Heterogeneous Subgroup Identification in Observational Studies

Jianshen Chen and Bryan Keller

College Board and Columbia University

Manuscript Under Review

Author Note

The authors contributed equally.
Abstract

Despite the overwhelming focus on the overall average treatment effect in the methodological and statistical literature, in many cases the efficacy of an educational program or intervention might vary based on unit background characteristics. The identification of subgroups for which a particular educational intervention is particularly effective or, on the other hand, has no effect or is possibly harmful, may have important practical implications, especially in terms of allocation of resources. We propose a five-step approach using propensity score matching and regression trees to identify subgroups with heterogeneous treatment effects in observational studies. Results of a Monte Carlo simulation study indicate that the proposed approach can accurately identify heterogeneous subgroups in observational studies while maintaining Type I error rate. In a case study with ECLSK data we find the effect of special education services varies based on a student’s kindergarten reading achievement score and whether the student attends public or private school.
Heterogeneous Subgroup Identification in Observational Studies

Introduction

Much effort has been devoted to the identification and estimation of overall average treatment effects with observational data through conditioning strategies such as propensity score analysis, nonparametric regression estimation, or doubly robust combinations of selection and outcome models. Despite the overwhelming focus on the overall average treatment effect in the methodological and statistical literature, in many cases the efficacy of an educational program or intervention might vary based on unit background characteristics. The identification of subgroups for which a particular educational intervention is particularly effective or, on the other hand, has no effect or is possibly harmful, may have important practical implications, especially in terms of allocation of resources. Recent advances in statistics and causal inference have focused on understanding such conditional average treatment effects. Nevertheless, methods for detecting and describing treatment effect heterogeneity with observational data have not entered the mainstream in education or behavior sciences. In this paper we describe an exploratory approach for identification of treatment effect heterogeneity based on propensity score matching, a method that is familiar to most educational researchers. We then use regression trees, a simple nonparametric regression technique, to identify subgroups based on background characteristics that best explain how units respond to treatments.

The organization of this paper is as follows. After a brief review of propensity score analysis and regression trees, we describe the proposed approach for subgroup identification. Next, we run a Monte Carlo simulation study to examine (a) the sensitivity and specificity of the method for detecting treatment effect heterogeneity and (b) the accuracy of estimated conditional average effects. We then apply the approach in a case study to search for heterogeneous subgroups in the context of an observational study designed to examine the average causal effect of special education on math achievement in fifth grade. We conclude with some discussion.
Propensity Score Analysis

In this paper, we consider the simple case of a binary treatment condition, in which individuals self-select into a “treatment group” or a “control group”. Rosenbaum and Rubin (1983) defined the propensity score as the conditional probability that a subject is assigned to the treatment group, given a vector of covariates measured prior to the treatment selection. Following the general notation of the Rubin causal model (Rubin, 1974), let $Y_i^1$ be the potential outcome for individual $i$ if the individual received the treatment, and $Y_i^0$ be the potential outcome for individual $i$ if the individual did not receive the treatment. The treatment indicator is denoted as $T_i$, where, for individual $i$, $T_i = 1$ if the individual was exposed to the treatment, and $T_i = 0$ if the individual was not exposed to the treatment. The propensity score for unit $i$, given a vector of observed covariates $X$ is be expressed as $e_i(X) = P(T_i = 1|X_i)$. The average treatment effect (ATE) for the population is defined as the average of the individual treatment effects:

$$ATE = E(Y_i^1 - Y_i^0) = E(Y_i^1) - E(Y_i^0).$$

The conditional average treatment effect is defined as the average treatment effect for units with a particular value, or ranges of values, on one or more covariates, $CATE = E(Y_i^1 - Y_i^0|X_i = x_i)$.

For identification of the ATE and CATE, propensity score techniques and other conditioning strategies rely on an important, multipart assumption known as strong ignorability (Rosenbaum & Rubin, 1983):

1. **ignorability**: the potential outcomes are independent of the treatment assignment given observed covariates $X$; that is, $\{Y^0, Y^1\} \perp T|X$,

2. **reliable measurement**: observed covariates $X$ have been reliably measured (Steiner, Cook, & Shadish, 2011), and

3. **positivity**: the propensity score for each unit lies strictly between zero and one; that is, $0 < e_i(X) < 1$ for all $i$. 
When the strong ignorability assumption holds, participants with the same propensity score have identical distributions on the set of covariates; that is, \( X \perp \perp T|e(X) \). Thus, by conditioning on the propensity score, rather than the full set of covariates, it is possible to restore covariate balance across treatment and control groups to what would have been expected under random assignment. Furthermore, under strong ignorability, the potential outcomes are independent of treatment assignment, \( \{Y^0, Y^1\} \perp \perp T|e(X) \), so that unbiased treatment effect estimates can be obtained.

