HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2017 December 28.

Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1199–1209. doi:10.1111/biom.12676.

A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring

Shuai Chen[1], **Lu Tian**[2], **Tianxi Cai**[3], and **Menggang Yu**[1,*]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53792, U.S.A

[2]Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305, U.S.A

[3]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A

Summary

Many statistical methods have recently been developed for identifying subgroups of patients who may benefit from different available treatments. Compared with the traditional outcome-modeling approaches, these methods focus on modeling interactions between the treatments and covariates while by-pass or minimize modeling the main effects of covariates because the subgroup identification only depends on the *sign* of the interaction. However these methods are scattered and often narrow in scope. In this paper, we propose a general framework, by weighting and A-learning, for subgroup identification in both randomized clinical trials and observational studies. Our framework involves minimum modeling for the relationship between the outcome and covariates pertinent to the subgroup identification. Under the proposed framework, we may also estimate the *magnitude* of the interaction, which leads to the construction of scoring system measuring the individualized treatment effect. The proposed methods are quite flexible and include many recently proposed estimators as special cases. As a result, some estimators originally proposed for randomized clinical trials can be extended to observational studies, and procedures based on the weighting method can be converted to an A-learning method and vice versa. Our approaches also allow straightforward incorporation of regularization methods for high-dimensional data, as well as possible efficiency augmentation and generalization to multiple treatments. We examine the empirical performance of several procedures belonging to the proposed framework through extensive numerical studies.

Keywords

A-learning; Individualized treatment rules; Observational studies; Propensity score; Regularization

* meyu@biostat.wisc.edu.

Supplementary Materials

Sample R codes for implementing the proposed method and Web Appendices referenced in Sections 2 – 5 are available with this paper at the Biometrics website on Wiley Online Library.

1. Introduction

With increasing numbers and types of treatments for many conditions, it is now well known that the benefits of many treatments differ substantially across different patient subpopulations. A key focus of recent research is to match patients with the most effective treatments to improve treatment efficacy when there is substantial heterogeneity of treatment effectiveness (Gabriel and Normand, 2012). To optimize treatment selection for individual patients, an important strategy is to use patients' baseline covariates to form a system for ranking or scoring their individualized treatment effects (ITEs). Statistical methods for estimating ITEs or constructing optimal individualized treatment rules (ITRs) often require investigation of treatment by covariate interactions. These treatment-modifying covariates are known as treatment-moderators. They need to be contrasted with prognostic covariates which lead to poorer or better outcomes under all treatment options. For example, higher tumor stage may be generally associated with worse prognosis for all treatments under considerations. Therefore if the goal is only for treatment selection, the relevant task is to identify treatment-moderators but ignore prognostic covariates.

More concretely, let $T = \pm 1$, $\mathbf{Z}$ and Y denote the treatment assignment, baseline covariates and outcome of interest for a patient, respectively. Algebraically, we can express $E(Y|T, \mathbf{Z}) = m(\mathbf{Z}) + T\Delta(\mathbf{Z})$, where $m(\mathbf{Z}) = 0.5\{E(Y|T = 1, \mathbf{Z}) + E(Y|T = -1, \mathbf{Z})\}$ is a function that reflects the main effect of $\mathbf{Z}$ and $\Delta(\mathbf{Z}) = 0.5\{E(Y|T = 1, \mathbf{Z}) - E(Y|T = -1, \mathbf{Z})\}$ is a contrast function that reflects treatment effects given $\mathbf{Z}$. Therefore modeling $E(Y|T, \mathbf{Z})$ is equivalent to modeling both $m(\mathbf{Z})$ and $\Delta(\mathbf{Z})$ as functions of covariates. Variables involved in $m(\mathbf{Z})$ are prognostic variables, while those in $\Delta(\mathbf{Z})$ are treatment moderators.

Traditional approaches to developing optimal ITRs model $m(\mathbf{Z})$ and $\Delta(\mathbf{Z})$ simultaneously to predict the outcomes and then estimate ITEs using these model based estimates. This approach requires correct specification for both $m(\mathbf{Z})$ and $\Delta(\mathbf{Z})$, even though in the end only the latter is used to guide the treatment selection. In this approach, the main effect $m(\mathbf{Z})$ becomes a nuisance parameter, whose specification, however, may affect the estimation of treatment contrast function $\Delta(\mathbf{Z})$. This is especially problematic since in practice there are often many prognostic variables but far fewer treatment moderators that actually alter treatment recommendation (Kraemer, 2013). However, if our goal is ranking ITE or developing ITR, then only ranks or signs of the contrast function $\Delta(\mathbf{Z})$ matter. Hence, it is desirable to have a robust estimate of ITEs without necessitating the estimation of $m(\cdot)$.

Several robustness approaches to subgroup identification have been proposed in recent years. In the randomized clinical trial (RCT) setting, it has been shown that mis-specification of the main effects $m(\mathbf{Z})$ has limited effects on estimating the treatment by covariate interaction (Qian and Murphy, 2011; McKeague and Qian, 2014; Lu et al., 2013; Ciarleglio et al., 2015), especially from the perspective of A-learning (Murphy, 2003; Robins, 2004). In the observational study setting, double robust procedure was introduced (Zhang et al., 2012). In addition, Vansteelandt et al. (2008) developed multiply robust estimators for interaction parameters, leaving the main effects unspecified. Beyond this robustness from the main effect misspecification, it has been found that masking observation such as noninformative censoring also only has a limited impact on subgroup identification (Xu et al., 2015).

Shifting from the outcome prediction framework, a modified covariate method was proposed by Tian et al. (2014) without the need of explicitly modeling main effects for data from RCTs. Furthermore, nonparametric approaches based on the regression tree were proposed to separate the main effects from the covariate-treatment interaction effects, either through sequential testing (Su et al., 2008) or prediction strategies (Foster et al., 2011; Loh et al., 2015). Lastly, converting the subgroup identification to a classification problem, outcome weighted learning methods were developed (Qian and Murphy, 2011; Zhang et al., 2012; Zhao et al., 2012; Xu et al., 2015).

All aforementioned methods appear to be very diverse: some are fairly ad-hoc and some are heavily model-dependent. The validity of the latter often relies on different parametric or non-parametric assumptions for $\Delta(\mathbf{Z})$. Furthermore, most existing methods focus on the treatment difference $\Delta(\mathbf{Z})$ as the metric for summarizing ITE. However, the choice of the metric may affect the analysis results and conclusions about ITE. For example, for non-negative Y, one may choose to use the ratio $E(Y|T=1,\mathbf{Z})/E(Y|T=-1,\mathbf{Z})$ to measure the ITE instead of the difference $E(Y|T=1,\mathbf{Z})-E(Y|T=-1,\mathbf{Z})$. The patients rankings based on $E(Y|T=1,\mathbf{Z})/E(Y|T=-1,\mathbf{Z})$ and $E(Y|T=1,\mathbf{Z})-E(Y|T=-1,\mathbf{Z})$ can potentially be quite different although these two metrics may lead to the same ITRs with patients assigned to treatment 1 when $E(Y|T=1,\mathbf{Z})>E(Y|T=-1,\mathbf{Z})$. On the other hand, one may employ ITE metrics such as $E\{U(Y)|T=1,\mathbf{Z}\}-E\{U(Y)|T=-1,\mathbf{Z}\}$ based on a monotone transformation $U(\cdot)$, which would alter both the ranking and the optimal treatment recommendation. Although different methods can be developed for specific choices of the ITE metric, we propose in this paper a unified framework that can allow for ITEs or ITRs estimation under different metrics.

