Variable Selection for Causal Effect Estimation: Nonparametric Conditional Independence Testing with Random Forests

Bryan Keller

Teachers College, Columbia University

Author Note

This paper is in press at the *Journal of Educational and Behavioral Statistics*. I thank Xavier de Luna, Jenny Häggström, and Peter Steiner for helpful discussion on an earlier draft of this paper.

Abstract

Widespread availability of rich educational databases facilitates the use of conditioning strategies to estimate causal effects with non-experimental data. With dozens, hundreds, or more potential predictors, variable selection can be useful for practical reasons related to communicating results, and for statistical reasons related to improving the efficiency of estimators. Background knowledge should take precedence in deciding which variables to retain. However, with many potential predictors, theory may be weak, such that functional form relationships are likely to be unknown. In this paper, I propose a nonparametric method for data-driven variable selection based on permutation testing with conditional random forest variable importance. The algorithm automatically handles nonlinear relationships and interactions in its naive implementation. Through a series of Monte Carlo simulation studies, and a case study with ECLSK data, I find that the method performs well across a variety of scenarios where other methods fail.

Keywords: Nonparametric conditional independence test, causal inference, variable selection, average treatment effect, random forest, permutation test

Variable Selection for Causal Effect Estimation: Nonparametric Conditional Independence Testing with Random Forests

Introduction

Widespread availability of high-quality educational databases such as the Early Childhood Longitudinal Study (ECLSK; Tourangeau, Nord, Lé, Pollack, & Atkins-Burnett, 2006) has made it easier for educational researchers to search for answers to causal research questions with non-randomized data through conditioning strategies. To identify an average causal effect with observational data through a conditioning strategy, such as regression estimation or propensity score analysis, one must assume that the potential outcomes are independent of the assignment mechanism, conditional on the observed pretreatment variables, $\mathbf{X}$. This empirically untestable assumption, often referred to as *ignorability* (Rosenbaum & Rubin, 1983), essentially implies that all confounding variables have been observed and reliably measured (Steiner, Cook, & Shadish, 2011). Ignorability is more likely to be satisfied if the set of conditioning variables is comprehensive and large.

From a practical perspective, a researcher using a large database may wish to winnow down the set of predictors from the dozens, hundreds, or thousands available to a more manageable number to (a) facilitate interpretation and communication of results[1], (b) allow parametric models to be stably fit, or (c) allow nonparametric algorithms to converge in a reasonable amount of time. From a statistical perspective, the indiscriminate inclusion of conditioning variables that are not associated with the potential outcomes will decrease the efficiency of causal effect estimators, even if ignorability is satisfied (Hahn, 1996, 2004). Substantive knowledge should play a prominent role in guiding which variables should be retained for conditioning, with an eye toward selecting multiple variables from heterogeneous domains (Steiner, Cook, Li, & Clark, 2015). However, especially when the number of candidate variables is large and/or the theory is weak, data-driven approaches may serve as a useful complement to theory-based variable selection.

In this paper, I develop a computationally feasible algorithm for nonparametric conditional independence testing that is particularly useful for variable selection in the context of estimating

[1]While the predictors in a propensity score or outcome model are not of primary interest when estimating an average causal effect, they are important for a number of reasons relevant to interpretation and communication of results such as balance checking and reporting, detecting effect heterogeneity and estimating conditional average treatment effects, and characterizing units deleted for considerations regarding lack of overlap.

causal effects by conditioning. The algorithm, implemented in R package **rfvarsel** (to be submitted to CRAN), is designed to select important predictors through conditional independence testing based on permutation of conditional random forest variable importance. Following description of the algorithm, it is tested through Monte Carlo simulations, which reveal that it performs well (a) in selecting important variables in its naive implementation even when the relationships between predictors and outcome are nonlinear and (b) in reducing bias and variability when paired with a variety of causal effect estimators. The algorithm is then applied to causal effect estimation in a large-scale educational data set, and some connections are made between the case study findings and expectations based on simulation results and theoretical considerations. The paper concludes with some discussion.

Theoretical Framework

Let T_i be the dichotomous exposure indicator and let Y_i^0 and Y_i^1 represent the potential outcomes for unit i in the event that $T_i = 0$ or 1, respectively. For notational convenience, I will suppress the index i going forward. I define *variable selection* as the process by which a subset of predictor variables $\mathbf{X}_S \subset \mathbf{X}$ is identified with the aim of satisfying unconfoundedness. That is,

$$\left\{Y^0, Y^1\right\} \perp\!\!\!\perp T \mid \mathbf{X} \quad \Longrightarrow \quad \left\{Y^0, Y^1\right\} \perp\!\!\!\perp T \mid \mathbf{X}_S.$$

To estimate an average treatment effect with a conditioning strategy, one or both of $\Pr(T \mid \mathbf{X})$ or $Y^j \mid \mathbf{X}, T = j$ for $j \in \{0, 1\}$ may be modeled. Although many data-driven approaches to variable selection have been proposed for causal inference applications, I use the framework proposed by de Luna, Waernbaum, and Richardson (2011) herein because it is aligned with the definition of variable selection given above, and provides theoretical guarantees regarding identification of the average causal effect.

It has been shown both analytically (Battacharya & Vogt, 2007) and through simulation (Austin, Grootendorst, & Anderseon, 2007; Brookhart et al., 2006) that conditioning on treatment-only predictors (i.e., instrumental variables; IVs) can decrease the efficiency of propensity score-based estimators, while conditioning on outcome-only predictors can increase efficiency. Furthermore, conditioning on an IV or a collider variable can amplify bias (Steiner &

Kim, 2016; Battacharya & Vogt, 2007; Wooldridge, 2009; Pearl, 2010). Although de Luna et al. (2011) proposed a number of target subsets for variable selection, I focus on the subset they refer to as $\mathbf{X}_Y$, because outcome-only predictors are retained in $\mathbf{X}_Y$, which, when conditioned on, result in more efficient estimation of average causal effects (Persson, Häggström, Waernbaum, & de Luna, 2017; Häggström, 2018). The procedure for selecting $\mathbf{X}_Y$ is carried out as follows.

Step 1. For $j \in \{0, 1\}$, define $\mathbf{X}_j \subseteq \mathbf{X}$ of minimum cardinality such that

$$Y^j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_j \mid \mathbf{X}_j, T = j \tag{1}$$

where $\mathbf{X} \setminus \mathbf{X}_j$ denotes the set of predictors that are in $\mathbf{X}$ but not in $\mathbf{X}_j$.

Step 2. Let $\mathbf{X}_Y = \mathbf{X}_0 \cup \mathbf{X}_1$.

If ignorability holds with $\mathbf{X}$, conditional independence of T and $\{Y^0, Y^1\}$ will also hold with $\mathbf{X}_Y$ (de Luna et al., 2011).

As an example, consider the graphical models for the potential outcomes, Y^0 and Y^1, displayed in Figure 1. To apply the algorithm, begin with the model for Y_0 and find the subset, $\mathbf{X}_0$, of minimum size such that Y^0 is conditionally independent of all variables not in $\mathbf{X}_0$, given the variables in $\mathbf{X}_0$. Note that the fact that the algorithm also calls for conditioning on $T = 0$ implies that it may be used with observed data because, for example, Y_i^0 is only observed for units that also have $T_i = 0$. Here, the minimum subset is $\mathbf{X}_0 = \{X_3, X_4, X_6\}$. Repeat the process for Y^1, and find that $\mathbf{X}_1 = \{X_3, X_4, X_5\}$. Finally, $X_Y = \mathbf{X}_0 \cup \mathbf{X}_1 = \{X_3, X_4, X_5, X_6\}$.

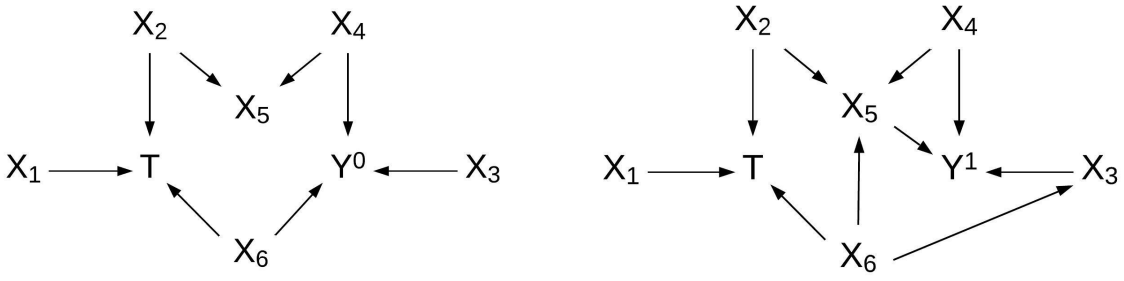

Figure 1. Directed acyclic graphs for potential outcomes Y^0, and Y^1.