Methods for conditioning on the propensity score include stratification, inverse weighting, matching and regression adjustment (see, e.g., Stuart, 2010 for a review). In this paper, we focus on propensity score matching because it stratifies the units into many groups, within which the conditional average treatment effect may be approximated. Various matching methods have been explored in the context of propensity score analysis such as nearest-neighbor matching, caliper matching, Mahalanobis distance matching and optimal matching (see e.g. Coca-Perraillon, 2007; Hansen & Klopfer, 2006). Optimal matching is “optimal” in the sense that each match can be revisited to minimize a global measure of distance between all matches. This is in contrast to greedy matching methods, which typically give different solutions based on different starting points (Rosenbaum, 1989). Optimal matching performs well when compared with other propensity score conditioning strategies (Kaplan & Chen, 2012; Steiner & Cook, 2013). Under optimal matching, units may be matched one-to-one, one-to-many or many-to-many. We use propensity score optimal full matching with many-to-many matching between the treatment group and the comparison group.

**Tree-based Methods**

**Regression Trees**

A *regression tree* is an algorithmic tool invented by Breiman, Friedman, Olshen, and Stone (1984) that models the relationship between an outcome variable, \( Y \), and predictors,
X_1, \ldots, X_p$, by iteratively splitting the units into subgroups based on predictor values. Every split creates two subgroups, called *nodes*, and any node that is unsplit is called a *terminal node*. Within each terminal node, the tree-predicted value for each member of the node is simply the mean outcome score for all units in the node. Thus, for unit $i$ in terminal node $t$, where $N_t$ represents the set of units in $t$, the tree-predicted value for unit $i$ is $\hat{Y}_i = \frac{1}{|N_t|} \sum_{i \in N_t} Y_i$. At each iteration, the mean squared prediction error, MSPE $= \sum_i \left( \hat{Y}_i - Y_i \right)^2$, is determined before splitting. Then, every possible split on every variable is considered, and the split that results in the largest decrease in MSPE is selected.

If left unchecked, regression trees would continue to split until each terminal node contained only one point, yielding a perfect fit to the data. To prevent overfitting of this sort, which would result in poor prediction accuracy, a stopping rule is imposed. The usual approach for determining the stopping rule is based on adding a penalty to the squared error based on the number of terminal nodes in the tree. This approach, referred to as *cost-complexity pruning*, is implemented in the `rpart` package (Therneau, Atkinson, & Ripley, 2015) in R (R Core Team, 2018), which we use to fit regression trees. The implementation of cost-complexity pruning requires an additional tuning parameter, which is typically selected through cross-validation.

Because regression trees model a response surface with discrete jumps, they can only roughly approximate a smooth surface, which results in bias. While it is possible to lower the minimum number of units per terminal node to get a better fit, this leads to overfitting and results in higher variance. More complex tree-based methods overcome this shortfall by pooling results based on many trees. Bootstrap aggregation (bagging), random forests, and boosted trees all benefit from the nonparametric flexibility of regression trees, but combine them in unique ways to improve the predictive performance.

The main advantage to using a single regression tree over other, more complex methods such as those mentioned above, is that the output from a single tree is easily interpretable because each split results in a partition on some variable. Thus, for
meaningfully segmenting a sample based on variable characteristics (which is our goal in identifying heterogeneous subgroups), single trees are a useful, if somewhat crude, tool.

**Random Forests**

Random forests were invented by Breiman (2001). The goal of random forests is to turn what would otherwise be a rather weak learner (a single tree) into a strong learner by repeatedly fitting individual regression trees and combining them in an intelligent way. Random forests are based on aggregated results from fitting regression trees to $B$ bootstrap samples of the data. A regression tree is grown for each bootstrap sample by recursively splitting until each terminal node is as small as possible without being smaller than a prespecified minimum node size. Importantly, each split uses a random sample of only $m$ predictors, where $m << p$, the total number of predictors. The resultant set of trees constitute the “random forest”. The out-of-bag (OOB) cases for replication $b$ are the cases that were not part of the $b$th bootstrap sample. Regression predictions are calculated as the average of the predicted values across the $B$ trees.

For our purposes, random forests are useful because they provide a quantitative summary of each variable’s importance in predicting the outcome. This metric proves extremely useful in winnowing down a large set of covariates (such as the 34 in our case study) to a more manageable number of important variables on which we can search for heterogeneous treatment effects. We use a permutation-based approach for variable selection with random forest variable importance described in Keller & Zhang, 2018, March.

