3. In Section 4, we present simulation studies to evaluate finite sample performance of the proposed method. The method is then illustrated using data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study in Section 5. Some concluding remarks are given in Section 6. Technical results are relegated to the supplementary material.

2 Method

2.1 Background and preliminaries

We consider the estimation of the optimal ITR in the single-stage setting. We observe $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$, comprising n independent and identically distributed triplets of $(\boldsymbol{X}, A, Y)$, where $\boldsymbol{X} \in \mathcal{X}$ denotes the patient's baseline variables; $A \in \mathcal{A} = \{-1, 1\}$ denotes the assigned

treatment; Y denotes the clinical outcome of interest, coded so that the larger the better. The data comes from either randomized trials or observational studies. An ITR is a map $d: \mathcal{X} \mapsto \mathcal{A}$ such that a patient presenting with $\boldsymbol{X} = \boldsymbol{x}$ will receive treatment $d(\boldsymbol{x})$.

Let $\mathcal{D}$ denote a class of ITRs of interest. To formally define the optimal ITR, d^{opt}, we adopt the potential outcome framework (Rubin, 1974). Let $Y(a)$ denote the potential outcome under treatment $a \in \{-1, 1\}$. The potential outcome under any ITR, d, can be defined as $Y(d) = Y(1)I\{d(\boldsymbol{X}) = 1\} + Y(-1)I\{d(\boldsymbol{X}) = -1\}$, where $I\{\cdot\}$ is the indicator function. Here we suppress the dependence of $Y(d)$ on $\boldsymbol{X}$. The performance of d is measured by the marginal mean outcome $V(d) \triangleq E\{Y(d)\}$, the so-called value function associated with the rule d. In other words, the value function $V(d)$ represents the overall population mean if treatment were to be assigned according to d. The optimal ITR, d^{opt}, is a rule that maximizes $V(d)$ among $\mathcal{D}$, i.e., $V(d^{\text{opt}}) \geq V(d)$ for all $d \in \mathcal{D}$.

In order to connect the potential outcomes with the observed data, we make the following assumptions: (i) consistency, $Y = Y(1)\, I(A = 1) + Y(-1)I(A = -1)$; (ii) positivity, $P(A = a|\boldsymbol{X}) > 0$ for $a = \pm 1$ and for all $\boldsymbol{X}$; (iii) no unmeasured confounding, $A \perp \{Y(-1), Y(1)\}|\boldsymbol{X}$. These are standard and well-studied assumptions in causal inference (Imbens and Rubin, 2015). Assumption (iii) is trivial in a randomized trial but unverifiable in an observation study (Robins et al., 2000).

Define $Q_0(\boldsymbol{x}, a) \triangleq E(Y|\boldsymbol{X} = \boldsymbol{x}, A = a)$, then under the aforementioned assumptions, it can be shown that

$$V(d) = E_{\boldsymbol{X}}[Q_0(\boldsymbol{X}, 1)I\{d(\boldsymbol{X}) = 1\} + Q_0(\boldsymbol{X}, -1)I\{d(\boldsymbol{X}) = -1\}],$$

where the outer expectation $E_{\boldsymbol{X}}[\cdot]$ is taken with respect to the marginal distribution of $\boldsymbol{X}$. The above formulation implies that $d^{\text{opt}}(\boldsymbol{x}) = \text{argmax}_{a \in \{-1,1\}} Q_0(\boldsymbol{x}, a)$. One approach is to posit a regression model $Q(\boldsymbol{X}, A; \boldsymbol{\beta})$ for $Q_0(\boldsymbol{X}, A)$, and estimate the nuisance parameter $\boldsymbol{\beta}$ by some $\hat{\boldsymbol{\beta}}$; e.g. least squares. Subsequently the optimal ITR is estimated by $\hat{d}(\boldsymbol{x}) = \text{argmax}_{a \in \{-1,1\}} Q(\boldsymbol{x}, a; \hat{\boldsymbol{\beta}})$ (Qian and Murphy, 2011). This is usually referred to as an indirect approach, which could lead to inconsistent estimators of d^{opt} when the posited model $Q(\boldsymbol{X}, A; \boldsymbol{\beta})$ is incorrect.

To alleviate the above issue, value-search or direct-maximization methods attempt to estimate d^{opt} by directly maximizing an estimator of the value function over the class $\mathcal{D}$. The key step is to construct a consistent and robust estimator of the value function, say $\hat{V}(\cdot)$. Then d^{opt} is estimated by $\hat{d} = \text{argmax}_{d \in D} \hat{V}(d)$. Let $\pi_0(a, \boldsymbol{x}) \triangleq P(A = a|\boldsymbol{X} = \boldsymbol{x})$ denote the true propensity score, so the value function can be rewritten as (Qian and Murphy, 2011; Zhao et al., 2012)

$$V(d) = E\left[\frac{Y}{\pi_0(A, \boldsymbol{X})} I\{A = d(\boldsymbol{X})\}\right].$$

In an observation study, $\pi_0(A, \boldsymbol{X})$ is unknown. A parametric model $\pi(A, \boldsymbol{X}; \boldsymbol{\gamma})$ may be posited; for example, a logistic regression model $\pi(1, \boldsymbol{X}; \boldsymbol{\gamma}) = \{1 + \exp(-\boldsymbol{X}^\top \boldsymbol{\gamma})\}^{-1}$, $\boldsymbol{X} = (1, \boldsymbol{X}^\top)^\top$. Let $\hat{\boldsymbol{\gamma}}$ denote the maximum likelihood estimator for $\boldsymbol{\gamma}$ based on $\{(A_i, \boldsymbol{X}_i)\}_{i=1}^n$, an inverse-probability weighted estimator (IPWE) for $V(d)$ is

$$\hat{V}^{\text{IPWE}}(d; \hat{\boldsymbol{\gamma}}) \triangleq \mathbb{P}_n\left[\frac{Y}{\pi(A, \boldsymbol{X}; \hat{\boldsymbol{\gamma}})} I\{A = d(\boldsymbol{X})\}\right],$$

where $\mathbb{P}_n$ is the empirical measure. It is straightforward to show that the IPWE is consistent for $V(d)$ if $\pi(A, \boldsymbol{X}; \boldsymbol{\gamma})$ is correctly specified, that is, $\pi_0(A, \boldsymbol{X}) = \pi(A, \boldsymbol{X}; \boldsymbol{\gamma}_0)$ for some $\boldsymbol{\gamma}_0$, but may not be otherwise.

Following ideas from Robins et al. (1994), an AIPWE can be constructed:

$$\hat{V}^{\text{AIPWE}}(d; \hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) \triangleq \mathbb{P}_n\left[\frac{Y I\{A = d(\boldsymbol{X})\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\}} - \frac{I\{A = d(\boldsymbol{X})\} - \pi\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\}} Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\}\right]. \quad (1)$$

By adding an augmentation term that involves both estimated propensity scores and regression models, the AIPWE improves efficiency and provides additional protection against model misspecification. The AIPWE is doubly robust in that it consistently estimates $V(d)$ as long as one of the nuisance working models is correctly specified, i.e., $\pi(A, \boldsymbol{X}; \hat{\boldsymbol{\gamma}}) \xrightarrow{p} \pi_0(A, \boldsymbol{X})$ or $Q(\boldsymbol{X}, A; \boldsymbol{\beta}) \xrightarrow{p} Q_0(\boldsymbol{X}, A)$.

Throughout the paper, we focus on the AIPWE, and suppress the superscript 'AIPWE' in $\hat{V}(d; \hat{\boldsymbol{\gamma}}, \boldsymbol{\beta})$. We refer to the estimator (1), with $\boldsymbol{\gamma}$ estimated by maximum likelihood and $\boldsymbol{\beta}$ estimated by least squares, as the usual doubly robust estimator from Zhang et al. (2012). However, when the propensity score model is correctly specified, but the outcome model is not, it is inefficient to adopt the least squares estimates of $\boldsymbol{\beta}$ in (1), where $\hat{V}(d; \hat{\boldsymbol{\gamma}}, \boldsymbol{\beta})$ could have a large variation. This motivates us to develop improved DR estimators with desirable efficiency properties.

2.2 Improved doubly robust estimators when the propensity score is fully specified

We first consider a fully specified propensity score model $\pi(A, \boldsymbol{X})$, say, involving no nuisance parameters. We will relax this shortly. Here, the specified propensity score model $\pi(A, \boldsymbol{X})$ may or may not be the same as the true propensity $\pi_0(A, \boldsymbol{X})$. For a fixed treatment regime d, the class of AIPW estimators for $V(d)$ is

$$\hat{V}(d; \boldsymbol{\beta}) \triangleq \mathbb{P}_n\left[\frac{Y I\{A = d(\boldsymbol{X})\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}} - \frac{I\{A = d(\boldsymbol{X})\} - \pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}} Q\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\}\right]. \quad (2)$$

We use $\hat{V}(d;\boldsymbol{\beta})$ to emphasize that the estimator does not involve the nuisance parameters related to propensity score, and varying the choice of $\boldsymbol{\beta}$ leads to different DR estimators with potentially very different behaviors. In the following, we will derive an estimator for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}^{\text{opt}}$, such that the resulting value function estimator $\hat{V}(d;\boldsymbol{\beta}^{\text{opt}})$ satisfies two properties:

(i) Doubly robust. $\hat{V}(d;\boldsymbol{\beta}^{\text{opt}})$ consistently estimates $V(d)$ when either the propensity score or the outcome model is correctly specified.

(ii) If the propensity score is correctly specified, it achieves the smallest asymptotic variance among all estimators of form (2), regardless of the specification of the outcome model.

