A NEW MULTILEVEL CART ALGORITHM AND ITS APPLICATION IN PROPENSITY SCORE ANALYSIS

A Dissertation

by

SHUQIONG LIN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Wen Luo
Committee Members,	Oi-Man Kwok
	Myeongsun Yoon
	Lei-Shih Chen
Head of Department,	Shanna Hagan-Burke

May 2018

Major Subject: Educational Psychology

ABSTRACT

The logistic regression model is the most commonly used analysis method for modeling binary data. Unbiased estimation using logistic regressions heavily depends on strong model assumptions which are often violated in reality. The classification and regression tree (CART) algorithm gains its popularity to replace the logistic regression, because CART does not require model assumptions and can model complex relationships automatically. However, only limited studies developed multilevel CART (M-CART) algorithms for modeling multilevel data with binary outcomes. Therefore, in the first study, a new M-CART algorithm was proposed for modeling multilevel data with binary outcomes which combines the multilevel logistic regression (M-logit) and the single-level CART (S-CART) using an expectation-maximization algorithm. This proposed algorithm allows inclusion of covariates at all levels, depends on no model assumptions, and captures interaction and nonlinearity in an automatic way. The performance of the proposed M-CART was compared with M-CART, S-CART, and single-level logistic regression (S-logit) in terms of prediction accuracy. Results from simulation study showed that M-CART lead to higher classification accuracy, sensitivity, specificity and Klecka's tau values than all other three methods.

In the second study, the proposed M-CART algorithm was applied in propensity score analysis (PSA) when having multi-site non-randomized control trials (non-RCTs). PSA is the most popular statistical technique that estimates the casual effect of a treatment by eliminating the systematic differences of pre-treatment covariates between individuals who receive treatment and individuals who do not receive treatment. M-logit and S-CART have been applied to estimate propensity scores, while no study has explored the performance of using M-CART for

estimation. Thus, in the second study, the performance of the proposed M-CART was compared with M-logit, S-CART, and S-logit in terms of covariate balance and treatment effect estimation. Results indicated that M-CART was more stable than the M-logit, S-CART and S-logit on achieving pre-treatment covariate balances and always yielded reasonable covariate balances over all conditions. Results further showed that, regardless of the PS conditioning approaches, M-CART yielded the least relative biases in the treatment effect estimations across all simulated conditions than other methods.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Wen Luo, and my committee members, Dr. Oi-man Kwok, Dr. Myeongsun Yoon, and Dr. Lei-Shih Chen, for their guidance and support throughout the course of this dissertation. I especially want to express my appreciation from the bottom of my heart to Dr. Luo for all her contributions of time, ideas and help on my Ph.D. study.

Also, I would like to express my gratitude and love to my mother and father for their unconditional support and all the sacrifices they made for me, especially during tough times in the Ph.D. pursuit.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Wen Luo, Oi-man Kwok, and Myeongsun Yoon of the Department of Educational Psychology, and Professor Lei-Shih Chen of the Department of Health and Kinesiology.

All work for the dissertation was completed independently by the student.

Funding Sources

Graduate study was supported by the College of Education and Human Development Strategic Research Award from Texas A&M University.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

During the past few decades, the use of multilevel models (Goldstein, 2011; Hox, Moerbeek, & van de Schoot, 2010) becomes an important trend for research in the behavioral science. For many situations, data naturally contain multilevel structures, with examples like students nested within classrooms and schools, employees within different organizations, and repeated test scores nested within individuals. Because observations in the same clusters share the same cluster-level characteristics, the assumption of independent observation required for single-level models is violated. Therefore, multilevel models are recommended and widely used nowadays to address the dependency issue in multilevel data (Aitkin & Longford, 1986). Using multilevel models, researchers can examine how covariates measured at each level affect outcome variables (Guo & Zhao, 2000). In addition, multilevel models control for clustering effects (Goldstein, 2011; Snijders, 2011) and yield more accurate parameter and standard error estimations than single-level models (Goldstein, 2011; Hox, 1998; Snijders, 2011). As a result, multilevel models provide more accurate confidence intervals and significance tests in general (Guo & Zhao, 2000).

In many situations, multilevel data have binary outcomes, such as binary test results (i.e., passing or failing a test) of children who are nested within teachers, college enrollments for high school students who are nested within different high schools, and retention status of undergraduates nested within various majors. The multilevel logistic regression model (e.g., Rumberger, 1995; Sideridis, Antoniou & Padeliadu, 2008) is the most commonly used analysis method for this type of data. Unbiased parameter estimation using multilevel logistic regression

models heavily depends on very strong model assumptions, such as linearity between log odds of outcome variables and predictors, and non-collinearity among predictors (Hox, Moerbeek, & van de Schoot, 2010; McMahon, Pouget, & Tortu, 2006; Rodríguez, 2008). However, these assumptions are often violated in practice, especially in social and behavioral sciences.

To overcome drawbacks of traditional statistical models such as logistic regression models, the classification and regression tree algorithms (CART, Breiman, Friedman, Olshen, & Stone, 1984), one of the most well-known data mining techniques, were recently introduced and applied in social and behavioral sciences (Lemon, Roy, Clark, Friedmann, & Rakowski, 2003). To apply to technique in multilevel data, several researchers extended single-level CART to multilevel CART (M-CART) which do not depend on model assumptions and control for clustering effects simultaneously (e.g., Lee, 2005; Zhang, 1998; Zhang & Ye, 2008). However, the majority of the existing M-CARTs do not allow the use of first-level covariates as predictors for modeling and are only suitable for multilevel data with repeated measures nested individuals (Hajjem, Bellavance, & Larocque, 2011; Sela & Simonoff, 2012). A few advanced M-CARTs are developed to overcome this limitation, but are only able to model multilevel continuous data (e.g., Hajjem, Bellavance, & Larocque, 2011; Sela & Simonoff, 2012). Therefore, the first goal of this dissertation is to develop a new M-CART which allows the inclusion of level-one covariates for modeling multilevel data with binary outcomes.

The advantages of the new M-CART developed in this dissertation, such as the ability to handle a large number of covariates and minimal assumption requirement, make it a good choice for estimating propensity scores in multi-site non-randomized controlled trials (non-RCTs). In multi-site non-RCTs, individuals within each cluster are assigned non-randomly to either treatment or control groups (Dziak, Nahum-Shani, & Collins, 2012) which yields systematic

differences on pretreatment covariates between the two groups. With this systematic differences before treatment, it is invalid to conclude that the observed difference between two groups on the outcome variable after receiving intervention is only due to treatment effects. Thus, to draw causal inferences of treatment effects in such trials, the main issue is how to adjust for the pretreatment imbalance on covariates between the treated and controlled groups.

Propensity score analysis (PSA) is the most popular statistical technique that estimates the casual effect of a treatment, policy, or other intervention by accounting for the imbalance of pre-treatment covariates between individuals who receive treatment and individuals who do not receive treatment (Austin, 2011). When employing PSA in multi-site non-RCTs (M-PSA), researchers often adopt either multilevel logistic regressions or single-level CARTs for estimating propensity scores (e.g., Thoemmes & West, 2011; Westreich, Lessler, & Funk, 2010), and use estimated propensity scores to balance pre-treatment covariates. However, multilevel logistic regressions may suffer from assumption violations and single-level CARTs ignore the clustering effects, both of which may result in biased propensity score estimations and lead to biased treatment effect estimates.

The proposed M-CART algorithm overcomes the limitations of multilevel logistic regression models and single-level CART, therefore it is a good candidate to be considered in propensity score estimation in multi-site non-RCTs. Therefore, the second goal of the dissertation is to evaluate the performance of the proposed M-CART in propensity score estimation in multi-site non-RCT.

To achieve the two goals, I conduct two studies in the dissertation. In the first study, I propose a multilevel CART algorithm for modeling multilevel data with binary outcomes which combines the multilevel logistic regression model and the single-level CART using expectation-

maximization algorithm. This proposed algorithm controls for clustering effects and allows the inclusion of covariates at all levels. The performance of the proposed M-CART is compared with multilevel logistic regression, single-level CART, and single-level logistic regression in terms of prediction accuracy. In the second study, I apply the proposed M-CART to propensity score analysis in multi-site non-RCTs, and compare the performance of M-CART with multilevel logistic regressions, single-level CART, and single-level logistic regressions in terms of covariate balance and treatment effect estimation.

CHAPTER II

A NEW MULTILEVEL CART ALGORITHM FOR MULTILEVEL DATA WITH BINARY OUTCOMES

Introduction

As the thrust of data mining, a few well-known data mining techniques have been introduced and applied to research in social sciences as alternatives to traditional statistic models recently (e.g., Baker, 2010; Finch, 2014). For example, the classification and regression tree (CART, Breiman, Friedman, Olshen, & Stone, 1984), a recursive partitioning method, is often used as an alternative to logistic regressions when predicting binary outcomes (Lee, Lessler, & Stuart, 2010).

Compared to logistic regressions, CART has several desirable properties when handling binary data. First, CART does not require strong assumptions such as linearity between log odds of dependent variables and predictors, and non-collinearity among predictors. Second, it is a data-driven approach which automatically includes the significant variables and remove non-significant ones (Timofeev, 2004), and is able to automatically capture non-linear and interaction terms (Lee, Lessler, & Stuart, 2010; Steinberg & Colla, 2009). Third, CART is invariant to monotonic transformations of variables, such as logarithm or square root of variables (McLachlan, 2004; Steinberg & Colla, 2009). Fourth, CART is able to handle missing data using surrogate splits without extra imputation procedure (Deconinck, Hancock, Coomans, Massart, & Vander; 2005; Feelders, 1999; Verbyla, 1987).

However, the majority of existing CART algorithms were developed only for single-level binary data. In education and other social sciences, multilevel data with binary outcomes are

even more common. For example, when predicting college students' retention status, the data is likely to have a multilevel structure with students nested in various majors. Single-level CARTs (S-CARTs) do not take clustering effects in multilevel data into consideration. A limited number of studies attempt to extend S-CARTs to multilevel CARTs (M-CARTs) for modeling multilevel data with binary outcomes (e.g., Lee, 2005; Zhang, 1998). These existing M-CART algorithms mainly work for longitudinal data with repeated binary measures nested within individuals and do not allow time-varying (i.e. level-1) covariates (e.g., Lee, 2005; Zhang, 1998; Zhang & Ye, 2008). This limits the use of the existing M-CARTs in multilevel data with individuals nested within groups, because, for person-in-group data, it is often that individual-level (level -1) covariates are significant predictors for modeling outcome variables.

To fill in the gap and allow the use of individual-level covariates for prediction, I propose a new M-CART algorithm which combines the features of S-CARTs and multilevel logistic models (M-logits) using the expectation-maximization (EM) algorithm. Using Monte Carlo simulations, I evaluate the performance of the proposed M-CART algorithm for modeling multilevel data with binary outcomes. Specifically I am interested in the following research questions:

1. Does the newly proposed M-CART algorithm yield better prediction than the M-logit, S-CART, and S-logit?

2. How do different intra-class correlations (*ICC*s), sample sizes, and degrees of non-linearity and interaction between outcomes and predictors impact the predictive performance of the M-CART, M-logit, S-CART, and S-logit?

In the following sections of the paper, I first briefly review M-logit, S-CART, and M-CART algorithms. I then introduce the algorithm of the proposed M-CART and present the simulation study. Finally, I discuss the findings, implications, and limitations.

Theoretical Framework

Multilevel Logistic Regression

In a given M-logit, a binary outcome Y with *N* observations nested within *H* clusters follows a Bernoulli distribution, conditional on random effects u. The probability of *Y*, indexed by π, satisfies the function below:

$$E(Y\,|u) = g(\pi) = X\beta + Zu, \tag{2.1}$$

where β is the vector of fixed effects. X is the design matrix for β, u is the vector of random effects which is assumed to have a multivariate normal distribution with mean (*M*) 0 and variance-covariance matrix Ω, *Z* is the design matrix for *u*, and *g*(.) is the logistic link function. In this M-logit, parameters are estimated by maximizing the marginal likelihood function as shown below

$$L(\beta, \Omega\,|Y) = \int P(Y|\beta, u\,)\Phi(u|\,\Omega)\,du, \tag{2.2}$$

where $P(Y|\beta, u)$ is the conditional probability distribution of *Y*, $\Phi(u|\,\Omega)$ represents the normal density function of u.

M-logits require strong model assumptions (Hox, Moerbeek, & van de Schoot, 2010; McMahon, Pouget, & Tortu, 2006; Rodríguez, 2008): (a) Observations between clusters are independent, whereas observations within clusters share auto-correlations. (b) Error terms at all levels are uncorrelated with predictors. (c) No multicollinearity exists among predictors. (d) Linearity should be found between logit of the outcome variable and predictors.

M-logits mathematically constrain probabilities in the [0, 1] range and are easily to be interpreted and understood (Kleinbaum & Klein, 2002). However, criticisms of using M-logits increase recently. First, as mentioned above, unbiased parameter estimations of fitting M-logits are based on strong statistic assumptions (Drake, 1993; McMahon, Pouget, & Tortu, 2006). In addition, M-logits are not applicable when the sample size is small, especially when the number of predictors are larger than the sample size (Jorgensen, 1983; Moineddin, Matheson, & Glazier, 2007). Moreover, researchers need to pre-define M-logit equations by including all considerable covariates (Bursac, Gauss, Williams, & Hosmer, 2008; D'Agostino Jr, 1998), which is problematic. Researchers who do not know the true models (the true models are never known in reality) might overlook significant predictors and potential non-linear and interaction terms when they pre-specify model equations. Thus, researchers should repeatedly revise models, which is a complex and error-prone process.

CART

A CART algorithm makes prediction by recursively partitioning data into groups based on a set of covariates. In general, it includes three steps. First, a given CART algorithm searches for every allowable splitting covariate and its cut-off value, and then selects the optimal one to categorize data into two most homogenous groups (i.e., two child nodes). This optimal splitting covariate and its cut-off value are chosen to guarantee that they maximize the similarity within nodes. A CART algorithm makes use of impurity-based measures to compute the magnitude of within-node similarity. Among various impurity-based measures (e.g., Breiman et al., 1984; Kearns & Mansour, 1999; Quinlan, 1987), Gini impurity (Breiman et al., 1984) is the most common choice for splitting when the outcome variable is binary (Hastie, Tibshirani &

Friedman, 2009; Timofeev, 2004). Gini impurity is based on the Gini gain function (Steinberg & Colla, 1995; Breiman et al., 1984) as below.

$$\Delta i(t) = i(t_p) - p_l[i(t_l)] - p_r[i(t_r)], \quad (2.3)$$

where $i(t) = 1 - \sum_{m=0}^{1} p\,(m|t)^2$. $p(m|\,t)$ is the conditional probability of outcome class m (m=0 or 1 for binary data) in node t, t_l and t_r are the child nodes of the parent node t_p, p_l and p_r are probabilities of being assigned into the t_l and t_r over the t_p. The covariate and cut-off value that maximize the Gini gain function are considered as the optimal ones for node partitioning.

Second, the first step above is iteratively reapplied to each of the newly developed child nodes, creating a classification tree, until a given stopping criterion is triggered. A tree growing phase can stop when observations within every node all share the same outcomes values; the maximum tree depth are reached, the number of observations in a node is less than the minimum number preset, or the splitting result is not better than a certain error threshold pre-specified (Hayes, Usami, Jacobucci, & McArdle, 2015; Rokach & Maimon, 2005). In a classification tree, the first node is called the root node and nodes with no further child nodes are called terminal nodes or leaves.

Third, to avoid overfitting, a fully grown tree is often pruned from leaves to the root in order to get the most parsimonious tree. The most commonly used pruning method for CART is the cost-complexity pruning (Breiman et al., 1984) which estimates the cost of trimming a set of subtrees using equation 2.4 and replaces subtrees with simple terminal nodes if trimming reduces cost.

$$R_\alpha(W)=R(W) + \alpha|W|, \quad (2.4)$$

where $R(W)=\sum_{t=1}^{|W|} r(t)p(t)$, $r(t)=\frac{E_t}{N_t}$, $p(t)=\frac{N_t}{N}$. In this function, E_t is the number of misclassified observations in node *t*, N_t is the number of observations in node *t*, *N* is the total number of observations in tree *W*, |*W*| indicates the number of terminal nodes in tree *W*, and the complexity parameter α represents the cost of adding an extra node. For each observation in the final tree, the predicted probability of being assigned into the target class equals to the observed proportion of target classes within the terminal node it belongs to.

