Journal of Data Science **10**(2012), 757-776

Tree-Structured Assessment of Causal Odds Ratio with Large Observational Study Data Sets

Joseph Kang[1]*, Xiaogang Su[2] and Kiang Liu[1]
[1]*Northwestern University and* [2]*University of Alabama at Birmingham*

Abstract: Observational studies of relatively large data can have potentially hidden heterogeneity with respect to causal effects and propensity scores–patterns of a putative cause being exposed to study subjects. This underlying heterogeneity can be crucial in causal inference for any observational studies because it is systematically generated and structured by covariates which influence the cause and/or its related outcomes. Addressing the causal inference problem in view of data structure, machine learning techniques such as tree analysis can be naturally necessitated. Kang, Su, Hitsman, Liu and Lloyd-Jones (2012) proposed Marginal Tree (MT) procedure to explore both the confounding and interacting effects of the covariates on causal inference. In this paper, we extend the MT method to the case of binary responses along with a clear exposition of its relationship with established causal odds ratio. We assess the causal effect of dieting on emotional distress using both a real data set from the Lalonde's National Supported Work Demonstration Analysis (NSW) and a simulated data set from the National Longitudinal Study of Adolescent Health (Add Health).

Key words: Binary potential outcomes, causal inference, maximum likelihood tree, propensity scores.

1. Introduction

1.1 Background

Observational studies of relatively large data can possess potential heterogeneities with respect to causal effects and/or propensity scores. Here the term "causal effects" is the same as the one under the framework of potential outcomes (or counterfactual outcomes) (Rubin, 2005; Robins, Hernán and Brumbac, 2000), and the propensity scores are probabilities of being exposed to a putative cause

*Corresponding author.

(typically a treatment) given study subjects' baseline information (Rosenbaum and Rubin, 1983).

The basic concept of causal inference starts with potential outcomes that are observable based on the way the treatment is assigned to study subjects. For example, as in Section 4.2 of this paper, suppose that one could choose to do either dieting or non-dieting. This person's certain outcome, which is related to dieting, can be observed under dieting or non-dieting. Such observable outcomes are called potential outcomes (Rubin, 2005). True causal effect of dieting vs. non-dieting is to compare these two potential outcomes for the same person (not among different people). However due to time and space limit, only one of these two potential outcomes can be actually observed. Such difficulty is addressed as the fundamental problem of causal inference (Paul, 1986). During the past decades, there have been many methods developed to address this problem in modern epidemiology, psychology, biostatistics, and sociology (Robins *et al.*, 2000; Schafer and Kang, 2008; Rubin, 2005; Morgan and Winship, 2007). Yet these causal inference methodologies all assessed causal effects without extensively exploring heterogenous subgroups in the entire study sample.

The underlying heterogeneous data structures of observational study data sets can be crucial in drawing causal inference because the heterogeneous structures are systematically generated by confounding covariates and are able to modify the degree and direction of the causal effects. Addressing the causal inference problem in view of data structure, machine learning techniques such as tree analysis can be necessitated. Health sciences have extensively benefitted from decision tree methods over the past decades due to their intuitive clarity of hierarchical presentation (Breiman, Friedman, Stone and Olshen, 1984; Zhang and Singer, 1999; Su, Wang and Fan, 2004). All tree methods adjust for the covariates by nonparametrically partitioning a whole data set into subgroups (also known as terminal nodes of a decision tree) of subjects who share a homogeneous response probability. That is, once the terminal nodes are optimally identified with resultant cost-complexity functions, the distribution of the outcome within each of the terminal nodes is no longer influenced by covariates. Adopting this beneficial idea of adjusting for influences of covariates, Kang *et al.* (2012) proposed a regression tree method called "Marginal Tree (MT)" that uses a likelihood-based decision rule to create terminal nodes where subjects share a homogeneous distribution for causal effect and its related constant propensity score.

The tree by Kang *et al.* (2012) was called "Marginal" because as the tree divides the data set, the joint distribution of the cause and its related outcome is marginalized over the space spanned by measured covariates so as to produce causal effects. The MT method used a continuous outcome and hence their tree method required a Gaussian normal distribution to build their tree.

As most tree methods (Breiman *et al.*, 1984; Zhang and Singer, 1999) distinguish differential properties of the continuous outcome and the categorical outcome, we extend the MT method to the case of binary outcome. There have been some recent advances in developing tree methods to adjust for the confounding covariates in estimating causal effects by (Su, Tsai, Wang, Nickerson and Li, 2009; Su, Zhou, Yan, Fan and Yang, 2008). To our knowledge, however, no current tree methodologies assemble the propensity scores into a binary outcome tree in a unified manner. Thus, in this paper, we jointly model the propensity scores and the conditional distribution of observed dichotomous outcome given treatment variable so that the causal effect can be correctly evaluated at resultant terminal nodes, within each of which confounding influences of covariates on both propensity score and the causal effect are to be removed. For this we entail the relationship between the MT model parameters in the case of the binary outcome and causal parameters under the potential outcome framework (Rubin, 2005).

The ultimate goal of the MT method is to identify subgroups that show differential effects which are assessed without any biases. With a posteriori determined subgroup-level effects using the MT method, population-level effects are to be more clearly explained and interventions can be tailored to subgroups of study subjects. One example of the usage of the MT method for the case of the binary outcome is to explore the existence of differential effects of dieting on adolescent girls' emotional distress using a simulated data set from the National Longitudinal Study of Adolescent Health (Add Health) (Harris, 2009). For this particular application, we search to answer the question: "Are there any subgroups of girls whose dieting causes differently emotional distress?"

2. Introduction to the Marginal Tree (MT) for the Binary Outcome

2.1 Model Specification

Consider a study cohort which contains i-th subject for $i = 1, \cdots, n$, each of which is associated with a vector containing covariates $\boldsymbol{x}_i$, a treatment T_i, and its related observed outcome Y_i. We assume that Y_i is a binary random variable and T_i is a dichotomous random variable with their respectively observed values y_i and t_i.

