Statistics in the Biosciences manuscript No.
(will be inserted by the editor)

Defining the Study Population for an Observational Study to Ensure Sufficient Overlap: A Tree Approach

Mikhail Traskin · Dylan S. Small

Received: date / Accepted: date

Abstract The internal validity of an observational study is enhanced by only comparing sets of treated and control subjects which have sufficient overlap in their covariate distributions. Methods have been developed for defining the study population using propensity scores to ensure sufficient overlap. However, a study population defined by propensity scores is difficult for other investigators to understand. We develop a method of defining a study population in terms of a tree which is easy to understand and display, and that has similar internal validity as that of the study population defined by propensity scores.

Keywords Observational study · Propensity score · Overlap

1 Introduction

An observational study attempts to draw inferences about the effects caused by a treatment when subjects are not randomly assigned to treatment or control as they would be in a randomized trial. A typical approach to an observational study is to attempt to measure the covariates that affect the outcome, and then adjust for differences between the treatment and control groups in these covariates via propensity score methods [25,16], matching methods [21,13] or regression; see [18], [22] and [27] for surveys.

Mikhail Traskin
Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
Tel.: +215-898-8231
Fax: +215-898-1280
E-mail: mtraskin@wharton.upenn.edu

Dylan S. Small
Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
E-mail: dsmall@wharton.upenn.edu

A quantity that is often of central interest in an observational study is the *average treatment effect*, the average effect of treatment over the whole population. However, in many observational studies, there is a *lack of overlap*, meaning that parts of the treatment and control group's covariate distributions do not overlap. For example, in studies of the comparative effectiveness of medical treatments, virtually all patients of certain types may receive a particular treatment and virtually no patients of certain other types may receive this treatment, but there may be a marginal group of patients who may or may not receive the treatment depending upon circumstances such as where the patient lives, patient preference or the physician's opinion [24]. When there is a lack of overlap, inferences for the average treatment effect rely on extrapolation. This is because extrapolation is needed to estimate the treatment effect for those subjects in the treatment group whose covariates differ substantially from any subjects in the control group (or vice versa). Rather than rely on extrapolation, it is common practice to limit the study population to those subjects with covariates that lie in the overlap between the treatment and control groups. In comparative effectiveness studies, these subjects in the overlap are the marginal patients for whom there is some chance that they would receive either treatment or control based on their covariates. Focusing on the average treatment effect for the marginal patients rather than all patients enhances the internal validity of the study and is more informative for deciding how to treat patients for whom there is currently no definitive standard of care. Knowing the average treatment effect on the currently marginal patients may also shift the margin and over a period of time, a sequence of studies may gradually shift the consensus of which patients should be treated [24].

A number of approaches have been developed for defining a study population that has overlap between the treated and control groups. An often used approach is to discard subjects whose propensity score values fall outside the range of propensity scores in the subsample with the opposite treatment [9, 27]; the propensity score is the probability of receiving treatment given the measured covariates [25]. This approach seeks to make the study population as large as possible while maintaining overlap. However, if there are areas of limited overlap, that is parts of the covariate space where there are limited numbers of observations for the treatment group compared to the control group or vice versa, the average treatment effect estimate on this study population may have large variance. It may be better to consider a more restrictive study population for which there is *sufficient overlap*. The goal is to define a study population which is as inclusive as possible but for which there is enough overlap that, not only is extrapolation not needed, but also the average treatment effect can be well estimated. Crump, Hotz, Imbens and Mitnik [8] and Rosenbaum [24] develop criterion by which to compare different choices of study populations according to this goal and then choose the study population to optimize these criteria. Crump et al.'s criterion is the variance of the estimated average treatment effect on the study population. They show that, under some conditions, the optimal study population by this criterion is those subjects whose propensity scores lie in an interval $[\alpha, 1-\alpha]$ with the optimal interval determined by the marginal distribution of the propensity score and usually well approximated by $[0.1, 0.9]$. Rosenbaum's criterion involves balancing (i) the sum of distances between the propensity scores (or related quantities) of matched treated and control subjects in the study population and (ii) the size of the study population; the criterion will be explained in detail in Section 2.3. Rosenbaum develops an algorithm that uses the optimal assignment algorithm for choosing the optimal study population according to his criterion.

All of the above approaches define the study population in terms of the propensity score and related quantities. A difficulty with these approaches is that it is hard to have a clear understanding of a study population defined in terms of propensity scores. Rosenbaum, in his book *Design of Observational Studies* [23], states, "Rather than delete individuals one at a time based on extreme propensity scores, it is usually better to go back to the covariates themselves, perhaps redefining the population under study to be a subpopulation of the original population of subjects. A population defined in terms of [the propensity score] is likely to have little meaning to other investigators, whereas a population defined in terms of one or two familiar covariates will have a clear meaning."

Our goal in this paper is to develop an approach to defining a study population that has sufficient overlap and is good by Crump et al.'s [8] criterion or Rosenbaum's [24] criterion for judging study populations, but that is also easily described. Our approach is to use a classification tree [4] to define the study population in a way that approximates a propensity score based rule for defining the study population. The resulting study population is easily described by a tree diagram. Figure 2 provides an example of a study population that is defined in terms of a classification tree.

Our paper is organized as follows. Section 2 provides the framework and assumptions that we will use as well as reviewing Crump et al.'s (Section 2.2) and Rosenbaum's (Section 2.3) approaches to defining the study population. Section 3 describes our method. Sections 4 and 5 presents examples. Section 6 provides discussion.

2 Framework

2.1 Assumptions and Notation

The framework we use is that of [25]. We have a random sample of size N from a large population. For each subject i in the sample, let D_i denote whether or not the treatment of interest was received, with $D_i = 1$ if subject i receives the treatment of interest and $D_i = 0$ if subject i receives the control. Let Y_i^1 be the outcome that subject i would have if she received the treatment and Y_i^0 be the outcome that subject i would have if she received the control; these are called potential outcomes. The treatment effect for subject i is $\tau_i = Y_i^1 - Y_i^0$.

Let $\tau(\mathbf{X})$ denote the average treatment effect for subjects with covariates $\mathbf{X}$, $\tau(\mathbf{X}) = E[Y_i^1 - Y_i^0 | \mathbf{X}_i = \mathbf{X}]$.

We observe D_i and Y_i, where $Y_i = Y_i^{D_i}$. In addition, we observe a vector of pre-treatment covariates denoted by $\mathbf{X}_i$, where the support of the covariates is $\mathbb{X}$. The propensity score $e(\mathbf{X})$ for a subject with covariates $\mathbf{X}$ is the probability of selection into the treatment given $\mathbf{X}$, $e(\mathbf{X}) = P(D_i = 1|\mathbf{X}_i = \mathbf{X})$.

We make the assumption that all of the confounders have been measured [25]. This means that, conditional on the covariates $\mathbf{X}$, the treatment indicator is independent of the potential outcomes:

$$\text{Assumption 1:} D_i \perp\!\!\!\perp (Y_i^0, Y_i^1)|\mathbf{X}_i.$$

Assumption 1 is called the strongly ignorable treatment assignment assumption [25] or the unconfoundedness assumption [18].

2.2 Minimum Variance Approach to Defining the Study Population

Crump et al. [8] seek to choose the study population that allows for the most precise estimation of the average treatment effect within the study population. They show that, under some conditions, this leads to discarding observations with propensity scores outside an interval $[\alpha, 1-\alpha]$ with the optimal cut-off value determined by the marginal distribution of the propensity score. Their approach is consistent with the common practice of dropping subjects with extreme values of the propensity score with two differences. First, the role of the propensity score in the selection rule is not imposed a priori, but emerges as a consequence of the criterion, and second, there is a principled way of choosing the cutoff value α. They show that the precision gain from their approach can be substantial with most of the gain captured by using a rule of thumb to discard observations with estimated propensity score outside the range $[0.1, 0.9]$.

