2013

A data-adaptive strategy for inverse weighted estimation of causal effects

Yeying Zhu, *Penn State University*
Debashis Ghosh, *Penn State University*
Bhramar Mukherjee, *University of Michigan - Ann Arbor*
Nandita Mitra, *University of Pennsylvania*

A data-adaptive strategy for inverse weighted estimation of causal effects

Yeying Zhu[1], Debashis Ghosh[1], Nandita Mitra[2] and Bhramar Mukherjee[3]

[1] Department of Statistics, Penn State University, [2] Department of Biostatistics and Epidemiology, University of Pennsylvania and [3]Department of Biostatistics, University of Michigan

Abstract

In most nonrandomized observational studies, differences between treatment groups may arise not only due to the treatment but also because of the effect of confounders. Therefore, causal inference regarding the treatment effect is not as straightforward as in a randomized trial. To adjust for confounding due to measured covariates, the average treatment effect is often estimated by using propensity scores. In this article, we focus on the use of inverse probability weighted (IPW) estimation methods. Typically, propensity scores are estimated by logistic regression. More recent suggestions have been to employ nonparametric classification algorithms from machine learning. In this article, we propose a weighted estimator combining parametric and nonparametric models. Some theoretical results regarding consistency of the procedure are given. Simulation studies are used to assess the performance of the newly proposed methods relative to existing methods, and a data analysis example from the Surveillance, Epidemiology and End Results (SEER) database is presented.

Key words and phrases: Boosting algorithms; Causal inference; Logistic regression; Observational data; Random forests.

1 Introduction

In many medical and scientific studies, investigators are interested in making causal statements about the effect of a treatment on outcomes. For a well-designed randomized study, we assume that any covariates that may influence the outcome are distributed the same among different treatment groups. Consequently, the treatment is the only factor that may cause differences in the outcome. However, in an observational study, where the treatment assignment is not controlled by a randomization scheme, there usually exists a set of confounders that may influence both the outcome and the treatment assignment. In this case, any causal inference failing to account for the confounders will lead to biased estimates of the treatment effect.

In the recent clinical literature, a common framework for assessing causal effects of the treatment effect on the response is based on the potential outcomes advocated by Rubin (1974) and Rosenbaum and Rubin (1983). In the latter paper, they proposed the concept of propensity scores, which is defined as the probability of receiving the treatment given the covariates. They further demonstrated that conditioning on the propensity score, the observed outcomes from an observational study can be viewed as coming from a randomized study. However, to claim this, one needs to assume that all confounders in the study are measured. In practice, sensitivity analyses are often conducted to investigate how unmeasured covariates could affect the inferred causal effect.

There are a variety of approaches to adjusting for the propensity score, summarized nicely in an overview by Lunceford and Davidian (2004). In this article, we focus on the use of inverse probability weighted (IPW) estimation, which we define in Section 2. While this has been a popular approach to the estimation of causal effects, Kang and Schafer (2007) argued against the use of such methods due to the fact that causal effects estimated using IPW were sensitive to observations with large weights. However, most of these observations/weights are informative to the analysis, so they cannot be completely discarded. For example, if a treated subject has a low

propensity score, the observed outcome of this subject is highly informative about the missing potential outcome for those in the control (untreated) group. An open question is how to deal with extreme weights in IPW estimation procedures.

Another theme addressed in the article is the choice of modeling procedure for propensity scores. It is obvious that different methods for estimating the propensity score will lead to different estimates of the treatment effect. In the statistical literature, propensity scores have been typically estimated by logistic regression. Recently, several studies employed machine learning methods as an alternative to logistic regression for modeling propensity scores (Setoguchi et al., 2008; Lee et al., 2010; McCaffrey et al., 2004).

In this work, we propose a class of weighted propensity score estimators that combine both parametric and nonparametric estimators. Unlike previous work in the area (Setoguchi et al., 2008; Lee et al., 2010; McCaffrey et al., 2004), we are able to prove some theoretical properties of these estimators. These estimators have the overall effect of shrinking extreme weights in the IPW estimation procedure, which leads to better finite-sample performance of the average causal effect estimators; this will be seen through a sequence of simulation studies in Section 5. To our knowledge, this is the first time that the estimation of propensity scores has been proposed in this manner, although similar ideas for hybrid or weighted estimation have appeared in other research areas. For example, Olkin and Spiegelman (1987) proposed a semiparametric approach to density estimation which combines the parametric maximum likelihood estimator and the nonparametric kernel estimator. Kouassi and Singh (1997) developed a semiparametric estimator for the hazard function with randomly censored survival data. In their example, they combined the estimator of a Weibull parametric model and kernel hazard estimator. Nottingham and Birch (2000) extended the seimparametric approach to quantal dose-response data which combines a quantal parametric model with a local linear regression estimator. Finally, inspired by a regression example where a parametric model is insufficient to fit the entire data set, Mays et al. (2001) introduced several semiparametric approaches to improve the

fit, one of which combines the regression fit of ordinary least squares and the fit of local linear regression. In all the above-mentioned approaches, the proposed estimators can be written in a unified manner. Assume the parametric estimator for the targeted estimand is denoted as y^P and the corresponding nonparametric estimator is y^{NP}, the semiparametric estimator can be written as:

$$y^{SP} = \lambda y^P + (1 - \lambda) y^{NP}, \tag{1}$$

where λ is a smoothing parameter estimated from the observed data. While our proposed estimator also combines parametric and nonparameric estimates, our approach is essentially different from the previous literature. First of all, the previous approaches are all one-stage methods, while we are focusing on a two-stage procedure. That is, we apply the combined estimator at stage one (the estimation of propensity scores) and improve IPW to estimate the treatment effect at stage two. We wish to understand the properties of the second-stage estimator. In the previous literature, the smoothing parameter λ in (1) can be regarded as the weight placed on the parametric estimator, which is estimated by minimizing/maximizing a certain objective function, such as MSE or PRESS of y^{SP}. This idea is not directly applicable to the modeling of propensity scores because propensity scores are nuisance parameters and a prediction model with improved MSE for propensity scores does not necessarily lead to better causal inference in the second stage (Lunceford and Davidian, 2004; Setoguchi et al., 2008). In our proposed methods, weights are data-adaptive and locally calculated. Essentially, we are able to shrink extreme weights to more reasonable values so the bias and variance of the inverse weighted estimation of causal effects can be reduced. Thirdly, while the nonparametic component in the above-mentioned literature is estimated by kernel estimator or local linear/polynomial estimator, the nonparametric component in our approach is estimated by one of the machine learning algorithms, such as tree-based classifiers.

The layout of the paper is as follows. In Section 2, we review the potential outcomes framework. In Section 3, we describe two classes of methods for estimating

propensity scores, parametric methods and nonparametric ones. We then propose a data-adaptive approach to estimate propensity scores, which is described in Section 4. There, we also present some consistency results. In Section 5, we present a simulation study and show that the newly proposed method is superior in terms of reducing bias and variance of the causal effect estimates. Also, we demonstrate how the proposed procedure demonstrates a statistically principled approach to downweighting extreme obserations in IPW estimation procedures. In Section 6, we illustrate our method by comparing treatments for cholangiocarcinomas, a cancer of the bile ducts using data collected through the Surveillance, Epidemiology and End Results (SEER) database.

2 A Review of the Potential Outcomes Framework

Let Y denote the response of interest and $\mathbf{X}$ be a p-dimensional vector of covariates. Let Z be a binary indicator of treatment exposure. We assume that Z takes the values $\{0, 1\}$: $Z = 1$ if treated, $Z = 0$ if control. Let the observed data be represented as $(Y_i, \mathbf{X}_i, Z_i)$, $i = 1, \ldots, n$, a random sample from $(Y, \mathbf{X}, Z)$.

When Y refers to the outcome under the receipt of a certain level of the treatment, we further define $\{Y(0), Y(1)\}$ to be the potential outcomes for subject i if control or treated. What we observe is $Y_i = Y_i(Z_i)$ $(i = 1, \ldots, n)$, which implies that $Y(0)$ and $Y(1)$ can not be observed simultaneously, i.e. one of them is missing. The relationship between Y and $\{Y(0), Y(1)\}$ can be summarized as

$$Y = Y(1) \times Z + Y(0) \times (1 - Z).$$

Two possible parameters of interest are the average causal effect:

$$ACE = E[Y(1) - Y(0)], \tag{2}$$

and the average causal effect among the treated:

$$ACET = E[Y(1) - Y(0)|Z = 1]. \tag{3}$$

ACET is of particular interest when the population of the study are those who actually receive the treatment. For example, a researcher from a smoking cessation counseling

tries to persuade the smokers to quit smoking and his research question is as follows: for those who actually smoke, what is the difference in the expected life expectancy if they did not smoke? In this example, the researcher is interested in estimating ACET.

In a randomized study, the treatment assignment is completely determined by randomization. Therefore, we have $Z \perp \{Y(0), Y(1)\}$. Consequently, an unbiased estimator for ACE (and ACET) is given by

$$\widehat{ACE} = \frac{\sum_{i=1}^{n} Y_i Z_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n} Y_i (1 - Z_i)}{\sum_{i=1}^{n} (1 - Z_i)}.$$

In an observational study, the vector of covariates $\mathbf{X}$ could be related to both the outcome and the treatment assignment. Since both Z and the potential outcomes $\{\mathrm{Y}(0),\mathrm{Y}(1)\}$ are affected by $\mathbf{X}$, $Z \perp \{Y(0), Y(1)\}$ will not hold. We refer to $\mathbf{X}$ as the set of confounders. For example, in a study of survival time among patients with cholangiocarcinomas, a patient's age will affect his/her choice of the treatment type due to economic or physical constraints. Meanwhile, it will also affect the patient's survival time. Therefore, age is often treated as a confounder in such studies with survival outcomes.

An important assumption made by Rosenbaum and Rubin (1983) for non-randomized observational studies is termed strongly ignorable treatment assignment:

$$Z \perp \{Y(0), Y(1)\} | \mathbf{X}. \tag{4}$$

This assumption says that treatment assignment is conditionally independent of the set of potential outcomes given the covariates. In other words, conditioning on the same value of $\mathbf{X}$, we can pretend that the observed outcomes are from a randomized trial. However, conditioning on a p-dimensional vector suffers from the "curse of dimensionality", especially when the dimension is high. Rosenbaum and Rubin (1983) further proposed the concept of propensity scores, which is defined as the probability of receiving the treatment given the covariates:

$$e(\mathbf{X}) \equiv P(Z = 1 | \mathbf{X})$$

Rosenbaum and Rubin (1983) show that if (4) holds, the following property is also true:

$$Z \perp \{Y(0), Y(1)\} | e(\mathbf{X}). \tag{5}$$

The fact that $e(\mathbf{X})$ is a scalar quantity greatly facilitates causal inference because we only need to condition on a single random variable.

Based on the assumption of (5), causal inference is a two-stage modeling process. In the first stage, the propensity score is estimated as a function of covariates. In the second stage, ACE in (2) or ACET in (3) is estimated as a function of the treatment indicator (sometimes with other covariates), adjusted by the propensity score. Lunceford and Davidian (2004) presented a thorough review of the methods for propensity score adjustment; here, we focus on using IPW estimation. The estimators for the ACE and ACET are given by

$$\widehat{ACE} = \frac{\sum_{i=1}^{n} Y_i Z_i / \hat{e}(\mathbf{X}_i)}{\sum_{i=1}^{n} Z_i / \hat{e}(\mathbf{X}_i)} - \frac{\sum_{i=1}^{n} Y_i (1 - Z_i)/(1 - \hat{e}(\mathbf{X}_i))}{\sum_{i=1}^{n} (1 - Z_i)/(1 - \hat{e}(\mathbf{X}_i))} \tag{6}$$

and

$$\widehat{ACET} = \frac{\sum_{i=1}^{n} Y_i Z_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n} Y_i (1 - Z_i) \hat{e}(\mathbf{X}_i)/(1 - \hat{e}(\mathbf{X}_i)}{\sum_{i=1}^{n} (1 - Z_i) \hat{e}(\mathbf{X}_i)/(1 - \hat{e}(\mathbf{X}_i))}, \tag{7}$$

respectively. We will also refer to the first stage and second stage throughout the paper.