We specify a model that explains the relationship between Y, T and X as $P(Y_i, T_i|\boldsymbol{x}_i)$ for $i = 1, \cdots, n$ because X_i may simultaneously influence both Y_i and T_i. By Bayes' rule, $P(Y_i, T_i|\boldsymbol{x}_i)$ can be decomposed as $P(Y_i|T_i, \boldsymbol{x}_i)P(T_i|\boldsymbol{x}_i)$. The MT model (Kang *et al.*, 2012) seeks a way of splitting data so that the resultant K subsets will implicitly explain all possible influence of covariates on the relationship between Y and T. To explain this, let S_k denotes subset k (or terminal node k of the MT). Within S_k, it holds $P(Y_i, T_i|\boldsymbol{x}_i) = P(Y_i, T_i)$ and

hence it follows that

$$\prod_{i=1}^{n} P\left(Y_i, T_i|\boldsymbol{x}_i\right) = \prod_{k=1}^{K}\prod_{i\in S_k} P\left(Y_i, T_i\right) = \prod_{k=1}^{K}\prod_{i\in S_k} P\left(Y_i|T_i\right) P\left(T_i\right). \tag{1}$$

According to (1), subjects in S_k have a homogenous joint distribution $P(Y_i, T_i)$, (whence $P(Y_i|T_i)$ and $P(T_i)$) regardless of their covariate values. Namely, the association between Y_i and T_i in the subgroup S_k is no longer influenced by $\boldsymbol{x}_i$, implying that the association is marginalized over covariates $\boldsymbol{x}_i$. This is the reason why the MT method was named "Marginal". Note that k is simply a function of $\boldsymbol{x}$, but we use k instead of $k(\boldsymbol{x})$ for notational convenience. Nevertheless, finding such subsets as above is a formidable task without resorting to tree modeling (Zhang and Singer, 1999). A tree model offers a natural grouping method. By applying appropriate splitting rules, we put forward a tree method like the MT method that has the flexibility to integrate aspects concerning both differential causal effects from $P(Y_i|T_i)$ and heterogeneous propensity $P(T_i)$ simultaneously.

To clarify our inferential goal of estimating parameters of the MT for binary outcomes, suppose for those in the kth subset ($i \in S_k$) that $P(Y_i|T_i)$ can be modeled with parameters β_{0k} and β_{1k} such that

$$\begin{aligned} P(Y_i|T_i) &= P(Y_i = 1|T_i; \beta_k)^{y_i} P(Y_i = 0|T_i; \beta_k)^{1-y_i} \\ &= (\exp it(\beta_{0k} + \beta_{1k} t_i))^{y_i} \left(1 - \exp it(\beta_{0k} + \beta_{1k} t_i)\right)^{1-y_i}, \end{aligned} \tag{2}$$

where $\exp it(a)$ indicates $(1 + \exp(-a))^{-1}$. Then β_{1k} defines the measure of an effect for the kth subset, which will be proved as the causal effect (causal odds ratio) in the next section.

2.2 Relationship with Causal Parameters under Potential Outcome Framework

It may be worthy to note that β_{1k} of (2) is a causal parameter under the potential outcome framework (Rubin, 2005). Let $Y_i(t)$ denote a binary potential outcome that would be observed under the condition t (where $t = 0, 1$) and let $\phi_i(t)$ denote $P(Y_i(t) = 1)$. Then the causal effect is expressed as causal odds ratio $\theta = (\phi_i(1)/(1-\phi_i(1)))/(\phi_i(0)/(1-\phi_i(0)))$. In other words, $\phi_i(t) = \exp it(\alpha + t\cdot\theta)$.

Now consider $\phi_i(t) = \exp it(\alpha_k + t \cdot \theta_k)$ within the kth subgroup ($\forall i \in S_k$). In (2), $P(Y_i = 1|T_i = t_i)$ was denoted as $\exp it(\beta_{0k} + \beta_{1k} t_i)$. Because $P(Y_i|T_i, \boldsymbol{x}_i) = P(Y_i|T_i)$, it holds that $P(Y_i{=}1|T_i = t_i,\, \boldsymbol{x}_i) = \exp it(\beta_{0k} + \beta_{1k} t_i)$. Also because Y_i is expressed as $T_i Y_i(1) + (1 - T_i) Y_i(0)$, it is true that

$$P(T_i Y_i\left(1\right) + \left(1 - T_i\right) Y_i\left(0\right) = 1|T_i = t_i, \boldsymbol{x}_i) = \exp it(\beta_{0k} + \beta_{1k} t_i). \tag{3}$$

If $t_i = 1$, then (3) becomes $P(Y_i(1) = 1|T_i = 1, \boldsymbol{x}_i) = \text{expit}(\beta_{0k} + \beta_{1k})$. And because of the unconfounded assumption that makes $Y_i(c)$ independent from T_i given measured covariates $\boldsymbol{x}_i$ for $c = 0$ or 1, it gets to hold that $P(Y_i(1) = 1|\boldsymbol{x}_i) = \text{expit}(\beta_{0k} + \beta_{1k})$. Finally since $\text{expit}(\beta_{0k} + \beta_{1k})$ does not involve any covariates $\boldsymbol{x}_i$, $P(Y_i(1) = 1) = \text{expit}(\beta_{0k} + \beta_{1k})$. In the same way, it can be shown that $P(Y_i(0) = 1) = \text{expit}(\beta_{0k})$. Therefore, β_{1k} is precisely the causal parameter θ_k.

Intuitively, within the kth subset ($\forall i \in S_k$), because $P(Y_i|T_i, \boldsymbol{x}_i) = P(Y_i|T_i)$, the vector of confounding covariates would not modify the relationship between T_i and Y_i in that subset and would not affect Y_i directly. In addition, because $P(T_i|\boldsymbol{x}_i) = P(T_i)$, the influences of confounding covariates on T_i are blocked and hence selection bias (Rosenbaum and Rubin, 1983) is to be removed in the kth subset. Figure 1 explains this in a figurative way with three types of confounders: 1) prognostic factor that directly influences the outcome; 2) treatment-confounder which affects propensity scores; and 3) effect-modifier that changes the degree and direction of causal effects. These three types of confounding effects disappear within each of the subsets of the MT model due to the absence of covariates in right side of (1).