The specifics of Crump et al.'s approach are as follows. Let $\mathbb{A}$ be a subset of the covariate space $\mathbb{X}$. For sets $\mathbb{A} \subset \mathbb{X}$, let $I_{\mathbf{X}_i \in \mathbb{A}}$ be an indicator for the event that $\mathbf{X}_i$ is an element of the set $\mathbb{A}$, and define the subsample average treatment effect $\tau_{\mathbb{A}}$:

$$\tau_{\mathbb{A}} = \frac{1}{N_{\mathbb{A}}} \sum_{i:\mathbf{X}_i \in \mathbb{A}} \tau(\mathbf{X}_i), \quad N_{\mathbb{A}} = \sum_{i=1}^{N} I_{\mathbf{X}_i \in \mathbb{A}}.$$

In words, $\tau_{\mathbb{A}}$ is the average treatment effect for the study population of subjects with covariates in $\mathbb{A}$, where the distribution of covariates in $\mathbb{A}$ is assumed to be the same as that in the sample. Under the assumption that the conditional variance of potential outcomes given $\mathbf{X}$ is constant, i.e., $Var(Y_i^0|\mathbf{X}_i = \mathbf{X}) = Var(Y_i^1|\mathbf{X}_i = \mathbf{X}) = \sigma^2$, the asymptotic variance of an efficient estimate of $\tau_{\mathbb{A}}$ is the following [12,20,16]:

$$V(\mathbb{A}) = \frac{1}{P(\mathbf{X} \in \mathbb{A})} E\left\{ \frac{\sigma^2}{e(\mathbf{X})} + \frac{\sigma^2}{1 - e(\mathbf{X})} | \mathbf{X} \in \mathbb{A} \right\}. \tag{1}$$

Crump et al. seek to choose the subpopulation $\mathbb{A}$ which minimizes the asymptotic variance. Focusing on estimands that average the treatment effect only over a subpopulation rather than the whole population has two effects on the asymptotic variance, pushing it in opposite directions. First, excluding subjects with covariate values outside the set $\mathbb{A}$ reduces the effective sample size in expectation from N to $NP(\mathbf{X} \in \mathbb{A})$, increasing the asymptotic variance by a factor of $1/P(\mathbf{X} \in \mathbb{A})$. Second, discarding subjects with high values for $\frac{\sigma^2}{e(\mathbf{X})} + \frac{\sigma^2}{1-e(\mathbf{X})}$ lowers the conditional expectation $E\left\{\frac{\sigma^2}{e(\mathbf{X})} + \frac{\sigma^2}{1-e(\mathbf{X})} | \mathbf{X} \in \mathbb{A}\right\}$. Optimally choosing $\mathbb{A}$ involves balancing these two effects.

Crump et al. show that the optimal choice of $\mathbb{A}$ is

$$\mathbb{A}^* = \{\mathbf{X} \in \mathbb{X} : \alpha \leq e(\mathbf{X}) \leq 1 - \alpha\}, \tag{2}$$

where, if

$$\sup_{\mathbf{X} \in \mathbb{X}} \frac{1}{e(\mathbf{X})\{1 - e(\mathbf{X})\}} \leq 2E\left[\frac{1}{e(\mathbf{X})\{1 - e(\mathbf{X})\}}\right],$$

then $\alpha = 0$ and otherwise α is a solution to

$$\frac{1}{\alpha(1-\alpha)} = 2E\left[\frac{1}{e(\mathbf{X})(1 - e(\mathbf{X}))} \mid \frac{1}{e(\mathbf{X})(1 - e(\mathbf{X}))} \leq \frac{1}{\alpha(1-\alpha)}\right].$$

The optimal set $\mathbb{A}^*$ depends only on the distribution of the covariates $\mathbf{X}$, so it can be constructed without looking at the outcome data. This avoids potential biases associated with using outcome data to define the study population. The set defined by (2) is optimal under homoskedasticity (the conditional variance of potential outcomes given $\mathbf{X}$ is constant) but even when homoskedasticity does not hold, the set (2) may be a useful approximation [8].

To implement their proposed criterion, Crump et al. first estimate the propensity score, e.g., via logistic regression, and then solve for the smallest value $\hat{\alpha} \in [0, 1/2]$ that satisfies

$$\frac{1}{\alpha(1-\alpha)} \leq 2\frac{\sum_{i=1}^{N} I_{\hat{e}(\mathbf{X}_i)(1-\hat{e}(\mathbf{X}_i)) \geq \alpha(1-\alpha)}/[\hat{e}(\mathbf{X}_i)\{1 - \hat{e}(\mathbf{X}_i)\}]}{\sum_{i=1}^{N} I_{\hat{e}(\mathbf{X}_i)(1-\hat{e}(\mathbf{X}_i)) \geq \alpha(1-\alpha)}}, \tag{3}$$

and use the set

$$\hat{\mathbb{A}} = \{\mathbf{X} \in \mathbb{X} : \hat{\alpha} \leq \hat{e}(\mathbf{X}) \leq 1 - \hat{\alpha}\} \tag{4}$$

Given this definition $\hat{\mathbb{A}}$ of the study population, one can use any standard method for estimation of, and inference for, average treatment effects, such as those surveyed in [18], [22] or [27], ignoring the uncertainty in $\hat{\mathbb{A}}$. Crump et al. use a version of the Horvitz-Thompson estimator [17] that is detailed in [16]. Specifically, Crump et al. first estimate the propensity score on the selected study population $\hat{\mathbb{A}}$ using the full set of covariates. They then estimate the average treatment effect for the study population by

$$\hat{\tau}_{\hat{\mathbb{A}}} = \sum_{i=1}^{N} \frac{D_i Y_i I_{\mathbf{X}_i \in \hat{\mathbb{A}}}}{\hat{e}(\mathbf{X}_i)} / \sum_{i=1}^{N} \frac{D_i I_{\mathbf{X}_i \in \hat{\mathbb{A}}}}{\hat{e}(\mathbf{X}_i)} - \sum_{i=1}^{N} \frac{(1 - D_i) Y_i I_{\mathbf{X}_i \in \hat{\mathbb{A}}}}{1 - \hat{e}(\mathbf{X}_i)} / \sum_{i=1}^{N} \frac{(1 - D_i) I_{\mathbf{X}_i \in \hat{\mathbb{A}}}}{1 - \hat{e}(\mathbf{X}_i)} \tag{5}$$

Crump et al. estimate the variance of $\hat{\tau}_{\hat{\mathbb{A}}}$ by using the bootstrap.

2.3 Balance for Treatment on Treated Effect Approach to Defining Study Population

A commonly used design for observational studies is to match each treated subject to a control subject with similar covariates [23]. This design seeks to estimate the treatment on treated effect, the average effect of treatment among those who receive treatment. The treatment on treated effect for the whole population is $E[Y_i^1 - Y_i^0 | D_i = 1]$. When some treated subjects are almost sure to receive treatment based on their covariates, it would be difficult to estimate the treatment on treated effect for the whole population without extrapolation, since there are no similar control subjects to compare these subjects to. One may instead want to only focus on estimating the treatment on treated effect for a subset of subjects. The treatment on treated effect for the subset of subjects with covariates $\mathbf{X} \in \mathbb{A}$ is $E[Y_i^1 - Y_i^0 | D_i = 1, \mathbf{X}_i \in \mathbb{A}]$.

Rosenbaum develops an algorithm for choosing an optimal subset of treated subjects to match and then optimally pair matching them to controls. There are two goals, which are at odds with one another: (i) to match as many treated subjects as possible, recognizing that some treated subjects may be too extreme to match and (ii) to match as closely as possible on the propensity score and other variables with a view to balancing many covariates. The algorithm makes three simultaneous decisions: (i) how many treated subjects to match; (ii) which specific treated subjects to match and (iii) which controls to pair to which treated subjects.

Let $\delta(\mathbf{X}_i, \mathbf{X}_j)$ denote a distance between the covariates $\mathbf{X}_i, \mathbf{X}_j$ and let $\boldsymbol{\Delta}$ denote the matrix which records all distances between treated and control subjects. Commonly used distances are the absolute difference in propensity scores, Mahalanobis distance or Mahalanobis distance with a caliper on the propensity score; see Rosenbaum [23], Chapter 8, for discussion of distance functions for observational studies. Rosenbaum's [24] optimality criterion depends on two parameters. One is the minimum number of treated subjects which we would like to match. The other is a critical distance $\tilde{\delta}$ that we would like our matched pairs to have a distance less than. Among two matchings which each have at least the minimum number of treated subjects, we prefer the matching with more treated subjects if its average increase in distance for the extra matches is less than the critical distance $\tilde{\delta}$ and otherwise prefer the matching with less treated subjects. Among matchings with the same number of treated subjects, we prefer the one with the smallest total distance. Rosenbaum [24] provides an algorithm for finding the optimal matching for this criterion that uses the assignment algorithm on an augmented distance matrix.