Random Forests and Variable Importance

The $\mathbf{X}_Y$ variable selection procedure is expressed with nonparametric statements of conditional independence. To maintain freedom from parametric and functional form assumptions, a method for conditional independence testing that relies on minimal assumptions is also desired. Random forests are useful because they are based on regression trees (Breiman, Friedman, Olshen, & Stone, 1984), which nonparametrically handle higher-order interactions and non-linear relationships.

Regression Trees

A *regression tree* is an algorithmic tool invented by Breiman et al. (1984) that models the relationship between a continuous outcome variable, Y, and predictors, $X_1, \ldots, X_p$, by iteratively splitting the data into subgroups based on predictor values. Every split creates two subgroups, called *nodes*, and any node that is unsplit is called a *terminal node* or *leaf node*. The tree-predicted value for each member of a node is simply the mean outcome score for all units in that node. Thus, for unit i in terminal node t, where N_t represents the set of units in t, and $|N_t|$ represents the number of units in t, the tree-predicted value for unit i is

$$\hat{Y}_i = \frac{1}{|N_t|} \sum_{i \in N_t} Y_i.$$

The adequacy of a regression tree fit may be measured by the *deviance*, defined for tree S as the sum of squared prediction errors, or by the closely related mean squared prediction error (MSPE):

$$\text{dev}(S) = \sum_i \left(Y_i - \hat{Y}_i\right)^2$$

$$\text{MSPE}(S) = \text{dev}(S) / \sum_i 1.$$

At each iteration, the tree-fitting algorithm considers every possible split on every variable, and the single split that results in the largest decrease in deviance is selected. If left unchecked, regression trees will continue to split until each terminal node contains only one unit, yielding a perfect fit to the data. A typical approach to prevent overfitting is to fit a complex tree and then

"prune" it by dropping nodes according to a regularized solution based on adding a term to the deviance that penalizes additional nodes (see, e.g., Hastie, Tibshirani, & Friedman, 2009 or Venables & Ripley, 2002 for details). The value of the penalty term coefficient is typically chosen by cross validation. See Appendix A for an example in which regression trees are applied to ECLSK data.

Random Forests and Variable Importance

A single regression tree yields a rather noisy fit to the data. That is, minor changes to the data can result in drastic changes to the predicted values for certain units. *Bagging*, short for *bootstrap aggregating*, helps to reduce the variability of tree-based predictions by averaging over a collection of fits produced on bootstrap samples of the data. A *random forest* (RF) is a collection of B bagged trees (i.e., B trees grown on bootstrap samples) where, importantly, at each split, a random sample of m predictors, where $m << p$, the total number of predictors, is used in determining the split. Randomly sampling predictors reduces the pairwise correlation between trees, which further reduces variability. The *out-of-bag* (OOB) cases for replication b are the cases that were not part of the bth bootstrap sample. Predicted values are calculated as averages of predicted values across the B trees.

Breiman (2001) introduced the permutation measure of variable importance. For each predictor X_j, $j \in \{1, \ldots, p\}$, the RF permutation variable importance is the average change in OOB MSPE after randomly permuting (i.e., shuffling) the values of only X_j. If X_j is truly unrelated to the outcome, one would expect, on average, no difference between the accuracy of the predictions before and after permuting. On the other hand, if X_j is strongly related to the outcome, one would expect a large difference.

Let $B^{(s)}$ represent the set of OOB cases for tree s. Using notation motivated by Strobl and Zeiles (2008), the variable importance for predictor X_j for regression tree $s \in 1, \ldots, ntree$ may be defined as follows:

$$\mathrm{VI}^{(s)}(X_j) = \frac{\sum_{i \in B^{(s)}} \left(Y_i - \hat{Y}^{(s)}_{i,\pi_j}\right)^2}{\left|B^{(s)}\right|} - \frac{\sum_{i \in B^{(s)}} \left(Y_i - \hat{Y}^{(s)}_{i}\right)^2}{\left|B^{(s)}\right|}, \tag{2}$$

where $\hat{Y}_i^{(s)}$ is the predicted value for unit i before and $\hat{Y}_{i,\pi_j}^{(s)}$ is the predicted value after permuting the values of X_j. The variable importance for predictor X_j is the average, taken over trees in the random forest.

$$\overline{\text{VI}}(X_j) = \frac{1}{ntree} \sum_{s=1}^{ntree} VI^{(s)}(X_j). \tag{3}$$

The importance method described in Equations 2 & 3 will be referred to as *traditional RF permutation importance*. A problematic feature of the traditional permutation scheme is that variables may attain a high level of importance either for their relationship with the outcome Y, or for their relationships with other predictor variables that are related to the outcome. Traditional permutation importance cannot, therefore, disentangle marginal association between predictor and outcome from conditional association given the other predictors.

Conditional Permutation Importance

Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008) proposed *conditional permutation importance* to disentangle marginal and conditional relations. For each regression tree in the random forest and for a given predictor variable, X_j, the idea is to isolate the predictive power of X_j for Y by permuting X_j within fixed levels of the other predictors, $\mathbf{X} \setminus X_j$. For example, suppose there are three predictors, i.e., $\mathbf{X} = \{X_1, X_2, X_3\}$, such that X_1 is measured on a continuous scale and X_2 and X_3 are dichotomous. Assuming enough observed cases, one could randomly permute the values of X_1 within the four subgroups defined by the other two predictors: $X_2 = 1, X_3 = 1$; $X_2 = 1, X_3 = 0$; $X_2 = 0, X_3 = 1$; and $X_2 = 0, X_3 = 0$. This approach would break the association between X_1 and the outcome, as desired. However, by permuting X_1 within categories based on the levels of X_2 and X_3, any predictive power carried by X_1 based solely on its association with the other predictors would still be present after permutation. Because importance of X_1 is defined as the difference in OOB MSPE before and after permutation, the conditional importance of X_1 in this case would reflect the unique contribution of X_1 to the reduction in OOB MSPE after conditioning on X_2 and X_3: the portion not unique to X_1 would be differenced out.

With more than a handful of predictors, sparsity due to the "curse of dimensionality" renders the exact approach implausible. In place of exact conditioning, one could condition on any balancing score (cf. Rosenbaum & Rubin, 1983) for the target predictor. For example, if the target predictor (i.e., the one to be permuted) is dichotomous, it may be permuted within strata based on the propensity score estimated by regressing the target predictor on the remaining predictors. For continuous or multicategory variables, generalized versions of the propensity score could be used (e.g., Fong, Hazlett, & Imai, 2018). The approach suggested by Strobl et al. (2008) is to permute within subgroups based on the random forest regression tree splits. This latter approach has the benefit of being fully nonparametric.

Nonparametric Conditional Independence Testing

A test of conditional independence, as opposed to one that fails to distinguish between marginal and conditional association, is particularly important for variable selection in causal inference applications. Consider the example depicted in Figure 2, where T is the dichotomous assignment, Y^0 is the potential outcome under $T = 0$, X_1 is a treatment-only predictor, X_3 is an outcome-only predictor, X_6 is a confounding variable, and X_5 is a collider variable on the backdoor path from $T \leftarrow X_2 \rightarrow X_5 \leftarrow X_4 \rightarrow Y_0$. Following Equation 1, $\mathbf{X}_0 = \{X_3, X_4, X_6\}$. Note, however, that X_5 is marginally associated with Y_0 through X_4. Furthermore, the path $X_1 \rightarrow T \leftarrow X_6 \rightarrow Y_0$, which is blocked at T, becomes unblocked when T is conditioned on, rendering X_1 marginally associated with Y_0. Thus, in this case, if tests of marginal (as opposed to conditional) independence are used, two spurious variables, the instrument, X_1, and the collider, X_5, are incorrectly retained.

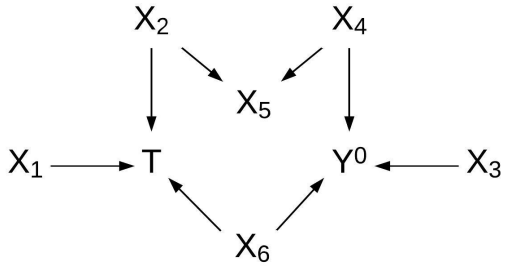

Figure 2. If tests of marginal, rather than conditional, independence are used to determine $\mathbf{X}_0$ with this directed acyclic graph, X_1 and X_5 are incorrectly retained.

Permutation Testing with RF Permutation Importance

Permutation testing provides a useful framework for nonparametrically testing individual variables based on RF importance measures. Rodenburg et al. (2008) used permutation of traditional RF permutation importance for variable selection and Altmann, Toloşi, Sander, and Lengauer (2010) summarized the approach, which can be described in four steps.

Step 1. Fit a random forest to calculate baseline importance for all predictors.

Step 2. Create R permuted copies of the predictor matrix, $\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^R$ by randomly shuffling rows.

Step 3. Run a random forest for each of the R replicates to generate R importance estimates for each variable.

Step 4. Compare each variable's baseline importance to its permutation distribution; Each variable that exceeds the (1 - α) percentile is retained.