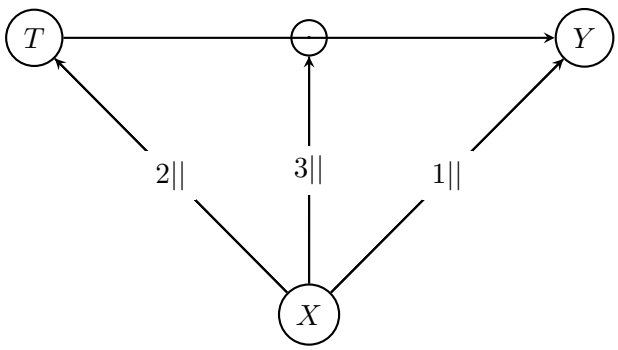

Figure 1: Three different types of confounding factors are controlled at each node of the MT model: "1||" indicates prognostic factor X not influencing Y; "2||" indicates treatment-confounder X not influencing T; and "3||" indicates effect-modifier X not modifying the relationship between T and Y

Therefore parameter β_{1k} is the causal effect for S_k ($\forall i \in S_k$) due to the fact that $P(Y_i|T_i) = P(Y_i|T_i, \boldsymbol{x}_i)$ and $P(T_i) = P(T_{ik}|\boldsymbol{x}_i)$.

Though the purpose of the MT is to explore the heterogeneities with respect to causal effects and propensities, the MT also computes average causal effects (ACE) for the entire study cohort. The ACE for the entire population can be estimated by combining the cell counts from the kth subset as the estimates proposed by Robins, Breslow and Greenland (1986). To explain this, let n_{rsk} denote cell count for $Y_i = r$ and $T_i = s$ at the kth subset; $R_k = (n_{11k} + n_{00k})/N_k$;

$Q_k = (n_{10k}+n_{01k})/N_k$; $R_k = (n_{11k}\cdot n_{00k})/N_k$; $S_k = (n_{10k}\cdot n_{01k})/N_k$. The Mantel-Haenszel estimator of the causal odds ratio $\hat{\theta}$ is R_+/S_+, where $R_+ = \sum_{k=1}^{K} R_k$ and $S_+ = \sum_{k=1}^{K} S_k$. Then Robins' variance estimator of $\hat{\theta}$ is

$$\widehat{var}\left(\log\hat{\theta}\right) = \frac{1}{2}\sum_{k=1}^{K}\left[\frac{P_kR_k}{R_+^2} + \frac{P_kS_{k+}Q_kR_k}{R_+S_+} + \frac{Q_kS_k}{S_+^2}\right], \tag{4}$$

with $\widehat{var}(\hat{\theta}) = \hat{\theta}^2 \cdot \widehat{var}(\log\hat{\theta})$ (Robins *et al.*, 1986).

2.3 The Tree Growing Procedure

The main idea of growing the tree is to find a data point that maximizes the sum of loglikelihood of the models for the left and the right child nodes (Su *et al.*, 2004). Without a loss of generality, we only consider binary splits induced by rules such as $\{X_j \leq A?\}$ for a continuous variable X_j or $\{X_j \in A?\}$ for a categorical variable, where A is a characteristic subset of distinct levels of X_j. We seek a tree model that accounts for the possible effect of $\boldsymbol{x}$ on the joint distribution of (Y_i, T_i) so that $P(Y_i, T_i|\boldsymbol{x}_i) = P(Y_i, T_i)$ within each of K terminal nodes.

Start with node S_k with binary exposures and binary responses. The corresponding likelihood function for data in k node is

$$\begin{aligned} L_k &= \prod_{i\in S_k} P(Y_i, T_i|\boldsymbol{x}_i) = \prod_{i\in S_k} P(Y_i|T_i, \boldsymbol{x}_i)P(T_i|\boldsymbol{x}_i) = \prod_{i\in S_k} P(Y_i|T_i)\,P(T_i) \\ &= \prod_{i\in S_k} (\exp it(\beta_{0k} + \beta_{1k}t_i))^{y_i}(1 - \exp it(\beta_{0k} + \beta_{1k}t_i))^{1-y_i} \cdot \pi_{ik}^{t_i}(1-\pi_{ik})^{1-t_i}, \end{aligned} \tag{5}$$

where $\pi_{ik} = P(T_i = 1|\boldsymbol{x}_i) = P(T_i = 1) = \pi_k$ is a constant within S_k. The log-likelihood function becomes

$$\begin{aligned} l_k = &\prod_{i\in S_k} (y_i \log(\exp it(\beta_{0k} + \beta_{1k}t_i)) + (1-y_i)\log(1 - \exp it(\beta_{0k} + \beta_{1k}t_i))) \\ &+ n_{1k}\log\pi_k + n_{0k}\log(1-\pi_k), \end{aligned} \tag{6}$$

where $n_{1k} = \prod_{i\in S_k} T_{ik}$ and $n_{0k} = \prod_{i\in S_k}(1 - T_{ik})$. The involved parameters $(\beta_{0k}, \beta_{1k}, \pi_k)$ are all specific to node τ, as well as (n_{1k}, n_{0k}). Their Maximum Likelihood Estimates (MLE) can be obtained by maximizing l_k. Plugging the MLE's for $(\beta_{0k}, \beta_{1k}, \pi_k)$, the maximized log-likelihood can be shown to be,

$$\begin{aligned} \widehat{l}_k = &\prod_{i\in S_k} (y_i \log(\exp it(\hat{\beta}_{0k} + \hat{\beta}_{1k}t_i)) + (1-y_i)\log(1 - \exp it(\hat{\beta}_{0k} + \hat{\beta}_{1k}t_i))) \\ &+ \log\frac{n_{1k}^{n_{1k}} \cdot n_{0k}^{n_{0k}}}{n_k^{n_k/2}}, \end{aligned} \tag{7}$$

where $\hat{\beta}_{0k}$ and $\hat{\beta}_{1k}$ respectively denote the log of relative risk and the log of odds ratio estimates in S_k.

