Communications for Statistical Applications and Methods
2016, Vol. 23, No. 1, 1–20

http://dx.doi.org/10.5351/CSAM.2016.23.1.001
Print ISSN 2287-7843 / Online ISSN 2383-4757

Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores

Joseph Kang[a], Wendy Chan[1,b], Mi-Ok Kim[c], Peter M. Steiner[d]

[a]Department of Preventive Medicine, Northwestern University, USA
[b]Department of Statistics, Northwestern University, USA
[c]Department of Pediatrics, University of Cincinnati and
Cincinnati Children's Hospital Medical Center, USA
[d]Department of Educational Pscychology, University of Wisconsin, USA

Abstract

Causal inference methodologies have been developed for the past decade to estimate the unconfounded effect of an exposure under several key assumptions. These assumptions include, but are not limited to, the stable unit treatment value assumption, the strong ignorability of treatment assignment assumption, and the assumption that propensity scores be bounded away from zero and one (the positivity assumption). Of these assumptions, the first two have received much attention in the literature. Yet the positivity assumption has been recently discussed in only a few papers. Propensity scores of zero or one are indicative of deterministic exposure so that causal effects cannot be defined for these subjects. Therefore, these subjects need to be removed because no comparable comparison groups can be found for such subjects. In this paper, using currently available causal inference methods, we evaluate the effect of arbitrary cutoffs in the distribution of propensity scores and the impact of those decisions on bias and efficiency. We propose a tree-based method that performs well in terms of bias reduction when the definition of positivity is based on a single confounder. This tree-based method can be easily implemented using the statistical software program, R. R code for the studies is available online.

Keywords: causal inference, propensity score, positivity assumption, classification and regression tree, jackknife resampling, inverse propensity weighting, random forest

1. Introduction

Randomized experiments have often been considered the gold standard by which causal relationships between variables can be established. When a treatment is randomly assigned to subjects in a study, randomization minimizes the potential confounding of treatment effects from any pre-treatment variables and the systematic differences induced by these variables. This desirable feature of randomized experiments strengthens the internal validity of the inferences from the experimental study. However, randomized experiments are not always feasible in practice and interest in causal relationships in observational and quasi-experimental studies has led to a well established suite of methods for causal inference in these contexts. When treatment is not randomized, as in the observational study, assumptions are needed to derive unbiased estimates of the treatment effect. One such assumption is known

[1] Corresponding author, Co-first author: Department of Statistics, Northwestern University, Evanston, IL 60208, USA.
E-mail: wendychan2016@u.northwestern.edu

Published 31 January 2016 / journal homepage: http://csam.or.kr

as the positivity assumption (Westreich and Cole, 2010). In nonrandomized studies, the positivity assumption addresses the probability of receiving treatment, but this probability must be modeled in the absence of randomization. Propensity score methods are commonly used to estimate probabilities of receiving treatment by conditioning on observable confounders that are considered relevant to the potential treatment effects (Rosenbaum and Rubin, 1983). Formally, the positivity assumption requires that the propensity score, for all values of the treatment and all combinations of values of the confounders, be strictly between 0 and 1. In any study in which the positivity assumption is not met, estimates of treatment effects may be biased when the analysis includes subjects whose probabilities of receiving treatment violate this assumption (Hong, 2010; Schafer and Kang, 2008). For nonrandomized experiments, the problem of positivity implies that there exist subjects of one treatment group that have no comparable subjects in an alternative treatment condition so that estimates of causal effects for these individuals is questionable.

Violations of the positivity assumption are indicative of overlap problems in the distributions of the propensity scores, and consequently, of the observed confounders, among the treatment groups. Inference for subjects that lie off of the common support of the propensity scores is difficult and often requires extrapolation methods. A common method of addressing this overlap concern and identifying "non-positivity" subjects is to truncate the distribution of propensity scores so that only those subjects whose propensity scores lie on the common support are included in the analysis. Typically, this truncation is done using specified calipers or optimal caliper widths on the distribution of propensity scores (Austin, 2011). Crump *et al.* (2009) proposed an alternative method that characterizes optimal subsamples by which the average treatment effect can be precisely estimated using the distributions of the propensity scores. Further work on the assessment of the positivity assumption has been done in previous studies (Petersen *et al.*, 2012; Westreich and Cole, 2010) using parametric models with the propensity score being randomly close to values of 0 or 1.

Current methods of addressing violations in the positivity assumption concern the detection and identification of "non-positivity" subjects and this identification is inherently connected with the estimation of propensity scores. While these current methods of addressing "non-positivity" perform well in practice, they depend on the specification of the propensity score model. Because the true propensity scores in observational studies are unknown in practice, estimation of the propensity scores is sensitive to the assumptions of the model in model-based approaches. This paper seeks to contribute to the current methods of detecting "non-positivity" by considering cases where there exist subjects with propensity scores that are systematically 0 and 1. In practice, identification of "non-positivity" subjects in these contexts is challenging because model based estimation of propensity scores may not predict a propensity score of 0 or 1. We consider a new, yet simple, tree-based method to detect "non-positivity" subjects that facilitates estimation of causal effects by excluding systematic "non-positivity" subgroups using classification and regression trees (CART) and Jackknife resampling methods. An advantage of tree-based methods is that propensity scores of 0 and 1 are accommodated. The proposed method shares similarities with established methods in Crump *et al.* (2009), but utilizes machine learning methods to precisely identify systematic "non-positivity" subjects. We evaluate combinations of various propensity score models, both parametric and nonparametric, with several causal inference methodologies such as matching with propensity scores, inverse propensity weighting (IPW), and regression-based G-computation methods in the presence of systematic "non-positivity" subjects. For each combination of causal inference methods and propensity score model, arbitrary cutoffs in the propensity scores were used to assess the performance of each method in terms of bias and efficiency. We illustrate that this method performs adequately when "non-positivity" is defined by a single confounder, specifically with a categorical confounder.

The paper is organized as follows. In Section 2, we begin with a brief review of the causal inference framework and propensity score methods. In Section 3, we review current tree-based methods such as the CART method, boosted regression and random forest used to estimate propensity scores. Section 4 details the new proposed method of removing "non-positivity" subjects using tree based methods to estimate the propensity scores. Section 5 reviews various causal inference methods based on estimated propensity scores and these methods include IPW, matching with propensity scores, and G-computation. Section 6 provides the simulation results of the conventional and new methods. Section 7 concludes with a discussion.

2. Review of causal inference framework and propensity scores

2.1. Defining causal effects

Rubin (1974, 1976, 1977, 1980a, 1980b, 1986) introduced the framework for defining causal effects and provided extensions of this framework to different study designs. Throughout this paper, we consider the simple case where we assume that the exposure, T, and outcome, Y, are binary random variables. Variables denoted in capital letters indicate random variables and lower case letters are their observed values. Consider a clinical study of the causal effect of smoking (denoted by $T = 1$ if smoking, $T = 0$ otherwise) on a binary heart outcome ($Y = 1$ if adverse lung condition; $Y = 0$ otherwise). In the interest of notational simplicity, we suppress i, the indicator for each study subject. Each study subject can have two potential outcomes, that is, $Y(1)$ the potential outcome under $T = 1$ and $Y(0)$ under $T = 0$. Important covariates include measured confounders X and possible unmeasured confounders U as shown in Table 1. The measured confounders X include pretreatment variables such as patient history. In the case of continuous outcomes, one expression of the causal effect is $\mathrm{E}(Y(1)) - \mathrm{E}(Y(0))$, where $\mathrm{E}(A)$ is the mathematical expectation of a random variable A. For binary outcomes, an analogous expression of the causal effect is given by the log odds ratio equation:

$$\mathrm{logit}(\mathrm{E}(Y(t))) = \alpha + \beta \cdot t, \tag{2.1}$$

where $\mathrm{logit}(A)$ is $\log(A/(1-A))$, $\mathrm{E}(Y(1))$ is $P(Y(1) = 1)$, and β is $\mathrm{logit}(\mathrm{E}(Y(1))) - \mathrm{logit}(\mathrm{E}(Y(0)))$, the causal log odds ratio.