The permutation approach operationalized in Steps 1 through 4 does not provide a test of conditional dependence for each predictor variable. There are two potential solutions to correct the procedure so that it provides tests of conditional independence between predictors and outcome, as desired. The first, proposed by Hapfelmeir and Ulm (2013), is to modify Steps 2 and 3 as follows, where p is the number of predictors:

Step 2′. For each j in $\{1, 2, \ldots, p\}$, create R copies of the predictor matrix by only shuffling the values of X_j.

Step 3′. For each j in $\{1, 2, \ldots, p\}$, run a random forest for each of the R replicates to generate p variable-specific permutation distributions.

This modification requires that $p \times R$ random forests be fit. For moderate to large values of p, especially when coupled with sample sizes that are not small, this approach is computationally intractable. The second solution, which, to my knowledge, has not been implemented or studied elsewhere, is to couple the original permutation approach (Steps 1 through 4) with a measure of conditional RF permutation importance. If conditional RF permutation importance is used in place of traditional RF permutation importance, shuffling all rows in Step 2 no longer invalidates the permutation variable selection procedure as a test of conditional independence because the

baseline importance of each predictor variable only reflects the improvement in OOB MSPE due to X_j given $\mathbf{X} \setminus X_j$.

I use this latter approach, and build up conditional RF permutation importance as follows:

1. Fit a random forest to the un-permuted data with *ntree* trees and compute OOB MSPE.
2. For each tree in the random forest,
   (a) For j in $\{1, 2, \ldots, p\}$, fit a regression tree with X_j predicted by $\mathbf{X} \setminus X_j$, and create a grouping factor, Z_j that assigns each unit a group number based on terminal node membership.
   (b) Let $\hat{Y}^{(s)}_{i,\pi_j|Z_j}$ denote the tree-predicted value for unit i in tree s after permuting the values of X_j within the levels of the factor Z_j. The tree-specific importance for each variable is determined as in Equation 2, except that $\hat{Y}^{(s)}_{i,\pi_j}$ is replaced by $\hat{Y}^{(s)}_{i,\pi_j|Z_j}$, reflecting that shuffling is now conditional on the grouping factor Z_j.
3. The overall conditional RF importance for X_j is the average over all trees, as in Equation 3.

The calculation of conditional RF importance requires significantly more computational time than the calculation of traditional RF importance. To reduce overall runtime, permutation testing with conditional RF importance is implemented only after first dropping predictors based on permutation testing with traditional RF importance. In this way, predictors that have no marginal association with the potential outcomes are dropped quickly, thereby reducing the dimension of the space over which conditional RF importance is run.

Simulation Studies

To my knowledge, the only other non-parametric approaches that have been studied for conditional independence testing within the $\mathbf{X}_Y$ framework are kernel smoothing and Bayesian networks under discrete mutual information. The kernel smoothing method, described in Persson et al. (2017) and available in package **CovSel** (Häggström, Persson, Waernbaum, & de Luna, 2015), is computationally demanding. I implemented kernel smoothing with the ECLSK data but the algorithm had not moved past the first of five "multistarts" after 24 hours running on a mid 2012 Mac Pro with 2 × 3.06 GHz 6-Core Intel Xeon processors and 40 GB RAM, so it was

abandoned. Gaussian Bayesian networks (GBNs) use mutual information based on Gaussian log-likelihoods for conditional independence testing, so there is an implicit assumption that relationships between variables are linear. Häggström (2018) proposed the use of Bayesian networks with tests of conditional independence based on discrete mutual information (Cover & Thomas, 2006). To avoid Gaussian distributional assumptions, the method requires that all continuous variables be discretized. Häggström (2018) found that discrete Bayesian networks resulted in accurate variable selection for large samples for both linear and nonlinear data-generation processes.

Variable Selection and Effect Estimation

Simulation studies were planned for four data-generation processes, each based on 100 replications, to probe the performance of the conditional RF importance permutation approach (RFC) relative to three other methods for high-dimensional variable selection: traditional RF importance under permutation (RFT), Gaussian Bayesian networks run in package **bnlearn** (Scutari, 2010; GBN), and Häggström's (2018) implementation of discrete Bayesian networks from packege **CovSelHigh** (Häggström, 2017; CSH). Both RFC and RFT were implemented in package **rfvarsel**. The first outcome evaluated for each of the four data-generation processes was each algorithm's accuracy in selecting the variables in the subset $\mathbf{X}_Y$ with data sets of size $N = 500$, $N = 2000$, and $N = 5000$. Then, given the 100 subsets actually selected by each algorithm for each sample size, four different methods to estimate average treatment effects were used to explore the impact of variable selection on bias and efficiency.

First, propensity score analysis was run with propensity scores estimated by generalized boosted regression models via package **twang** (Ridgeway, McCaffrey, Morral, Ann, and Burgette (2015); GBM). Second, propensity score analysis was run with propensity scores estimated by the covariate balancing propensity score via package **CBPS** (Fong, Ratkovic, Hazlett, & Imai, 2015; Imai & Ratkovic, 2014). For both propensity score methods, the average treatment effect was estimated by inverse propensity weighting. Third, Bayesian additive regression trees (BART) were run in package **bartMachine** (Kapelner & Bleich, 2016). For a given data set, BART was used to estimate the average causal effect by running a single BART model regressing the outcome on predictors and the treatment indicator. Potential outcomes were then imputed using

predicted values, and the average difference in imputed potential outcomes was reported as the estimate of the average causal effect. Fourth, targeted maximum likelihood estimation (TMLE) was used in conjunction with BART via package **tmle** (Gruber & van der Laan, 2012; van der Laan & Gruber, 2010). BART was used to estimate propensity scores and impute potential outcomes, which were then passed to TMLE to create a doubly robust estimator.

For variable selection, performance was measured by the proportion of replications for which variables in the target set, $\mathbf{X}_Y$, were correctly identified, and by the frequency with which variables outside the target set were incorrectly retained. For causal effect estimation, performance was measured by average absolute bias and simulation standard deviation, defined as follows,

$$\text{Average Absolute Bias} = \frac{1}{R}\sum_{r=1}^{R}|\hat{\tau}_r - \tau|$$

$$\text{Simulation Standard Deviation} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\tau}_r - \bar{\hat{\tau}}\right)^2}$$

where τ is the overall population average treatment effect, $\hat{\tau}_r$ is the estimate based on the rth replication, and $\bar{\hat{\tau}}$ is the average of the estimates over the R total replications.

Data Generation

The data-generating processes (DGPs) for Y^0 are displayed graphically in Figure 3; the DGPs for Y^1 were identical with the exception of a constant treatment effect of 0.5 units. For the linear case, X_1, X_2, and X_3 have a direct linear effect on both the exposure indicator T and the potential outcome, Y^0; X_4 and X_6 have a direct linear effect only on T; X_5 and X_7 have a direct linear effect only on Y^0; and X_8 through X_{100} are noise. X_1, X_6, X_7, and half the noise variables were simulated as dichotomous. The remaining variables were simulated as standard normal. The outcome model coefficients were set to 1.0, which yielded semi-partial correlations with the outcome of .2 for dichotomous variables and .4 for continuous variables, which correspond roughly with Cohen's (1988) guidelines for moderate effect sizes for Pearson correlations. Propensity score model coefficients were set to $log(2) \approx 0.69$, so that a one unit increase in each predictor was associated with a doubling of the odds of assignment to the exposed group.

The nonlinear DGP was identical with the exception that variables X_2 and X_3 influence the exposure and potential outcomes only through their interactions, denoted by filled circles in Figure 3. The third DGP is identical to the linear case with the exception that X_8 is a collider variable on the back door path from $T \leftarrow X_6 \rightarrow X_8 \leftarrow X_7 \rightarrow Y^0$. The fourth DGP combines the nonlinear and collider cases. For all DGPs in Figure 3, $\mathbf{X}_Y = \{X_1, X_2, X_3, X_5, X_7\}$.