Since the true values of propensity scores are unknown, it is necessary to estimate them in the first stage. The estimation of propensity scores sometimes involves a high-dimensional vector of covariates. Traditionally, this is done by logistic regression. Recently, machine learning methods have been proposed to estimate the propensity scores, such as classification trees or generalized boosted regression (Setoguchi et al., 2008; Lee et al., 2010; McCaffrey et al., 2004). Many simulation studies have shown that different estimation methods employed in the first stage will affect the finite-sample properties of the estimated treatment effect in the second stage. For example, Lee et al. (2010) show that when there is a moderate misspecification in the logistic regression model, ensemble machine learning methods (random forests and

generalized boosted regression) yield smaller bias and variance, more consistent 95% confidence interval coverage.

3 Propensity Score Modeling

3.1 The traditional method: logistic regression

In much of the literature, the estimation of the propensity score is usually done by logistic regression. Logistic regression assumes that

$$e(\mathbf{X}) \equiv e(\mathbf{X}, \beta) = \frac{1}{1 + \exp\{-\mathbf{X}^T\beta\}}. \tag{8}$$

The estimation of β is achieved by maximum likelihood. Notice that we ignore the information of Y in the first stage. Using logistic regression to estimate propensity score can be achieved by almost any statistical software. However, logistic regression is not without drawbacks. First of all, we specify a parametric form of $e(X)$ in (8). In most cases, only including main effects into the model is not adequate, but it is also hard to determine which interaction terms should be included, especially when the vector of covariates is high-dimensional. Finally, logistic regression is not resistant to outliers (Kang and Schafer, 2007; Pregibon, 1982). In particular, Kang and Schafer (2007) shows when the logistic regression model is misspecified, IPW leads to large bias.

In the simulation study in Section 5, we also show the performance of another parametric approach for estimating propensity scores: linear discriminant analysis (LDA). In LDA, we assume

$$e(\mathbf{X}) \equiv e(\mathbf{X}, \mu_{\mathbf{0}}, \mu_{\mathbf{1}}, \Sigma) = \frac{f_1(\mathbf{X})\pi_1}{f_0(\mathbf{X})\pi_0 + f_1(\mathbf{X})\pi_1},$$

where $f_i(\mathbf{X})$ is the posterior probability density function for class i and follows a multivariate normal distribution:

$$f_i(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(\mathbf{X}-\mu_i)^T\Sigma^{-1}(\mathbf{X}-\mu_i)/2}, \quad i = 0, 1.$$

In practice, the mean vectors μ_0, μ_1 and the common covariance matrix Σ, as well as the prior class probabilities π_0, π_1 are estimated from training data using sample

statistics. Unlike logistic regression, LDA requires the normal assumption. However, Hastie et al. (2009) claim that in most situations, the two models give very similar results, which is also the case in our simulation study.

3.2 Machine learning techniques

If we regard the estimation of propensity scores as a classification problem, we may extend our methods to the nonparametric machine learning realm. These methods include classification and regression trees (CART) and its various extensions, such as pruned CART, bagged CART, random forests and boosting. Other methods include support vector machines (SVM) and K-nearest neighbors (KNN).

Except for SVM and KNN, all of the above-mentioned algorithms are based on the construction of tree classifiers. In general, a tree classifier works as follows: beginning with a training data set $(\mathbf{X}_i, Z_i)$, $i = 1, \ldots, n$, a tree classifier repeatedly splits nodes based on one of the covariates in $\mathbf{X}$, until it stops splitting by some stopping criteria (for example, the terminal node only contains training data from one class). Each terminal node is then assigned a class label by the majority of Z_i that falls in that terminal node. Once a testing data point with a covariate vector $\mathbf{X}$ is introduced, the data point is run from the top of the tree until it reaches one of the terminal nodes. The prediction is made by the class label of that terminal node. Compared to parametric algorithms, tree-based algorithms have several advantages. First of all, there is no need to assume any parametric model for a tree: in constructing a tree, the algorithms only need to determine the criterion for splitting a node, and when to stop splitting (Breiman et al., 1984). By splitting a tree based on different covariates at different nodes, the algorithm automatically includes interaction terms in the model. Second, because the algorithm is nonparametric, the tree classifier can pick important covariates (in a stepwise manner) even when $\mathbf{X}$ is high dimensional or most of the covariates are highly correlated (McCaffrey et al., 2004). Moreover, a standard tree classifier is usually very fast to fit and robust to outliers (Breiman et al., 1984).

One of the biggest issues of a standard tree classifier is its tendency to overfit. That is, the constructed tree is usually too adaptive to the training data, and hence yields high prediction errors for testing data. To solve the over-fitting problem, pruned CART was proposed (Breiman et al., 1984). In pruned CART, a tree is fully grown and then pruned back until some stopping criteria are met. For example, the cross-validation error rate of the pruned tree reaches the minimum. Compared to a standard tree classifier, the pruned tree is smaller in size and yields lower prediction errors.

Another class of tree-based algorithms is called random forests, which was first introduced by Breiman (2001). It belongs to the category of so-called ensemble methods: instead of generating one classification tree, it generates many trees. At each node of a tree, a random subset of the covariates are selected and the node is split based on the best split among the selected covariates. For a testing data point with a covariate vector $\mathbf{X}$, each tree votes for one of the classes and the prediction can be made by the majority votes among the trees. In the first stage of causal inference, if we apply random forests algorithm, the propensity score could be estimated as the proportion of trees that vote for class 1. Biau et al. (2008) proved the consistency of random forests estimator in terms of predicting the class label. In that paper, they also commented that random forests are among the most accurate general-purpose classifiers available.

Bagged CART (Breiman, 1996) also belongs to the category of ensemble tree classifiers. In this algorithm, a bootstrap sample of the original training data is generated with replacement for multiple times and each bootstrap sample produces one classification tree. The bootstrap sample size is usually taken to be the same as the original data set. For a testing data point with a covariate vector $\mathbf{X}$, the propensity score can be estimated in the same way as in random forests.

Boosting is another class of algorithms that represents ensemble classifiers. Instead of taking a bootstrap sample from the training data with equal probabilities each time, the AdaBoost algorithm (Freund and Schapire, 1997) suggests giving more weights to observations that were misclassified more often by the previous trees. Each tree is a

weak classifier and the final classifier is a weighted average of all the trees. Generalized boosted model (GBM) (Ridgeway, 1999) is an extension of boosting which can directly produce estimates of propensity scores. In GBMs, let $g(\mathbf{X}) = \log[e(\mathbf{X})/(1 - e(\mathbf{X}))]$ and the maximum likelihood function is:

$$l(g) = \sum_{i=1}^{n} Z_i g(\mathbf{X}_i) - \log\{1 + \exp[g(\mathbf{X}_i)]\}. \tag{9}$$

To maximize $l(g)$ in (9), $g(\mathbf{X})$ is updated at each iteration with $g(\mathbf{X}) + h(\mathbf{X})$ where $h(\mathbf{X})$ is the fitted value from a regression tree which models $\gamma_i = Z_i - 1/\{1 + \exp[-g(\mathbf{X}_i)]\}$, the largest increase in (9). McCaffrey et al. (2004) provides a detailed algorithm for estimating propensity scores using GBM.

Support vector machines (SVM) is also a nonparametric classification method. The objective of SVM is to find the optimal hyperplane that maximizes the margin between two classes (Cortes and Vapnik, 1995). If we recode Z as -1 and 1 ($\tilde{Z} = 2Z - 1$), the optimization problem is equivalent to the maximization of the following objective function:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} \tilde{Z}_i \tilde{Z}_{i'} K(\mathbf{X}_i, \mathbf{X}_{i'}),$$

subject to $\alpha_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} \alpha_i \tilde{Z}_i = 0$. K is the kernel function that projects original data to higher dimensions and the simplest kernel is $K(\mathbf{X}_i, \mathbf{X}_{i'}) = \mathbf{X}_i \cdot \mathbf{X}_{i'}$. The above algorithm is used when misclassification in the training data is not allowed. Cortes and Vapnik (1995) extended SVM to allow misclassification in the training data, and the objective function becomes:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} \tilde{Z}_i \tilde{Z}_{i'} (K(\mathbf{X}_i, \mathbf{X}_{i'}) + \frac{1}{C} \delta_{ii'}),$$

where $\delta_{ii'} = 1$ if $i = i'$ and 0 otherwise, and $C > 0$ is a user-selected constant allowing for misclassifications. SVM does not estimate propensity scores directly but by fitting a logistic regression model to $(\mathbf{X}_i, \hat{Z}_i), i = 1, \ldots, n$, where $\hat{Z}_i$ is the estimated class label by SVM.

One of the simplest nonparametric classification algorithms is called K-nearest neighbors. It works as follows: for a testing data point, finding the K nearest points

in the training set in terms of some distance measure, e.g. Euclidean distance. The class label for the testing data point is assigned to the majority class among the selected K data points. In the first stage of causal inference, the propensity score for a testing data point could be estimated as the proportion of its K nearest neighbors that vote for class 1.

4 A Data-Adaptive Approach for Modeling Propensity Scores

4.1 The proposed method: combining logistic regression with nonparametric machine learning methods

As described in the previous section, there are plenty of models to estimate propensity scores from both parametric and nonparametric perspectives. It is understood that there is no uniformly "best" algorithm for all the data sets and we can always employ some model selection criteria to select the best one for a particular data set. However, doing so ignores the randomness and uncertainty in the model selection procedure. In the literature, model-combining/ model-averaging techniques are often used to account for the uncertainty in the model selection procedure. Examples are Hoeting et al. (1999), Yang (2001), Yuan and Yang (2005) and Ghosh and Yuan (2008).

We now develop ways to combine parametric models with nonparametric models to estimate propensity scores. Let $e_1(\mathbf{X})$ be the estimate of the propensity score from a logistic regression model and $e_2(\mathbf{X})$ be the estimate of the propensity score from a nonparametric algorithm, such as a random forests model or generalized boosted model (GBM). We denote the proposed estimate as:

$$\hat{e}(\mathbf{X}) = \frac{a}{a+b} e_1(\mathbf{X}) + \frac{b}{a+b} e_2(\mathbf{X}). \tag{10}$$

$\hat{e}(\mathbf{X})$ is a weighted average of logistic regression and the nonparametric estimator with weights a and b. Intuitively, we would think that if $e_1(\mathbf{X})$ is more accurate, we would give more weight to logistic regression and if $e_2(\mathbf{X})$ is more accurate, we would like to give more weight to the nonparametric model. Therefore, given $(Y_i, \mathbf{X}_i, Z_i)$,

$i = 1, \ldots, n$, we propose the following data-based weights:

$$a_i = e_1(\mathbf{X}_i)^{Z_i}[1 - e_1(\mathbf{X}_i)]^{(1-Z_i)}, \tag{11}$$

and

$$b_i = e_2(\mathbf{X}_i)^{Z_i}[1 - e_2(\mathbf{X}_i)]^{(1-Z_i)}. \tag{12}$$

That is, we use the estimated Bernoulli likelihood at each sample point as its weight. Essentially, the choice of weights in (11) and (12) reflects the following principle: if $Z_i = 1$, $\hat{e}(\mathbf{X}_i)$ gives more weight to the higher value of $e_1(\mathbf{X}_i)$ and $e_2(\mathbf{X}_i)$; on the other hand, if $Z_i = 0$, $\hat{e}(\mathbf{X}_i)$ gives more weight to the lower value of $e_1(\mathbf{X}_i)$ and $e_2(\mathbf{X}_i)$. It can be easily shown that the weights are the posterior likelihood given a uniform prior to each model.

After comparing various non-parametric techniques, we recommend $e_2(\mathbf{X})$ be estimated from the random forests model or GBM. The reason can be seen clearly from our simulation results, which will be presented in Section 5.