Building upon the weighting approach considered in Tian et al. (2014) for RCT, we propose both propensity score weighting and A-learning methods for subgroup identification that are applicable for both observational studies and RCTs. Our framework is flexible and includes many aforementioned estimators in the literature as special cases, despite their clearly different origins. The rest of the paper is organized as follows. In Section 2, we demonstrate that minimizers of a class of convex loss functions can recover the optimal ITR. With properly chosen loss functions, our proposed estimator can not only recover the ITR but also the magnitude of ITE. In Section 3, we show that many recently proposed estimators can be represented as special cases within our frameworks. In Sections 4 and 5, we compare the finite-sample properties of several estimators and their extensions within our frameworks via simulation studies and real data examples. Finally we conclude the paper with some discussions including extensions to multiple treatment groups in Section 6.

2. Methods

2.1 Notations and assumptions

We adopt the notation based on the potential outcome framework in causal inference (Rubin, 2005). $Y^{(1)}$ and $Y^{(-1)}$ are the potential outcomes if the patient receives a new treatment $T=1$ and a standard treatment $T=-1$, respectively. We also assume that only one of the potential outcomes $Y^{(1)}$ and $Y^{(-1)}$ can be observed for each patient, i.e, $Y=I(T=1)Y^{(1)}+I(T=-1)Y^{(-1)}$, where $I(\cdot)$ is the indicator function. We further assume that T is independent of

$(Y^{(1)}, Y^{(-1)})$ given the covariates $\mathbf{Z}$, that is the "strongly ignorable assumption" (Rosenbaum and Rubin, 1983; Rubin, 2005). For the treatment assignment, we assume that $\Pr(T = 1|\mathbf{Z}) = \pi(\mathbf{Z})$, where the propensity score $\pi(\mathbf{Z})$ is typically known and free of $\mathbf{Z}$ in randomized trials; but is unknown and needs to be estimated (e.g. via regression modeling) in observational studies. The observed data $\{(Y_i, T_i, \mathbf{Z}_i), i = 1, \cdots, n\}$ consist of n independent identically distributed (i.i.d) copies of $(Y, T, \mathbf{Z})$.

Our goal is to construct a personalized benefit scoring system $f(\mathbf{Z})$ based on the covariates $\mathbf{Z}$ via both a weighting approach and an A-learning approach such that the new treatment shall be recommended for the patients based on $f(\mathbf{Z})$, which is often $\text{sign}\{f(\mathbf{Z})\}$. We demonstrate in Section 2.2 that the optimality of such a rule under different scenarios. In addition, we demonstrate in Section 2.4 that the proposed approach to estimating an optimal $f(\cdot)$ is also useful for quantifying the magnitude of treatment benefit.

We consider a loss function $M(y, v)$ satisfying two conditions:

- **A.** $M_v(y, v) = \partial M(y, v)/\partial v$ is increasing in v for any given y;
- **B.** $U(y) \equiv M_v(y, 0)$ is monotone in y.

Here, condition A ensures that $M(y, v)$ is convex in v, which allows us to "order" the expected utility under the comparative treatments to form an ITR. Condition B is simply to make the transformed quantity, i.e. $U(Y)$, an interpretable endpoint. For example, $M(y, v)$ can be the squared loss function $(y - v)^2$, which clearly satisfies aforementioned two conditions with $M_v(y, v) = 2v - 2y$ and $U(y) = -2y$. When Y only taking non-negative values and $\Pr(Y^{(t)} > 0|\mathbf{Z} = \mathbf{z}) > 0$ for $t = \pm 1$ and any $\mathbf{z}$, we may let $M(y, v) = y \log\{1 + \exp(-v)\}$ with $U(y) = -y/2$.

A "Fisher-consistent" ITR $d_0(\mathbf{z}) \in \{1, -1\}$ can be constructed via M. Specifically, $d_0(\cdot)$ maximizes the value function $\mathbb{V}_U(d) = -E[\{U(Y^{(1)}) - U(Y^{(-1)})\} d(\mathbf{Z})]$. Note that direct maximization of $\mathbb{V}_U(d)$ is not feasible both statistically and numerically due to the discrete nature of $d(\cdot)$, whereas minimization of smooth loss functions $\ell(f)$ with respect to $f(\cdot)$ overcomes such difficulties. Here $\ell(\cdot)$ is constructed based on M as detailed in the next section. When $U(y)$ linear and decreasing in y, the maximizer of $\mathbb{V}_U(d)$ is the same as the maximizer of the standard value function $E[Y^{(1)}I\{d(\mathbf{Z}) = 1\} + Y^{(-1)}I\{d(\mathbf{Z}) = -1\}]$ employed in the literature (Qian and Murphy, 2011; Zhao et al., 2012). The use of a broader class of $M(\cdot, \cdot)$ along with its corresponding $U(\cdot)$ enables us to consider alternative metrics to quantify treatment benefits. Throughout, we first assume that $\pi(\mathbf{Z})$ is known and provide discussions on estimating $\pi(\mathbf{Z})$ in Section 2.3.

2.2 Weighting and A-learning Approaches to Subgroup Identification

Weighting Method—For a given $M(\cdot, \cdot)$ and covariate level $\mathbf{z}$, we first consider the loss function $\ell_W(f) = E\{\ell_W(f, \mathbf{Z})\}$ and let $f_{W0} = \text{argmin}_f \ell_W(f)$, where

$$\ell_W(f, \mathbf{z}) = E\left[\frac{M\{Y, Tf(\mathbf{Z})\}}{T\pi(\mathbf{Z}) + (1-T)/2}|\mathbf{Z} = \mathbf{z}\right]$$
$$= E[M\{Y, f(\mathbf{Z})\}|T = 1, \mathbf{Z} = \mathbf{z}] + E[M\{Y, -f(\mathbf{Z})\}|T = -1, \mathbf{Z} = \mathbf{z}].$$

We next show that $d_0(\mathbf{Z}) = \text{sign}\{ f_{W0}(\mathbf{Z})\}$ maximizes the value function $\mathbb{V}_U(d)$. For any $\mathbf{z}$, the first order condition of the minimization is

$$E[M_v\{Y, f_{W0}(\boldsymbol{Z})\}|T=1, \boldsymbol{Z}=\mathbf{z}]=E[M_v\{Y, -f_{W0}(\boldsymbol{Z})\}|T=-1, \boldsymbol{Z}=\mathbf{z}]. \quad (1)$$

Consequently, for a patient with a negative score (that is, $f_{W0}(\mathbf{z}) < 0$), we have

$$\begin{aligned} &E\{U(Y^{(1)})|\boldsymbol{Z}=\mathbf{z}\}=E\{M_v(Y,0)|T=1, \boldsymbol{Z}=\mathbf{z}\} \\ &>E[M_v\{Y, f_{W0}(\boldsymbol{Z})\}|T=1, \boldsymbol{Z}=\mathbf{z}]=E[M_v\{Y, -f_{W0}(\boldsymbol{Z})\}|T=-1, \boldsymbol{Z}=\mathbf{z}] \quad (2) \end{aligned}$$

$$>E[M_v(Y,0)|T=-1, \boldsymbol{Z}=\mathbf{z}\}=E\{U(Y^{(-1)})|\boldsymbol{Z}=\mathbf{z}\}. \quad (3)$$

The inequalities in (2) and (3) follow from the fact that $M_v(y, v)$ is increasing in v (Condition A) and the equality in (2) is the consequence of the first order condition (1). Similarly, for a patient with a positive score (that is $f_{W0}(\mathbf{z}) > 0$), we have $E\{ U(Y^{(1)})|\mathbf{Z} = \mathbf{z}\} < E\{ U(Y^{(-1)})|\mathbf{Z} = \mathbf{z}\}$. Hence, $d_0(\mathbf{z}) = \text{sign}\{ f_{W0}(\mathbf{z})\}$ is an optimal ITR that maximizes $\mathbb{V}_U(d)$.