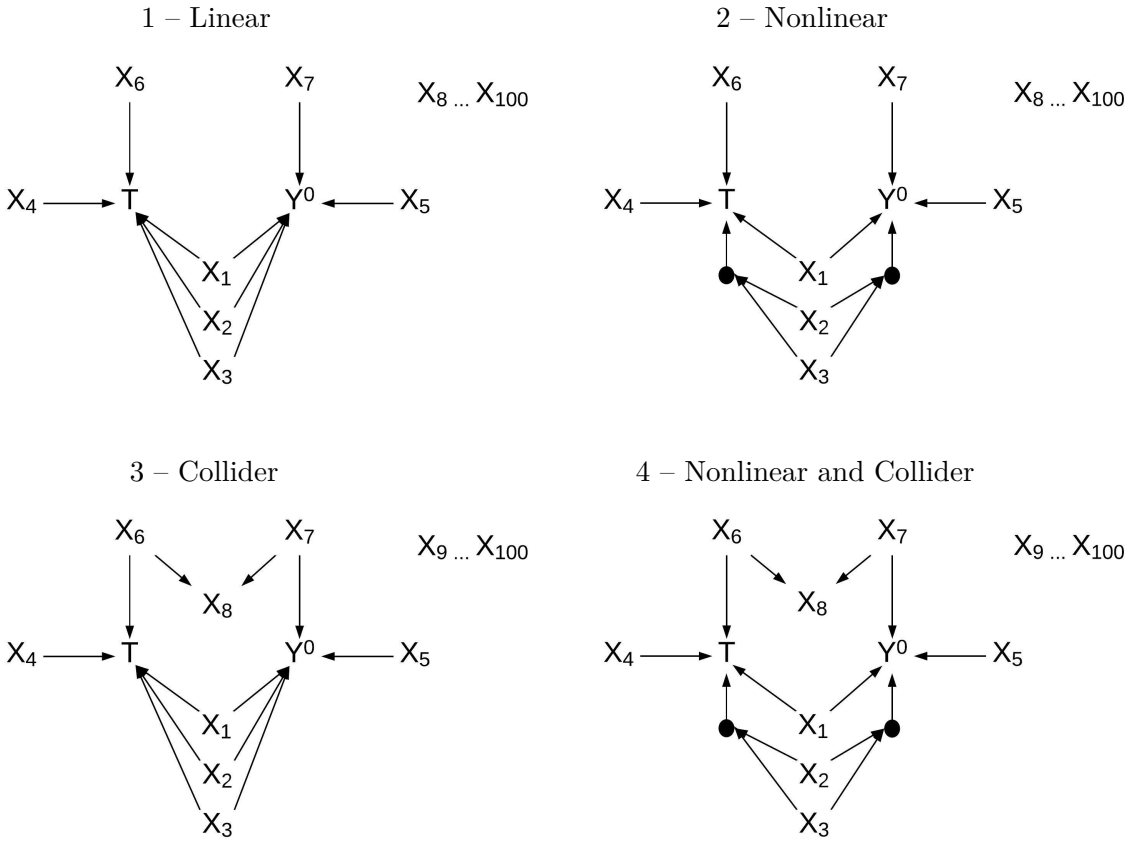

Figure 3. Data generating processes for Y^0 for the four simulation studies. Data generation for Y^1 is identical with the exception of a constant treatment effect.

Motivating considerations for the four data-generating processes are related to expected differences in performance across the methods with respect to variable selection and estimation accuracy and precision. Considerations for variable selection:

1. All methods are expected to perform well when the data-generating process is linear.
2. Methods based on RF importance should outperform GBNs and CSH in correctly identifying variables involved in nonlinear relationships with the potential outcomes. GBNs

use mutual information based on Gaussian log-likelihoods for conditional independence testing, so there is an implicit assumption that relationships between variables are linear. The CSH algorithm is based on the discrete version of mutual information, which does not assume linearity. However, because continuous variables must be discretized, information loss will likely result in less sensitivity to detect relationships that will be especially noticeable for small to moderate sample sizes.

3. RFT, which is the only method of the four that fails to distinguish marginal from conditional independence, will fail to eliminate collider variables because they are marginally associated with the potential outcomes.

Considerations for estimation precision and accuracy:

1. IVs and noise variables decrease precision of estimation when conditioned on; thus, all methods are expected to perform more poorly with respect to finite sample bias and efficiency when all 100 predictors are included, relative to when only the target set X_Y is included.

2. Conditioning on an IV or a collider variable has the potential to amplify bias; thus, large biases are expected in cases where the incorrect retention of an instrument or collider co-occurs with the incorrect omission of key confounding variables.

Results - Variable Selection

Variable selection results are shown in Table 4. The bars represent the proportion of 100 replications that included each variable from X_1 to X_8. Solid bars are used to denote variables that are part of the target subset X_Y; striped bars denote variables that were incorrectly retained. The numbers reported in the rows labeled "D" and "C" give the total number of noise variables (X_9 through X_{100}) retained across all replications for dichotomous and continuous predictors, respectively. For example, adding together the values 6 and 7, reported for GBN for the linear DGP, the combined rate of noise retention was 13/9200, where the number 9200 represents the total number (100 replications $\times$ 92 noise variables) of opportunities to retain noise variables for each cell of the simulation design.

The overall patterns of results are largely as expected. First, note that GBN performs remarkably well when relationships are linear (either with or without a collider variable), and very poorly when relationships are nonlinear. For the nonlinear DGPs, IVs X_4 and X_6 were selected with increasing frequency as sample size increased. This is expected because when the nonlinear terms X_2 and X_3 are incorrectly omitted and T is conditioned on, as required by Equation 1, paths from $\{X_4, X_6\} \rightarrow T \leftarrow \{X_2, X_3\} \rightarrow Y_0$ become unblocked, and the resultant associations are detected with more sensitivity at higher sample sizes. As expected, RFT failed to exclude the collider variable X_8. An increase was observed in the incorrect frequency of inclusion of X_4 and X_6 for RFT as sample size increases, likely due to marginal associations induced between the IVs and the potential outcomes due to conditioning on T.

The RFC method performed well across all scenarios. For the moderate and large sample size conditions, with $N = 2000$ and $N = 5000$, RFC correctly retained variables in X_Y in 100% of replications across all four data-generation scenarios. Comparing the results for the third and fourth DGPs across RFT and RFC, it is clear that the use of conditional RF importance permitted RFC to eliminate the IVs and collider that were incorrectly retained by RFT. For $N = 500$, RFT was more sensitive than RFC in correctly retaining important predictors in all scenarios (e.g., 89% vs 81% for X_1 with the linear DGP) [2]. RFC was uniformly more sensitive than CSH across all four DGPs, though CSH selected a lower proportion of IVs and colliders for some DGPs (e.g., 0% vs 5% for X_8 with the fourth DGP with $N = 2000$).

Results - Estimation

In addition to variables selected by GBN, RFT, RFC, and CSH, the cases of no variable selection (ALL) and correct selection of the target set X_Y were also considered. For each of the (6 variable selection methods $\times$ 3 sample size conditions $\times$ 4 DGPs) 72 cells of the simulation design, four methods were used to estimate the average causal effect for each of the 100 replications. Results for absolute bias and simulation standard deviations are shown in Figures 5 and 6, respectively. RMSE is not shown because it was dominated by the bias.

[2] This difference is due to the fact that the implementation of RFC involved a preliminary step wherein RFT was run first, so some predictors were tested twice. This relative loss of sensitivity could be avoided by leaving out the preliminary step at the cost of increased computational time, or by using a less stringent Type I error rate for the first round of testing.

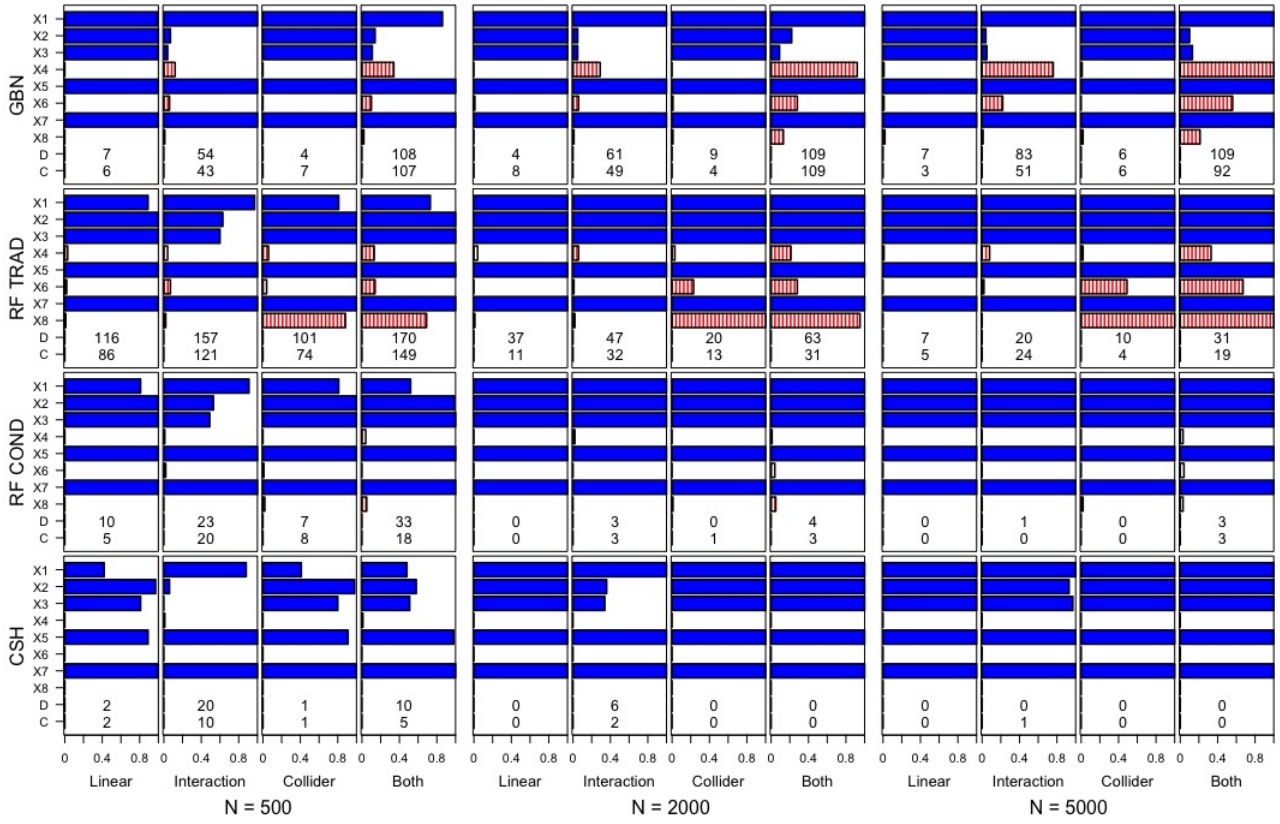

Figure 4. Bars represent Monte Carlo simulation study results on the proportion of 100 replications each variable was retained. Solid bars denote variables that were correctly retained (because they are in X_Y); striped bars denote variables that were incorrectly retained. For each cell, rows labeled D and C give the number of dichotomous and continuous noise variables incorrectly retained, respectively. GBN = Gaussian Bayesian networks, RF TRAD = traditional RF permutation importance under permutation, RF COND = conditional RF permutation importance under permutation, CSH = discrete Bayesian networks.

