Biometrika (2020), **103**, 1, *pp*. 1–17
Printed in Great Britain Advance Access publication on 31 July 2018

Targeted Optimal Treatment Regime Learning Using Summary Statistics

BY J. CHU, W. LU, S. YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive,
Campus Box 8203, Raleigh, North Carolina 27695, U.S.A.
jchu3@ncsu.edu, wlu4@ncsu.edu, syang24@ncsu.edu

SUMMARY

Personalized decision-making, aiming to derive optimal treatment regimes based on individual characteristics, has recently attracted increasing attention in many fields, such as medicine, social services, and economics. Current literature mainly focuses on estimating treatment regimes from a single source population. In real-world applications, the distribution of a target population can be different from that of the source population. Therefore, treatment regimes learned by existing methods may not generalize well to the target population. Due to privacy concerns and other practical issues, individual-level data from the target population is often not available, which makes treatment regime learning more challenging. We consider the problem of treatment regime estimation when the source and target populations may be heterogeneous, individual-level data is available from the source population, and only the summary information of covariates, such as moments, is accessible from the target population. We develop a weighting framework that tailors a treatment regime for a given target population by leveraging the available summary statistics. Specifically, we propose a calibrated augmented inverse probability weighted estimator of the value function for the target population and estimate an optimal treatment regime by maximizing this estimator within a class of pre-specified regimes. We show that the proposed calibrated estimator is consistent and asymptotically normal even with flexible semi/nonparametric models for nuisance function approximation, and the variance of the value estimator can be consistently estimated. We demonstrate the empirical performance of the proposed method using simulation studies and a real application to an eICU dataset as the source sample and a MIMIC-III dataset as the target sample.

Some key words: Covariate shift; Double robustness; Empirical likelihood; Entropy balancing; Multi-source policy learning.

1. INTRODUCTION

Personalized decision-making, a pseudo intelligence paradigm tailored to an individual's characteristics, has recently attracted a great deal of attention in many fields, such as precision medicine, social services, economics, and recommendation system. An individualized treatment rule (ITR) formalizes treatment decisions as a function mapping from patient information to a recommended treatment. An optimal ITR is the one that leads to the greatest expected outcome in the population of interest, known as the value function.

A variety of approaches have been developed for estimating optimal ITRs. One class of approaches is model-based as they directly model the conditional mean outcome given covariates and treatment, known as the Q-function, and then use the estimated Q-function to infer the optimal ITR. Such methods include Q-learning (Qian & Murphy, 2011) and its semiparametric extension, A-leaning (Murphy, 2003), where only the contrast function is modeled while the baseline mean function is completely unspecified. Alternatively, direct value search methods have been developed and extensively studied recently (e.g. Zhang et al., 2013; Luckett et al., 2020; Athey & Wager, 2021). These methods learn the optimal treatment regime by regime evaluation. They first establish a flexible estimator of the value function, such as the augmented inverse probability weighted (AIPW) estimator, and the optimal ITR is then estimated by maximizing the estimated value function within a class of pre-specified ITRs, such as linear decision rules and tree-based decision rules. The AIPW value estimator possesses the double robustness property, i.e., it is consistent for the value function if either the Q-function or the propensity score model is correctly specified.

Though the double robustness of the AIPW value estimator is appealing, it's only maintained when the source and target populations are identical. In other words, when there exists heterogeneity between the source and target populations, the AIPW value estimator obtained based on the source sample may no longer be consistent for the value function of the target population. Thus, the optimal ITR learned from the source data may not be optimal for the target population. In many real-world applications, the value function of an ITR over the distribution of the target population is of significant interest, which can be different from that of the source population. For example, in medical studies, it is known that the results of a randomized controlled trial cannot be directly transported because the covariate distribution in a target population may be different (Cole & Stuart, 2010). Due to study design and inclusion/exclusion criteria, the source sample can be unrepresentative of the target population we are interested in. When there is heterogeneity between the source and target populations, an estimated optimal ITR from the source sample may not generalize well to the target population (Lee et al., 2021). Such problems gain increasing attention in the ITR learning fields recently. Zhao et al. (2019) and Mo et al. (2021) proposed different collections of possible target covariate distributions and estimated the optimal ITR by optimizing the worst-case quality assessment among the collection. Uehara et al. (2020) considered a nonparametric estimator for the density ratio of the covariate distributions of the source and target populations and constructed a weighted estimator for the target value function based on the estimated density ratio. However, all these methods require the availability of individual-level data from both the source and target populations, which may be unrealistic in many applications. For example, while large-population based databases, such as the Surveillance, Epidemiology and End Results database, can provide reliable summary statistics for covariates, such as means and medians, and overall survival statistics for the disease population, critical information about individual factors that influence the choice of treatment and clinical outcomes of interest may not be available (Huang & Qin, 2020; Chen et al., 2021). Moreover, due to privacy and confidentiality concerns, comprehensive individual-level data is often prohibited to share with researchers. In contrast, summary statistics of patient characteristics of the target population are often available and can be easily shared for research purposes.

In this paper, we consider the targeted optimal treatment regime learning where we have individual-level data from the source sample but only a few summary statistics of covariate distributions from the target population. As we alluded to previously, when there is heterogeneity in covariate distributions between the source and target populations, the estimated optimal ITR obtained by maximizing the value estimator constructed based on the source sample may not be optimal for the target population. One way to address this issue is to assign different subject weights to the source sample and calibrate the source covariate distributions to the target covariate distributions. Calibration weighting is widely used to integrate auxiliary information in survey sampling and causal inference, such as empirical likelihood based methods (Qin &

Zhang, 2007), entropy-based covariate balancing methods (Hainmueller, 2012), and quadratic loss based covariate balancing methods (Zubizarreta, 2015). Such weighting methods allow adjusting covariate distributions of the source sample using various summary statistics of covariates in the target population, such as means, variances, correlations, and quantiles. We propose a calibrated AIPW estimator of the value function using summary statistics from the target population and then search for the optimal ITR for the target population by maximizing the calibrated AIPW value estimator over a pre-specified class of ITRs. Here, the subject weights for the source sample are estimated by solving a general convex optimization problem with constraints. The objective function in the optimization problem can be chosen from the Cressie-Read family (Cressie & Read, 1984), while the constraints force the weighted summary statistics of source covariates to be the same as that from the target population. We show that the calibrated AIPW estimator for the target value function is consistent, asymptotically normal, and has the double robustness property if the estimated weight function converges to the density ratio of covariate distributions between the two populations. The double robustness entails that the value estimator remains root-n consistent if any one of the two parametric models for the propensity score and outcome mean is correctly specified or if both models are estimated nonparametrically satisfying a certain rate condition for convergence. Interestingly, if the source and target populations have the same covariate distribution, the calibrated optimal value estimator gains efficiency over the uncalibrated one by utilizing additional summary information. However, in general, the weights learned from the calibration methods may not consistently estimate the density ratio. Under such general cases, the proposed calibrated AIPW estimator can still converge to the value function of a pseudo population that may be closer to the target distribution compared with the source population. As such, it can give a more accurate estimator for the value function of the target population than the uncalibrated value estimator, and the optimal ITR obtained by maximizing the calibrated AIPW value estimator can be more favorable for the target population.

2. Statistical Framework

2.1. *Value Function and Optimal ITR*

In a randomized trial or observational study, suppose there are two treatment options, labeled as control/treatment 0 and experimental treatment/treatment 1. Let A taking values 0 or 1 in accordance with the two options, denote the treatment received. Let $X \in \mathbb{R}^p$ be a vector of baseline covariates and Y be the observed outcome of interest. We assume larger values of Y are preferred by convention. The observed data are then $\{O_i = (Y_i, A_i, X_i), i = 1, \ldots, n\}$, which are independent and identically distributed. Define the potential outcomes $Y^*(0)$ and $Y^*(1)$ as the outcomes that would be observed if a subject received treatment 0 or 1, respectively. As is customary in causal inference (Rubin, 1978), we make the following assumptions.

Assumption 1. (A1) $Y = Y^*(1)A + Y^*(0)(1 - A)$, (A2) $\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A \mid X$, and (A3) $0 < \mathrm{pr}(A = 1 \mid X = x) < 1$ for all x such that $\mathrm{pr}(X = x) > 0$.