Multilevel CART (M-CART)

As an extension of S-CART, M-CART has a short history and mainly focuses on longitudinal data. The key idea of currently available M-CART algorithms is to modify impurity-based measures by adding corrections for correlated observations within clusters. Segal (1992) and Larsen and Speckman (2004) build a M-CART for longitudinal continuous data by using Mahalanobis distance within each node as the impurity-based measure to compute within-node similarity for node partition. De'Ath (2002) and Abdolell, LeBlanc, Stephens, and Harrison (2002) adopt multivariate sum-of-square deviance within each node as the impurity measure in their M-CARTs. Zhang and his colleagues utilize log-likelihood of multivariate binary distributions as impurity measure for longitudinal binary data modeling (Zhang, 1998; Zhang & Ye, 2008). Lee (2005) apply Pearson residuals estimated from fitting marginal regression model into the impurity-based measure to find optimal split when having longitudinal binary outcomes.

An important characteristic of the aforementioned methods is that they adjust auto-correlations within nodes, which requests the repeated measures (correlated observations) of the same individuals being classified into the same nodes and therefore no time-varying covariates (level-1 covariates) being allowed (Segal, 1992; Hajjem, Bellavance & Larocque, 2011). That is,

repeated measures under the same individuals must be classified into the same nodes and assigned the same predicted outcomes. This characteristic severely limits the use of these M-CART algorithms in modeling person-in-group multilevel data, because it is often necessary to predict the outcome using level-1 (i.e., individual-level) covariates and individuals in the same clusters are likely have different values on the outcome variable. It is unrealistic to classify all persons within the same clusters into the same nodes and assigned the same predicted outcomes.

To overcome this problem and enable covariates at individual-level to be candidates in splitting process, a few researchers proposed M-CARTs that combine S-CARTs and multilevel linear models under the EM framework. This idea was initially launched by Hajjem, Bellavance, and Larocque (2011), and Sela and Simonoff (2012), and then modified by other researchers who replace CARTs with other decision tree algorithms which is beyond the scope of this study (detailed information can be found in Eo & Cho, 2014; Hajjem, Bellavance & Larocque, 2014; Loh & Zheng, 2013). These M-CARTs are more appropriate for person-in-group multilevel data, because person-level predictors can be used for splitting, and persons within the same clusters are allowed to be categorized into different nodes. Furthermore, studies have shown that these methods outperform multilevel linear regressions (Hajjem, Bellavance, & Larocque, 2011; Sela & Simonoff, 2012). However, these M-CARTs can only be applied to multilevel data with continuous outcomes. A M-CART algorithm that can overcome the limitation aforementioned for person-in-group multilevel data with binary outcomes is yet to be developed.

The Proposed M-CART Algorithm

Built on the work of Sela and Simonoff (2012), the proposed M-CART algorithm decompose a multilevel binary outcome into the fixed and the random components which are estimated using the S-CART and M-logit respectively. The estimated fixed and random

components are then combined and updated iteratively under the EM framework until convergence. The details of the proposed M-CART algorithm are described below.

In a two level data set with *N (j*=1,2,3,…, *N)* observations nested within *H* clusters (*j*=1,1,3,…, *H*), there is a binary outcome Y that follows a Bernoulli distribution, conditional on random effects u. The probability of *Y*, π, is modeled as $E(Y\,|u) = g^{-1}(\pi) = X\beta + Zu$. To predict *Y*, the proposed M-CART involves the following steps:

1) Random effect component u is initialized with a vector of values calculated as deviances between the grand mean ($\bar{Y}$) and cluster means ($\bar{Y}_j$).

2) Fixed effects $X\beta$ is extracted by subtracting the random effect component *u* from the outcomes, $Y - Zu$. Then the isolated $X\beta$ part is modeled using S-CART with all covariates *X*, which creates a categorical indicator, *Ind*, to represent terminal nodes in the S-CART based tree, *W*. In the S-CART, Gini impurity and cost-complexity pruning are adopted.

3) The *Ind* estimated in step 2 is used as a new and only covariate in the M-logit (equation 2.5) to model the probability of *y*=1. Log-likelihood values are calculated using Laplacian approximation (Raudenbush, Yang, & Yosef, 2000). The random effect u estimated in this step is then used in step 2 to update the fixed effect $X\beta$.

$$E(Y\,|u) = g^{-1}(\pi) = Ind\;\lambda + Zu, \tag{2.5}$$

4) Step 2 and 3 are iteratively executed until the change of log-likelihood values between two iterations is smaller than the pre-set tolerance value or the maximum iteration number is achieved. The tree and model parameters estimated in the last iteration, W', λ', and u', are finalized as the final parameter estimations.

5) When predicting the outcome values of new cases, it needs to distinguish between two types of predictions. One type of prediction concerns a new unit in an existing cluster, and the other type concerns a new unit in a new cluster. When predicting a new unit in an existing cluster, the new case is classified into terminal nodes according to the output tree developed in the last iteration, W', with a new terminal node indicator, Ind'_{new}, for those new cases. Then, the probability of being classified into the target class Y =1 for the new case is calculated using equation 2.5 with $Ind = Ind'_{new},\ \lambda = \lambda', and\ u = u'$. When predicting a new unit in a new cluster, the marginal expectation can be used assuming the new cluster is sampled randomly. In this case, the random effect in equation 2.5 is replaced with 0 because the population mean of the random effects is zero.

Simulation Study

To compare the performance of the proposed M-CART algorithm with the M-logit, S-CART, and S-logit, a simulation study was conducted. I generated data based on 32 data conditions (4 sample sizes * 2 *ICC*s * 4 degrees of non-linearity and interaction) with 300 replications in each condition. For each generated dataset, the 4 methods were employed, resulting in a total of 128 conditions (32 data conditions* 4 estimation methods). All data were generated and analyzed using R 3.2.4 (R Core Team, 2016).

Data Generation

Data sets were generated based on a two level random intercept model with a binary outcome variable Y_{ij} (Y_{ij}= 0 or 1), three level-1 predictors X_1, X_2, X_3, three level-2 predictors X_4, X_5, X_6, and their squared terms and interaction terms ${X_1}^2$, X_1X_2 , ${X_4}^2$, and X_4X_5. The relationship between Y_{ij} and covariates was defined as shown in equation 2.6.

$$logit\left(Y_{ij}\right) = \beta_{0j\_T} + \beta_{1j\_T}X_1 + \beta_{2j\_T}X_2 + \beta_{3j\_T}X_3 + \beta_{4j\_T}{X_1}^2 + \beta_{5j\_T}X_1X_2 \quad (2.6)$$

$$\beta_{0j\_T} = \gamma_{00\_T} + \gamma_{01\_T}X_4 + \gamma_{02\_T}X_5 + \gamma_{03\_T}X_6 + \gamma_{04\_T}{X_4}^2 + \gamma_{05\_T}X_4X_5 + u_{0j\_T}$$

$$\beta_{1j\_T} = \gamma_{10\_T}$$

$$\beta_{2j\_T} = \gamma_{20\_T}$$

$$\vdots$$

$$\beta_{5j\_T} = \gamma_{50\_T}.$$

Predictors X_1 and X_4 were continuous variables generated from a standard normal distribution, *N* (0, 1). Predictors X_2 and X_5 were also continuous variables and generated from a similar normal distribution with a slightly larger variance, *N* (0, 2). Predictors X_3 and X_6 were binary variables generated from a binominal distribution with the expected probability of .5. Slightly positive correlations between X_3 and X_6 ($r_{23} = .2$) and between X_2 and X_4 ($r_{46} = .2$) were allowed as Lee, Lessler, and Stuart (2011) set in their study. Level-2 random effects u_{0j_T} was generated from a normal distribution with mean of 0 and variance $\sigma^2_{u0j\_T}$ of either 0.37 or 1.41 depending on the different *ICC*s conditions which are explained in the following section.

For model parameters (see Table 1), the intercept γ_{00_T} was fixed to 0 for simplicity. Regression coefficients γ_{01_T}, γ_{02_T}, γ_{03_T}, γ_{10_T}, γ_{20_T}, and γ_{30_T} were set as 0.3 to reflect moderate effects on outcome *Y* (Martin, 2015; Yu, 2012). Regression coefficients for all nonlinear and interaction terms, γ_{04_T}, γ_{05_T}, γ_{40_T}, and γ_{50_T}, were set to be the same with values ranging from 0 to 0.9 to indicate different degrees of interaction and non-linearity.

Table 1
Model Parameters Used across Different Degrees of Interaction and Non-linearity Scenarios

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4
γ_{00_T}	0	0	0	0
γ_{01_T}	0.3	0.3	0.3	0.3
γ_{02_T}	0.3	0.3	0.3	0.3
γ_{03_T}	0.3	0.3	0.3	0.3
γ_{04_T}	0	0.3	0.6	0.9
γ_{05_T}	0	0.3	0.6	0.9
γ_{10_T}	0.3	0.3	0.3	0.3
γ_{20_T}	0.3	0.3	0.3	0.3
γ_{30_T}	0.3	0.3	0.3	0.3
γ_{40_T}	0	0.3	0.6	0.9
γ_{50_T}	0	0.3	0.6	0.9

Design Factors

Sample size. Sample size was manipulated via the number of clusters (*Nc*) and cluster size (*Ns*). *Nc* was set to be 30 or 50 which are commonly used in previous multilevel data modeling studies (Finch & French, 2011; Kwok, Luo & West, 2010; Maas & Hox, 2005; Jak, Oort & Dolan, 2013). The cluster size was set to be 65 and 125, in which 80% of data was used for training purpose (*Ns-t was 52* and *100*) and 20% for validation purpose (*Ns-v was 13* and *25*). These cluster sizes were chosen based on previous simulation studies (e.g., Finch & French, 2011; Kwok, Luo & West, 2010; Maas & Hox, 2005) and the requirement of minimum cluster size of 50 for M-logits (Moineddin, Matheson, & Glazier, 2007). Combining the 4 sample size conditions, the total sample size *N* ranges from 1950 (65*30=1950) to 6250 (125*50=6250), covering a wide range of sample sizes.

Conditional ICC. Conditional *ICC* was set to be .10 and .30, representing small and large clustering effects in educational settings (Hedges & Hedberg, 2007; Hox, Moerbeek, & van de

Schoot, 2010; Luo, Cappaert, & Ning, 2015). Based on the selected *ICC* values, the level-2 variance $\sigma^2_{u0j\_T}$ was computed using the equation $ICC = \frac{\sigma^2_{u0j\_T}}{\sigma^2_{u0j\_T}+\sigma^2_{e\_T}}$ (Maas & Hox, 2005) where $\sigma^2_{e\_T} = \frac{\pi^2}{3}$ for M-logits (Snijders & Bosker, 1999). Therefore, $\sigma^2_{u0j\_T}$ was set to be 0.37 and 1.41 for *ICC* of 0.1 and 0.3 respectively.

Degree of non-linearity and interaction. Regression coefficients for the nonlinear and non-additive terms, γ_{04_T}, γ_{05_T}, $\gamma_{40\_}T$, and γ_{50_T}, were set to have four values 0, 0.3, 0.6 to 0.9 to represent various degrees of nonlinearity and interaction effects (Table 1). When γ_{04_T} = γ_{05_T} = γ_{40_T} = γ_{50_T} = 0, the true model was a linear regression model with only main effects. As the effects increased, the true model was increasingly dominated by the squared and interaction terms.

Analysis

Every generated data set was partitioned into training and validation sets using random resampling without replacement within clusters. The training data sets were used for developing the M-CART, M-logit, S-CART, and S-logit models. The validation data sets were not exposed to model development, but only used to evaluate the output trees and regression models created in the training step.

For all M-CARTs, S-CARTs, M-logits and S-logits, only the first-order terms (i.e., main effects) of the 6 predictors were used as independent variables. It means when true data generation models contained the nonlinear and interaction effects, the estimation models used for analyses were over-simplified models. This was to mimic real data modeling situations in which researchers do not have prior knowledge of nonlinear and interaction effects and tend to include

main effects only. It also allowed us to examine whether and how much the proposed M-CART is able to capture unspecified non-linear and interaction effects in an automatic way.

For M-CARTs, the tolerance of change on log-likelihood values was set as 0.0001; the maximum number of iterations was 1000; the minimum number of observations that must exist in a node was 2; the complexity parameter α used in pruning procedure was 0.01. These settings were also applied to S-CARTs, except for the tolerance of log-likelihood change which is not applicable for S-CARTs. For M-logit modeling, a random intercept model was used and estimated with Maximum likelihood with Laplacian approximation.

Evaluation Criteria

After obtaining the estimated trees and logistic regressions, the validation data sets were used for evaluating the performance of the estimated models. Four measures, classification accuracy, sensitivity, specificity, and Klecka's tau (τ, Klecka, 1980), were considered. Classification accuracy refers to the proportion of correctly classified observations. Sensitivity refers to the ability to correctly classify observations with positive outcomes (i.e., y=1). It equals the proportion of correctly classified y=1 observations out of all y=1 observations. Specificity measures the ability to correctly classify observations having negative outcomes (i.e., y=0). It equals the proportion of correctly classified y=0 observations out of all y=0 observations. Finally, Klecka's tau, ranging from 0 to 1, measures the degree of improvement on classification accuracy over a random allocation. Below is the equation for Klecka's tau:

$$\tau = \frac{n_{cor} - \sum_{m=0}^{1} p_m n_m}{N_v - \sum_{m=0}^{1} p_m n_m} \quad (2.7)$$

where n_{cor} is the number of observations correctly classified, n_m is the number of observations belonging to class m (m=0 or 1), N_v is the number of total observations in the validation data set, and p_m is the prior probability of class membership by chance which is 0.5 in our study.

A factorial ANOVA with all main effects and interactions, and the effect size eta-squared (η^2) were also computed to investigate the impacts of these designed factors on the performance of different estimation models in terms of classification accuracy, sensitivity, specificity, and tau. Eta-squared values 0.01, 0.06, and 0.14 were used to indicate small, moderate, and large effect sizes respectively (Cohen, 1998).

Results

Classification Accuracy

The distribution of classification accuracy was shown in Figure 1. In total, 66.45% of the cases (M = 0.6645) were correctly classified averaging over all data conditions with standard deviation (SD) of 0.0619. ANOVA results showed that only the four main effects and the interaction between the degree of nonlinearity and interaction factor and the estimation method factor had meaningful effect sizes. Specifically, the estimation model factor had the largest effect on classification accuracy [$F(3, 38268) = 2199.6523$, $p < 0.001$] with effect size η^2 of 0.1213. On average, as shown in Table 2, the M-CART algorithm produced higher classification accuracy than the M-logit model ($accuracy_{M\text{-}CART} = 0.6869 > accuracy_{M\text{-}logit} = 0.6791$), while the S-logit had the lowest classification accuracy ($accuracy_{S\text{-}logit} = 0.6308$).

ICC had a small to medium effect on classification accuracy [$F(1, 38268) = 3033.1327$, $p < 0.001$, $\eta^2 = 0.0527$]. Shown in Table 2, as the increase of *ICC*, the classification accuracy increased for all estimation methods. For instance, when *ICC* increased from 0.1 to 0.3, the classification accuracy based on using the M-CART went up from 0.6728 to 0.7010.

The sample size factor yielded a small impact on classification accuracy [$F(3, 38268) = 365.0096$, $p < 0.001$] with η^2 equaled to 0.0190. Generally, the classification accuracy increased as the number of clusters and/or the cluster size raised for all estimation models (Table 2).