A-learning Method—We next demonstrate that the optimal ITR can be equivalently obtained via a different loss function constructed via A-learning ideas (Murphy, 2003; Robins, 2004; Lu et al., 2013; Ciarleglio et al., 2015). Specifically, consider the loss function $\ell_A(f) = E\{\ell_A(f, \mathbf{Z})\}$ and let $f_{A0} = \text{argmin}_f \ell_A(f)$, where

$$\begin{aligned} &\ell_A(f,\mathbf{z})=E(M[Y,\{(T+1)/2-\pi(\boldsymbol{Z})\}\times f(\boldsymbol{Z})]|\boldsymbol{Z}=\mathbf{z}) \\ &=E(I(T=1)M[Y,\{1-\pi(\boldsymbol{Z})\}f(\boldsymbol{Z})]|\boldsymbol{Z}=\mathbf{z})+E[I(T=-1)M\{Y,-\pi(\boldsymbol{Z})f(\boldsymbol{Z})\}|\boldsymbol{Z}=\mathbf{z}] \\ &=\pi(\mathbf{z})E(M[Y,\{1-\pi(\mathbf{z})\}f(\mathbf{z})]|T=1,\boldsymbol{Z}=\mathbf{z})+\{1-\pi(\mathbf{z})\}E[M\{Y,-\pi(\mathbf{z})f(\mathbf{z})\}|T=-1,\boldsymbol{Z}=\mathbf{z}]. \end{aligned}$$

Then for any $\mathbf{z}$ with $\pi(\mathbf{z}) \in (0, 1)$, the first order condition for f_{A0} is

$$E(M_v[Y,\{1-\pi(\mathbf{z})\}f_{A0}(\mathbf{z})]|T=1,\boldsymbol{Z}=\mathbf{z})=E[M_v\{Y,-\pi(\mathbf{z})f_{A0}(\mathbf{z})\}|T=-1,\boldsymbol{Z}=\mathbf{z}]. \quad (4)$$

Hence, for a patient with negative score (that is $f_{A0}(\mathbf{z}) < 0$), we can have

$$\begin{aligned} &E\{U(Y^{(1)})|\boldsymbol{Z}=\mathbf{z}\}=E\{M_v(Y,0)|T=1,\boldsymbol{Z}=\mathbf{z}\} \\ &>E(M_v\{Y,\{1-\pi(\boldsymbol{Z})\}f_{A0}(\boldsymbol{Z})\}|T=1,\boldsymbol{Z}=\mathbf{z}]=E[M_v\{Y,-\pi(\boldsymbol{Z})f_{A0}(\boldsymbol{Z})\}|T=-1,\boldsymbol{Z}=\mathbf{z}] \quad (5) \end{aligned}$$

$$>E\{M_v(Y,0)|T=-1,\boldsymbol{Z}=\mathbf{z}\}=E\{U(Y^{(-1)})|\boldsymbol{Z}=\mathbf{z}\}. \quad (6)$$

The inequalities in (5) and (6) follow from Condition A and the equality in (5) is from the first order condition (4). Similarly, for a patient with a positive score, we can have $E\{U(Y^{(1)})|\mathbf{Z}=\mathbf{z}\} < E\{U(Y^{(-1)})|\mathbf{Z}=\mathbf{z}\}$. Thus, the minimizer f_{A0} can also be used for subgroup identification with $d_0(\mathbf{Z}) = \text{sign}\{f_{A0}(\mathbf{Z})\}$ also maximizing the value function $\mathbb{V}_U(d)$.

2.3 Implementation

In this section, we provide some details on how to implement the proposed procedures in practice. Since most of the discussions apply for both the weighting and A-learning methods, we use ★ =W and A to index these two approaches respectively for conciseness. To approximate the minimizer $f_{\star 0}$ with observed data, one may first estimate the loss functions $\ell_\star(f)$ empirically. Specifically, it is straightforward to show that $\ell_W(f)$ and $\ell_A(f)$ can be respectively estimated by

$$L_W(f)=\frac{1}{n}\sum_{i=1}^{n}\frac{M\{Y_i,T_i f(\boldsymbol{Z}_i)\}}{T_i\pi(\boldsymbol{Z}_i)+(1-T_i)/2} \text{ and} \quad (7)$$

$$L_A(f)=\frac{1}{n}\sum_{i=1}^{n}M[Y_i,\{(T_i+1)/2-\pi(\boldsymbol{Z}_i)\}\times f(\boldsymbol{Z}_i)]. \quad (8)$$

Since the form of $f_{\star 0}$ is unknown, direct maximization of $L_\star(f)$ among all functional spaces is not feasible. In practice, model assumptions can be imposed to restrict the search space of $f_{\star 0}(\cdot)$. For example, a simple but useful approach is to assume that $f_{\star 0}(\cdot)$ can be approximated by a linear combination of a set of basis functions given *a priori*. That is, $f_{\star 0}(\mathbf{z}) \approx \sum_{k=1}^{K}\beta_k B_k(\mathbf{z})$, where $\{B_k(\mathbf{z}), k = 1, \cdots, K\}$ are K basis functions such as B-spline bases (Ruppert et al., 2003). One may then find $(\hat{\beta}_1, \cdots, \hat{\beta}_K)$ to minimize the loss function $L\{\sum_{k=1}^{K}\beta_k B_k(\cdot)\}$ or its penalized counterpart and let $\hat{f}(\cdot)=\sum_{k=1}^{K}\hat{\beta}_k B_K(\cdot)$ be the estimated benefit score. Alternatively, one may employ machine learning algorithms such as boosting to construct $\hat{f}(\cdot)$ based on $L_\star(f)$ (Hastie et al., 2009).

In many modern applications, the number of covariates is large but typically only a small subset is relevant to the treatment selection. Therefore, it is desirable to incorporate variable selection in subgroup identification using penalization approaches such as lasso (Hastie et al., 2009). For our proposed framework, it is easy to apply appropriate regularization to minimize the penalized loss function $L_\star(f) + \lambda(f)$ where the penalty term $\lambda(f)$ can be chosen to screen out noise features or encourage specific structure of the benefit score $f(\cdot)$.

It is also important to note that in observational studies, the propensity scores $\pi(\mathbf{Z}_i)$ are unknown and need to be replaced by their consistent estimators in constructing $L_\star(f)$. When $\mathbf{Z}$ is discrete or low dimensional, non-parametric estimators can be used for $\pi(\cdot)$. When the dimension of $\mathbf{Z}$ is not small, regression models such as logistic regression can be imposed for $\pi(\cdot)$.

To improve estimation efficiency, we may add possible augmentation to $M(\cdot, \cdot)$ function, while still preserve the same interpretation of obtained benefit scores. In Web Appendices A and B, we provide the justification and implementation for the efficiency augmentation.

2.4 Estimating the Magnitude of Individualized Treatment Effect

We have shown above that the benefit score, defined as the minimizer of the appropriately constructed loss function, can be used for subgroup identification since the sign of the score is consistent with the direction of the treatment effect. In this section, via several examples, we will demonstrate that often the value of the benefit score can also be used to approximate the size of the ITE. By choosing different M, the corresponding minimizers may reflect ITE quantified by different metrics. For example, when Y is non-negative, one may summarize the ITE given $\mathbf{Z}$ as $E(Y^{(1)}|\mathbf{Z}) - E(Y^{(-1)}|\mathbf{Z})$ or $E(Y^{(1)}|\mathbf{Z})/E(Y^{(-1)}|\mathbf{Z})$ (VanderWeele and Knol, 2014). Both metrics are widely used when investigating treatment covariate interactions and the preference of one over the other seems to be quite problem-specific (VanderWeele and Knol, 2014). In the traditional outcome prediction approach, one needs to employ seemingly different regression models to estimate such ITEs. Here, we will show that the ITEs under different metrics can be naturally unified under our proposal by considering different $M(\cdot, \cdot)$.