In every cell of the design and across all estimation methods, the absolute bias based on conditioning on all 100 predictors (ALL) was larger than the bias due to conditioning only on the target set, X_Y. Relative to no variable selection, RFT and RFC always helped. The results for GBN and CSH were mixed, with substantial bias reduction for linear scenarios but potential for poor bias reduction or even bias inflation when coupled with nonlinear DGPs. For $N = 500$, variable selection with CSH lead to bias increases in some cells due to lack of sensitivity. Performance with CSH improved markedly for larger sample size conditions.

The variability of estimators after variable selection, as measured by simulation standard deviation, contributed relatively little to the RMSE; nevertheless, there were some trends with

respect to variability. After variable selection with CSH, nonparametric estimators (i.e., BART, TMLE with BART, and propensity score analysis with GBM) were associated with larger variances. As an example, consider the interaction DGP at $N = 5000$. For that cell of the design, 91 of 100 replications led to estimates between 0.4 and 0.6 (the true value was 0.5) but the remaining nine estimates were larger than 1.0. In those nine cases, the CSH procedure had resulted in the incorrect omission of either X_2, X_3, or both. GBN, on the other hand, performed poorly with respect to variable selection for all 100 replications, which led to all 100 estimates being larger than 1.0, and, therefore, a tight spread around a biased value.

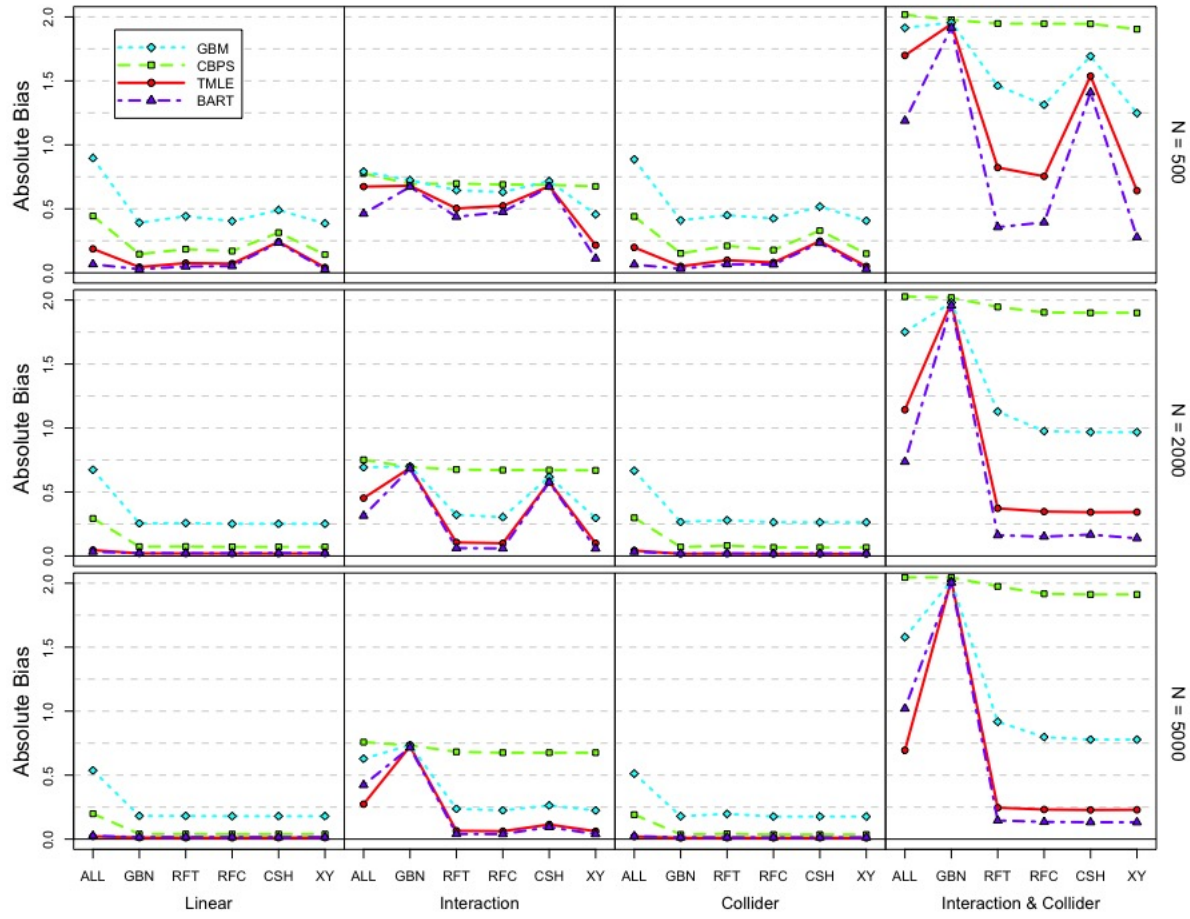

Figure 5. Each point represents the simulation-based average absolute bias for a given sample size, variable selection method, and estimation method. **Variable selection methods**: ALL = no variable selection, GBN = Gaussian Bayesian networks, RFT = traditional RF permutation importance under permutation, RFC = conditional RF permutation importance under permutation, CSH = discrete Bayesian networks, XY = correct selection of the variables in X_Y.

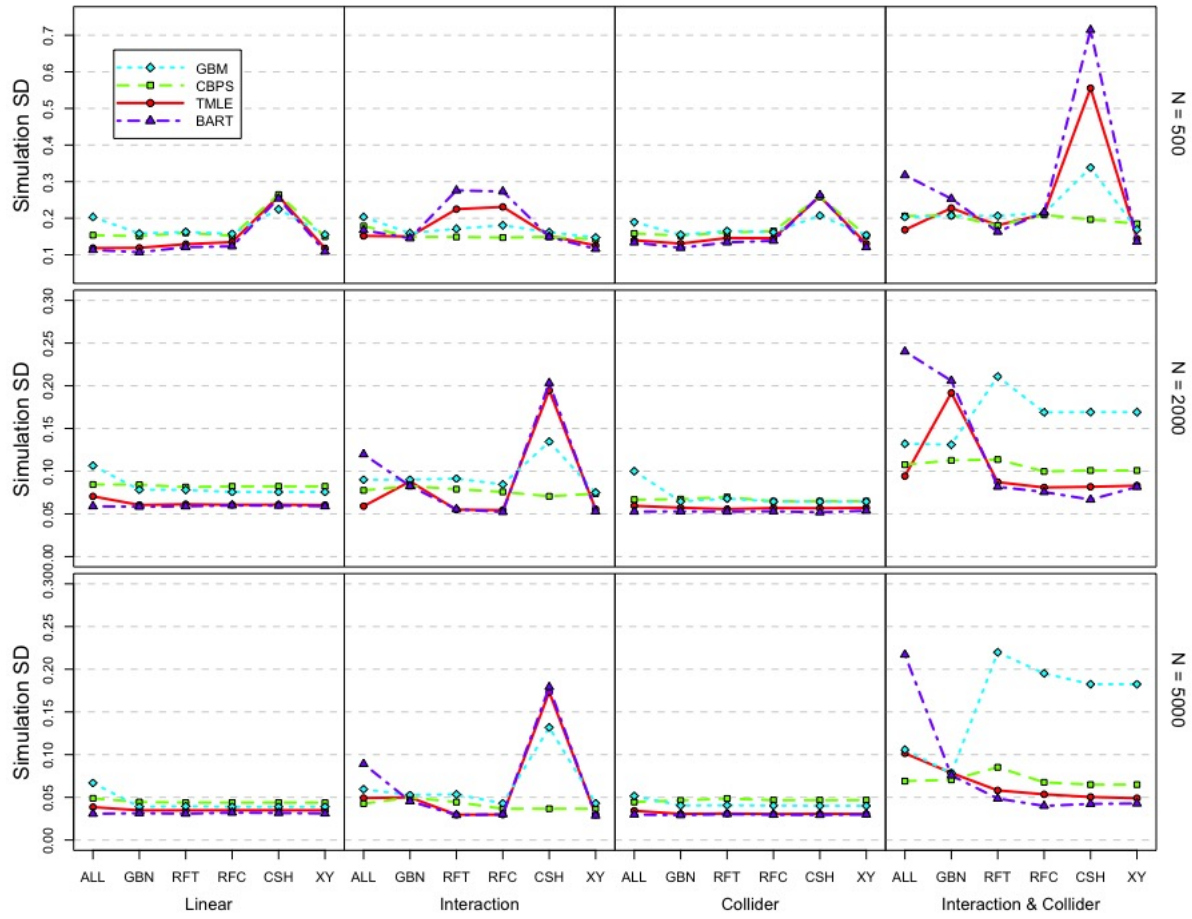

Figure 6. Each point represents the simulation standard deviation for a given sample size, variable selection method, and estimation method. ALL = no variable selection, GBN = Gaussian Bayesian networks, RFT = traditional RF permutation importance under permutation, RFC = conditional RF permutation importance under permutation, CSH = discrete Bayesian networks, XY = correct selection of the variables in X_Y.