**Identification of Subgroups with Heterogeneous Treatment Effects**

The CATE, $E(Y_i^1 - Y_i^0|X_i = x_i)$, essentially specifies a nonparametric regression of the individual treatment effects on the covariate set. If there is an interaction b
Randomized Trials

Effect heterogeneity is characterized by interactions between one or more covariates and the treatment variable. Regression trees handle interactions in their naïve implementation, so it is not surprising that tree-based methods have played a prominent role in the literature. In particular, a number of tree-based methods have been proposed to identify treatment effect heterogeneity in randomized experiments. Su, Tsai, Wang, Nickerson, and Li (2009) used regression trees with a customized loss function designed to measure the strength of treatment by subgroup interactions. Foster and colleagues (Foster, Taylor, & Ruberg, 2011; Foster, Taylor, Kaciroti, & Nan, 2015) used random forests to identify covariate regions for binary and continuous outcomes. Imai and Strauss (2011) fit regression trees followed by Bayesian models to estimate group-specific treatment effects.

Approaches for identifying effect heterogeneity in randomized experiments based on methods other than trees have been proposed as well. Imai and Ratkovic (2013) introduced an algorithm based on regularized support vector machines. Ding, Feller, and Miratrix (2016) developed a randomization approach to evaluate effect variations that are unexplained by observed covariates. In a 2017 issue of the Journal of Research on Educational Effectiveness devoted to variation in treatment effects in randomized trials, Bloom, Raudenbush, Weiss, and Porter (2017) used mixed effects models to study cross-site variation in program efficacy and Harding, Morris, and Hill (2017) demonstrated the use of principal stratification to identify local average treatment effects.

Observational Studies

Despite the aforementioned work in the context of randomized experiments, fewer studies have aimed to identify effect heterogeneity in observational study settings. Xie, Brand, and Jann (2012) proposed several methods to detect treatment by propensity score interactions, but did not consider heterogeneity based on individual covariates. Hsu, Small, and Rosenbaum (2013) and Hsu, Zubizarreta, Small, and Rosenbaum (2015) developed a
framework for sensitivity analysis in the presence of effect heterogeneity by identifying subgroups of matched-pairs formed using exact matching and leveraged the method for identification of heterogeneity. Exact matching is ideal because it requires no functional form or parametric assumptions; however, with many covariates or a handful of continuously measured covariates, exact matching is not feasible because the number of possible combinations of covariate values grows exponentially with the addition of each new covariate.

Wager and Athey (2018) derived asymptotic results for random forests and applied them to forests based on trees that are tuned to treatment effect estimation. A strength of their causal forest algorithm, which directly estimates CATEs, is the recourse to asymptotically-valid confidence intervals for treatment effect estimates. Like other methods that have been used to estimate individual treatment effects, such as Bayesian additive regression trees (BART; Hill, 2011), causal forests do not explicitly address identification of heterogeneous subgroups. The output must be further processed to identify groups for which the treatment is differentially effective.

The Proposed Method

In this paper, we describe and evaluate an exploratory approach for heterogeneous subgroup identification with observational data. First, optimal full propensity score matching is used to estimate individual treatment effects within matched strata. Second, covariates that are important in predicting variation in the estimated individual treatment effects are identified via a pruned regression tree fit. Criticisms have characterized the practice of subgroup identification as suffering from a lack of planned statistical approaches that are both valid and sufficiently powered (e.g., Assman, Pocock, Enos, & Kasten, 2000). To that end, we propose the systematic use of permutation-based Type I error rate studies to ensure valid inferences. Furthermore, we examine the power properties of the proposed method through simulations.
1. *Estimate propensity scores.* A logit model is often used, but nonparametric methods, such as random forests or boosted CART, align well with our approach because they automatically handle interactions.

2. *Stratify units.* Units are grouped into strata via many-to-many optimal matching based on estimated propensity scores, and the stratum-specific treatment effect estimate for each stratum is used as an estimate of the individual treatment effect for each unit in that stratum.

3. *Type I error rate permutation study.* The cost-complexity tuning parameter for the regression tree is selected based on a permutation-based study of Type I error rates for different values of the tuning parameter.

4. *Identify subgroups.* A single regression tree is fit to the stratum-specific treatment effects, using the covariates as predictors. Any subgroup splits identified by the regression tree are interpreted as evidence for effect heterogeneity on the variable or variables involved in the splits.

5. *Estimate subgroup-specific effects.* Perform many-to-many optimal matching again, subject to the additional constraint that matching takes place within heterogeneous subgroups identified in Step 4, and estimate average and conditional average treatment effects.

**Design of the Simulation Study**

We designed a Monte Carlo simulation study to (a) examine the Type I error rate and power of the subgroup identification approach in detecting treatment effect heterogeneity and (b) assess the accuracy of estimation of conditional average treatment effects based on those subgroups. Data generation is described next.