Hence, when the outcome model is correctly specified, but the propensity score may not be, the desired $\boldsymbol{\beta}^{\text{opt}}$ must converge in probability to $\boldsymbol{\beta}_0$, where $Q(\boldsymbol{X}, A;\boldsymbol{\beta}_0) = Q_0(\boldsymbol{X}, A)$. On the other hand, when the propensity score is correctly specified, $\text{Var}\{\hat{V}(d;\boldsymbol{\beta}^{\text{opt}})\} \le \text{Var}\{\hat{V}(d;\boldsymbol{\beta})\}$ for any $\boldsymbol{\beta}$.

Lemma 1.—*Let $\boldsymbol{\beta}$ be any root-n consistent estimator converging in probability to some $\boldsymbol{\beta}^*$, i.e., $\boldsymbol{\beta} - \boldsymbol{\beta}^* = O_p(n^{-1/2})$. When the propensity score is correct, $\pi(A, \boldsymbol{X}) = \pi_0(A, \boldsymbol{X})$, but $Q(\boldsymbol{X}, A;\boldsymbol{\beta})$ may or may not be, the influence function for $\hat{V}(d;\boldsymbol{\beta})$ is*

$$\frac{YI\{A = d(\boldsymbol{X})\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} - \frac{I\{A = d(\boldsymbol{X})\} - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} Q\{\boldsymbol{X}, d(\boldsymbol{X});\beta^*\} - V(d)\,. \tag{3}$$

A proof is given in Appendix A. The preceding result shows that when the propensity score is correct, the asymptotic variance of $\hat{V}(d;\boldsymbol{\beta})$ does not depend on the sampling variation of $\boldsymbol{\beta}$ but only on its limit in probability $\boldsymbol{\beta}^*$. Based on Lemma 1 and the law of total variance, its asymptotic variance is proportional to

$$\begin{aligned} &E\left(\text{Var}\left[\frac{YI\{A = d(\boldsymbol{X})\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} - \frac{I\{A = d(\boldsymbol{X})\} - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} Q\{\boldsymbol{X}, d(\boldsymbol{X});\beta^*\} \mid \boldsymbol{X}\right]\right) \\ &+ \text{Var}\left(E\left[\frac{YI\{A = d(\boldsymbol{X})\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} - \frac{I\{A = d(\boldsymbol{X})\} - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} Q\{\boldsymbol{X}, d(\boldsymbol{X});\beta^*\} \mid \boldsymbol{X}\right]\right) \\ &= (I) + (II)\,. \end{aligned} \tag{4}$$

Notice that (II) in (4) equals $\text{Var}[Q_0\{\boldsymbol{X}, d(\boldsymbol{X})\}]$, which does not depend on $\boldsymbol{\beta}^*$. Furthermore, it can be shown that (see Appendix A for details)

$$\begin{aligned} (I) = E\left[\frac{1 - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} Q^2\{\boldsymbol{X}, d(\boldsymbol{X});\boldsymbol{\beta}^*\}\right] + E\left[\frac{YI\{A = d(\boldsymbol{X})\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} - Q_0\{\boldsymbol{X}, d(\boldsymbol{X})\}\right]^2 \\ - 2E\left[\frac{1 - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} Q_0\{\boldsymbol{X}, d(\boldsymbol{X})\} \cdot Q\{\boldsymbol{X}, d(\boldsymbol{X});\boldsymbol{\beta}^*\}\right] \end{aligned}$$

Denote the minimizer of (4) as $\boldsymbol{\beta}^{\text{opt}}$. By taking the derivative of (4) with respect to $\boldsymbol{\beta}^*$ and setting it equal to zero, $\boldsymbol{\beta}^{\text{opt}}$ is the solution to

$$E\left(\frac{1-\pi_0\{d(\boldsymbol{X}),\boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}),\boldsymbol{X}\}}[Q_0\{\boldsymbol{X},d(\boldsymbol{X})\}-Q\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\}]Q_{\boldsymbol{\beta}}\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\}\right)=0, \tag{5}$$

where $Q_{\boldsymbol{\beta}}(\boldsymbol{X},A;\boldsymbol{\beta}) \triangleq \partial Q(\boldsymbol{X},A;\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

Hence, if the outcome model $Q(\boldsymbol{X},A;\boldsymbol{\beta})$ is correct; that is, $Q(\boldsymbol{X},A;\boldsymbol{\beta}_0)=Q_0(\boldsymbol{X},A)$ for some $\boldsymbol{\beta}_0$, then in fact $\boldsymbol{\beta}^{\text{opt}}=\boldsymbol{\beta}_0$. If the outcome model is incorrect, such $\boldsymbol{\beta}^{\text{opt}}$ still exists and minimizes (4). However, the usual least squares estimates $\boldsymbol{\beta}^{\text{LS}}$ solving $\mathbb{P}_n\left[\{Y-Q(\boldsymbol{X},A;\boldsymbol{\beta})\}Q_{\boldsymbol{\beta}}(\boldsymbol{X},A;\boldsymbol{\beta})\right]=0$ does not converge to this $\boldsymbol{\beta}^{\text{opt}}$. This explains why the usual DR estimator $\hat{V}(d;\boldsymbol{\beta}^{\text{LS}})$ is sub-optimal when the outcome regression model is misspecified.

In the following, we will propose two different forms of estimators $\boldsymbol{\beta}^{\text{opt}}$, which converge in probability to $\boldsymbol{\beta}^{\text{opt}}$ and satisfy (i) and (ii) simultaneously. We first consider $\boldsymbol{\beta}^{\text{opt1}}$ as the solution to the following inverse-probability weighted estimating equation

$$\mathbb{P}_n\left(\frac{I\{A=d(\boldsymbol{X})\}}{\pi\{d(\boldsymbol{X}),\boldsymbol{X}\}}\frac{1-\pi\{d(\boldsymbol{X}),\boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}),\boldsymbol{X}\}}[Y-Q\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\}]Q_{\boldsymbol{\beta}}\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\}\right)=0. \tag{6}$$

This can be viewed as a weighted least squares based on subjects whose treatment assignments coincide with those recommended by d, with weights $[1-\pi\{d(\boldsymbol{X}),\boldsymbol{X}\}]/\pi^2\{d(\boldsymbol{X}),\boldsymbol{X}\}$. When the propensity score is correct, but the outcome regression may not be, the left-hand side of (6) converges in probability to the left-hand side of (5), hence $\boldsymbol{\beta}^{\text{opt1}} \xrightarrow{p} \boldsymbol{\beta}^{\text{opt}}$. On the other hand, when the outcome regression is correct but the propensity score may not be, the left-hand side of (6) converges in probability to

$$E\left(\frac{\pi_0\{d(\boldsymbol{X}),\boldsymbol{X}\}[1-\pi\{d(\boldsymbol{X}),\boldsymbol{X}\}]}{\pi^2\{d(\boldsymbol{X}),\boldsymbol{X}\}}[Q_0\{\boldsymbol{X},d(\boldsymbol{X})\}-Q\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\}]Q_{\boldsymbol{\beta}}\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\}\right),$$

which equals 0 when $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, thus $\boldsymbol{\beta}^{\text{opt1}} \xrightarrow{p} \boldsymbol{\beta}_0$. The following lemma formally establishes the improved doubly robust property of the proposed estimator $\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}})$. See Appendix A for the proof.

Lemma 2.— *$\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}}) \xrightarrow{p} V(d)$ when either the propensity score or the outcome regression model is correctly specified. In addition, when the propensity score model is correct, $\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}})$ achieves the smallest asymptotic variance among all estimators of form (2).*

The estimating equation (6) only utilizes the subjects whose treatment assignments coincide with those recommended by d. Since we need to search for the best treatment rule in a large class of ITRs, $\mathscr{D}$, it is possible that for some d, there are very few subjects satisfying

$A = d(\boldsymbol{X})$. This leads to highly unstable $\boldsymbol{\beta}^{\text{opt1}}$ and could be problematic, in particular when the sample size n is very small. To address this issue, we propose an *augmented* inverse-probability weighted estimating equation, denoted by $\boldsymbol{\beta}^{\text{opt2}}$, which is the solution to

$$(*) \quad -\mathbb{P}_n\left(\frac{I\{A = d(\boldsymbol{X})\} - \pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}\cdot \frac{1-\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}\left[\widehat{Q}_0\{\boldsymbol{X}, d(\boldsymbol{X})\} - Q\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\}\right]\mathcal{Q}_{\boldsymbol{\beta}}\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\}\right) = 0\,.$$

Here, (*) is the left hand side of (6), and $\hat{Q}_0\left\{\boldsymbol{X}, d(\boldsymbol{X})\right\} = \hat{Q}_0(\boldsymbol{X}, 1)I\left\{d(\boldsymbol{X}) = 1\right\} + \hat{Q}_0(\boldsymbol{X}, -1)I\left\{d(\boldsymbol{X}) = -1\right\}$. $\hat{Q}_0(\boldsymbol{X}, a)$ is the estimator for $E(Y \mid \boldsymbol{X}, A = a)$. We propose to use nonparametric techniques for obtaining $\hat{Q}_0(\boldsymbol{X}, a)$, which provides flexibility in model specification. For continuous $\boldsymbol{X}$, we apply the kernel regression method, i.e.,

$$\hat{Q}_0(\boldsymbol{X}, a) = \frac{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{X})I(A_i = a)Y_i}{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{X})I(A_i = a)},$$

where $K_{\boldsymbol{H}}(\cdot) = |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}\cdot)$ is a multivariate kernel with a bandwidth matrix $\boldsymbol{H}$. When $\boldsymbol{X}$ contains both continuous and categorical variables, the 'generalized product kernels' from Racine and Li (2004) is used. Under some regularity conditions, $\hat{Q}_0\left\{\boldsymbol{X}, d(\boldsymbol{X})\right\}$ is a consistent estimator for $Q_0\{\boldsymbol{X}, d(\boldsymbol{X})\}$. As a consequence, $\boldsymbol{\beta}^{\text{opt2}} \xrightarrow{p} \boldsymbol{\beta}^{\text{opt}}$ when the propensity score is correct; and $\boldsymbol{\beta}^{\text{opt2}} \xrightarrow{p} \boldsymbol{\beta}_0$ when the outcome regression is correct. The following lemma formally establishes the improved doubly robust property of the estimator $\hat{V}(d; \boldsymbol{\beta}^{\text{opt2}})$. The technical conditions and the proofs are provided in Appendix A.