Note that for each subject, only one potential outcome can be observed and the unobserved potential outcome is missing depending on the condition of the exposure T. Thus, T determines which potential outcome is observed. Suppose that T is randomized, so that T "discloses" the potential outcomes purely at random, regardless of (un)measured confounders. Then, because of randomization, the missing potential outcome of $Y(1)$ has the same distribution as the observed $Y(1)$ so that $\mathrm{E}(Y(1)) = \mathrm{E}(Y(1)|T = 1) = \mathrm{E}(Y|T = 1)$. Thus, randomization enables the observed log odds ratio

$$\mathrm{logit}(\mathrm{E}(Y(1)|T = 1)) - \mathrm{logit}(\mathrm{E}(Y(0)|T = 0)) \tag{2.2}$$

to be causal in estimating β in Equation (2.1).

However, if T is not randomized, the treatment effect is confounded (correlated) with the confounders X and U. Thus, the potential outcomes $Y(1)$ and $Y(0)$ are dependent on T, and also on X and U as well. In most causal inference studies, researchers try to measure all confounders such that all or at least most of the confounding bias is removed. Thus far, there is no method developed to test this assumption on bias removal, but sensitivity analyses exist to analyze the effect of unmeasured confounders if it exists (Lin *et al.*, 1998; Rosenbaum, 2002; Brumback *et al.*, 2004; Shen *et al.*, 2011).

Table 1: Potential outcomes framework

Subject	X	U	T	Y	$Y(1)$	$Y(0)$
$i = 1$	x_1	u_1	1	Observed	Observed	Missing
$i = 2$	x_2	u_2	1	Observed	Observed	Missing
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = m$	x_m	u_m	1	Observed	Observed	Missing
$i = m + 1$	x_{m+1}	u_{m+1}	0	Observed	Missing	Observed
$i = m + 2$	x_{m+2}	u_{m+2}	0	Observed	Missing	Observed
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = n$	x_n	u_n	0	Observed	Missing	Observed

2.2. Propensity scores

In the absence of treatment randomization, conditioning on observable confounders creates a pseudo-randomized data set where T is disassociated with X. However, when X includes multiple confounders, the matching problem by which units in the treatment group, $T = 1$, are compared to those in the control, $T = 0$, quickly becomes a multidimensional problem. Propensity scores are scalar valued summaries of multi-dimensional confounders, where the propensity score of receiving treatment is defined as $e(X) \equiv P(T = 1|X = x)$. In a randomized study, $e(x)$ is a constant, usually 0.5, regardless of measured and unmeasured confounders. In non-randomized experiments, the true propensity scores are unknown since the mechanism by which subjects receive treatment is unknown. The propensity scores therefore must be estimated from observed data, and in an observational study, $e(x)$ can range from 0 to 1. Propensity scores are balancing scores where matching by the propensity scores is equivalent to matching by all the confounders used to estimate the propensity scores (Rosenbaum and Rubin, 1983). In matching, this reduces the multidimensional nature of the problem to a unidimensional nature.

2.3. Assumptions for causal inference with propensity scores

Because nonrandomized studies lack the desirable properties of randomized treatment assignment, assumptions are needed to derive unbiased estimates of causal effects in observational studies. While the positivity assumption is the focus of this study, additional assumptions are necessary to define treatment effects in the propensity score framework.

Assumption 1. (Stable unit treatment value assumption (SUTVA) (Rubin, 2005)) *If SUTVA holds, the potential response of an individual is independent of the mechanism used to assign treatment and is independent of the treatments assigned to other individuals. Specifically, consider a study with N individuals and a binary treatment T indexed by $T = 0, 1$ and the outcome of interest, Y. Using this notation, let $Y_i(T)$ denote the potential outcome of individual i when exposed to treatment T, $T = 0, 1$. Under SUTVA, if individual i, $i = 1, \ldots, N$ is exposed to treatment T, the observed value of Y will be $Y_i(T)$ (Rubin, 1980a). Note that under SUTVA, $Y_i(t)$ depends only on the treatment T_i that individual i received and not on the treatment that another individual i' received, where $i \neq i'$. In addition, SUTVA implies that there are no unrepresented versions of the treatments so that $Y_i(T)$ is independent of which version of T was administered (Rubin, 1980a).*

Assumption 2. (Consistency (Cole and Frangakis, 2009)) *Under this assumption, an individual's potential outcome under the observed exposure is the individual's observed outcome. That is, $Y_i^{\text{obs}} = Y_i(t)$ if $T_i = t$.*

Assumption 3. (Strong ignorability of treatment assignment (Rosenbaum and Rubin, 1983)) *Let $e(X)$ be the estimated propensity score and $T = 1$ be the indicator of receiving the treatment condition. Under strong ignorability, $(Y(1), Y(0)) \perp T|e(X)$ and $0 < \mathrm{P}(T = 1|e(X)) < 1$. Strong ignorability is typically met in studies where treatment is randomly assigned to individuals. Importantly, the strong ignorability assumption implies that X contains all confounders that explain variation in the potential treatment effects and that the probability of any individual i receiving a treatment T_i is nonzero.*

In addition to these assumptions and positivity, unbiased estimation of the treatment effects also requires correct specification of the causal model. Given that these assumptions hold, various methods have been proposed to derive unbiased estimates of the average causal effect. These include matching (Rosenbaum, 2010), stratification (Rosenbaum, 2010), and inverse of propensity weighted methods (Robins *et al.*, 2000), all of which will be reviewed briefly in the later sections. Before discussing these methods of estimation of causal effects, we begin with a summary of existing tree-based methods of estimation of propensity scores that can predict propensity scores of 0 and 1.

3. Tree-based methods for estimation of propensity scores

3.1. Classification and regression trees (CART)

Decision trees (Breiman *et al.*, 1984) are a class of nonparametric methods used in machine learning to solve classification and regression problems. Tree-based methods partition an entire data set into different subgroups based on values of significant predictors. For example, consider a regression problem with continuous outcome Y and related binary predictors X_1, X_2 (0 or 1). Let S_1 denote an entire undivided sample and LS_1 denote the sum of the squared differences between Y and its mean, namely $LS_1 = \sum_{i \in S_1}(y_i - \overline{y}_{S_1})^2$, for subject i who belongs to S_1. As usual, the sum of squared differences assesses the distance between Y and its mean. The goal of CART is to reduce this distance by partitioning the data so that

$$\sum_{i \in S_1} (y_i - \overline{y}_i)^2 >> \sum_{k=1}^{K} \sum_{i \in S_k} \left(y_i - \overline{y}_{S_k}\right)^2, \tag{3.1}$$

where k indicates subgroups with K being its total number and "$A >> B$" indicates that B is minimally smaller than A to the point that further division of the data is spurious. The distance measure $\sum_{i \in S_k}(y_i - \overline{y}_i)^2$ in (3.1), which CART aims to optimize, is called the "impurity function" (Breiman *et al.*, 1984) and can easily be replaced by likelihood (Kang, 2012). Now given predictors X_j, $j = 1, 2$, CART will evaluate $\sum_{i \in \{X_j = 0\}}(y_i - \overline{y}_i)^2 + \sum_{i \in \{X_j = 1\}}(y_i - \overline{y}_i)^2$, which is denoted by LS_{X_j}. Here LS_{X_1} indicates the sum of the sums of squared differences between Y and its mean for two groups divided by the X_1 values (0 or 1). If X_1 is more predictive of and correlated to Y compared to X_2, then LS_{X_1} will be much smaller than LS_{X_2}. That is, X_1 divides the data set into more homogeneous subgroups compared to X_2 so that the average of the $y's$ in subgroups divided by X_1 have smaller variance (or standard deviation) from real data compared to the average of the $y's$ in subgroups by X_2 (Breiman *et al.*, 1984). This can be generalized to $p > 2$ confounders: $X_1, X_2, \ldots, X_p$, where CART evaluates all these confounders as described above. The case with continuous confounders is defined analogously. Because different values of the same confounder can be used to determine each splitting, CART can evaluate the same confounder again, recursively. For this reason, the CART algorithm is sometimes called "recursive partitioning" (Zhang and Singer, 1999). This algorithm can be implemented in the R package RPART.