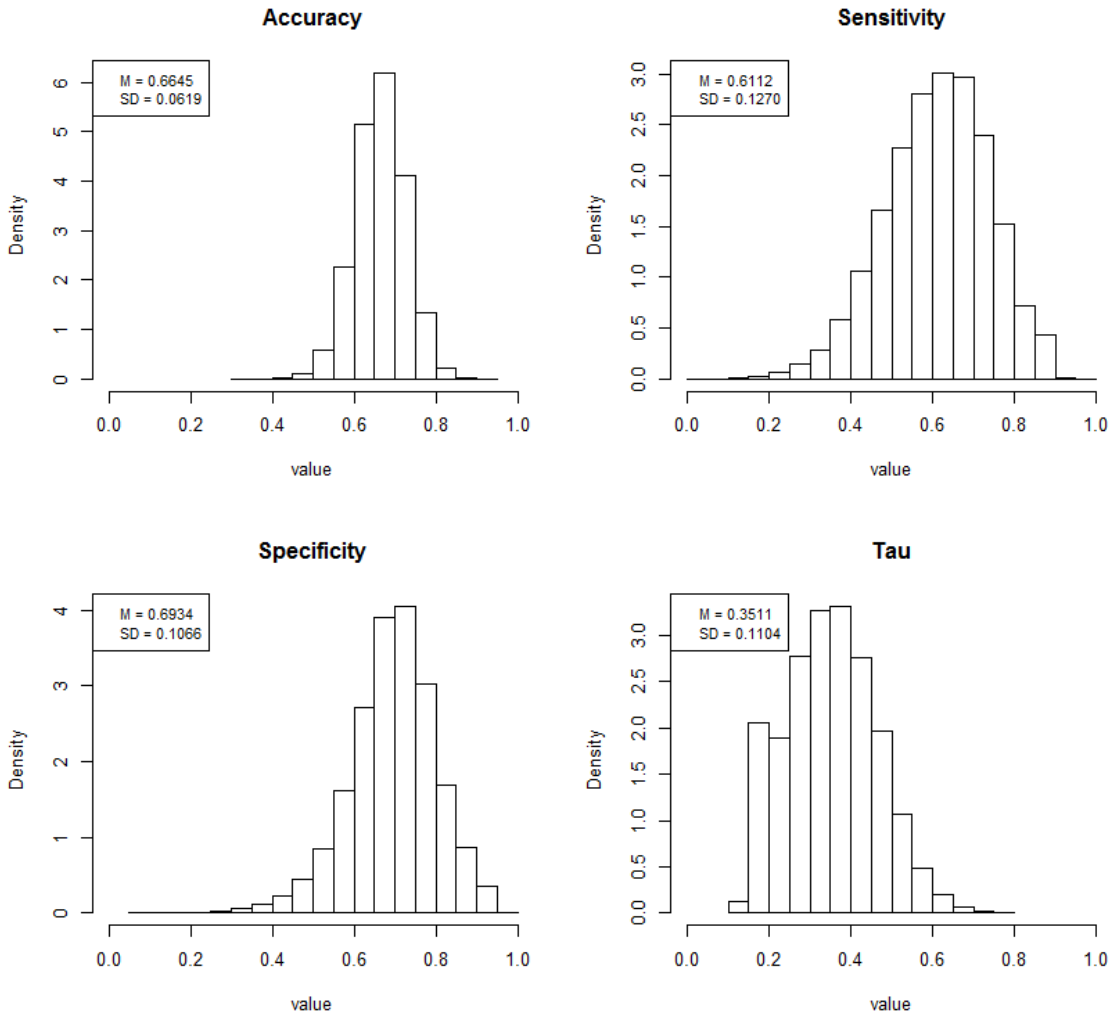

Figure 1. The overall performance based on four criteria values.

Even though the degree of nonlinearity and interaction factor also only had a small main effect [$F(3, 38268) = 1656.7281$, $p < 0.001$ with η^2 equaled to 0.0288], but interaction between the degree of nonlinearity and interaction factor and the estimation model factor had a moderate to large effect size [$F(3, 38268) = 2048.7180$, $p < 0.001$, $\eta^2 = 0.1129$]. When non-linearity and interaction effects were both zero, the M-logit and S-logit produced higher accuracy values on average than the corresponding CART algorithms ($accuracy_{M\text{-}logit} = 0.6922 > accuracy_{Slogit} = 0.6614 > accuracy_{M\text{-}CART} = 0.6473 > accuracy_{S\text{-}CART} = 0.6284$). When the true models had small

Table 2
Classification Accuracy Mean Values by Different Design Factors

Conditions	M-cart		M-logit		S-cart		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ICC								
0.1	0.6728	0.0616	0.6659	0.0520	0.6501	0.0582	0.6269	0.0545
0.3	0.7010	0.0582	0.6924	0.0551	0.6682	0.0587	0.6348	0.0570
Nonlinearity & interaction								
0	0.6473	0.0606	0.6922	0.0524	0.6284	0.0529	0.6614	0.0501
0.3	0.6772	0.0544	0.6795	0.0560	0.6413	0.0531	0.6396	0.0500
0.6	0.7012	0.0561	0.6754	0.0595	0.6755	0.0541	0.6132	0.0530
0.9	0.7220	0.0497	0.6694	0.0503	0.6915	0.0524	0.6091	0.0537
Sample size								
30-65	0.6759	0.0719	0.6685	0.0697	0.6484	0.0691	0.6259	0.0680
30-125	0.6897	0.0631	0.6826	0.0517	0.6626	0.0582	0.6329	0.0564
50-65	0.6854	0.0581	0.6760	0.0505	0.6558	0.0549	0.6283	0.0536
50-125	0.6967	0.0510	0.6896	0.0438	0.6699	0.0513	0.6362	0.0418
Average	0.6869	0.0618	0.6791	0.0553	0.6592	0.0592	0.6308	0.0558

nonlinearity and interaction effects (i.e., 0.3), the M-CART and M-logit had similar classification accuracy ($accuracy_{M\text{-}CART}$ = 0.6772, $accuracy_{M\text{-}logit}$ = 0.6795) and both were more accurate than the S-CART and S-logit ($accuracy_{S\text{-}CART}$ = 0.6413, $accuracy_{S\text{-}logit}$ = 0.6396). As the nonlinearity and interaction effects increased to the highest level (i.e., 0.9), the M-CART showed higher classification accuracy than all the other three estimation methods ($accuracy_{M\text{-}CART}$ = 0.7220 > $accuracy_{M\text{-}logit}$ = 0.6694, $accuracy_{S\text{-}CART}$ = 0.6915 and $accuracy_{S\text{-}logit}$ = 0.6091).

Sensitivity

The distribution of sensitivity was shown in Figure 1. On average, 61.12% of cases whose observed outcome were *y*=1 were correctly predicted as *y*=1 across all data (*M* = 0.6112, *SD* = 0.1270). ANOVA results showed that only the four main effects and the interaction between the degree of nonlinearity and interaction factor and the estimation method factor had at least small effect sizes.

The impact of the degree of nonlinearity and interaction factor had the largest effect size with $\eta^2 = 0.1375$ [$F(3, 38268) = 7005.5968$, $p < 0.001$]. As the true data generation models became more nonlinear and non-additive, the sensitivity increased for all estimation models. For example, $sensitivity_{M\text{-}CART}$ increased from 0.5617 to 0.7193 and $sensitivity_{M\text{-}logit}$ increased from 0.5822 to 0.6657 when the degree of nonlinearity and interaction increased from 0 to 0.9 (Table 3).

The estimation model factor had a moderate effect size on sensitivity with η^2 equaled to 0.0508 [$F(3, 38268) = 834.1415$, $p < 0.001$]. Averaging across all design factors, the M-CART had the highest sensitivity, followed by the M-logit, S-CART, and S-logit ($sensitivity_{M\text{-}CART}$ = 0.6446 > $sensitivity_{M\text{-}logit}$ = 0.6211 > $sensitivity_{S\text{-}CART}$ = 0.6108 > $sensitivity_{S\text{-}logit}$ = 0.5703, Table 3).

The sample size [$F(3, 38268) = 241.8885$, $p < 0.001$, $\eta^2 = 0.0142$] and *ICC* [$F(1, 38268) = 913.1552$, $p < 0.001$, $\eta^2 = 0.0179$] had only small effect sizes. As the sample size and *ICC* increased, the sensitivity increased slightly across all estimation models (Table 3).

Table 3
Sensitivity Mean Values by Different Design Factors

Conditions	M-cart		M-logit		S-cart		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ICC								
0.1	0.6237	0.1204	0.6009	0.1166	0.5954	0.1193	0.5580	0.1307
0.3	0.6656	0.1088	0.6414	0.1136	0.6251	0.1143	0.5825	0.1518
Nonlinearity & interaction								
0	0.5617	0.1151	0.5822	0.1243	0.5280	0.1096	0.5425	0.1232
0.3	0.6133	0.1036	0.6007	0.1087	0.5730	0.1001	0.5516	0.1270
0.6	0.6842	0.0882	0.6360	0.1128	0.6495	0.0946	0.5709	0.1479
0.9	0.7193	0.0860	0.6657	0.0979	0.6905	0.0887	0.6161	0.1551
Sample size								
30-65	0.6195	0.1334	0.5899	0.1308	0.5887	0.1332	0.5634	0.1515
30-125	0.6540	0.1132	0.6316	0.1178	0.6194	0.1238	0.5754	0.1492
50-65	0.6403	0.1091	0.6247	0.1106	0.6046	0.1071	0.5658	0.1361
50-125	0.6647	0.1042	0.6384	0.0983	0.6283	0.1004	0.5765	0.1293
Average	0.6446	0.1168	0.6211	0.1168	0.6103	0.1180	0.5703	0.1420

The small interaction effect between the degree of nonlinearity and interaction factor and the estimation model factor was found [$F(3, 38268) = 241.8885$, $p < 0.001$, $\eta^2 = 0.0142$]. When the true models had non-linearity & interaction equaled to 0, using the M-logit and S-logit produced slightly higher sensitivity means than the M-CART and S-CART respectively

(sensitivity$_{\text{M-logit}}$ = 0.5822 > sensitivity$_{\text{M-CART}}$ = 0.5617; sensitivity$_{\text{S-logit}}$ = 0.5425 > sensitivity$_{\text{S-CART}}$ = 0.5280). Once the true relationship contains non-linearity and interaction, the M-CART outperformed the M-logit and S-CART, and the S-logit yielded the lowest sensitivity values.

Specificity

The distribution of specificity was shown in Figure 1. Across all data sets, around 69.34% of cases whose observed outcome were $y = 0$ were correctly predicted as $y = 0$ (M = 0.6934, SD = 0.1066). Same as before, only the four main effects and the interaction effect between the degree of nonlinearity and interaction factor and the estimation model factor were not trivial.

The sample size, *ICC*, and estimation method factors affected specificity performance on the same way as they impacted classification accuracy (Table 4), but the effect sizes were all small [$\eta^2 = 0.0190$ for sample size, $\eta^2 = 0.0093$ for *ICC*, and $\eta^2 = 0.0248$ for estimation method]. The impact of the interaction on specificity was also the same as that on classification accuracy, but with a larger effect size [$F(3, 38268) = 4437.3146$, $p < 0.001$, $\eta^2 = 0.0920$].

The impact of degrees of nonlinearity and interaction on specificity had a moderate effect size ($F(3, 38268) = 14.5058$, $p < 0.001$, $\eta^2 = 0.0882$), but this impact was different from that on classification accuracy (Table 4). For specificity, linear and additive models were associated the highest specificity values across all estimation models. As the models were increasingly dominated by non-linear and non-additive terms, the benefit of using CART algorithms, especially M-CART, increased. When having medium and large non-linearity and interaction, S-CART even had larger specificity means than M-logits.

Tau

The distribution of tau was shown in Figure 1. Averaging over conditions, prediction using models improved 35.11% predictive accuracy (M = 0.3511, SD = 0.1104) over using random allocations. Same as other criterion, only the main effects and the interaction between estimation methods and degrees of non-linearity and interaction had remarkable effect sizes.

Table 4
Specificity Mean Values by Different Design Factors

Conditions	M-cart		M-logit		S-cart		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ICC								
0.1	0.6917	0.0846	0.6981	0.1060	0.6801	0.0766	0.6730	0.1257
0.3	0.7158	0.0872	0.7150	0.1105	0.6954	0.0885	0.6848	0.1461
Nonlinearity & interaction								
0	0.7242	0.0801	0.7834	0.0759	0.6924	0.0821	0.7663	0.0910
0.3	0.6887	0.0942	0.7148	0.0965	0.6779	0.0794	0.6994	0.1105
0.6	0.6943	0.0854	0.6805	0.1052	0.6861	0.0832	0.6538	0.1277
0.9	0.7079	0.0827	0.6476	0.1064	0.6947	0.0866	0.5963	0.1472
Sample size								
30-65	0.6929	0.1062	0.6876	0.1276	0.6775	0.1018	0.6617	0.1495
30-125	0.7053	0.0903	0.7101	0.1138	0.6919	0.0853	0.6861	0.1439
50-65	0.6998	0.0787	0.7027	0.1005	0.6826	0.0756	0.6734	0.1326
50-125	0.7172	0.0645	0.7258	0.0853	0.6991	0.0631	0.6946	0.1137
Average	0.7048	0.0867	0.7066	0.1088	0.6878	0.0831	0.6789	0.1363

Small effect sizes were found for both the sample size factor [$F(3, 38268) = 339.8893$, $p < 0.001$, $\eta^2 = 0.0188$] and the degree of non-linearity and interaction factor [$F(3, 38268) = 1208.3781$, $p < 0.001$, $\eta^2 = 0.0223$]; Moderate effect sizes were found for the estimation method

factor [$F(3, 38268) = 1564.7292$, $p < 0.001$, $\eta^2 = 0.0879$] and the ICC factor [$F(1, 38268) = 2234.3002$, $p < 0.001$, $\eta^2 = 0.0418$]; the interaction term had a large effect size withη^2 equaled to 0.1141 [$F(3, 38268) = 2031.5126$, $p < 0.001$]. Even though the magnitudes of these effect sizes were different from that on classification, the impact of these designed factors on tau (Table 5) shared the same pattern as that on classification accuracy.

Table 5
Tau Mean Values by Different Design Factors

Conditions	M-cart		M-logit		S-cart		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ICC								
0.1	0.3629	0.1112	0.3362	0.0931	0.3225	0.1056	0.2790	0.0874
0.3	0.4033	0.1099	0.4123	0.1007	0.3568	0.1082	0.3178	0.1025
Nonlinearity & interaction								
0	0.3038	0.1038	0.4012	0.1021	0.2650	0.0907	0.3535	0.0987
0.3	0.3714	0.0971	0.3746	0.1078	0.3152	0.0900	0.3069	0.0890
0.6	0.4103	0.1031	0.3646	0.1039	0.3735	0.1023	0.2776	0.0887
0.9	0.4469	0.0983	0.3567	0.0971	0.4049	0.0978	0.2556	0.0870
Sample size								
30-65	0.3609	0.1282	0.3562	0.1172	0.3203	0.1191	0.2773	0.1020
30-125	0.3893	0.1143	0.3799	0.1065	0.3447	0.1091	0.3072	0.1058
50-65	0.3773	0.1071	0.3700	0.1014	0.3318	0.1022	0.2918	0.0921
50-125	0.4048	0.0936	0.3909	0.0861	0.3619	0.0974	0.3173	0.0821
Average	0.3831	0.1124	0.3742	0.1041	0.3397	0.1081	0.2984	0.0966

Discussion

In this study, I proposed and evaluated a multilevel CART algorithm for predicting binary outcomes in the person-within-cluster design. The proposed method can handle covariates at all levels, accounts for the clustering effects in multilevel data and automatically captures the nonlinearity and interaction, therefore enhances prediction accuracy.

Results indicated that M-CART outperformed S-CART in terms of classification accuracy, sensitivity, specificity, and tau, which is consistent with previous research that compared M-logits vs. S-logits (e.g., Guo & Zhao, 2000; Hox, 1998; Wong & Mason, 1985) and multilevel regression trees vs. single level regression trees (e.g., Hajjem, Bellavance & Larocque, 2011; Hajjem, Bellavance & Larocque, 2014; Martin, 2015; Sela & Simonoff, 2012). In particular, Hajjem, Bellavance and Larocque (2011) found that multilevel regression trees with either random intercepts or random slopes outperformed single level regression trees. All of these findings highlight the importance of explicitly modeling higher-level random effects when making predictions of outcomes in multilevel data.

Compared to multilevel logistic regression models, the proposed M-CART accounts for omitted nonlinear and interaction effects and has higher prediction accuracy as the degree of nonlinearity and interaction becomes high. The new M-CART proposed in this study inherits the advantage of CART algorithm which tests all possible nonlinearities and interaction terms, identifies significant ones, and prunes away the insignificant ones during the tree building process using Gini index and cost-complexity measure (Lee, Lessler, & Stuart, 2010; Steinberg & Colla, 2009; Timofeev, 2004). Our findings are corroborated by previous simulation studies which focus on continuous outcomes in multilevel data. Sela and Simonoff (2012) found that multilevel regression trees fitted multilevel continuous data better than multilevel linear

regressions when the data generation process is based on tree structures with non-linear and interaction relationship between predictors and outcomes. Hajjem, Bellavance & Larocque (2014) compared random forest algorithm with linear regression models, and found that random forest algorithms yielded smaller prediction errors than linear regressions. Furthermore, CART algorithms are not only found to yield more accurate predictions than linear regression models, but also found to outperform linear regressions with stepwise covariate selection procedure or nonlinear regressions (e.g., Lemon, Roy, Clark, Friedmann, & Rakowski, 2003; Razi & Athappilly, 2005).