An ITR is a function $d(\cdot)$ that maps values of X to $\{0, 1\}$, so that a subject with covariates value $X = x$ would receive treatment 1 if $d(x) = 1$ and treatment 0 if $d(x) = 0$. For any arbitrary ITR $d(\cdot)$, we can define the potential outcome as $Y^*(d) = Y^*(1)d(X) + Y^*(0)\{1 - d(X)\}$, which would be observed if a randomly chosen individual had been assigned a treatment according to $d(\cdot)$, where we suppress the dependence of $Y^*(d)$ on X. We then define the value function under $d(\cdot)$ as the expectation of the potential outcome as $V(d) = E\{Y^*(d)\} = E\left[Y^*(1)d(X) + Y^*(0)\{1 - d(X)\}\right].$

Suppose $\mathcal{D}$ is a class of ITRs of interest. Then we define the optimal ITR as $d^{\text{opt}}(X) = \text{argmax}_{d \in \mathcal{D}} V(d)$. In clinical practice, it may be desirable to consider a class of ITRs indexed by a vector of parameters β for feasibility and interpretability. We denote such a class of rules as $\mathcal{D}_\beta$ and its element as $d(X;\beta)$. For example, we can consider a class of linear ITRs $\{d(X;\beta) = I(\beta^{\mathrm{T}}\tilde{X} > 0) : \beta \in \mathbb{R}^{p+1}, \|\beta\|_2 = 1\}$, where $\tilde{X} = (1, X^{\mathrm{T}})^{\mathrm{T}}$. Given a linear ITR $d(X;\beta)$, we use a shorthand to write its value function $V(d)$ as $V(\beta)$. Let $\beta^* = \text{argmax}_\beta V(\beta)$. Then, the optimal linear ITR is $d_\beta^{\text{opt}} = d(X;\beta^*)$. The true optimal ITR d^{opt} may not be in $\mathcal{D}_\beta$. Thus, d_β^{opt} may not be the same as d^{opt}. However, when attention focuses on the feasible class $\mathcal{D}_\beta$, estimation of d_β^{opt} is still of considerable interest. In this paper, we focus on linear ITRs.

2.2. *Source and Target Populations*

The difference between covariate distributions in the source and target populations is called a covariate shift (Sugiyama & Kawanabe, 2012). In this paper, we assume that there is a pooled population $\mathbb{P}$ consisting of both the source population $\mathbb{P}^{\mathrm{s}}$ and the target population $\mathbb{P}^{\mathrm{t}}$. Let S be a binary indicator for selection action: $S = 1$ if the individual comes from the source population and $S = 0$ if the individual comes from the target population. A covariate shift results from the situation where $\text{pr}(S = 1 \mid X) \neq \text{pr}(S = 0 \mid X)$.

For the source population $\mathbb{P}^{\mathrm{s}}$, we observe individual-level data $\{O_i = (Y_i, A_i, X_i), i = 1, \ldots, n\}$. For the target population $\mathbb{P}^{\mathrm{t}}$, only summary statistics of covariate distributions, such as mean, variance or quantiles are available. For $\mathbb{P}^{\mathrm{s}}$, denote the density or probability mass function of covariates as $f^{\mathrm{s}}(X)$ and its associated expectation as E; for $\mathbb{P}^{\mathrm{t}}$, we use the notation $f^{\mathrm{t}}(X)$ and E^{t} correspondingly. The summary statistics from the target population $\mathbb{P}^{\mathrm{t}}$ are denoted as $\mu_{g0} = E^{\mathrm{t}}\{g(X)\}$, where $g(X) = \{g_1(X), g_2(X), \ldots, g_q(X)\}^{\mathrm{T}}$ is a $q \times 1$ specified function. For example, a common choice is $g(X) = (X_1, X_2, \ldots, X_p)$, and μ_{g0} gives the mean of all covariates in the target population. We assume that summary statistics from the target population are derived from large databases so that their uncertainty are negligible. With only the summary statistics, targeted ITR learning is impossible without further assumptions in order to borrow information from the source sample.

Assumption 2. (A4) $E\{Y(a) \mid X\} = E^{\mathrm{t}}\{Y(a) \mid X\}$, and (A5) $\text{pr}(S = 1 \mid X) > 0$.

Assumption (A4) implies that true Q-functions are identical in both the source and target populations (Dahabreh et al., 2019). A stronger version of this assumption is the ignorability assumption that $\{Y(1), Y(0)\} \perp\!\!\!\perp S \mid X$ (Buchanan et al., 2018). Assumption (A5) implies that the support of the target X distribution must be covered by the support of the source X distribution.

3. PROPOSED METHOD

3.1. *Calibrated AIPW Estimator*

For the source population, define the propensity score as $\pi(X) = \text{pr}(A = 1 \mid X, S = 1)$ and conditional mean outcome model as $\mu(X, A) = E(Y \mid X, A)$. In practice, $\pi(\cdot)$ and $\mu(\cdot)$ can be estimated from the observed source data based on some posited parametric models $\pi(X;\eta)$ and $\mu(X, A;\theta)$, respectively. Alternatively, they can also be estimated nonparametrically, e.g. using kernel regression or random forest. Given an ITR $d(X;\beta)$, Zhang et al. (2012) proposed an AIPW estimator of the value function $V(\beta)$ as

$$\widehat{V}^{\mathrm{o}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I\{A_i = d(X_i;\beta)\}}{\varrho(A_i \mid X_i;\widehat{\eta})}\{Y_i - \mu_d(X_i;\beta,\widehat{\theta})\} + \mu_d(X_i;\beta,\widehat{\theta})\right],$$

where the superscript o is a shorthand for original, $\varrho(A_i \mid X_i; \widehat{\eta}) = \pi(X_i; \widehat{\eta})A_i + \{1 - \pi(X_i; \widehat{\eta})\}(1 - A_i)$, $\mu_d(X_i; \beta, \widehat{\theta}) = \mu(X_i, 1; \widehat{\theta})I\{d(X_i; \beta) = 1\} + \mu(X_i, 0; \widehat{\theta})I\{d(X_i; \beta) = 0\}$, and $\widehat{\eta}$ and $\widehat{\theta}$ are the estimates of η and θ, respectively, based on some posited parametric models.

Denote the value function of the target population under the ITR $d(X; \beta)$ as $V^{\mathrm{t}}(\beta)$. If the source and target populations have the same covariate distributions, i.e. $f^{\mathrm{s}}(X) = f^{\mathrm{t}}(X)$, then $V^{\mathrm{t}}(\beta)$ can be consistently estimated by $\widehat{V}^{\mathrm{o}}(\beta)$ based on the source sample. However, since covariate distributions between the source and target populations often differ in practice, $\widehat{V}^{\mathrm{o}}(\beta)$ may be biased for $V^{\mathrm{t}}(\beta)$. To reduce the bias, a natural approach is to consider calibration weights, i.e., to assign different weights to individual data points in the source sample so that the weighted data is more representative of the target distribution. Specifically, we consider the following calibrated AIPW estimator

$$\widehat{V}^{\mathrm{c}}(\beta) = \sum_{i=1}^{n} w_i \left[\frac{I\{A_i = d(X_i; \beta)\}}{\varrho(A_i \mid X_i; \widehat{\eta})} \{Y_i - \mu_d(X_i; \beta, \widehat{\theta})\} + \mu_d(X_i; \beta, \widehat{\theta}) \right],$$

where the superscript c is a shorthand for calibrated, w_i's are calibration weights satisfying $\sum_{i=1}^{n} w_i = 1$ and other constraints. Using the summary statistics μ_{g0} from the target population, we can utilize methods, such as empirical likelihood (Qin & Zhang, 2007) and entropy balancing (Hainmueller, 2012; Zhao & Percival, 2017), to learn the weights. In the next section, we propose a general framework to estimate the weights. 170

3.2. *Weights Estimation*

Let $h\,(w)$ denote a generic convex distance function between a scalar w and n^{-1}. We consider the following optimization problem $\min_{w_1,\ldots,w_n} \sum_{i=1}^{n} h\,(w_i)$ under the constraints $\sum_{i=1}^{n} w_i\{g(X_i) - \mu_{g0}\} = 0$, $\sum_{i=1}^{n} w_i = 1$, and $w_i \geq 0$. 175

In the considered optimization problem, the function $h(w)$ plays a role in quantifying the discrepancy of calibration weights and the uniform distribution n^{-1}. We choose the function $h(w)$ from the Cressie-Read family of discrepancies (Cressie & Read, 1984). The Cressie-Read family is defined through a class of additive convex functions that encompasses a broad family of distance functions. Specifically, 180

$$CR(\gamma) = \sum_{i=1}^{n} h\,(w_i) = \sum_{i=1}^{n} \{\gamma(\gamma + 1)\}^{-1}\{(nw_i)^{\gamma+1} - 1\}.$$

Three special cases with $\gamma \in \{-1, 0, 1\}$ are popular. In particular, $CR(-1) = \sum_{i=1}^{n} -\log(nw_i)$ and $CR(0) = \sum_{i=1}^{n}(nw_i)\log(nw_i)$. Minimizing $CR(-1)$ is equivalent to maximizing $\sum_{i=1}^{n} \log(w_i)$, leading to the maximum empirical log-likelihood objective function. Minimizing $CR(0)$ is equivalent to maximizing $-\sum_{i=1}^{n} w_i \log(w_i)$, leading to the maximum empirical exponential likelihood or entropy. Finally, minimizing $CR(1)$ is equivalent to minimizing the sum of squares $\sum_{i=1}^{n}(w_i - n^{-1})^2$. To be consistent with the existing literature, we call the weight estimation method as the empirical likelihood method for $\gamma = -1$ and the entropy balancing method for $\gamma = 0$. We summarize the correspondence between γ and the form of $h(w)$ in Table 1. 185 190

The first constraint is referred to as the balancing constraint, which calibrates the covariate distribution of the source sample to the target population in terms of $g(X)$. As a common premise to solve the above optimization problem, μ_{g0} should fall within the convex hull of $\{g(X_i), i = 1, ..., n\}$. Then, the optimization problem can be solved using the method of La- 195