1. Covariates $x_1, x_2, \ldots, x_{10}$ were independently generated from the following distributions, which were selected to reflect the natural variety of predictor types
encountered in practice (i.e., continuous, dichotomous and count). Generated samples were of size \( n = 1000 \).

\[
x_1 \sim N(0, 1); \quad x_6 \sim Bernoulli(0.3)
\]
\[
x_2 \sim Poisson(2); \quad x_7 \sim N(-1, 3)
\]
\[
x_3 \sim Bernoulli(0.5); \quad x_8 \sim N(2, 2)
\]
\[
x_4 \sim N(0, 2); \quad x_9 \sim N(1, 0.8)
\]
\[
x_5 \sim Bernoulli(0.6); \quad x_{10} \sim N(2, 1)
\]

2. Let \( X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9}, x_{i10})' \) be the vector of covariate values for unit \( i \). The propensity score was modeled as follows,

\[
e_i(X) = \frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)},
\]

where the generating coefficients are \( \alpha = 0 \) and \( \beta = (0.6, 0.1, -0.3, -0.4, 0.2, -0.3, -0.2, 0.2, 0.5, 0.3) \). These coefficients range in magnitude from small to moderate, are both positive and negative, and yield sufficient propensity score overlap.

3. Values of the treatment assignment vector, \( T \), were determined by comparing the propensity score \( e_i(X) \) to a random uniform variable on \([0,1]\), \( U_i \), such that \( T_i = 1 \) if \( U_i \leq e_i(X) \) and \( T_i = 0 \) otherwise.

4. Potential outcomes were generated as follows,

\[
Y^1_i = \rho X_i + \gamma + \lambda I(x_{i1}) + \epsilon_{i1}
\]
\[
Y^0_i = \rho X_i + \epsilon_{i0},
\]

where \( \rho = (0.2, 0.1, 0.2, -0.1, -0.2, 0.2, -0.2, 0.1, 0.2, 0.1) \), \( \epsilon_1, \epsilon_0 \sim N(0, 1) \) and \( \gamma \), the
treatment main effect, was set equal to 1.0. The indicator was defined such that
\[ I(x_{i1}) = 1 \text{ if } x_{i1} > 0.5 \text{ and } I(x_{i1}) = 0 \text{ if } x_{i1} \leq 0.5. \]
That is, two heterogeneous subgroups are defined by covariate \( x_1 \) with the cutoff value at 0.5. The observed outcome \( Y_i \) was determined by the potential outcomes as follows:
\[
Y_i = Y_i^1T_i + Y_i^0(1 - T_i).
\]
5. Data = \( \{(Y_i, x_i, T_i), i = 1, \ldots, n\} \).
6. Replicate the above steps for \( R = 1000 \) replications.

The coefficient of the treatment by \( x_1 \) interaction, denoted by \( \lambda \) in Equation 3, was set to 1, 2, or 3. For those values, semi-partial correlations between the interaction term and the the outcome variable were, respectively, .19, .34, and .46, which correspond loosely to Cohen’s (1988) guidelines of .1, .3, and .5 for small, moderate, and large effect sizes for Pearson correlation. For the Type I error rate study, we set \( \lambda = 0 \). The purpose of the Type I error rate study is to determine a cost-complexity tuning parameter such that Type I error rate is held to \( \alpha = .05 \). Thus, we varied the value of the tuning parameter and, for each value, counted the number of times, out of 1000 replications, that the algorithm mistakenly identified heterogeneity when none existed.

For the power studies, the true value of the overall average treatment effect may be determined as follows:

\[
\text{ATE} = E[Y_i^1 - Y_i^0]
= E[\rho X_i + \gamma + \lambda I_i(x_1) + \epsilon_{i1}] - E[\rho X_i + \epsilon_{i0}]
= \gamma + \lambda E[I_i(x_1)]
= \gamma + \lambda(1 - \Phi(0.5))
= 1 + \lambda(0.31),
\]
where the second-to-last equality follows from the fact that \( x_1 \sim N(0, 1) \). Thus, ATE =
1.31, 1.62, and 1.93, for $\lambda = 1$, 2, and 3, respectively. Bias, simulation standard deviation, and mean squared error are used to measure the adequacy of the ATE estimators over the $R$ simulation replications as follows.

$$
\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} (\hat{\text{ATE}}_r - \text{ATE}(\lambda))
$$

$$
\text{SD} = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} (\hat{\text{ATE}}_r - \bar{\hat{\text{ATE}}}_r)^2}
$$

$$
\text{MSE} = \frac{1}{R} \sum_{r=1}^{R} (\hat{\text{ATE}}_r - \text{ATE}(\lambda))^2,
$$

where $\text{ATE}(\lambda) = 1 + \lambda(0.31)$.