Rosenbaum suggests examining a few choices of critical distance $\tilde{\delta}$, in particular, $\tilde{\delta} = \infty$ which will result in using all treated subjects, $\tilde{\delta}$ equal to the 5% quantile of distances in the distance matrix $\boldsymbol{\Delta}$ and $\tilde{\delta}$ equal to the 20% quantile of distances in $\boldsymbol{\Delta}$. The success of a match of treated and control subjects is judged by whether or not it balances the covariates in the treated and control groups. Measures of balance include the standardized differences between the treated and control group's covariates and propensity scores and a comparison of the p values for testing for a difference between the treated and control group means in covariates to what would be expected in a completely randomized experiment [23]. The standardized difference is the difference in means divided by a pooled standard deviation before matching. The pooled standard deviation is the square root of the unweighted average of the variances in the treated and control groups before matching. An absolute standardized difference less than 0.2 is considered adequate balance with a value less than 0.1 being ideal [5,6]. The goal in Rosenbaum's [24] approach is to obtain an internally valid estimate of the treatment on treated effect on as large a subset of the whole population as possible. Consequently, the match with the most treated subjects that has adequate balance is chosen.

3 Tree Method for Defining the Study Population

3.1 Minimum Variance Criterion

Let $r(\mathbf{X})$ be a definition of a study population, where $r(\mathbf{X}) = 1$ if a subject with covariates $\mathbf{X}$ is in the study population and 0 if not. In this section we consider $r(\mathbf{X})$ as Crump et al.'s [8] propensity-score based definition (3)–(4) of the study population described in Section 2.2. We seek to find a definition of a study population r', which is similar to r, but which can be easily described in terms of a few covariates. We propose to use classification trees to do this. Specifically, for Crump et al.'s criterion, let $s(\mathbf{X})$ classify whether $\mathbf{X}$'s propensity score is too low to be in the study population defined by (3)–(4), in the study population or too high to be in the study population: $s(\mathbf{X}) = \mathit{Low}$ if $\hat{e}(\mathbf{X}) < \hat{\alpha}$, $s(\mathbf{X}) = \mathit{In}$ if $\hat{\alpha} \leq \hat{e}(\mathbf{X}) \leq 1 - \hat{\alpha}$ and $s(\mathbf{X}) = \mathit{High}$ if $\hat{e}(\mathbf{X}) > \hat{\alpha}$. We build a classification tree to classify the variable $s(\mathbf{X})$ and then the study population consists of those values of $\mathbf{X}$ that are classified as In by the classification tree.

Classification trees are a nonparametric method of classifying a categorical outcome variable based on covariates that results in a classification rule that can be displayed as a tree [4]. The classification tree partitions the covariate space and then the classification for each set in the partition is the most likely category among the observations that fall into the set. A classification tree is built using binary recursive partitioning [4,28]. At each step of the construction, a split is made in some set of the current partition between higher and lower values of one variable (or for categorical variables, between one subset of the values and another subset). The split is chosen to optimize some criterion for how well the resulting partition classifies the variable. Following [4], we use the Gini index, which is defined as follows. Let p_{jk} denote the proportion in class k at node j of the tree. Then the Gini index is $\sum_j \sum_{k \neq k'} p_{jk} p_{jk'}$.

We use the R library `rpart` to build classification trees in R. Example code is provided in supplementary materials for our paper.

To control the complexity of the tree's definition of the study sample, we can limit the maximum depth of the tree. For a given maximum depth of the tree, the R function `rpart` uses cross validation to choose the tree of that depth which minimizes the probability of misclassification of the tree on a future sample.

Let T_0 be the unpruned tree generated by the `rpart` function. When pruning, we search for a tree $T \subseteq T_0$. The tree quality is measured by the misclassification rate. Since there is no independent sample to estimate the misclassification rate, `rpart`'s algorithm uses a penalized misclassification rate of the form

$$L_\gamma(T) = \sum_{m=1}^{T} n_m L_m(T) + \gamma|T|,$$

where $|T|$ is the number of leaves in the tree T, n_m is the number of observations in the leaf m, $L_m(T)$ is the training misclassification rate, which is defined as the proportion of points with label different from the leaf's label, and γ is the regularization parameter. If $\gamma = 0$, then the tree is not pruned. The regularization γ is chosen to try to minimize the misclassification rate. `rpart`, by default, uses 10-fold cross-validation to choose γ. Specifically, `rpart` generates 10 trees using 9 parts of the data as training set and 10th as a test set to estimate effect of different γ values on the tree's misclassification rate. We used "1 SE rule" (see [15]) to choose the γ that gives the "best" average result. See [4] for more details on the cross-validation procedure used by `rpart`. Once the estimate $\hat{\gamma}$ of the optimal γ parameter is chosen, the generated tree T_0 is pruned by looking for $T \subseteq T_0$ such that $L_{\hat{\gamma}}(T)$ is minimized.

We used the complexity parameter `cp` $= 0$ to allow growing large trees. The complexity parameter `cp` determines which splits are considered by the rpart algorithm when growing trees[1].

In usual applications of trees, the goal is to best classify the outcome on a future sample and the depth of the tree is chosen by cross validation with this goal in mind. In our application, our goal is to choose a tree which defines a study population that is easily understood but also allows for estimating the treatment effect with close to as small a variance as possible (where (1) shows the variance for a given study population $\mathbb{A}$). For the best tree (as chosen by `rpart` through cross validation) of various depths, we create a table of some measure of how small the variance of the estimated average treatment effect is and choose a depth which strikes a good balance between providing an easily understood study population and a small variance of the estimated average treatment effect.

For Crump et al.'s criterion for defining the study population from Section 2.2, we use the following ratio as measure of how well a tree that defines a study population $\tilde{A}$ approximates the best study population:

$$\frac{V(\tilde{\mathbb{A}})}{V(\hat{\mathbb{A}})}, \tag{6}$$

[1] Setting `cp` to 0 means that all possible splits will be considered. The default value of 0.01 means that those splits that result in the decrease of Gini criterion less than 0.01 are not considered. In particular, this implies that if for a given terminal node any split results in the Gini decrease of less than 0.01, then `rpart` does not split this node any further

i.e., the ratio of the estimated asymptotic variance of $\hat{\tau}_{\tilde{\mathbb{A}}}$ by using the study population $\tilde{\mathbb{A}}$ defined by the tree to the estimated asymptotic variance of $\hat{\tau}_{\hat{\mathbb{A}}}$ for the study population $\hat{\mathbb{A}}$ defined by (3)–(4). When calculating this ratio, σ^2 cancels out so does not need to be estimated. A ratio of 1.5 implies that for the given set $\tilde{\mathbb{A}}$ the estimated variance of the average treatment effect is 1.5 times greater than the variance of the estimated treatment effect over the "best" set $\hat{\mathbb{A}}$.

For estimating the ratio (6), we use a cross-validation procedure. The cross-validation procedure we use is the based on the one used by Breiman [3]. Unlike usual k-fold cross-validation, this procedure gives us more flexibility in choosing the train and test sizes and allows us to decrease the variance of the estimate. The procedure can be described as follows.

1. Split the original data set into the training and test parts so that training part contains 80% of the original data and test part contains the remaining 20%. A relatively large size of the test part (20%) was chosen to ensure that with high probability categorical variables in both training and test sets have the same number of levels.
2. Use the training part to fit the propensity score model and then, using the propensity score cutoff that was determined from the whole sample, classify all the observations in the training set into the *Low*, *In* and *High* classes.
3. Use the `rpart` function with the desired `maxdepth` parameter set to find an approximation to the *In* set.
4. Use the test set to estimate the ratio of the variance estimate for the set obtained with the tree to the variance estimate obtained using the logistic regression propensity score model estimates obtained on the training set.
5. Repeat steps 1–4 100 times and compute the median of the ratio of the variance estimate for the set obtained with the tree to the variance estimate obtained using the logistic regression propensity score model estimates obtained on the training set. We used the median instead of the mean because the distribution of ratios is right-skewed.