CART determines the total number of subgroups of a tree by pruning similar subgroups using a cost-complexity function

$$R_\alpha(T) = R(T) + \alpha|\breve{T}|, \tag{3.2}$$

where $R(T)$ is the average within group sum of squares $\sum_k^K \sum_{i \in S_k}(y_i - \bar{y}_i)^2$ and $|\breve{T}|$ is the number of subgroups of the subtree. The parameter α is the complexity parameter that contributes a penalty term for each additional subgroup (Deconinck *et al.*, 2005). The parameter α is estimated with cross-validation techniques (Deconinck *et al.*, 2005).

In the case of a binary outcome, CART uses the entropy function $-p \cdot \log(p) - (1-p) \cdot \log(1-p)$, where p is the proportion of the binary response, as an impurity function to be used in equation (3.1). The Bayes error, $\min(p, 1-p)$, as well as the Gini index, $p \cdot (1-p)$ are also used. All these impurity functions will be zero if p is zero (Zhang and Singer, 1999). In particular, the entropy function is 0 because $0 \cdot \log(0)$ is defined to be 0 (Zhang and Singer, 1999). This is an important property of CART in terms of excluding subjects who always defy or accept a certain treatment.

Despite the many merits of CART, it has received criticism on its stability and other well-known issues (Hastie *et al.*, 2001). Though stability is a concern when interpreting the model, the focus of this paper is in the ability of a single CART tree to detect systematic "non-positivity" subjects. In the following section, we review the established methods that bypass some of the limitations of a single tree.

3.2. Random forest

Random Forest (RF) is a method that builds a forest of CART trees from bootstrap samples of observed data. RF is closely related to the concept of bagging. "Bagging" uses bootstrap aggregation to aggregate predictions that are repeatedly fitted based on bootstrap samples. It is well known that a bagging estimator gives better prediction than one single prediction model (Hastie *et al.*, 2001, p.303). Unlike traditional single CART analysis, RF does not require pruning. Yet building a forest of many trees can bypass criticisms of the single CART method and enhance the prediction of a single CART tree model by averaging the entire forest of trees (Hastie *et al.*, 2001, p.600). While mathematical properties of RF are still under study (Biau, 2012), the clear idea of building many trees in RF is to improve unbiased prediction solely by reducing variance (Hastie *et al.*, 2001, p.600). In particular, RF uses randomly selected subsets of predictors at each splitting. Given a p-dimensional vector of predictor variables, a subset of variables is used to determine the best split in the tree (Breiman, 2001). The R package randomForest, which was used in our study, has default values: $\sqrt{p}$ for categorical responses and $p/3$ for continuous responses.

One criticism that is still under debate against RF is its potential overfitting (Hastie *et al.*, 2001, p.596). RF, while it does not prune, builds trees that are as different from amongst themselves as possible so that correlations among the trees become smaller and hence the overall variance of the prediction becomes smaller. Generally the performance of RF has been assessed with various measures for prediction errors such as misclassification errors and average absolute errors (Hastie *et al.*, 2001, pp. 589–593). While these measures evaluate the general performance of RF over the entire training or test data set, we would like to focus on the particular subgroups where propensity scores (predictions) are strictly zero or one. In the next section, we explain the boosting method as a competitor model that has been used in observational studies.

3.3. Boosting method (Generalized Boosted Model; GBM)

One of the most widely used parametric methods of estimating propensity scores is with the logistic

regression model. However, since the inception of the boosted regression model, which can be implemented in R packages GBM and TWANG, the boosting method has become readily available to the public as a well-performing nonparametric procedure to estimate the true propensity score model (McCaffrey *et al.*, 2004). The estimation goal of the logistic regression and that of the boosted model (option "`bernoulli`" in GBM package) is the same: both of them aim to achieve the maximum Bernoulli probability likelihood. Yet, they differ greatly: logistic regression requires that the functional form be known, that is, the main and/or interaction effect terms, while the boosting method does not make this requirement. The boosting model utilizes a collection of simple regression tree models that are added together to estimate the propensity scores (McCaffrey *et al.*, 2004).

Boosting is a simple algorithm that iteratively and additively updates the candidate model to yield the best prediction model. To illustrate this, consider a typical algorithm that R package GBM estimates, $p(x) = P(Y = 1|x)$ with its default "`Bernoulli`" option for a binary response. The clear goal of the "`gbm`" function with the default option "`bernoulli`" is to maximize the log-likelihood of the Bernoulli distribution with $g(x) = \text{logit}(p(x))$:

$$\sum_i \left[y \cdot g(x) - \log\left(1 + \exp\left(g\left(x\right)\right)\right)\right], \tag{3.3}$$

where the functional form of $p(x)$ in $g(x)$ is relaxed and in fact, unspecified. The log-likelihood can be maximized with respect to $g(x)$, which results in $r(x) = y - (1 + \exp(-g(x)))^{-1}$. Note that the nonparametric way in which the confounders, x, enter the prediction model is an interesting feature of boosted regression. This is seen in greater detail in the boosting algorithm (Freund *et al.*, 1999).

Note that compared to a single tree from CART, the aggregation and inclusion of interactions among the variables in the boosting algorithm and model may restrain $p(x) = \text{expit}(\sum_{b=1}^{B} g(x)^{(b)})$, for b iterations, away from extreme values of 0 and 1 because of the aggregated nature. This constraint away from zero and one may or may not be beneficial. It may be beneficial because when $p(x)$ is used as the propensity model in the IPW method, the estimation is relatively more stable as shown in our simulation study. But it may not be beneficial in detecting extreme propensity values as it approaches 0 or 1 compared to CART. We investigate the performance of this popular boosting method, packaged in GBM, to estimate the propensity model and binary outcome model in causal inference for observational studies with falsely included subjects whose propensity scores are 0 or 1.

4. JCART

In the earlier section regarding CART, we introduced a simple idea of detecting subgroups of subjects whose positivity assumption is systematically violated. These non-positivity subgroups, which were identified by one single tree, may affect the analysis and lead to spurious conclusions due to the instability of a single CART. To avoid such spuriosity, we use multiple CART estimates based on a randomly divided Jackknife resampling scheme (Phillip, 2001). This method is referred to as the Jackknifed CART (JCART) method.

Let θ denote the target causal estimand, which, in the case of a binary outcome, may be the causal odds ratio, $\text{logit}(\text{E}(Y(1)) - \text{logit}(\text{E}(Y(0)))$. The JCART method resamples $\text{E}(Y(t))$, $t = 0, 1$ using the following steps:

1. Randomly divide the whole data into J groups. Each group has $(1/J) \cdot 100\%$ of the original data set.

2. Delete group j out of J randomly divided groups and use the rest of the $(J - 1)$ groups, denoted

as L_j, for growing the CART with 10 folds cross-validation ("`xval=10`," the default option in R package RPART).

3. Repeat Steps 1–2 for each of the J random samples so as to get J estimates, $\hat{\theta}_1, \ldots, \hat{\theta}_J$, where $\hat{\theta}_j$ is the causal estimator of L_j. For each L_j group, let s_j indicate subgroup s of sample L_j and $P(s_j)$ be the proportion of subjects in subgroup s_j of L_j. Note that the subgroups s_j are the homogeneous subgroups partitioned by the CART method in which the treatment effect is most precisely estimated. The causal estimator of L_j is given by $\text{logit}(\text{E}(Y_{s_j}(1))) - \text{logit}(\text{E}(Y_{s_j}(0)))$, and $\text{E}(Y_{s_j}(1))$ is estimated by $\sum_{s_j} E(Y_{s_j}|t_{s_j} = 1)P(s_j)$, where $E(Y_{s_j}|t_{s_j})$ can be simply modeled by $\text{expit}(\hat{\alpha}_{s_j} + \hat{\beta}_{s_j} \cdot I(T_{s_j}{=}t_{s_j}))$. One method of estimating the coefficients $\hat{\alpha}_{s_j}, \hat{\beta}_{s_j}$ is through logistic regression using the subjects in subgroup s_j.