To this end, we first consider $M(y, v) = (y - v)^2$ and the corresponding propensity score weighted empirical loss function is

$$L_W(f) = \frac{1}{n}\sum_i \frac{\{Y_i - T_i f(\mathbf{Z}_i)\}^2}{T_i \pi(\mathbf{Z}_i) + (1 - T_i)/2}. \quad (9)$$

Assuming that f_{W0} minimizes $\ell_W(f) = E\{L_W(f)\}$, the first order condition given in (1) leads to $2 f_{W0}(\mathbf{z}) = E(Y^{(1)}|\mathbf{Z} = \mathbf{z}) - E(Y^{(-1)}|\mathbf{Z} = \mathbf{z})$. Hence once an estimator $\hat{f}(\cdot)$ of $f_{W0}(\cdot)$ is obtained, we can use $2\hat{f}(\mathbf{Z})$ to approximate the ITE. Similarly, we can consider the A-learning loss function corresponding to the quadratic loss,

$$L_A(f) = \frac{1}{n}\sum_i [Y_i - \{(T_i + 1)/2 - \pi(\mathbf{Z}_i)\} \times f(\mathbf{Z}_i)]^2. \quad (10)$$

The first order condition (4) implies that $f_{A0}(\cdot) = \text{argmin}_f E\{L_A(f)\} = E(Y^{(1)}|\mathbf{Z} = \mathbf{z}) - E(Y^{(-1)}|\mathbf{Z} = \mathbf{z})$. Therefore, we also can approximate ITE by constructing appropriate estimator for the minimizer of $E\{L_A(f)\}$. Thus, the quadratic loss $M(y, v) = (y - v)^2$

recovers treatment benefit scores that approximate treatment benefit measured by mean differences.

Next, we consider the exponential loss $M(y, v) = y\exp(-v)$. The corresponding empirical loss functions are

$$L_W(f)=\frac{1}{n}\sum_i \frac{Y_i e^{-T_i f(\boldsymbol{Z}_i)}}{T_1\pi(\boldsymbol{Z}_i)+(1-T_i)/2} \text{ and } L_A(f)=\frac{1}{n}\sum_i Y_i e^{-\{(T_i+1)/2-\pi(\boldsymbol{Z}_i)\}\times f(\boldsymbol{Z}_i)}.$$

Similarly, the first order conditions (1) and (4) imply that

$$\exp\{2f_{W0}(\boldsymbol{Z})\}=E(Y^{(1)}|\boldsymbol{Z})/E(Y^{(-1)}|\boldsymbol{Z}) \text{ and } \exp\{f_{A0}(\boldsymbol{Z})\}=E(Y^{(1)}|\boldsymbol{Z})/E(Y^{(-1)}|\boldsymbol{Z}),$$

respectively. Thus, the exponential loss leads to benefit scores that recover the ITE measured by the ratio of the expect outcomes under two different treatments.

3. A Review of Several Methods and Their Relationship with Our Framework

Tian et al. (2014) proposed a method for RCTs, which is a special case of our weighted loss function $L_W(f)$ in (7). Particularly, three different types of M, were described in their paper for continuous, binary, and survival type of outcomes, respectively. For continuous outcomes, $M(y, v) = (y - v)^2$. For binary outcomes, $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$. For survival outcomes

$$M(y,v)=-\left\{\int_0^\tau (v-\log[E\{e^v I(X\geq u)\}])\, dN(u)\right\},$$

where $y = (X, \delta) = \{\tilde{X} \wedge C, I(\tilde{X} \leq C)\}$, $\tilde{X}$ is the survival time, C is the censoring time, $N(t) = I(\tilde{X} \leq t)\delta$ and τ is a fixed point such that $P(X \geq \tau) > 0$. However, the interpretation of $U(y) = M_v(y, 0)$ is trickier due to the two-dimensional outcomes and Tian et al. (2014) proved that $U(y)$ is a monotone transformation of survival time $\tilde{X}$ given additional conditions. Besides, the optimal efficiency augmentation forms proposed by Tian et al. (2014) can also be viewed as special cases of our efficiency augmentation.

For $d(\mathbf{Z}) = \pm 1$, an outcome weighted estimator (OWE) finds the optimal decision rule by $d_{opt}(Z) = \text{argmin}_d E[\{T\pi(\mathbf{Z}) + (1 - T)/2\}^{-1} YI\{Td(\mathbf{Z}) < 0\}]$ (Qian and Murphy, 2011; Zhao et al., 2012; Zhang et al., 2012). However, since the 0–1 loss $I(v < 0)$ is neither convex nor continuous, it needs to be replaced by a convex and continuous surrogate loss operationally to overcome the computational obstacle, e.g., replacing $yI(v < 0)$ by $M(y, v) = y\phi(v)$. Xu et al. (2015) used the logistic loss function $\phi(v) = \log\{1 + \exp(-v)\}$ and Zhao et al. (2012) used the hinge function $\phi(v) = (1 - v)^+$, where $x^+ = \max(x, 0)$. With those surrogate loss functions, it is clear that the outcome weighted estimation procedure is equivalent to ours based on the loss function $L_W(f)$ with the corresponding $M(\cdot, \cdot)$. Although the aforementioned justification of our proposal is based on differentiable $M(\cdot, \cdot)$, we show that

it can be extended to non-differentiable hinge function for subgroup identification in Web Appendix C.

Moreover, negative outcomes may cause ill-behaved OWES. One way to deal with this problem is to shift all outcomes to positive values. However, the estimation efficiency may be compromised after such a shift. On the other hand, one may employ a flipping transformation: for negative outcome Y with treatment assignment T: we can change its outcome and treatment assignment to $-Y$ and $-T$, respectively. This flipping transformation does not change the 0–1 loss based on the original data, but in general affects the losses based on the surrogate function $\phi(v)$ with unclear consequences in final estimation. However, within the proposed framework, it is equivalent to using a flipping version of $M(\cdot, \cdot)$ function. For example, the flipping version for outcome-weighted logistic loss function $M(y, v) = y \log\{1 + \exp(-v)\}$ used by Xu et al. (2015) is $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ with $M_v(y, v) = -y[1 + \exp\{\text{sign}(y)v\}]^{-1}$. It is not hard to verify that this flipping version of $M(\cdot, \cdot)$ satisfies the two conditions mentioned in Section 2 and thus can be used to yield a valid estimator for the benefit score. In addition, we show that a doubly robust AIPWE estimator proposed by Zhang et al. (2012) can be obtained using a generalized augmented loss in Web Appendix F.

Lu et al. (2013) and Ciarleglio et al. (2015) proposed an A-learning estimator for the semiparametric outcome model: $E(Y|T, \mathbf{Z}) = m(\mathbf{Z}) + TC_G(\mathbf{Z}; \beta)$, where $m(\cdot)$ is unspecified and β is a finite dimensional vector. Let $f(\mathbf{Z}) = C_G(\mathbf{Z}; \beta)$, their proposed A-learning estimator is equivalent to minimizing our A-learning type loss function $L_A(f)$ with $M(\cdot, \cdot)$ being the squared loss $M(y, v) = (y - v)^2$. According to our justification, this A-learning method can also be extended to other $M(\cdot, \cdot)$, and we will illustrate the logistic loss in numerical studies.