4.2 Some intuitive insights

The choice of the weights in our proposed method can be explained from a decision-theoretic perspective. Similar to Ghosh and Yuan (2008), we argue that the objective is to decide whether the parametric model or the nonparametric model is a good approximation to the true propensity score model. Define $e_i \equiv \mathrm{I}\{\text{the } i\text{th model is a good model}\}$, $v_i \equiv \Pr(\text{the } i\text{th model is a good model})$ and $s_i \equiv \mathrm{I}\{\text{the } i\text{th model is selected as a good one}\}$, $i = 1, 2$. Then, the number of false-positive (FP) decisions are:

$$FP(s, e) = \sum_{i=1}^{2} s_i(1 - e_i),$$

and the number of false-negative (FN) decisions are:

$$FN(s, e) = \sum_{i=1}^{2} (1 - s_i)e_i.$$

Consequently, the expected count of false-positive (FP) decisions and false-negative (FN) decisions are given by:

$$FP(s,v) = E[FP(s,e)] = \sum_{i=1}^{2} s_i(1-v_i),$$

and

$$FN(s,v) = E[FN(s,e)] = \sum_{i=1}^{2} (1-s_i)v_i.$$

Ghosh and Yuan (2008) proved that the optimal decision with respect to the loss function $L(s,v) = cFP + FN$ is:

$$s_i = I\{v_i \geq t^*\}$$

where $t^* = c/(c+1)$. When $c = 1$, $t^* = 1/2$. As we can see, the optimal decision only depends on the probability of being a good model, which inspires us to use the posterior Bernoulli likelihood to assign weights in our proposed method.

4.3 Consistency of the proposed estimator

In this section, we prove some results regarding consistency of the proposed estimator in (10) when $e_1(\mathbf{X})$ is estimated by logistic regression and $e_2(\mathbf{X})$ is estimated by GBM. We first show the consistency of GBM (i.e., $e_2(\mathbf{X})$). Most of the proof follows Zhang and Yu (2005). Then, we show that the logistic regression estimator ($e_1(\mathbf{X})$) is still consistent even when there is a misspecification of the parametric model. The assumptions and proof are given by White (1982). In the end, we show the consistency of the proposed estimator which is a weighted average of the two.

Lemma 4.3.1: GBM is one form of greedy boosting defined in Algorithm 2.1. of Zhang and Yu (2005).

Proof: Following the previous notations, let $Z = 0$ or 1, $Z^* = 2Z - 1$, and $e(\mathbf{X}) = P(Z = 1|\mathbf{X}) = P(Z^* = 1|\mathbf{X})$. As described by McCaffrey et al. (2004), GBM aims to minimize

$$\begin{aligned} -El(Z, e(\mathbf{X})) &= -E\{Z\log\, e(\mathbf{X}) + (1-Z)\log(1-e(\mathbf{X}))|\mathbf{X}\} \\ &= E\{\log(1+e^{-f(\mathbf{X})Z^*})|\mathbf{X}\}, \end{aligned}$$

where $f(\mathbf{X}) = \log\{e(\mathbf{X})/1 - e(\mathbf{X})\}$. Define $\phi(f, z) = \log(1 + e^{-fz})$ and

$$A(f) = E_{X,Z^*}\phi(f(\mathbf{X}), Z^*). \tag{13}$$

Equivalently, GBM aims to find solution to the following optimization problem

$$\inf_{f \in span(S)} A(f)$$

where $span(S) = \{\sum_{j=1}^{m} w^j f^j : f^j \in S, w^j \in R, m \in Z^+\}$ is a linear function space. For GBM, f is fitted by an additive model with basis functions consisting of regression trees. Therefore, $S = \{I_{(-\infty,a_1]\times\cdots\times(-\infty,a_p])} : a_1, \ldots, a_p \in R\}$, which is the indicator of rectangular regions in the feature space. More specifically, given f_k at the kth step, the algorithm aims to find $\alpha \in R$ and $h(\mathbf{X}) \in S$ such that $f_{k+1} = f_k + \alpha h(\mathbf{X})$ approximately minimizes $A(f)$. Based on the definition of Algorithm 2.1. of Zhang and Yu (2005), GBM is one form of greedy boosting.

Lemma 4.3.2: a) The function A(f) in Lemma 4.3.1, (13) is convex and differentiable. b) A(f) is second-order differentiable and for GBM, the second-order derivative satisfies $A''_{f,g}(0) \leq 1$ where $A_{f,g}(h) = A(f + hg)$. c) $\phi(f, z)$ in Lemma 4.3.1 satisfies the Lipschitz condition.

Proof: Using the definition of functional derivative, we have

$$A'_{f,g}(0) = \lim_{h\to 0} \frac{1}{h}[A(f + hg) - A(f)] = -E_{X,Z^*} \frac{e^{-f(\mathbf{X})Z^*} Z^*}{1 + e^{-f(\mathbf{X})Z^*}} g(\mathbf{X})$$

and

$$\begin{aligned} A''_{f,g}(0) &= \lim_{h\to 0} \frac{1}{h^2}[A(f + hg) - 2A(f) + A(f - hg)] \\ &= E_{X,Z^*} \frac{(Z^*)^2}{(1 + e^{f(\mathbf{X})Z^*})(1 + e^{-f(\mathbf{X})Z^*})} g^2(\mathbf{X}) \end{aligned}$$

For GBM, we have $|g| = |(Z^* + 1)/2 - e(\mathbf{X})| \leq 2$, $(Z^*)^2 = 1$, $(1 + e^{f(\mathbf{X})Z^*})(1 + e^{-f(\mathbf{X})Z^*}) = 2 + e^{f(\mathbf{X})Z^*} + e^{-f(\mathbf{X})Z^*} \geq 4$. Therefore, $0 \leq A''_{f,g}(0) \leq 1$. $A(f)$ is convex. Let $f' = fz$. Then through some simple calculation, we can prove that $\phi(f, z) = \phi(f')$ satisfies the Lipschitz condition.

Given the dataset $D_1^n = \{(X_1, Z_1^*), \ldots, (X_n, Z_n^*)\}$, we follow the definitions in Zhang and Yu (2005): let $\hat{A}(f) = n^{-1} \sum_{i=1}^n \phi(f(\mathbf{X}_i), Z_i^*)$ be the empirical risk. Define $R_n(G, D_1^n)$ as the Rademacher complexity of G:

$$R_n(G, D_1^n) = E_\sigma \sup_{g \in G} \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i, Z_i^*),$$

where $\sigma_i = \pm 1$ with probability $\frac{1}{2}$.

Proposition 4.3.1: Assume the following:

(A1) There exists a unique f^* such that $A(f^*) = \inf_{f \in span(S)} A(f)$.

(A2) For any sequence f_m, $A(f_m) \xrightarrow{p} A(f^*)$ implies $f_m \xrightarrow{p} f^*$.

(A3) Consider two sequences of sample independent numbers k_n and β_n such that $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} \beta_n R_n(S) = 0$, where $R_n(S) = E_{D_1^n} R_n(S, D_1^n)$ is the expected Rademacher complexity for GBM. We assume the algorithm (GBM) stops at step $\hat{k}$ such that $\hat{k} \leq k_n$ and $||\hat{f}_{\hat{k}}||_1 \leq \beta_n$.

Then, we have

$$\hat{e}_{\hat{k}}(\mathbf{X}) \xrightarrow{p} e^*(\mathbf{X}), \quad as\ n \to \infty, \tag{14}$$

where $e^*(\mathbf{X}) = (1 + \exp\{-f^*\})^{-1}$.

Proof: Based on Lemma 4.3.1. and 4.3.2, we find GBM satisfies Assumption 3.1 and 3.2 in Zhang and Yu (2005). In addition, since for tree-based classifiers, we have: $R_n(S) \leq \tilde{C}(d/n)^{1/2} \to 0$, where d is the Vapnik-Chervonenkis (VC) dimension and $\tilde{C}$ is a constant, we can always find k_n and β_n, both of $o(n^{\frac{1}{2}})$ such that $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} \beta_n R_n(S) = 0$. According to Theorem 3.1 in Zhang and Yu (2005), as long as we stop GBM at step $\hat{k}$ such that $\hat{k} \leq k_n$ and $||\hat{f}_{\hat{k}}||_1 \leq \beta_n$, we have

$$\lim_{n\to\infty} E_{D_1^n} A(\hat{f}_{\hat{k}}) = A(f^*) \tag{15}$$

From (15), we have the L_1 convergence:

$$\lim_{n\to\infty} E_{D_1^n} |A(\hat{f}_{\hat{k}}) - A(f^*)| = 0.$$

Consequently, we have that $A(\hat{f}_{\hat{k}}) \xrightarrow{p} A(f^*), \quad as\ n \to \infty$. This implies that $\hat{f}_{\hat{k}}(\mathbf{X}) \xrightarrow{p} f^*(\mathbf{X}), \quad as\ n \to \infty$, which implies $\hat{e}_{\hat{k}}(\mathbf{X}) \xrightarrow{p} e^*(\mathbf{X}), \quad as\ n \to \infty$, the desired result.

Next, we show the consistency of the logistic regression estimator of propensity scores. The consistency will hold no matter whether the parametric model is the true underlying distribution or not. Define $e_\beta(\mathbf{X}) = (1 + \exp\{-\mathbf{X}^T\beta\})^{-1}$, $\beta \in R^p$, $f_\beta(\mathbf{X}) = \log\{e_\beta(\mathbf{X})/1 - e_\beta(\mathbf{X})\}$. The method of maximum likelihood aims to solve

$$\inf_{f \in span(S')} A(f)$$

where $S' = \{X^T\beta, \beta \in R^P\}$.

Proposition 4.3.2: Assume the following assumptions from White (1982):

(B1) Conditional on $\mathbf{X}_i$ $(i = 1, \ldots, n)$, the Z_i have a joint distribution G on a parameter space Ω, with a Radon-Nikodým density $g = dG/d\nu$ with respect to a dominating measure ν.

(B2) The parameter β is in a compact subset $B \subset R^p$. The logistic likelihood

$$f(z, \beta|\mathbf{x}) = e_\beta(\mathbf{x})^z[1 - e_\beta(\mathbf{x})]^{(1-z)},$$

is measurable in z for every β in B. Furthermore, it is continuous in β for every z. ever

(B3) $E\{\log g(Z_i)\}$ exists and $|\log f(z, \beta|\mathbf{x})| \leq m(z)$ for all $\beta \in B$, where m is integrable with respect to G.

(B4) Define $I(g, f|\beta) \equiv \int \log g(u|\mathbf{x})dG(u) - \int \log f(u, \beta|\mathbf{x})dG(u)$ as the Kullback-Leibler distance between g and f. Assume that $I(g, f|\beta)$ has a unique minimum at β_0.

Denote $\hat{\beta}$ as the maximum likelihood estimator of β in logistic regression and $\hat{e}_\beta(\mathbf{X}) = (1 + \exp\{-\mathbf{X}^T\hat{\beta}\})^{-1}$. Under the assumptions (B1)-(B4), White (1982) shows that

$$\hat{e}_\beta(\mathbf{X}) \xrightarrow{p} e_{\beta_0}(\mathbf{X}), \quad as\ n \to \infty$$

where $e_{\beta_0}(\mathbf{X}) = (1 + \exp\{-X^T\beta_0\})^{-1}$. β_0 is the so-called least false parameter in the sense that it minimizes the Kullback-Leibler distance between the true model and the parametric model.

Theorem 4.3.1: Under the conditions listed in Proposition 4.3.1 and 4.3.2, the proposed estimator in (10) is consistent if $f^*(\mathbf{X}), f_{\beta_0}(\mathbf{X}) \in span(S) \cap span(S')$.