Lemma 3.—$\hat{V}(d; \boldsymbol{\beta}^{\text{opt2}}) \to V(d)$ *when either the propensity score or the outcome regression model is correctly specified. In addition, when the propensity score is correct,* $\hat{V}(d; \boldsymbol{\beta}^{\text{opt2}})$ *achieves the smallest asymptotic variance among all estimators of form (2).*

2.3 Scenario where there is a nuisance parameter in the propensity score model

In practice, if the propensity scores are unknown, we can posit a parametric propensity score model $\pi(A, \boldsymbol{X}; \boldsymbol{\gamma})$ involving some nuisance parameters. To construct an improved DR estimator for the value function, we must take into account the effect of estimating $\boldsymbol{\gamma}$. Consider the class of AIPW estimators presented in (1). Let $\hat{\boldsymbol{\gamma}}$ be the maximum likelihood estimator of $\boldsymbol{\gamma}$ based on $\{(A_i, \boldsymbol{X}_i)\}_{i=1}^{n}$. We aim to find $\boldsymbol{\beta}^{\text{opt}}$ such that $\hat{V}(d; \hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}^{\text{opt}})$ is doubly robust, and has the smallest asymptotic variance among the class of estimators (1) when the propensity score is correctly specified.

Since $\hat{\boldsymbol{\gamma}}$ is the maximizer of the binomial likelihood

$$\prod_{i=1}^{n} \pi(1, \boldsymbol{X}_i; \boldsymbol{\gamma})^{I(A_i = 1)} \{1 - \pi(1, \boldsymbol{X}_i; \boldsymbol{\gamma})\}^{I(A_i = -1)},$$

the score vector for $\boldsymbol{\gamma}$ is

$$S_{\boldsymbol{\gamma}}(A, \boldsymbol{X}, \boldsymbol{\gamma}) = I(A = 1) \frac{\pi_{\boldsymbol{\gamma}}(1, \boldsymbol{X}; \boldsymbol{\gamma})}{\pi(1, \boldsymbol{X}; \boldsymbol{\gamma})} - I(A = -1) \frac{\pi_{\boldsymbol{\gamma}}(1, \boldsymbol{X}; \boldsymbol{\gamma})}{1 - \pi(1, \boldsymbol{X}; \boldsymbol{\gamma})},$$

where $\pi_{\boldsymbol{\gamma}}(1, \boldsymbol{X}; \boldsymbol{\gamma}) = \partial \pi(1, \boldsymbol{X}; \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$. When $\pi(A, \boldsymbol{X}; \boldsymbol{\gamma})$ is correctly specified, i.e. $\pi(A, \boldsymbol{X}; \boldsymbol{\gamma}_0) = \pi_0(A, \boldsymbol{X})$ for some $\boldsymbol{\gamma}_0$, and $\boldsymbol{\beta}$ converging in probability to $\boldsymbol{\beta}^*$, the influence functions corresponding to estimators of the form (1) have the following expression

$$\tilde{\varphi}(Y, A, \boldsymbol{X}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}^*) - \Gamma_0(\boldsymbol{\beta}^*) \Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}, 0}^{-1} S_{\boldsymbol{\gamma}}(A, \boldsymbol{X}, \boldsymbol{\gamma}_0), \tag{7}$$

where $\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}, 0} = E\{S_{\boldsymbol{\gamma}}(A, \boldsymbol{X}, \boldsymbol{\gamma}_0) S_{\boldsymbol{\gamma}}^{\top}(A, \boldsymbol{X}, \boldsymbol{\gamma}_0)\}, \Gamma_0(\boldsymbol{\beta}) = -E\{\partial \tilde{\varphi}(Y, A, \boldsymbol{X}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}) / \partial \boldsymbol{\gamma}^{\top}\}$, and $\tilde{\varphi}(Y, A, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{Y I\{A = d(\boldsymbol{X})\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}\}} - \frac{I\{A = d(\boldsymbol{X})\} - \pi\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}\}} Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\} - V(d)$.

Compared with (3), the influence functions (7) involve an additional term due to estimation of $\boldsymbol{\gamma}$. However, this additional term disappears when both models are correct. Define $\Phi_0(\boldsymbol{\beta}) \triangleq -E\{\partial^2 \tilde{\varphi}(Y, A, \boldsymbol{X}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}) / \partial \boldsymbol{\gamma}^{\top} \partial \boldsymbol{\beta}\}$. In a slight abuse of notation, denote the minimizer of the variance of (7) as $\boldsymbol{\beta}^{\text{opt}}$. It is the solution to

$$\begin{aligned} &E\left(\frac{1 - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}} \left[Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\} + \Phi_0(\boldsymbol{\beta}) \Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}, 0}^{-1} \frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}_0\}}{1 - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}\right]\right. \\ &\left. \cdot \left[Q_0\left\{\boldsymbol{X}, d(\boldsymbol{X})\right\} - Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\} - \Gamma_0(\boldsymbol{\beta}) \Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}, 0}^{-1} \cdot \frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}_0\}}{1 - \pi_0\{d(\boldsymbol{X}), \boldsymbol{X}\}}\right]\right) = 0. \end{aligned} \tag{8}$$

Detailed derivations of the influence function and its variance are deferred to Appendix A.

To compress notations, we write to $R \triangleq I\{A = d(\boldsymbol{X})\}$. Consider $\boldsymbol{\beta}^{\text{opt3}}$ as the solution to

$$\begin{aligned} &\mathbb{P}_n\left(\frac{R[1 - \pi\left\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\right\}}{\pi^2\left\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\right\}} \left[Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\} + \hat{\Phi}(\boldsymbol{\beta}) \hat{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} \cdot \frac{\pi_{\boldsymbol{\gamma}}\left\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\right\}}{1 - \pi\left\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\right\}}\right]\right. \\ &\left. \left[Y - Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\} - \hat{\Gamma}(\boldsymbol{\beta}) \hat{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} \frac{\pi_{\boldsymbol{\gamma}}\left\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\right\}}{1 - \pi\left\{d(\boldsymbol{X}), \boldsymbol{X}; \hat{\boldsymbol{\gamma}}\right\}}\right]\right) = 0, \end{aligned} \tag{9}$$

where $\hat{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \mathbb{P}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\left\{S_{\boldsymbol{\gamma}}(A, \boldsymbol{X}, \hat{\boldsymbol{\gamma}}) S_{\boldsymbol{\gamma}}^{\top}(A, \boldsymbol{X}, \hat{\boldsymbol{\gamma}})\right\}, \hat{\Gamma}(\boldsymbol{\beta}) = -\mathbb{P}_n\left\{\partial \tilde{\varphi}(Y, A, \boldsymbol{X}, \hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) / \partial \boldsymbol{\gamma}^{\top}\right\}$, and $\hat{\Phi}(\boldsymbol{\beta}) = -\mathbb{P}_n\left\{\partial^2 \tilde{\varphi}(Y, A, \boldsymbol{X}, \hat{\gamma}, \boldsymbol{\beta}) / \partial \boldsymbol{\gamma}^{\top} \partial \boldsymbol{\beta}\right\}$. In Appendix A, we show that $\hat{V}(d; \hat{\gamma}, \boldsymbol{\beta}^{\text{opt3}})$ is doubly robust, and achieves the smallest asymptotic variance with $\boldsymbol{\beta}^{\text{opt3}} \xrightarrow{p} \boldsymbol{\beta}^{\text{opt}}$ when the propensity score is correct.

Correspondingly, we can construct an *augmented* inverse-probability weighted estimating equation and consider $\boldsymbol{\beta}^{\text{opt4}}$ as the solution to

$$\left(**\right) - \mathbb{P}_n\left(\frac{[R-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}][1-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}]}{\pi^2\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}} \cdot \left[Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\right\} + \hat{\Phi}(\boldsymbol{\beta})\hat{\Sigma}_{\gamma\gamma}^{-1}\cdot\frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}}{1-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}}\right] \cdot \left[\hat{Q}_0\left\{\boldsymbol{X},d(\boldsymbol{X})\right\} - Q\left\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\right\} - \hat{\Gamma}(\boldsymbol{\beta})\hat{\Sigma}_{\gamma\gamma}^{-1}\frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}}{1-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\hat{\boldsymbol{\gamma}}\}}\right]\right) = 0, \tag{10}$$

where (**) is the left hand side of (9). Using a similar argument, $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt4}})$ satisfies (i) and (ii), and is improved doubly robust.