4. Simulation

4.1 Continuous outcomes

We conducted extensive numerical studies with both continuous and binary outcomes. We generated a $p = 50$ dimensional covariate vector $\mathbf{Z} = (Z_1, \cdots, Z_p)'$ from a mean-zero multivariate normal distribution with variance 1 and covariance ρ, where ρ is set to be either 0 for the independent setting or 1/3 for the correlated setting. The treatment assignment T was generated from a simple logistic regression model $\text{logit}\{\pi(\mathbf{Z})\} = -1 + Z_1$. The outcome Y was simulated from nonlinear model

$$Y = \left(\beta_0 + \sum_{j=1}^{10} \beta_j Z_j\right)^2 + T\left(\gamma_0 + \sum_{j=1}^{4} \gamma_j Z_j + 0.8Z_1^2 + 0.8Z_2^2\right) + \varepsilon,$$

where $\varepsilon \sim N(0, 2)$ and $(\gamma_0, \cdots, \gamma_4) = (0.4, 0.8, -0.8, 0.8, -0.8)$. The coefficients for the main effects were set as either (i) $\beta_0 = 6^{-1/2}$, $\beta_1 = \beta_2 = 0$, $\beta_j = 0.5 \times 6^{-1/2}$, $j = 3, \cdots, 10$, representing moderate main effects; or (ii) $\beta_0 = 3^{-1/2}$, $\beta_1 = \beta_2 = 0$, $\beta_j = 0.5 \times 3^{-1/2}$, $j = 3, \cdots, 10$, representing large main effects. Throughout, we let the training sample size $n = 300$

and tested the performances of methods using an independently generated test data with a sample size of 10,000. For each simulation scenario, results were summarized based on 500 datasets. For all methods, we center the outcome Y by its sample average before model fitting.

We considered two functional classes of $f(\cdot)$ when minimizing the loss functions: (i) a linear model with $f_{\text{lin}}(\boldsymbol{Z})=\beta_0+\sum_{i=1}^{p}\beta_i Z_i$ where the lasso regularization was used to estimate the β_i's and the tuning parameter of lasso was chosen by 5-fold cross-validation (CV); (ii) an additive model with $f_{\text{add}}(\boldsymbol{Z})=\sum_{i=1}^{p} f_i(Z_i)$, where $f_i(\cdot)$, $i = 1, \cdots, p$, are nonlinear functions to be estimated. In fitting the additive model, we first screened the covariates by applying lasso regularization to the simple linear additive model. Then, the B-Spline method was implemented based on the selected variables (Ruppert et al., 2003). Operationally, we capped the maximum number of selected covariates in the first step to meet the requirement of R package mgcv, which was used to fit the additive model. The propensity score function $\pi(\cdot)$ was treated unknown and estimated by a fitting a lasso-regularized logistic regression with tuning parameter also selected via 5-fold CV.

We also considered various choices of $M(y, v)$. Specifically, we considered the following 7 methods: (1) **Full**: Full regression by regressing Y on $\mathbf{Z}$, $(T + 1)/2$ and $(T + 1)/2 \times \mathbf{Z}$ and then use the estimated treatment-covariate interaction terms to construct ITE; (2)$\mathbf{W}_{\text{sq–L}}$:Weighting method with the squared loss $M(y, v) = (y - v)^2$ and $f = f_{\text{lin}}$. This is a generalization of the modified covariate model proposed by Tian et al. (2014); (3) $\mathbf{W}_{\text{sq–A}}$: Weighting method with the squared loss $M(y, v) = (y - v)^2$ and $f = f_{\text{add}}$; (4) $\mathbf{W}_{\text{flo–L}}$: Weighting method with the flipping version of the outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and $f = f_{\text{lin}}$. This is a variant of the estimator proposed by Xu et al. (2015) using flipping transformation; (5) $\mathbf{A}_{\text{sq–L}}$: A-learning method with the squared loss $M(y, v) = (y - v)^2$ and $f = f_{\text{lin}}$. This is the A-learning estimator proposed in Lu et al. (2013); (6) $\mathbf{A}_{\text{sq–A}}$: A-learning method with the squared loss $M(y, v) = (y - v)^2$ and $f = f_{\text{add}}$; (7) $\mathbf{A}_{\text{flo–L}}$: A-learning method with the flipping version of the outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and $f = f_{\text{lin}}$. This is the extended version of the simple outcome-weighted logistic loss under the A-learning framework.

Figure 1 shows the boxplots for the rank correlation coefficients between the estimated scores and true treatment effects $\Delta(\mathbf{Z})$ in the test set. Higher rank correlation coefficients should indicate better performance. Here we used $\Delta(\mathbf{Z}) = E(Y^{(1)} - Y^{(-1)}|\mathbf{Z})$ as the ITE metric. We further evaluate performances regarding subgroup identification, that is in identifying the subgroup of patients $\{i|E(Y_i^{(1)}|\boldsymbol{Z}_i)>E(Y_i^{(-1)}|\boldsymbol{Z}_i)\}$. Figure 2 shows the average receiver operating characteristic (ROC) curves among 500 runs. The full regression has the worst performance among these methods, especially with correlated covariates, and other methods approximating the benefit score with a linear function have comparable performance. More flexible nonlinear additive models outperform their linear counterparts as expected. The A-learning method with the squared loss $M(y, v) = (y - v)^2$ performs slightly worse than the weighting method with correlated covariates, while A-learning and weighting with $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ have similar performances. When

there are big main effects, the performances of all methods become slightly worse than the scenarios with moderate main effects, especially when the covariates are correlated, likely due to the fact that main effect may mask the interactions of interest.

We also checked the performance of efficiency augmentation and possible influence of incorrect propensity score model. These additional results are in Web Appendix D.

4.2 Binary outcomes

For binary outcomes, we simulated the outcome by dichotomizing a continuous latent response:

$$Y=I\left\{(\beta_0+\sum_{j=1}^{10}\beta_j Z_j)^2+T(\gamma_0+\sum_{j=1}^{4}\gamma_j Z_j+0.8Z_1^2+0.8Z_2^2)+\varepsilon>0\right\},$$

and all other settings were the same as those for continuous outcomes. To improve estimation efficiency, we also subtracted 0.5 from all Ys before the analysis for flipping version of outcome-weighted logistic loss. This subtraction was used because when $y = 0$, $M(y, v) = 0$ for some choices of M. Although subjective, this shift was quite helpful for efficiency gain. Parallel to the settings with continuous outcome, we first implemented the traditional full logistic regression with both main effects and interaction effects. Secondly, we employed both weighting and A-learning methods with logistic likelihood $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ ($\mathbf{W}_{\text{lo}}$ and $\mathbf{A}_{\text{lo}}$) proposed by Tian et al. (2014), and the flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ ($\mathbf{W}_{\text{flo}}$ and $\mathbf{A}_{\text{flo}}$). The latter case was of particular interest since the shifted outcome may take negative values. The benefit score was approximated by either the simple linear or the nonparametric additive function. The lasso regularization was used for feature selection. Figures 3 and 4 show the corresponding results for rank coefficients and ROC curves, respectively. The nonlinear methods outperform their linear counterparts in terms of rank correlations with the underlying ITE but similarly based on ROC curves. On the other hand, the A-learning method seems to perform better than the weighting method in terms of the ROC curves but similarly based on rank correlations.

5. Real Data Example (Mammography Screening Study)

This is a randomized study for female participants who were non-adherent to mammography screening guidelines at the study baseline. One primary interest of the study was to compare the intervention effects of phone counseling on mammography screening (phone intervention) versus usual care at 21 months post-baseline. The outcome is whether the subject took mammography screening during this time period. We conduct outcome shift by subtracting 0.5 from all binary outcomes for flipping version of outcome weighted logistic loss. There are 530 subjects with 259 in the phone intervention group and 271 in the usual care group. 16 binary baseline covariates, including sociodemographics, health belief variables, and stage of readiness to undertake mammography screening, and 1 categorical variable, number of years had a mammogram in past 2 to 5 years, are available in the study.

Considering the covariates' first and second order interactions, there are 204 features in total.