Proof: If $f^*(\mathbf{X}), f_{\beta_0}(\mathbf{X}) \in span(S) \cap span(S')$, we have both $\hat{e}_{\hat{k}}(\mathbf{X})$, $\hat{e}_{\beta}(\mathbf{X})$ converge to $e^*(\mathbf{X}) = e_{\beta_0}(\mathbf{X})$ as $n \to \infty$. Following the notations in Proposition 4.3.1 & 4.3.2, the proposed estimator in (10) can be rewritten as:

$$\hat{e}(\mathbf{X}) = \frac{a}{a+b}\hat{e}_{\hat{k}}(\mathbf{X}) + \frac{b}{a+b}\hat{e}_{\beta}(\mathbf{X}),$$

where $a = \hat{e}_{\hat{k}}(\mathbf{X})^Z[1 - \hat{e}_{\hat{k}}(\mathbf{X})]^{(1-Z)}$ and $b = \hat{e}_{\beta}(\mathbf{X})^Z[1 - \hat{e}_{\beta}(\mathbf{X})]^{(1-Z)}$. When $Z = 1$,

$$\hat{e}(\mathbf{X}) = \frac{\hat{e}^2_{\hat{k}}(\mathbf{X}) + \hat{e}^2_{\beta}(\mathbf{X})}{\hat{e}_{\hat{k}}(\mathbf{X}) + \hat{e}_{\beta}(\mathbf{X})} \xrightarrow{p} e^*(\mathbf{X})$$

as $n \to \infty$. When $Z = 0$,

$$\hat{e}(\mathbf{X}) = \frac{(1 - \hat{e}_{\hat{k}}(\mathbf{X}))\hat{e}_{\hat{k}}(\mathbf{X}) + (1 - \hat{e}_{\beta}(\mathbf{X}))\hat{e}_{\beta}(\mathbf{X})}{2 - \hat{e}_{\hat{k}}(\mathbf{X}) - \hat{e}_{\beta}(\mathbf{X})} \xrightarrow{p} e^*(\mathbf{X})$$

as $n \to \infty$. Therefore, the proposed estimator $\hat{e}(\mathbf{X})$ is consistent if $e_1(\mathbf{X})$ is estimated by logistic regression and $e_2(\mathbf{X})$ is estimated by GBM.

Furthermore, if the true log odds of the propensity score can be approximated arbitrarily close by functions lying in the intersection of $span(S)$ and $span(S')$, we have that the proposed estimator converge to the true value of propensity scores. Consequently, $\widehat{ACE}$ in (6) and $\widehat{ACET}$ in (7) are consistent estimators for ACE and ACET.

4.4 The sandwich variance estimator

Since both $\widehat{ACE}$ in (6) and $\widehat{ACET}$ in (7) can be written as solutions to estimating equations, we use the sandwich formula based on the theory of M-estimation (Stefanski and Boos, 2002) to get the variance of the estimated treatment effects. Let $\hat{e}_i$ be the proposed estimator of propensity score for subject i. Denote $\hat{\mu}_{1,ACE} = \frac{\sum_{i=1}^n Y_i Z_i/\hat{e}_i}{\sum_{i=1}^n Z_i/\hat{e}_i}$

as an estimator of E[Y(1)] and $\hat{\mu}_{0,ACE} = \frac{\sum_{i=1}^{n} Y_i(1-Z_i)/(1-\hat{e}_i)}{\sum_{i=1}^{n}(1-Z_i)/(1-\hat{e}_i)}$ as an estimator of E[Y(0)], we have

$$\widehat{Var}(\widehat{ACE}) = \sum_{i=1}^{n} \frac{Z_i(Y_i - \hat{\mu}_{1,ACE})^2}{\hat{e}_i^2} / (\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i})^2 + \sum_{i=1}^{n} \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,ACE})^2}{(1-\hat{e}_i)^2} / (\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}_i})^2.$$

Denote $\hat{\mu}_{1,ACET} = \frac{\sum_{i=1}^{n} Y_i Z_i}{\sum_{i=1}^{n} Z_i}$ as an estimator of $E[Y(1)|Z=1]$ and $\hat{\mu}_{0,ACET} = \frac{\sum_{i=1}^{n} Y_i(1-Z_i)\hat{e}_i/(1-\hat{e}_i)}{\sum_{i=1}^{n}(1-Z_i)\hat{e}_i/(1-\hat{e}_i)}$ as an estimator of $E[Y(0)|Z=1]$, we have

$$\widehat{Var}(\widehat{ACET}) = \sum_{i=1}^{n} \frac{Z_i(Y_i - \hat{\mu}_{1,ACET})^2}{(\sum_{i=1}^{n} Z_i)^2} + \sum_{i=1}^{n} \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,ACET})^2 \hat{e}_i^2}{(1-\hat{e}_i)^2} / (\sum_{i=1}^{n} \frac{(1-Z_i)\hat{e}_i}{1-\hat{e}_i})^2.$$

5 Simulation Studies

5.1 Methodology comparison

To examine the performance of our proposed method, we conducted extensive simulation studies. In the first set of simulations, we compared the proposed combined estimators with different parametric and nonparametric algorithms for estimating the propensity score. We followed a slightly modified simulation structure as in Setoguchi et al. (2008) and Lee et al. (2010). Denote the vector of covariates to be $\mathbf{X}$. $\mathbf{X}$ is a 11-dimensional vector with X_0 being the intercept term, (X_1, X_2, X_3, X_4) being confounders, (X_5, X_6, X_7) only related to the treatment assignment, and (X_8, X_9, X_{10}) only related to the potential outcomes. In the simulation, $X_1 - X_{10}$ are first generated from $MVN(0, \Sigma)$ where Σ is a non-identity covariance matrix. Then, X_1, X_3, X_6, X_8, X_9 are dichotomized into 0-1 variables. The treatment indicator Z is generated from a Bernoulli distribution with p a function of the covariates. We use the following eight true propensity models to generate the treatment indicator (Lee et al., 2010):

- Scenario A: main effects only;
- Scenario B: main effects and one quadratic term;
- Scenario C: main effects and three quadratic terms;

- Scenario D: main effects and three two-way interaction terms;
- Scenario E: main effects, three two-way interaction terms and one quadratic term;
- Scenario F: main effect and ten two-way interaction terms;
- Scenario G: main effects, ten two-way interaction terms and three quadratic terms.

The response Y is then generated by:

$$Y = \alpha^T X + \gamma Z + \epsilon, \epsilon \sim N(0, \sigma^2),$$

with $\gamma = -0.4$, $\sigma = 0.1$ (25 percent of the effect of exposure) and

$$\alpha = (-3.85, 0.3, -0.36, -0.73, -0.2, 0, 0, 0, 0.71, -0.19, 0.26)^T.$$

According to Setoguchi et al. (2008), the effect of exposure (-0.4) is built on the effect of hormone replacement therapy on fracture or colorectal cancer. The coefficients for each predictor in the propensity models and the outcome model are based on the claims for the use of statins. For our simulation study, we generated $n = 1000$ sample points each time and repeat the procedure $N = 1000$ times.

In the first stage, we compared both parametric and nonparametric estimation algorithms. These algorithms include two parametric choices: logistic regression and linear discriminant analysis (LDA); six nonparametric choices: CART, pruned CART, bagged CART, random forests (RF), generalized boosted model (GBM), support vector machines (SVM) and K-nearest neighbors (KNN). We include all ten covariates as predictors and only consider main effects while fitting logistic regression since in practice, researchers do not know the true propensity score model and it is natural for them to assume a linear and additive model.

In the second stage, we estimate both ACE and ACET by inverse propensity weighting. The weights are chosen as follows: when estimating ACE, the treated person receives a weight of $1/\hat{e}(\mathbf{X})$ while the control person receives a weight of

$1/(1-\hat{e}(\mathbf{X}))$; when the objective is to estimate ACET, the treated person receives a weight of 1 while the control person receives a weight of $\hat{e}(\mathbf{X})/(1-\hat{e}(\mathbf{X}))$, where $\hat{e}(\mathbf{X})$ is the estimated propensity score from the first stage. The estimated ACE or ACET is then calculated as the difference in the weighted average among the treatment group and the control group.

5.2 Results

To compare the newly proposed method with existing methods in the literature, we need to evaluate its comparative performance. One way to evaluate the performance is to see how close the estimates are to the true propensity scores using simulations. However, Lunceford and Davidian (2004) showed that conditioning on the estimated propensity score rather than the true propensity score can yield smaller variance of the estimated ACE or ACET. That is, even when the propensity score is estimated more accurately in the first stage, it does not necessarily yield better causal inference in the second stage. Therefore, we should focus on the quality of the estimates in the second stage rather than the first stage. In our simulation study, we mainly look at the absolute bias (the percentage of the absolute value of the bias compared to -0.4), standard error and 95% confidence interval coverage of the estimated causal effect from various methods. In addition, we also report the average standardized absolute mean distance (ASAM), which is calculated in the following ways: for each covariate, calculate the absolute value of the difference (d_j) in the weighted means between the treatment group and the control group after applying the weights; then, divide d_j by the standard deviation of the covariate in the treatment group and average d_j over all the covariates (McCaffrey et al., 2004) . Tables 13 and 14 in the Appendix shows the performance metrics for estimating ACE and ACET, respectively.

There are several conclusions from the simulation results. First of all, parametric methods tend to yield lower bias but higher variance than nonparametric methods. In general, both parametric methods (LR and LDA) perform quite well. Among all nonparametric methods that we tested, random forests and GBM tend to perform the

best in terms of bias and variance, and coverage probabilities. These two conclusions validate our proposed method. In other words, due to the trade-off between bias and variance among parametric and nonparametric methods, it is reasonable to combine the parametric method with the nonparametric method. In addition, because of the superior performance of the random forests algorithm and GBM over other nonparametric algorithms, we choose them as the nonparametric component in our newly proposed estimator. In the performance metrics, "Proposed 1" is the combination of logistic regression with random forests and "Proposed 2" is the combination of logistic regression with GBM.

In terms of estimating ACET, we find that the newly proposed method, "Proposed 1", yields the smallest bias and variance in any of the eight scenarios compared to logistic regression and random forests; the same thing happens to "Proposed 2" while comparing it to logistic regression and GBM. In terms of estimating ACE, random forests yields much smaller biases than GBM. Therefore, it looks more beneficial to combine logistic regression with random forests rather than GBM. The newly proposed method ("Proposed 1") beats logistic regression and random forests in yielding the smallest variance. When the propensity score model becomes more complicated, the proposed method tends to give the smallest bias among the three.

In terms of computing time, we find that all the algorithms we tried are quite fast, with the relative exception of GBM. We evaluated the computing time for each algorithm to calculate the propensity scores based on one data set ($n = 1000$). We obtained the following times (in seconds): Logistic: 0.07, LDA: 0.19, CART: 0.25, PRUNE: 0.32, BAG: 2.42, RF: 1.72, GBM: 63.63, SVM: 1.7, KNN: 0.11, Proposed 1: 1.84, and Proposed 2: 63.75. As can be seen, random forests is much faster than GBM. Due to the savings in computing time and sometimes superior performance, we prefer the combination of LR and RF ("Proposed 1") to the combination of LR and GBM ("Proposed 2"). However, we have no proof of the consistency of the combination of LR and RF; that is currently under investigation.

In our simulation study, the two-stage causal inference also fits the model structure

discussed by Brookhart and van der Laan (2006). Using their notation, if we denote ACE or ACET as ψ, and the propensity score as η, ψ is hence the parameter of interest and η is the nuisance parameter. The issue of choosing an optimal model to estimate propensity scores can be restated as follows: assuming we have K different candidate models for estimating η, which model is optimal? Denote the resulting estimates of ψ (ACE or ACET) from the K candidate models as $\hat{\psi}_1(\mathbf{X}),\ldots,\hat{\psi}_K(\mathbf{X})$, and let $\hat{\psi}_0(\mathbf{X})$ be an approximately unbiased but highly variable estimate for ψ. The model used to estimate η in $\hat{\psi}_0(\mathbf{X})$ is regarded as the reference model. To account for the fact that there is a trade-off between bias and variance while estimating ψ, the authors proposed a cross-validation criterion for selecting the optimal estimator of the nuisance parameter among the K candidate models. Let X_v^0 be the training sample and X_v^1 be the testing sample in the v-th iteration of the Monte-Carlo cross-validation, the criterion function is defined as follows:

$$C_v(k) = \frac{1}{V}\sum_{v=1}^{V}(\hat{\psi}_k(X_v^0) - \hat{\psi}_0(X_v^1))^2.$$

The optimal model for estimating propensity scores is then chosen to be the one which leads to the smallest C_v among the K models. Brookhart and van der Laan (2006) proved that the optimal model selected by the Monte Carlo cross-validation criteria leads to the smallest mean square error of the parameter of interest.