The ability of CART to capture omitted nonlinearity and interaction effects have some ramifications. Our results showed that when the true relationship between covariates and the outcome was linear and additive, M-CART and S-CART resulted in worse prediction than the corresponding logistic regression models. Similar phenomenon was found in the study of Sela and Simonoff (2012) which showed that multilevel linear models had lower prediction error than multilevel regression trees when there was no non-linear or interaction effects. However, it is worth mentioning that in our study M-CART was close to M-logits in terms of sensitivity (2% difference) when the relationship is linear and additive, indicating M-CART can even recover the linear and additive true relationship fairly and has very strong capability to identify these cases with positive outcome (y =1).

Implications and Limitations

Overall, I demonstrated the usefulness of the proposed M-CART algorithm and therefore recommended to use this method when data conditions are similar to the ones simulated in this study, especially when the true relationships between covariates and outcomes are not as simple as linear and additive. Even in situations where the true relationships between covariates and

outcomes are more likely to be linear and additive, researchers can still use the proposed M-CART algorithm because M-CART and M-Logits performed similarly in terms of the accuracy in predicting the occurrence of an event (i.e., sensitivity).

All findings and implications should be considered in light of the limitations. The proposed M-CART algorithm has a few drawbacks. First, like all CART algorithms, the proposed M-CART has covariates selection bias. Covariates with more potential cut-off points are more likely to be chosen as splitting candidates. Hence, continuous covariates are more likely to be chosen than categorical covariates. Second, the structure of random effect component should be pre-defined. That is, when fitting M-CART, researchers need to make decision about the structure of random effect matrix.

CHAPTER III

PROPENSITY SCORE ESTIMATION USING MULTILEVEL CART

Introduction

Causal inference has been a growing area of educational research. Randomized controlled trials (RCTs), the gold standard for drawing casual inference (Austin, 2011; Meldrum, 2000), are difficult and sometimes impossible to implement in educational research due to ethical and practical reasons. On the contrary, non-randomized controlled trials (non-RCTs), such as observational studies, are more widely used. Unlike RCTs, samples in non-RCTs are not randomly assigned into treatment and control groups, and therefore systematic difference might exist between treated and controlled groups.

Traditionally, researchers rely heavily on the use of regression models to account for imbalances on pretreatment covariates (e.g, Altman, 2005; Pocock, Assmann, Enos, & Kasten, 2002; Senn, 1989). Since the late 1990s, propensity score analysis (PSA, Rosenbaum & Rubin, 1983) become the prevalent method for eliminating the imbalances (Thoemmes & Kim, 2011). PSA estimates treatment effects by conditioning on propensity scores (PSs) which refer to the conditional probabilities of assigning individuals into the treated group given a series of observed pretreatment covariates (Rosenbaum & Rubin, 1983). PSA has its advantages over using traditional regression models, such as allowing to summarize all covariates imbalances simultaneously into one single score (Perkins, Tu, Underhill, Zhou, & Murray, 2000) and isolating outcome prediction from modeling treatment assignment mechanism (McCaffrey, Griffin, Almirall, Slaughter, Ramchand, & Burgette, 2013; Rickles, 2012).

In the past, non-RCT studies where casual inferences are drawn from have been limited to single-level settings. Nowadays, increasing researchers are aware that the nature structure of data collected in those studies are often multilevel. To apply PSA into multi-site non-RCTs (M-PSA), it is important to take into account clustering effects of data when estimating PSs. In single-level non-RCTs, estimating casual treatment effects via PSA assumes that the given treatment effect is stable across individuals regardless of what treatments are received by other individuals and how treatments are assigned (Stable-Unit-Treatment-Value assumption, Cox, 1985; Rubin, 1978, 1980). This assumption becomes inevitably violated under M-PSA. For example, in a multi-site non-RCT, the impact of a new teaching strategy on students' academic achievement is affected by not only different teachers who deliver the instruction (i.e., different intervention fidelities across teachers), but also students' classmates due to the collective learning process (i.e., peer effects). That is, in multi-site non-RCTs, even though a given treatment effect is stable across individuals within the same clusters, it is more than often that the treatment effect varies across clusters. Failure to account for clustering effects in such case yields biased PSs and treatment effect estimations. However, M-PSA has not received enough attention from both a methodological as well as an applied perspective (Bellara, 2013). The estimation of PSs in multi-site non-RCTs (M-PSs) remains as using either single level logistic regressions (S-logits) or multilevel logistic regressions (M-logits) for a long time without much development.

As data mining techniques become increasingly popular, a few researchers recently started to adopt Classification and Regression Tree algorithm (CART, Breiman, Friedman, Olshen, & Stone, 1984), a promising data mining approach, for estimating PSs (e.g., Lee, Lessler, & Stuart, 2010; Westreich, Lessler, & Funk, 2010). CART overcomes drawbacks of logistic regressions (Timofeev, 2004; Westreich, Lessler, & Funk, 2010) and is more appropriate

than some other data mining techniques for PSA (Lee, Lessler, & Stuart, 2010). However, the use of CART for PSs estimation is still at its infancy. More important, no multilevel CART algorithm (M-CART) has been applied for estimating M-PSs. To fill in this research gap, in this study, the M-CART algorithm proposed in chapter II is employed for estimating M-PSs and its performance is examined. In the following sections, the related theoretical framework is presented first, including the treatment effects and PSA in multi-site trials. Then, a simulation study is conducted to compare the performance of four M-PS estimation methods, M-CART, M-logit, single level CART (S-CART), and S-logit, in M-PSA across different simulation conditions. This study serves to answer the following questions:

1. How do M-PS estimation methods interacting with simulated conditions impact the balances of covariates for different M-PS estimation methods?
2. How do M-PS estimation methods interacting with simulated conditions impact the estimations of treatment effects conditioning on different types of M-PS estimations?

Theoretical Framework

Rubin's Causal Model in Multi-site Trials

Rubin's causal model (Rubin, 1974, 1976, 1978, 1980, 2005) is the most popular model for defining casual treatment effects for a wide variety of disciplines including education (Rickles & Seltzer, 2014). Based on Rubin's casual model, the casual treatment effect of a given intervention for an individual is measured as the difference of potential outcomes if this individual would participate in both treatment ($T = 1$) and control ($T = 0$) conditions (assume two intervention assignments).

To extend the Rubin's casual model to the multi-site trials, I focus on a two-level structure where N individuals ($i = 1, 2,\ldots, N$) are nested within H clusters ($j = 1, 2,\ldots, H$). A

individual *i* nested within *j* cluster has pretreatment covariates X_{ij} measured at both the individual level (U_{ij}) and the cluster level (V_j) [i.e., $X_{ij} = (U_{ij}, V_j)$], and two potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$ associated with two treatment assignments $T = 1$ and $T = 0$. The population average treatment effect (ATE) is expressed as

$$\delta_{ATE} = E\{E\,[Y_{ij}(1)|\, j - Y_{ij}\,(0)|\, j]\}, \tag{3.1}$$

where $Y_{ij}(1)|j$ is the potential outcome value of the *i* individual nested within cluster *j* if this individual receive the treatment and $Y_{ij}\,(0)|\, j$ is the potential outcome value of the same individual if he/she do *not* receive the treatment.

Since a given individual can only be exposed to one experimental condition, $Y_{ij}(1)$ and $Y_{ij}(0)$ can never be measured at the same time for any individual, which causes a problem in estimation. In multi-site RCTs where the treated and controlled groups are homogeneous on pre-treatment covariates *X* due to randomization, $E\,[Y_{ij}(1)|\, j - Y_{ij}\,(0)|\, j]$ can be simply considered as the expected difference on outcomes between treatment and control groups within cluster *j*. So in multi-site RCTs, ATE can be defined as

$$\delta_{ATE} = E\,\{E\,[Y(1)|\, j] - E[Y(0)|\, j]\}. \tag{3.2}$$

However, in multi-site non-RCTs, individuals in treatment and control groups are heterogeneous on *X*. Therefore, researchers should first adjust for the imbalances of *X* between groups in order to estimate treatment effects using equation 3.2.

Propensity Score Analysis in Multi-site Non-RCTs

In multi-site non-RCTs, an M-PS is defined as the conditional probability π_{ij} of an individual *i* nested within cluster *j* receives the treatment intervention, given pretreatment covariates X_{ij} [$X_{ij} = (U_{ij}, V_j)$]. That is, $\pi_{ij} = p\,(T = 1|\, X_{ij})$, where $0 < \pi_{ij} < 1$ (Hong, 2004). The

estimated M-PSs are considered as balancing scores, meaning that treated and control individuals with the same M-PSs are considered as homogeneous given X. M-PSA is built on two assumptions. The first assumption is strongly ignorable treatment assignment, indicating that the treatment assignment is not a function of that individual's potential outcome conditional on covariates, $T\perp [Y(1), Y(0)]|X$, (Rosenbaum & Rubin, 1983). The second assumption is multilevel exchangeability assumption, assuming that treatment effects are exchangeable not only within but also across treatment conditions (Hong, 2004; Hong & Raudenbush, 2006). M-PSA, following the similar procedure used in single level PSA, includes covariates selection, PS estimation, PS conditioning, balance diagnosis, and treatment effect estimation.

Covariates selection. A PSA can only be as good as the covariates that are presented to researchers (Thoemmes & Kim, 2011). Many studies have explored the question of what covariates are important to be included in estimating PSs, and the most fundamental suggestion they provide is to include all confounding variables which are significantly correlated with both outcomes and treatment assignments (e.g., Austin, 2011; Emsley, Lunt, Pickles, & Dunn, 2008; Harder, Stuart, & Anthony, 2010; Millimet & Tchernis, 2009; Rosenbaum & Rubin, 1984).

PS estimation. In existing literature, three types of logistic regression models are commonly used for estimating M-PSs. (a) Single level logistic regression models with cluster-level covariates (e.g., Rosenbaum, 1986). This type of models is widely used in the past but has been demonstrated to be ineffective on achieving covariate balances and estimating treatment effects because it fails to control for clustering effects (Thoemmes & West, 2011; Kim & Seltzer, 2007). (b) Fixed effect logistic regressions with dummy-coded cluster memberships as covariates (e.g., Leite et al., 2015; Rosenbaum, 1986; Thoemmes & West, 2011). This type of regressions accounts for clustering influence and therefore outperforms single level logistic regressions

(Arpino & Mealli, 2011; Yu, 2012). However, the increase of cluster numbers causes the degrees of freedom to be decreased which would be troublesome (Rosenbaum, 1986). (c) Multilevel logistic regressions (e.g., Arpino & Mealli, 2011; Hong & Raudenbush, 2005; Kelcey, 2009; Kim & Seltzer, 2007, Yu, 2012). Previous studies show that M-logits produce more accurate M-PSs and less biased treatment effects than the other two estimation methods (e.g., Kelcey, 2009; Li, Zaslavsky, & Landrum, 2013; Rickles, 2012 ; Su & Cortina, 2009; Thoemmes & West, 2011; Yu, 2012), because M-logits control for the effect of clustering (Thoemmes & West, 2011), adjust standard errors properly (Raudenbush & Bryk, 2002), and allow modeling complex treatment assignment mechanisms to reflect realistic study designs (Kelcey, 2009).

The use of logistic regressions has its attractions of a) mathematically constraining probabilities in the range [0, 1] (Kleinbaum & Klein; 2002); b) easily converging on parameter estimations (Westreich, Lessler, & Funk, 2010); (c) being implemented in most statistical packages (Westreich, Lessler, & Funk, 2010); (d) being widely known and understood by researchers. However, using logistic regressions for M-PSA has been criticized. First, logistic regressions require strong model assumptions, especially multilevel logistic regressions. For example, the log odds of the outcome should be linearly related with covariates; random effects should be uncorrelated with covariates at all levels (Hox, Moerbeek, & van de Schoot, 2010; McMahon, Pouget, & Tortu, 2006). If any model assumption is violated, covariate balances and unbiased treatment effects estimation may not be achieved (Drake, 1993). Second, M-PSA via logistic regressions might not be applicable when the number of pretreatment covariates is larger than the number of samples (Breiman, 2001). Third, using logistic regressions requires researchers to predefine model equations (Bursac, Gauss, Williams, & Hosmer, 2008; D'Agostino Jr; 1998; Lee, Lessler, & Stuart, 2010). In other word, when logistic regressions are

chosen to estimate M-PSs, researchers need to clearly know the true treatment assignment mechanism and correctly specify the relationship between treatment assignment and covariates, which are unknown in reality.

The CART algorithm is a relatively new method for replacing logistic regressions in PSA (Westreich, Lessler and Funk, 2010). It does not request strong model assumptions that are required for logistic regressions, such as independence of residuals, non-collinearity among predictors. Also, it is a data-driven method which automatically includes the significant variables and remove non-significant ones (Timofeev, 2004), and is able to automatically capture non-linear and non-additive terms (Lee, Lessler, & Stuart, 2010; Steinberg & Colla, 2009). Except for these well-known advantages of using CART, CART also has its specific merits when estimating PSs. First, CART is widely found to outperform logistic models when dealing with a large number of covariates, which is an importance feature since PSA typically requires many pretreatment covariates (Westreich, Lessler, & Funk, 2010). Second, CART explicitly reports the probability (i.e., PS) of an observation being assigned into the treatment group (Westreich, Lessler, & Funk, 2010). Not all data mining techniques can export probability values as clear as CART.

So far, only a few simulation studies (e.g., Pirracchio, Petersen, & van der Laan, 2014; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Lee, Lessler, & Stuart, 2010) compare the performance of logistic regressions and CART when conducting PSA. These studies generally support that logistic regressions outperform CART only when the true association between covariates and assignment is additive and linear (i.e., models with only main effects). As models become more and more complex, CART results in more accurate estimations of treatment effects compared to logistic regressions. In reality, the relationship between treatment

assignment and covariates is typically much more complicated than linear and additive association due to the complexity of treatment allocation mechanism. While, researchers who never know the true relationship tend to create overly simplified models by falsely assuming a linear and/or additive relationship between covariates and treatment in PSA (Thoemmes & Kim, 2011). Therefore, the CART algorithm is considered as a new and promising approach for PSA. However, all of these comparison studies are under single-level settings. The comparison on performance based on using multilevel CART and multilevel logistic regressions in M-PSA is rarely examined.

PS conditioning. Conditioning is a step of utilizing estimated M-PSs to adjust the systematic difference between treatment and control groups on *X*. There are two main categories of conditioning. One is to make an adjustment prior to calculating treatment effects, such as matching (Arpino, B., & Mealli, 2011; Rosenbaum & Rubin, 1985) and stratification (Rosenbaum & Rubin, 1984; Xiang & Tarasawa, 2015). The other category is to balance pretreatment covariates while estimating treatment effects, such as covariates adjustment using PS (Rosenbaum & Rubin, 1983) and inverse probability of weighting (IPTW, Rosenbaum, 1987; Li, Zaslavsky, & Landrum, 2013). There is no conditioning approach that outperforms others across all situations. Propensity score matching excels when the treated group is contained within a larger control pool (Hade & Lu, 2014), but any unmatched individuals would be discarded and removed from treatment effect estimations. Stratification keeps all individuals, but it has been found to have greater biases than matching and IPWT on treatment effect estimations (Austin, 2009; Austin, Grootendorst, & Anderson, 2007; Austin & Mamdani, 2006; Lunceford & Davidian, 2004). Covariate adjustment approach also includes all individuals and is preferred only when treated and control groups have narrow overlap on PS distribution (Hade & Lu,

2014). IPTW leads to large biases when individuals have extreme low or high PSs (Austin, 2011). Overall, matching and covariates adjustment using PS are the two most frequently used techniques according to reviews (Hade & Lu, 2014; Thoemmes & Kim, 2011).