The Jackknife point and variance estimates, respectively, are given by

$$\hat{\theta}_{\text{Jack}} = \frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_j, \tag{4.1}$$

$$\text{Var}\left(\hat{\theta}_{\text{Jack}}\right) = \frac{J-1}{J} \sum\nolimits_j \left(\hat{\theta}_j - \hat{\theta}_{\text{Jack}}\right)^2. \tag{4.2}$$

Note that the Jackknife estimator $\hat{\theta}_{\text{Jack}}$ is the average of $\hat{\theta}_j$, $j = 1, \ldots, J$. An advantage of the Jackknife method over bootstrapping is that if one bootstrap sample is cross-validated with a $9:1$ ratio, the training sample of this bootstrap has only about 56.7% ($= 63 \cdot 9/10$) of the original data (Kleiner *et al.*, 2012). Using the bootstrap sample would lower the statistical power to assess true causal estimands and would produce possibly unnecessarily parsimonious CART models that may subsequently result in the under-adjustment of confounding variables. Alternatively, the Jackknife method gives about 87.75% ($= 39/40 \cdot 100 \cdot 9/10$) of the original data when $J = 40$, an idea that was also used in recent work by Su *et al.* (2012). We use $J = 40$ for the simulation study in subsequent sections. JCART uses the nonparametric methods of CART to estimate propensity scores, with propensities of 0 and 1 possible, while also inheriting the properties of unbiased point estimation and efficiency in variance estimation of the Jackknife resampling method. Table 2 summarizes the desirable features of each of the tree-based methods of propensity score estimation used in the following study.

5. Existing methods of causal inference using propensity scores

In this section, we review several causal inference methods used to estimate treatment effects based on estimated propensity scores.

5.1. The IPW method

IPW is a method that was proposed to reduce bias in the estimation of population quantities (Horvitz and Thompson, 1952). IPW has also been used in epidemiology to draw causal conclusions (Robins *et al.*, 2000). Let $\mathbf{X}$ denote a vector of confounders that contains some measured baseline information; T an exposure (or an intervention) variable; and Y a health outcome. Consider $\text{E}(Y = 1|T = t)$, which can simply be modeled using

$$\text{E}(Y = 1|T = t) = \text{expit}\left(\alpha^* + \beta^* \cdot I\left(T = t\right)\right) \tag{5.1}$$

where expit(a) is exp(a)/(1+exp(a)).

Table 2: Comparison of tree based methods of propensity score estimation

Method	Features	Drawbacks
CART	• Partitions dataset based on values of specified predictors • Can predict propensity scores of 0 and 1 using entropy function	• Small changes in the target data set can cause instability in the model • Selection of only one predictor at each node of the CART tree
RF	• Partitions data using multiple CART trees from bootstrap samples • Uses subsets of predictors for division of homogenous groups	• Potential overfitting
GBM	• Estimates propensity scores using an aggregated collection of regression trees • Functional form of propensity score model may be left unspecified	• Aggregated nature of method may not predict propensity scores of 0 and 1
JCART	• Uses CART on Jackknife subsets of data to estimate propensity scores • Estimates are aggregated over all Jackknife samples	• Selection of only one predictor at each node of the CART tree

CART = Classification and Regression Trees, RF = Random Forest, GBM = Generalized Boosted Model, JCART = Jackknifed CART

Since $\mathbf{X}$ confounds the relation between Y and T, β^* is subject to bias. The IPW method weights estimating equation (5.1) with the inverse of the propensity score in order for β^* to be the causal parameter β in equation (2.1). In particular, the weight for the causal effect of T is:

$$\text{sw} = \frac{P(T = t)}{P(T = t|X = x)}. \tag{5.2}$$

Weight sw in equation (5.2) is called a stabilized weight because the numerator probability makes the weight stabilized (or smaller) and makes the conditional expectation of the weights equal to the numerator (Robins *et al.*, 2000).

5.2. Matching

Matched sampling was first introduced as a method to treat an observational study data set as a pseudo-randomized data set where groups are comparable or balanced with respect to baseline confounding factors (Rosenbaum and Rubin, 1983). Matching and the IPW method are tools commonly used to achieve the goal of turning an observational data set into a pseudo-randomized data set. In other words, both matching and IPW methods aim to disassociate the exposure variable from measured baseline confounding variables.

However, matching methods do not necessarily weight a data set in general (except Full matching (Rosenbaum, 2010)) while the IPW method weights each individual with the inverse probability. The matching method subsamples subjects from the exposure group ($T = 1$) and the control group ($T = 0$) based on the baseline confounding variables. However, because of the multidimensional nature of the confounding variables, summary measures such as the propensity score have also been used in creating matched groups. If the propensity score is used, it serves as a balancing score to balance the distributions of the confounders among the comparison groups. In this paper, three matching methods that are used are "full matching," "nearest-neighbor matching," and "coarsened exact matching." Matching methods were implemented using the R package MATCHIT with options "`cem`," "`nearest`," and "`full`" for coarsened exact matching, nearest neighbor matching, and full matching, respectively.

5.3. G-computation (regression estimates) of the interaction effects

As Table 1 in Section 2 shows, $Y(t)$ has missing data because both potential outcomes cannot simultaneously be observed. But $\mathrm{E}(Y(t))$ can be predicted (or imputed) with $\mathrm{E}(Y(t)|T = t, x; \theta_t)$, where θ_t is a vector of coefficients of the logistic regression model $\mathrm{E}(Y|T = t, x; \theta_t) = \mathrm{expit}(x\theta_t)$. In particular, $\mathrm{E}(Y|T = t, x)$ is called the Q-model in the G-computation literature (Snowden *et al.*, 2011). Note that t and x in $\mathrm{E}(Y|t, x)$ can have complex interactions. The interaction terms make the coefficient of t difficult to interpret as a causal effect and subsequently facilitate the reluctance of considering interaction terms for t and x in clinical studies. Yet the Q-model ($\mathrm{E}(Y|t, x)$) allows these interaction terms because the Q-model is used for predicting potential outcomes.

The Q-model is built in the following way. We fit a logistic regression model to estimate θ_t and predict $\mathrm{E}(Y(t)|x; \theta_t) = \pi(t; \theta_t)$. Let $\pi(t; \theta_t)$ denote $\mathrm{E}(\mathrm{expit}(x\theta_t)) = \int \mathrm{expit}(x\theta_t)dF(x)$, which is estimated by $\widetilde{\pi}(1; \hat{\theta}_t) = (1/n)\sum_{i=1}^{n}(\mathrm{expit}(x\hat{\theta}_t))$. This technique is called G-computation (Robins, 1986; Taubman *et al.*, 2009; Westreich *et al.*, 2012) or regression estimation. From this, the causal effect (causal odds ratio) is estimated by

$$\eta = \frac{\widetilde{\pi}\left(1; \hat{\theta}_1\right) \cdot \left(1 - \widetilde{\pi}\left(0; \hat{\theta}_0\right)\right)}{\widetilde{\pi}\left(0; \hat{\theta}_0\right) \cdot \left(1 - \widetilde{\pi}\left(1; \hat{\theta}_1\right)\right)}. \tag{5.3}$$

Because the Q-model is the most important basis for predicting potential outcomes, the relation between T and X should be correctly modeled. With logistic regression, this is done by using linear combinations of T and X. Targeted maximum likelihood estimation (TMLE) is another approach that yields an efficient estimator, especially in the case when the starting density is incorrectly specified (van der Laan, 2013). Tree-based machine learning models consider nonparametric forms of interactions between T and X. Modeling $\mathrm{E}(Y|T = t, x; \hat{\theta}_t)$ requires careful effort because the lack of overlap in the support of x among the two groups defined by the binary conditions of T would result in the extrapolation of the imputation. If the positivity assumption were violated, the extrapolation over propensity scores of zero or one is overt. Results from our simulation study report biases with the G-computation method when systematic "non-positivity" subjects are falsely included.

6. Simulation study designs and analysis results

To analyze the performance of the tree-based methods of propensity score estimation in detecting systematic "non-positivity" subjects, we performed four simulation studies. The simulation studies were constructed using combinations of two factors, number of confounders and definition of "non-positivity." For the latter, "non-positivity" was considered in terms of the values of a specific confounder since the JCART method partitions the data based on the values of individual predictors. Because the partitioning procedure of tree-based methods performs well when the specific predictor is categorical, we explore the performance of each tree based method when the definition of "non-positivity" has a simple definition (based on one predictor) and a slightly complex definition (based on more than one predictor). In addition, since the detection of subjects with $e(X) = 0, 1$ is the goal of this paper, we hypothesize that propensity score estimation methods that predict such propensity scores will yield estimates of treatment effects that are the most efficient.