To compare different methods, we randomly selected 80% participants and set the rest as a test set to evaluate the performance of the estimated benefit scores for ITEs. Specifically, following Xu et al. (2015), we evaluate the performance of a treatment recommendation rule $\hat{d}(Z)=\text{sign}\{\widehat{f_0}(Z)\}$ by the enhanced treatment effects $E[\Delta\{\mathbf{Z}, t, \hat{d}(\cdot)\}] = E(Y|\hat{d}(\mathbf{Z}) = t, T = t) - E(Y|\hat{d}(\mathbf{Z}) = t, T = -t)$, which can be estimated by the empirical weighted averages in the test set. This quantity measures the difference in the outcome between participants received the recommended intervention and those didn't. If both $E[\Delta\{\mathbf{Z}, 1, \hat{d}(\cdot)\}]$ and $E[\Delta\{\mathbf{Z}, -1, \hat{d}(\cdot)\}]$ are positive, then the benefit score-based recommendation of the intervention is helpful for the participants in the study population. When coupled with the inverse probability weighting technique, the enhanced treatment effects score is still a valid measure when the data are from observational study. The procedures were repeated for 200 random splits and the mean enhanced treatment effects (and estimated standard errors from these 200 splits) for different methods are reported in Table 1, where larger average enhanced treatment effect indicates better performance of the estimator. The benefit scores are approximated by simple linear functions since most features are binary. The full regression performs the worst among all methods, and the flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ seems slightly better than the logistic likelihood loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ under the same setting, while comparable with the efficiency-augmented logistic likelihood loss. Additional real data analysis for a national supported work study can be found in Web Appendix E.

6. Discussion

In this article, we proposed a flexible framework for treatment scoring in both observational studies and RCTs, based on weighting and A-learning methods. The proposed methods are quite flexible and many recently proposed estimators can be represented as special cases within our frameworks.

A very practical issue of applying our proposal is the choice of the $M(\cdot, \cdot)$ and $f(\cdot)$ functions. As we demonstrate in Section 2.4, the choices may depend on the preference of the ITE metrics. For example, if ITE is quantified by $E(Y^{(1)}|\mathbf{Z})/E(Y^{(-1)}|\mathbf{Z})$ instead of $E(Y^{(1)}|\mathbf{Z}) - E(Y^{(-1)}|\mathbf{Z})$, suitable M needs to be constructed accordingly. We also note that different M can have equivalent $U(Y)$ and ITR. For example, $M(y, v) = (y - v)^2$ and $M(y, v) = y \log\{1 + \exp(-v)\}$ both have $U(Y)$ as a linear transformation of Y, and therefore should identify the same subgroup. However, they can have fairly different performances in finite-sample studies as demonstrated in numerical studies.

When specifying the class of functions for f, one also needs to balance the bias variance tradeoff. A simple linear form may be appropriate for a specific data with one type of ITE while nonlinear bases functions might be needed to adequately approximate the ITE for other cases. With sufficiently large sample size, one may use cross-validation or sample split to select an optimal set of basis for a given dataset with a given M. When the number of covariates or the number of basis functions is large, one may overcome overfitting by

employing popular lasso or elastic net regularization to help with variable selection and stabilize model fitting.

Suitable efficiency augmentation such as outcome shift may help to reduce the variability and enhance the robustness of relevant estimators. Thus it is crucial to withhold an independent test set to objectively examine the performance of resulting ITR estimators based on different combinations of M, f, regularization procedure and efficiency augmentation method (Zhao et al., 2013). An appropriate choice of the estimation procedure may be made by considering the complexity, clinical interpretability and computational cost associated with the estimation.

When the models for the propensity and/or f are misspecified, the rank of the estimated benefit scores can still be informative. To account for mis-specification in f, one may nonparametrically calibrate the treatment effect estimator according to the rank of the scores as in Cai et al. (2011) and construct the corresponding ITR. Mis-specification in the propensity score may lead to sub-optimal estimation of the benefit scores, but one may use such scoring systems for future clinical trials to more accurately determine ITR.

Although our framework described in earlier sections focuses on binary treatments, the generalization to multiple treatments is feasible. In Web Appendix G, we extend our framework to multiple treatments with an additional assumption on M using the weighting method. Tao and Wang (2016) proposed a method for multi-treatment selection by generalizing the OWE to settings with more than two treatment arms. Their multi-treatment OWE could be viewed as a special case of our generalized framework for multiple treatments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1409-21219 for Chen and Yu) and by grants from the National Institute of Health (R01 HL089778-05 and U54 HG007963 for Cai and Tian). The views in this publication do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

References

Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12:270–282. [PubMed: 20876663]

Ciarleglio A, Petkova E, Ogden RT, Tarpey T. Treatment decisions based on scalar and functional baseline covariates. Biometrics. 2015; 71:884–94. [PubMed: 26111145]

Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in medicine. 2011; 30:2867–2880. [PubMed: 21815180]

Gabriel S, Normand S. Getting the methods right–the foundation of patient-centered outcomes research. N Engl J Med. 2012; 367:787–90. [PubMed: 22830434]

Hastie, TJ., Tibshirani, RJ., Friedman, JH. Springer series in statistics. Springer; New York: 2009. The elements of statistical learning: data mining, inference, and prediction.

Kraemer HC. Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. Statistics in Medicine. 2013; 32:1964–1973. [PubMed: 23303653]

Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Statistics in medicine. 2015; 34:1818–1833. [PubMed: 25656439]

Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. Statistical methods in medical research. 2013; 22:493–504. [PubMed: 22116341]

McKeague IW, Qian M. Estimation of treatment policies based on functional predictors. Statistica Sinica. 2014; 24:1461–1485. [PubMed: 25165416]

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

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

Robins, JM. Optimal structural nested models for optimal sequential decisions. Proceedings of the second seattle Symposium in Biostatistics; Springer; 2004. p. 189-326.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.

Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association. 2005; 100:322–331.

Ruppert, D., Wand, M., Carroll, R. Semiparametric Regression. 1. Cambridge University Press; 2003.

Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. The international journal of biostatistics. 2008; 4:2.

Tao Y, Wang L. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. Biometrics. 2016; doi: 10.1111/biom.12539

Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association. 2014; 109:1517–1532. [PubMed: 25729117]

VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiologic Methods. 2014; 3:33–72.

Vansteelandt S, VanderWeele TJ, Tchetgen EJ, Robins JM. Multiply robust inference for statistical interactions. Journal of the American Statistical Association. 2008; 103:1693–1704. [PubMed: 21603124]

Xu Y, Yu M, Zhao YQ, Li Q, Wang S, Shao J. Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics. 2015; 71:645–53. [PubMed: 25962845]

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012; 1:103–114. [PubMed: 23645940]

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

Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association. 2013; 108:527–539. [PubMed: 24058223]