The values of subgroup-specific CATEs depend on the cutoff values selected for $x_1$ for each replication, as well as $\lambda$ values. Recall that the data-generating cutoff value was set to be 0.5. However, because of sampling variability, for any given replication, the regression tree algorithm may identify a cutoff close to, but not identical to, 0.5. For example, suppose the cutoff selected by the regression tree was 0.4. In that case, the two subgroup-specific CATEs may be calculated as follows:

$$
\text{CATE}_{\{x_1 \leq 0.4\}} = E[Y^1 - Y^0|x_1 \leq 0.4] = 1
$$

$$
\text{CATE}_{\{x_1 > 0.4\}} = E[Y^1 - Y^0|x_1 > 0.4]
$$

$$
= E[Y^1 - Y^0|0.4 < x_1 \leq 0.5] + E[Y^1 - Y^0|x_1 > 0.5]
$$

$$
= \frac{Pr(0.4 < x_1 \leq 0.5)}{Pr(x_1 > 0.4)} * 1 + \frac{Pr(x_1 > 0.5)}{Pr(x_1 > 0.4)}(1 + \lambda)
$$

$$
= \frac{\Phi(0.5) - \Phi(0.4)}{1 - \Phi(0.4)} + \frac{1 - \Phi(0.5)}{1 - \Phi(0.4)}(1 + \lambda)
$$

$$
= 0.10 + 0.90(1 + \lambda).
$$

Thus, for $\lambda = 2$, $\text{CATE}_{\{x_1 > 0.4\}} = 2.80$; calculations follow the same logic for other cutoff
values. Let $c_r$ be the cutoff value for $x_1$ selected for replication $r \in \{1, \ldots, R\}$. For the CATE estimators, bias and mean squared error are defined as follows.

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} \left( \text{CATE}_r - \text{CATE}(\lambda, c_r) \right)$$

$$\text{MSE} = \frac{1}{R} \sum_{r=1}^{R} \left( \text{CATE}_r - \text{CATE}(\lambda, c_r) \right)^2,$$

where $\text{CATE}(\lambda, c_r)$ is determined as described in Equations 4 and 5.

**Results of the Simulation Study**

Cost-complexity tuning parameter values .01, .02 and .03 yielded Type I error rates of .74, .15 and .05, respectively. Thus, we set the cost-complexity parameter at .03 for subsequent analyses. Results of the power study are presented in Table 1, which displays the proportion of replications for which the method identified subgroups. When the treatment by covariate interaction was small (i.e., $\lambda = 1.0$), the correct covariate, $x_1$, was correctly identified as the only predictor of subgroup heterogeneity in only 144 out of 1000 replications. For moderate and large magnitudes (i.e., $\lambda = 2.0$ and $3.0$), the number of correct decisions improved to 728 and 923 out of 1000, respectively. For $\lambda = 1.0$, the average cut score on $x_1$ for the 144 replications for which the correct covariate was identified was 0.40, with a rather large standard deviation of 0.41; recall the data-generating cut score was 0.50. For the moderate heterogeneity condition with $\lambda = 2$, the average cut score of $x_1$ for the 728 relevant replications was 0.49, with a standard deviation of 0.18. For the large heterogeneity condition with $\lambda = 3$, the average cut score was 0.50, with a standard deviation of 0.10.

Table 2 summarizes the accuracy of estimators of ATEs and subgroup-specific CATEs through bias, simulation standard deviation, and mean squared error. For each estimator,
Table 1
Subgroup Identification Performance across 1000 Replications in the Simulation Study

<table>
<thead>
<tr>
<th>Case</th>
<th>$\lambda = 1$</th>
<th>$\lambda = 2$</th>
<th>$\lambda = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$ only (correct decision)</td>
<td>144</td>
<td>728</td>
<td>923</td>
</tr>
<tr>
<td>$x_1$ and another variable</td>
<td>5</td>
<td>35</td>
<td>54</td>
</tr>
<tr>
<td>Split but not on $x_1$</td>
<td>31</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>No subgroups identified</td>
<td>820</td>
<td>224</td>
<td>18</td>
</tr>
</tbody>
</table>

$\lambda$ is the heterogeneity coefficient described in Equation 3

there is a row of the table for the “Naive” version and the “Step (5)” version. The “Naive” versions of estimators are based directly on the regression tree output from Step (4), whereas the “Step (5)” estimators are based on stratified matching within subgroups, as specified in Step (5). It is also important to note that when the magnitude of the treatment by $x_1$ interaction was small (i.e., $\lambda = 1$), our approach only correctly identified heterogeneity about 14% of the time. Thus, estimates for $\lambda = 1$ in Table 2 are based on only 144 replications, so they are less stable than for the other two values of $\lambda$, which used 732 and 923 replications, respectively.

The overall ATE was estimated without bias across all three values of $\lambda$ and for both the naive and Step (5) estimators. As for the conditional effects, the naive estimators carried large and significant bias which increased as the strength of the treatment effect modification increased. In contrast, the Step (5) estimators, which matched propensity scores within subgroups, were close to unbiased with MSE values over an order of magnitude smaller than the naive method.