We further performed a small simulation study to test our proposed method according to the cross-validation criterion. Based on the same data generation procedure as in the previous simulation study, we simulate 100 data sets with sample size n=1000. The Monte-Carlo cross-validation is performed $V = 25$ times with 50% of the data used in the training set each time. The reference model is set to be the logistic regression model. The cross-validation values are calculated and shown in Table 1.

As can be seen from the table, in almost all scenarios, the best models are the two proposed models which yield the smallest C_v values. The only exception is in Scenario G, where the values of C_v for the proposed methods are slightly larger than

Table 1: Monte-Carlo cross validation (C_v) for average causal effect among the treated(ACET).LDA = linear discriminant analysis; CART = classification and regression trees; PRUNE = pruned CART; BAG = bagged CART; RF = random forests; GBM=generalized boosted model; SVM = support vector machines; KNN = k-nearest neighbors.

Method	A	B	C	D	E	F	G
Logistic	0.0104	0.0085	0.0077	0.0126	0.0117	0.0111	0.0100
LDA	0.0103	0.0085	0.0077	0.0125	0.0114	0.0111	0.0099
CART	0.0149	0.0132	0.0200	0.0196	0.0176	0.0210	0.0291
PRUNE	0.0187	0.0181	0.0232	0.0254	0.0222	0.0237	0.0331
BAG	0.0095	0.0090	0.0116	0.0130	0.0121	0.0126	0.0178
RF	0.0093	0.0083	0.0086	0.0117	0.0110	0.0114	0.0149
GBM	0.0086	0.0084	0.0091	0.0122	0.0113	0.0120	0.0155
SVM	0.0118	0.0133	0.0107	0.0179	0.0184	0.0187	0.0194
KNN,k=3	0.0240	0.0225	0.0213	0.0357	0.0288	0.0389	0.0343
KNN,k=10	0.0107	0.0099	0.0088	0.0140	0.0124	0.0137	0.0144
Proposed 1	0.0084	0.0073	0.0059	0.0108	0.0099	0.0099	0.0117
Proposed 2	0.0083	0.0077	0.0070	0.0113	0.0104	0.0107	0.0129

logistic regression and LDA. We also tried dividing 66.7% of the data into the training set and the results are very similar.

5.3 A further simulation study: trimming large weights

Previous literature has shown that treatment effect estimates obtained by inverse propensity weighting are greatly influenced by subjects who receive the treatment but $\hat{e}(\mathbf{X}) \approx 0$ and those who receive control but $\hat{e}(\mathbf{X}) \approx 1$ (in both cases, the weights will be extremely large). Kang and Schafer (2007) argued that even mild lack of fit of the propensity scores in these two regions will lead to large bias. However, it has been shown that logistic regression model often fails to estimate correctly in these regions (Pregibon, 1982). Our method works well in the simulation study because it avoids the above two extreme cases and is able to shrink outlying weights to more sensible values. That is, if $Z = 1$ and $e_1(\mathbf{X}) \approx 0$, $\hat{e}(\mathbf{X}) \approx e_2(\mathbf{X})$ in the proposed estimator; on the other hand, if $Z = 0$ and $e_1(\mathbf{X}) \approx 1$, $\hat{e}(\mathbf{X}) \approx e_2(\mathbf{X})$ in the proposed estimator. To have a closer look at the weights in inverse propensity weighting, Figure 1 shows the boxplot of the weights calculated from logistic regression, random forests and the

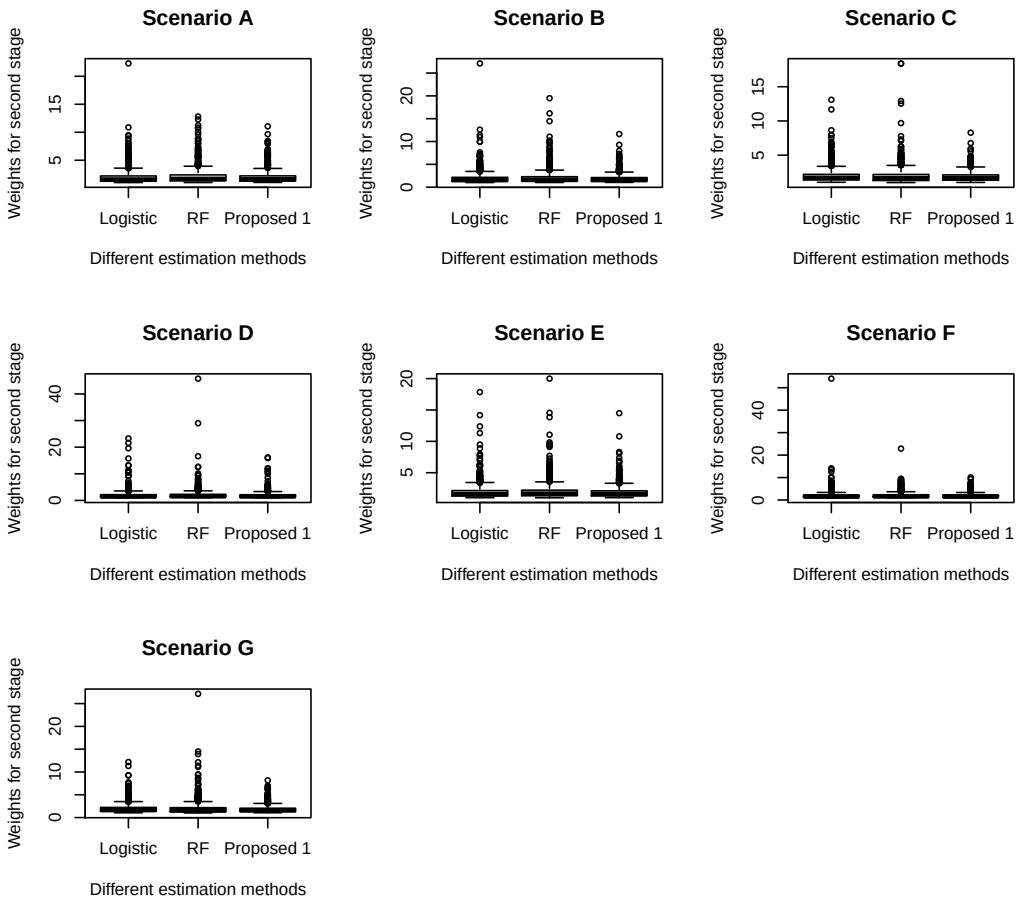

Figure 1: Distribution of propensity score weights used in inverse propensity weighting for different estimation methods: LR, RF, Proposed 1.

proposed method based on one simulated data set for all seven scenarios. As can be seen, the extremely large weights by logistic regression or random forests disappear in the newly proposed method, which is one of the reasons why it works well.

Figure 2 shows the boxplot of propensity score weights for logistic regression, GBM and the weighted estimator of the two. The conclusion is consistent with Figure 1.

To further illustrate the above viewpoint, we revisited a simulation study by Kang and Schafer (2007), which resembles a real study they encountered. The purpose of the study is to estimate a population mean with the existence of missing data. The authors argue that to estimate the population mean, propensity score based methods perform badly partially because of the extreme weights when there is a

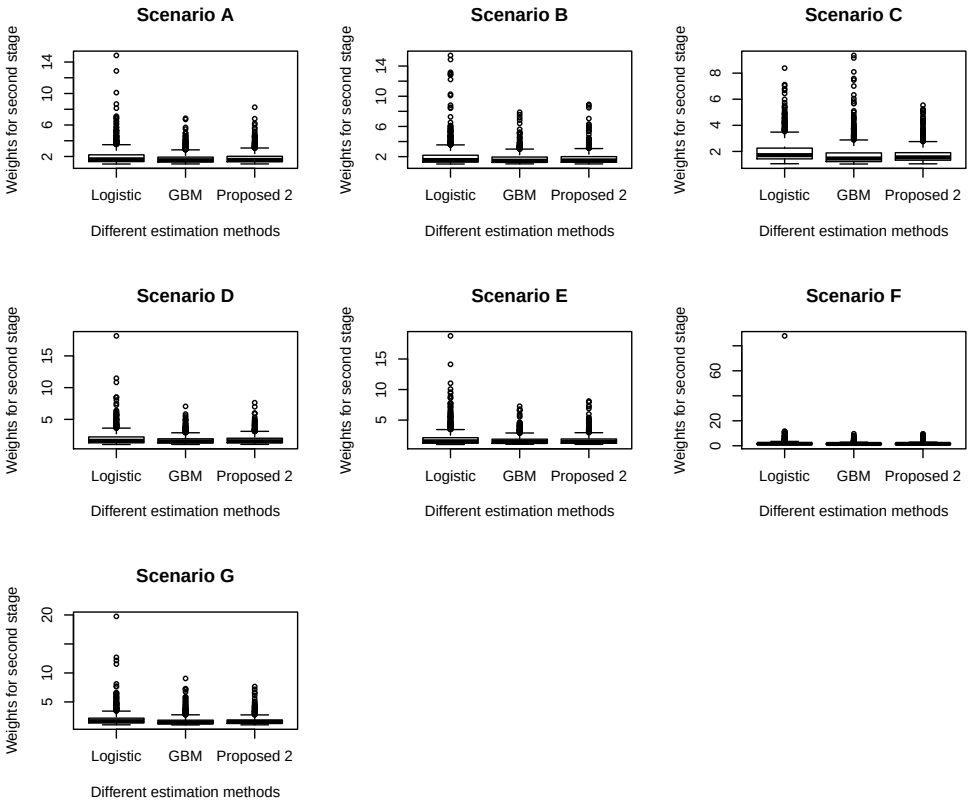

Figure 2: Distribution of propensity score weights used in inverse propensity weighting for different estimation methods: LR, GBM, Proposed 2.

misspecification of the propensity score model. We re-run the simulation to check whether our proposed method can improve the performance of propensity score based methods.

The simulation setup is as follows: the covariates $\mathbf{X} = (X_1, X_2, X_3, X_4)$ are generated from $MVN(0, I_{4\times4})$ and the response Y is generated from

$$Y = 210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + \epsilon, \epsilon \sim N(0, 1)$$

and the true propensity scores are:

$$e(\mathbf{X}) = P(Z = 1|\mathbf{X}) = \frac{1}{\exp\{-(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)\}},$$

where $z_i = 1$ if y_i is observed and $z_i = 0$ if y_i is missing. In addition, the authors assume that, instead of observing $x_{ij}, i = 1, \ldots, n, j = 1, \ldots, 4$, the following covariates are observed: $m_{i1} = \exp(x_{i1}/2)$, $m_{i2} = x_{i2}/(1+\exp(x_{i2}))+10$, $m_{i3} = (x_{i1}x_{i3}/25+0.6)^3$, $m_{i4} = (x_{i2} + x_{i4} + 20)^2$. The objective is to estimate $\mu = E(Y)$ based on those respondents whose $Z = 1$ using inverse propensity weighting (IPW), stratification and bias-corrected (double-robust) estimators. Details of the calculation for the three methods can be found in Kang and Schafer (2007). All three methods rely on the estimation of propensity scores. In the simulation, we fit the propensity score model by logistic regression, random forests and the weighted average of the two, based on x_{ij} and m_{ij}, separately. To be noticed, the true propensity score model is the logistic regression model of z_i on x_{ij}.