Table 1. *The formulation of $\rho(x)$ for the empirical likelihood, entropy balancing, and Cressie-Read family*

Method	Empirical Likelihood	Entropy Balancing	Cressie-Read
γ	-1	0	γ
$h(w)$	$-\ln(nw)$	$nw\ln(nw)$	$\frac{(nw)^{\gamma+1}-1}{\gamma(\gamma+1)}$
$\rho(x)$	$(1-x)^{-1}$	$\exp(x)$	$(1+\gamma x)^{1/\gamma}$

grangian multipliers with the loss function

$$L = \{\gamma(\gamma+1)\}^{-1}\sum_{i=1}^{n}\{(nw_i)^{\gamma+1}-1\} - n\lambda^{\mathrm{T}}\sum_{i=1}^{n} w_i\left\{g(X_i)-\mu_{g0}\right\} + n\varphi\left(1-\sum_{i=1}^{n} w_i\right). \tag{1}$$

As noted in Newey & Smith (2004), by minimizing (1), the estimator for w_i is

$$w(X_i;\widehat{\lambda}) = \rho\left[\widehat{\lambda}^{\mathrm{T}}\{g(X_i)-\mu_{g0}\}\right]\Big/\sum_{j=1}^{n}\rho\left[\widehat{\lambda}^{\mathrm{T}}\{g(X_j)-\mu_{g0}\}\right], \tag{2}$$

where the function $\rho(x)$ for different γ values are summarized in Table 1, and $\widehat{\lambda}$ solves the equation $\sum_{i=1}^{n}\rho\left[\lambda^{\mathrm{T}}\{g(X_i)-\mu_{g0}\}\right]\{g(X_i)-\mu_{g0}\}=0$.

Let $W(X;\lambda)=nw(X;\lambda)$. The proposed calibrated AIPW estimator is then

$$\widehat{V}^{\mathrm{c}}(\beta) = \frac{1}{n}\sum_{i=1}^{n} W(X_i;\widehat{\lambda})\left[\frac{I\{A_i=d(X_i;\beta)\}}{\varrho(A_i\mid X_i;\widehat{\eta})}\{Y_i-\mu_d(X_i;\beta,\widehat{\theta})\}+\mu_d(X_i;\beta,\widehat{\theta})\right].$$

Thus, the regime learning procedure can be summarized as a 3-step algorithm:

Step 1. Estimate calibration weights, e.g., using the empirical likelihood method or entropy balancing method.

Step 2. Estimate the propensity score $\pi(\cdot)$ and the conditional outcome mean $\mu(\cdot)$ using either parametric models or nonparametric models.

Step 3. Construct the calibrated AIPW estimator with the components estimated in Steps 1 and 2, and obtain the optimal ITR by maximizing the calibrated AIPW estimator within a class of pre-specified ITRs, such as linear decision rules.

Before delving into theoretical analysis, it is important to define the underlying population for which $\widehat{V}^{\mathrm{c}}(\beta)$ is targeting unambiguously. Toward this end, let λ^{*} be the limit of $\widehat{\lambda}$ and

$$W^{*}(X;\lambda) = \rho\left[\lambda^{\mathrm{T}}\{g(X)-\mu_{g0}\}\right]\big/E\left(\rho\left[\lambda^{\mathrm{T}}\{g(X)-\mu_{g0}\}\right]\right).$$

In general, the calibration weights are not guaranteed to be non-negative. As pointed out in Schennach (2007), when $\gamma \le 0$, the estimated weights are non-negative by construction. It can be shown that $f^{+}(X) \propto f^{\mathrm{s}}(X)W^{*}(X;\lambda^{*})$ is a valid density or probability mass function when $\gamma = -1$ or 0. Therefore, it defines a pseudo population $\mathbb{P}^{+}$. While for $\gamma > 0$, the calibration weights can take on negative values, and thus the corresponding $f^{+}(X)$ is not always a valid density or probability mass function. Therefore, we focus on $\gamma = -1, 0$ for illustration. It is expected that $\widehat{V}^{\mathrm{c}}(\beta)$ will converge to the value function under the ITR $d(X;\beta)$ for the pseudo population $\mathbb{P}^{+}$, when either the propensity score or the conditional mean outcome model is correctly specified.

Moreover, when $W(X;\lambda^*) \propto f^{\mathrm{t}}(X)/f^{\mathrm{s}}(X)$, we have $f^+(X) = f^{\mathrm{t}}(X)$. Then, $\widehat{V}^{\mathrm{c}}(\beta)$ is also a consistent estimator of the value function for the target population. Denote the density or probability mass function of covariates in the pooled population $\mathbb{P}$ as $q(X)$. Then $f^{\mathrm{s}}(X)$ and $f^{\mathrm{t}}(X)$ can be described as $q(X \mid S=1)$ and $q(X \mid S=0)$, respectively. By Bayesian Theorem, we have, 220

$$\frac{f^{\mathrm{t}}(X)}{f^{\mathrm{s}}(X)} = \frac{q(X \mid S=0)}{q(X \mid S=1)} \propto \frac{\mathrm{pr}(S=0 \mid X)}{\mathrm{pr}(S=1 \mid X)}.$$

If $\mathrm{pr}(S=0 \mid X)$ follows a logistic regression with covariates $g(X)$, we have $\mathrm{pr}(S=0 \mid X)/\mathrm{pr}(S=1 \mid X) \propto \exp\{\alpha^{\mathrm{T}} g(X)\}$. Moreover, based on (2), we have $W(X;\widehat{\lambda}) \propto \exp\{\widehat{\lambda}^{\mathrm{T}} g(X)\}$ when $\gamma = 0$. Therefore, the weights obtained by the entropy balancing method satisfy $W(X;\lambda^*) \propto f^{\mathrm{t}}(X)/f^{\mathrm{s}}(X)$ under the logistic regression model for $\mathrm{pr}(S=0 \mid X)$. Similarly, if $\mathrm{pr}(S=0 \mid X)$ can be represented by the following model

$$\mathrm{pr}(S=0 \mid X) = \frac{\kappa_0}{1-\alpha^{\mathrm{T}}\{g(X)-\mu_{g0}\}} \Bigg/ \left[1 + \frac{\kappa_0}{1-\alpha^{\mathrm{T}}\{g(X)-\mu_{g0}\}}\right],$$

where κ_0 is a positive constant and α satisfies $1-\alpha^{\mathrm{T}}\{g(X)-\mu_{g0}\} > 0$, we have $\mathrm{pr}(S=0 \mid X)/\mathrm{pr}(S=1 \mid X) \propto [1-\alpha^{\mathrm{T}}\{g(X)-\mu_{g0}\}]^{-1}$. Therefore, under the above model, the weights obtained by the empirical likelihood method satisfy $W(X;\lambda^*) \propto f^{\mathrm{t}}(X)/f^{\mathrm{s}}(X)$. 225

In general, the calibration weights can not lead to a pseudo population with exactly the same covariate distribution as for the target population. However, it is expected that with more constraints based on summary statistics from the target population, the covariate distribution of the pseudo population will get closer to that of the target population. Therefore, the optimal ITR obtained based on $\widehat{V}^{\mathrm{c}}(\beta)$ would be better than that obtained based on $\widehat{V}^{\mathrm{o}}(\beta)$. Let $V^+(\beta)$ denote the value function under the ITR $d(X;\beta)$ for the pseudo population $\mathbb{P}^+$ and define $\beta^* = \mathrm{argmax}_\beta V^+(\beta)$. Then, the true optimal linear ITR for $\mathbb{P}^+$ is $d(X;\beta^*)$ and the estimated optimal linear ITR is $d(X;\widehat{\beta}^{\mathrm{c}})$, where $\widehat{\beta}^{\mathrm{c}} = \mathrm{argmax}_\beta \widehat{V}^{\mathrm{c}}(\beta)$. Similarly, define $\widehat{\beta}^{\mathrm{o}} = \mathrm{argmax}_\beta \widehat{V}^{\mathrm{o}}(\beta)$. The estimated optimal linear ITR without calibration is $d(X;\widehat{\beta}^{\mathrm{o}})$. 230 235

4. Theoretical Properties

In this section, we establish the asymptotic properties of the calibrated AIPW estimator $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$. The proofs of all theorems are given in the supplementary material.

We first consider the case when the propensity score model $\pi(x)$ and conditional mean outcome model $\mu(x,a)$ are estimated based on some posited parametric models $\pi(X;\eta)$ and $\mu(X,A;\theta)$, respectively. Denote the estimating equations for λ, θ, η and $V^+(\beta^*)$ as 240

$$\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \rho\left[\lambda^{\mathrm{T}}\{g(X_i)-\mu_{g0}\}\right]\{g(X_i)-\mu_{g0}\} \\ C(X_i,A_i,Y_i;\theta) \\ S(X_i,A_i;\eta) \\ W(X_i;\lambda)\psi(X_i,A_i,Y_i;\beta,\theta,\eta) - V^+(\beta^*) \end{pmatrix} = 0,$$

where

$$\psi(X_i,A_i,Y_i;\beta,\theta,\eta) = \frac{I\{A_i = d(X_i;\beta)\}}{\varrho(A_i \mid X_i;\eta)}\{Y_i - \mu_d(X_i;\beta,\theta)\} + \mu_d(X_i;\beta,\theta).$$

Let $\widehat{\lambda}$, $\widehat{\theta}$ and $\widehat{\eta}$ denote the estimators of λ, η and θ obtained from the above equations and let λ^*, θ^* and η^* denote the limits of $\widehat{\lambda}$, $\widehat{\theta}$ and $\widehat{\eta}$, respectively. To establish the asymptotic properties of $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$, we impose the following regularity conditions.