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

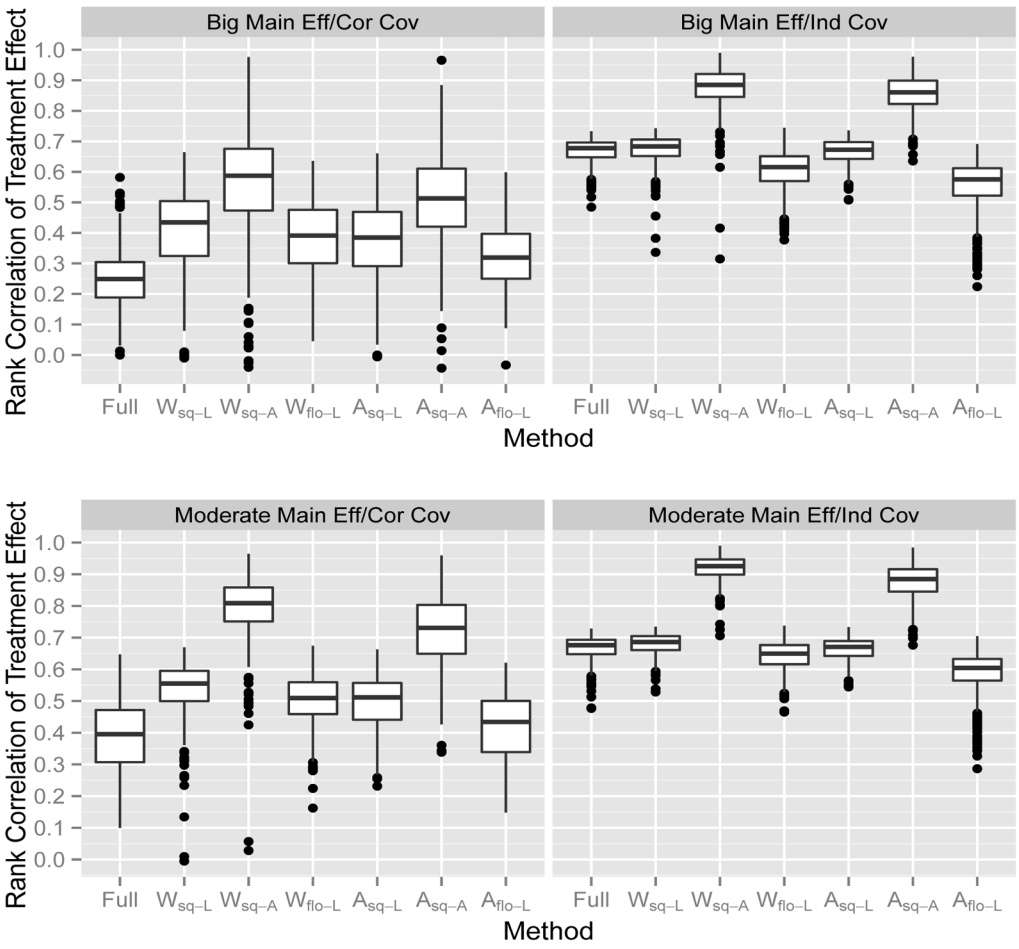

Figure 1.

Boxplots for the rank correlation coefficients between the estimated benefit scores and true treatment effects for continuous outcomes. Method "Full" uses the full regression; Method "W_{sq-L}" uses the weighting method with squared loss $M(y, v) = (y - v)^2$ and a linear f; Method "W_{sq-A}" uses the weighting method with squared loss $M(y, v) = (y - v)^2$ and a nonparametric additive f; Method "W_{flo-L}" uses the weighting method with flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f; Method "A_{sq-L}" uses the A-learning method with $M(y, v) = (y - v)^2$ and a linear f; Method "A_{sq-A}" uses the A-learning method with $M(y, v) = (y - v)^2$ and a nonparametric additive f; Method "A_{flo-L}" uses the A-learning method with flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f.

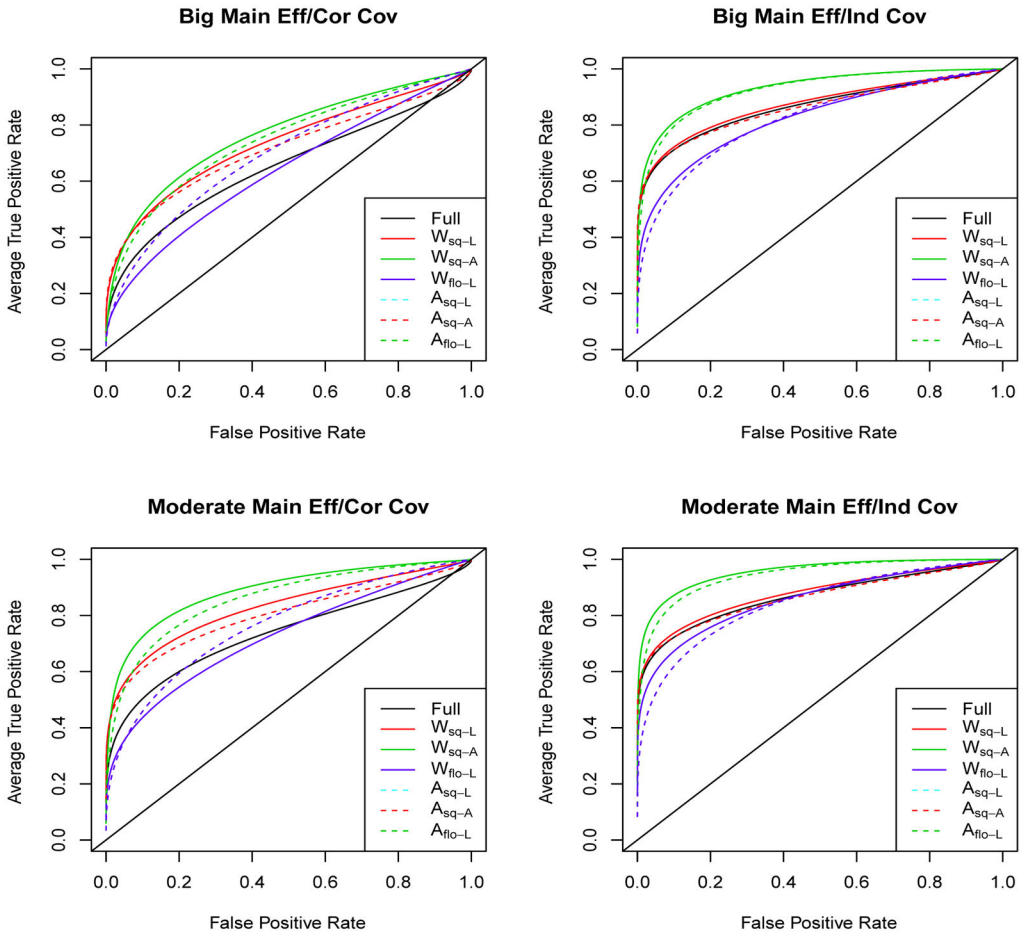

Figure 2.

ROC curves of estimated benefit scores for subgroup identification when the outcomes are continuous. Method "Full" uses the full regression; Method "W_{sq-L}" uses the weighting method with squared loss $M(y, v) = (y - v)^2$ and a linear f;Method "W_{sq-A}" uses the weighting method with squared loss $M(y, v) = (y - v)^2$ and a nonparametric additive f;Method "W_{flo-L}" uses the weighting method with flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f; Method "A_{sq-L}" uses the A-learning method with $M(y, v) = (y - v)^2$ and a linear f; Method "A_{sq-A}" uses the A-learning method with $M(y, v) = (y - v)^2$ and a nonparametric additive f; Method "A_{flo-L}" uses the A-learning method with flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f.