When S_k is partitioned into the two child nodes by split s, the left child node S_{k_L} (answering 'Yes' to the binary question) and the right child node S_{k_R} (answering 'No'), the likelihood function becomes $L_{k_L} + L_{k_R}$ owning to the independence of data, where L_{k_L} and L_{k_R} both have analogous forms to L_k in (5). This likelihood can be viewed as a likelihood function for split s as well as those involved parameters. The best split $s^\star$ yields maximum likelihood. Practically, the best split is obtained as the one that corresponds to the largest maximized log-likelihood $\widehat{l}_{k_L} + \widehat{l}_{k_R}$, which can be given as, up to a constant,

$$\begin{aligned}
&\prod_{i\in\{S_{k_R}\}} (y_i \log(\exp it(\hat{\beta}_{0k_R} + \hat{\beta}_{1k_R} t_i)) + (1-y_i)\log(1-\exp it(\hat{\beta}_{0k_R} + \hat{\beta}_{1k_R} t_i))) \\
&+ \prod_{i\in\{S_{k_L}\}} (y_i \log(\exp it(\hat{\beta}_{0k_L} + \hat{\beta}_{1k_L} t_i)) + (1-y_i)\log(1-\exp it(\hat{\beta}_{0k_L} + \hat{\beta}_{1k_L} t_i))) \\
&+ \log\{n_{L1}^{n_{L1}} \cdot n_{L0}^{n_{L0}} / n_L^{n_L/2}\} + \log\{n_{R1}^{n_{R1}} \cdot n_{R0}^{n_{R0}} / n_R^{n_R/2}\},
\end{aligned} \tag{8}$$

where, n_{L1} denotes the total number of observations in the treatment group of the left child node. Subsequently, either child node is optimally partitioned in the same way. This splitting process is then continued until a terminal node is claimed under some released stopping rules. This procedure results in a large initial tree $\mathcal{T}_0$, which will be optimally pruned subsequently.

2.4 Pruning with Akaike Information Criteria (AIC)

The large initial tree structure $\mathcal{T}_0$ is an over-fitted model which contains both true and spurious splits. The final tree structure is a subtree of $\mathcal{T}_0$. Though we can investigate all its subtrees and select the best, there are too many subtrees to examine. The pruning idea of CART (Breiman *et al.*, 1984) is to selectively narrow down the subtree choices.

To prune marginal trees, it is convenient to adopt the AIC pruning developed by Su *et al.* (2004). To explain, we first introduce some notations. Let $\widetilde{\mathcal{T}}$ represent the set of all terminal nodes $\{S_k\colon\ k = 1, \cdots, K$, as in the previous section$\}$ for tree $\mathcal{T}$ and $\mathcal{T} - \widetilde{\mathcal{T}}$ represent the set of its internal nodes. $\mathcal{T}_h$ denotes the branch tree with h as its root node. Also the cardinality notation $|\cdot|$ indicates the number of nodes in a set of a tree, i.e., $|\widetilde{\mathcal{T}}|$, the number of terminal nodes in tree $\mathcal{T}$. Note that for a given tree structure $\mathcal{T}$, the corresponding maximized log-likelihood function can be given as

$$\widehat{l}_{\mathcal{T}} = -\sum_{\tau\in\widetilde{\mathcal{T}}} \widehat{l}_\tau. \tag{9}$$

To measure the performance of tree $\mathcal{T}$, AIC (Akaiki, 1974) can be used:

$$\mathrm{AIC}^{(\mathcal{T})} = -2 \cdot l^{(\mathcal{T})} + 2 \cdot 3 \cdot |\widetilde{\mathcal{T}}|, \tag{10}$$

where $3 \cdot |\widetilde{\mathcal{T}}|$ is the total number of parameters in (6). A smaller AIC is associated with a more preferable tree model among competing ones.

The AIC pruning starts with the large initial tree $\mathcal{T}_0$. For any internal node h of $\mathcal{T}_0$, let $\mathcal{T}_0 - \mathcal{T}_h$ indicate the subtree with h truncated and compute $\mathrm{AIC}^{(\mathcal{T}_0 - \mathcal{T}_h)}$. Then the internal node h^* is the weakest link that minimizes $\mathrm{AIC}^{(\mathcal{T}_0 - \mathcal{T}_h)}$ over all possible values of h. Here, h^* is "weakest" because without the internal node h^* the resultant subtree best fits the data compared to the other internal nodes. Let $\mathcal{T}_1 = \mathcal{T}_0 - \mathcal{T}_{h^*}$. Subsequent pruning in this manner will produce a decreasing sequence of subtrees $\mathcal{T}_M \prec \cdots \prec \mathcal{T}_1 \prec \mathcal{T}_0$, where $\mathcal{T}_M$ is the root node that represents the undivided entire data set and the notation $\prec$ means "is subtree of".

2.5 Selecting the Best-Sized Subtree

In order to determine the optimally pruned tree, we divide the entire sample into two groups: the learning sample $\mathcal{L}_1$ and the test sample $\mathcal{L}_2$ with a ratio of 2 : 1 as in (Su *et al.*, 2004). Using the learning sample$\mathcal{L}_2$. Let $l_{test}^{(\mathcal{T}_m)}$ be the validated log-likelihood computed with parameter estimates obtained from the learning sample $\mathcal{L}_1$and data values of y and t used from the test sample $\mathcal{L}_2$. Then

$$\mathrm{AIC}^{(\mathcal{T}_m)} = -2 \cdot l_{test}^{(\mathcal{T}_m)} + 2 \cdot 3 \cdot |\widetilde{\mathcal{T}}_m|, \tag{11}$$

so that the best-sized subtree $\mathcal{T}^*$ may be defined with minimum AIC:

$$\mathrm{AIC}^{(\mathcal{T}^*)} = \min_{\{\mathcal{T}_m | 0 \le m \le M\}} \mathrm{AIC}^{(\mathcal{T}_m)}. \tag{12}$$

After the most optimal tree structure is identified using $\mathcal{L}_1$ and $\mathcal{L}_2$, more accurate estimates for the node parameters can be computed by applying the tree model to the entire data set. Also the tree $\mathcal{T}^*$ can be further pruned using likelihood ratio tests because the space of parameters in (6) can be shown to be nested in those of their parental nodes.

When the sample size is moderate, the log-likelihood $l^{(\mathcal{T}_m)}$ in AIC can be validated via V-fold cross validation. In this approach, we randomly divide the data into V equally-sized folds and estimate the parameters (β_0, β_1, π) using $(V-1)$ folds of data with the v-th fold excluded. Then we calculate the log-likelihood by plugging the observations in the v-th fold (denote it as $l_v^{(\mathcal{T}_m)}$). The cross-validated log-likelihood for the m-th subtree is given by $l_{\mathrm{CV}}^{(\mathcal{T}_m)} = \sum_{v=1}^{V} l_v^{(\mathcal{T}_m)}$.

3. Simulation Study Designs and Analysis Results

3.1 Designs of Simulation Studies

Before assessing the effect of dieting on adolescent girls' emotional distress, we evaluate the MT method through simulated samples. There are two different sample sizes: one with 600 subjects and the other with 1200 subjects.