We run the simulation over 1000 samples with different sample sizes ($n = 200, 500, 1000$) and report the bias, standard deviation (SD), square root of the mean squared error (RMSE) and median of the absolute error (MAE) for $\hat{\mu}$. The first thing we notice is that, there are quite a few respondents whose propensity scores are estimated as 0 by random forests in each round of the simulation. As a result, their inverse propensity weights are infinite, which makes it impossible to produce an estimate of μ if we only employ random forests. However, our proposed method can still work because in equation (10), if $e_2(\mathbf{X}) = 0$ and $Z = 1$, the weight $b = 0$ and $\hat{e}(\mathbf{X}) = e_1(\mathbf{X})$.

Table 2: Simulation results for inverse propensity weighting estimators of μ

Sample size	PS model	Method	Bias	SD	RMSE	MAE
n=200	Fit with x_{ij}	LR	-0.181	3.924	3.926	2.332
		Proposed	-1.344	3.391	3.646	2.471
	Fit with m_{ij}	LR	1.718	9.716	9.862	3.237
		Proposed	-2.112	4.022	4.541	3.062
n=500	Fit with x_{ij}	LR	-0.071	2.614	2.614	1.570
		Proposed	-1.089	2.056	2.326	1.626
	Fit with m_{ij}	LR	3.742	10.733	11.361	2.709
		Proposed	-1.612	2.682	3.128	2.159
n=1000	Fit with x_{ij}	LR	0.008	1.779	1.778	1.165
		Proposed	-0.881	1.479	1.721	1.158
	Fit with m_{ij}	LR	5.011	11.107	12.180	2.441
		Proposed	-1.207	3.022	3.252	1.597

Table 3: Simulation results for stratified estimators of μ

Sample size	PS model	Method	Bias	SD	RMSE	MAE
n=200	Fit with x_{ij}	LR	-1.182	3.066	3.284	2.271
		Proposed	-1.372	3.224	3.502	2.362
	Fit with m_{ij}	LR	-2.954	3.273	4.408	3.248
		Proposed	-2.236	3.432	4.095	2.900
n=500	Fit with x_{ij}	LR	-1.195	1.829	2.184	1.500
		Proposed	-1.204	1.923	2.268	1.579
	Fit with m_{ij}	LR	-2.973	1.908	3.532	2.986
		Proposed	-1.804	2.003	2.695	1.959
n=1000	Fit with x_{ij}	LR	-1.084	1.304	1.695	1.204
		Proposed	-1.040	1.309	1.672	1.176
	Fit with m_{ij}	LR	-2.909	1.376	3.218	2.936
		Proposed	-1.560	1.398	2.094	1.644

Therefore, in the following analysis, we only compare the results of logistic regression and the proposed method. The results are shown in Table 2, 3 & 4, respectively.

For IPW estimators (see Table 2), when the propensity score is estimated from the true covariates x_{ij}, our proposed method yields higher bias but smaller variance, compared to logistic regression. The RMSE values by our proposed method are consistently smaller than logistic regression. When the propensity score is fitted with m_{ij}, IPW estimates by logistic regression are biased and lead to high variances. This is partially due to "occasional highly erratic estimates produced by a few enormous weights" (Kang and Schafer, 2007). In this case, our proposed method greatly reduces

Table 4: Simulation results for double-robust estimators of μ

Sample size	PS model	Method	Bias	SD	RMSE	MAE
n=200	Fit with x_{ij}	LR	-0.066	2.608	2.608	1.758
		Proposed	-0.066	2.608	2.608	1.753
	Fit with m_{ij}	LR	-4.567	7.863	9.090	3.724
		Proposed	-1.625	3.356	3.727	2.506
n=500	Fit with x_{ij}	LR	0.077	1.617	1.618	1.074
		Proposed	0.078	1.615	1.616	1.074
	Fit with m_{ij}	LR	-6.010	9.319	11.085	4.089
		Proposed	-1.657	2.323	2.852	1.881
n=1000	Fit with x_{ij}	LR	0.016	1.188	1.188	0.782
		Proposed	0.017	1.189	1.188	0.784
	Fit with m_{ij}	LR	-8.239	12.458	14.931	4.867
		Proposed	-1.847	2.824	3.373	1.735

the variance and yields less bias when the sample size gets larger. The RMSE values by our proposed method are only 46.0% ($n = 200$), 27.5% ($n = 500$), and 26.7% ($n = 1000$) of the RMSEs by logistic regression.

For stratified estimators based on true covariates x_{ij} (see Table 3), the performance of logistic regression and our proposed method are very close to each other. Particularly, when $n = 1000$, our proposed method is slightly better than the true model. When the propensity score model is fitted with m_{ij}, the proposed method reduces the RMSE by 7% ($n = 200$), 23.7% ($n = 500$), and 34.9% ($n = 1000$) , compared to logistic regression.

For bias corrected estimators (see Table 4), when both the regression model and propensity score model are fitted by misspecified covariates, logistic regression yields large values of bias and RMSE while our proposed method greatly reduces RMSEs by 59% ($n = 200$), 74.3% ($n = 500$), and 77.4%($n = 1000$) .

6 Data Analysis Example

6.1 Survival analysis using propensity score adjusted model

In this section, we apply the proposed method to a case study. The data set was obtained from a study of 3894 patients with intrahepatic cholangiocarcinomas (IHC,

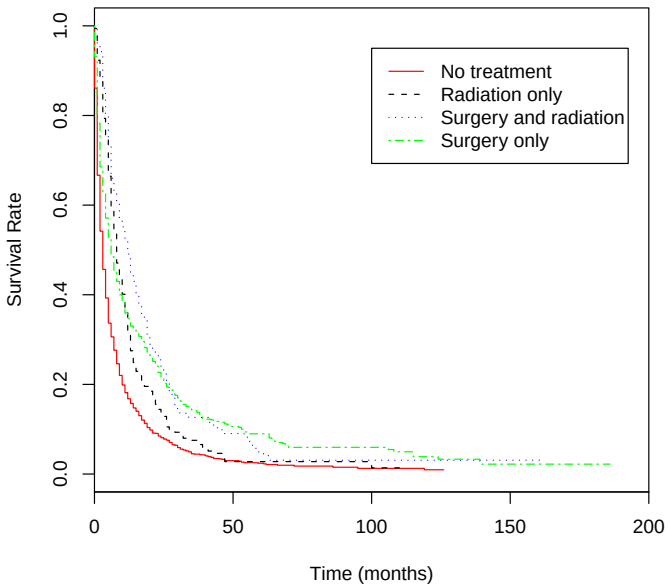

Figure 3: Kaplan-Meier estimators of survival functions by different treatment groups.

Shinohara et al. (2008)). In this medical study, the response variable is the survival time after diagnosis with IHC and the treatments are: surgery and radiation, surgery only, radiation only and no treatment. Among the patients, more than 50% of them receive no treatment because those patients have advanced disease so it is difficult to remove the tumor (have the surgery) and it is unclear whether or not radiation really helps. The Kaplan-Meier estimators of survival functions separated by treatment groups are displayed in Figure 3. Table 5 and Table 6 display some baseline characteristics and tumor characteristics of study patients. P-values for categorical variables are computed from chi-squared tests while p-values for continuous variables are calculated from analysis of variance. As we can see, the distributions of age, race, tumor stage, tumor grade and year of diagnosis are significantly different among different treatment groups. In other words, these variables could be potential confounders for the treatments.

We next fit a sequence of survival analysis models to check whether any treatment, baseline characteristics or tumor characteristics have a statistically significant effect

Table 5: Descriptive statistics of baseline characteristics by different treatment groups

Variable	Surgery and radiation ($N = 262$)	Surgery ($N = 956$)	Radiation ($N = 343$)	No treatment ($N = 2333$)	p-value
	Mean (SD)				
Age (year)	62.67 (13.50)	65.49 (13.86)	66.27 (11.33)	71.94 (13.23)	$< .0001$
	Count (Percent%)				
Race/ethnicity					
Caucasian	210(80.15)	792(83.11)	248(72.30)	1883(80.88)	
African American	12(4.58)	45(4.72)	22(6.41)	178(7.65)	
Other	40(15.27)	116(12.17)	73(21.28)	267(11.47)	$< .0001$
Gender					
Male	148(56.49)	509(53.24)	190(55.39)	1195(51.22)	
Female	114(43.51)	447(46.76)	153(44.61)	1138(48.78)	.2072

Table 6: Descriptive statistics of tumor characteristics by different treatment groups

Variable	Surgery and radiation ($N = 262$)	Surgery ($N = 956$)	Radiation ($N = 343$)	No treatment ($N = 2333$)	p-value
	Count (Percent%)				
Stage					
Distant	68(25.95)	296(30.96)	81(23.62)	615(26.36)	
Localized	61(23.28)	269(28.14)	81(23.62)	423(18.13)	
Regional	92(35.11)	217(22.70)	101(29.45)	439(18.82)	
Unknown	41(15.65)	174(18.20)	80(23.32)	856(36.69)	$< .0001$
Grade					
Grade I	22(8.40)	85(8.89)	22(6.41)	82(3.51)	
Grade II	62(23.66)	198(20.71)	30 (8.75)	188(8.06)	
Grade III	51(19.47)	184(19.25)	46 (13.41)	214(9.17)	
Grade IV	0(0.00)	18(1.88)	2(0.58)	16(0.69)	
Unknown	127(48.47)	471(49.27)	243(70.85)	1833(78.57)	$< .0001$
Year of diagnosis					
1988-1992	80(30.53)	248(25.94)	47(13.70)	280(12.00)	
1993-1997	117(44.66)	465(48.64)	47(13.70)	471(20.19)	
1998-2003	65(24.81)	243(25.42)	249(72.59)	1582(67.81)	$< .0001$

Table 7: Hazard ratios in univariate and multivariate Cox PH models

Variable	Univariate HR(95% CI)	p-value	Multivariate HR(95% CI)
Age	1.02(1.01-1.02)	< 0.0001	1.01(1.01-1.02)
Treatment			
No treatment	1		1
Radiation only	0.61(0.54-0.68)		0.63(0.55-0.72)
Surgery only	0.59(0.55-0.64)		0.47(0.42-0.52)
Surgery and radiation	0.48(0.42-0.55)	< 0.0001	0.37(0.31-0.43)
Gender			
Female	1		
Male	1.00(0.93-1.070)	0.98	
Race/ethnicity			
White	1		1
African American	1.21(1.06-1.39)		1.38(1.18-1.62)
Other	0.96(0.86-1.06)	0.0138	1.14(1.01-1.29)
SEER location		0.2497	
Grade			
Grade I	1		
Grade II	1.08(0.90-1.30)		
Grade III	1.55(1.48-1.61)		
Grade IV	1.86(1.57-2.21)	< 0.0001	
Stage			
Distant	1		1
Localized	0.49(0.45-0.55)		0.50(0.46-0.56)
Regional	0.63(0.57-0.69)	< 0.0001	0.66(0.60-0.72)
Year of diagnosis			
1988-1992	1		1
1993-1997	0.86(0.78-0.95)		0.86(0.76-0.98)
1998-2003	0.76(0.69-0.83)	< 0.0001	0.52(0.46-0.59)

on the survival time of patients from IHC. The second and third columns of Table 7 display the hazard ratios and the corresponding p-values by fitting a Cox proportional hazard model with only one predictor at a time. As we can see, Age, Treatment, Race/ethnicity, Grade, Stage, and Year of diagnosis are all significant predictors of survival from IHC. We further fit a multivariate Cox proportional hazard model with all the significant predictors found in the univariate setting. Since Grade has too many missing values (68.67%), we excluded this variable in our multivariate model. The variable Stage has 29.56% missing rate, but we kept it in our model. From the fourth column in Table 7, we see that all predictors included in the multivariate Cox proportional hazards model are significant. Therefore, we take Age, Race/ethnicity, Stage,and Year of diagnosis to be potential confounders in this study.