**Design of the Case Study**

The Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K; NCES, 2001), is a national longitudinal study focused on child development and early school experiences. The treatment effects of interest for this case study are related to the impacts of receiving special education services versus not receiving special education services on reading and mathematics scores measured during fifth grade (year 2004). The data are
Table 2

<table>
<thead>
<tr>
<th>Estimator</th>
<th>λ = 1</th>
<th>λ = 2</th>
<th>λ = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
</tr>
<tr>
<td>Cut-off value on $x_1$</td>
<td>-0.10</td>
<td>0.41</td>
<td>24.98</td>
</tr>
<tr>
<td>ATE (Naive)</td>
<td>-0.01</td>
<td>0.11</td>
<td>58.19</td>
</tr>
<tr>
<td>ATE (Step 5)</td>
<td>-0.01</td>
<td>0.11</td>
<td>58.54</td>
</tr>
<tr>
<td>$\text{CATE}_{{x_1 \leq c_r}}$ (Naive)</td>
<td>0.07</td>
<td>0.17</td>
<td>4.37</td>
</tr>
<tr>
<td>$\text{CATE}_{{x_1 \leq c_r}}$ (Step 5)</td>
<td>-0.12</td>
<td>0.19</td>
<td>6.37</td>
</tr>
<tr>
<td>$\text{CATE}_{{x_1 &gt; c_r}}$ (Naive)</td>
<td>-0.16</td>
<td>0.24</td>
<td>11.50</td>
</tr>
<tr>
<td>$\text{CATE}_{{x_1 &gt; c_r}}$ (Step 5)</td>
<td>0.18</td>
<td>0.31</td>
<td>11.74</td>
</tr>
</tbody>
</table>

$\lambda$ is the heterogeneity coefficient described in Equation 3; note that for the lines that report on $\text{CATE}$ estimates, the true value of the subgroup-specific $\text{CATE}$ depends on the estimated cutoff, as described in the last paragraph of the last section; $c_r$ is the cut value on $x_1$ selected for replication $r$; estimates are based on 144, 728 and 932 replications for $\lambda = 1$, 2 and 3, respectively; “Naive” estimates are obtained from the regression tree output after Step (4); “Step 5” estimates are obtained after stratified matching within subgroups.

motivated by Morgan, Frisco, Farkas, and Hibel (2010) and were described in Keller and Tipton (2016). Morgan et al. (2010) identified 34 covariates, measured at either kindergarten or first grade, based on theoretical considerations or prior empirical research results, distributed across six domains: demographic (three variables), academic (nine variables), school composition (five variables), family context (six variables), health (four variables), and parent rating of child (seven variables). The covariates are described in greater detail, including ECLS-K code names, in Keller and Tipton (2016). After eliminating cases with missing data, the final analytic sample included 7,362 children, among whom 429 children had received special education services and 6,933 children had not.

Using the five-step procedure, we first estimated propensity scores using a logit model with interactions\(^1\). For the second step, we matched children receiving special education

\(^1\)The usual recommended steps for the specification of the propensity score, including iterative respecification of the propensity score model to achieve optimal balance, and an examination of overlap on the logit of the propensity score, are important here, though we do not describe them in detail because our focus is on the detection of heterogeneous subgroups. See, e.g., Keller & Tipton, 2016 for a summary of recommended steps in propensity score analysis.
services to those not receiving special education services, but having similar propensities to receive these services. We obtained 371 matched strata through optimal full matching. The number of children within each matched stratum ranged from 2 to 501 with a median of 6 and mean of 19. Eighty percent of matched strata had fewer than 20 children and ninety percent of matched strata had fewer than 41 children in each stratum. We then calculated stratum-specific treatment effect estimates for reading and mathematics scores, respectively. For children within the same matched stratum, we assumed their covariate profiles were similar (recall, under strong ignorability, \( X \perp T \mid e(X) \)) and, thus, we used the stratum-specific treatment effect estimates as “individual” treatment effect estimates.