M-PS conditioning can be conducted either within (conditioning within clusters, CWC) or across clusters (conditioning across clusters, CAC). CWC indicates that the conditioning procedure is conducted within each cluster, while CAC allows an individual in a cluster be paired with a similar individual from other clusters. Generally, CWC is ideal because it ensures covariate balances and unbiased treatment effects within clusters and provides a good control for cluster-level covariates simultaneously (Lingle, 2009; Thoemmes, 2009; Thoemmes & West, 2011).

Balance diagnosis. After conditioning, it needs to examine whether the balance on every covariate between treated and controlled groups is achieved. Balances diagnosis is often assessed by measures of effect size. Standard mean difference (Δ, Cohen, 1988), also known as Cohen's *d*, is the most commonly used effect size measure for diagnosing balance (Rosenbaum & Rubin, 1985). It is defined as the mean difference of a given covariate between treated and controlled groups, divided by standard deviation. Guidelines indicate that $\Delta = 0.25$ represents reasonable cut-off for acceptable standardized biases (Bellara, 2013; Rosenbaum & Rubin, 1985; Stuart, 2010).

Treatment effect estimation. Treatment effect estimations are various according to the PS conditioning approaches chosen. For example, when matching or stratification conditioning approaches is used, treatment effects can be estimated directly as the mean difference of outcome variable *Y* between paired treated and controlled individuals (Imbens, 2004). When IPTW is selected, the treatment effect is calculated as the weighted mean difference between all treated

and controlled individuals (Lanza, Moore, & Butera, 2013). When using covariate adjustment approach, the outcome variable is regressed on the estimated M-PSs and an indicator variable denoting the treatment assignment, so that the casual effect of treatment effect is estimated controlling for M-PSs (Austin, 2011).

Simulation Study

In this simulation study, I compared the performance of M-CART algorithm proposed in previous chapter with M-logit, S-CART, and S-logit for estimating PSs when having multi-site non-RCTs. Data sets were generated to mimic two-level multi-site non-RCTs with 24 data generation conditions (2 conditional intra-class correlations * 3 degrees of nonlinearity and non-additivity relationships between X and T *4 sample sizes). For each generated data set, the four M-PS estimation methods combined with 2 different conditioning approaches were applied. So, in total, 192 situations (24 data conditions * 4 estimation methods * 2 conditioning approaches) with 300 replications in each situation were analyzed. Both data generation and analysis were conducted via R 3.2.4 (R Core Team, 2016).

Models and Fixed Parameters for Data Generation

Data generation in this study included generating the treatment assignment T, true M-PSs π, outcome variable Y, and pretreatment covariates X in multi-site non-RCT design. Parameters used in data generation followed the simulation studies of Bellara (2013) and Lee, Lessler, and Stuart (2010).

Treatment assignment. Each simulated dataset contained N number of individuals (i = 1,2,…, N) who were nested within H number of clusters (j =1,2,…, H). Because in multi-site non-RCTs, the treatment assignment should be assigned at individual level, N numbers of T were generated from a Bernoulli distribution with expected probability of 0.5 which ensured that, on

average, half of samples received the treatment intervention (T =1) and the other half received the control intervention (T =0).

Covariates. 10 covariates were simulated including both individual and cluster level variables (Table 6). The individual level covariates contained 6 covariates (U_1, U_2, …, U_6) correlated to both treatment and outcome variables, and 1 covariate U_7 only associated with the

Table 6
All Variables Generated in This Simulation Study

Variable	Distribution	Covariate Level	Relationship
T	Binominal (p=.5)	individual level	
Y	Normal (M=0, SD=1)	individual level	
U_1	Binominal (p=.5)	individual level	associated with both T and Y
U_2	Normal (M=0, SD=1)	individual level	associated with both T and Y
U_3	Normal (M=0, SD=1)	individual level	associated with both T and Y
U_4	Normal (M=0, SD=1)	individual level	associated with both T and Y
U_5	Normal (M=0, SD=1)	individual level	associated with both T and Y
U_6	Normal (M=0, SD=1)	individual level	associated with both T and Y
U_7	Normal (M=0, SD=1)	individual level	associated with Y
V_1	Normal (M=0, SD=1)	cluster level	associated with both T and Y
V_2	Binominal (p=.5)	cluster level	associated with both T and Y
V_3	Normal (M=0, SD=1)	cluster level	associated with Y

outcome. The cluster level had 2 covariates (V_1, V_2) correlated to both treatment and outcome variables and 1 outcome-associated covariate (V_3). In these 10 covariates, U_1 and V_2 were dichotomous variables generated from a binominal distribution with an expected probability of .5. The rest 8 covariates were continuous variables falling into standardized normal distributions.

True M-PS. A two-level logistic regression model with random slopes u was used for generating the true M-PSs, in which main effects, quadratic terms, and interaction terms of U_1, $U_2, \ldots, U_6$, V_1, and V_2, were used as covariates:

$$logit\left(T_{ij}\right) = \beta_{0j\_T} + \beta_{1j\_T}U_{1ij} + \beta_{2j\_T}U_{2ij} + \cdots + \beta_{6j\_T}U_{6ij} \quad (3.3)$$
$$+ \beta_{7j\_T}U_{2ij}{}^2 + \beta_{8j\_T}U_{3ij}{}^2 + \beta_{9j\_T}U_{4ij}{}^2$$
$$+\beta_{10j\_T}U_{1ij}U_{2ij} + \beta_{11j\_T}U_{3ij}U_{4ij} + \beta_{12j\_T}U_{5ij}U_{6ij} + \beta_{13j\_T}U_{2ij}U_{4ij}U_{6ij}$$

$$\beta_{0j\_T} = \gamma_{00\_T} + \gamma_{01\_T}V_{1j} + \gamma_{02\_T}V_{2j} + \gamma_{03\_T}V_{1j}{}^2 + \gamma_{04\_T}V_{1j}V_{2j} + u_{0j\_T}$$

$$\beta_{1j\_T} = \gamma_{10\_T} + \gamma_{11\_T}V_{1j} + \gamma_{12\_T}V_{2j} + \gamma_{13\_T}V_{1j}{}^2 + u_{1j\_T}$$

$$\beta_{2j\_T} = \gamma_{20\_T} + \gamma_{21\_T}V_{1j} + \gamma_{22\_T}V_{2j} + \gamma_{23\_T}V_{1j}V_{2j} + u_{2j\_T}$$

$$\beta_{3j\_T} = \gamma_{30\_T} + u_{3j\_T}$$

$$\beta_{4j\_T} = \gamma_{40\_T}$$

$$\vdots$$

$$\beta_{13j\_T} = \gamma_{13\ 0\_T}.$$

The grand mean γ_{00_T} was fixed as 0. All coefficients of main effects for level-1 covariates were fixed as 0.2 ($\gamma_{10\_T} = \gamma_{20\_T} = \gamma_{30\_T} = \gamma_{40\_T} = \gamma_{50\_T} = \gamma_{60\_T}$ = 0.2, Table 7) and coefficients of main effects for level-2 covariates were set to be 0.4 (γ_{01_T} = γ_{02_T} = γ_{11_T} = γ_{12_T} = γ_{21_T} = γ_{22_T} = 0.4). Coefficient values of quartic and interaction terms varied to indicate different degrees of nonlinearity and interaction between *T* and *X* (details are described in the next section). The level-2 random effect *u* was generated from a multivariate normal distribution,

MVN ~ (0, $\begin{bmatrix} \sigma^2_{u0j\_T} = 0.37/1.41 & 0 & 0 & 0 \\ 0 & \sigma^2_{u1j\_T} = 0.5 & 0 & 0 \\ 0 & 0 & \sigma^2_{u2j\_T} = 0.5 & 0 \\ 0 & 0 & 0 & \sigma^2_{u3j\_T} = 0.5 \end{bmatrix}$). The $\sigma^2_{uo\_T}$ was set to be either 0.37 or 1.41 to represents different conditional intra-class correlations (*ICC*s).

Table 7
Values of Parameters Used in Different Degrees of Non-additivity and Non-linearity

Parameter	Scenario 1	Scenario 2	Scenario 3
γ_{00_T}	0	0	0
γ_{01_T}	0.4	0.4	0.4
γ_{02_T}	0.4	0.4	0.4
γ_{03_T}	0	0.4	0.6
γ_{04_T}	0	0.4	0.6
γ_{10_T}	0.2	0.2	0.2
γ_{11_T}	0	0.4	0.6
γ_{12_T}	0	0.4	0.6
γ_{13_T}	0	0.4	0.6
γ_{20_T}	0.2	0.2	0.2
γ_{21_T}	0	0.4	0.6
γ_{22_T}	0	0.4	0.6
γ_{23_T}	0	0.4	0.6
γ_{30_T}	0.2	0.2	0.2
γ_{40_T}	0.2	0.2	0.2
γ_{50_T}	0.2	0.2	0.2
γ_{60_T}	0.2	0.2	0.2
γ_{70_T}	0	0.2	0.4
γ_{80_T}	0	0.2	0.4
γ_{90_T}	0	0.2	0.4
$\gamma_{10\ 0\_T}$	0	0.2	0.4
$\gamma_{11\ 0\_T}$	0	0.2	0.4
$\gamma_{12\ 0\_T}$	0	0.2	0.4
$\gamma_{13\ 0\_T}$	0	0.2	0.4

Outcome variable. The *Y* was generated from a two-level random slope model to reflect the impact of the *T* and *X* on the *Y*. In this model, *Y* was regressed on 2 covariates that only associated with *Y*, and the true π. The π was included in this model to make sure that all the main, quadratic and interaction terms defined in the M-PS generating model (equation 3.3) impacted both *T* and *Y* and therefore were true confounders.

$$Y_{ij} = \beta^{Y}_{oj\_T} + T * \delta + \beta^{Y}_{1j\_T} U_{7ij} + \beta^{Y}_{2j\_T} \pi_{ij} + e^{Y}_{ij\_T} \quad (3.4)$$

$$\beta^{Y}_{oj\_T} = \Upsilon^{Y}_{00\_T} + \Upsilon^{Y}_{01\_T} V_{3j} + u^{Y}_{0j\_T}$$

$$\beta^{Y}_{1j\_T} = \Upsilon^{Y}_{10\_T} + u^{Y}_{1j\_T}$$

$$\beta^{Y}_{2j\_T} = \Upsilon^{Y}_{20\_T}$$

The grand mean $\Upsilon^{Y}_{00\_T}$ was 0. The coefficient for the level-1 pretreatment covariate,$\Upsilon^{Y}_{10\_T}$, was set as 0.2 and the coefficient for the level-2 pretreatment covariate, $\Upsilon^{Y}_{01\_T}$ was 0.4. The impact of M-PSs on *Y* was fixed as 0.5 (i.e., $\Upsilon^{Y}_{20\_T}$ = 0.5). The true treatment effect δ was fixed as 0.5. The level-1 residual was synthetized from a standard normal distribution. The level-2 random effect was generated from a multivariate normal distribution, MVN ~ (0,

$$\begin{bmatrix} \sigma^2_{u^{Y}_{0j\_T}} = 0.25 & 0 \\ 0 & \sigma^2_{u^{Y}_{1j\_T}} = 0.25 \end{bmatrix}$$

), so that the conditional *ICC* = .20 indicating moderate clustering effects existed in *Y*.

Design Factors

Conditional ICC in the M-PS model. Conditional *ICC*s in the M-PS model were manipulated through the variance of random effects σ^2_{u0j}. *ICC*s were set as .10 and .30, which represented small and large clustering effects in educational settings (Hedges & Hedberg, 2007;

Hox, Moerbeek, & van de Schoot, 2010; Luo, Cappaert, & Ning, 2015). Since $ICC = \frac{\sigma^2_{u0j}}{\sigma^2_{u0j}+\sigma^2_e}$ (Maas & Hox, 2005) and $\sigma^2_{e\_T} = \frac{\pi^2}{3}$ which is the default scaling value for all M-logits (Snijders & Bosker, 1999), $\sigma^2_{u0j\_T}$ was set to be 0.37 and 1.41 to guarantee that *ICC* could be 0.1 and 0.3 respectively.

Degree of nonlinearity and interaction. For the M-PS generation model (equation 3.3), three scenarios were considered to indicate three types of relationship between *T* and *X* (Table 3). The M-PS model used in scenario 1 contained only main effects with all coefficients for nonlinear and interaction terms as 0 and coefficients for main effects were defined as mentioned before. In scenario 2, small degree of nonlinearity and interaction was allowed. In this scenario, coefficients for main effects kept the same, coefficients for level-1 and level-2 nonlinear and interaction terms were set to be 0.2 and 0.4 respectively. In scenario 3, large degree of nonlinearity and interaction was simulated. Coefficients for main effects were still the same, but the coefficients for level-1 and level-2 nonlinear and interaction terms increased to 0.4 and 0.6 respectively. From scenario 1 to 3, the relationship between *X* and *T* was increasingly dominated by nonlinear and interaction terms.

Sample size. Different number of clusters (*Nc*) and cluster sizes (*Ns*) were simulated. *Nc* were 15 and 30 to indicate small and moderate number of clusters (e.g., Finch & French, 2011; Kwok, Luo & West, 2010; Maas & Hox, 2005). *Ns* were 20 and 50 to represent small and moderate cluster sizes (e.g., Finch & French, 2011; Maas & Hox, 2005; Jak, Oort & Dolan, 2013; Peugh & Enders, 2010). Combining *Ns* and *Nc* conditions, 4 different sample size conditions with total sample size *N* were used in this study from 300 (20*15=300) to 1500 (50*30=1500).

Analysis Procedure

Covariates selection. All covariates associated with treatment assignment *T* were selected to estimate M-PSs. That is, 8 pretreatment covariates U_1, U_2, …, U_6, V_1, and V_2 were used.

M-PS estimation. Every generated data set was fitted using four estimation methods: M-CART, M-logit, S-CART, and S-logit. For all estimation models, only the first-order terms (i.e., main effects) of the 8 selected pretreatment covariates were used as predictors. Only first-order terms were used because majority of existing PS studies chose models with only main effects (Thoemmes & Kim, 2011) since researchers do not have prior knowledge of nonlinear and interaction effects and tend to used over-simplified models. This setting allowed us to examine whether and how much the proposed M-CART is able to capture unspecified non-linear and interaction effects in M-PS estimation.

For applying M-CART, the tolerance of change of log-likelihood values was set as 0.0001; the maximum number of iterations was 1000; the minimum number of observations in a node was 2. It should be noticed that no pruning procedure was needed in this case, because generalization and overfitting were no longer a concern when estimating M-PSs (Rubin, 1997). These settings were also applied to S-CART, except for the tolerance of log-likelihood change which was not applicable for S-CART.

The M-logit also only contained the main effects of 8 predictors as shown below:

$$logit\left(T_{ij}\right) = \beta_{0j\_Mlogit} + \beta_{1j\_Mlogit}U_{1ij} + \beta_{2j\_Mlogit}U_{2ij} + \cdots + \beta_{6j\_Mlogit}U_{6ij} \qquad (3.5)$$

$$\beta_{0j\_Mlogit} = \gamma_{00\_Mlogit} + \gamma_{01\_Mlogit}V_{1j} + \gamma_{02\_Mlogit}V_{2j} + u_{0j\_Mlogit}$$

$$\beta_{1j\_Mlogit} = \gamma_{10\_Mlogit} + u_{1j\_Mlogit}$$

$$\beta_{2j\_Mlogit} = \gamma_{20\_Mlogit} + u_{2j\_Mlogit}$$

$$\beta_{3j\_Mlogit} = \gamma_{30\_Mlogit} + u_{3j\_Mlogit}$$

$$\beta_{4j\_Mlogit} = \gamma_{40\_Mlogit}$$

$$\beta_{5j\_Mlogit} = \gamma_{50\_Mlogit}$$

$$\beta_{6j\_Mlogit} = \gamma_{60\_Mlogit}.$$

In this two-level random slope logit regression model, intercept and the first three regression coefficients were allowed to be random, which was the same as the M-PS generating model. Maximum likelihood estimation via Laplacian approximation was used as the estimator. Similarly, main effects of 8 predictors were used for fitting S-logits as shown in equation 3.6:

$$logit\ (T) = \beta_{0\_Slogit} + \beta_{1\_Slogit}U_1 + \beta_{2\_Slogit}U_2 + \cdots + \beta_{6\_Slogit}U_6 + \beta_{7\_Slogit}V_1 + \beta_{8\_Slogit}V_2 + e_{Slogit}. \quad (3.6)$$

Estimated probabilities from fitting these M-PS models were estimated M-PSs (i.e., $\hat{\pi}$).