Since steps 1–5 can be done for various values of the maxdepth parameter, this gives us a graphical representation of the dependence of the ratio on this parameter.

Given all the above discussion, our approach of growing a tree to describe the study population can be formalized as follows.

1. Use the original data set to fit a propensity score model.
2. Use Crump et al. [8] method[2], summarized by (3)–(4), to estimate set $\mathbb{A}$.
3. Use the propensity score model of step 1 and definition of the study population of step 2 to classify all the observations in the original data set into either *Low*, *In* or *High* categories.

[2] In fact, our method of growing a tree will work for any method that excludes from the study population observations with either too low or too high propensity score.

4. Use the cross-validation procedure described above and subject-specific knowledge to determine the optimal tree depth.
5. Use the classification of step 3 and desired tree depth to grow a tree[3] that describes the study population.
6. Use the obtained tree to define the study population on which the treatment effect is estimated.

In summary, our tree approach is easy to implement with functions in R provided in the supplementary materials and makes the study population easy to understand compared to the Crump et al. [8] approach.

An important feature of our approach, like that in Crump et al. [8], is that the tree that defines the study population is chosen only by looking at the covariates $\mathbf{X}$ and treatment D, and not looking at the outcomes Y. Thus, the choice of study population is done before looking at the outcome data, avoiding potential biases associated with using outcome data when defining the study population.

In our experiments we noticed that pruning did not have a considerable effect of the tree size, probably because we were growing small trees to begin with. Also, the ratio 6 was not noticeably affected by pruning. As a result, when considering an alternative approach of Section 3.2 to defining a study population, we did not using pruning.

3.2 Balance for Treatment on Treated Effect Criterion

In this section, we discuss how to use a tree to define a study population for estimating the treatment on treated effect from a matched pair design. Our approach is to approximate the optimal subset chosen by Rosenbaum's criterion described in Section 2.3 by a tree:

1. Choose an optimal set of treated subjects using Rosenbaum's approach. Let $r_i = 1$ if treated subject i is in the optimal set of treated subjects and $r_i = 0$ if not.
2. Fit a tree to predict r_i based on the covariates for all treated subjects. Let $s(\mathbf{X}_i) = 1$ if i is predicted to have $r_i = 1$ based on the tree and covariates $\mathbf{X}_i$, and $s(\mathbf{X}_i) = 0$ if i is predicted to have $r_i = 0$. Define the study population as the set of subjects with covariates $\mathbf{X}$ such that $s(\mathbf{X}) = 1$, i.e., those subjects who are predicted to have $r = 1$ based on the tree.
3. Find the optimal pair match of the subjects with $s_i = 1$. This can be done using the `optmatch` package in R [14].
4. For continuous outcomes, the treatment effect can be estimated by the difference in mean outcomes between the treated and control subjects in

[3] We did not use weighting of observations when growing trees. Weighting can be used to ensure that, for example, tree is more eager to exclude observations from the study, hence increasing the chance that observations with either too high or too low propensity scores are excluded from the study. If weighting is used to grow a tree, then a similar weighting scheme should be used in the cross-validation step.

Table 1 Treated and Control Means and Standardized Differences (treated mean - control mean)/square root of average of within group variances) for the Job Training Study.

Covariate	Treated Mean	Control Mean	Standardized Difference
Age	24.63	34.85	-1.17
Education	10.38	12.12	-0.69
African American	0.80	0.25	1.32
Hispanic	0.09	0.03	0.25
Married	0.17	0.87	-1.94
No High School Degree	0.73	0.31	0.94
1975 Earnings	3066	19,063	-1.57

the pairs and a matched pair t-test can be used to make inferences. For binary outcomes, inferences for the treatment effect can be based on the methods for matched pair designs described in [1].

We consider trees of various maximum depths, aiming to find a tree that yields a study population that (i) has acceptable balance between the treated subjects and their matched control subjects; (ii) is easy to understand; and (iii) is as large as possible given the constraints (i) and (ii). To achieve these goals, we can also consider fitting a tree that has a smaller or larger loss for misclassifying subjects with $r_i = 1$ to 0 as compared to misclassifying subjects with $r_i = 0$ to 1.

4 Example 1: Job Training Program

We consider the National Supported Work Demonstration (NSW) job training program, which was designed to help disadvantaged workers lacking basic job skills to move into the labor market by giving them work experience and counseling in a sheltered environment [19]. The data set was originally constructed by [19] and subsequently used by [9] and [26] among others. The particular sample we use here is the one used by [9]. The treatment group is drawn from an experimental evaluation of the job training program. The control group is a sample drawn from the Panel Study of Income Dynamics. The treatment group contains 297 subjects and the control group contains 2490 subjects. The job training program took place in 1976–1977. The outcome is earnings in 1978. The covariates are age, education, a dummy variable for being African American, a dummy variable for being Hispanic, a dummy variable for being married, a dummy variable for having no high school degree and earnings in 1975. The control and treatment group's covariate distributions differ substantially as seen in Table 1.

4.1 Minimum Variance Criterion

We consider Crump et al.'s minimum variance criterion for defining the study population described in Section 2.2. We estimated the propensity score by

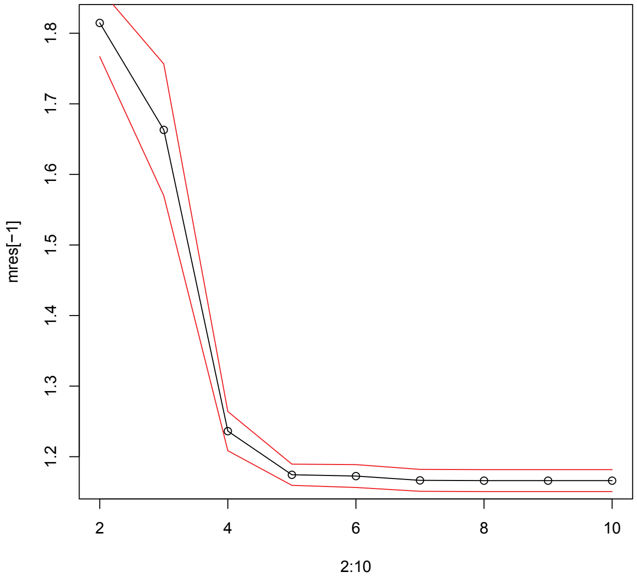


Fig. 1 For the job training data, estimated ratio of variance of best trees of various depths to that of the estimated optimal subpopulation by Crump et al.'s [8] criterion: $\hat{\mathbb{A}} = \{\mathbf{X} \in \mathbb{X} : 0.066 \leq \hat{e}(\mathbf{X}) \leq 0.934\}$

a logistic regression on the covariates. The estimated optimal cutoff point from (3) is $\hat{\alpha} = 0.066$. After fitting the propensity score model using logistic regression and thresholding the estimated probabilities at 0.066 and 0.934 we get the following classification counts:

Low	*In*	*High*
2187	584	16

Hence, only 26.7% of observations should be included into the study according to Crump et al.'s [8] criterion for defining the study population. Using the study population based on this estimated optimal cutoff point produces a more than million-fold decrease in the asymptotic variance (1) compared to using all subjects.

Figure 1 shows the ratio of the estimated asymptotic variance of the best trees of various depths to that of the optimal subpopulation $\hat{\mathbb{A}} = \{\mathbf{X} \in \mathbb{X} : \hat{\alpha} \leq \hat{e}(\mathbf{X}) \leq 1 - \hat{\alpha}\}$ using the cross-validation approach described in Section 3.1.

Figure 1 shows there is a big gain from going from a tree of depth 1 to a tree of depth 2; the variance of the estimated treatment effect is reduced by more than 50%. There is not much gain from going from a tree of depth 2 to a tree of depth 3. Going from a tree of depth 2 to a tree of depth 4

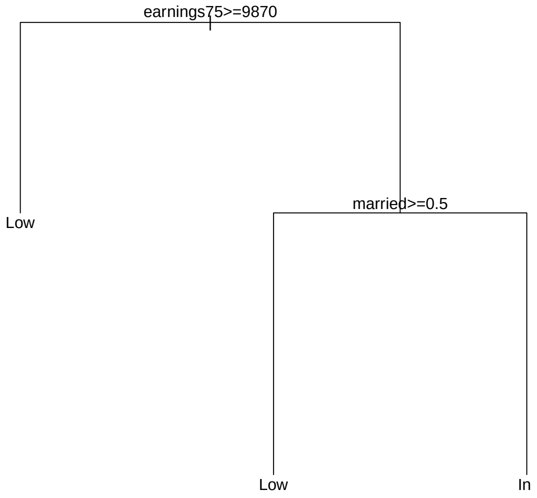


Fig. 2 Best tree of depth 2 for the minimum variance criterion for the job training study.

reduces the variance by around 30%. There is not much gain in going beyond a tree of depth 4. Thus, a tree of depth 2 or depth 4 are the most reasonable choices to define the study population, and the choice between them depends on a researcher's tradeoff between complexity of the definition of the study population vs. a small variance of the estimated treatment effect.