In the above discussion, we proposed improved DR estimators of $V(d)$ for a fixed treatment regime d. Notice that by (8), the optimal value $\boldsymbol{\beta}^{\text{opt}}$ is d-dependent, i.e., different d's correspond to different $\boldsymbol{\beta}^{\text{opt}}$. Rigorously speaking, we should write $\boldsymbol{\beta}^{\text{opt}}(d)$, and $\boldsymbol{\beta}^{\text{opt}}(d)$ for the nuisance parameter estimates. To estimate d^{opt}, we first identify the corresponding $\boldsymbol{\beta}^{\text{opt}}(d)$ and the $\hat{V}\{d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(d)\}$, for each $d\in\mathscr{D}$. We then find the optimal d among the class $\mathscr{D}$ that leads to the largest $\hat{V}\{d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(d)\}$, i.e., $\hat{d}=\text{argmax}_{d\in\mathscr{D}}\hat{V}\{d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(d)\}$. In practice, the ITR is often indexed by a set of parameters, for instance, $d(\boldsymbol{x})=\text{sign}\{\boldsymbol{x}^\top\boldsymbol{\eta}\}$, where $\boldsymbol{x}=(1,\boldsymbol{x}^\top)^\top$. Since $\hat{V}\{d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(d)\}$ is a nonsmooth function of $\boldsymbol{\eta}$, standard optimization methods can be problematic. We used a genetic algorithm discussed by Goldberg (1989), which is available in the R package rgenoud (Mebane Jr and Sekhon, 2011). In the rest of the paper, we suppress the letter d in $\boldsymbol{\beta}^{\text{opt}}(d)$ and $\boldsymbol{\beta}^{\text{opt}}(d)$ when there is no confusion.

3 Theoretical results

In this section, we establish asymptotic normality of the proposed estimators and the usual doubly robust estimator of $V(d)$. We do not discuss the situation when both propensity score and outcome models are misspecified, given that the resulting estimator is not consistent for $V(d)$. We first consider the case where propensity score is fully specified. We have the following result.

Theorem 1.

(Asymptotic normality when propensity score model is full specified). When either the propensity score or the outcome model is correct,

$$\sqrt{n}\left\{\hat{V}(d;\boldsymbol{\beta}^{\text{LS}})-V(d)\right\}\xrightarrow{D}N(0,U_1(\boldsymbol{\theta}_0^{\text{LS}})),$$

$$\sqrt{n}\left\{\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}})-V(d)\right\}\xrightarrow{D}N(0,U_2(\boldsymbol{\theta}_0^{\text{opt1}})),$$

$$\sqrt{n}\left\{\hat{V}(d;\boldsymbol{\beta}^{\text{opt2}}) - V(d)\right\} \xrightarrow{D} N(0, U_3(\boldsymbol{\theta}_0^{\text{opt2}})).$$

See Appendix C for detailed expressions of $U_1(\boldsymbol{\theta}), U_2(\boldsymbol{\theta}), U_3(\boldsymbol{\theta})$. *The true parameters are*

$\boldsymbol{\theta}_0^{\text{LS}} = (\boldsymbol{\beta}_{\text{LS}}^{*\top}, V(d))^\top$ *where* $\boldsymbol{\beta}_{\text{LS}}^*$ *satisfies* $E\left[Q_{\boldsymbol{\beta}}(\boldsymbol{X}, A; \boldsymbol{\beta}_{\text{LS}}^*)\left\{Y - Q(\boldsymbol{X}, A; \boldsymbol{\beta}_{\text{LS}}^*)\right\}\right] = \boldsymbol{0}$.

$\boldsymbol{\theta}_0^{\text{opt1}} = (\boldsymbol{\beta}_{\text{opt1}}^{*\top}, V(d))^\top$ *where* $\boldsymbol{\beta}_{\text{opt1}}^*$ *satisfies*

$$E\left(\frac{I\{A = d(\boldsymbol{X})\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}} \cdot \frac{1 - \pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}\left[Y - Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}_{\text{opt2}}^*\right\}\right] Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}_{\text{opt1}}^*\right\}\right) = \boldsymbol{0}.$$

$\boldsymbol{\theta}_0^{\text{opt2}} = (\boldsymbol{\beta}_{\text{opt2}}^{*\top}, V(d))^\top$ where $\boldsymbol{\beta}_{\text{opt2}}^*$ satisfies

$$\begin{aligned} &E\left(\frac{I\{A = d(\boldsymbol{X})\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}} \cdot \frac{1 - \pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}\left[Y - Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}_{\text{opt2}}^*\right\}\right] Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}_{\text{opt2}}^*\right\}\right) \\ &- E\left(\frac{I\{A = d(\boldsymbol{X})\} - \pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}} \cdot \frac{1 - \pi\{d(\boldsymbol{X}), \boldsymbol{X}\}}{\pi\{d(\boldsymbol{X}), \boldsymbol{X}\}} \cdot \right. \\ &\left.\left[Q_0\left\{\boldsymbol{X}, d(\boldsymbol{X})\right\} - Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}_{\text{opt2}}^*\right\}\right] Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}_{\text{opt2}}^*\right\}\right) = \boldsymbol{0}. \end{aligned}$$

The estimators $\hat{V}(d;\boldsymbol{\beta}^{\text{LS}})$ and $\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}})$ involve solving jointly a set of M-estimating equations (Stefanski and Boos, 2002). Thus, the asymptotic variance of $\hat{V}(d;\boldsymbol{\beta}^{\text{LS}})$ and $\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}})$ can be calculated based on standard M-estimation theory. The estimator $\hat{V}(d;\boldsymbol{\beta}^{\text{opt2}})$ is obtained by solving a set of estimating equations where some infinite dimensional parameters, in this case, $Q_0(\boldsymbol{X}, a), a = \pm 1$, are estimated nonparametrically in the first stage, which is referred to as semiparametric M-estimators (Andrews, 1994; Newey, 1994; Chen et al., 2003; Ichimura and Lee, 2010). In Appendix B, we establish asymptotic normality of such semiparametric M-estimators by extending Theorem 2 in Chen et al. (2003). The detailed proof of Theorem 1 can be found in Appendix C.

Remark 1.

When propensity score is correct, it can be shown that $U_1(\boldsymbol{\theta}) = U_2(\boldsymbol{\theta}) = U_3(\boldsymbol{\theta})$ *which equals to (4), the asymptotic variance of the influence function for* $\hat{V}(d;\boldsymbol{\beta})$. *In addition, when propensity score is correct but outcome model incorrect,* $\boldsymbol{\beta}_{\text{opt1}}^* = \boldsymbol{\beta}_{\text{opt2}}^* = \boldsymbol{\beta}^{\text{opt}}$. *Recall that* $\boldsymbol{\beta}^{\text{opt}}$ *is defined in (5), which minimizes the asymptotic variance (4). However,* $\boldsymbol{\beta}_{\text{LS}}^*$, *the limit of least squares estimates, is different from* $\boldsymbol{\beta}^{\text{opt}}$. *Consequently,* $\hat{V}(d;\boldsymbol{\beta}^{\text{opt1}})$ *and* $\hat{V}(d;\boldsymbol{\beta}^{\text{opt2}})$ *have the same asymptotic variance, which is smaller than that of* $\hat{V}(d;\boldsymbol{\beta}^{\text{LS}})$. *Though, in small sample size scenarios,* $\hat{V}(d;\boldsymbol{\beta}^{\text{opt2}})$ *is preferred since it utilizes the complete data, and could lead to a more stable estimate. When both models are correct,* $\boldsymbol{\beta}^{\text{opt}} = \boldsymbol{\beta}_{\text{LS}}^* = \boldsymbol{\beta}_0$, *where* $\boldsymbol{\beta}_0$ *satisfies* $Q(\boldsymbol{X}, A; \boldsymbol{\beta}_0) = Q_0(\boldsymbol{X}, A)$. *As a result, all three estimators have the same asymptotic variance. When the outcome model is correct but propensity score incorrect, it is not possible to directly compare the asymptotic variances of these estimators.*

The following theorem presents asymptotic properties of $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt3}})$ and $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt4}})$, the estimators for the value function of d where there is a nuisance parameter in the propensity score model.

Theorem 2.

(Asymptotic normality when there is a nuisance parameter in the propensity score model). When either the propensity score or the outcome model is correct,

$$\sqrt{n}\left\{\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{LS}}) - V(d)\right\} \xrightarrow{D} N(0, U_4(\boldsymbol{\theta}_0^{\text{LS2}}))\,.$$

The true values arewhere $\boldsymbol{\theta}_0^{\text{LS2}} = (\boldsymbol{\gamma}^{*\top}, \boldsymbol{\beta}_{\text{LS}}^{*\top}, V(d))^{\top}$ *where* $\boldsymbol{\gamma}^*$ *satisfies* $E\left\{S_{\boldsymbol{\gamma}}(A, \boldsymbol{X}; \boldsymbol{\gamma}^*)\right\} = \boldsymbol{0}$.