PS conditioning. The estimated M-PSs were then used in matching and covariate-adjustment using PS. Those two conditioning approaches were chosen because they were the top two commonly used methods (Hade & Lu, 2014; Thoemmes & Kim, 2011) and could be good representatives of the two types of conditioning methods. One-to-one matching within clusters was applied. Each treated individual within cluster *j* was paired with only one controlled individual within the same cluster if their $\hat{\pi}$ were similar or the same. Matching with replacement was allowed, meaning one controlled individual would be used for pairing more than one treated individuals. For covariate-adjustment using PS, the $\hat{\pi}$ was used as a covariate to predict the *Y* using the model shown in the treatment effect estimation procedure (equation 3.8).

Treatment effect estimation. Given using matching, the ATE was estimated as the average score of estimated within-cluster ATE ($\widehat{\delta_{ATE\ j}}$):

$$\widehat{\delta_{ATE}} = \sum_{j=1}^{H}(\frac{N_j}{N}\widehat{\delta_{ATE}}_j), \tag{3.7}$$

where $\widehat{\delta_{ATE}}_j = E[Y(1)|j] - E[Y(0)|\,j]$, N_j is the cluster size in cluster j. For covariate-adjustment, $\widehat{\delta_{ATE}}$ equaled to the estimated β_{1j_ATE} by fitting the following multilevel logistic regression model.

$$Y_{ij} = \beta_{0j\_ATE} + \beta_{1j\_ATE}T_{ij} + \beta_{2j\_ATE}U_{7ij} + \beta_{3j\_ATE}\hat{\pi}_{ij} + e_{ij\_ATE} \tag{3.8}$$

$$\beta_{0j\_ATE} = \gamma_{00\_ATE} + \gamma_{01\_ATE}V_{3j} + u_{0j\_ATE}$$

$$\beta_{1j\_ATE} = \gamma_{10\_ATE}$$

$$\beta_{2j\_ATE} = \gamma_{20\_ATE}$$

Evaluation Criterion

Two criterion were used to evaluate the performance of applying M-CART, M-logit, S-CART, and S-logit in M-PSA. The first one was standardized mean difference (Δ) used to measure the performance of balancing covariates. Ways of estimating Δ were also determined by conditioning methods. For matching within cluster conditioning, standardized mean difference for a given continuous covariate *x* (Δ_x) equaled to the average score of within-cluster Δ ($\Delta_{x,j}$),

$$\Delta_x = E(\Delta_{x,j}) = E(\frac{\bar{X}_{T=1|x,j} - \bar{X}_{T=0|x,j}}{\sqrt{\frac{(n_{T=1|x,j})s^2_{T=1|x,j} + (n_{T=0|x,j})s^2_{T=0|x,j}}{n_{T=1|x,j} + n_{T=0|x,j}}}}) \tag{3.9}$$

where $\bar{X}_{T=1|x,j}$ and $\bar{X}_{T=0|x,j}$ are the mean x for treatment and control groups given cluster *j* respectively. The denominator is the pooled standardized deviation. Δ_x for a dichotomous covariate *x* was estimated as

$$\Delta_x = E(\Delta_{x,j}) = E(\frac{prev_{T=1|x,j} - prev_{T=0|x,j}}{\sqrt{prev_{x,j}\,(1 - prev_{x,j})}}) \tag{3.10}$$

where $prev_{T=1|x,j}$ and $prev_{T=0|x,j}$ are the prevalence of $x = 1$ for treated and controlled groups within cluster j respectively.

For covariate-adjustment using PS, $\Delta_x$was estimated using a multilevel model (equation 3.11) in which the x is the outcome and $\hat{\pi}$ and T are predictors.

$$f(X_{ij}) = \beta_{0j\_\Delta} + \beta_{1j\_\Delta}T_{ij} + \beta_{2j\_\Delta}\hat{\pi}_{ij} + e_{ij\_\Delta} \quad (3.11)$$

$$\beta_{0j\_\Delta} = \gamma_{00\_\Delta} + u_{0j\_\Delta}$$

$$\beta_{1j\_\Delta} = \gamma_{10\_\Delta}$$

$$\beta_{2j\_\Delta} = \gamma_{20\_\Delta}$$

In this model, $f(.)$ indicate an identity link function when x is continuous and a logit link function when x is dichotomous. γ_{10_Δ} measures the unstandardized difference between treated and controlled groups on x. Thus, the Δ_x could be defined based on the t statistics of γ_{10_Δ} using equation 3.12 (Rosenthal & Rosnow, 1991),

$$\Delta_x = \frac{2t}{\sqrt{df}}, \quad (3.12)$$

where t is the t-test score of γ_{00_Δ}, df is the degree of freedom. Smaller standardized mean differences indicate the samples are approaching a state of balance. The value 0.25 was used as the cut-off, meaning that $\Delta_x \leq 0.25$ was considered as achieving balance on x.

The more important criteria was the relative bias (RB) of $\hat{\delta}$ used to indicate the accuracy of treatment effect estimations which is the goal of conducting PSA. The relative bias was defined as

$$\text{RB} = \frac{\hat{\delta} - \delta}{\delta} \quad (3.13)$$

where δ is the true treatment effect 0.5. Due to the importance of this criteria, a factorial ANOVA with all main effects and interactions, and the effect size eta-squared (η^2) were furtherly computed to investigate the impacts of the designed factors on RB. Eta-squared values 0.01, 0.06, and 0.14 were used to indicate small, moderate, and large effect sizes respectively (Cohen, 1998).

Results

M-PS Estimation

The distributions of true and estimated M-PSs based on four M-PS estimation methods over all replications were shown in Figure 2. The true M-PS distribution for treated and controlled groups showed that around one half of treated and controlled individuals had overlapped M-PSs. It is similar to the M-PS distribution simulated in Bellara's study (2013) from which the data generation setting of this study was borrowed. As shown in Figure 2, the M-CART produced nearly identical M-PS distribution to the true M-PS distribution for both treatment and control groups. The M-PS distribution based on S- CART was closer to the true M-PS distribution than the M-PS distributions based on the M-logit and S-logit.

Standardized Mean Difference

In every data set, matching and covariate-adjustment using PS resulted in two standardized mean difference values for each of 8 selected covariates (i.e., U_1, U_2, …, U_6, V_1, and V_2). For each selected covariates, *M* and *SD* were aggregated over replications by M-PS estimation methods and design factors in Table 8 to Table 15. All estimated Δ based on matching conditioning in Table 14 and Table 15 were 0, because matching was conducted within clusters and thus cluster level covariates V_1 and V_2 were the same between two groups.

Conditioning approach. Under matching conditioning, notable in these tables was all estimated Δ smaller than 0.25, indicating the overall acceptable performance of covariate balances for all M-PS estimation methods across design factors. The M-logit yielded the smallest Δ scores, followed by the S-logit and M-CART, and the S-CART produced the largest Δ scores.

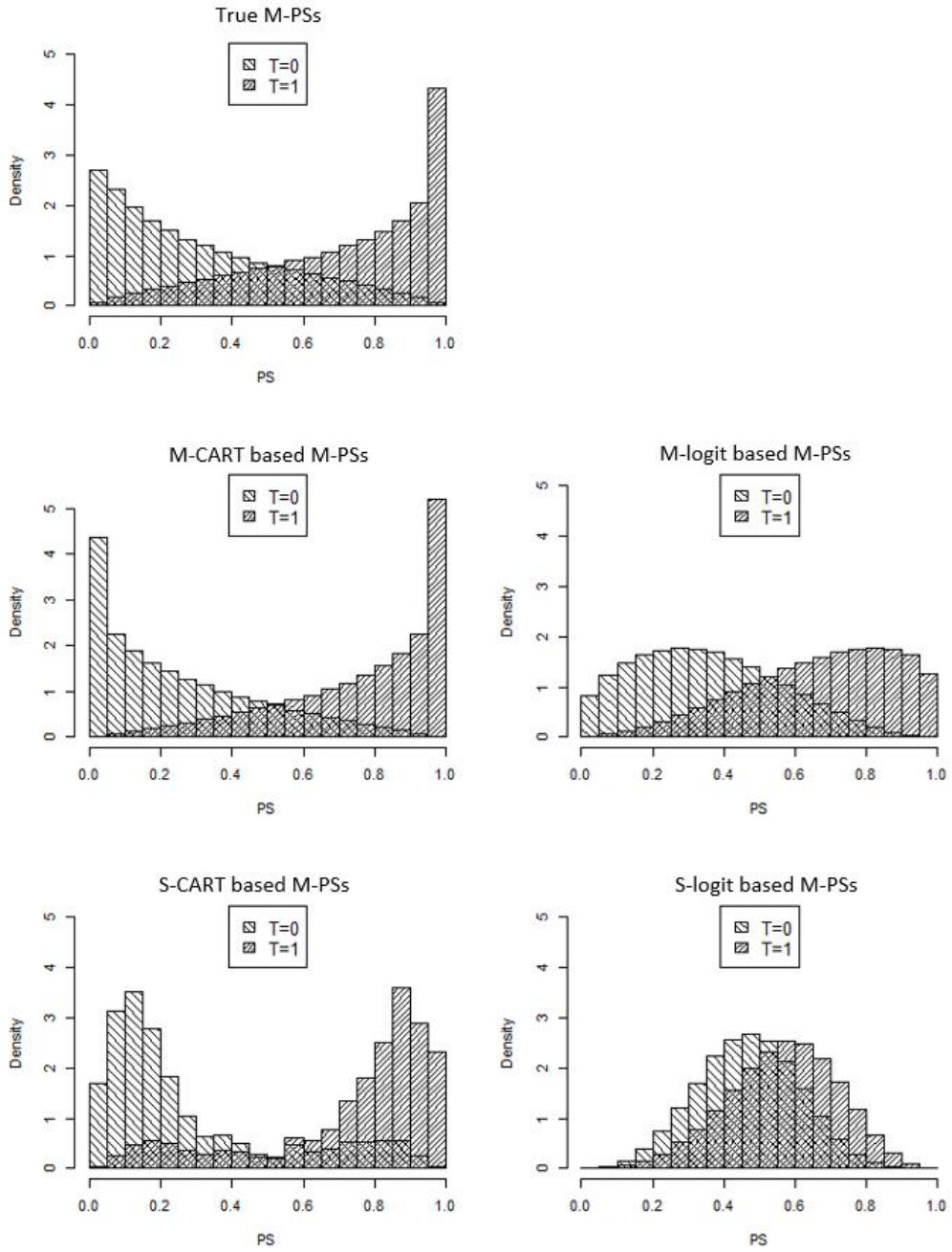

Figure 2. Distributions of M-PSs for treated and controlled groups.

For example, for covariate U_1 when *ICC* = 0.1, $\Delta_{U1|\,M-logit}$ = 0.0331 < $\Delta_{U1|\,S-logit}$ = 0.0809 < $\Delta_{U1|\,M-CART}$ = 0.1111 < $\Delta_{U1|\,S-CART}$ = 0.1354 (Table 8). Unlike the results based on the matching, covariate-adjustment using PS performed better under the S-logit and M-CART with Δ values smaller than 0.25 for all covariates. While the M-logit and S-CART had a few Δ values larger than the cut-off value (bold numbers in Table 8 to Table 15). These findings were consistent across various *ICC*s, degrees of nonlinearity and interaction, and sample sizes.

ICC. Larger *ICC*s resulted in higher Δ values for all estimation methods across all 6 individual level covariates. For instance, when using matching conditioning, $\Delta_{U2|\,M-CART}$ increased from 0.0670 to 0.0722 when the *ICC* increased from 0.1 to 0.3 (Table 9). The increase of Δ was strengthen when using covariate-adjustment adjustment using PS conditioning approach.

Sample size. Increasing number of clusters and/or cluster size resulted in smaller Δ values generally. For example, when covariate-adjustment using PS conditioning was used, $\Delta_{U3\,|M-CART}$ = 0.0568 when *Nc*= 15 and *Ns*=20 > $\Delta_{U3\,|M-CART}$ = 0.0343 when *Nc*= 30 and *Ns*=20; $\Delta_{U3\,|M-CART}$ = 0.0568 when *Nc*= 15 and *Ns*=20 > $\Delta_{U3\,|M-CART}$ = 0.0087 when *Nc*= 15 and *Ns*=50 (Table 10). This pattern was found in all conditioning methods and estimation methods for individual level covariates.

Degrees of nonlinearity and interaction. Larger degrees of nonlinearity and interaction were associated with higher Δ values. For example, when matching conditioning was used, $\Delta_{U2\,|M-CART}$ increased from 0.0524 to 0.0825 as the degrees of nonlinearity and interaction

increased (Table 9). This result was found for all estimation methods and conditioning methods across all individual level covariates.

Table 8
Standardized Mean Difference Values for U_1

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.1111	0.1063	0.0331	0.0697	0.1354	0.1130	0.0809	0.0700
0.3	0.1113	0.1102	0.0434	0.0846	0.1439	0.1254	0.0935	0.0780
Nonlinearity and Interaction								
0 - 0	0.0520	0.1102	0.0175	0.0802	0.0546	0.1130	0.0372	0.0788
0.2 - 0.4	0.1109	0.1159	0.0335	0.0740	0.1468	0.1195	0.1005	0.0748
0.4 - 0.6	0.1709	0.1138	0.0637	0.0772	0.2175	0.1250	0.1239	0.0685
Sample Size								
15-20	0.1312	0.1167	0.0543	0.1019	0.1664	0.1298	0.1107	0.1056
15-50	0.0857	0.1177	0.0212	0.0792	0.1210	0.1124	0.0628	0.0783
30-20	0.0938	0.1086	0.0326	0.0732	0.1385	0.1408	0.0706	0.0688
30-50	0.0758	0.0801	0.0148	0.0542	0.1127	0.0822	0.0588	0.0433
Covariate-adjustment using PS								
ICC								
0.1	0.0168	0.0923	0.1400	0.1490	0.0869	0.2198	0.0042	0.0120
0.3	0.0318	0.2600	0.4430	0.2317	0.3406	0.3821	0.0414	0.0594
Nonlinearity and Interaction								
0 - 0	0.0071	0.2654	0.2360	0.3274	0.1389	0.3555	0.0036	0.0214
0.2 - 0.4	0.0097	0.0953	**0.2690**	0.0968	0.1972	0.1792	0.0144	0.0273
0.4 - 0.6	0.0160	0.1676	**0.4190**	0.1469	**0.2952**	0.3183	0.0504	0.0586
Sample Size								
15-20	0.0501	0.1440	**0.4199**	0.2495	**0.3648**	0.3545	0.0403	0.0704
15-50	0.0356	0.0551	0.0737	0.0458	0.0599	0.0919	0.0017	0.0072
30-20	0.0495	0.1305	**0.4452**	0.2679	**0.2930**	0.3062	0.0451	0.0533
30-50	0.0120	0.1750	0.2273	0.1982	0.0574	0.2514	0.0041	0.0121

Table 9
Standardized Mean Difference Values for U_2

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.0670	0.1074	0.0263	0.0646	0.0817	0.1049	0.0469	0.0635
0.3	0.0722	0.1169	0.0390	0.0751	0.0888	0.1185	0.0549	0.0737
Nonlinearity and Interaction								
0 - 0	0.0524	0.1205	0.0139	0.0779	0.0576	0.1088	0.0382	0.0758
0.2 - 0.4	0.0719	0.1063	0.0351	0.0683	0.0814	0.1068	0.0586	0.0659
0.4 - 0.6	0.0825	0.1097	0.0438	0.0633	0.0927	0.1195	0.0560	0.0641
Sample Size								
15-20	0.0852	0.1484	0.0363	0.0951	0.1392	0.1431	0.0776	0.0792
15-50	0.0663	0.1073	0.0225	0.0704	0.0895	0.0991	0.0405	0.0684
30-20	0.0704	0.1133	0.0319	0.0653	0.0914	0.1173	0.0592	0.0598
30-50	0.0565	0.0797	0.0200	0.0485	0.0690	0.0872	0.0364	0.0469
Covariate-adjustment using PS								
ICC								
0.1	0.0267	0.0992	0.1289	0.1466	0.0692	0.1695	0.0047	0.0124
0.3	0.1199	0.1626	0.2198	0.1311	0.1537	0.1523	0.0216	0.0405
Nonlinearity and Interaction								
0 - 0	0.0541	0.1710	0.1012	0.1366	0.0605	0.1910	0.0054	0.0273
0.2 - 0.4	0.0738	0.1050	0.1071	0.0873	0.0868	0.1623	0.0078	0.0133
0.4 - 0.6	0.1019	0.1666	0.2308	0.1126	0.1272	0.1794	0.0262	0.0389
Sample Size								
15-20	0.1479	0.2558	**0.3488**	0.2529	0.1476	0.2044	0.0220	0.0356
15-50	0.0468	0.0554	0.1464	0.0470	0.0532	0.0872	0.0095	0.0098
30-20	0.1415	0.2390	0.2215	0.2541	0.1551	0.2068	0.0218	0.0586
30-50	0.0206	0.1734	0.1107	0.2013	0.0281	0.2452	0.0092	0.0019