Next, for step three, we ran a permutation-based Type I error rate study to select the value of the cost-complexity tuning parameter to use with regression trees. The key for permutation here is to randomly shuffle yoked outcome/treatment pairs, leaving covariate values fixed. This scheme ensures that the raw treatment effect is retained for each permutation, while any relationships between covariates and treatment effects are destroyed. By inducing a lack of systematic association between covariate values and treatment effects, we ensure that the permuted data align with the permutation null hypothesis of no effect. Then, permutation testing may be run with varying values of the regression tree cost-complexity tuning parameter to assess Type I error rates. Testing based on permutation is particularly useful because, under ignorability and a no interference assumption, it permits valid testing without reliance on parametric assumptions (Rubin, 1980; Keller, 2012). One thousand permuted data sets were created for each cost-complexity value. Type I error rate was calculated as a proportion over the 1000 randomizations for which our method resulted in regression trees with more than one node (i.e., at least one split), thereby spuriously indicating covariate-dependent treatment effect heterogeneity. We found the cost-complexity parameter value of 0.01 was associated with a Type I error rate of 0.01 and a cost-complexity value of 0.005 was associated with a Type I error rate of 0.06 for both reading and math outcomes.
In the fourth step, regression trees were fit using cost-complexity parameter values of 0.01 and 0.005, with the vector of individual treatment effect estimates used as the outcome variable, to identify subgroups with heterogeneous treatment effects in two ways: (1) using all 34 covariates as predictors and (2) using only the covariates identified as important by the random forest variable selection algorithm (Keller & Zhang, 2018, March). The variable selection results, displayed in Table 3, show that public vs private schooling (S2KPUPRI), kindergarten reading score (RIRT) and parent rating of child’s communication skills (P1PRONOU) were selected as important predictors of variation in individual treatment effects for both math and reading outcomes. Finally, for step five, we reran optimal full propensity score matching with the added constraint that matches must be made within subgroups to estimate the overall ATE and the subgroup-specific CATEs.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Variable Selected</th>
<th>RF Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>S2KPUPRI</td>
<td>28.7</td>
</tr>
<tr>
<td></td>
<td>RIRT</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>P1PRONOU</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td>C1FMOTOR</td>
<td>12.1</td>
</tr>
<tr>
<td>Mathematics</td>
<td>S2KPUPRI</td>
<td>21.5</td>
</tr>
<tr>
<td></td>
<td>RIRT</td>
<td>16.9</td>
</tr>
<tr>
<td></td>
<td>P1PRONOU</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>P1DISABL</td>
<td>7.8</td>
</tr>
</tbody>
</table>

Table 3
Variables Selected as Important Predictors of Treatment Effect Heterogeneity by Random Forest Variable Selection; S2KPUPRI = public or private school indicator; RIRT = kindergarten reading achievement score; P1PRONOU = parent’s rating of verbal communication; C1FMOTOR = fine motor skills; P1DISABL = indicator of parent’s rating of child’s disability status

Results of the Case Study

For both math and reading outcomes, setting the cost-complexity parameter at 0.01, the more conservative value, yielded two subgroups with heterogeneous treatment effects based on kindergarten reading pretest (RIRT) with identical cutoffs at RIRT = 40; plots
are displayed in Figure 1. Children whose average IRT reading scores at kindergarten were at least 40 had higher treatment effect estimates, implying that students with higher initial reading level tended to benefit more from special education services. The finding that RIRT is an important predictor of effect heterogeneity is consistent with the variable selection results, which identified kindergarten reading score as the second most important covariate for predicting individual treatment effect estimates for both reading and mathematics (see Table 3).

![Figure 1](image)

*Figure 1.* Heterogeneous subgroups for the reading outcome variable (left panel) and the math outcome variable (right panel) identified by regression trees with complexity parameter set to 0.01.

To interpret the plot for the reading outcome variable, shown in the left panel of Figure 1, note that the full sample of 7049 students was split into two subgroups, depending on whether the reading pretest score was less than 40 or not. For the 3579 students with reading pretest scores less than 40, the naive CATE estimate was -3.9. For the remaining 3754 students, with reading pretest scores at or above 40, the naive CATE estimate was 3.6. The plot for the math outcome, shown in the right panel of Figure 1 may be interpreted similarly. After stratified matching, the Step (5) CATE estimates for reading were, respectively, -9.25 and 3.46 and the CATE estimates for math were, respectively, -8.22 and 8.31. Both analyses suggest that enrollment in special education...
services had a net positive effect for students with kindergarten reading scores higher than 40 and a net negative effect for students with kindergarten reading scores lower than 40.

Setting the cost-complexity parameter to 0.005 resulted in the regression trees plotted in Figure 2, which show that, in addition to initial reading level at kindergarten, private vs public school attendance plays a role in the efficacy of special education services on academic achievement in reading in math. Again, this finding was consistent with the variable selection results, which flagged kindergarten reading (RIRT) and the public/private indicator variable (S2KPUPRI) as the two most important predictors of effect heterogeneity. In addition to the plots, results for both naive and Step 5 estimates are shown in Table 4. Interestingly, conclusions based on the naive and Step 5 estimates do not always agree. For example, for cp = 0.005, the naive estimates suggest that private school attendance is associated with improved efficacy of special education services relative to public school attendance, whereas, the Step 5 estimates suggest the opposite. The discrepancy is due to two factors, both of which favor the interpretation based on the Step 5 estimates. First, the Step 5 estimates only allow treatment/comparison group matches within identical subgroups, which we found through simulation (cf., Table 2) led to more accurate and precise estimation. Second, due to our use of a caliper, any units that had no counterpart within two tenths of a pooled SD of the propensity score logit and also within the same subgroup were eliminated. Note, for example, in the lower right panel of Table 4, the estimates for CATE_{RIRT ≥ 40; Private} changed rather drastically from 11.87 to 1.65. Also note, however, the drastic change in sample size from 887 to 367. That is, 520 comparison cases that had been counted toward the estimation of the CATE under the naive estimator were eliminated as no longer comparable when matches were constrained to be made both within subgroups and subject to the caliper constraint.