Assumption 3. Assume the following regularity conditions hold: (A6) The supports of X and Y are bounded. (A7) The function $\mu(x,a)$ is smooth and bounded for all (x,a). (A8) The weight function $W(x;\lambda)$ is smooth and bounded away from ∞, and it has bounded first derivatives with respect to λ. (A9) The value function $V^+(\beta)$ is twice continuously differentiable in a neighborhood of β^*. (A10) There exist some constants $\delta_0 > 0$ such that $\mathrm{pr}(|\tilde{X}^{\mathrm{T}}\beta^*| \leq \delta) = O(\delta)$, where the big-O term is uniform in $0 < \delta \leq \delta_0$. (A11) (i) $\sqrt{n}(\widehat{\lambda} - \lambda^*) = O_p(1)$, (ii) $\sqrt{n}(\widehat{\theta} - \theta^*) = O_p(1)$, and (iii) $\sqrt{n}(\widehat{\eta} - \eta^*) = O_p(1)$.

Conditions (A6) - (A9) are standard regularity conditions used to establish the uniform convergence results. Condition (A10) excludes the situation with $\mathrm{pr}(\tilde{X}^{\mathrm{T}}\beta^* = 0) > 0$ and ensures the true targeted optimal ITR is uniquely defined, known as the margin condition, which is often assumed to derive a sharp convergence rate for the value function under the estimated optimal ITR (e.g. Luedtke & Van Der Laan, 2016). Condition (A11) assumes the $\sqrt{n}$-convergence rates of parameter estimates in the calibration weight function, propensity score model, and conditional mean outcome model, which usually hold under mild conditions for posited parametric models, for example, a logistic or probit regression model for the propensity score, a linear model for the conditional mean outcome, and weights obtained using the empirical likelihood method or entropy balancing method.

Define

$$\xi_{i1} = W(X_i;\lambda^*)\psi(Y_i,A_i,X_i;\beta^*,\theta^*,\eta^*) - V^+(\beta^*), \qquad \xi_{i3} = H_\theta^{\mathrm{T}} G_\theta^{-1} C(X_i,A_i,Y_i;\theta^*),$$
$$\xi_{i2} = H_\lambda^{\mathrm{T}} G_\lambda^{-1} \rho\left[(\lambda^*)^{\mathrm{T}}\{g(X_i)-\mu_{g0}\}\right]\{g(X_i)-\mu_{g0}\}, \quad \xi_{i4} = H_\eta^{\mathrm{T}} G_\eta^{-1} S(X_i,A_i;\eta^*),$$

where

$$\begin{aligned}
H_\lambda &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial W(X_i;\lambda^*)}{\partial\lambda}\right\}\psi(Y_i,A_i,X_i;\beta^*,\theta^*,\eta^*),\\
H_s &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} W(X_i;\lambda^*)\frac{\partial\psi(Y_i,A_i,X_i;\beta^*,\theta^*,\eta^*)}{\partial s} \ \ (s=\theta,\eta),\\
G_\lambda &= -\mathbb{E}\left(\rho'\left[(\lambda^*)^{\mathrm{T}}\{g(X)-\mu_{g0}\}\right]\{g(X)-\mu_{g0}\}\{g(X)-\mu_{g0}\}^{\mathrm{T}}\right),\\
G_\theta &= -\mathbb{E}\left\{\partial C(X,A,Y;\theta^*)/\partial\theta^{\mathrm{T}}\right\}, G_\eta = -\mathbb{E}\left\{\partial S(X,A;\eta^*)/\partial\eta^{\mathrm{T}}\right\}.
\end{aligned}$$

Note that ξ_{i2}, ξ_{i3} and ξ_{i4} are the terms in the inference function of $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$ due to estimators $\widehat{\lambda}$, $\widehat{\theta}$ and $\widehat{\eta}$, respectively.

THEOREM 1. *Assume either $\pi(X;\eta)$ or $\mu(X,A;\theta)$ is correctly specified. Under (A1)-(A11), we have, as $n\to\infty$, $\sqrt{n}\{\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}}) - V^+(\beta^*)\} \longrightarrow N(0,\sigma_1^2)$, in distribution, where $\sigma_1^2 = E\left\{(\xi_{i1}+\xi_{i2}+\xi_{i3}++\xi_{i4})^2\right\}$. In addition, σ_1^2 can be estimated by replacing expectation with empirical sum and true values $V^+(\beta^*)$, λ^*, θ^*, and η^* with $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$, $\widehat{\lambda}$, $\widehat{\theta}$, and $\widehat{\eta}$, respectively.*

Next, we consider the case when both propensity score model $\pi(x)$ and conditional mean outcome model $\mu(x,a)$ are estimated by flexible semi/nonparametric models with certain convergence rates. For example, $\pi(x)$ and/or $\mu(x,a)$ are estimated using kernel regression or random forest. Let $\widehat{\pi}(x)$ and $\widehat{\mu}(x,a)$ denote the corresponding estimators. The calibrated AIPW estimator $\widehat{V}^{\mathrm{c}}(\beta)$ can be similarly defined by replacing $\pi(x;\widehat{\eta})$ and $\mu(x,a;\widehat{\theta})$ with $\widehat{\pi}(x)$ and $\widehat{\mu}(x,a)$, respectively. To derive the asymptotic distribution of $\widehat{V}^{\mathrm{c}}(\beta)$, we need the following modified condition.

(A11') (i) $\sqrt{n}(\widehat{\lambda}-\lambda^*)=O_p(1)$; (ii) $\left[P\{\widehat{\pi}(X)-\pi(X)\}^2\right]^{\frac{1}{2}}\sum_{a=0}^{1}\left[P\{\widehat{\mu}(X,a)-\mu(X,a)\}^2\right]^{\frac{1}{2}}=o_p(n^{-1/2})$, where $P\{f(X)\}=\int f(x)^2 dF_X(x)$.

Condition (A11') (ii) is commonly imposed in the causal inference literature to derive the asymptotic distribution of the AIPW estimators when the nuisance functions are estimated with certain convergence rates (Kennedy, 2016; Farrell et al., 2021). For example, if $\pi(x)$ is estimated based on a correctly specified parametric model, $\widehat{\pi}(x)$ is $\sqrt{n}$-consistent. Then it only requires $\widehat{\mu}(x,a)$ to be consistent for (A11') to hold. This can be easily achieved by most nonparametric regression methods. However, when both $\mu(x,a)$ and $\pi(x)$ are estimated nonparametrically, it usually requires both terms to be estimated with the rate $o_p(n^{-1/4})$. This can be achieved by some nonparametric methods, such as kernel regression or random forest under certain conditions. With Condition (A11') (ii), we can establish the $\sqrt{n}$-consistency of $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$. In addition, the asymptotic variance of $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$ will not depend on the variances of estimates $\widehat{\pi}(x)$ and $\widehat{\mu}(x,a)$. The results are summarized in the following theorem.

THEOREM 2. *Under (A1)-(A10) and (A11'), we have, as* $n\to\infty$, $\sqrt{n}\{\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})-V^{+}(\beta^*)\}\longrightarrow N(0,\sigma_2^2)$, *in distribution, where* $\sigma_2^2=E\left\{(\xi_{i1}+\xi_{i2})^2\right\}$. *Here,* ξ_{i1} *and* ξ_{i2} *are defined the same as in Theorem 1 but replacing* $\pi(x;\eta)$ *and* $\mu(x,a;\theta)$ *with* $\pi(x)$ *and* $\mu(x,a)$, *respectively. In addition,* σ_2^2 *can be estimated by replacing expectation with empirical sum and true values* $V^{+}(\beta^*)$, λ^*, $\pi(x)$ *and* $\mu(x,a)$ *with* $\widehat{V}^{\mathrm{c}}(\widehat{\beta}^{\mathrm{c}})$, $\widehat{\lambda}$, $\widehat{\pi}(x)$ *and* $\widehat{\mu}(x,a)$, *respectively.*

Remark 1. The theorems established above focus on the inference for the optimal value function. In the proof of Theorems 1 and 2, we show that $\widehat{\beta}^{\mathrm{c}}$ has the cubic root convergence rate. In addition, the asymptotic distribution of $\widehat{\beta}^{\mathrm{c}}$ can be established and its associated inference can be done by bootstrap-based methods (e.g. Cattaneo et al., 2020).

Finally, we compare the efficiency of $\widehat{V}^{\mathrm{o}}(\beta)$ and $\widehat{V}^{\mathrm{c}}(\beta)$ when the source and target populations have the same covariate distributions, i.e. $\mathbb{P}^{\mathrm{s}}=\mathbb{P}^{\mathrm{t}}$. Under such case, $V^{+}(\beta)=V^{\mathrm{t}}(\beta)$, the value function under the ITR $d(X;\beta)$ for the target population.