The trees of depths 2 and 4 are shown in Figures 2 and 3 respectively. The way in which to read these trees to see if a subject is in the study population is that we follow the logical statements given in the boxes. When we arrive at a leaf of the tree, if the classification of the leaf is *In*, the subject is in the study population and if the classification of the leaf is *Low* (too low a propensity score) or *High* (too high a propensity score), the subject is not in the study population. The tree of depth 2 defines the study population as subjects whose 1975 earnings were less than $9,870 and are not married. The tree of depth 4 defines the study population as subjects who fall into any of the following groups: (i) 1975 earnings less than $9,870, not married and 17.5 or older; (ii) 1975 earnings less than $9,870, not married, younger than 17.5 and 1975 earnings greater than $616; (iii) 1975 earnings less than $ 9,870, married, not black and Hispanic; (iv) 1975 earnings less than $ 9,870, married, black and younger than 38.5; (v) 1975 earnings less than $17,520, not married and black.

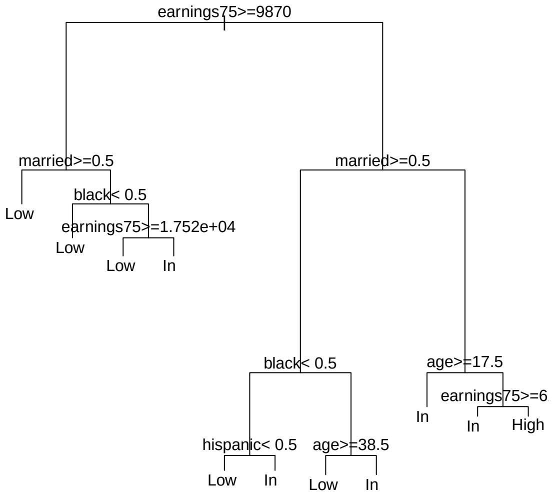


Fig. 3 Best tree of depth 4 for the minimum variance criterion for the job training study.

Table 2 shows the estimated effect of job training programs (as well as standard errors and the % of all subjects in the study sample) using the Horvitz-Thompson estimator (5) for the study populations defined by the tree of depth 2 in Figure 2, the tree of depth 4 in Figure 3 and the Crump et al. criterion. The standard errors were estimated by the bootstrap as in [8]. Specifically, using those subjects in the sample for the the given study population definition, we resample subjects with replacement and estimate the treatment effect by (5).

The estimates for all of the study populations is that the job training program has a negative effect; the effect is significant for the study populations defined by Figure 3 and the Crump et al. criterion. The estimates for the study population defined by Figure 2 differs substantially from that for the study population defined by Figure 3 and the Crump et al. criterion. This suggests that the treatment effect is not constant. A valuable feature of the tree approach is that it gives us some idea about what population an estimates refers to and can potentially be generalized to.

4.2 Balance for Treatment on Treated Effect Criterion

We consider choosing a study population for estimating the treatment on treated effect from a matched pair design as discussed in Section 2.3. For the

Table 2 Estimated Average Treatment Effect of Job Training for Different Study Populations.

Study Population	Treatment Effect	Standard Error (Bootstrap Estimate)	% of All Subjects in Study Sample
Tree of Depth 2	-360	882	13.1
Tree of Depth 4	-2293	843	19.9
Crump et al.	-2509	819	20.9

distance matrix $\boldsymbol{\Delta}$, we use the rank-based Mahalanobis distance described by [23], between two observations, where the original observations are replaced by their ranks, and a penalty is added if two observations have propensity scores that differ by more than 0.2 standard deviations of the propensity score. This distance takes into account the goals of forming close individual matches (by using the Mahalanobis distance), obtaining overall balance (by having a penalty if the propensity scores are too far apart) and robustness (by replacing the original data by its ranks). See Chapter 8 of Rosenbaum (2010) [23] for further discussion of this distance matrix and its rationale.

Figure 4(a) shows box plots of the standardized differences for the covariates and the propensity score between a treated and a matched pair control group for (i) all treated subjects, (ii) the optimal subset of treated subjects when the critical distance is the 25% quantile of all distances in $\boldsymbol{\Delta}$ and (iii) the optimal subset of treated subjects when the critical distance is the 20% quantile of all distances in $\boldsymbol{\Delta}$. Using all treated subjects produces poor balance with most standardized differences being greater than 0.2. Using the optimal subset of treated subjects when the critical distance is the 25% quantile improves the balance but one standardized difference is still greater than 0.2. Using the optimal subset of treated subjects when the critical distance is the 20% quantile provides good balance with all standardized differences less than 0.2 and most standardized differences less than 0.1. This optimal subset uses 154 out of the original 297 treated subjects.

We consider fitting a tree to the classification of subjects as *In* or *Out* of the optimal subset of treated subjects when the critical distance is the 20% quantile. The resulting standardized differences are shown in Figure 4(b). Figure 5 compares the p-values from two sample tests to the uniform distribution, so worse covariate balance than in a completely randomized experiment leads to points below the diagonal line, whereas better balance leads to points above the diagonal line. A tree of depth 1 has unacceptable balance with standardized differences above 0.2. A tree of depth 2 has all standardized differences less than 0.2, but some standardized difference are quite close to 0.2. A tree of depth 3 has all standardized differences less than 0.1. Also, Figure 5 shows that the treatment and control groups defined by the tree of depth 3 have covariate balance that is better than that expected from a randomized experiment. The tree of depth 3 uses a similar, but smaller number of treated subjects (124), compared to the optimal subset (154).

Table 3 Estimated Treatment on Treated Effect of Job Training for Different Study Populations.

Study Population	Treatment Effect	Standard Error	Number of Pairs
Optimal Subset with Critical Distance = 20% Quan.	-3450	1089	154
Tree of Depth 3	-4637	1285	140

The tree of depth 3 is plotted in Figure 6. The study population associated with this tree is as follows (where note that the job training data we consider only involves so all possible study populations only involve men): (i) married men with earnings less than \$2308; (ii) men at least 21 years old who earned at least \$2308 in 1975; (iii) unmarried men who are at least 38 years old and earned less than \$2,308 in 1975; (iv) men who are less than 21 years old and earned at least \$8,256 in 1975. Table 3 shows the estimated treatment on treated effect for this study population based on a matched pair t-test, along with that of the optimal subset that the tree of depth 3 sought to approximate. The job training program is estimated to have a substantial negative effect, $-4,637$ (p-value $< .001$) on the treated subjects in the study population defined by the tree of depth 3. Note that this is the treatment on treated effect, so that even though men who have high earnings in 1975 are included in the study population, they are unlikely to be treated and hence this treatment effect says little about what their treatment effect would be if they were treated.

The standard error for the treatment on treated effect estimate for the tree of depth 3 study population is only 18% higher than that of the optimal subset study population. In this example, the tree approach has provided an easily understood study population that yields a treatment effect estimate with a similar standard error as that of the harder to understand optimal subset study population.

The study population defined by the tree of depth 3 is fairly easily described, but we might want to limit ourselves to study populations defined by trees of depth 2. Unfortunately, from Figure 4, the study population defined by the tree of depth 2 lacks good covariate balance with a standardized difference of almost 0.2. If we want to have a tree of depth 2 and are willing to reduce the number of treated subjects (i.e., reduce the sample size), we can consider fitting a tree that puts a higher loss on misclassifying a subject who is out of the optimal subset than a subject who is in the optimal subset. For example, we consider a tree of depth 2 that has double the loss for misclassifying a subject who is out of the optimal subset than a subject who is in the optimal subset. The tree is shown in Figure 7. This tree uses 85 treated subjects, compared to the 124 treated subjects used by the tree of depth 3 in Figure 6. The study population defined by this tree is (i) married men whose 1975 earnings are less than \$2,840 and (ii) men whose 1975 earnings are at least \$2,840 and are at least 22 years old. The treatment effect for this study population is estimated to be that job training reduces earnings by \$5,265 with a standard error of \$1,766, which is a 37% higher standard error than for the study population defined by the tree of depth 3 in Figure 6.