When either the propensity score or the outcome model is correct,

$$\sqrt{n}\left\{\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt3}}) - V(d)\right\} \xrightarrow{D} N(0, U_5(\boldsymbol{\theta}_0^{\text{opt3}})),$$

$$\sqrt{n}\left\{\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt4}}) - V(d)\right\} \xrightarrow{D} N(0, U_6(\boldsymbol{\theta}_0^{\text{opt4}}))\,.$$

The true parameters are $\boldsymbol{\theta}_0^{\text{opt3}} = (\boldsymbol{\gamma}^{*\top}, \boldsymbol{\zeta}_{\text{opt3}}^{*\top}, \boldsymbol{\beta}_{\text{opt3}}^{*\top}, V(d))^{\top}$ *where* $(\boldsymbol{\zeta}_{\text{opt3}}^*, \boldsymbol{\beta}_{\text{opt3}}^*)$ *is the solution to the following set of equations:*

$$\begin{aligned}
&\boldsymbol{\alpha} + E\{\partial\tilde{\varphi}(Y, A, \boldsymbol{X}, \boldsymbol{\gamma}^*, \boldsymbol{\beta})/\partial\boldsymbol{\gamma}\} = \boldsymbol{0},\\
&[\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_q] - E\left\{S_{\boldsymbol{\gamma}}(A, \boldsymbol{X}, \boldsymbol{\gamma}^*)S_{\boldsymbol{\gamma}}^{\top}(A, \boldsymbol{X}, \boldsymbol{\gamma}^*)\right\} = \boldsymbol{0},\\
&[\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q] + E\left\{\partial^2\tilde{\varphi}(Y, A, \boldsymbol{X}, \boldsymbol{\gamma}^*, \boldsymbol{\beta})/\partial\boldsymbol{\gamma}^{\top}\partial\boldsymbol{\beta}\right\} = \boldsymbol{0},
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
&E\left(\frac{R[1-\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}^*\}]}{\pi^2\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}^*\}}\left[Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\}\right.\right.\\
&\left.+ [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q][\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_q]^{-1}\frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}^*\}}{1-\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}^*\}}\right]\\
&\left.\cdot\left[Y - Q\left\{\boldsymbol{X}, d(\boldsymbol{X}); \boldsymbol{\beta}\right\} - \boldsymbol{\alpha}^{\top}[\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_q]^{-1}\frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}^*\}}{1-\pi\{d(\boldsymbol{X}), \boldsymbol{X}; \boldsymbol{\gamma}^*\}}\right]\right) = \boldsymbol{0}\,.
\end{aligned} \tag{12}$$

The true parameters are $\boldsymbol{\theta}_0^{\text{opt4}} = (\boldsymbol{\gamma}^{*\top}, \boldsymbol{\zeta}_{\text{opt4}}^{*\top}, \boldsymbol{\beta}_{\text{opt4}}^{*\top}, V(d))^{\top}$ *where* $(\boldsymbol{\zeta}_{\text{opt4}}^*, \boldsymbol{\beta}_{\text{opt4}}^*)$ *is the solution to (11) and*

$$(***) - E\left(\frac{[R-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}][1-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}]}{\pi^2\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}} \cdot \left[Q_{\boldsymbol{\beta}}\left\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\right\} + [\phi_1,\ldots,\phi_q][\psi_1,\ldots,\psi_q]^{-1}\frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}}{1-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}}\right] \cdot \left[Q_0\left\{\boldsymbol{X},d(\boldsymbol{X})\right\} - Q\left\{\boldsymbol{X},d(\boldsymbol{X});\boldsymbol{\beta}\right\} - \boldsymbol{\alpha}^\top[\psi_1,\ldots,\psi_q]^{-1}\frac{\pi_{\boldsymbol{\gamma}}\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}}{1-\pi\{d(\boldsymbol{X}),\boldsymbol{X};\boldsymbol{\gamma}^*\}}\right]\right) = \boldsymbol{0},$$

where (***) is the left hand side of (12). See Appendix D for detailed expressions of $U_4(\boldsymbol{\theta}), U_5(\boldsymbol{\theta}), U_6(\boldsymbol{\theta})$ *and the definitions of* $\boldsymbol{\zeta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\psi}^\top, \phi^\top)^\top$, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^\top, \ldots, \boldsymbol{\psi}_q^\top)^\top$, $\phi = (\phi_1^\top, \ldots, \phi_q^\top)^\top$. *Note that here we use* $[\phi_1, \ldots, \phi_q]$ *to represent a matrix with j-th column being* ϕ_j, *similarly for* $[\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_q]$.

Remark 2.

When the propensity score model is correct, observe that $\theta_0^{\text{opt3}} = \theta_0^{\text{opt4}}$ *where* $\boldsymbol{\beta}^*_{\text{opt3}} = \boldsymbol{\beta}^*_{\text{opt4}} = \boldsymbol{\beta}^{\text{opt}}$, *where* $\boldsymbol{\beta}^{\text{opt}}$ *is the the minimizer of the variance of (7) in this case. Furthermore, it can be shown that* $U_4(\theta_0^{\text{LS2}})$, $U_5(\theta_0^{\text{opt3}})$, $U_6(\theta_0^{\text{opt4}})$ *equal to the variance of (7) evaluated at* $\boldsymbol{\beta} = \boldsymbol{\beta}^*_{\text{LS}}$ *and* $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{\text{opt}}$, *respectively. Therefore, when the propensity score is correct but outcome incorrect,* $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt3}})$ *and* $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt4}})$ *are asymptotically equivalent and more efficient than* $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{LS}})$. *When both models are correct, all three estimators are asymptotically equivalent.*

Remark 3.

An estimator for $V(d^{\text{opt}})$, *the overall population mean under the optimal regime, may be found as* $\hat{V}\{\hat{d};\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(\hat{d})\}$. *Following Zhang et al. (2012),* $n^{1/2}\left[\hat{V}\{\hat{d};\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(\hat{d})\} - V(d^{\text{opt}})\right] = n^{1/2}\left[\hat{V}\{d^{\text{opt}};\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(d^{\text{opt}})\} - V(d^{\text{opt}})\right] + o_p(1)$.

Thus, the asymptotic variance of $\hat{V}\{\hat{d};\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(\hat{d})\}$ *can be approximated by that of* $\hat{V}\{d^{\text{opt}};\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt}}(d^{\text{opt}})\}$, *which by Theorem 2 can be estimated using the usual sandwich technique.*

4 Simulation Studies

We conducted several simulation studies to evaluate the finite sample performance of our proposed method. The following six methods were compared: *Q*-learning based on linear regression (QL-LR, Qian and Murphy (2011)); *Q*-learning based on kernel regression (QL-KR); maximizing $\hat{V}^{\text{IPWE}}(d;\hat{\boldsymbol{\gamma}})$ within a pre-specified class of ITRs (IPWE); maximizing $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta})$ where standard maximum likelihood estimators are used for the nuisance parameters (Usual-DR, Zhang et al. (2012)); maximizing $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt3}})$ where $\boldsymbol{\beta}^{\text{opt3}}$ solves the IPW estimating equation (9) (Improved-DR); maximizing $\hat{V}(d;\hat{\boldsymbol{\gamma}},\boldsymbol{\beta}^{\text{opt4}})$ where DR). $\boldsymbol{\beta}^{\text{opt4}}$ solves the *augmented* IPW estimating equation (10) (Aug-Improved-DR).

The simulation set up is similar to Kang and Schafer (2007) with some modifications. $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4)$ was generated as standard multivariate normal, and $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ was defined as $X_1 = \exp(Z_1/2), X_2 = Z_2/\{1 + \exp(Z_1)\} + 10, X_3 = (Z_1 Z_3/25 + 0.6)^3, X_4 = (Z_2 + Z_4 + 20)^2$, so that $\boldsymbol{Z}$ can be expressed in terms of $\boldsymbol{X}$. The treatment A was generated from $\{-1, 1\}$ according to the model $P(A = 1|\boldsymbol{X}) = \exp\{l(\boldsymbol{X})\}/[1 + \exp\{l(\boldsymbol{X})\}]$, where $l(\boldsymbol{x}) = -z_1 + 0.5z_2 - 0.25z_3 - 0.1z_4$ in Scenario 1, and $l(\boldsymbol{x}) = 0.5z_1 - 0.5$ in Scenario 2. The response variable was normally distributed with $Y = 10 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4 + A(-1 - 10Z_1 + 10Z_2) + \epsilon$, where $\epsilon \sim N(0, 1)$. It is straightforward to deduce that $d^{\text{opt}}(\boldsymbol{x}) = \text{sign}(-1 - 10z_1 + 10z_2)$. Via Monte Carlo simulation with 10^6 replicates, we obtained $E\{Y(d^{\text{opt}})\} = 21.32$. The following modeling choices are considered for the propensity and outcome regression models.

CCA correctly specified logistic regression model for $\pi_0(A; \boldsymbol{X})$ with $\boldsymbol{Z}$ as predictors in both scenarios, and a correctly specified model for $Q_0(\boldsymbol{X}, A)$ with $\boldsymbol{Z}$, A, $\boldsymbol{Z}A$ as predictors in both scenarios.

CI A correctly specified logistic regression model for $\pi_0(A; \boldsymbol{X})$ with $\boldsymbol{Z}$ as predictors in both scenarios, and an incorrectly specified model for $Q_0(\boldsymbol{X}, A)$ with $\boldsymbol{X}$, A, $\boldsymbol{Z}A$ as predictors in both scenarios.

IC An incorrectly specified logistic regression model for $\pi_0(A; \boldsymbol{X})$ with $\boldsymbol{X}$ as predictors in Scenario 1, and without any predictors in Scenario 2, and a correctly specified model for $Q_0(\boldsymbol{X}, A)$ with $\boldsymbol{Z}$, A, $\boldsymbol{Z}A$ as predictors in both scenarios.

II An incorrectly specified logistic regression model for $\pi_0(A; \boldsymbol{X})$ with $\boldsymbol{X}$ as predictors in Scenario 1, and without any predictors in Scenario 2, and an incorrectly specified model for $Q_0(\boldsymbol{X}, A)$ with $\boldsymbol{X}$, A, $\boldsymbol{Z}A$ as predictors in both scenarios.

For IPWE, we use C. and I. to denote correct and incorrect propensity models, respectively. For QL-LR, we use.C and.I to denote correct and incorrect linear regression models. For QL-KR, we use .C and .I to denote kernel regression based on ($\boldsymbol{Z}$, A) and based on ($\boldsymbol{X}$, A), respectively. In all direct-maximization methods (IPWE, Usual-DR, Improved-DR, Aug-Improved-DR), we choose $\mathscr{D} = \{\text{sign}(\eta_0 + \eta_1 z_1 + \eta_2 z_2 + \eta_3 z_3 + \eta_4 z_4)\}$ so that $d^{\text{opt}} \in \mathscr{D}$. By imposing $\|\boldsymbol{\eta}\| = 1$, d^{opt} corresponds to $(\eta_0, \eta_1, \eta_2, \eta_3, \eta_4) = (-0.07, -0.71, 0.71, 0, 0)$.