Case Study

In this section, variable selection methods are applied to ECLSK data to estimate the average causal effect of exposure to special education services on math achievement in fifth grade. The data set, which is described in greater detail in Keller and Tipton (2016), includes variables motivated by Morgan, Frisco, Farkas, and Hibel (2010), who examined the effect of student exposure to special education services on later social and academic outcomes. There are 429 exposed cases and 6933 comparison cases; the outcome variable and 34 potential predictor variables are summarized in Table A1. Note that these analyses are mean to be illustrative; resultant estimates should not be interpreted as robust estimates of the causal effect.

Results - Variable Selection

Variable selection results are shown in Table 1. Of 34 potential predictors, 12 were retained by GBN, 10 by RFT, 8 by RFC, and 5 by CSH. The five variables selected by CSH were also selected by all other methods. Because these data were not simulated, the true value of the treatment effect is not known. Nevertheless, patterns in the variable selection results correspond with expectations based on theory and simulation results. For example, RIRT (kindergarten reading score) was dropped by all methods except RFT. Recall that RFT is the only method of the four that fails to discriminate between conditional and marginal dependence. The marginal linear relationship between RIRT and MIRT (kindergarten math score) is quite strong, with an estimated Pearson correlation of .71. It is plausible that the impact of RIRT on the outcome, 5th grade math score, although marginally very strong, was conditionally weak after accounting for MIRT, which would explain why it was dropped by the other algorithms. Another example involves P1FIRKDG (an indicator for first time in kindergarten) and P1AGEENT (child's age when starting kindergarten); both were selected by the RF procedures but not by GBN nor by CSH. Based on exploratory linear regression modeling, it is found that P1FIRKDG and P1AGEENT are involved in highly significant and strong two-way interactions. The forest-based methods may have picked up on these variables because of their strong nonlinear relationships with the outcome.

Results - Estimation

Average causal effect estimates and 95% confidence intervals are displayed in Figure 7. For GBM and CBPS, 95% confidence intervals are based on sandwich standard errors from package **survey** (Lumley, 2004); for TMLE, asymptotic intervals produced by the **tmle** package are reported; for BART, intervals were calculated using a nonparametric bootstrap (percentile method) based on 50 replications. The naive, unadjusted estimate of the average causal effect of special education services on fifth grade math achievement is -19.2 points, with 95% confidence interval given by (-21.5, -17.0).

Adjusting for all 34 covariates drastically and significantly reduced the estimated average effect to between -5.4 and -1.8, depending on the estimation method, with three of the four 95%

Variable	GBN	RFT	RFC	CSH
DEMOGRAPHIC				
GENDER	X	X	X	X
WKWHITE	X			
WKSESL	X	X	X	X
ACADEMIC				
RIRT		X		
MIRT	X	X	X	X
P1FIRKDG		X	X	
P1AGEENT		X	X	
apprchT1	X	X	X	X
P1HSEVER	X			
SCHOOL COMPOSITION				
S2KMINOR	X			
FAMILY CONTEXT				
P1FSTAMP	X	X	X	
P1HMAFB	X			
HEALTH				
C1FMOTOR	X	X	X	X
wt_ounces	X			
PARENT RATING				
P1SOLVE	X	X		
TOTAL	12	10	8	5

Table 1
Subset of ECLSK variables retained by at least one variable selection method. GBN = Gaussian Bayesian networks, RFT = traditional RF permutation importance under permutation, RFC = conditional RF permutation importance under permutation, CSH = discrete Bayesian networks. Boldface variable names indicate dichotomous predictors.

confidence intervals excluding zero and, thus, still suggesting a significant negative impact of exposure to special education services on 5th grade math achievement. Conditioning only on variables selected by GBN gave estimates between -5.0 and -3.6, also with three of four intervals excluding zero. Conditioning on variables selected by CSH yielded estimates between -6.4 and -3.8, with all intervals excluding zero. The use of RFT lead to estimates between -2.4 and 0.1, with only one interval that excluded zero. Estimates based on RFC were between -3.8 and -1.6 with two of four intervals excluding zero.

There are no randomized experimental results to use as benchmarks for comparison because of obvious ethical issues. Without recourse to the true causal effect it is not possible to say which methods were most unbiased. Nevertheless, it is interesting to note that the two RF-based methods tended to produce estimates closer to zero than any of the other methods, including conditioning on all 34 available predictors.

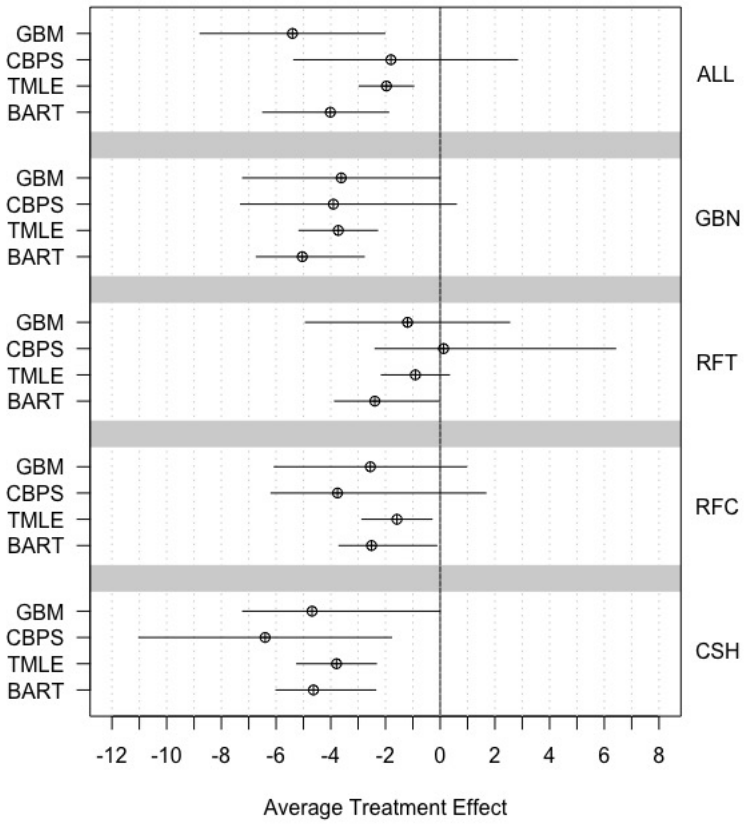

Figure 7. ECLSK case study results including average treatment effect estimates (points with cross-hairs) and 95% confidence intervals for the effect of special education services on math achievement in fifth grade. For effect estimation: GBM = propensity score analysis via generalized boosted modeling, CPBS = propensity score analysis via covariate balancing propensity score, TMLE = targeted maximum likelihood, BART = Bayesian additive regression trees. For variable selection: ALL = no variable selection, GBN = Gaussian Bayesian networks, RFT = permutation testing with traditional RF permutation importance, RFC = permutation testing with conditional RF permutation importance, CSH = discrete Bayesian networks.

Discussion

By using conditional RF importance in a permutation testing approach, one gets the best of both worlds: (a) a non-parametric variable selection algorithm that automatically handles nonlinear relationships and (b) one that tests the null hypothesis of *conditional* independence that is typically desired for causal applications. RF permutation methods were compared with Gaussian Bayesian networks and discrete Bayesian networks (CSH) through simulation and in a case study. Unlike GBN, which may fail to detect nonlinear relationships no matter the sample size, CSH is model free and has ideal properties in the limit. What the simulation results underscore, however, is that for small to moderate sample sizes, RF-based algorithms can be more

powerful, especially for detecting nonlinear relationships.

A general limitation with any variable selection method is related to the potential to drop weak confounding variables due to lack of sensitivity. This could be problematic in a situation in which there are many very weak confounders that, in aggregate, reduce a non-trivial amount of bias, but individually are too weak to be detected. The simulations show that sensitivity increases as sample size increases, as expected. Nevertheless, future work that more specifically addresses sample size and effect size considerations would be useful.