THEOREM 3. *Assume (A1)-(A10) and (A11') hold. When* $\mathbb{P}^{\mathrm{s}}=\mathbb{P}^{\mathrm{t}}$, *we have that both* $\sqrt{n}\{\widehat{V}^{\mathrm{o}}(\beta)-V^{\mathrm{t}}(\beta)\}$ *and* $\sqrt{n}\{\widehat{V}^{\mathrm{c}}(\beta)-V^{\mathrm{t}}(\beta)\}$ *are asymptotically normal with mean zero, while the latter one has the same or smaller asymptotic variance.*

Theorem 3 implies that even when the source and target populations have the same covariate distributions, the calibrated AIPW value estimator can be more efficient than the original AIPW value estimator without calibration. The efficiency gain of the calibrated estimator comes from the constraints imposed based on available summary statistics of the covariate distribution for the target population.

5. SIMULATION STUDIES

We have carried out extensive simulation studies to evaluate the performance of the proposed methods. Here we focus on two methods for computing the weights: empirical likelihood ($\gamma=-1$) and entropy balancing ($\gamma=0$). The results for $\gamma=1$ are provided in the supplementary material. For illustration, we only considered means of all covariates as the summary statistics from the target population. The corresponding pseudo populations are denoted as $\mathbb{P}_{EB}^{+}$ for $\gamma=0$ and $\mathbb{P}_{EL}^{+}$ for $\gamma=-1$, respectively. Table 2 defines additional notation for the simulation. Since

Table 2. *Additional notation used in the simulation studies*

Population	Value	Optimal ITR	Estimators
$\mathbb{P}^{\mathrm{t}}$	$V^{\mathrm{t}}(\beta)$	$d(X;\beta^{\mathrm{t}})$; $\beta^{\mathrm{t}} = \mathrm{argmax}_\beta\, V^{\mathrm{t}}(\beta)$	
$\mathbb{P}^{+}_{EB}$	$V^{+}_{EB}(\beta)$	$d(X;\beta^{*}_{EB})$; $\beta^{*}_{EB} = \mathrm{argmax}_\beta\, V^{+}_{EB}(\beta)$	$\widehat{V}^{\mathrm{c}}_{EB}(\beta)$; $\widehat{\beta}^{\mathrm{c}}_{EB} = \mathrm{argmax}_\beta\, \widehat{V}^{\mathrm{c}}_{EB}(\beta)$
$\mathbb{P}^{+}_{EL}$	$V^{+}_{EL}(\beta)$	$d(X;\beta^{*}_{EL})$; $\beta^{*}_{EL} = \mathrm{argmax}_\beta\, V^{+}_{EL}(\beta)$	$\widehat{V}^{\mathrm{c}}_{EL}(\beta)$; $\widehat{\beta}^{\mathrm{c}}_{EL} = \mathrm{argmax}_\beta\, \widehat{V}^{\mathrm{c}}_{EL}(\beta)$

estimated value functions are non-smooth and non-convex in β, following Zhang et al. (2012), we used the genetic algorithm (Whitley, 1994) to find $\widehat{\beta}^{\mathrm{o}}$, $\widehat{\beta}^{\mathrm{c}}_{EB}$, and $\widehat{\beta}^{\mathrm{c}}_{EL}$. The optimization was implemented using the function `genoud` in the R package `rgenoud` (Mebane Jr & Sekhon, 2011).

For the source sample, outcomes are generated from the model $Y = \mu(X, A) + \epsilon$, where $X = (X_1, X_2, X_3)^{\mathrm{T}}$,

$$\mu(X, A) = \exp\left\{2 - 0.1X_1 - 0.2X_2 + 0.2X_3 + A\frac{2\mathrm{sign}(X_3 - X_2^2 + 1)}{2 + |X_3 - X_2^2 + 1|}\right\},$$

and ϵ is generated from a normal distribution with mean 0 and variance 0.25. In addition, we considered two different propensity score models for the treatment indicator A: $\pi(X) = 0.5$, which represents a randomization study; $\mathrm{logit}\{\pi(X)\} = 0.5X_1 - 0.5X_2 + 0.5X_3$, which represents an observational study.

We considered four different scenarios of the covariate distributions for $\mathbb{P}^{\mathrm{s}}$ and $\mathbb{P}^{\mathrm{t}}$, which are summarized in Table 3 and Table 4. In Scenario 1, the covariate distributions of the source and target populations are the same. We have $\mathbb{P}^{\mathrm{s}} = \mathbb{P}^{+}_{EB} = \mathbb{P}^{+}_{EL} = \mathbb{P}^{\mathrm{t}}$. In Scenario 2, the ratio $f^{\mathrm{t}}(X)/f^{\mathrm{s}}(X)$ can be written as $\exp\{\ln(0.4) + \ln(4)X_1\}$ or $1/\{1 - 1.875(X_1 - 0.8)\}$. It can be shown that $W(X;\lambda^*) \propto f^{\mathrm{t}}(X)/f^{\mathrm{s}}(X)$ for both calibration methods. Therefore, we have $\mathbb{P}^{+}_{EB} = \mathbb{P}^{+}_{EL} = \mathbb{P}^{\mathrm{t}}$ even if we only use means of covariates as the summary statistics from the target population. This implies $V^{+}_{EB}(\beta) = V^{+}_{EL}(\beta) = V^{\mathrm{t}}(\beta)$, and both $\widehat{V}^{\mathrm{c}}_{EB}(\widehat{\beta}^{\mathrm{c}}_{EB})$ and $\widehat{V}^{\mathrm{c}}_{EL}(\widehat{\beta}^{\mathrm{c}}_{EL})$ are consistent estimators of $V^{\mathrm{t}}(\beta^{\mathrm{t}})$ when either the propensity score or conditional mean outcome model is correctly specified. However, in Scenarios 3 and 4, $W(X;\lambda^*)$ is no longer proportional to $f^{\mathrm{t}}(X)/f^{\mathrm{s}}(X)$. Thus, $\widehat{V}^{\mathrm{c}}_{EB}(\widehat{\beta}^{\mathrm{c}}_{EB})$ and $\widehat{V}^{\mathrm{c}}_{EL}(\widehat{\beta}^{\mathrm{c}}_{EL})$ are doubly robust estimators only for the value functions of their corresponding pseudo populations, but not for that of the target population.

Table 3. *Covariate distributions for* $\mathbb{P}^{\mathrm{s}}$ *and* $\mathbb{P}^{\mathrm{t}}$ *used in the simulation studies*

Scenario	$f^{\mathrm{s}}(X)$	$f^{\mathrm{t}}(X)$
1	$X_1 \sim Bernoulli(0.5)$ $(X_2, X_3)^{\mathrm{T}} \sim N((-1, 0)^{\mathrm{T}}, \Sigma_1)$	$X_1 \sim Bernoulli(0.5)$ $(X_2, X_3)^{\mathrm{T}} \sim N((-1, 0)^{\mathrm{T}}, \Sigma_1)$
2	$X_1 \sim Bernoulli(0.5)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 1 \sim N((1, -1)^{\mathrm{T}}, \Sigma_1)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 0 \sim N((-1, 1)^{\mathrm{T}}, \Sigma_2)$	$X_1 \sim Bernoulli(0.8)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 1 \sim N((1, -1)^{\mathrm{T}}, \Sigma_1)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 0 \sim N((-1, 1)^{\mathrm{T}}, \Sigma_2)$
3	$X_1 \sim Bernoulli(0.7)$ $(X_2, X_3)^{\mathrm{T}} \sim N((0.1, -0.2)^{\mathrm{T}}, \Sigma_1)$	$X_1 \sim Bernoulli(0.8)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 1 \sim N((1, -1)^{\mathrm{T}}, \Sigma_1)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 0 \sim N((-1, 1)^{\mathrm{T}}, \Sigma_2)$
4	$X_1 \sim Bernoulli(0.6)$ $(X_2, X_3)^{\mathrm{T}} \sim N((0, 0)^{\mathrm{T}}, \Sigma_1)$	$X_1 \sim Bernoulli(0.8)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 1 \sim N((1, -1)^{\mathrm{T}}, \Sigma_1)$ $(X_2, X_3)^{\mathrm{T}} \mid X_1 = 0 \sim N((-1, 1)^{\mathrm{T}}, \Sigma_2)$

$$\Sigma_1 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}.$$

Table 4. *Summary statistics of X_1, X_2, X_3 in different scenarios*

Population	Statistics	Scenario			
		1	2	3	4
$\mathbb{P}^{\mathrm{s}}$	Mean	$0.5, -1, 0$	$0.5, 0, 0$	$0.7, 0.1, -0.2$	$0.6, 0, 0$
	Variance	$0.25, 1, 1$	$0.25, 2, 2$	$0.21, 1, 1$	$0.24, 1, 1$
$\mathbb{P}^{\mathrm{t}}$	Mean	$0.5, -1, 0$	$0.8, 0.6, -0.6$		
	Variance	$0.25, 1, 1$	$0.16, 1.64, 1.64$		