For each scenario, we considered four sample sizes for training datasets: n = 100, 250, 500 or 1000, and repeated the simulation 500 times. The ITRs are constructed based on the training set and then evaluated on a large and independent test set (size 10000) based on two criteria: value function, i.e., the overall population mean when we apply the estimated optimal ITR to the test dataset; the misclassification error rate of the estimated optimal ITR from the true optimal ITR, i.e., $\mathbb{P}_n^* \left[I\{\hat{d}(\boldsymbol{X}) \neq d^{\text{opt}}(\boldsymbol{X})\} \right]$. Here $\mathbb{P}_n^*$ denotes the empirical measure using the test data.

Results for Scenario 1 are presented in Figure 1, where we draw boxplots of the value functions over 500 replications. Here we only report the results for $n = 250$ or 1000 (see Appendix E for further results, e.g., $n = 100$ or 500). As expected, Q-learning works the best if the outcome model is correctly specified but has relatively poor performance if this model is incorrect. When the outcome model is correct (CC, IC), Aug-Improved-DR and Usual-DR have similar performance. This is not surprising. Recall that when the outcome model is correct, the proposed nuisance parameter estimate $\boldsymbol{\beta}^{\text{opt4}}$ converges in probability to $\boldsymbol{\beta}_0$, the same limit of the least squares estimates. When the propensity model is correct but the outcome regression model is misspecified (CI), Aug-Improved-DR dominates Usual-DR, evidenced by larger value functions and smaller variance in value functions, e.g., the mean (sd) of value functions for Aug-Improved-DR are 21.07 (0.43) and 21.23 (0.39) when the sample size is 250 and 1000, respectively. Comparatively, for Usual-DR, the mean (sd) of value functions are 20.52 (0.76) and 21.07 (0.40). In addition, note that Improved-DR and Aug-Improved-DR have almost identical performance when the sample size is large ($n = 1000$). However, Improved-DR is unstable under small sample size ($n = 250$). This justifies the need to construct *augmented* IPW estimating equations to estimate the nuisance parameters, as we discussed in the method section.

To better demonstrate the superior performance of our proposed method under the CI setting, we focus on the comparison between Aug-Improved-DR and Usual-DR in terms of the misclassification rates. Results for Scenario 1 are shown in Figure 2 with sample sizes ranging from 100 to 1000. Notice that Aug-Improved-DR produced much smaller misclassification rates as well as smaller variations. In particular, it outperforms the usual DR estimator by a large margin when the sample size is small.

Simulation results for Scenario 2 are provided in Figure 3 and Appendix E. Again, the proposed method outperforms other competing methods in both value functions and misclassification rates. In Appendix E, we also report the mean squared errors (MSE) of different methods in terms of estimating $\boldsymbol{\eta}$. Aug-Improved-DR has smaller MSE than its competitors.

5 Application to the STAR*D Study

We apply the proposed method to analyze data from the STAR*D Study (Rush et al., 2004). Funded by the National Institute of Mental Health, the study was conducted to compare various treatment options for major depressive disorder when patients fail to respond to the initial treatment of citalopram (CIT). From 2001 to 2006, a total of 4041 outpatients with nonpsychotic depression, aged 18–75, were enrolled from 41 clinical sites in the U.S. The score on the 16-item Quick Inventory of Depressive Symptomatology (QIDS) was the primary outcome. The QIDS score ranges from 0 to 27, where higher scores indicate more severe depression.

The trial had four levels (see Fig. 1 in Rush et al. (2004)). Here, we focused on the first two levels. At level-1, patients received CIT for 12 to 14 weeks. Those who achieved clinically meaningful response (total QIDS score under 5) were remitted from future treatments. At level-2, participants without a satisfactory response to CIT had the option to either switch to

a different medication, or to augment their existing citalopram. Those in the "switch" group were randomly assigned to bupropion (BUP), cognitive therapy (CT), sertraline (SER), or venlafaxine (VEN). Those in the "augment" group were randomly assigned to CIT+BUP, CIT+buspirone (BUS), or CIT+CT. If a patient had no preference, he/she was assigned to any of the above treatments.

We use the QIDS score at the end of level-2 as the clinical outcome Y and compared two categories of treatments: (i) treatment with selective serotonin reputake inhibitors (SSRI): CIT+BUP, CIT+BUS, CIT+CT, and SER; (ii) non-SSRI: BUP, CT, and VEN. Denote $A = 1$ for SSRI and $A = -1$ for non-SSRI. Since patients in the "augment" group were all treated with SSRIs (violating the positivity assumption), we exclude these subjects from our analysis, which leaves a total of 817 subjects. Among them, 656 and 161 patients were in the "switch" and "no preference" group, respectively. 296 patients received SSRI treatments, while 521 patients received non-SSRI treatments. Comparisons using t-test show that there is no significant difference between the SSRI and the non-SSRI category with respect to QIDS scores.

We applied four methods to estimate the optimal ITR for those patients who had entered level-2. Prognostic variables $\boldsymbol{X}$ include QIDS score at the start of level-2, change of QIDS score during the level-1 period, preference regarding level-2 treatment, and other demographic variables such as gender, race, age, education level and employment status. The propensity scores $\pi_0(A, \boldsymbol{X})$ estimated by empirical proportions based on preferring to switch or no are preference. We used a linear regression of Y given $(\boldsymbol{X}, A, \boldsymbol{X}A)$ for the outcome model. For all methods, we randomly split the data into training and test set with 1:1 ratio. The estimated ITR was obtained using the training set, and then evaluated on the test set by $\mathbb{P}_n^*\left[YI\{A = \hat{d}(\boldsymbol{X})\}/\hat{\pi}_0(A, \boldsymbol{X})\right]/\mathbb{P}_n^*\left[I\{A = \hat{d}(\boldsymbol{X})\}/\hat{\pi}_0(A, \boldsymbol{X})\right]$. This procedure is repeated 500 times. Results for IPWE, QL-LR, Usual-DR and Aug-Improved-DR are displayed in Figure 4, where lower scores are desirable. The estimated QIDS score by using Aug-Improved-DR is 9.62 (sd = 0.37), which is smaller than IPWE (10.15, sd = 0.36), QL-LR (9.87, sd = 0.38), and Usual-DR (9.65, sd = 0.40). In addition, Aug-Improved-DR outperformed the one-size-fits-all approaches (QIDS score of 9.98 for SSRI and 10.12 for non-SSRI).

6 Discussion

In this article, we proposed an improved DR estimator for the optimal ITRs by directly maximizing an AIPWE of the marginal mean outcome over a class of ITRs. Our estimator is doubly robust, and designed to be more efficient than other DR estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. As shown in the numerical studies, the proposed method achieves better performance compared to other existing methods. The proposed method is appealing, given that in many practical applications, correct specification of the outcome model can be challenging, while the propensity score is either known by design or more likely to be correctly specified.

There are several important ways this work may be extended. The first is to extend it to the multi-stage decision setting. Zhang et al. (2013) proposed a doubly robust estimator for the optimal DTR where the nuisance parameters indexing the outcome models are estimated iteratively by a sequence of least squares regressions. More efficient DR estimators could be obtained if we use IPW or augmented IPW estimating equations to estimate these nuisance parameters. This is the direction we are currently pursuing.

Another future direction is to consider biased-reduced doubly robust estimation, i.e., estimate the nuisance parameters so as to minimize the bias of the DR estimator under misspecification of both working models. Vermeulen and Vansteelandt (2015) proposed biased-reduced DR estimators for several missing data and causal inference models. It would be interesting to investigate whether this principle can be adapted to the context of estimating optimal treatment regimes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge support by R01DK108073 awarded by the National Institutes of Health.

References

Andrews DW (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica, 62:43–72.

Bang H. and Robins JM (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973. [PubMed: 16401269]

Blatt D, Murphy SA, and Zhu J. (2004). A-learning for approximate planning. Ann Arbor, 1001:48109–2122.

Cao W, Tsiatis AA, and Davidian M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika, 96(3):723–734. [PubMed: 20161511]

Chan IS and Ginsburg GS (2011). Personalized medicine: progress and promise. Annual review of genomics and human genetics, 12:217–244.

Chen X, Linton O, and Van Keilegom I. (2003). Estimation of semiparametric models when the criterion function is not smooth. Econometrica, 71(5):1591–1608.

Collins FS and Varmus H. (2015). A new initiative on precision medicine. New England journal of medicine, 372(9):793–795.

Goldberg DE (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

Hamburg MA and Collins FS (2010). The path to personalized medicine. New England Journal of Medicine, 363(4):301–304.

Ichimura H. and Lee S. (2010). Characterization of the asymptotic distribution of semiparametric M-estimators. Journal of Econometrics, 159(2):252–266.

Imbens GW and Rubin DB (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Kang JD and Schafer JL (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science, 22(4):523–539.

Laber EB, Linn KA, and Stefanski LA (2014). Interactive model building for Q-learning. Biometrika, 101(4):831–847. [PubMed: 25541562]

Liu Y, Wang Y, Kosorok MR, Zhao Y, and Zeng D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. Statistics in medicine, 37(26):3776–3788. [PubMed: 29873099]

Mebane WR Jr and Sekhon JS (2011). Genetic optimization using derivatives: the rgenoud package for R. Journal of Statistical Software, 42(11):1–26.