6.1. Simulation setup

The simulation used in this study comprised of four separate simulation studies, denoted by $\mathrm{Sim}_1, \mathrm{Sim}_2, \mathrm{Sim}_3, \mathrm{Sim}_4$. The sample size for each simulation was set at $N = 1000$ and the generated confounders,

Table 3: Description of simulations

Simulation	Number of Confounders	Definition of "Non-positivity"
Sim_1	4	X_1
Sim_2	4	$X_{12} = X_1 + X_2$
Sim_3	40	X_1
Sim_4	40	$X_{12} = X_1 + X_2$

Table 4: Assignment of non-positivity subjects

Propensity scores ($e(X)$)	$\text{Sim}_1, \text{Sim}_3$	$\text{Sim}_2, \text{Sim}_4$
$e(X) = 0$	$X_1 = 0.1$	$X_{12} \leq Q(0.10, X_{12})$
$e(X) = 1$	$X_1 = 1.0$	$X_{12} \geq Q(0.90, X_{12})$

Note: $Q(p, \lambda)$ denotes the p quantile of the quantity λ.

Table 5: Propensity structure for our simulation study

$e(X)$	$\text{E}(Y(1))$	$\text{E}(Y(0))$
$0 < e(x) < 1$	0.6	0.4
$e(x) = 0$	0.1	0.9
$e(x) = 1$	0.1	0.9

X_i, $i = 1, \ldots, 40$, were all categorical. Sim_1 and Sim_2 were completed with four categorical confounders and $\text{Sim}_3, \text{Sim}_4$ were completed with 40 categorical confounders. For each pair of simulations, the "non-positivity" subjects were defined by one categorical confounder, X_1, for Sim_1 and Sim_3, and by the sum of two categorical confounders, $X_{12} = X_1 + X_2$ for Sim_2 and Sim_4. The 10% and 90% quantiles of X_{12} were used to define the "non-positivity" subjects for Sim_2 and Sim_4. A description of the simulations is given in Table 3. The X_i were randomly sampled, with equal probability, from the interval $[0.1, 1]$. Table 4 describes the structure used to define the "non-positivity" group for each simulation, Sim_i, $i = 1, \ldots, 4$.

The true propensity scores, $e(X)$, were generated by the model $\text{expit}(\sum_{i=1}^{J/2} X_i - \sum_{i=J/2+1}^{J} X_i)$, where $J = 4$ for $\text{Sim}_1, \text{Sim}_2$ and $J = 40$ for $\text{Sim}_3, \text{Sim}_4$, and the propensity scores were used to generate the binary exposure variable T. The binary potential outcomes, $Y(t)$, $t = 0, 1$, were sampled using the probability structure given in Table 5. For example, subjects with $0 < e(x) < 1$, will have their realized binary potential outcome $Y(1)$ ssampled with fixed probability 0.6 for $E(Y(1))$ and probability 0.4 for $E(Y(0))$. Also, subjects with $e(x) = 0$ will have realized potential outcomes $Y(1)$ and $Y(0)$ sampled with respective probabilities 0.1 and 0.9. The probabilities were equivalent for both groups of "non-positivity" subjects so that both groups would have the same level of extreme outcomes. Note that these potential outcomes, $Y(1), Y(0)$, were sampled with fixed probabilities within the subgroups defined by the propensity scores. The subgroups themselves were defined based on the ranges of the confounders X_i as described in Table 4. For example, under Sim_1 and Sim_3, if the covariate X_1 had values 0.1 or 1, the potential outcome $Y(1)$ for these subjects was sampled with fixed probability 0.1 whereas if the covariate X_1 had values between 0.1 and 1, the potential outcome $Y(1)$ was sampled with probability 0.6. While the potential outcomes are typically functions of the covariates, studies have also been done where the potential outcomes are functions of propensity scores (Little and An, 2004). In the current context, it is impossible in this data set to estimate $\text{E}(Y(1))$ for subjects with $e(x)=0$ so that they should be removed from the study. Similarly, subjects with $e(x) = 1$ should also be removed. The observed outcome Y, which was used for analysis, was defined as $Y = T \cdot Y(1) + (1 - T) \cdot Y(0)$.

The performance of JCART was compared with IPW estimation, G-computation with RF, and

Table 6: Description of propensity score and estimation models and methods

Methods	Description
· JCART	The propensity score model is estimated using CART.
· $\text{JCART}_{[0.025,0.975]}$	The prediction model is based on Jackknife subsamples from the data set.
· $\text{JCART}_{[0.05,0.95]}$	
· $\text{JCART}_{[0.1,0.9]}$	
· Match.near.logit	The propensity scores are estimated using matching methods with distance criterion
· Match.cem.logit	determined by "nearest-neighbor," "coarsened exact matching," and "full matching."
· Match.full.logit	The logistic regression model ("logit") and Random Forest ("rpart") are used in the pro-
· Match.near.rpart	pensity score model.
· Match.cem.rpart	
· Match.full.rpart	
· IPW.glm	Propensity scores are estimated using logistic regression.
· $\text{IPW.glm}_{[0.025,0.975]}$	The estimation model uses the IPW method to estimate causal effects.
· $\text{IPW.glm}_{[0.05,0.95]}$	
· $\text{IPW.glm}_{[0.1,0.9]}$	
· IPW.gbm	Propensity scores are estimated using GBM.
· $\text{IPW.gbm}_{[0.025,0.975]}$	The estimation model uses the IPW method to estimate causal effects.
· $\text{IPW.gbm}_{[0.05,0.95]}$	
· $\text{IPW.gbm}_{[0.1,0.9]}$	
· IPW.rf	Propensity scores are estimated using RF.
· $\text{IPW.rf}_{[0.025,0.975]}$	The estimation model uses the IPW method to estimate causal effects.
· $\text{IPW.rf}_{[0.05,0.95]}$	
· $\text{IPW.rf}_{[0.1,0.9]}$	
· GC.glm	Propensity scores are estimated using a generalized linear model.
· $\text{GC.glm}_{[0.025,0.975]}$	Treatment effects are estimated using the G-computation method.
· $\text{GC.glm}_{[0.05,0.95]}$	
· $\text{GC.glm}_{[0.1,0.9]}$	
· GC.gbm	Propensity scores are estimated using GBM.
· $\text{GC.gbm}_{[0.025,0.975]}$	Treatment effects are estimated using the G-computation method.
· $\text{GC.gbm}_{[0.05,0.95]}$	
· $\text{GC.gbm}_{[0.1,0.9]}$	
· GC.rf	Propensity scores are estimated using RF.
· $\text{GC.rf}_{[0.025,0.975]}$	Treatment effects are estimated using the G-computation method.
· $\text{GC.rf}_{[0.05,0.95]}$	
· $\text{GC.rf}_{[0.1,0.9]}$	

CART = Classification and Regression Trees, JCART = Jackknifed CART, IPW = Inverse Propensity Weighting, GBM = Generalized Boosted Model, RF = Random Forest, GC = G-computation.

GBM in estimating the causal odds ratio. For each tree-based method, a total of 1000 trees were used. Interactions of order 2 and 4 were considered in the GBM model, both of which are higher than the default. However, only the results for interactions of depth order 2 were presented since results under order 4 were similar. To evaluate these methods with "non-positivity" subjects, we assessed the bias and efficiency using combinations of the methods with various truncation ranges in the propensity score values. Cole and Hernán (2008) explored the bias-variance tradeoff by considering several truncations of the weights in IPW. Using this recommendation, the truncated ranges of propensity scores ($[0.025, 0.975]$, $[0.05, 0.95]$, and $[0.1, 0.9]$) were used to assess the ability of the proposed methods to detect and remove subjects whose propensities were zero or one from the analysis. Note that the range $[0.1, 0.9]$ coincides with the recommended range proposed in Crump *et al.* (2009).