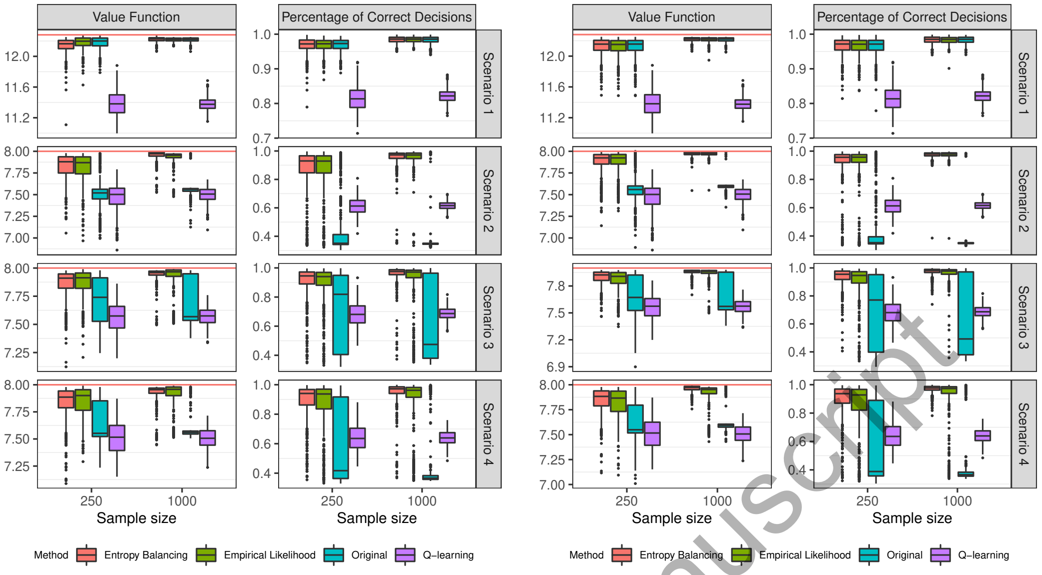
We considered a source sample with size $n = 250, 1000$. For each setting, we conducted 500 replications. In our implementation, the propensity score and conditional mean outcome models are estimated using two methods:

(I) Both are estimated based on posited parametric models. In particular, the propensity score is estimated using a correctly specified logistic regression model, while the conditional mean outcome is estimated using a linear model with all the covariates and covariate-treatment interactions, which is a misspecified model.

(II) The propensity score is estimated nonparametrically using a generalized additive model, and the conditional mean outcome model $\mu(x, a)$ is estimated nonparametrically using the random forest for $a = 0$ and 1, separately.

We also implemented Q-learning as a benchmark for comparison. Specifically, we fitted linear models for Q-functions and inferred optimal linear ITRs from the estimated Q-functions. An ITR estimated by Q-learning is denoted as $d(x; \widehat{\beta}_Q)$. To evaluate and compare the performance of estimated optimal ITRs obtained from the original AIPW estimator, proposed calibrated AIPW estimators, and Q-learning, we compute the corresponding value functions and percentages of correct decisions for the target population. Specifically, we generate covariates X^{t} for a large sample with size $N = 10^5$ from the target population. The value function of an estimated ITR $d(x; \widehat{\beta})$, where $\widehat{\beta} = \widehat{\beta}^{\mathrm{o}}$, $\widehat{\beta}^{\mathrm{c}}_{EB}$, $\widehat{\beta}^{\mathrm{c}}_{EL}$, or $\widehat{\beta}_Q$ is computed by $V^{\mathrm{t}}(\widehat{\beta}) = N^{-1} \sum_{i=1}^{N} \mu\{X^{\mathrm{t}}_i, d(X^{\mathrm{t}}_i; \widehat{\beta})\}$, and its associated percentage of correct decisions is $1 - N^{-1} \sum_{i=1}^{N} |d(X^{\mathrm{t}}_i; \widehat{\beta}) - d(X^{\mathrm{t}}_i; \beta^{\mathrm{t}})|$. Here, the true optimal ITR $d(X; \beta^{\mathrm{t}})$ for the target population is obtained by maximizing $V^{\mathrm{t}}(\beta)$ over β using the grid-search method. We report the values and percentages of correct decisions results of $d(x; \widehat{\beta}^{\mathrm{o}})$, $d(x; \widehat{\beta}^{\mathrm{c}}_{EB})$, $d(x; \widehat{\beta}^{\mathrm{c}}_{EL})$, and $d(x; \widehat{\beta}_Q)$ for the observational study in Fig. 1(a) (method I) and Fig. 1(b) (method II). Similar results for the randomization study are provided in the supplementary material.

We have the following observations. In Scenario 1, the optimal ITR estimated by Q-learning has poor performance in terms of value and percentage of correct decisions, due to the misspecification of Q-function. All other three estimated optimal ITRs have good and comparable performance in terms of values and percentages of correct decisions, which is expected since $\mathbb{P}^{\mathrm{s}} = \mathbb{P}^{+}_{EB} = \mathbb{P}^{+}_{EL} = \mathbb{P}^{\mathrm{t}}$. In addition, as the sample size increases, the means of value functions become closer to the true optimal value for the target population, percentages of correct decisions get closer to 1, and the standard deviations of value functions and percentages of correct decisions become smaller. However, in Scenarios 2-4 where $\mathbb{P}^{\mathrm{s}} \neq \mathbb{P}^{\mathrm{t}}$, the estimated optimal ITR obtained using the original method has poor performance: the means of value functions are much smaller than the true optimal value for the target population and percentages of correct decisions are also much smaller than 1. This implies that the estimated optimal ITR obtained using the original method may not generalize well to the target population when $\mathbb{P}^{\mathrm{s}} \neq \mathbb{P}^{\mathrm{t}}$. The optimal ITR estimated by Q-learning still yields poor performance. However, the estimated optimal ITRs obtained using the proposed calibration methods still have competitive performance similar to those



(a) Implementation method I (b) Implementation method II

Fig. 1. The value and percentage of correct decisions results of estimated optimal ITRs for the observational study with implementation methods I and II. The red lines are the values of the true optimal ITRs for the target population.

observed in Scenario 1. This supports that the proposed calibration using summary statistics can improve the treatment decision for the target population.

Next, we study the estimation and inference results of $\widehat{V}^{c}_{EB}(\widehat{\beta}^{c}_{EB})$ and $\widehat{V}^{c}_{EL}(\widehat{\beta}^{c}_{EL})$. For implementation method I, the asymptotic variances of $\widehat{V}^{c}_{EB}(\widehat{\beta}^{c}_{EB})$ and $\widehat{V}^{c}_{EL}(\widehat{\beta}^{c}_{EL})$ were estimated using the results established in Theorem 1, while for implementation method II, the corresponding asymptotic variances were estimated using the results established in Theorem 2 because Condition (A11') holds. In our simulations, we observed that the empirical likelihood method may produce a few extreme calibration weights in Scenarios 3 and 4. These extreme weights usually do not inflate the biases of $\widehat{V}^{c}_{EL}(\widehat{\beta}^{c}_{EL})$, but they do lead to overestimated standard errors due to the instability in variance estimation. To control the effects of these extreme weights, we stabilize the weights by reducing the large weights $\widehat{w}_i$ $(> a_n^{-1})$ to $\tilde{w}_i$ according to $(\tilde{w}_i)^{-1} = (\widehat{w}_i)^{-1} + a_n$ for $a_n = o_p(n)$. Such a stabilization leads all weights to be no larger than a_n^{-1}. The rationale for considering a_n to be $o_p(n)$ is that because $w_i \propto n^{-1}$, the stabilization does not affect the weights asymptotically. In the simulation study, we take $a_n = 12 \log n$. Based on our numerical studies, such a stabilization doesn't affect the biases of $\widehat{V}^{c}_{EL}(\widehat{\beta}^{c}_{EL})$ much but can give a reasonable standard error estimate. On the other hand, the calibration weights computed using the entropy balancing method do not have any extreme values in all four scenarios. Such an observation is consistent with the findings in the literature since the entropy balancing loss tends to penalize the deviation of the estimated weights $\widehat{w}_i$'s from n^{-1} more than the empirical likelihood method. We report the mean and standard deviation of $\widehat{V}^{c}_{EB}(\widehat{\beta}^{c}_{EB})$ and $\widehat{V}^{c}_{EL}(\widehat{\beta}^{c}_{EL})$, the mean of estimated standard errors and the empirical coverage probability (CP) of 95% Wald-type confidence intervals. The true optimal values $V^{+}_{EB}(\beta^{*}_{EB})$ and $V^{+}_{EL}(\beta^{*}_{EL})$ are computed using the grid-search method based on a large dataset generated from the corresponding pseudo populations similar to

Table 5. *Simulation results for the observational study with implementation method I*