There are two limitations associated specifically with RF permutation methods for variable selection. First, RFT and RFC take more time to run than GBN or CSH because permutation approaches are computationally expensive. Second, as is typically the case with nonparametric methods, if parametric and functional form assumptions are tenable, methods such as GBN that capitalize on those assumptions will have slightly better finite-sample properties. This second point is not a strong limitation, however, because data-driven variable selection, especially with high-dimensional data, is typically used precisely because there is no strong theory regarding how variables are related. Furthermore, the consequences of incorrectly making strong functional form assumptions can be dire, as seen in the biases associated with GBN in nonlinear scenarios.

As expected, RFC correctly eliminated collider variables and instrumental variables that were incorrectly retained by RFT. The implications for bias, however, were generally modest, with relatively small gains in accuracy for RFC over RFT. From a causal perspective, the inclusion of a collider variable (like X_8 in simulations 3 and 4) will only cause a problem in terms of identification of the average causal effect if the backdoor path from T to the potential outcome is not blocked some other way. Given that ignorability was assumed with the complete set of predictor variables, RFT will also identify the variable that joins the collider to the potential outcome (X_7 in simulations 3 and 4), thereby blocking the confounding effect of including the collider. Differences in finite sample bias associated with RFC and RFT may be more drastic in situations in which ignorability is not satisfied given the observed pretreatment variables $\mathbf{X}$, but such cases were not investigated; it would be an interesting topic for future research.

With respect to the simulation standard deviations, shown in Figure 6, the simulation SDs appear to be tied to the variable selection results via the bias; higher variability occurs in tandem with higher bias and poorer variable selection accuracy. This raises questions about the need to

account for variability due to the variable selection procedure in confidence intervals and standard errors. The most straightforward methods to do so involve resampling, which, unfortunately, compounds computational time.

Note that RFC is an *explicit* method for variable selection. That is, the RFC procedure is designed be used as a pre-processing step to select important variables for further analyses. This is in contrast to methods that select variables *implicity* by upweighting important variables and downweighting unimportant ones on the fly. Some examples of methods that do implicit variable selection are penalized versions of regression such as ridge, lasso, and elastic net (Hastie et al., 2009) and all methods based on regression trees, which, as described in Appendix A, only make use of variables that maximize prediction when splitting. BART and GBM are both based on regression trees and, therefore, do variable selection implicitly by only splitting on important variables. Nevertheless, the use of RFT and RFC for variable selection always led to less bias than using the full set of variables with BART and GBM for the simulated data sets examined herein. As is the case with any simulation study, however, it is not appropriate to make strong generalizations to data scenarios that are very different.

Expert substantive knowledge and results from prior literature should take precedence in guiding the selection of important conditioning variables in an attempt to satisfy the ignorability assumption. However, there are bound to be variables, a large majority in some cases, for which the theory is weak. It is precisely for variables for which the theory is weak that questions about conditional independence and functional form relationships are likely to be unanswered. Simulation results discussed herein confirmed that dropping instrumental variables, colliders, and spurious noise variables prior to causal effect estimation improved finite sample behavior. The RFC algorithm performed well in both an absolute sense and relative to other methods for data-driven variable selection across a variety of conditions. Importantly, no tuning is required; RFC handles nonlinear relationships and interactions in its naive implementation.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*, 1340 – 1347.

Austin, P. C., Grootendorst, P., & Anderseon, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*, 734–753.

Battacharya, J., & Vogt, W. (2007). *Do instrumental variables belong in propensity scores?* (Technical Working Paper No. 343). National Bureau of Economic Research.

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

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

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149–1156.

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

Cover, T., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

de Luna, X., Waernbaum, I., & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, *98*, 861–875.

Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics*, *12*, 156 – 177.

Fong, C., Ratkovic, M., Hazlett, C., & Imai, K. (2015). Cbps: Covariate balancing propensity score [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=CBPS` (R package version 0.10)

Gruber, S., & van der Laan, M. J. (2012). tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, *51*(13), 1–35. Retrieved from `http://www.jstatsoft.org/v51/i13/`

Häggström, J. (2017). Covselhigh: Model-free covariate selection in high dimensions [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=CovSelHigh` (R package version 1.1.1)

Häggström, J. (2018). Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*, *74*, 389–398.

Häggström, J., Persson, E., Waernbaum, I., & de Luna, X. (2015). CovSel: An R package for covariate selection when estimating average causal effects. *Journal of Statistical Software*, *68*, 1–20.

Hahn, J. (1996). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*, 315–331.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics*, *86*, 73–76.

Hapfelmeir, A., & Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics and Data Analysis*, *60*, 50–69.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, New York: Springer.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B*, *76*, 243–263.

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, *70*(4), 1–40. doi: 10.18637/jss.v070.i04

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, *41*, 326-348.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*, 236–254.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In P. Grunwald & P. Spirtes (Eds.), *Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence.* Corvallis, OR: AUAI.

Persson, E., Häggström, J., Waernbaum, I., & de Luna, X. (2017). Data-driven algorithms for

dimension reduction in causal inference. *Compuational Statistics & Data Analysis*, *105*, 280–292.

Ridgeway, G., McCaffrey, D., Morral, A., Ann, B., & Burgette, L. (2015). twang: Toolkit for weighting and analysis of nonequivalent groups [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=twang` (R package version 1.4-9.3)

Rodenburg, W., Heidema, A. G., Boer, J. M. A., Bovee-Oudenhoven, I. M. J., Feskens, E. J. M., Mariman, E. C. M., & Keijer, J. (2008). A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes. *Physiological Genomics*, *33*, 78âĂŞ-90.

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

Scutari, M. (2010). Learning bayesian newtorks with the bnlearn R package. *Journal of Statistical Software*, *35*, 1-22.

Steiner, P. M., Cook, T. D., Li, W., & Clark, M. H. (2015). Bias reduction in quasi-experiments with little selection theory but many covariates. *Journal of Research on Educational Effectiveness*, *8*, 552–576.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*, 213–236.

Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*, *4*, 2193–3677.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*.

Strobl, C., & Zeiles, A. (2008). Danger: High power! – exploring the statistical properties of a test for random forest variable importance. In P. Brito (Ed.), *Compstat 2008 – proceedings in computational statistics* (Vol. II). Heidelberg, Germany: Physica Verlag.

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=rpart` (R package version 4.1-9)

Tourangeau, K., Nord, C., Lé, T., Pollack, J. M., & Atkins-Burnett, S. (2006). Early childhood

longitudinal study, kindergarten class of 1998-1999 (ecls-k), combined user's manual for the ecls-k fifth-grade data files and electronic codebooks (nces 2006-032) [Computer software manual]. Retrieved from `http://www.nces.ed.gov/ecls`. (U.S. Department of Education. Washington, DC: National Center for Education Statistics)

van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*, Article 17.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). New York: Springer.

Wooldridge, J. (2009, July). *Should instrumental variables be used as matching variables?* (Tech. Rep.). East Lansing, MI: Michigan State University.

Appendix A

Table A1

Table A1

Variable Names, Descriptions, and Standardized Mean Differences (d) and Variance Ratios (r) by Special Education Status

Variable Name	Description of Variable	Values	*d*	*r*
DEMOGRAPHIC				
GENDER	Male	0, 1	0.38	0.88
WKWHITE	White	0, 1	0.17	0.79
WKSESL	Socioeconomic Status	[-4.8, 2.8]	-0.29	0.89
ACADEMIC				
RIRT	Kindergarten Reading Score	[23.17, 139.36]	-0.65	0.53
MIRT	Kindergarten Math Score	[11.9, 99.0]	-0.71	0.77
S2KPUPRI	Public School	0, 1	0.44	0.25
P1EXPECT	Parental Expectations	Integers 1–6	-0.32	1.22
P1FIRKDG	First-Time Kindergartener	0, 1	-0.41	3.26
P1AGEENT	Child's Age at K Entry (Months)	[54, 79]	0.08	1.08
apprchT1	Approaches to Learning Rating	Integers 1–4	-0.70	1.20
P1HSEVER	Attended Head Start	0, 1	0.19	1.42
chg14	Ever Changed Schools	0, 1	0.02	1.09
SCHOOL COMPOSITION				
avg_RIRT	Reading IRT	[27.9, 80.0]	-0.23	0.79
avg_MIRT	Math IRT	[16.1, 66.1]	-0.18	0.82
avg_SES	SES	[-2.2, 2.5]	-0.16	0.88
avg_apprchT1	Approaches to Learning	[1.5, 4.0]	-0.14	0.80
S2KMINOR	Percent Minority Students	Integers 1–5	-0.20	0.77
FAMILY CONTEXT				
P1FSTAMP	Received Food Stamps	0, 1	0.12	1.26
ONEPARENT	One-Parent Family	0, 1	0.13	1.22
STEPPARENT	Stepparent Family	0, 1	0.05	1.19
P1NUMSIB	Number of Siblings	[0, 10]	0.16	1.17
P1HMAFB	Mother's Age at First Birth	Years [12, 45]	-0.26	1.00
WKCAREPK	Nonparental Pre-K Child Care	0, 1	-0.07	1.14
HEALTH				
P1EARLY	Number of Days Premature	[0, 112]	0.19	2.05
wt_ounces	Birth Weight (Ounces)	[17, 214]	-0.11	1.24
C1FMOTOR	Fine Motor Skills	Integers 0–9	-0.63	1.27
C1GMOTOR	Gross Motor Skills	Integers 0–8	-0.43	1.54
PARENT RATING OF CHILD				
P1HSCALE	Overall Health	Integers 1–5	0.12	1.17
P1SADLON	Sad/Lonely	Integers 1–4	0.10	1.32
P1IMPULS	Impulsive	Integers 1–4	0.41	1.55
P1ATTENI	Attentive	Integers 1–4	0.72	1.45
P1SOLVE	Problem Solving	Integers 1–4	0.68	1.55
PSPRONOU	Verbal Communication	Integers 1–4	0.86	1.51
P1DISABL	Child has Disability	0, 1	0.82	2.38
EXPOSURE VARIABLE				
F5SPECS	Special Education Services	0, 1	NA	NA
OUTCOME VARIABLE				
C6R4MSCL	Fifth Grade Math Score	[50.9, 170.7]	-0.77	1.40