6.2. Performance of JCART using different propensity score ranges

A complete list of the methods used to estimate the average causal effect and the propensity scores

Table 7: Results for Simulation Sim_1

Method	Bias (mean)	Bias (median)	RMSE
JCART	0.00	0.00	0.03
$\text{JCART}_{[0.025,0.975]}$	0.00	0.00	0.03
$\text{JCART}_{[0.05,0.95]}$	0.00	0.00	0.03
$\text{JCART}_{[0.1,0.9]}$	0.00	0.00	0.03
Match.near.logit	−0.74	−0.74	0.75
Match.cem.logit	0.10	0.07	0.75
Match.full.logit	−0.54	−0.54	0.56
Match.near.rpart	0.00	0.00	0.05
Match.cem.rpart	0.10	0.07	0.75
Match.full.rpart	−0.02	−0.03	0.28
IPW.glm	−0.57	−0.57	0.57
$\text{IPW.glm}_{[0.025,0.975]}$	−0.57	−0.57	0.57
$\text{IPW.glm}_{[0.05,0.95]}$	−0.57	−0.57	0.57
$\text{IPW.glm}_{[0.1,0.9]}$	−0.55	−0.55	0.55
IPW.gbm	−0.57	−0.57	0.58
$\text{IPW.gbm}_{[0.025,0.975]}$	−0.57	−0.57	0.58
$\text{IPW.gbm}_{[0.05,0.95]}$	−0.57	−0.57	0.58
$\text{IPW.gbm}_{[0.1,0.9]}$	−0.57	−0.57	0.58
IPW.rf	−0.64	−0.64	0.65
$\text{IPW.rf}_{[0.025,0.975]}$	−0.05	−0.06	0.07
$\text{IPW.rf}_{[0.05,0.95]}$	0.00	0.01	0.05
$\text{IPW.rf}_{[0.1,0.9]}$	0.01	0.01	0.08
GC.glm	−0.42	−0.42	0.42
$\text{GC.glm}_{[0.025,0.975]}$	−0.42	−0.42	0.42
$\text{GC.glm}_{[0.05,0.95]}$	−0.42	−0.42	0.42
$\text{GC.glm}_{[0.1,0.9]}$	−0.41	−0.41	0.42
GC.gbm	−0.56	−0.56	0.56
$\text{GC.gbm}_{[0.025,0.975]}$	−0.56	−0.56	0.56
$\text{GC.gbm}_{[0.05,0.95]}$	−0.56	−0.56	0.56
$\text{GC.gbm}_{[0.1,0.9]}$	−0.54	−0.54	0.54
GC.rf	−0.46	−0.46	0.46
$\text{GC.rf}_{[0.025,0.975]}$	−0.46	−0.46	0.46
$\text{GC.rf}_{[0.05,0.95]}$	−0.46	−0.46	0.46
$\text{GC.rf}_{[0.1,0.9]}$	−0.45	−0.45	0.46

Note: Four confounders used with nonpositivity subjects defined by X_1 only.
RMSE = Root Mean Square Error, JCART = Jackknifed Classification and Regression Trees,
IPW = Inverse Propensity Weighting, GC = G-computation.

is given in Table 6. In Table 6, the first column, "Method", describes the method used to compute each estimate and is composed of two parts: the estimation method and the propensity score model. The names of the estimates are: "JCART" for the JCART method, "Match.near," "Match.cem," and "Match.full" for the matching methods using "nearest-neighbor," "coarsened exact matching," and "full matching," respectively, with propensity scores estimated by logistic regression and RF. "IPW.glm," "IPW.gbm," and "IPW.rf" indicate the IPW methods using logistic regression, GBM, and RF for the propensity score model, respectively. Finally, "GC.glm," "GC.gbm," and "GC.rf" represent the G-computation method with propensity model based on logistic regression, GBM, and RF, respectively. The ranges in the subscripts attached to each name indicate the propensity score range by which the method was used to estimate the response. All methods, with exception to matching, were used to estimate the response on three truncated ranges of the propensity score: $[0.025, 0.975]$, $[0.05, 0.95]$, and $[0.1, 0.9]$. Methods without a range indicate estimation on the full $[0, 1]$ propensity score range.

Tables 7–10 show the performance of the estimates of the IPW, matching, G-computation, and

Table 8: Results for Simulation Sim_2

Method	Bias (mean)	Bias (median)	RMSE
JCART	−0.54	−0.55	0.56
$\text{JCART}_{[0.025,0.975]}$	−0.51	−0.52	0.52
$\text{JCART}_{[0.05,0.95]}$	−0.46	−0.46	0.47
$\text{JCART}_{[0.1,0.9]}$	−0.30	−0.29	0.32
Match.near.logit	−0.87	−0.86	0.88
Match.cem.logit	0.02	0.05	0.79
Match.full.logit	−0.61	−0.61	0.63
Match.near.rpart	−0.92	−0.94	0.93
Match.cem.rpart	0.02	0.05	0.79
Match.full.rpart	−0.55	−0.54	0.69
IPW.glm	−0.67	−0.67	0.67
$\text{IPW.glm}_{[0.025,0.975]}$	−0.67	−0.67	0.67
$\text{IPW.glm}_{[0.05,0.95]}$	−0.66	−0.66	0.67
$\text{IPW.glm}_{[0.1,0.9]}$	−0.62	−0.61	0.62
IPW.gbm	−0.77	−0.77	0.77
$\text{IPW.gbm}_{[0.025,0.975]}$	−0.77	−0.77	0.77
$\text{IPW.gbm}_{[0.05,0.95]}$	−0.77	−0.77	0.77
$\text{IPW.gbm}_{[0.1,0.9]}$	−0.77	−0.77	0.77
IPW.rf	−0.78	−0.78	0.78
$\text{IPW.rf}_{[0.025,0.975]}$	−0.32	−0.32	0.32
$\text{IPW.rf}_{[0.05,0.95]}$	−0.17	−0.18	0.18
$\text{IPW.rf}_{[0.1,0.9]}$	−0.06	−0.06	0.10
GC.glm	−0.35	−0.35	0.35
$\text{GC.glm}_{[0.025,0.975]}$	−0.35	−0.35	0.35
$\text{GC.glm}_{[0.05,0.95]}$	−0.35	−0.35	0.35
$\text{GC.glm}_{[0.1,0.9]}$	−0.34	−0.34	0.35
GC.gbm	−0.70	−0.70	0.71
$\text{GC.gbm}_{[0.025,0.975]}$	−0.70	−0.70	0.71
$\text{GC.gbm}_{[0.05,0.95]}$	−0.70	−0.69	0.70
$\text{GC.gbm}_{[0.1,0.9]}$	−0.65	−0.64	0.65
GC.rf	−0.57	−0.57	0.57
$\text{GC.rf}_{[0.025,0.975]}$	−0.56	−0.57	0.57
$\text{GC.rf}_{[0.05,0.95]}$	−0.56	−0.55	0.57
$\text{GC.rf}_{[0.1,0.9]}$	−0.52	−0.51	0.53

Note: Four confounders used with nonpositivity subjects defined by $X_{12} = X_1 + X_2$.
RMSE = Root Mean Square Error, JCART = Jackknifed Classification and Regression Trees,
IPW = Inverse Propensity Weighting, GC = G-computation.

JCART methods. The tables are organized as follows, "Bias (mean)" indicates the average difference between estimates of causal effects and the true effect of log(2.25); "Bias (median)" indicates the median of differences between estimates of causal effects and the true effect; RMSE (Root Mean Square Error) is the square root of the average squared difference between causal estimates and the true effect.