Method	Scenario	1		2		3		4	
	n	250	1000	250	1000	250	1000	250	1000
	$V^{\mathrm{t}}(\beta^{\mathrm{t}})$	12.28		8.00		8.00		8.00	
Entropy Balancing	$V^+_{EB}(\beta^*_{EB})$	12.28		8.00		7.99		7.99	
	Mean	12.34	12.32	8.20	8.06	8.22	8.08	8.24	8.08
	SD	0.42	0.21	0.36	0.17	0.31	0.16	0.45	0.17
	SE	0.47	0.24	0.41	0.19	0.35	0.17	0.43	0.19
	CP$^+$	96.8	95.6	95.4	95.6	95.0	95.2	94.6	94.8
	CP$^{\mathrm{t}}$	96.8	95.6	95.4	95.6	95.2	96.0	94.8	96.0
Empirical Likelihood	$V^+_{EL}(\beta^*_{EL})$	12.28		8.00		8.14		8.16	
	Mean	12.36	12.31	8.17	8.06	8.09	8.19	7.90	8.22
	SD	0.41	0.22	0.37	0.16	0.35	0.20	0.43	0.26
	SE	0.38	0.20	0.41	0.19	0.32	0.20	0.37	0.27
	CP$^+$	93.6	96.0	96.4	97.0	92.0	96.8	81.2	94.0
	CP$^{\mathrm{t}}$	93.6	96.0	96.4	97.0	93.2	88.4	87.0	93.4

Table note: Mean, the average of estimates; SD, the empirical standard deviation of estimates; SE, the mean of estimated standard errors; CP$^+(\%)$, the empirical coverage probability of a 95% confidence interval for $V^+_{EB}(\beta^*_{EB})$ or $V^+_{EL}(\beta^*_{EL})$; CP$^{\mathrm{t}}(\%)$, the empirical coverage probability of a 95% confidence interval for $V^{\mathrm{t}}(\beta^{\mathrm{t}})$.

Table 6. *Simulation results for the observational study with implementation method II*

Method	Scenario	1		2		3		4	
	n	250	1000	250	1000	250	1000	250	1000
	$V^{\mathrm{t}}(\beta^{\mathrm{t}})$	12.28		8.00		8.00		8.00	
Entropy Balancing	$V^+_{EB}(\beta^*_{EB})$	12.28		8.00		7.99		7.99	
	Mean	12.42	12.32	8.12	8.03	8.15	8.06	8.19	8.06
	SD	0.50	0.22	0.33	0.15	0.29	0.13	0.34	0.15
	SE	0.54	0.25	0.38	0.18	0.30	0.14	0.37	0.17
	CP$^+$	97.4	97.0	96.8	96.2	93.8	95.4	94.6	95.4
	CP$^{\mathrm{t}}$	97.4	97.0	96.8	96.2	94.6	95.6	95.0	96.0
Empirical Likelihood	$V^+_{EL}(\beta^*_{EL})$	12.28		8.00		8.14		8.16	
	Mean	12.42	12.32	8.11	8.03	8.06	8.18	7.86	8.16
	SD	0.50	0.23	0.33	0.15	0.31	0.17	0.43	0.23
	SE	0.46	0.21	0.37	0.17	0.29	0.17	0.35	0.24
	CP$^+$	95.4	95.4	96.6	96.0	91.6	95.4	74.4	93.2
	CP$^{\mathrm{t}}$	95.4	95.4	96.6	96.0	94.0	84.0	83.4	96.2

Notation is defined in Table 5.

the computation of $V^{\mathrm{t}}(\beta^{\mathrm{t}})$. In addition, we consider two types of CP: (1) CP$^+$ for the optimal values $V^+_{EB}(\beta^*_{EB})$ or $V^+_{EL}(\beta^*_{EL})$ of the corresponding pseudo population; (2) CP$^{\mathrm{t}}$ for the optimal values $V^{\mathrm{t}}(\beta^{\mathrm{t}})$ of the target population. Simulation results for the observational study are summarized in Table 5 with implementation method I and Table 6 with implementation method II. Similar results for the randomization study are provided in the supplementary material.

We have the following observations. In Scenarios 1 and 2, since $\mathbb{P}^+_{EB} = \mathbb{P}^+_{EL} = \mathbb{P}^{\mathrm{t}}$, we have $V^+_{EB}(\beta^*_{EB}) = V^+_{EL}(\beta^*_{EL}) = V^{\mathrm{t}}(\beta^{\mathrm{t}})$. Both calibrated value estimators are nearly unbiased. The mean of estimated standard errors is close to the standard deviation of the estimators, and the empirical coverage probability of 95% confidence intervals is close to the nominal level for all settings. In Scenarios 3 and 4, $V^+_{EB}(\beta^*_{EB})$ or $V^+_{EL}(\beta^*_{EL})$ is no longer equal to $V^{\mathrm{t}}(\beta^{\mathrm{t}})$. However, we can see that both $V^{+}_{EB}(\beta^{*}_{EB})$ and $V^{+}_{EL}(\beta^{*}_{EL})$ are close to $V^{\mathrm{t}}(\beta^{\mathrm{t}})$. In particular, the difference between $V^{+}_{EB}(\beta^{*}_{EB})$ and $V^{\mathrm{t}}(\beta^{\mathrm{t}})$ is nearly negligible. This implies that both calibration methods give good approximation of the target population, while the entropy balancing method is better than the empirical likelihood method for the considered Scenarios 3 and 4. A possible explanation is that the probability $\mathrm{pr}(S = 0 \mid X)$ can be well approximated by a logistic regression under Scenarios 3 and 4 so that the entropy balancing calibration method can approximate the target population very well. Moreover, the entropy balancing estimators are nearly unbiased, the mean of estimated standard errors is close to the standard deviation of the estimators, and the empirical coverage probabilities of 95% confidence intervals for both $V^{+}_{EB}(\beta^{*}_{EB})$ and $V^{\mathrm{t}}(\beta^{\mathrm{t}})$ are close to the nominal level for all settings. For the empirical likelihood method, as n increases, the mean of estimators get closer to its true value $V^{+}_{EL}(\beta^{*}_{EL})$, the mean of estimated standard errors get closer to the standard deviation of estimators, and the empirical coverage probability of 95% confidence intervals for $V^{+}_{EL}(\beta^{*}_{EL})$ get closer to the nominal level as expected. However, because of the difference between $V^{+}_{EL}(\beta^{*}_{EL})$ and $V^{\mathrm{t}}(\beta^{\mathrm{t}})$, the empirical coverage probability of 95% confidence intervals for $V^{\mathrm{t}}(\beta^{\mathrm{t}})$ is lower than the nominal level for some settings even when n increases to 1000. Finally, standard deviations of the estimators reported in Table 6 for implementation method II are generally smaller than the corresponding values reported in Table 5 for implementation method I. Such efficiency gains are mainly due to the nonparametric fit of the conditional mean outcome model in implementation method II compared with the misspecified parametric conditional mean outcome model used in implementation method I.

We also compared the original AIPW estimator without calibration with the calibrated AIPW estimators under Scenario 1, where the source and target populations are identical. As expected, all three estimators are consistent for the optimal value of the target population. The standard deviations of the original AIPW estimator for observational study with implementation method II are $0.55, 0.27$ for $n = 250, 1000$. These values are larger than the corresponding values of calibrated estimators in Table 6, which supports the results established in Theorem 3.

6. REAL DATA ANALYSIS

We illustrate the proposed method using an application to data from the eICU collaborative research database (eICU-CRD) (Goldberger et al., 2000; Pollard et al., 2018, 2019) and the MIMIC-III clinical database (Goldberger et al., 2000; Johnson et al., 2016, 2019). Specifically, we use the eICU dataset as the source population, while treating the MIMIC-III dataset as the target population. Both MIMIC-III and eICU data consist of patients who suffered from sepsis. The eICU-CRD is a multi-center ICU database comprising de-identified health-related data associated with over 200,000 admissions to ICUs across the US between 2014-2015. The MIMIC-III database is a single-center ICU database comprising de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. It is likely that the populations in the two databases have some heterogeneity.

Both eICU and MIMIC-III data collected information from ICU patients with sepsis disease, and thus contain common baseline covariates and treatment. In our study, we consider $p = 7$ baseline covariates in both samples: age (years), admission weights (kg), admission temperature (Celsius), glucose level (mg/dL), blood urea nitrogen (BUN) amount (mg/dL), creatinine amount (mg/dL), white blood cell (WBC) count (K/uL). Here, treatment is coded as 1 if receiving the vasopressor, and 0 if receiving other medical supervisions such as IV fluid resuscitation. We consider the cumulative balance (mL) as the outcome of interest. A positive cumulative balance indicates that a patient's fluid input is higher than their output. The condition describing

Table 7. *Mean and standard deviations (in parenthesis) of baseline characteristics in the source and target datasets*

	Source	Target
$-\lvert$Cumulative Balance$\rvert$ (Y)	-1746.6 (1561.3)	-1785.0 (1246.6)
Age (X_1)	65.7 (15.1)	66.5 (16.5)
Admission Weights (X_2)	80.0 (22.9)	79.7 (20.7)
Admission Temperature (X_3)	36.5 (1.1)	36.8 (0.8)
Glucose (X_4)	158.6 (102.7)	145.6 (72.4)
BUN (X_5)	31.7 (20.3)	27.9 (18.4)
Creatinine (X_6)	1.8 (1.5)	1.5 (1.4)
WBC (X_7)	14.4 (8.4)	12.0 (6.5)

excess fluid is known as hypervolaemia or fluid overload. In critically ill patients, fluid overload is related to increased mortality and also leads to several complications like pulmonary edema, cardiac failure, tissue breakdown, and impaired bowel function (Claure-Del Granado & Mehta, 2016). A negative cumulative balance indicates that a patient's fluid output is higher than their input. The condition describing inadequate fluid is known as hypovolaemia. Patients with severe hypovolemia can develop ischemic injury of vital organs, leading to multi-system organ failure (Taghavi & Askari, 2021). We use $Y = -\lvert$cumulative balance$\rvert$ as the outcome, so a larger value of the outcome is better. After removing abnormal values, the MIMIC-III dataset consists of 10746 subjects, among which 2242 patients were treated with the vasopressor, while the rest were treated with other medical supervisions. The MIMIC-III data is treated as the target population. We sample $n = 1000$ subjects from the eICU dataset as the source sample, among which 271 patients were treated with the vasopressor, while the rest were treated with other medical supervisions. Table 7 summarizes the mean and standard deviation of the outcome and covariates in the source and target samples. We can see some differences in the means of some covariates, such as glucose level, blood urea nitrogen amount, and WBC count.