First, three covariates X_1, X_2 and X_3 are generated independently and identically from integers 1, 2, 3, 4 and 5, with each element having an equal probability of 0.2. The model configuration is shown in Table 1, which can be viewed as a Venn diagram. There are three subsets–W_1: $X_1 \leq 0.5$ & $X_2 \leq 0.5$; W_2: $X_1 \leq 0.5$ & $X_2 > 0.5$; and W_3: $X_1 > 0.5$. Our choice of the median value is for the sake of simplicity.

Table 1: Simulation designs with Cases I-III

	Case I	Case II	Case III
W_1	Propensity: 0.5 Odds Ratio: 0.1	Propensity: 0.2 Odds Ratio: 0.9	Propensity: 0.2 Odds Ratio: 0.1
W_2	Propensity: 0.4 Odds Ratio: 1.0	Propensity: 0.5 Odds Ratio: 1.0	Propensity: 0.5 Odds Ratio: 1.0
W_3	Propensity: 0.6 Odds Ratio: 5.0	Propensity: 0.7 Odds Ratio: 1.1	Propensity: 0.7 Odds Ratio: 5.0

In Case I of Table 1, propensities for the three subgroups are also close to or equal to 0.5. Propensity of 0.5 usually indicates a completely randomized treatment assignment. While propensities in Case I are close to 0.5, the effect sizes (odds ratios) are distinguishable from one another: W_1 subset has a negative effect, W_2 subset null effect, and W_3 subset positive effect. Unlike Case I, Case II has discernible propensities for the three subsets-0.2, 0.5 and 0.7- while the effect sizes are all close to 1. For example, it is hard to distinguish odds ratios of 0.9 and 1. Finally Case III has all differential propensities and effect sizes.

3.2 A Measure of Tree Performance

In this section we describe a way of measuring the ability to separate each pair of observations of an empirical tree model when compared to the true tree model underlying the data. Consider a $n \times n$ proximity matrix $\boldsymbol{D}(\mathcal{T})$ for the data based on tree $\mathcal{T}$ such that $d_{ii'} = 1$ if the i-th observation falls into a different terminal node from i'-th observation and 0 otherwise. $\boldsymbol{D}(\mathcal{T})$ is a symmetric matrix with diagonal elements being 0. First compute the corresponding proximity matrix $\boldsymbol{D}$ for the true tree structure $\mathcal{T}^{(0)}$ from which the data were generated. Likewise, the proximity matrix $\boldsymbol{D'}$ can be computed for the final tree structure $\mathcal{T}$ developed

via the MT model from a simulated data set. Let $g_{ii'} = d_{ii'} - d'_{ii'}$. Notice that $g_{ii'} = 0$ if and only if $\mathcal{T}$ and $\mathcal{T}^{(0)}$ agree; 1 if $\mathcal{T}$ separates a pair that should be together according to $\mathcal{T}^{(0)}$; and -1 if $\mathcal{T}$ fails to separate a pair that should be separated according to $\mathcal{T}^{(0)}$. We define the disagreement measure κ between an empirical tree structure developed from the data and the true tree structure as the total proportion of disagreements

$$\kappa = \frac{\sum_{i=1}^{n} \sum_{i'>i} \left| d_{ii'} - d'_{ii'} \right|}{\{n(n-1)/2\}} = \frac{\sum_{i=1}^{n} \sum_{i'>i} |g_{ii'}|}{\{n(n-1)/2\}}. \tag{13}$$

In assessing the performance of $\mathcal{T}$, both the number of positive ones and the number of negative ones can be reported . In other words,

$$\kappa_+ = \frac{\sum_{i=1}^{n} \sum_{i'>i} g_{ii'} \cdot 1_{(g_{ii'}>0)}}{\{n(n-1)/2\}}, \tag{14}$$

and

$$\kappa_- = \frac{\sum_{i=1}^{n} \sum_{i'>i} g_{ii'} \cdot 1_{(g_{ii'}<0)}}{\{n(n-1)/2\}}. \tag{15}$$

We expect a large κ_+ if tree $\mathcal{T}$ over-split the data while κ_- is large if tree $\mathcal{T}$ under-split the data. Apparently, κ resembles the agreement coefficient of (Cohen, 1960) in some way and can be illustrated using a 2×2 contingency table. The results of this misclassification measures are reported in the bottom part of Tables 2 and 3. The misclassification rate has three components: for κ_+, "Over-misclassification rate" indicates the frequency of $g_{ii'} = 1$ divided by $((n-1) \cdot n/2)$; for κ_-, "Under-misclassification rate" indicates the frequency of $g_{ii'} = -1$ divided by $((n-1) \cdot n/2)$; and "Misclassification rate" is just the sum of both "Over-misclassification rate" and "Under-misclassification rate". n can take a sample size of 600 (or 1200). These misclassification results in Tables 2 and 3 are percentages rounded at the first decimal point.

3.3 Performance of the Trees for 200 Simulated Samples

Figure 2 shows that the MT optimally identifies the underlying true subsets of Table 1. Numbers inside the nodes indicate node numbers and character strings below the nodes indicate true subsets. Taking an example of Case I in Table 1, the tree diagram in Figure 2 indicates that variable X_1 plays the most determining role in splitting the population because the median split of X_1 variable makes the maximum values of $P(Y_L|T_L)P(T_L) + P(Y_R|T_R)P(T_R)$, where the subscript '$_L$' indicates a subgroup whose description is $X_1 \leq 0.5$, and '$_R$' indicates a subgroup with $X_1 > 0.5$. Whether it comes to propensities or effect sizes, as long as either of the two quantities are differentiable, the MT correctly identifies the three subsets.