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

Newey WK (1994). The asymptotic variance of semiparametric estimators. Econometrica, 62(6):1349–1382.

Qian M. and Murphy SA (2011). Performance guarantees for individualized treatment rules. Annals of statistics, 39(2):1180. [PubMed: 21666835]

Racine J. and Li Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics, 119(1):99–130.

Robins J, Sued M, Lei-Gomez Q, and Rotnitzky A. (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. Statistical Science, 22(4):544–559.

Robins JM (2004). Optimal structural nested models for optimal sequential decisions. In Proceedings of the second seattle Symposium in Biostatistics, pages 189–326. Springer.

Robins JM, Hernan MA, and Brumback B. (2000). Marginal structural models and causal inference in epidemiology.

Robins JM and Rotnitzky A. (2001). Comment on the bickel and kwon article, “inference for semiparametric models: Some questions and an answer”. Statistica Sinica, 11(4):920–936.

Robins JM, Rotnitzky A, and Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688.

Rubin DB and van der Laan MJ (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. The International Journal of Biostatistics, 4(1).

Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM, et al. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Controlled clinical trials, 25(1):119–142. [PubMed: 15061154]

Scharfstein DO, Rotnitzky A, and Robins JM (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448):1096–1120.

Schulte PJ, Tsiatis AA, Laber EB, and Davidian M. (2014). Q-and A-learning methods for estimating optimal dynamic treatment regimes. Statistical science, 29(4):640. [PubMed: 25620840]

Stefanski LA and Boos DD (2002). The calculus of M-estimation. The American Statistician, 56(1):29–38.

Tan Z. (2007). Comment: Understanding OR, PS and DR. Statistical Science, 22(4):560–568.

Tan Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika, 97(3):661–682.

Tsiatis A. (2007). Semiparametric theory and missing data. Springer Science & Business Media.

Tsiatis AA and Davidian M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science, 22(4):569. [PubMed: 18516239]

Tsiatis AA, Davidian M, and Cao W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. Biometrics, 67(2):536–545. [PubMed: 20731640]

Van der Laan MJ and Robins JM (2003). Unified methods for censored longitudinal data and causality. Springer Science & Business Media.

Vermeulen K. and Vansteelandt S. (2015). Bias-reduced doubly robust estimation. Journal of the American Statistical Association, 110(511):1024–1036.

Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, Slamon DJ, Murphy M, Novotny WF, Burchmore M, et al. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of her2-overexpressing metastatic breast cancer. Journal of Clinical Oncology, 20(3):719–726. [PubMed: 11821453]

Wallace MP and Moodie EE (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. Biometrics, 71(3):636–644. [PubMed: 25854539]

Zhang B, Tsiatis AA, Laber EB, and Davidian M. (2012). A robust method for estimating optimal treatment regimes. Biometrics, 68(4):1010–1018. [PubMed: 22550953]

Zhang B, Tsiatis AA, Laber EB, and Davidian M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika, 100(3):681–694.

Zhao Y-Q, Laber EB, Ning Y, Saha S, and Sands B. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. Journal of Machine Learning Research. In press.

Zhao Y-Q, Zeng D, Laber EB, and Kosorok MR (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. Journal of the American Statistical Association, 110(510):583–598. [PubMed: 26236062]

Zhao Y-Q, Zeng D, Rush AJ, and Kosorok MR (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107(499):1106–1118. [PubMed: 23630406]

Zhou X, Mayer-Hamblett N, Khan U, and Kosorok MR (2017). Residual weighted learning for estimating individualized treatment rules. Journal of the American Statistical Association, 112(517):169–187. [PubMed: 28943682]

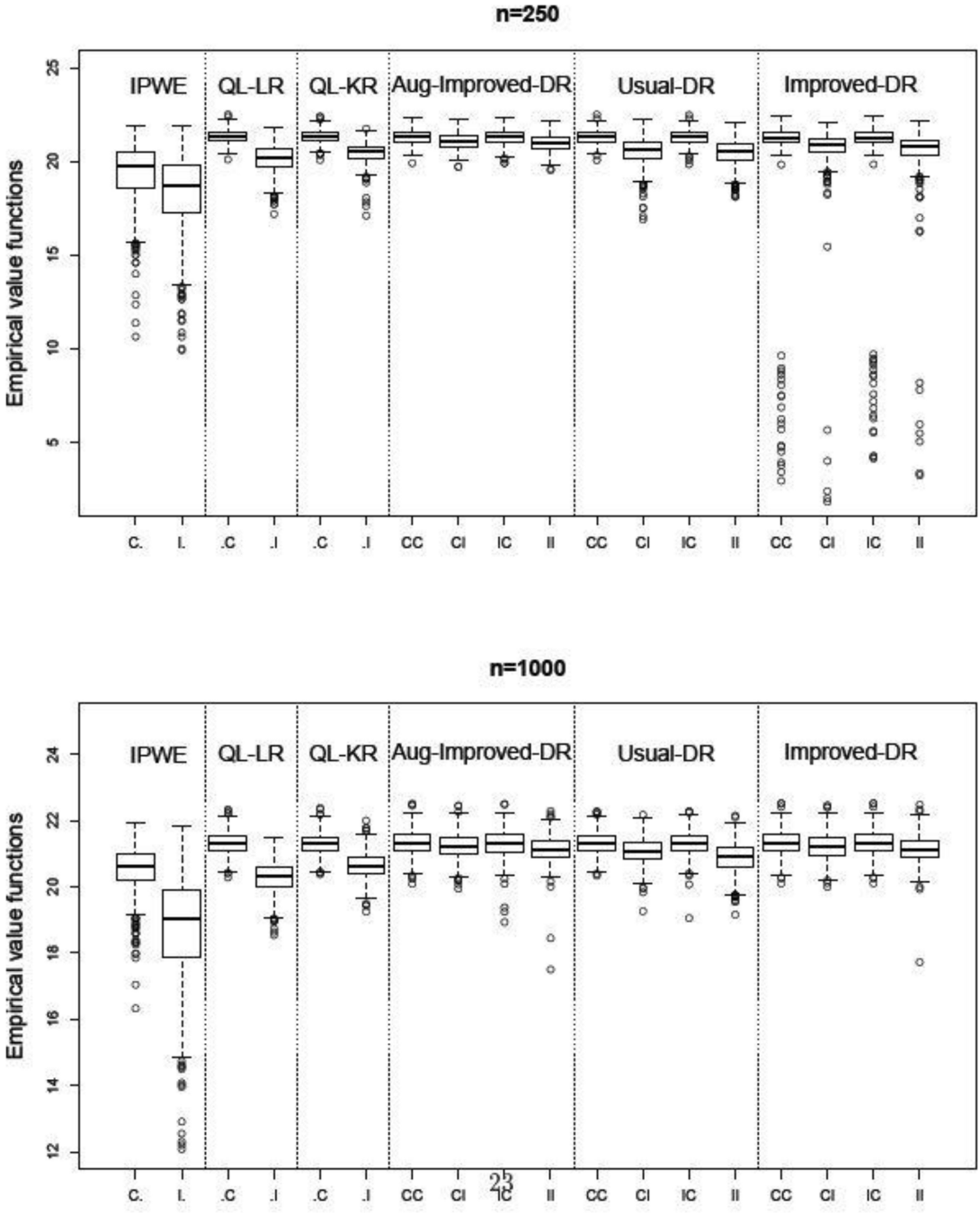

Fig. 1.

Simulation results for Scenario 1. Value functions over 500 replications. The optimal value is $E\{Y(d^{\text{opt}})\} = 21.32$.