Table 10
Standardized Mean Difference Values for U_3

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.0414	0.1010	0.0107	0.0505	0.0553	0.1241	0.0601	0.0605
0.3	0.0434	0.1157	0.0124	0.0660	0.0592	0.1267	0.0758	0.0789
Nonlinearity and Interaction								
0 - 0	0.0420	0.1073	0.0061	0.0660	0.0485	0.1006	0.0129	0.0611
0.2 - 0.4	0.0466	0.1083	0.0129	0.0710	0.0520	0.1145	0.0250	0.0688
0.4 - 0.6	0.0476	0.1095	0.0155	0.0850	0.0613	0.1160	0.0460	0.0747
Sample Size								
15-20	0.0634	0.1376	0.0246	0.1039	0.0874	0.1580	0.0364	0.0999
15-50	0.0408	0.1046	0.0008	0.0677	0.0447	0.0994	0.0170	0.067
30-20	0.0554	0.1118	0.0208	0.0777	0.0652	0.1048	0.0212	0.0623
30-50	0.0322	0.0794	0.0004	0.0467	0.0357	0.0793	0.0124	0.0437
Covariate-adjustment using PS								
ICC								
0.1	0.0283	0.1143	0.1127	0.1520	0.0533	0.2165	0.0025	0.0097
0.3	0.0606	0.3012	0.1753	0.1990	0.1338	0.5029	0.0065	0.0168
Nonlinearity and Interaction								
0 - 0	0.0530	0.1691	0.0990	0.2640	0.0529	0.3734	0.0044	0.0219
0.2 - 0.4	0.0714	0.1348	0.1160	0.0616	0.0851	0.1921	0.0020	0.0059
0.4 - 0.6	0.0966	0.2192	**0.2869**	0.1009	0.1126	0.3136	0.0070	0.0120
Sample Size								
15-20	0.0568	0.2104	0.2251	0.1923	0.1983	0.3668	0.0064	0.0183
15-50	0.0087	0.0576	0.0833	0.0409	0.0631	0.2908	0.0019	0.0035
30-20	0.0343	0.2604	0.1544	0.2578	0.1453	0.2833	0.0091	0.0215
30-50	0.0079	0.1724	0.0631	0.2109	0.0343	0.0980	0.0006	0.0097

Table 11
Standardized Mean Difference Values for U_4

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.0413	0.1132	0.0135	0.0718	0.0499	0.1022	0.0223	0.0614
0.3	0.0484	0.1101	0.0157	0.0787	0.0571	0.1031	0.0263	0.0713
Nonlinearity and Interaction								
0 - 0	0.0188	0.1074	0.0052	0.0701	0.0289	0.1046	0.0077	0.0657
0.2 - 0.4	0.0432	0.1176	0.0110	0.0694	0.0493	0.0983	0.0240	0.0652
0.4 - 0.6	0.0875	0.1101	0.0126	0.0861	0.0923	0.1052	0.0413	0.0682
Sample Size								
15-20	0.0649	0.1476	0.0172	0.1059	0.0698	0.1364	0.0415	0.0893
15-50	0.0442	0.1025	0.0125	0.0708	0.0479	0.0933	0.0220	0.0610
30-20	0.0552	0.1157	0.0161	0.0708	0.0667	0.1992	0.0324	0.0708
30-50	0.0361	0.0808	0.0103	0.0533	0.0416	0.0718	0.0213	0.0444
Covariate-adjustment using PS								
ICC								
0.1	0.0619	0.2139	0.1159	0.1229	0.0650	0.1899	0.0024	0.0085
0.3	0.1485	0.2417	0.2231	0.1500	0.1794	0.4404	0.0081	0.0171
Nonlinearity and Interaction								
0 - 0	0.0327	0.1564	0.0596	0.0567	0.0318	0.1596	0.0025	0.0058
0.2 - 0.4	0.0729	0.2793	0.1038	0.0841	0.0355	0.2757	0.0084	0.0117
0.4 - 0.6	0.2146	0.5478	**0.3451**	0.2685	**0.2947**	0.3102	0.0048	0.0208
Sample Size								
15-20	0.2232	0.4932	**0.2916**	0.1772	0.2490	0.4802	0.0102	0.0201
15-50	0.0778	0.2644	0.1563	0.1792	0.0893	0.2524	0.0022	0.0108
30-20	0.1093	0.4661	0.1978	0.1564	0.1096	0.4444	0.0080	0.0169
30-50	0.0167	0.0876	0.0323	0.0329	0.0349	0.0835	0.0005	0.0033

Table 12
Standardized Mean Difference Values for U_5

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.0604	0.1385	0.0098	0.0838	0.0654	0.1278	0.0146	0.0676
0.3	0.0603	0.1363	0.0124	0.0840	0.0688	0.1324	0.0140	0.0762
Nonlinearity and Interaction								
0 - 0	0.0513	0.1304	0.0081	0.0806	0.0606	0.1452	0.0108	0.0728
0.2 - 0.4	0.0630	0.1325	0.0098	0.0845	0.0619	0.1315	0.0150	0.0701
0.4 - 0.6	0.0675	0.1274	0.0153	0.0866	0.0695	0.1356	0.0194	0.0728
Sample Size								
15-20	0.0642	0.1859	0.0125	0.1054	0.0684	0.1803	0.0171	0.0994
15-50	0.0539	0.1379	0.0106	0.0872	0.0630	0.1177	0.0158	0.0694
30-20	0.0616	0.1232	0.0110	0.0800	0.0643	0.1274	0.0184	0.0696
30-50	0.0519	0.1026	0.0101	0.0630	0.0626	0.0850	0.0154	0.0491
Covariate-adjustment using PS								
ICC								
0.1	0.0280	0.0538	0.0500	0.0731	0.1555	0.2631	0.0012	0.0058
0.3	0.0751	0.1145	0.1075	0.1355	**0.3000**	0.3190	0.0036	0.0149
Nonlinearity and Interaction								
0 - 0	0.0230	0.1263	0.0274	0.1854	0.1005	0.2240	0.0009	0.0155
0.2 - 0.4	0.0465	0.0501	0.0562	0.0471	**0.2977**	0.2603	0.0018	0.0055
0.4 - 0.6	0.0852	0.0760	0.1527	0.0805	**0.3352**	0.2389	0.0047	0.0100
Sample Size								
15-20	0.0847	0.0904	0.1538	0.142	**0.3984**	0.2233	0.0059	0.0152
15-50	0.0390	0.0727	0.0829	0.1284	0.1549	0.3046	0.0010	0.0104
30-20	0.0710	0.1373	0.0642	0.1144	**0.3200**	0.2193	0.0023	0.0125
30-50	0.0115	0.0361	0.0140	0.0326	0.0379	0.1171	0.0004	0.0030

Table 13
Standardized Mean Difference Values for U_6

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.1077	0.1111	0.0143	0.0841	0.1203	0.1184	0.0250	0.0635
0.3	0.1171	0.1190	0.0189	0.0823	0.1281	0.1274	0.0396	0.0749
Nonlinearity and Interaction								
0 - 0	0.0794	0.1116	0.0128	0.0863	0.0895	0.1222	0.0290	0.0682
0.2 - 0.4	0.1100	0.1184	0.0165	0.0821	0.1283	0.1186	0.0337	0.0687
0.4 - 0.6	0.1477	0.1152	0.0204	0.0812	0.1747	0.1578	0.0342	0.0707
Sample Size								
15-20	0.1206	0.1543	0.0312	0.1193	0.1276	0.1635	0.0402	0.0958
15-50	0.1056	0.1177	0.0100	0.0778	0.1229	0.1259	0.0139	0.0642
30-20	0.1196	0.1041	0.0182	0.0770	0.1275	0.1122	0.0419	0.0714
30-50	0.1037	0.0843	0.0069	0.0587	0.1127	0.0899	0.0133	0.0454
Covariate-adjustment using PS								
ICC								
0.1	0.1357	0.2622	0.1261	0.1099	0.1574	0.2327	0.0028	0.0078
0.3	**0.2870**	0.2980	**0.3232**	0.1605	**0.3843**	0.5034	0.0141	0.0240
Nonlinearity and Interaction								
0 - 0	0.1247	0.2436	0.2014	0.2265	**0.2975**	0.2862	0.0035	0.0184
0.2 - 0.4	0.1699	0.0631	**0.2528**	0.1875	**0.2999**	0.1958	0.0047	0.0095
0.4 - 0.6	0.2293	0.0989	**0.3798**	0.2263	**0.4152**	0.2222	0.0171	0.0198
Sample Size								
15-20	0.2017	0.2700	**0.3938**	0.1887	**0.3563**	0.2548	0.0174	0.0271
15-50	0.0916	0.2115	0.1882	0.1517	0.1186	0.2382	0.0017	0.0081
30-20	0.2149	0.2438	**0.2687**	0.1644	**0.2501**	0.2818	0.0138	0.0247
30-50	0.0473	0.0950	0.0479	0.0360	0.0585	0.0975	0.0009	0.0038

Table 14
Standardized Mean Difference Values for V_1

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Nonlinearity and Interaction								
0 - 0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.2 - 0.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.4 - 0.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sample Size								
15-20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15-50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
30-20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
30-50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Covariate-adjustment using PS								
ICC								
0.1	0.0615	0.1779	0.2200	0.2015	0.0513	0.1702	0.0054	0.0159
0.3	0.1707	0.2477	0.3002	0.2085	0.1574	0.2153	0.0280	0.0607
Nonlinearity and Interaction								
0 - 0	0.0366	0.0882	0.1355	0.0961	0.0217	0.1862	0.0062	0.0162
0.2 - 0.4	0.0709	0.1436	**0.2770**	0.1714	0.0320	0.2088	0.0112	0.0498
0.4 - 0.6	0.1408	0.1066	**0.3678**	0.1974	0.1092	0.2832	0.0328	0.0490
Sample Size								
15-20	0.1783	0.1906	**0.3749**	0.2332	0.0735	**0.3098**	0.0393	0.0818
15-50	0.0605	0.1593	0.243	0.2777	0.0539	0.2131	0.0206	0.0206
30-20	0.1051	0.1375	**0.2616**	0.2555	0.0631	**0.2850**	0.0157	0.0446
30-50	0.0206	0.0637	0.0610	0.0535	0.0269	0.0931	0.0012	0.0063

Table 15
Standardized Mean Difference Values for V_2

	M-CART		M-logit		S-CART		S-logit	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Matching								
ICC								
0.1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Nonlinearity and Interaction								
0 - 0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.2 - 0.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.4 - 0.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sample Size								
15-20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15-50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
30-20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
30-50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Covariate-adjustment using PS								
ICC								
0.1	0.0377	0.1726	0.1060	0.1568	0.1015	0.2363	0.0052	0.0274
0.3	0.0474	0.1989	0.1854	0.2857	0.1569	0.2118	0.0164	0.0676
Nonlinearity and Interaction								
0 - 0	0.0877	0.2512	**0.2570**	0.2050	0.1713	0.2109	0.0159	0.0783
0.2 - 0.4	0.0236	0.0798	0.1369	0.0983	0.1577	0.2099	0.0130	0.0430
0.4 - 0.6	0.0163	0.1263	0.0320	0.1603	0.0485	0.2014	0.0035	0.0211
Sample Size								
15-20	0.0600	0.2447	**0.2827**	0.2317	**0.2594**	0.6546	0.0203	0.0672
15-50	0.0432	0.2503	0.0972	0.2662	0.1269	0.8251	0.0097	0.0658
30-20	0.0537	0.1865	0.1755	0.2347	0.0893	0.3079	0.0117	0.0461
30-50	0.0133	0.0616	0.0273	0.0523	0.0411	0.1085	0.0016	0.0108

Treatment Effect

The estimated relative bias by two conditioning methods and three design factors across four M-PS estimation methods are listed in Table 16. ANOVA results showed that all main effects, and the interaction effect between degrees of nonlinearity and PS estimation methods had remarkable effect sizes.

Estimation method. The impact of PS estimation methods on RB was large with η^2= 0.1644 [F(3, 57409) = 487.4455, p < 0.001]. On average, M-CART yielded the most accurate ATE estimations with smallest RBs under both PS conditioning approaches. For example, when using matching approach, $RB_{M\text{-}CART} = 0.1852 < RB_{M\text{-}logit} = 0.1962 < RB_{S\text{-}CART} = 0.2440 < RB_{S\text{-}logit} = 0.2843$ on average.

***Conditioning approach*.** The conditioning approach had a small effect size with η^2= 0.0282 [F(3, 57409) = 65.8610, p < 0.001]. Averaging across simulation conditions, covariate-adjustment using PS yielded slightly lower RB than using matching approach. For instance, when using M-CART for data simulated, $RB_{M\text{-}CART}$ estimated based on covariate-adjustment using PS conditioning equaled to 0.1169 on average which was smaller than $RB_{M\text{-}CART}$ estimated based on matching condition method 0.1852; when using M-logit estimation method, mean $RB_{M\text{-}logit} = 0.1688$ based on covariate-adjustment using PS conditioning was smaller than mean $RB_{M\text{-}logit} = 0.1962$ based on matching conditioning.

***ICC*.** Increasing *ICC*s yielded slightly larger relative biases for all conditioning and estimation methods, which was found to have small effect size with η^2= 0.0094 [F(1, 57409) = 47.8953, p < 0.001]. For instance, $RB_{M\text{-}logit}$ increased from 0.1876 to 0.2040 when the *ICC* went up.

Table 16
Treatment Effects for Four Estimation Methods by Conditioning Methods and Simulated Conditions

	M-CART Relative Bias	M-logit Relative Bias	S-CART Relative Bias	S-logit Relative Bias
Matching				
ICC				
0.1	0.1801	0.1876	0.2350	0.2794
0.3	0.1903	0.2040	0.2530	0.2891
Nonlinearity and Interaction				
0 - 0	0.0909	0.0734	0.1516	0.1729
0.2 - 0.4	0.1686	0.1793	0.2329	0.2652
0.4 - 0.6	0.2962	0.3447	0.3474	0.4147
Sample Size				
15-20	0.2107	0.2115	0.2699	0.2988
15-50	0.1757	0.1912	0.2348	0.2897
30-20	0.2089	0.2202	0.2660	0.2869
30-50	0.1386	0.1602	0.2052	0.2616
Average	0.1852	0.1962	0.2440	0.2843
Covariate-adjustment using PS				
ICC				
0.1	0.1061	0.1559	0.2309	0.2996
0.3	0.1277	0.1805	0.2447	0.3213
Nonlinearity and Interaction				
0 - 0	0.0463	0.0280	0.1608	0.1926
0.2 - 0.4	0.1084	0.1565	0.2294	0.3008
0.4 - 0.6	0.1960	0.3301	0.3233	0.4380
Sample Size				
15-20	0.1443	0.1717	0.2649	0.3093
15-50	0.1095	0.1853	0.2280	0.3223
30-20	0.1345	0.1708	0.2489	0.3094
30-50	0.0792	0.1450	0.2095	0.3008
Average	0.1169	0.1688	0.2378	0.3105

***Degree of nonlinearity and interaction*.** Increasing degrees of nonlinearity and non-additivity caused larger RBs for four PS estimation methods in both conditioning methods (Table 16), which had a extremely large effect size with η^2= 0.3430 [F(2, 57409) = 1937.7111, p < 0.001]. Results also showed that the interaction between this factor and the estimation method factor had a small to moderate effect on RB values [F(6, 57409) = 83.7247, p < 0.001, η^2= 0.0465]. M-logits produced the smallest RBs (Table 16) when the M-PS estimation model was linear and additive. M-CART performed closely to M-logits in this condition. While, the two singe level estimation methods, especially S-logits, had larger RBs. For example, when using covariate-adjustment conditioning, $RB_{M\text{-}logit}$ = 0.0280 < $RB_{M\text{-}CART}$ = 0.0463 < $RB_{S\text{-}CART}$ = 0.1608 < $RB_{S\text{-}logit}$ = 0.1926. When having small nonlinearity and non-additivity (0.2- 0.4 condition), the two multilevel M-PS estimation methods still outperformed the corresponding single level M-PS estimation methods. However, M-CART started to perform better than M-logits with smaller RBs. S-logits, again, lead to the largest RBs. For example, when using matching conditioning, $RB_{M\text{-}CART}$ = 0.1686 < $RB_{M\text{-}logit}$ = 0.1793 < $RB_{S\text{-}CART}$ = 0.2329 < $RB_{S\text{-}logit}$ = 0.2652. When having large nonlinearity and non-additivity, M-CART had the smallest RBs, while S-logits produced the largest RBs. It should be noted that, in this case, S-CART performed almost identically to the M-logit with similar RBs (Table 16).