Table 7 lists the results for Sim_1, where four confounders were used in the study and one confounder, X_1, determined the "non-positivity" subjects. Table 8 is structured analogously for Sim_2 where the sum of two confounders, $X_{12} = X_1 + X_2$ was used to define the "non-positivity" subjects. From Table 7, when a single categorical confounder was used to define the subjects with $e(X) = 0, 1$, the JCART methods showed superior performance across all measures of bias. Furthermore, several of the other methods that yielded estimates with magnitudes of bias close to zero used the tree-based method RF to estimate the propensity scores. Closer analysis of the IPW estimates with propensities estimated by RF show that the magnitude of the bias becomes smaller as the range of propensity

Table 9: Results for Simulation Sim_3

Method	Bias (mean)	Bias (median)	RMSE
JCART	0.00	−0.01	0.17
$\text{JCART}_{[0.025,0.975]}$	0.00	−0.01	0.17
$\text{JCART}_{[0.05,0.95]}$	0.00	−0.01	0.17
$\text{JCART}_{[0.1,0.9]}$	0.00	0	0.17
Match.near.logit	−0.73	−0.75	0.76
Match.cem.logit	NA	NA	NA
Match.full.logit	−0.61	−0.60	0.65
Match.near.rpart	0.00	−0.02	0.17
Match.cem.rpart	NA	NA	NA
Match.full.rpart	0.00	−0.01	0.43
IPW.glm	−0.64	−0.64	0.66
$\text{IPW.glm}_{[0.025,0.975]}$	−0.62	−0.62	0.64
$\text{IPW.glm}_{[0.05,0.95]}$	−0.60	−0.59	0.62
$\text{IPW.glm}_{[0.1,0.9]}$	−0.58	−0.56	0.61
IPW.gbm	−0.58	−0.59	0.61
$\text{IPW.gbm}_{[0.025,0.975]}$	−0.58	−0.59	0.61
$\text{IPW.gbm}_{[0.05,0.95]}$	−0.58	−0.59	0.61
$\text{IPW.gbm}_{[0.1,0.9]}$	−0.58	−0.59	0.61
IPW.rf	−0.76	−0.77	0.78
$\text{IPW.rf}_{[0.025,0.975]}$	−0.76	−0.77	0.78
$\text{IPW.rf}_{[0.05,0.95]}$	−0.76	−0.77	0.78
$\text{IPW.rf}_{[0.1,0.9]}$	−0.43	−0.44	0.46
GC.glm	−0.55	−0.55	0.57
$\text{GC.glm}_{[0.025,0.975]}$	−0.57	−0.58	0.59
$\text{GC.glm}_{[0.05,0.95]}$	−0.57	−0.57	0.60
$\text{GC.glm}_{[0.1,0.9]}$	−0.57	−0.56	0.60
GC.gbm	−0.57	−0.58	0.60
$\text{GC.gbm}_{[0.025,0.975]}$	−0.54	−0.55	0.57
$\text{GC.gbm}_{[0.05,0.95]}$	−0.52	−0.52	0.55
$\text{GC.gbm}_{[0.1,0.9]}$	−0.49	−0.49	0.52
GC.rf	−0.64	−0.66	0.66
$\text{GC.rf}_{[0.025,0.975]}$	−0.61	−0.63	0.63
$\text{GC.rf}_{[0.05,0.95]}$	−0.59	−0.60	0.61
$\text{GC.rf}_{[0.1,0.9]}$	−0.55	−0.56	0.58

Note: 40 confounders used with nonpositivity subjects defined by X_1 only.
RMSE = Root Mean Square Error, JCART = Jackknifed Classification and Regression Trees,
IPW = Inverse Propensity Weighting, GC = G-computation.

scores is restricted. This is not surprising as truncation of the propensity score range removes subjects whose true propensities lie close to or fall at the extremes.

By comparison, Table 8 shows the results for Sim_2 when "non-positivity" is defined by a combination of two confounders. In this study, the performance of JCART is considerably worse compared to Table 7 where the average bias has a magnitude of 0.54 without truncation and improves as the range of propensity scores is truncated to $[0.1, 0.9]$ with a magnitude of 0.3 in mean bias. The performance of the IPW method with propensities estimated by RF is comparable to the JCART methods and even superior for certain truncations of the propensity score range. Here, when the definition of "non-positivity" includes more than one covariate, the estimator is biased, but the bias decreases with an increasingly restrictive propensity score range. Some of the other methods, such as G-computation with propensity estimated by RF show a small improvement compared to the JCART estimator. However, methods such as G-computation with propensity estimated by GBM yield estimates that are consistently biased, regardless of the truncation range. Although JCART estimates are biased under

this framework, the bias is smaller than a majority of the other methods considered.

When 40 confounders were used in estimating the propensity score, results from Table 9 show that the JCART estimates have small magnitudes of bias in terms of mean bias, but the RMSE is now much larger when compared to the case with only 4 confounders. The magnitude of the median bias is consistently at 0.01 across all truncation ranges for JCART with exception to $[0.1, 0.9]$, but compared to Matching and G-computation, JCART estimates are less biased. Unlike the results from Table 7, the bias from the IPW estimates with RF is consistent at -0.76 while the range of truncation becomes more restrictive from $[0, 1]$ to $[0.05, 0.95]$, and the bias decreases only for the range $[0.1, 0.9]$. The extent of bias becomes worse as the definition of the "non-positivity" subjects includes two confounders, as illustrated by the results from Table 10. The results from Table 10 illustrate that estimation of the causal effect becomes inconsistent as the estimates underestimate the true value, which was also the case for Sim_2 with 4 confounders. As shown in Table 9, the JCART estimates perform slightly better compared to the IPW.rf estimates, and though its performance is mixed compared to the other methods, the average mean bias continues to be smaller than a majority of the estimates considered.

7. Discussion

This paper is concerned with the performance of tree-based methods on various truncation ranges of propensity scores when subjects with systematic propensities of zero or one are falsely included in the study. The positivity assumption is assumed in all observational studies and it states that the propensities of being exposed to all possible exposure conditions for all combinations of observed confounders should be bounded away from zero and one. Subjects whose propensity scores are systematically zero or one, as determined by their baseline confounders, do not have a comparable set of subjects to be compared with in the study. Failure to remove such subjects in an observational study will introduce bias, as demonstrated in our simulation studies.

Different combinations of propensity score estimation and causal inference methods were evaluated in the simulation studies. Though matching methods are natural in that they tend to exclude subjects whose propensities do not overlap across comparison groups, our simulation studies show that even after discarding these subjects, matching methods did not necessarily demonstrate superior performance compared to the JCART method. Our simulation study assessed the ability of several propensity score models to identify subjects whose propensity scores were zero or one. In particular, we used tree-based methods such as GBM, RF, and the JCART method to fit the propensities. A surprising result was that GBM was unable to detect subjects with propensities being zero and one. Though these three methods have been compared extensively, this study is the first to compare them in the context of causal inference and the positivity assumption over different truncation ranges of the propensity score. Further study should be done beyond the scope of checking the positivity assumption with these machine learning tools, but we see the trend of the GBM consistently bounded away from 0 and 1, perhaps too much to the point of introducing biases as shown in the simulations. Further theoretical arguments will have to be developed for GBM's inability of precisely predicting the propensities which are (close to) zero or one.

The simulations used in the study only considered categorical confounders. Since JCART uses specific cutoff values to determine the sub-groups, categorical variables facilitate the partitioning so that assignment of a subject to a group is unambiguous. Though not done here, we suspect that similar partitioning with continuous confounders may lead to less homogeneous subgroups. This suggests that if there is a chance of having subjects whose positivity assumption is not valid, such subjects may be easier to detect if confounding factors are categorized rather than being left continuous.

Table 10: Results for Simulation Sim_4

Method	Bias (mean)	Bias (median)	RMSE
JCART	−0.56	−0.55	0.60
$\text{JCART}_{[0.025,0.975]}$	−0.52	−0.52	0.57
$\text{JCART}_{[0.05,0.95]}$	−0.45	−0.45	0.51
$\text{JCART}_{[0.1,0.9]}$	−0.30	−0.32	0.38
Match.near.logit	−0.81	−0.81	0.84
Match.cem.logit	NA	NA	NA
Match.full.logit	−0.62	−0.61	0.70
Match.near.rpart	−0.91	−0.93	0.95
Match.cem.rpart	NA	NA	NA
Match.full.rpart	−0.58	−0.56	0.77
IPW.glm	−0.68	−0.69	0.70
$\text{IPW.glm}_{[0.025,0.975]}$	−0.61	−0.62	0.63
$\text{IPW.glm}_{[0.05,0.95]}$	−0.56	−0.57	0.59
$\text{IPW.glm}_{[0.1,0.9]}$	−0.51	−0.52	0.55
IPW.gbm	−0.79	−0.79	0.82
$\text{IPW.gbm}_{[0.025,0.975]}$	−0.79	−0.79	0.82
$\text{IPW.gbm}_{[0.05,0.95]}$	−0.79	−0.79	0.82
$\text{IPW.gbm}_{[0.1,0.9]}$	−0.79	−0.79	0.82
IPW.rf	−0.93	−0.93	0.95
$\text{IPW.rf}_{[0.025,0.975]}$	−0.93	−0.93	0.95
$\text{IPW.rf}_{[0.05,0.95]}$	−0.92	−0.92	0.95
$\text{IPW.rf}_{[0.1,0.9]}$	−0.75	−0.77	0.78
GC.glm	−0.49	−0.51	0.52
$\text{GC.glm}_{[0.025,0.975]}$	−0.52	−0.54	0.55
$\text{GC.glm}_{[0.05,0.95]}$	−0.52	−0.53	0.55
$\text{GC.glm}_{[0.1,0.9]}$	−0.51	−0.52	0.55
GC.gbm	−0.77	−0.78	0.80
$\text{GC.gbm}_{[0.025,0.975]}$	−0.70	−0.71	0.73
$\text{GC.gbm}_{[0.05,0.95]}$	−0.65	−0.66	0.68
$\text{GC.gbm}_{[0.1,0.9]}$	−0.59	−0.59	0.63
GC.rf	−0.80	−0.79	0.82
$\text{GC.rf}_{[0.025,0.975]}$	−0.73	−0.73	0.75
$\text{GC.rf}_{[0.05,0.95]}$	−0.68	−0.68	0.71
$\text{GC.rf}_{[0.1,0.9]}$	−0.62	−0.61	0.66