We also carried out an analysis that included only the covariates identified by random forest variable selection (listed in Table 3) for subgroup identification. This method yielded the same results as including all 34 covariates in the single regression tree
model for both reading and mathematics, indicating that variable selection may be used to narrow down variables used for subgroup identification.

**Conclusion**

In this paper we described an approach that identifies subgroups of units in an observational study for which a treatment is differentially effective based on the values of one or more covariates. Our approach is accessible because it uses the mechanics of propensity score matching and simple regression trees, both of which are familiar to applied researchers in education and behavior science, and easy to implement with existing software packages. We found through Monte Carlo simulation that the method correctly flags covariates that interact with treatment with high probability when their interaction with treatment is moderately strong or stronger. Furthermore, we found that by selecting the right value of the cost-complexity tuning parameter for the regression tree fit, it is possible to control the proportion of false positives to a prespecified nominal level.

In a case study using data from ECLS-K, we identified kindergarten reading score as the most important moderator of the effect of special education services on 5th grade achievement in reading and math. In particular, we estimated the average effect of
Table 4

<table>
<thead>
<tr>
<th>Estimator</th>
<th>cp = 0.01</th>
<th>cp = 0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive</td>
<td>Step 5</td>
</tr>
<tr>
<td></td>
<td>$n$</td>
<td>Est</td>
</tr>
<tr>
<td>Overall ATE</td>
<td>7049</td>
<td>-0.29</td>
</tr>
<tr>
<td>$\text{CATE}_{{\text{RIRT}&lt;40}}$</td>
<td>3579</td>
<td>-4.13</td>
</tr>
<tr>
<td>$\text{CATE}_{{\text{RIRT}\geq40}}$</td>
<td>3470</td>
<td>3.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Math Achievement Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{CATE}_{{\text{RIRT}&lt;40; \text{Public}}}$</td>
</tr>
<tr>
<td>$\text{CATE}_{{\text{RIRT}\geq40; \text{Private}}}$</td>
</tr>
<tr>
<td>$\text{CATE}_{{40\leq\text{RIRT}&lt;47; \text{Public}}}$</td>
</tr>
<tr>
<td>$\text{CATE}_{{40\leq\text{RIRT}&lt;47; \text{Private}}}$</td>
</tr>
<tr>
<td>$\text{CATE}_{{40\leq\text{RIRT}\geq40; \text{Private}}}$</td>
</tr>
</tbody>
</table>

* Table 4
Overall and subgroup-specific estimates of the impact of exposure to special education services on students’ reading achievement in fifth grade; Total sample size $n = 7362$ is reduced to $n = 7049$ due to a caliper of 0.2 pooled SDs of the logit propensity score imposed on matches; Total sample size is further reduced for Step 5 due to the constraint that matches be made within subgroups.

The results of the case study demonstrate how heterogeneous subgroup identification
can provide meaningful and important results that are of immediate interest to stakeholders such as teachers, parents, and administrators. But how reliable are the results? First, we must underscore that the results will only be valid insofar as the assumptions required for identification and estimation of average causal effects through conditioning strategies, as laid out in Rosenbaum & Rubin, 1983, are upheld. In our case study, for example, we might be concerned that ignorability is not satisfied. Although a sensitivity analysis could be carried out to probe the robustness of estimates to unobserved confounding (Rosenbaum, 2002), we did not do so herein. Furthermore, the definition of the “treatment” is opaque. Nuances such as total duration, quality of services, staff training, etc., have not been considered and yet certainly varied from student to student and site to site. Finally, there may have been interference between students assigned to special education in the same schools and, aside from accounting for several school-level covariates, our analyses do not handle school-level nesting.

Concerns about the assumptions required for identification notwithstanding, if we assume they are met, are the results on the identification of heterogeneous subgroups likely to trustworthy using this method? Our simulation results are informative here because they demonstrate that by using intelligently chosen cost-complexity tuning parameter values, the Type I error rate may be held to a nominal level. Thus, we may be confident that any covariates identified by the procedure as moderators are real, subject to the usual caveats regarding the possibility of Type I errors. Although our simulation results suggest that subgroup-specific estimates based on Step 5 are likely to be accurate, a limitation of our method is that no standard errors or confidence intervals are provided for the CATEs. Future research aimed at proposing and evaluating resampling approaches such as the jackknife, bootstrap, and boosting to attain error bounds on estimated CATEs with this method would be useful.
References