5 Example 2: Right Heart Catheterization

Connors et al. [7] used a propensity score matching approach to study the effect of right heart catheterization (RHC) on mortality in an observational study. RHC is a diagnostic procedure for directly measuring cardiac function. The measures of cardiac function provided by RHC are useful for directing immediate and subsequent therapy. However, RHC can cause complications such as line sepsis, bacterial endocarditis and large vein thrombosis [7]. At the time of Connors et al.'s study, the benefits of RHC had not been demonstrated in a randomized controlled trial. The popularity of the procedure and the widespread belief that it is beneficial made conducting a randomized trial difficult and an attempt at a randomized trial was stopped because most physicians refused to allow their patients to be randomized. For an observational study of RHC, it is important to control for confounding variables that are associated with decisions to use RHC and outcome; for example, patients with low blood pressure are more likely to be managed with RHC and such patients are more likely to die. For Connors et al.'s study, a panel of 7 specialists in critical care specified the variables that would relate significantly to the decision to use or not to use RHC. The variables were age, sex, race (black, white, other), years of education, income, type of medical insurance (private, Medicare, Medicaid, private and Medicare, Medicare and Medicaid, or none), primary disease category, secondary disease category, 12 categories of admission diagnosis, activities of daily living (ADL) and Duke Activity Status Index (DASI) 2 weeks before admission, do-not-resuscitate status on day 1 of admission, cancer (none, localized, metastatic), an estimate of the probability of surviving 2 months, acute physiology component of the APACHE III score, Glasgow Coma Score, weight, temperature, mean blood pressure, respiratory rate, heart rate, PaO_2/FIO_2 ratio, $PaCO_2$, pH, WBC count, hematocrit, sodium, potassium, creatinine, bilirubin, albumin, urine output, 12 categories of comorbid illness and whether the patient transferred from another hospital. The outcome was survival at 30 days after admission. The treatment is that RHC was applied within 24 hours of admission and the control is that RHC was not applied within 24 hours of admission. The study contained patients admitted to intensive care units (ICUs) who met severity and other entry criteria and were in one or more of nine disease categories on admission; the disease categories are acute respiratory failure (ARF), chronic obstructive pulmonary disease (COPD), congestive heart failure (CHF), cirrhosis, nontraumatic coma, colon cancer metastatic to the liver, non-small cell cancer of the lung (stage III or IV) and multiorgan system failure (MOSF) with malignancy or sepsis. There are 2184 patients in the treatment group and 3551 in the control group.

Table 4 Treated and Control Means and Standardized Differences (treated mean − control mean)/square root of average of within group variances) for several important variables for the RHC Study.

Covariate	Treated Mean	Control Mean	Standardized Difference
Age	60.71	61.76	-0.06
Female	0.41	0.46	-0.09
APACHE Score	60.74	50.93	0.50
Multiple Organ System Failure with Sepsis	0.32	0.15	0.41

The control and treatment group's covariate distributions differ substantially on several key variables as seen in Table 4. The treated group has a substantially higher average APACHE (Acute Physiology and Chronic Health Evaluation) score, meaning that the treated group has higher severity of disease when being admitted. The treated group is also substantially more likely to be admitted with multiple organ system failure with sepsis. Multiple organ system failure with sepsis is a severe condition from which patients are highly likely to die.

5.1 Minimum Variance Criterion

Crump et al. [8] used the RHC study to illustrate their method. When the categorical covariates are broken into dummy variables and combined with the continuous covariates, there are 72 total covariates. Crump et al. estimated the propensity score by logistic regression on the 72 covariates. Based on the estimated propensity score, they calculated the optimal cutoff value α from (3) as $\alpha = 0.1026$ so that their study population consists of subjects with propensity scores between 0.1026 and 0.8974. This results in 82% of the original sample being in the study group. 16% of subjects are excluded because they have too low propensity scores and 2% are excluded because they have too high propensity scores. Since the propensity score describing the study population is based on 72 covariates, the study population is not that easily described.

Figure 8 shows the ratio of the estimated asymptotic variance of using the study populations defined by best trees of various depths to that of the study population from Crump et al.'s method described in the above paragraph. We used the cross-validation approach described in Section 3.1 to estimate this ratio. To carry out the cross validation, we had to do some minor preprocessing, removing the two patients with a secondary disease category of colon cancer as otherwise the cross-validation would fail since the training and test data sets would have a secondary disease category variable with a different number of levels.

Figure 8 shows that compared to using the whole population, which is equivalent to a tree of depth 0, the smaller study populations defined by trees of depths $1, \ldots, 10$ show only a small gain in variance reduction for the estimated average treatment effect, less than a 15% gain. Given the interpretability advantages of using the whole population, it seems best to use the whole population rather than the study population defined by a tree. The study population from Crump et al.'s method, i.e., subjects with propensity scores between 0.1026 and 0.8974, does show some gain in the estimated variance of the average treatment effect estimate over the whole population, about a 30% gain[4] It is not clear whether or not this gain in variance would make it worth using the harder to interpret study population defined by propensity scores between 0.1026 and 0.8974 rather than the whole population. The best study population to use might depend on the audience for the study.

5.2 Balance for Treatment on Treated Effect Criterion

As in Section 4.2, we consider choosing a study population for estimating the treatment on treated effect from a matched pair design. For the distance matrix $\boldsymbol{\Delta}$, we use the rank-based Mahalanobis distance as in Section 4.2. Figure 9 and Figure 10 shows that choosing the optimal subset with critical distance equal to the 5% quantile of distances in $\boldsymbol{\Delta}$ provides good balance; all standardized differences are less than 0.1 and the covariate balance is comparable to that of a randomized experiment[5].

We consider fitting a tree to the classification of subjects as in or out of the optimal subset of treated subjects when the critical distance is the 5% quantile. The resulting standardized differences are shown in Figure 9 along with those of the optimal subset. Figure 10 compares the p-values from two-sample tests to the uniform distribution, so worse covariate balance than in a completely randomized experiment leads to points below the diagonal line, whereas better balance leads to points above the diagonal line.

The trees of depths 1, 2 and 3 do not have acceptable balance, with some standardized differences greater than 0.2 (see Figure 9) and worse covariate balance than that expected in a completely randomized experiment (see Figure 10). The tree of depth 4 has acceptable balance, with no standardized differences much above 0.1 and covariate balance similar to that expected in a randomized experiment.

Figure 11 show the study population defined by the tree of depth 4. The study population includes the following subjects: (i) APACHE score greater than or equal to 61.5, respiratory rate less than 25.5 and a primary disease

[4] Crump et al. [8] used the bootstrap to assess the variance of the estimated average treatment effect and reported a similar 36% gain in the bootstrap variance of the estimated average treatment effect from their study population compared to the whole population. The bootstrap requires using the outcome data. We would like to select our study population before looking at the outcome data, which is why we use the cross validation procedure from Section 3.1 in estimating the ratio of the variance of estimated treatment effect for a given study population to the variance of estimated average treatment effect for the whole population.

[5] Rosenbaum [24] used only data on subjects under age 65. For this subset of subjects, Rosenbaum found that choosing the optimal subset with critical distance equal to the 5% quantile of distances in $\boldsymbol{\Delta}$ also provided good balance and focused on this study population.

Table 5 Estimated Treatment on Treated Effect of right heart catheterization for Different Study Populations.