Note: 40 confounders used with nonpositivity subjects defined by $X_{12} = X_1 + X_2$.
RMSE = Root Mean Square Error, JCART = Jackknifed Classification and Regression Trees,
IPW = Inverse Propensity Weighting, GC = G-computation.

However, the performance of JCART and whether it improves upon existing methods comes into question when the propensity score incorporates many confounders and when the definition of "non-positivity" is based on more than one categorical confounder. This was illustrated in simulations Sim_3 and Sim_4 as the number of confounders increased and in simulations Sim_2 and Sim_4 when two confounders were used to define subjects with $e(X) = 0, 1$. A possible explanation lies in the fact that the definition of "non-positivity" becomes more complex with additional confounders to the extent that precise identification of such subjects becomes difficult. Although the definition of "non-positivity" in some applications may be more complex than a simple sum of two confounders, we use the example of two confounders as a simple way to demonstrate the potential drawbacks of JCART. In spite of this, the magnitude of bias becomes significantly smaller as a more restrictive range of propensity scores is used and this is the case for almost all of the methods across all simulations.

Our main motivation for using CART in detecting subjects whose positivity assumption is not met is that CART itself produces zero and one propensities in its tree structure. Because one single tree

may not represent the whole data structure, we used the Jackknife resampling methods to account for the uncertainties of trees and to compute variance estimates of causal effects. The rationale behind the use of the Jackknife over the bootstrap method lies in the higher percentage of information retention of an entire data set for the former compared to the latter. Further research in the use of JCART should incorporate more than two conditions of the exposure in general. The R code for the simulation studies of this paper are available online so that users can familiarize themselves with tree based methods for causal inference.

8. Supplementary material

The R code for JCART simulations is available in the Supplementary Material section on the CSAM homepage (http://csam.or.kr).

References

Austin PC (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies, *Pharmaceutical Statistics*, **10**, 150–161.

Biau G (2012). Analysis of a random forests model, *Journal of Machine Learning Research*, 1063–1095.

Breiman L (2001). Random forests, *Machine learning*, **45**, 5–32.

Breiman L, Friedman JH, Olshen RA, and Stone CJ (1984). *Classification and Regression Trees*, Wadsworth and Brooks.

Brumback BA, Hernan MA, Haneuse SJ, and Robins JM (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures, *Stat Med*, **23**, 749–767.

Cole SR and Frangakis CE (2009). The consistency statement in causal inference: a definition or an assumption?, *Epidemiology*, **20**, 3–5.

Cole SR and Hernán MA (2008). Constructing inverse probability weights for marginal structural models, *American Journal of Epidemiology*, **168**, 656–664.

Crump RK, Hotz VJ, Imbens GW, and Mitnik OA (2009). Dealing with limited overlap in estimation of average treatment effects, *Biometrika*, asn055.

Deconinck E, Hancock T, Coomans D, Massart D, and Vander Heyden Y (2005). Classification of drugs in absorption classes using the classification and regression trees (CART) methodology, *Journal of Pharmaceutical and Biomedical Analysis*, **39**, 91–103.

Freund Y, Schapire R, and Abe N (1999). A short introduction to boosting, *Journal-Japanese Society For Artificial Intelligence*, **14**, 1612.

Hastie T, Tibshirani R, and Friedman J (2001). *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York.

Hong, G. (2010). Marginal mean weighting through stratification: adjustment for selection bias in multilevel data, *Journal of Educational and Behavioral Statistics*, **35**, 499–531.

Horvitz D and Thompson D (1952). A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association*, **47**, 663–685.

Kang J, Su X, Hitsman B, Liu K, and Lloyd-Jones D (2012). Tree-structured analysis of treatment effects with large observational data, *Journal of Applied Statistics*, **39**, 513–529.

Kleiner A, Talwalkar A, Sarkar P, and Jordan M (2012). The big data bootstrap, arXiv preprint arXiv: 1206.6415.

Lin D, Psaty BM, and Kronmal R (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics*, 948–963.

Little R and An H (2004). Robust likelihood-based analysis of multivariate data with missing values, *Statistica Sinica*, **14**, 949–968.

McCaffrey DF, Ridgeway G, and Morral AR (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies, *Psychological Methods*, **9**, 403–425.

Petersen ML, Porter KE, Gruber S, Wang Y, and van der Laan MJ (2012). Diagnosing and responding to violations in the positivity assumption, *Stat Methods Med Res*, **21**, 31–54.

Phillip SK (2001). The delete-a-group Jackknife, *Journal of Official Statistics*, **17**, 521–526.

Robins J (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect, *Math Modelling*, **7**, 1393–1512.

Robins JM, Hernán MA, and Brumbac B (2000). Marginal structural models and causal inference in epidemiology, *Epidemiology*, **11**, 550–560.

Rosenbaum PR (2002). *Observational Studies*, Springer, New York.

Rosenbaum PR (2010). *Observational Studies*, 2nd Ed., Springer, New York.

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

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

Rubin DB (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Rubin DB (1977). Assignment to treatment group on the basis of a covariate, *Journal of Educational and Behavioral Statistics*, **2**, 1–26.

Rubin DB (1980a). Discussion of paper by D. Basu, *Journal of the American Statistical Association*, **75**, 591–593.

Rubin DB (1980b). Comment, *Journal of the American Statistical Association*, **75**, 591–593.

Rubin DB (1986). Statistics and causal inference: comment: which ifs have causal answers, *Journal of the American Statistical Association*, **81**, 961–962.

Rubin DB (2005). Causal inference using potential outcomes: design, modeling, decisions, *Journal of the American Statistical Association*, **100**, 322–331.

Schafer JL and Kang J (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example, *Psychological Methods*, **13**, 279.

Shen C, Li X, Li L, and Were MC (2011). Sensitivity analysis for causal inference using inverse probability weighting, *Biometrical Journal*, **53**, 822–837.

Snowden JM, Rose S, and Mortimer KM (2011). Implementation of G-computation on a simulated data set: demonstration of a causal inference technique, *American Journal of Epidemiology*, **173**, 731–738.

Su X, Kang J, Fan JJ, Levine RA, and Yan X (2012). Facilitating score and causal inference trees for observational studies, *Journal of Machine Learning Research*, **13**, 2955–2994.

Taubman SL, Robins JM, Mittleman MA, and Hernán MA (2009). Intervening on risk factors for coronary heart disease: an application of the parametric g-formula, *International Journal of Epidemiology*, **38**, 1599–1611.

van der Laan M (2013). *Targeted maximum likelihood estimation*, US Patent, 8,438,126.

Westreich D and Cole SR (2010). Invited commentary: positivity in practice, *American Journal of Epidemiology*, **171**, 674–677.

Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, Gange SJ, and Hernán MA (2012).

The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death, *Statistics in Medicine*, **31**, 2000–2009.
Zhang H and Singer B (1999). *Recursive Partitioning in the Health Sciences*, Springer, New York.

Received April 20, 2015; Revised August 28, 2015; Accepted November 26, 2015