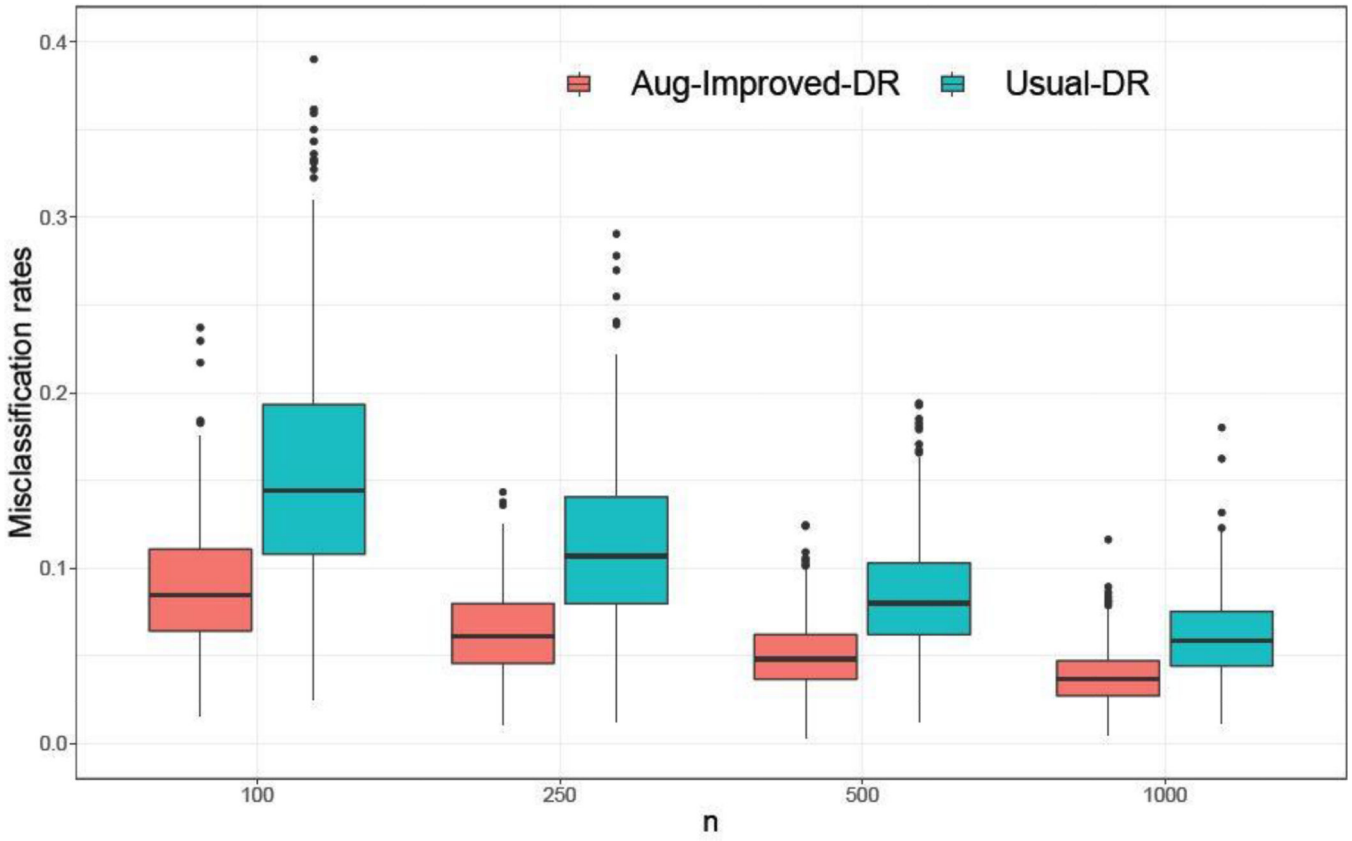

Fig. 2.

Simulation results for Scenario 1 under CI: propensity score correct, outcome model incorrect. Misclassification rates over 500 replications.

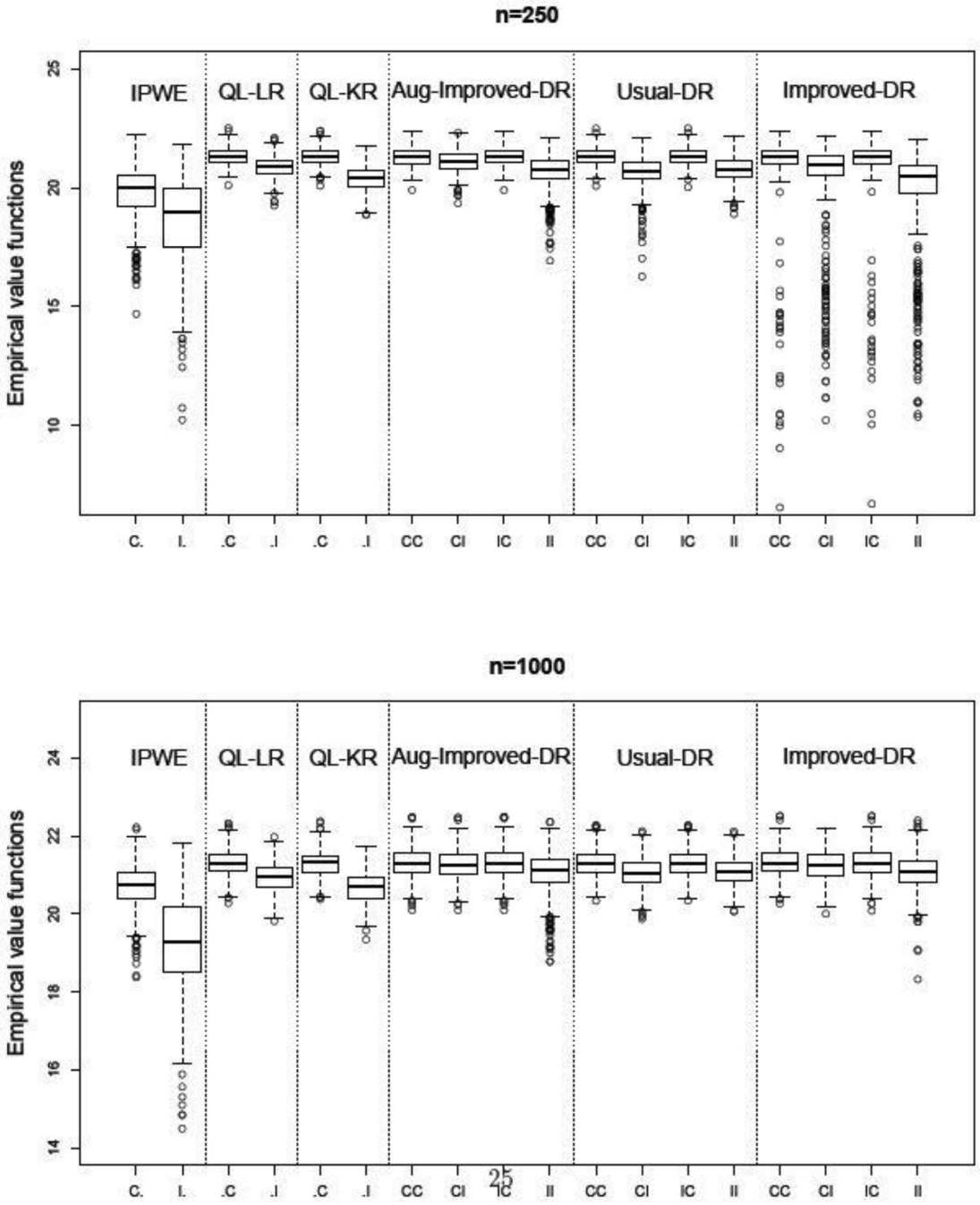

Fig. 3.

Simulation results for Scenario 2. Value functions over 500 replications. The optimal value is $E\{Y(d^{\text{opt}})\} = 21.32$.

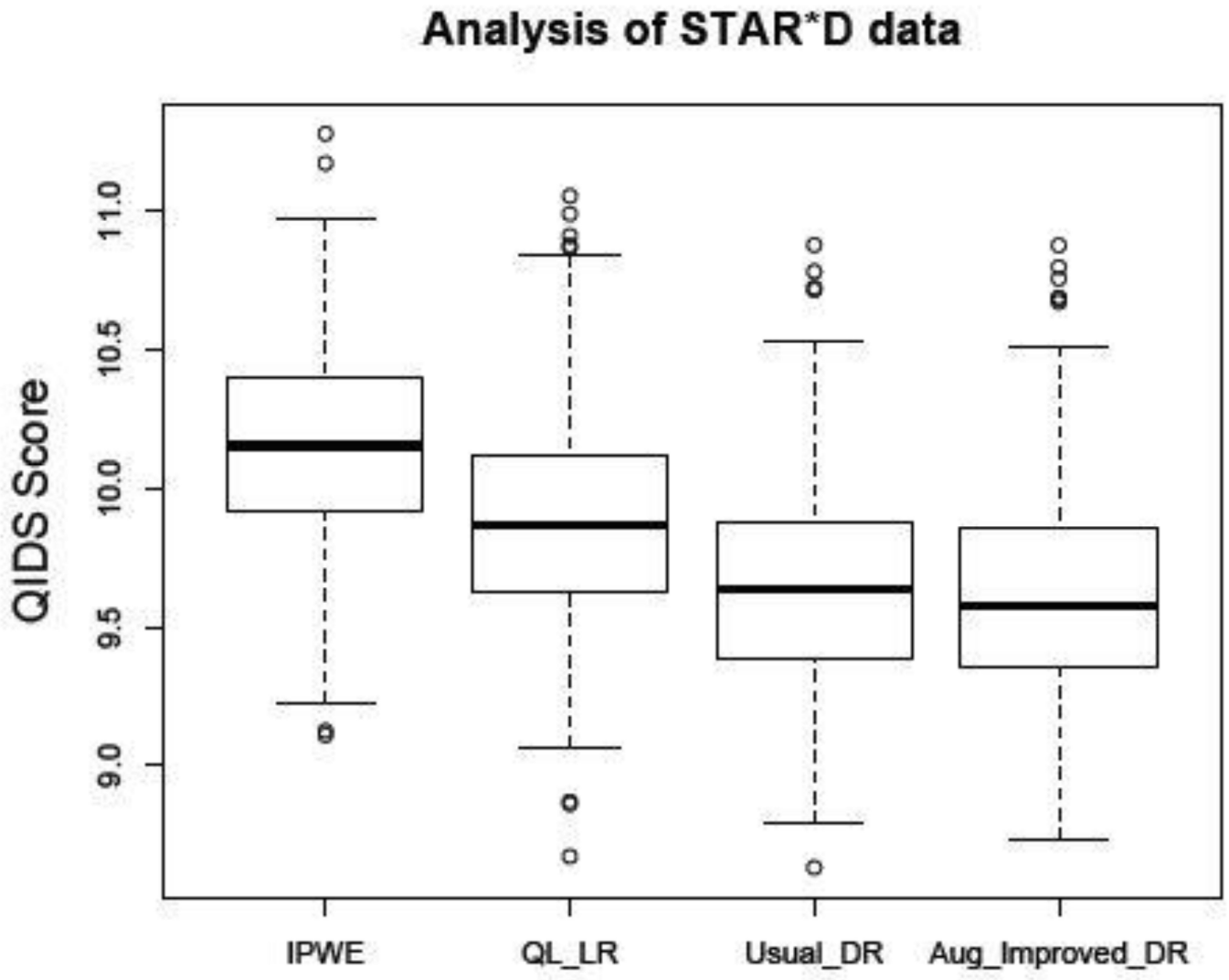

Fig. 4.
QIDS score based on 500 replications. Lower scores are more preferable.