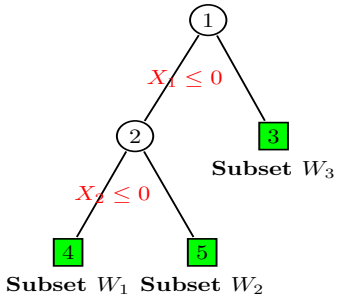

Figure 2: Tree Graph for Table 1

To assess the performance of the MT method for the three different Cases I-III in Table 1, we repeatedly drew 200 samples of size 600 (1200). In Tables 2 and 3, "BIAS" indicates the average difference between 200 estimates of causal effects and true values of effects; RMSE (Root Mean Square Error) is the square root of the average squared difference between 200 estimates and true values of causal effects. The true effects and propensities are computed from each sample with true subset memberships that were a priori set in Table 1. Misclassification rates are explained in the previous Section 3.2. In order for the estimated trees to identify the three true subsets of Cases I-III of Table 1, a terminal node, in which a majority of subjects show a certain true subset, was designated to estimate that particular true subset. In this way, the misclassification errors are also incorporated in estimating causal effect. Also BIAS and RMSE were computed for all subsets for only observed estimates because some samples did not produce estimates of effect for subgroup W_2 in general, which was revealed in the ignorable misclassification rates.

Table 2: Simulation results with Cases I-III for sample size 600

(Sample size = 600)	Case I		Case II		Case III	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Overall effect	0.009	0.021	0.001	0.022	0.009	0.031
Effect for subset W_1	0.057	0.109	0.004	0.203	0.061	0.159
Effect for subset W_2	0.003	0.037	0.003	0.079	0.037	0.566
Effect for subset W_3	0.000	0.002	0.000	0.000	0.005	0.064
Propensity for subset W_1	0.009	0.020	0.033	0.062	0.018	0.045
Propensity for subset W_2	0.000	0.003	0.002	0.012	0.000	0.008
Propensity for subset W_3	0.000	0.000	0.000	0.000	0.000	0.001
Over-misclassification (%)	0.0		0.1		0.1	
Under-misclassification (%)	5.0		5.4		2.7	
All-misclassification (%)	5.0		5.5		2.8	

Table 3: Simulation results with Cases I-III for sample size 1200

(Sample size = 1200)	Case I		Case II		Case III	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Overall effect	0.001	0.008	0.000	0.007	0.003	0.018
Effect for subset W_1	0.007	0.037	0.005	0.042	0.011	0.076
Effect for subset W_2	0.000	0.000	0.000	0.000	0.000	0.000
Effect for subset W_3	0.000	0.002	0.003	0.045	0.000	0.002
Propensity for subset W_1	0.001	0.007	0.006	0.027	0.003	0.021
Propensity for subset W_2	0.000	0.000	0.000	0.000	0.000	0.000
Propensity for subset W_3	0.000	0.000	0.000	0.001	0.000	0.000
Over-misclassification (%)	0.0		0.1		0.0	
Under-misclassification (%)	0.7		1.1		0.5	
All-misclassification (%)	0.7		1.1		0.5	

Tables 2 and 3 describe that the MT for Cases I-III of Table 1 worked fairly well in all criteria. As the sample size increased from 600 to 1200, the overall measures became more precise. This simulation indicates that the likelihood-based decision rule that has both the element of outcome model $P(Y|T)$ and propensity model $P(T)$ disclose the correct data structure when either the weak effect sizes or the subtle differences in propensities are present. Misclassification rates for the MT in Tables 2 and 3 all appeared to be insignificant, which indicates that the MT correctly identifies the imbedded heterogeneity of the differential subgroups.

4. Applications

4.1 Lalonde's National Supported Work Demonstration Analysis

The Lalonde's National Supported Work Demonstration analysis (NSW) treated people with a certain job training program and compared control subjects to see if there would be causal effects of job training on later real earnings (Dehejia and Wahba, 1999). By its design, this study is a non-randomized observational study.

A publicly available subsample of this dataset is called Lalonde and it is currently available in R software packages twang (Ridgeway, McCaffrey, Morral, Griffin and Burgette, 2012) and MatchIt (Ho, Imai, King and Stuart, 2012). The variables in the Lalonde dataset include participation in the job training program (treat, which is equal to 1 if participated in the program, and 0 otherwise), age, years of education, race (black, which is equal to 1 if black, and 0 otherwise; hispan, which is equal to 1 if hispanic, and 0 otherwise), marital status, high school degree (nodegree, which is equal to 1 if no degree, and 0 otherwise), 1974 real earnings, 1975 real earnings, and the main outcome variable, 1978 real earnings

($re78$). For the descriptive statistics of these covariates, refer to Ridgeway *et al.* (2012).

For the purpose of the binary MT case, we create a new binary outcome variable based on median real earnings of 1978 in the Lalonde dataset. That is, the new outcome was 1 if the participant has $re78$ greater than its median earning, and 0 otherwise. Our goal was to estimate causal effects of the job training program on the median real earnings of 1978.

The MT generated two terminal nodes using 614 Lalonde subjects: 243 blacks and 371 non-blacks. Both blacks and non-blacks had statistically insignificant causal log odds ratios (P-values of 0.221 and 0.444 for both blacks and non-blacks respectively). The Robins' formula for the MT's ACE (Section 2.2) computed causal odds ratio 0.917 with confidence interval $[0.633, 1.329]$ with P-value 0.648. This implies that in general there is no effect of the job training program for all 614 Lalonde subjects. In Figure 3, the estimated propensity scores were reported in curve parentheses next to the total sample sizes in the two terminal nodes.

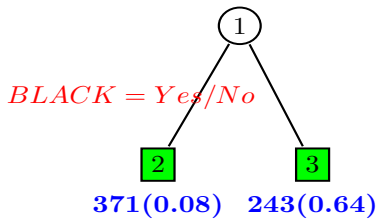

Figure 3: Tree Graph for the MT analysis of the Lalonde data

We also estimated the ACE using the Marginal Structural Model (MSM; Robins *et al.*, 2000). The propensity score was estimated generalized boosted regression, which can be readily estimated by the R software package "twang" (Ridgeway *et al.*, 2012). The MSM produced basically the same ACE estimates: causal odds ratio 1.016 with confidence interval [0.607, 1.702] with P-value 0.950. In conclusion, 1) both the MT and the MSM produced equivalent ACE estimates of the job training effect on the median real earnings and 2) there was no significant causal effect of the putative job training on median real earnings for 614 Lalonde subjects.

4.2 The Study of the Association between Dieting Behavior and Emotional Distress in Adolescent Girls

In this section, we apply the MT model to the study of the dieting effects on emotional distress among adolescent girls. Though some studies showed that dieting appeared to be associated with negative psychological outcomes including

depression and anxiety (Kovacs, Obrosky and Sherrill, 2003; Warren and Cooper, 1988), Johnson and Wardle (2005) found that dietary restraint had no significant effect on distress one year later after controlling for measures of body dissatisfaction. But statistical summaries of dieting effects from these researches principally depended on multiple regressions.