Sample size. For all conditioning and estimation methods, increasing *Nc* and/or *Ns* yielded lower RBs (Table 16), which had a small effect size with η^2= 0.0243 [F(3, 57409) = 58.4352, p < 0.001]. For example, when *Nc = 30* and covariate-adjustment conditioning was applied, $RB_{M\text{-}CART}$ decreased from 0.1345 to 0.0792 when *Ns* went up from 20 to 50; when controlling *Ns* as 50 and using matching conditioning, $RB_{M\text{-}logit}$ decreased from 0.1912 to 0.1602 when *Ns* went up from 15 to 30.

Discussion

This study contributed to existing literature by using a multilevel CART algorithm for estimating PSs when conducting PSA in multi-site non-RCTs. In this study, I compared the performance of using the M-CART with the M-logit, S-CART and S-logit to estimate PSs across different PS conditioning approaches (i.e., matching and covariate-adjustment using PS) and various simulated sample characteristics (i.e., magnitudes of clustering effect, covariate relationships to treatment assignment, and sample sizes).

Overall, M-CARTs achieved good covariate balances with all standardized mean difference scores less than the cut-off value 0.25 across all conditions, even though M-CARTs did not always yield the smallest standardized mean differences. In addition, M-CART was comparable to or better than the other three M-PS estimation methods in term of estimating ATE across all conditions. The performances of M-logit, S-CART and S-logit on balancing pretreatment covariates differences and estimating ATE were worse and heavily relied on PS conditioning approaches and simulated sample characteristics.

With matching conditioning approach, logistic regression estimation methods, especially the M-logit, produced smaller standardized mean differences and achieve better covariate balances than two CART algorithms. While, S-logits and M-CARTs performed noticeably better than M-logits and S-CARTs when covariate-adjustment using PS conditioning approach was employed. This finding was partially supported by Bellara's study (2013) in which researchers claimed that M-logits performed better on balancing covariate differences than S-logits under matching conditioning, but S-logits were found to have appreciably better covariate balances than M-logits when using covariate-adjustment using PS conditioning. Even though mixed performance of using M-CART for covariate balance according to different conditioning

approaches, M-CART produced more accurate treatment effect estimations than the other three estimation methods on average under both conditioning approaches.

Increasing *ICC* values yield slightly worse balance and less accurate ATE estimation for all M-PS estimation methods. Within each *ICC* condition, single level models (S-CART and S-logit) performed worse in comparison to multilevel models (M-CART and M-logit) in terms of balancing covariates and estimating ATE, because overlooking clustering effects when modeling multilevel data caused estimation biases (Gelman, 2006; Hox, 1998;). This finding was proved in previous propensity score analysis studies (e.g., Thoemmes, 2009).

Balance of covariates and ATE estimation accuracy decreased as the model became more complex and increasingly dominated by nonlinear and non-additive terms for all four M-PS estimation methods, which was consistent with previous studies (e.g., Bellara, 2013; Lee, Lessler, & Stuart, 2010). This is intuitive, because the estimation models excluded nonlinear and interaction terms which became less fit for data as degrees of nonlinearity and interaction increased.

Additionally, results showed that four estimation methods performed differently on estimating ATE according to different degrees of nonlinearity and interaction. When the relationship between covariates and treatment assignment is linear and additive, M-logits, not surprisingly, produced the most accurate ATE, as it was true models which should obviously provide the least biased ATE. However, M-CART functioned almost identically to M-logit and outperformed S-CART and S-logit in this case, meaning that M-CART could recover the linear and additive model well in terms of estimating ATE. As degrees of nonlinearity and non-additivity increased, the benefit of using CART algorithms was enlarged. M-CART started to yield the most accurate ATE estimations. M-CART worked the best was highly because it had

the ability to automatically capture the non-linear and non-addictive terms while controlling for the clustering effect in the simulated multilevel data. Even though the performance of using M-CART for M-PSA has not been explored before, these findings are also indirectly proved in existing studies. Previous researchers demonstrated that multilevel linear models produced less root-mean squared errors than multilevel CART algorithms for predicting multilevel data with continuous data when the true relationship between outcomes and covariates were linear and additive (Lingle, 2009; Sela & Simonoff, 2012; Thoemmes, 2009). When the relationship was nonlinear and non-additive, researchers found that CART algorithms yield less prediction errors than linear regression models for both multilevel (e.g., Sela & Simonoff, 2012; Hajjem, Bellavance & Larocque , 2011) and single level settings (e.g., Lee, Lessler, & Stuart, 2010; Steinberg & Colla, 2009).

Increasing number of clusters and cluster sizes assisted to improve balances and ATE estimation accuracy for all M-PS estimation methods using either matching or covariate-adjustment using PS conditioning, because larger sample sizes have been demonstrated to associated with less model prediction errors. This finding agreed with many previous research (e.g., Bellara, 2013; Lee, Lessler, & Stuart, 2010; Li, Zaslavsky, & Landrum, 2013; Lingle, 2009; Thoemmes, 2009).

To sum up, M-logit is recommended only when the linear and additive relationship between covariates and treatment assignment was found and matching method was applied, otherwise using the proposed M-CART for estimating PSs in multi-site non-RCTs is always a better option when conducting M-PSA. However, based on the design of this study, there were several generalizability limitations to consider. First, the treated and controlled groups were balanced on group size. The performance of using M-CART when having unbalanced groups

should be furtherly examined. Second, only two experimental conditions (i.e., treatment and control groups) were simulated in this study. Future research can also extend this by having multiclass treatment assignment in multi-site non-RCTs. Third, only matching and covariate-adjustment using PS were applied as conditioning approaches. The usefulness of M-CART under other commonly used conditioning approaches such as stratification and inverse probability of weighting had not been testified.

CHAPTER IV

CONCLUSIONS

This dissertation proposed and evaluated a new multilevel CART algorithm used for modeling multilevel data with binary outcomes and estimating propensity scores in multi-site non-RCTs. The new multilevel CART algorithm combines multilevel logistic regression and single-level CART using the EM algorithm. It overcomes the disadvantages of single-level CART and multilevel logistic regressions alone and inherits the advantages of both methods. Specifically, the new multilevel CART controls for clustering effects, allows inclusion of covariates at all levels, depends on no model assumptions, and captures interaction and nonlinearity in an automatic way. Study one showed that M-CART lead to higher prediction accuracy than the M-logit, S-CART, and S-logit in terms of classification accuracy, sensitivity, specificity and Klecka's tau. This benefit of applying the proposed M-CART for modeling multilevel data with binary outcomes was consistent across different data situations including different levels of clustering effects and sample sizes, and strengthened when the relationships between outcomes and predictors were nonlinear and non-additive.

In addition to making predictions, the new M-CART algorithm can also applied in propensity score analysis for multi-site non-RCTs. There are two keys for unbiased propensity score estimations when conducting M-PSA. First, the model used for estimating propensity scores should have the ability of controlling for clustering effects. Second, as a requirement of strongly ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983), all confounding covariates should be included as predictors for estimating propensity scores. It has been suggested in the literature that all observed confounding variables including nonlinear terms

and interactions of these variables should be included in the model. Thus, it is favorable to use the proposed M-CART algorithm that can account for clustering effects and capture all interaction and nonlinearity among the observed covariates. Study Two results indicated that M-CART was more stable than the M-logit, S-CART and S-logit on achieving pre-treatment covariate balance and always yield reasonable covariate balance over all conditions. Simulation results furtherly showed that, regardless of the PS conditioning approaches, M-CART yielded the least relative biases in the ATE estimates across all simulated conditions. Even when the relationship between outcomes and treatment assignment was linear and additive, M-CART still outperformed other estimation methods when covariate-adjustment using PS conditioning was applied.

REFERENCES

Abdolell, M., LeBlanc, M., Stephens, D., & Harrison, V. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in Medicine*, *21*(22), 3395-3409.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General), 149*(1), 1-43.

Altman, D. G. (2005). Covariate imbalance, adjustment for. *Encyclopedia of Biostatistics*. 1-6.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*(4), 1770-1780.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083-3107.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424.

Austin, P. C., Grootendorst, P., Normand, S. L. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine, 26*(4), 754-768.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*(12), 2084-2106.

Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112-118.

Bellara, A. P. (2013). *Effectiveness of propensity score methods in a multilevel framework: A Monte Carlo study*. (Doctoral dissertation). University of South Florida, Tampa, FL.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199-231.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, Florida: CRC press.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source code for Biology and Medicine*, *3*(1), 17.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, D. R. (1985). *Planning of experiments*. New York, NY: John Wiley & Sons.

D'Agostino Jr, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265-2281.

De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, *83*(4), 1105-1117.

Deconinck, E., Hancock, T., Coomans, D., Massart, D. L., & 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*(1), 91-103.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*(4), 1231-1236.

Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. *Psychological Methods, 17*(2), 153.

Emsley, R., Lunt, M., Pickles, A., & Dunn, G. (2008). Implementing double-robust estimators of causal effects. *Stata Journal, 8*(3), 334-353.

Eo, S. H., & Cho, H. (2014). Tree-structured mixed-effects regression modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, *23*(3), 740-760.

Feelders, A. (1999). Handling missing data in trees: Surrogate splits or statistical imputation?. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 329-334). Berlin, Heidelberg: Springer.

Finch, W. H. (2014). Recursive partitioning methods for prediction in education: Application to the identification of students at-risk for academic failure. *International Journal of Quantitative Research in Education*, *2*(2), 133-152.

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling*, *18*(2), 229-252.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432-435.

Goldstein, H. (2011). *Multilevel statistical models (Vol. 922)*. New York, NY: John Wiley & Sons.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, *26*(1), 441-462.

Hade, E. M., & Lu, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine*, *33*(1), 74-87.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters, 81*(4), 451-459.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation, 84*(6), 1313-1328.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The elements of statistical learning: Data mining, inference, and prediction,* New York: Springer.

Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. J. (2015). Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and Aging*, *30*(4), 911.

Hedges, L., & Hedberg, E.C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention* (Doctoral dissertation). University of Michigan, Ann Arbor, MI.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205-224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901-910.

Hox, J. J. (1998). Multilevel modeling: when and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, Data Analysis, and Data Highways* (pp. 147-154). Berlin: Springer.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications.* New York, NY: Routledge.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling, 20*(2), 265-282.

Jorgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika, 70*(1), 19-28.

Kearns, M., & Mansour, Y. (1999). On the boosting ability of top–down decision tree learning algorithms. *Journal of Computer and System Sciences, 58*(1), 109-128.

Kelcey, B. M. (2009). *Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings* (Doctoral dissertation). The University of Michigan, Ann Arbor, MI.

Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection processes vary across schools. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. Retrieved from http://files.eric.ed.gov/fulltext/ED495846.pdf

Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.

Kleinbaum, D. G., Klein, M. (2002). *Logistic regression. A self-learning text*. New York, NY: Springer.

Kwok, O., Luo, W., & West, S. G. (2010). Using modification indexes to detect turning points in longitudinal data: A Monte Carlo study. *Structural Equation Modeling*, *17*(2), 216-240.

Lanza, S. T., Moore, J. E., & Butera, N. M. (2013). Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American journal of community psychology*, *52*(3), 380-392.

Larsen, D. R., & Speckman, P. L. (2004). Multivariate regression trees for analysis of abundance data. *Biometrics*, *60*(2), 543-549.

Lee, S. K. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis*, *49*(4), 1105-1119.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337-346.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research*, *50*(3), 265-284.

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, *26*(3), 172-181.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373-3387.

Lingle, J. A. (2009). *Evaluating the performance of propensity scores to address selection bias in a multilevel context: A Monte Carlo simulation study and application using a national dataset* (Doctoral dissertation). Georgia State University, Atlanta, GA.

Loh, W. Y., & Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, *7*(1), 495-522.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine, 23*(19), 2937-2960.

Luo, W., Cappaert, K. J., & Ning, L. (2015). Modelling partially cross-classified multilevel data. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 342-362.

Maas, C. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86-92.

Martin, D. P. (2015). *Efficiently exploring multilevel data with recursive partitioning* (Doctoral dissertation). University of Virginia, Charlottesville, VA.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388-3414.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*. New York, NY: John Wiley & Sons.

McMahon, J. M., Pouget, E. R., & Tortu, S. (2006). A guide for multilevel modeling of dyadic data with binary outcomes using SAS PROC NLMIXED. *Computational Statistics & Data Analysis*, *50*(12), 3663-3680.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology Clinics of North America*, *14*(4), 745-760.

Millimet, D. L., & Tchernis, R. (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business & Economic Statistics*, *27*(3), 397-415.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*(1), 1-10.

Peugh, J. L., & Enders, C. K. (2010). Specification searches in multilevel structural equation modeling: A Monte Carlo investigation. *Structural Equation Modeling*, *17*(1), 42-65.

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2014). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, *181*(2), 108-119.

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X. H., & Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, *9*(2), 93-101.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*, *21*(19), 2917-2930.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-machine Studies*, *27*(3), 221-234.

R Core Team. (2016). R: *A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*(1), 141-157.

Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications, 29*(1), 65-74.

Rickles, J. (2012). *Using a two-stage propensity score matching strategy and multilevel modeling to estimate treatment effects in a multisite observational study* (Doctoral dissertation). University of California, Los Angeles, CA.

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics, 39*(6), 612-636.

Rodríguez, G. (2008). Multilevel generalized linear models. In *Handbook of multilevel analysis* (pp. 335-376). New York, NY: Springer.

Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165-192). New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational and Behavioral Statistics*, *11*(3), 207-224.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*(398), 387-394.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York, NY: McGraw-Hill Humanities.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34-58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*(371), 591-593.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(8), 757-763.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322-331.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*(3), 583-625.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, *87*(418), 407-418.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, *86*(2), 169-207.

Senn, S. J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, *8*(4), 467-475.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546-555.

Sideridis, G. D., Antoniou, F., & Padeliadu, S. (2008). Teacher biases in the identification of learning disabilities: An application of the logistic multilevel model. *Learning Disability Quarterly, 31*(4), 199-209.

Snijders, T. A. (2011). Multilevel analysis. In *International encyclopedia of statistical science* (pp. 879-882). New York, NY: Springer.

Snijders T., Bosker J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Beverly Hills, CA: Sage.

Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. In *The top ten algorithms in data mining* (pp. 179-201). New York, NY: CRC Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Su, Y. S., & Cortina, J. (2009). *What do we gain? Combining propensity score methods and multilevel modeling.* In *American Political Science Association Annual Meeting*, Toronto, Canada. Retrieved from https://ssrn.com/abstract=1450058

Thoemmes, F. J. (2009). *The use of propensity scores with clustered data: A simulation study*. Arizona State University, Tempe, Arizona.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118.

Thoemmes, F., & West, S.G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*(3), 514-543.

Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications* (Doctoral dissertation). Humboldt University, Berlin, German.

Verbyla, D. L. (1987). Classification trees: a new discrimination tool. *Canadian Journal of Forest Research, 17*(9), 1150-1152.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), 826-833.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*(391), 513-524.

Xiang, Y., & Tarasawa, B. (2015). Propensity score stratification using multilevel models to examine charter school achievement effects. *Journal of School Choice*, *9*(2), 179-196.

Yu, B. (2012). *Variable selection and adjustment in relation to propensity scores and prognostic scores: From single-level to multilevel data* (Doctoral dissertation). University of Toronto, Toronto, Canada.

Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, *93*(441), 180-193.

Zhang, H., & Ye, Y. (2008). A tree-based method for modeling a multivariate ordinal response. *Statistics and Its Interface, 1*(1), 169-178.