Study Population	Treatment Effect	Standard Error	Number of Pairs
Optimal Subset with Critical Distance = 5% Quan.	-0.068	0.0156	1563
Tree of Depth 4	-0.031	0.0164	1351

category of cirrhosis or COPD; (ii) APACHE score greater than or equal to 61.5 but less than 86.5, respiratory rate greater than or equal to 25.5 and a cardiovascular diagnosis at admission; (iii) APACHE score greater than or equal to 86.5, a respiratory rate greater than or equal to 25.5 and a mean blood pressure greater than or equal to 51.5; (iv) APACHE score less than 61.5, a $\mathrm{PaO_2/FIO_2}$ ratio greater than or equal to 117.57 and a primary disease category of acute respiratory failure, congestive heart failure, cirrhosis, colon cancer, coma, COPD or lung cancer; (v) APACHE score less than 61.5, a $\mathrm{PaO_2/FIO_2}$ ratio greater than or equal to 117.57, a primary disease category of multiple organ system failure (with malignancy or sepsis) and a respiratory rate greater than or equal to 61.5; (vi) APACHE score less than 53.5, a $\mathrm{PaO_2/FIO_2}$ ratio less than 117.57, not transferred from another hospital;

Table 5 shows the estimated treatment on treated effect of RHC on survival to 30 days after admission for the study population defined by the tree of depth 4, along with that of the optimal subset that the tree of depth 4 sought to approximate. The standard error was computed in a way that accounts for the matched nature of the sample using the procedure described in [1]. For both study populations, RHC is estimated to decrease survival; the effect is significant at the 5% level in the optimal subset population and at the 10% level in the study population defined by the tree of depth 4.

6 Discussion

Reliable estimation of a treatment effect requires that there be sufficient overlap between the treated and control group's covariate distributions. In this paper, we have developed a tree approach to choosing a study population definition that has sufficient overlap and is easily described. We have considered using the tree approach to find study population definitions that perform well according to either the criterion of minimum variance of estimated average treatment effect proposed by [8] or balance for treatment on treated effect (combined with as large a sample as possible) proposed by [24]. The tree approach could be used to find study populations that work well for other criteria, such as the criterion of subjects whose propensity scores overlap with those of the opposite treatment group.

For some studies, the tree approach can find a study population that performs similarly to the optimal one according to the minimum variance criterion or the balance of treatment on treated effect criterion, but is much more easily described. Examples are the job training study for both the minimum variance criterion (Section 4.1) and the balance of treatment on treated effect criterion (Section 4.2) and the right heart catheterization study for the balance of treatment on treated effect criterion (Section 5.2). However, the tree approach is not a panacea for finding easily described, close to optimal study populations. For some studies, the tree approach with a small maximum depth will lead to a study population definition that performs considerably worse than the optimal one defined by a criterion such as Crump et al.'s minimum variance criterion [8]. For such studies, we must decide whether the gain in interpretability from an easily described study population outweighs the higher variance. For such studies, we could use a tree with a larger depth, but then the number of groups that are included (excluded) becomes quite large and the study population consists of the union of many highly specific subpopulations.

A potential drawback of the tree method for determining a study population is that it might end up excluding groups that are already underrepresented in research, e.g., certain minority groups. With the propensity score approach to determining the study population, most, but not all, of such populations might be excluded. Future research could consider how to constrain the tree method to not entirely exclude underrepresented groups.

Example R code for implementing the methods developed in our paper is provided in supplementary materials.

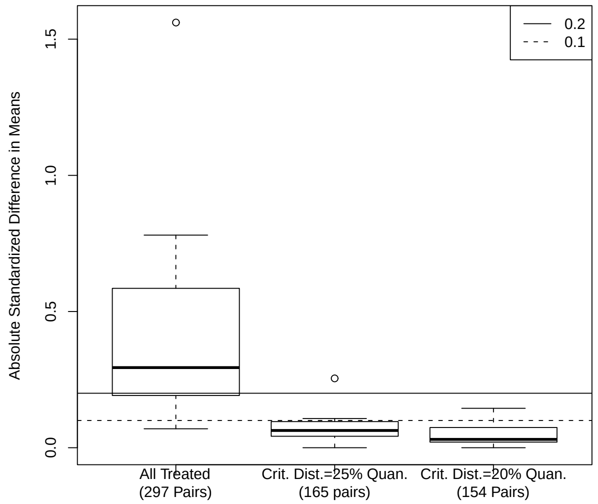
Acknowledgements The authors thank Jing Cheng for insightful suggestions and valuable comments on the paper, and thank Paul Rosenbaum for insightful suggestions. This work is supported by NSF grant SES-0961971.

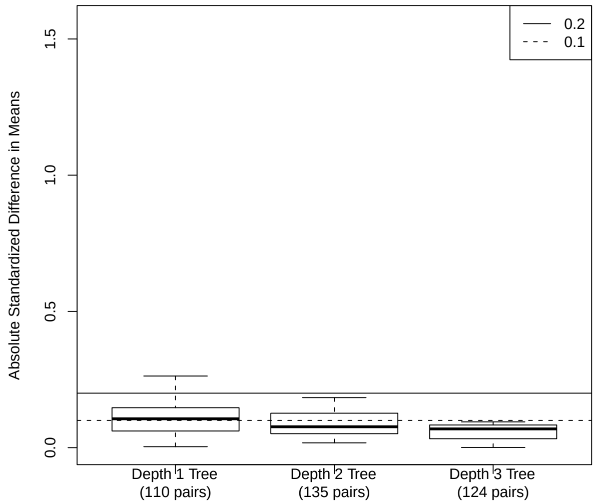
References

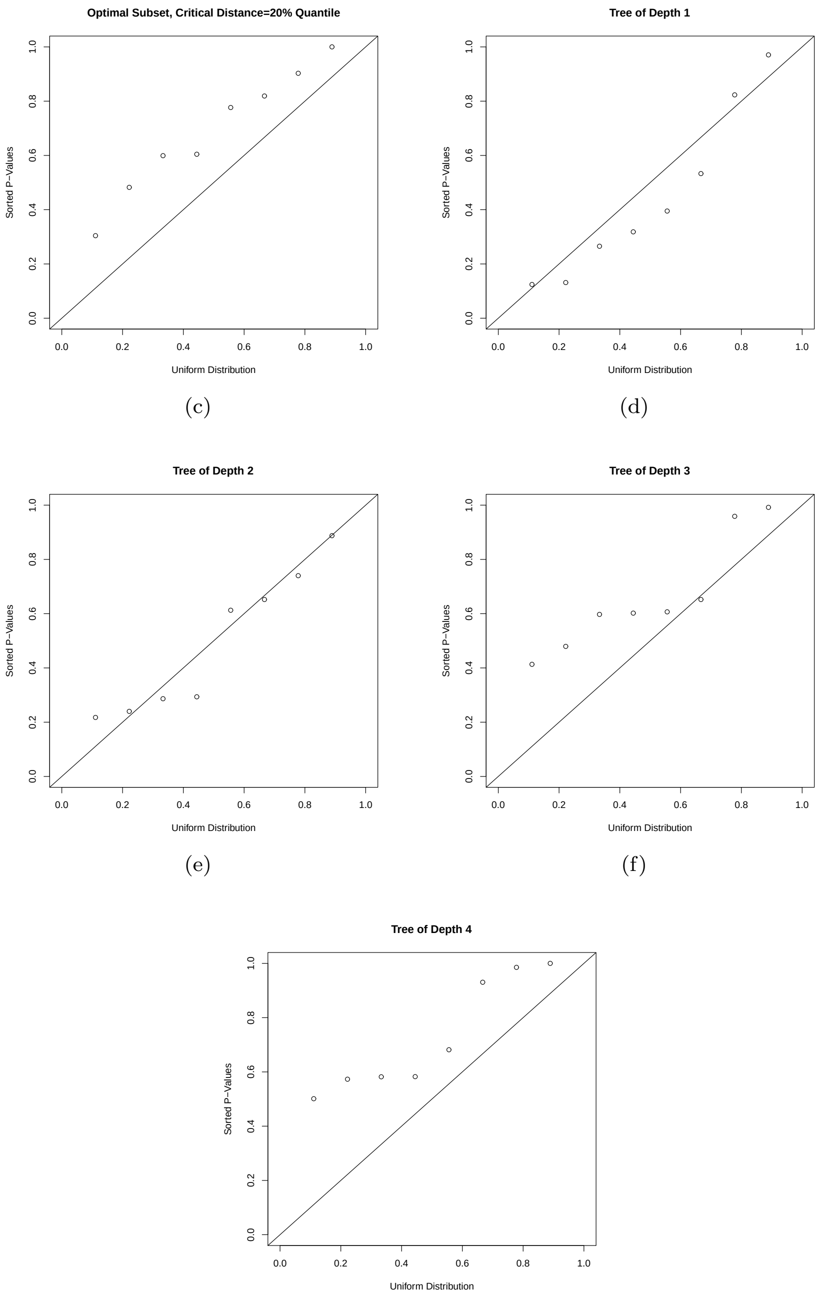
1. Agresti, A. and Min, Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Statistics in Medicine*, 23, 65–75 (2004).
2. Austin, P. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29, 2137–2148 (2010).
3. Breiman, L. Random Forests. *Machine Learning*, 45, 5–32 (2001).
4. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Classification and Regression Trees.* Wadsworth and Brooks/Cole, Monterey (1984).
5. Cochran, W.G. The effectiveness of an adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313 (1968).
6. Cochran, W.G. and Rubin, D.B. Controlling bias in observational studies: a review. *Sankhya Ser. A*, 35, 417–446 (1973).
7. Connors, A.F., Speroff T., Dawson, N.V., Thomas, C., Harrell, F.E., Wagner, D., Desbiens, N., Goldman, L., Wu, A.W., Califf, R.M., Fulkerson, W.J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J. and Knaus, W. A. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, 276, 889–97 (1996).
8. Crump, R., Hotz, V.J., Imbens, G. and Mitnik, O. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96, 187–199 (2009).
9. Dehejia, R. and Wahba, S. Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062 (1999).