Since the purpose of this study is to examine whether the additional use of radiation therapy would improve the mortality from IHC, we therefore conduct two separate analyses: in the first analysis, we compare surgery and radiation with surgery only and in the second analysis, we compare radiation only with no treatment. In this way, we only need to focus on causal inference based on binary treatments in each analysis. To be noticed, in the first analysis, surgery with radiation will be the treatment and surgery only will be the control. For each analysis, we wish to determine whether the use of radiation therapy is associated with longer survival. Due to the existence of confounders in the study, we need to employ a propensity score-adjusted model. Researchers usually analyse such data using the Cox proportional hazards (PH) model, which models the hazard function conditional on treatment and baseline covariates:

$$\lambda(t) = \lambda_0(t)\exp\{Z\gamma + \mathbf{X}^T\beta\}, \tag{16}$$

where $\lambda(t)$ is the hazard rate at time t, $\lambda_0(t)$ is the baseline hazard rate at time t, Z is the treatment indicator and $\mathbf{X}$ is the baseline covariates . In this situation, we are not performing causal inference strictly speaking. Although the strong ignorability treatment assumptions in (4) and (5) do not directly apply here, propensity scores can still be used as coarse balancing scores, which means the propensity score $e(\mathbf{X})$

Table 8: Deviance for different propensity score models

Model	Analysis 1	Analysis 2
Logistic Regression	1009.378	1387.896
Random Forests	1194.650	1817.991
Proposed	972.685	1335.162

satisfies $Z \perp \mathbf{X}|e(\mathbf{X})$. We use equation (10) to calculate the estimated propensity score for each patient, which is a weighted average of logistic regression estimates and random forests estimates. We can first evaluate the propensity score models using the deviance metric:

$$\text{Deviance} = -2\sum_{i=1}^{n}\{Z_i \log\, e_i(\mathbf{X}) + (1 - Z_i)\log[1 - e_i(\mathbf{X})]\}.$$

Table 8 shows the deviance of different propensity score models. To be noted, Analysis 1 is the analysis of radiation with surgery vs. surgery only while analysis 2 is the analysis of radiation only vs. no treatment. In both analyses, the proposed method yields the lowest deviance.

We then fit a Cox proportional hazards model with the treatment indicator as the predictor. We assign each patient a weight by the inverse propensity weighting scheme. That is, for the treated, the weight is $1/\hat{e}(\mathbf{X})$ and for the control, the weight is $1/(1-\hat{e}(\mathbf{X}))$, where $\hat{e}(\mathbf{X})$ is the estimated propensity score by our proposed method. For comparison, we also fit the propensity score model using logistic regression and random forests, separately. We notice, for some patients in treatment groups, their propensity scores are estimated to be 0 by random forests model, which makes their weights infinity. As a result, when we employ propensity score adjusted model by random forests to estimate the hazard ratio, we have to exclude these patients in the analysis. This ad hoc modification can be problematic. Therefore, we mainly focus on the comparison between the proposed method and logistic regression. Table 9 and 10 display the results. As can be seen from both analysis, the standard errors of γ in equation (16) by propensity score adjusted models are smaller than the multivariate model. In addition, our proposed method yields slightly smaller standard errors than the model using logistic regression to estimate propensity scores. In conclusion, the

Table 9: Radiation with surgery vs. surgery only

	$\hat{\gamma}$ (SE)	HR (95% CI)
Multivariate model	-0.3236(0.0865)	0.72 (0.61-0.86)
Propensity score adjusted model by LR	-0.2165(0.0493)	0.81(0.73-0.89)
Propensity score adjusted model by RF	-0.1108(0.0413)	0.90(0.83-0.97)
Propensity score adjusted model by Proposed	-0.2427(0.0492)	0.78(0.71-0.86)

Table 10: Radiation only vs. no treatment

	$\hat{\gamma}$ (SE)	HR (95% CI)
Multivariate model	-0.5189(0.0747)	0.60 (0.51-0.69)
Propensity score adjusted model by LR	-0.4646(0.0362)	0.63 (0.59-0.67)
Propensity score adjusted model by RF	-0.5288(0.0295)	0.59 (0.56-0.62)
Propensity score adjusted model by Proposed	-0.4801(0.0360)	0.62 (0.58-0.66)

use of radiation therapy is found to significantly improve the survival rate by either the multivariate model or the propensity score adjusted model. This finding is consistent with what have been reported in Shinohara et al. (2008).

6.2 Sensitivity Analysis

Next, we perform a sensitivity analysis for the propensity score adjusted model in each group, respectively. The sensitivity analysis is used to examine the impact of unknown confounders on the hazard ratio of the treatment (Lin et al., 1998; Mitra and Heitjan, 2007). Here we assume the unknown confounder is binary with imbalanced distributions between the treatment and the control group. More specifically, we assume the unknown confounder follows a Bernoulli distribution with success probability P_1 for the treatment group and P_0 for the control group. Assume the hazard ratio of the unknown confounder is e^{β_1} for the treatment and e^{β_0} for the control group. Denote the hazard ratio of the treatment as e^{γ^*} and e^{γ} with and without the confounder in the model, respectively. Lin et al. (1998) show that

$$\gamma^* \approx \gamma - \log \frac{e^{\beta_1} P_1 + (1 - P_1)}{e^{\beta_0} P_0 + (1 - P_0)}. \tag{17}$$

Based on equation (17), we are able to calculate the hazard ratio of the treatment after adjusting the unmeasured confounder. Notice that e^{γ} is the estimated hazard

Table 11: Sensitivity analysis

P_1	P_0	$e^{\beta_1} = e^{\beta_0}$	Radiation with surgery vs.surgery only e^{γ^*}(95% CI)	Radiation only vs. no treatment e^{γ^*}(95% CI)
0.1	0.2	1.25	0.80 (0.73-0.88)	0.64 (0.59-0.68)
0.1	0.2	1.5	0.82 (0.74-0.90)	0.65 (0.61-0.69)
0.1	0.2	2	0.85 (0.77-0.94)	0.68 (0.63-0.72)
0.1	0.3	1.25	0.82 (0.74-0.90)	0.65 (0.61-0.69)
0.1	0.3	1.5	0.85 (0.78-0.94)	0.68 (0.64-0.72)
0.1	0.3	2	0.92 (0.84-1.02)	0.73 (0.69-0.78)
0.1	0.4	1.25	0.84 (0.76-0.92)	0.67 (0.62-0.71)
0.1	0.4	1.5	0.89 (0.81-0.98)	0.71 (0.66-0.75)
0.1	0.4	2	0.99 (0.90-1.09)	0.79 (0.74-0.84)
0.2	0.3	1.25	0.80 (0.73-0.88)	0.63 (0.59-0.68)
0.2	0.3	1.5	0.82 (0.74-0.90)	0.65 (0.61-0.69)
0.2	0.3	2	0.85 (0.77-0.93)	0.67 (0.63-0.72)

ratio by our proposed method in Table 9 and 10. Table 11 shows the results for different scenarios. As we can see, for radiation with surgery versus surgery only group, the significant effect of radiation will be attenuated only when the prevalence of the unmeasured confounder in the untreated group is at least three times as likely in the treated group and the hazard ratio of the unmeasured confounder is at least two. For the radiation only versus no treatment group, the treatment remains significant in any scenario listed in the table. This sensitivity analysis demonstrates that the results obtained from our proposed method are robust against the unknown confounder which we fail to include in the model.

6.3 Simulated Data

In the IHC study, the true value of the treatment effect is unknown. To test the performance of our proposed methods for survival data, we further conduct a simulation study. The simulation setup is based on the analysis of surgery and radiation vs. surgery only in Section 6.1. We simulate 1000 data sets for sample sizes $n = 1218$. We generate four confounders: X_1 (Age), X_2 (Race), X_3 (Stage) and X_4 (Year of diagnosis). X_1 is generated from $N(64.9, 13.8^2)$; X_2, X_3 and X_4 are generated

Table 12: Simulation results for $\hat{\gamma}$

Method	Mean	Bias	SD	RMSE	MAE
LR	-0.2565	-0.0065	0.1244	0.1245	0.0784
RF	-0.2695	-0.0195	0.1648	0.1658	0.1113
GBM	-0.3110	-0.0610	0.0930	0.1112	0.0755
Proposed 1	-0.2804	-0.0304	0.1187	0.1225	0.0781
Proposed 2	-0.3006	-0.0506	0.0917	0.1082	0.0717

from multinomial distributions with three categories each and the event probabilities are $(0.823, 0.047, 0.13)$, $(0.299, 0.271, 0.43)$ and $(0.269, 0.478, 0.253)$, respectively. The treatment indicator Z is generated from a logistic regression model such that about 21.5% of the subjects receive the treatment. The survival time T is generated from an accelerated failure-time (AFT) model based on Weibull distribution with scale parameter $k = 1$. Based on the Cox PH model in (16), T can be generated from the following function (Bender et al., 2005):

$$T = -\frac{\log(\mathrm{U})}{\lambda}\exp(-Z\gamma - \mathbf{X}^T\beta),$$

where $U \sim \mathrm{Unif}(0, 1)$. The values of the coefficients are from the estimates of the Cox PH model in the data analysis example. In particular, $\lambda = 1/E(T) = 0.0709$ and the treatment effect $\gamma = -0.25$. Next, the censoring time C is generated from $U(0, a)$ where $a = 100$ is properly chosen such that the censoring rate is about 14.8%. The observed survival time is $\mathrm{Tobs} = \min(T, C)$. We then fit a AFT model with the treatment indicator as the predictor. We assign each patient a weight by the inverse propensity weighting scheme based on different propensity score estimation methods. The simulation results for estimating γ are displayed in Table 12. From the table, we find the standard deviation (SD), square root of mean squared error (RMSE) and median of the absolute error (MAE) of $\hat{\gamma}$ are reduced by the proposed methods if we compare "Proposed 1" to logistic regression and random forests, "Proposed 2" to logistic regression and GBM.

7 Discussion

In this article, we have developed a class of propensity score estimators that have desirable properties for average causal effect estimators. The approach involves combining the traditional parametric model with the more recent nonparametric models using machine learning methods in estimating propensity scores. We proposed a weighted average of logistic regression and a machine learning algorithm (random forests or generalized boosted model) with weights properly chosen. The first simulation shows that the newly proposed method reduces the variance of the estimates, as well as the bias in most cases. The second simulation study demonstrates when there is a misspecification in the logistic regression model, our proposed method can shrink large weights to produce less biased and variable estimates. When the machine learning algorithm fails to work because of infinite weights, both the second simulation study and the data analysis example show that our proposed method can still work properly.

It should be noted that the proposed procedure also is tailored to estimators of causal effects that are estimated using inverse propensity weighted procedures. In particular, as Kang and Schafer (2007) showed, IPW procedures can suffer from poor performance due to model misspecification. This will manifest itself in terms of observations with extreme weights. In this regard, the proposed methodology can be viewed as developing "robust" weights that incorporate all observations while simultaneously keeping weights from becoming too extreme. While practitioners using causal inference know that observations with extreme weights need to be downweighted in the analysis for estimating causal effects, many of the solutions proposed have been *ad hoc*, while our procedure is more principled.

Although the IHC study in our data analysis example has four treatments, we divided them into two groups and focused on causal inference for binary treatments. It is also desirable to explore how to improve causal inference in the regime of multilevel treatments and extend the work by Imai and van Dyk (2004) and Tchernis et

al. (2005).

Acknowledgements

The authors thank Brian Lee for making his code available. The work of Zhu and Ghosh was supported by the National Institute on Drug Abuse grant P50 DA010075-16 and NCI grant CA 129102. The work of Mukherjee was supported by NSF grant DMS-1007494 and NIH/NCI grant CA156608. The content of this manuscript is solely the responsibility of the author(s) and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health. Mitra would like to acknowledge Eric Shinohara, MD for making the cholangiocarcinoma data available to us.

Appendix

Table 13: Simulation results for average causal effect (ACE).