Thus using our MT method, we explored in identifying subgroups whose dieting effects are heterogeneous and in aggregating those effects as average causal dieting effects. Schafer and Kang (2008) generated a simulated data that infers psychological effects of dieting and that does not rely on typical parametric distributions such as Gaussian normal or binomial distributions. This data set was used to approximate the National Longitudinal Study of Adolescent Health (Add Health), a representative sample of American middle and high school students measuring a broad array of health characteristics, behaviors and attitudes (Harris, 2009). Throughout this section, we used this simulated data set. Detailed simulation procedures are described in Schafer and Kang (2008). The merits of the usage of this simulated data set are signified by no missing values (no dropout) and fully observed potential outcomes that define causal effects. Our simulated data is a simple random sample of $N = 6{,}000$ girls from http://www.stat.psu.edu/~jls/causal/. Using the MT model, we aimed to determine whether a dichotomous status of dieting (the putative cause variable) at wave I differentially caused emotional distress (the outcome variable) at wave II which was one year from wave I. The emotional distress is a composite measure of a 19-item mood scale with response categories ranging from 0 (never or rarely) to 3 (most or all of the time). The 19 items were averaged so that the average values may be continuously distributed between 0 and 3 in a similar way by Resnick, Bearman, Blum, Bauman, Harris, Jones, Tabor, Beuhring, Sieving, Shew, Ireland, Bearinger and Udry (1997). Once we obtained the continuous scale measure, we further dichotomized each of potential outcomes of the emotional distress at their respective medians. Namely, a potential emotional distress $Y(0)$ under no dieting ($T = 0$) was defined to be high ($Y(0) = 1$) when the distress value was bigger than its median 0.57, otherwise it was set to be low ($Y(0) = 0$). Also a potential emotional distress $Y(1)$ under dieting ($T = 1$) was defined to be high ($Y(1) = 1$) when the distress value was bigger than its median 0.55. Data analysis with the MT model used only observed potential outcomes ($TY(1) + (1 - T)Y(0)$). For the baseline emotional distress variable, we used its three quartiles (0.32, 0.58, 0.89) to define four resulting subgroups. The list of variables are found in the Table 4.

With the simulated 6,000 girls, the true causal effect with the known potential outcomes of the emotional distress was 1.03. Thus using the known potential outcomes, we see that there is no significant dieting effect at the entire sample

Table 4: Descriptions of variables used in simulated ADD health population

Name	Description	Mean	SD
DISTR.1	Emotional distress at Wave I (min = 0, max = 2.84)	0.64	0.42
BLACK	1 = Black, 0 = otherwise	0.24	0.43
NBHISP	1 = non-Black Hispanic, 0 = otherwise	0.15	0.36
GRADE	Grade in school at Wave I $(7, \cdots, 11)$	9.21	1.38
SLFHLTH	Self-rating of overall health (1 = excellent, 2 = very good, $\cdots$, 5 = poor)	2.23	0.93
SLFWGHT	Self-rating of weight (1 = very underweight, 2 = slightly under, $\cdots$, 5 = very over)	3.32	0.79
WORKHARD	"When you get what you want, it's usually because you worked hard for it" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	2.12	0.9
GOODQUAL	"You have lots of good qualities" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	1.81	0.68
PHYSFIT	"You are physically fit" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	2.3	0.93
PROUD	"You have a lot to be proud of" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	1.78	0.77
LIKESLF	"You like yourself just the way you are" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	2.18	1.02
ACCEPTED	"You feel socially accepted" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	2.18	1.02
FEELLOVD	"You feel socially accepted" (1 = strongly agree, $\cdots$, 5 = strongly disagree)	1.81	0.85

level. Now we would like to explore a question regarding whether there are subgroups of adolescent girls who show differential effects of dieting on their emotional distress. To explore this question, we fitted the MT model for the proposed data set. The $6,000$ girls were randomly divided into 3 groups. The first two groups were used to build series of nested subtrees and the last group was used to validate the proposed trees so as to obtain the most optimal tree. With the most optimal tree we used the $6,000$ girls to estimate dieting effects and propensities as shown in Figure 4. In Figure 4, the estimated propensity scores were reported in parentheses next to the total sample sizes in the terminal nodes.

In total, nine terminal nodes were identified by the MT. As in Figure 4, the tree first divided the data set at the median value 0.58 of the baseline emotional distress (DISTR). For those in the group with DISTR$\leq$ 0.58, they were further

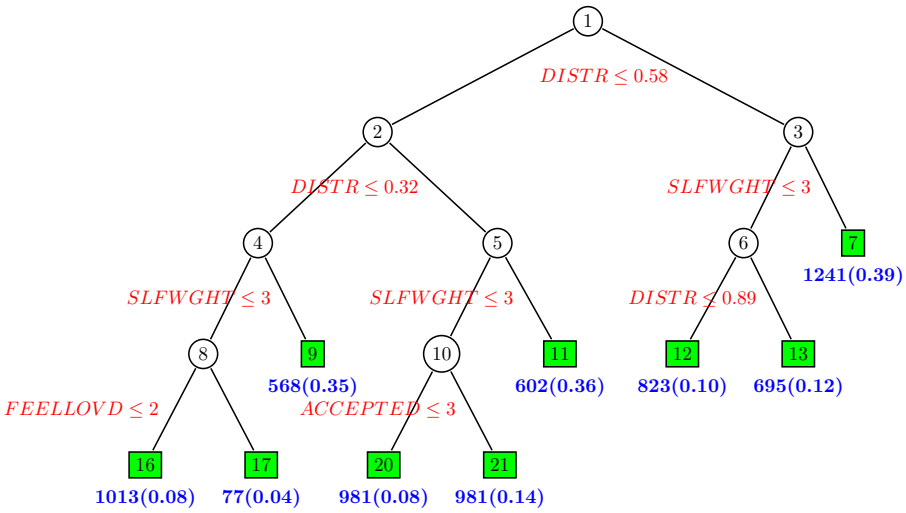

Figure 4: Tree Graph for the MT analysis of dieting data

divided into two groups at the first quartile (0.32) of the baseline emotional distress. And subsequently, three more confounding variables were introduced as the data set was further divided: SLFWGHT, FEELLOVD, ACCEPTED. The descriptions of these variables are listed in Table 3.