Appendix B

A Regression Tree Example with ECLSK Data

In the ECLSK data set, there are 429 exposed cases (i.e., students that received special education services) and 6933 comparison cases (i.e., students that did not receive special education services). Here a regression tree will be fit to the 429 exposed cases using package **rpart** in R. The first step is to install and then load the **rpart** package.

```
R> install.packages("rpart")
R> library(rpart)
```

Next, prepare the data. `eclsk` is the name of the data frame that contains the outcome, `C6R4MSCL`, the exposure indicator, `F5SPECS`, and 34 additional predictors. The data are then subsetted to only use the exposed cases (i.e., cases with `F5SPECS` value of 1).

```
R> eclsk_exposed <- eclsk[which(eclsk$F5SPECS == 1), ]
```

Then, since this will result in a subset for which the `F5SPECS` variable is 1 for all cases, that variable is dropped from the data frame.

```
R> eclsk_exposed <- eclsk_exposed[, -which(names(eclsk) == "F5SPECS")]
```

Now the regression tree algorithm may be run.

```
R> rpart1 <- rpart(formula = C6R4MSCL ~ .,
R>                 data = eclsk_exposed,
R>                 control = rpart.control(minsplit = 50, cp = .005))
```

The period on the right side of the formula specifies that the outcome, `C6R4MSCL`, is to be predicted by all remaining variables in the data frame. The `minsplit` value is the minimum number of cases required in a node for a split to be attempted. Thus, nodes with fewer than `minsplit` cases are terminal. Setting `minsplit` to a larger value is one way to prevent overfitting. The `cp` value is a complexity parameter that saves computational time by not considering splits that decrease overall misfit by `cp`. Setting `cp` to a lower value will trigger a more fine-grained cross-validation table at the expense of more computational time.

The cross validation plot based on the run shows that `cp` = .019 is associated with the lowest cross-validated error. The dashed horizontal line is drawn one SE above the minimum of the curve. In practice, especially with larger data sets, better predictive performance may be attained by selecting the first value of `cp` that is below the horizontal line, even if it is not the minimum.

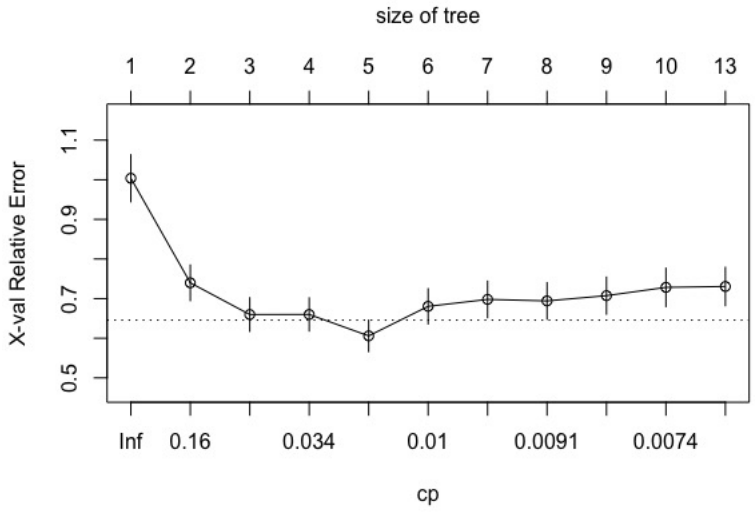

Figure B1. Cross validation plot for the regression tree error for ECLSK exposed cases

Having selected `cp` = .019, a pruned tree is then created. The output follows.

```
R> (rpart2 <- prune(rpart1, cp = .019))


n= 429


node), split, n, deviance, yval
      * denotes terminal node


 1) root 429 325357.50 131.6455
   2) MIRT< 31.465 332 188005.50 123.8792
     4) MIRT< 19.1725 66  25745.05 105.0488 *
     5) MIRT>=19.1725 266 133051.10 128.5515
      10) WKSESL< -0.275 103  44190.17 120.4778 *
```

```
      11) WKSESL>=-0.275 163  77904.28 133.6533 *
   3) MIRT>=31.465 97  48791.18 158.2267
     6) RIRT< 45.0375 67  27939.38 151.0906 *
     7) RIRT>=45.0375 30   9819.97 174.1640 *
```

The first line of output gives the sample size. The second line gives a key for interpreting the remaining output. Node number comes first, followed by the name and position of the variable on which a split was made, followed by the sample size remaining in each "daughter" node, followed by each node's contribution to the deviance, followed by the predicted value for the node, followed by an asterisk if the node is terminal. The root node represents the data before any splitting. Thus, the mean of 131.6 reported for the root node is the grand mean of the outcome variable for the 429 exposed cases. The deviance based on using the grand mean as the predicted value for all 429 cases is 325,358. The deviance for the final pruned tree can be calculated by summing each terminal node's contribution to the deviance: 25,745 + 44,190 + 77,904 + 27,939 + 9,820 = 185,598.

Package **rpart.plot**, which must be installed and loaded, will be used to plot the regression tree. The function `prp` creates the plot.

```
R> install.packages("rpart.plot")
R> library(rpart.plot)
R> prp(x = rpart4, type = 1, extra = 1,
      main = "Regression Tree for ECLSK Treated Cases")
```

The plot offers a succinct visual display of the information summarized in the output. The first split was on the kindergarten math pretest score, MIRT, at a score of 31.5. The second split was also made on MIRT at a score of 19.2, thus modeling a quadratic relationship (MIRT^2). The next split was made on the RIRT variable at a value of 45.0, thus modeling an interaction between MIRT and RIRT. The final split is on the socioeconomic status variable WKSESL, at a value of -0.3. The tree-predicted value for the 30 special education students that scored higher than 31 on the kindergarten math exam and higher than 45 on the kindergarten reading exam is 174; whereas, the tree-predicted value for the 66 special education students that scored lower than 19 on the kindergarten math exam is 105. Other nodes may be interpreted similarly.

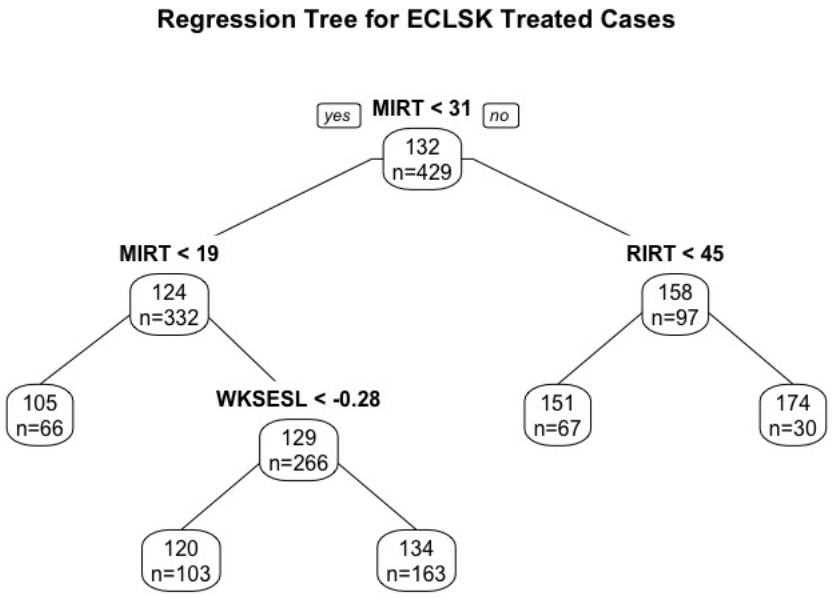

Figure B2. Regression tree plot for ECLSK exposed cases after pruning