Method	A	B	C	D	E	F	G
				ASAM			
Logistic	0.02	0.02	0.02	0.03	0.03	0.03	0.03
LDA	0.02	0.02	0.02	0.03	0.03	0.03	0.03
CART	0.15	0.15	0.13	0.17	0.17	0.14	0.15
PRUNE	0.17	0.17	0.13	0.18	0.18	0.15	0.16
BAG	0.15	0.14	0.13	0.17	0.17	0.14	0.15
RF	0.06	0.06	0.06	0.07	0.08	0.06	0.06
GBM	0.09	0.09	0.09	0.11	0.11	0.11	0.11
SVM	0.11	0.11	0.07	0.12	0.12	0.08	0.12
KNN,k=3	0.25	0.24	0.20	0.27	0.27	0.22	0.27
KNN,k=10	0.08	0.08	0.07	0.10	0.10	0.08	0.10
Proposed 1	0.05	0.05	0.06	0.06	0.07	0.06	0.06
Proposed 2	0.07	0.07	0.08	0.08	0.09	0.09	0.08
				Absolute bias (percent)			
Logistic	6.43	6.01	5.31	7.07	7.09	8.35	7.51
LDA	6.45	5.99	5.34	7.83	7.15	8.43	7.28
CART	17.59	15.46	13.55	17.34	16.02	16.96	19.57
PRUNE	21.45	18.47	14.18	19.21	17.58	18.00	20.66
BAG	14.29	12.47	11.58	15.32	14.52	12.44	14.91
RF	7.16	6.93	8.17	7.51	7.72	7.91	7.58
GBM	11.01	10.53	10.60	12.91	13.13	11.80	12.61
SVM	16.60	16.18	11.83	17.42	18.00	12.32	18.13
KNN,k=3	38.18	35.00	30.08	39.78	35.80	30.48	40.65
KNN,k=10	8.78	8.54	8.88	9.63	9.31	8.28	8.89
Proposed 1	6.85	6.57	6.98	7.77	8.44	7.36	7.30
Proposed 2	9.22	9.31	10.50	10.98	11.83	11.42	10.60
				Standard error			
Logistic	0.061	0.060	0.057	0.064	0.063	0.061	0.064
LDA	0.060	0.060	0.057	0.064	0.063	0.061	0.064
CART	0.056	0.056	0.059	0.057	0.057	0.060	0.058
PRUNE	0.055	0.055	0.059	0.056	0.056	0.059	0.057
BAG	0.053	0.053	0.055	0.053	0.053	0.054	0.054
RF	0.060	0.060	0.062	0.061	0.061	0.061	0.062
GBM	0.054	0.055	0.055	0.055	0.055	0.055	0.055
SVM	0.053	0.053	0.054	0.054	0.054	0.055	0.055
KNN,k=3	0.058	0.057	0.057	0.058	0.057	0.058	0.058
KNN,k=10	0.060	0.060	0.059	0.061	0.060	0.060	0.061
Proposed 1	0.057	0.056	0.054	0.058	0.057	0.055	0.058
Proposed 2	0.055	0.055	0.053	0.056	0.055	0.053	0.056
				95% CI coverage			
Logistic	100	100	100	100	100	100	100
LDA	100	100	100	100	100	100	100
CART	78.1	85.4	91.2	80.8	83.0	83.0	74.2
PRUNE	66.6	75.8	89.0	75.6	78.5	80.3	71.3
BAG	90.4	94.2	94.4	89.6	89.7	92.1	88.3
RF	99.7	100.0	99.9	99.9	99.9	99.9	99.9
GBM	98.5	98.9	97.8	96.9	96.1	96.9	96.5
SVM	87.9	88.1	95.9	85.5	83.5	95.0	85.4
KNN,k=3	23.5	30.3	43.4	20.9	28.7	45.0	18.1
KNN,k = 10	98.9	99.3	98.8	98.4	98.9	99.5	99.2
Proposed 1	99.8	100.0	100.0	100.0	99.3	99.9	99.7
Proposed 2	99.5	99.5	98.3	99.1	98.0	97.6	99.1

Table 14: Simulation results for average causal effect among the treated (ACET).

Method	A	B	C	D	E	F	G
				ASAM			
Logistic	0.04	0.04	0.05	0.06	0.06	0.06	0.08
LDA	0.04	0.04	0.05	0.06	0.06	0.07	0.08
CART	0.16	0.15	0.14	0.17	0.16	0.15	0.14
PRUNE	0.17	0.16	0.14	0.18	0.17	0.16	0.15
BAG	0.13	0.13	0.12	0.14	0.14	0.12	0.11
RF	0.07	0.07	0.07	0.08	0.08	0.07	0.07
GBM	0.06	0.06	0.06	0.07	0.07	0.06	0.07
SVM	0.09	0.09	0.06	0.09	0.08	0.09	0.07
KNN,k=3	0.20	0.19	0.16	0.22	0.20	0.22	0.17
KNN,k=10	0.09	0.09	0.08	0.11	0.10	0.10	0.09
Proposed 1	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Proposed 2	0.05	0.05	0.05	0.05	0.05	0.05	0.05
				Absolute bias (percent)			
Logistic	9.48	9.17	11.24	13.20	15.02	15.14	23.39
LDA	9.47	9.13	11.40	13.12	15.08	15.13	23.50
CART	18.87	14.52	17.06	16.30	15.02	21.31	18.37
PRUNE	22.01	17.34	17.65	17.82	16.07	22.06	19.12
BAG	11.73	9.43	11.33	9.76	8.78	10.92	10.22
RF	9.19	7.97	10.72	9.19	8.96	9.98	10.67
GBM	8.95	7.65	8.58	8.20	7.62	7.85	8.26
SVM	14.54	12.27	9.48	11.71	11.05	12.52	9.80
KNN,k=3	30.25	23.94	17.84	28.34	21.16	29.36	16.50
KNN,k=10	11.52	9.97	10.41	10.53	10.83	10.34	11.36
Proposed 1	8.05	6.80	7.04	7.97	8.00	8.30	8.73
Proposed 2	8.44	7.15	6.53	7.86	7.26	7.38	6.74
				Standard error			
Logistic	0.068	0.067	0.061	0.076	0.074	0.076	0.069
LDA	0.068	0.067	0.061	0.076	0.074	0.075	0.069
CART	0.060	0.060	0.068	0.062	0.062	0.062	0.067
PRUNE	0.059	0.059	0.067	0.061	0.060	0.061	0.066
BAG	0.057	0.056	0.062	0.058	0.057	0.058	0.060
RF	0.065	0.063	0.068	0.067	0.064	0.068	0.066
GBM	0.062	0.060	0.063	0.064	0.062	0.063	0.063
SVM	0.058	0.058	0.059	0.060	0.060	0.060	0.061
KNN,k=3	0.063	0.062	0.063	0.064	0.063	0.064	0.065
KNN,k=10	0.066	0.065	0.064	0.067	0.065	0.067	0.065
Proposed 1	0.062	0.060	0.057	0.065	0.063	0.065	0.060
Proposed 2	0.061	0.059	0.057	0.064	0.062	0.063	0.059
				95% CI coverage			
Logistic	99.5	99.6	99.2	99.6	98.3	97.9	86.5
LDA	99.6	99.5	99.1	99.6	98.3	97.7	85.6
CART	79.4	89.3	89.2	86.1	88.8	74	84.6
PRUNE	70.2	80.1	87.2	80.9	84.8	71.9	81.7
BAG	95.1	98.3	96.8	97.9	98.6	95.6	97.7
RF	99.3	99.6	99.3	99.8	99.7	99.5	99.5
GBM	99	99.8	99.7	100	99.9	99.7	99.7
SVM	92.9	96	98.4	95.8	96.9	96.3	97.6
KNN,k=3	49.7	72.2	84.3	55.9	78.1	57.2	87.6
KNN,k=10	97.9	98.8	98.6	98.9	99	98.4	97.4
Proposed 1	99.7	99.8	99.9	99.8	99.9	99.7	99.1
Proposed 2	99.9	99.7	99.9	100	99.9	99.8	100

References

Bender, R. and Augustin, T. and Blettner, M. (2005), "Generating Survival Times to Simulate Cox Proportional Hazards Models," *Statistics in Medicine*, **24**, 1713–1723.

Biau, G, Devroye, L. and Lugosi, G. (2008), "Consistency of Random Forests and Other Averaging Classifiers," *Journal of Machine Learning Research*, **9**, 2015–2033.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, **24**, 123 – 140.

Breiman, L. (2001), "Random Forests," *Machine Learning*, **45**, 5 – 32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth and Brooks/Cole.

Brookhart, M.A. and van der Laan, M.J. (2006), "A Semiparametric Model Selection Criterion with Applications to the Marginal Structural Model," *Computational Statistics and Data Analysis*, **50**, 475 – 498.

Cortes, C. and Vapnik, V. (1995), "Support-Vector Networks," *Machine Learning*, **20**, 273-297.

Freund, Y. and Schapire, R. E. (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, **55**, 119-139.

Ghosh, D. and Yuan, Z. (2008), "Combining Multiple Biomarker Models in Logistic Regression," *Biometrics*, **64**, 431 – 439.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition,* New York: Springer.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. A. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, **14**, 382 – 401.

Kouassi, D.A. and Singh, J. (1997), "A Semiparametric Approach to Hazard Estimation with Randomly Censored Observations," *Journal of the American Statistical Association*, **92**, 1351–1355.

Kang, J. D. Y. and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, **22**, 523 – 539.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010), "Improving Propensity Score Weighting using Machine Learning," *Statistics in Medicine* **29**, 337 – 346.

Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, **54**, 948 – 963.

Lunceford, J. K. and Davidian, M. (2004), "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, **23**, 2937 – 2960.

Imai, K. and van Dyk, D.A. (2004), "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, **99**, 854–866.

Mays, J.E. and Birch, J.B. and Starnes, B.A. (2001), "Model Robust Regression: Combining Parametric, Nonparametric, and Semiparametric Methods," *Journal of Nonparametric Statistics*, **13**, 245–277.

McCaffrey, D.F., Ridgeway, G. and Morral, A.R. (2004), "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, **9**, 403 – 425.

Mitra, N. and Heitjan, D. F. (2007), "Sensitivity of the Hazard Ratio to Nonignorable Treatment Assignment in an Observational Study," *Statistics in Medicine*, **26**, 1398 – 1414.

Nottingham, Q.J. and Birch, J.B. (2000), "A Semiparametric Approach to Analysing Dose–Response Data," *Statistics in Medicine*, **19**, 389–404.

Olkin, I. and Spiegelman, C.H. (1987), "A semiparametric Approach to Density Estimation," *Journal of the American Statistical Association*, **82**, 858–865.

Pregibon, D. (1982), "Resistant Fits for Some Commonly Used Logistic Models with Medical Applications," *Biometrics*, **38**, 485 – 498.

Ridgeway, G. (1999), "The State of Boosting," *Computing Science and Statistics*, **31**, 172 – 181.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, **70**, 41 – 54.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, **66**, 688 – 701.

Setoguchi, S., Schneeweiss, S., Brookhart, M.A., Glynn, R.J. and Cook, E.F. (2008), "Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study," *Pharmacoepidemiology and Drug Safety*, **17**, 546 – 555.

Shinohara, E.T., Mitra, N., Guo, M. and Metz, J.M. (2008), "Radiation Therapy is Associated with Improved Survival in the Adjuvant and Definitive Treatment of Intrahepatic Cholangiocarcinoma," *International Journal of Radiation Oncology Biology Physics* **72**, 1495 – 1501.

Stefanski, L.A. and Boos, D.D. (2002), "The Calculus of M-Estimation," *The American Statistician*, **56**, 29–38.

Tchernis, R., Horvitz-Lennon, M. and Normand, S.L.T. (2005), "On the Use of Discrete Choice Models for Causal Inference," *Statistics in Medicine*, **14**, 2197 – 2212.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica: Journal of the Econometric Society*, **50**, 1–25.

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, **96**, 574 – 588.

Yuan, Z. and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, **100**, 1202 – 1214.

Zhang, T. and Yu, B. (2005), "Boosting with Early Stopping: Convergence and Consistency," *The Annals of Statistics*, **33**, 1538 – 1579.