We used the means of all seven covariates of the target population as the summary statistics to estimate the calibration weights by the entropy balancing and empirical likelihood methods. We computed three optimal linear ITRs, $d(x; \widehat{\beta}^{\rm o})$, $d(x; \widehat{\beta}^{\rm c}_{EB})$, and $d(x; \widehat{\beta}^{\rm c}_{EL})$ by maximizing the original and calibrated AIPW value function estimators based on the source sample. In our implementation, the propensity score model was estimated using a logistic regression with all covariates and the conditional mean outcome model was estimated using the random forest for treatments 0 and 1 separately. To assess the performance of these three estimated optimal ITRs for treatment decisions in the target population, we apply them to random samples drawn from the target population. Specifically, we randomly sample $N = 1000$ subjects from the MIMIC-III data as the target sample and repeat this sampling procedure 100 times. We have individual-level data from the target population, which can be used as the benchmark for evaluation. For a given ITR $d(x;\beta)$, we computed the AIPW estimator of its value function based on the target sample by

$$\widehat{V}^{\rm t}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{I\{A_i^{\rm t} = d(X_i^{\rm t};\beta)\}}{\varrho^{\rm t}(A_i^{\rm t} \mid X_i^{\rm t};\widehat{\eta})}\{Y_i^{\rm t} - \widehat{\mu}_d^{\rm t}(X_i^{\rm t};\beta)\} + \widehat{\mu}_d^{\rm t}(X_i^{\rm t};\beta)\right],$$

where $\varrho^{\rm t}(A_i^{\rm t} \mid X_i^{\rm t};\widehat{\eta}) = \pi^{\rm t}(X_i^{\rm t};\widehat{\eta})A_i^{\rm t} + \{1 - \pi^{\rm t}(X_i^{\rm t};\widehat{\eta})\}(1 - A_i^{\rm t})$, $\widehat{\mu}_d^{\rm t}(X_i^{\rm t};\beta) = \widehat{\mu}^{\rm t}(X_i^{\rm t},1)I\{d(X_i^{\rm t};\beta) = 1\} + \widehat{\mu}^{\rm t}(X_i^{\rm t},0)I\{d(X_i^{\rm t};\beta) = 0\}$, the propensity score $\pi^{\rm t}(X_i^{\rm t};\widehat{\eta})$ was estimated using a logistic regression model, and the conditional mean outcome models $\widehat{\mu}^{\rm t}(X_i^{\rm t},a)$, $a = 0, 1$, were estimated using random forest.

Table 8. *Mean and standard deviations (in parenthesis) of the value estimators and percentages of correct decisions*

	Oracle	Entropy Balancing	Empirical Likelihood	Original
Value	-1674.1 (47.2)	-1752.0 (49.5)	-1773.8 (49.3)	-1945.3 (75.6)
Percentage of Correct Decisions	/	0.80 (0.1)	0.76 (0.1)	0.31 (0.1)

Let $\widehat{\beta}^{\text{oracle}} = \text{argmax}_{\beta}\, \widehat{V}^{\text{t}}(\beta)$. Then, $d(x;\widehat{\beta}^{\text{oracle}})$ is the optimal linear ITR for the target sample and $\widehat{V}^{\text{t}}(\widehat{\beta}^{\text{oracle}})$ is the associated optimal value, which can serve as the benchmark. We also computed the estimated values of three estimated ITRs $d(x;\widehat{\beta}^{\text{o}})$, $d(x;\widehat{\beta}^{\text{c}}_{EB})$, and $d(x;\widehat{\beta}^{\text{c}}_{EL})$ by $\widehat{V}^{\text{t}}(\widehat{\beta}^{\text{o}})$, $\widehat{V}^{\text{t}}(\widehat{\beta}^{\text{c}}_{EB})$, and $\widehat{V}^{\text{t}}(\widehat{\beta}^{\text{c}}_{EL})$, respectively, and their associated percentages of correct decisions, defined as $1 - N^{-1}\sum_{i=1}^{N} |d(X^{\text{t}}_i;\widehat{\beta}) - d(X^{\text{t}}_i;\widehat{\beta}^{\text{oracle}})|$ for an estimated ITR $d(x;\widehat{\beta})$. Table 8 summarize the means and standard deviations of the value estimators and percentages of correct decisions over 100 replications. We can see that the ITRs obtained using the proposed calibration methods have much better performance than the original AIPW estimator without calibration. Their estimated values are much closer to the optimal value computed using the target samples and the associated percentages of correct decisions are much closer to 1. Moreover, the ITR obtained using the entropy balancing method has slightly better performance than the one obtained using the empirical likelihood method in terms of both value and percentage of correct decisions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of theorems and additional simulation results.

REFERENCES

ATHEY, S. & WAGER, S. (2021). Policy learning with observational data. *Econometrica* **89**, 133–161.

BUCHANAN, A. L., HUDGENS, M. G., COLE, S. R., MOLLAN, K. R., SAX, P. E., DAAR, E. S., ADIMORA, A. A., ERON, J. J. & MUGAVERO, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. R. Statist. Soc.* A **181**, 1193–1209.

CATTANEO, M. D., JANSSON, M. & NAGASAWA, K. (2020). Bootstrap-based inference for cube root asymptotics. *Econometrica* **88**, 2203–2219.

CHEN, Z., NING, J., SHEN, Y. & QIN, J. (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics* **77**, 1024–1036.

CLAURE-DEL GRANADO, R. & MEHTA, R. L. (2016). Fluid overload in the icu: evaluation and management. *BMC Nephrology* **17**, 109.

COLE, S. R. & STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American Journal of Epidemiology* **172**, 107–115.

CRESSIE, N. & READ, T. R. (1984). Multinomial goodness-of-fit tests. *J. R. Statist. Soc.* B **46**, 440–464.

DAHABREH, I. J., ROBERTSON, S. E., TCHETGEN, E. J., STUART, E. A. & HERNÁN, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75**, 685–694.

FARRELL, M. H., LIANG, T. & MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89**, 181–213.

GOLDBERGER, A. L., AMARAL, L. A., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E., MOODY, G. B., PENG, C.-K. & STANLEY, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20**, 25–46.

HUANG, C.-Y. & QIN, J. (2020). A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Statistics in Medicine* **39**, 1573–1590.

JOHNSON, A., POLLARD, T. & MARK, R. (2019). Mimic-iii clinical database demo (version 1.4). *PhysioNet* .

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A. & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data* **3**, 1–9.

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

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D. & Cai, J. (2021). Improving trial generalizability using observational studies. *Biometrics* , doi:10.1111/biom.13609.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. & Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *J. Am. Statist. Assoc.* **115**, 692–706.

Luedtke, A. R. & Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44**, 713–742.

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

Mo, W., Qi, Z. & Liu, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *J. Am. Statist. Assoc.* **116**, 659–674.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Statist. Soc.* B **65**, 331–355.

Newey, W. K. & Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica* **72**, 219–255.

Pollard, T., Johnson, A., Raffa, J., Celi, L., Badawi, O. & Mark, R. (2019). Icu collaborative research database (version 2.0). *PhysioNet* .

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G. & Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data* **5**, 1–13.

Qian, M. & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39**, 1180–1210.

Qin, J. & Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc.* B **69**, 101–122.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6**, 34–58.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.* **35**, 634–672.

Sugiyama, M. & Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT Press.

Taghavi, S. & Askari, R. (2021). Hypovolemic shock. *StatPearls [Internet]* .

Uehara, M., Kato, M. & Yasui, S. (2020). Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems* **33**, 49–61.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing* **4**, 65–85.

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

Zhang, B., Tsiatis, A. A., Laber, E. B. & Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100**, 681–694.

Zhao, Q. & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference* **5**.

Zhao, Y.-Q., Zeng, D., Tangen, C. M. & LeBlanc, M. L. (2019). Robustifying trial-derived optimal treatment rules for a target population. *Electron. J. Statist.* **13**, 1717–1743.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Assoc.* **110**, 910–922.

[*Received on* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]