Since the nine subgroups were from one study sample, there were nine simultaneous tests and interval estimations to see if their respective causal odd ratios were significant (H_o: odds ratio is 1 vs. H_a: odds ration is not 1). The Bonferroni-adjusted confidence intervals for all subgroups of the MT included value 1 and their four p-values were bigger than the adjusted significance level of 0.005 ($\approx 0.05/9$). The MT's average causal odds ratio for the entire sample was 1.03[0.88-1.21], where values inside the "[]" indicate confidence intervals. Note that the MT's point estimate is the same as the true causal odds ratio 1.03 which was computed with fully known potential outcomes. We ran this analysis 5 more times with 5 different types of random division of the 6,000 girls. The 5 different runs of the MT consistently produced insignificant subgroups (terminal nodes). This may be largely due to the fact that the likelihood-based MT estimates with randomly divided samples are consistently same across the divided samples whose differences are random rather than systematic.

Finally, with fully known potential outcomes, the results from all nine subgroups were consistently verified that none of the true causal effects were significant for the subgroups. These findings lead to the conclusion that there is no significant causal dieting effect for the entire subgroups as well as the whole sample. This finding is consistent with Kang *et al.* (2012).

5. Discussion

In this paper, we extended the MT model to the case of binary outcomes. In particular, the relationship between the parameter estimates using the MT and those under the potential outcome framework was explained for the case of binary outcome. The MT method does require the unconfoundedness assumption (Rubin, 2005) in order to estimate causal parameters while the parametric assumptions both for the potential outcomes and for the propensity scores were relaxed in a sense because the MT divides the entire data set with the nonparametric recursive partitioning algorithm. Yet, the MT is not completely nonparametric but semiparametric because it clearly specifies simple likelihoods for 1) the association between the cause and its related outcome and 2) the propensities.

Because the purpose of the MT is rather exploratory, it would fit into a large observational study which is expected to have differential effects along with differential propensity scores. For the confirmatory verification of the identified significance of different subgroups, one can, for example, apply the MT model from one study cohort (one study site among multi-site research centers) to the other study cohort (the other study site) to see if those two models still produce the identical results. In this way external validity may be assessed across different study sites within the same large study cohort. Also the newly found characteristics of differential subgroups can serve to generate new scientific hypotheses, which will guide health scientists to explore whether or not these findings are clinically and epidemiologically meaningful.

The stability of the MT is a shortcoming which a single tree analysis has been criticized for. Instability implies that a small perturbation in data would cause dramatic changes in the final tree model. With such concern being ubiquitous in any single tree-related methodologies, the stability of the MT can be explained with the misclassification rates in Tables 2-3. It is because while the general usage of CART (Breiman *et al.*, 1984) is principally for the prediction purpose, the MT is most concerned about its ability to single out subgroups with differential treatment effects. Thus the stability of the MT can be judged by whether it successfully mines the correct subgroups that had been embedded in the simulated data sets. Tables 2-3 show that as sample size increases, the classification performance, which was measured by misclassification rates, improved. The reason for this stable enhancement is largely due to the fact that likelihood estimation, which the MT employs, is more precise with a larger sample. More extensive analysis of the stability can be addressed by the bagging approach or the random forests' paradigm, which may be an outstretching of this paper.

The generalizability of the MT to other biostatistical and/or epidemiological studies is immediate because the splitting criterion of the MT is likelihood-based. A remaining limitation of the MT is that the decision rules do not distinguish

variables by types that influence the outcome, the treatment assignment mechanism, and/or the modification of the causal effects due to the fact that the MT adjusts for both confounding effects in a simultaneous, yet implicit manner. One way to bypass this problem is to make separate regressions–one for outcome and the other for the propensity model–so that variables can be identified to influence both/either the propensity model and/or the outcome model.

The R codes that are used to build the MT model will be distributed upon request to the corresponding author.

References

Akaiki, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees.* Chapman & Hall, Boca Raton, Florida.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37-46.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053-1062.

Harris, K. (2009). The National Longitudinal Study of Adolescent Health (Add Health), Waves 1 and 2, 1994-1996; Wave 3, 2001-2002; Wave 4, 2007-2009. [*machine-readable data file and documentation*]. Carolina Population Center, University of North Carolina at Chapel Hill.

Ho, D., Imai, K., King, G. and Stuart, E. (2012). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. http://gking.harvard.edu/matchit/

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945-960.

Johnson, F. and Wardle, J. (2005). Dietary restraint, body dissatisfaction, and psychological distress: a prospective analysis. *Journal of Abnormal Psychology* **114**, 119-125.

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

Kovacs, M., Obrosky, D. S. and Sherrill, J. (2003). Developmental changes in the phenomenology of depression in girls compared to boys from childhood onward. *Journal of Affective Disorders* **74**, 33-48.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press, New York.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H. and Udry, J. R. (1997). Protecting adolescents from harm: findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association* **278**, 823-832.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A. and Burgette, L. (2012). twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. http://cran.r-project.org/web/packages/twang/index.html

Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**, 311-323.

Robins, J. M., Hernán, M. A. and Brumbac, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550-560.

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

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

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods* **13**, 279-313.

Su, X. G., Tsai, C. L., Wang, H. S., Nickerson, D. M. and Li, B. G. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* **10**, 141-158.

Su, X. G., Wang, M. and Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics* **13**, 586-598.

Su, X. G., Zhou, T., Yan, X., Fan, J. and Yang, S. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics* **4**, Article 2.

Warren, C. and Cooper, P. J. (1988). Psychological effects of dieting. *British Journal of Clinical Psychology* **27**, 269-270.

Zhang, H. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences.* Springer, New York.

Received February 24, 2012; accepted June 22, 2012.

Joseph Kang
Department of Preventive Medicine
Northwestern University
680 North Lake Shore Drive, Suite 1400, Chicago, IL 60611, USA
joseph-kang@northwestern.edu

Xiaogang Su
School of Nursing
University of Alabama at Birmingham
1530 3rd Avenue South, Birmingham, AL 35294-1210, USA
xgsu@uab.edu

Kiang Liu
Department of Preventive Medicine
Northwestern University
680 North Lake Shore Drive, Suite 1102, Chicago, IL 60611, USA
kiangliu@northwestern.edu