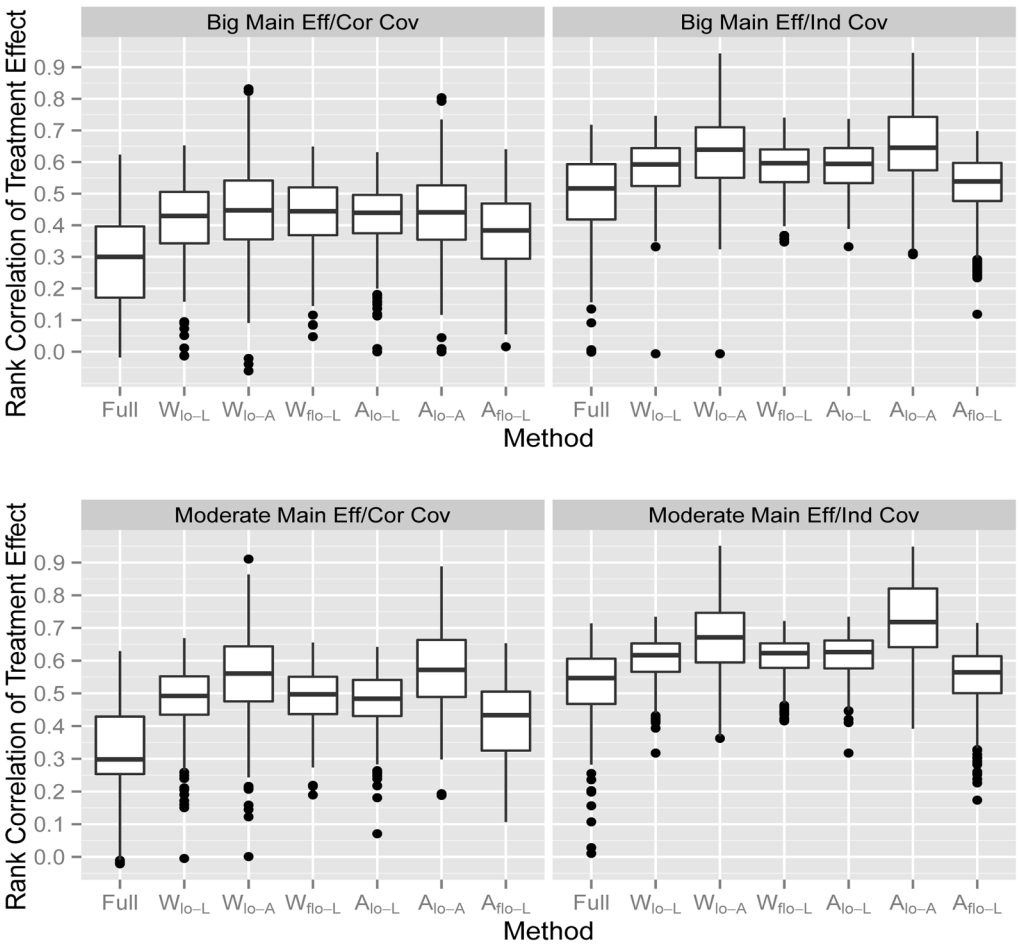

Figure 3.

Boxplots for the rank correlation coefficients between the estimated benefit scores and true treatment effects for binary outcomes. Method "Full" uses the full logistic regression; Method "$W_{lo–L}$" uses the weighting method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a linear f; Method "$W_{lo–A}$" uses the weighting method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a nonparametric additive f;Method "$W_{flo–L}$" uses the weighting method with flipping outcome weighted logistic loss $M(y, v) = |y| \log[1+\exp\{-\text{sign}(y)v\}]$ and a linear f;Method "$A_{lo–L}$" uses the A-learning method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a linear f; Method "$A_{lo–A}$" uses the A-learning method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a nonparametric additive f;Method "$A_{flo–L}$" uses the A-learning method with flipping outcome weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f.

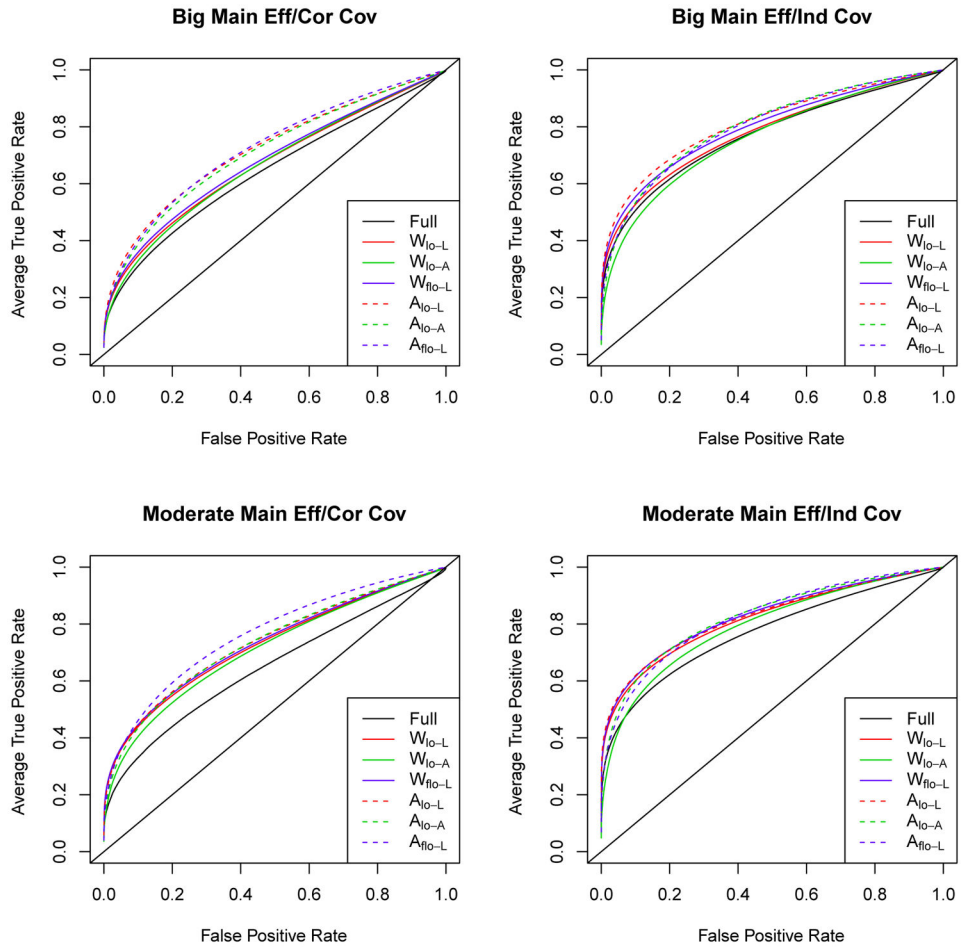

Figure 4.

ROC curves of estimated benefit scores for subgroup identification when outcomes are binary. Method "Full" uses the full logistic regression; Method "W_{lo-L}" uses the weighting method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a linear f; Method "W_{lo-A}" uses the weighting method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a nonparametric additive f; Method "W_{flo-L}" uses the weighting method with flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f; Method "A_{lo-L}" uses the A-learning method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a linear f; Method "A_{lo-A}" uses the A-learning method with logistic loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$ and a nonparametric additive f; Method "A_{flo-L}" uses the A-learning method with flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$ and a linear f.

Table 1

The average estimated enhanced comparative treatment effect (standard errors) and average subgroup sizes (proportions) in test data based on 200 random splits of mammography screening study

	$t = 1$		$t = -1$	
Method	**Mean (SE)**	**Subgroup Size**	**Mean (SE)**	**Subgroup Size**
Full	−0.005 (0.009)	62 (59%)	0.039 (0.011)	44 (41%)
W_{lo-L}	0.008 (0.009)	53 (50%)	0.058 (0.011)	53 (50%)
W_{loE-L}	0.020 (0.010)	49 (46%)	0.073 (0.010)	57 (54%)
W_{flo-L}	0.014 (0.010)	52 (49%)	0.070 (0.010)	54 (51%)
A_{lo-L}	0.002 (0.009)	52 (49%)	0.056 (0.011)	54 (51%)
A_{loE-L}	0.021 (0.010)	49 (47%)	0.072 (0.010)	57 (53%)
A_{flo-L}	0.012 (0.010)	52 (49%)	0.069 (0.011)	54 (51%)

NOTE: “Full” uses full logistic regression; “W_{lo-L}” uses weighted logistic likelihood loss $M(y, v) = -[yv - \log\{1 + \exp(v)\}]$; “$W_{loE-L}$” uses weighted efficiency-augmented logistic likelihood loss; “W_{flo-L}” uses weighted flipping outcome-weighted logistic loss $M(y, v) = |y| \log[1 + \exp\{-\text{sign}(y)v\}]$; “$A_{lo-L}$” uses A-learning method with logistic likelihood loss; “A_{loE-L}” uses A-learning method with efficiency-augmented logistic likelihood loss; “W_{flo-L}” uses A-learning method with flipping outcome-weighted logistic loss.