10. Grzybowski, M., Clements, E., Parsons, L., Welch, R., Tintinalli, A. and Ross, M. Mortality benefit of immediate revascularization of acute ST-segment elevation myocardial infarction in patients with contraindications to thrombolytic therapy: A propensity analysis. *Journal of the American Medical Association*, 290, 1891–1898 (2003).
11. Guyatt, G., Ontario Intensive Care Group. A randomized control trial of right-heart catheterization in critically ill patients. *Journal of Intensive Care Medicine*, 6, 91–95 (1991).
12. Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331 (1998).
13. Hansen, B. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99, 609–618 (2004).
14. Hansen, B. Optmatch. *R News*, 7, 18–24 (2007).
15. Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer (2001).
16. Hirano, K., Imbens, G. and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 1161–89 (2003).
17. Horvitz, D. and Thompson, D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 46, 663–85 (1952).
18. Imbens, G. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86, 1–29 (2004).
19. LaLonde, R. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76, 604–620 (1986).
20. Robins, J. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122–129 (1995).
21. Rosenbaum, P. Optimal matching in observational studies. *Journal of the American Statistical Association*, 84, 1024–1032 (1989).
22. Rosenbaum, P. *Observational Studies*, 2nd ed. Springer, New York (2001).
23. Rosenbaum, P. *Design of Observational Studies*. Springer, New York (2010).
24. Rosenbaum, P. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, in press.
25. Rosenbaum, P. and Rubin, D. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55 (1983).
26. Smith, J. and Todd, P. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125, 305–353 (2005).
27. Stuart, E. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25, 1–21 (2010).
28. Therneau, T. and Atkinson, E. An Introduction to Recursive Partitioning Using the rpart Routine. Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester (1997). URL http://www.mayo.edu/hsr/techrpt/61.pdf.
29. Vincent, J., Baron, J., Reinhart, K., Gattinoni, L., Thijs, L., Webb, A., Meier-Hellmann, A. Nollet, G. and Peres-Bota, D. Anemia and blood transfusion in critically ill patients. *Journal of the American Medical Association*, 288, 1499–1507 (2002).

Fig. 4 Absolute standardized difference in means between treated and matched control groups for 8 covariates, including the propensity score, for (a) all treated subjects and optimal subsets of treated subjects using a critical distance of the 25% or 20% quantile of distances in $\boldsymbol{\Delta}$ and (b) treated subjects who fall into a study population defined by trees of various maximum depths fit to the classification of treated subjects as in or out of optimal subset of treated subjects using a critical distance of the 20% quantile of distances in $\boldsymbol{\Delta}$.







(c)

(d)

(e)

(f)

(g)

Fig. 5 Quantile-quantile plots comparing the two-sample p-values for 8 covariates, including the propensity score, to the uniform distribution in five matched comparisons

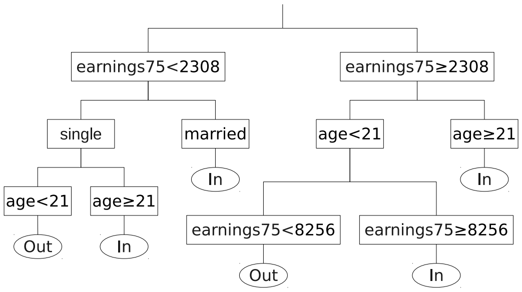


Fig. 6 Best tree of depth 3 for the job training study for the balance for treatment on treated effect criterion.

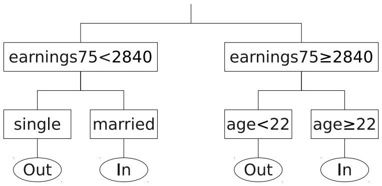


Fig. 7 Best tree of depth 2 with double loss of misclassifying subjects who are out of the optimal subset compared to those in the optimal subset for the job training study for the balance for treatment on treated effect criterion.

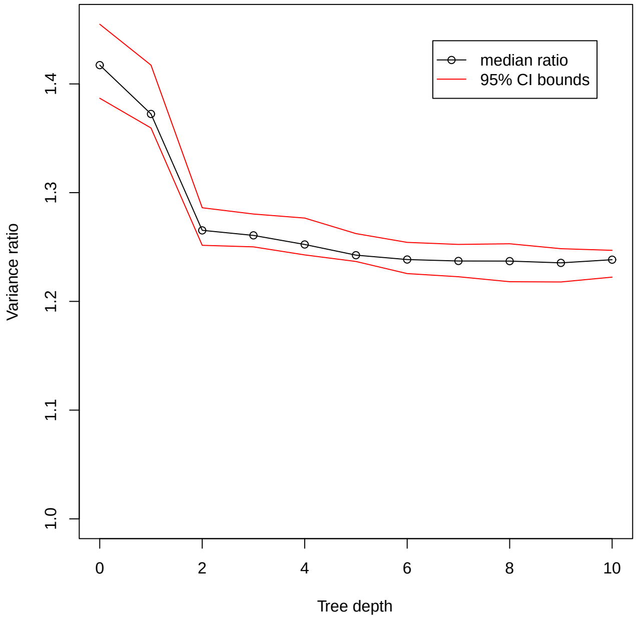
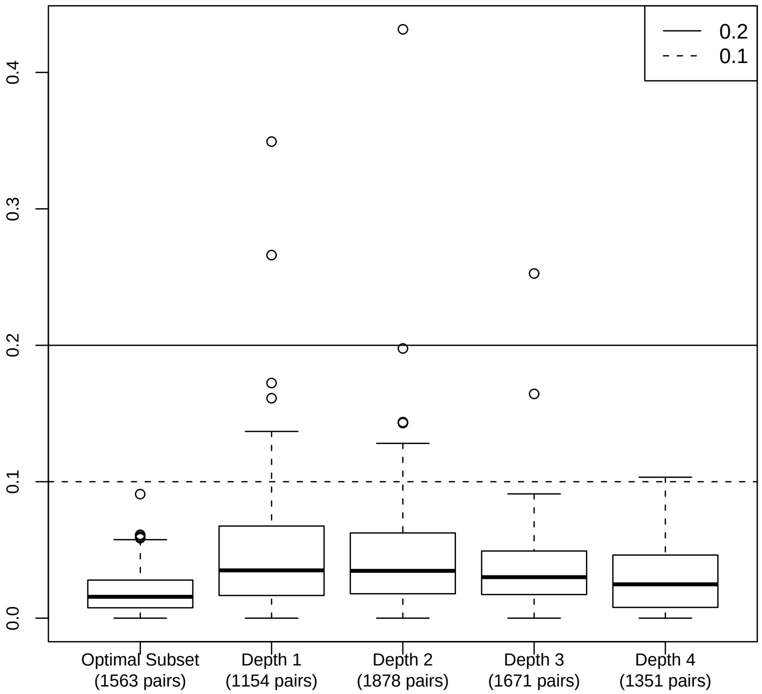
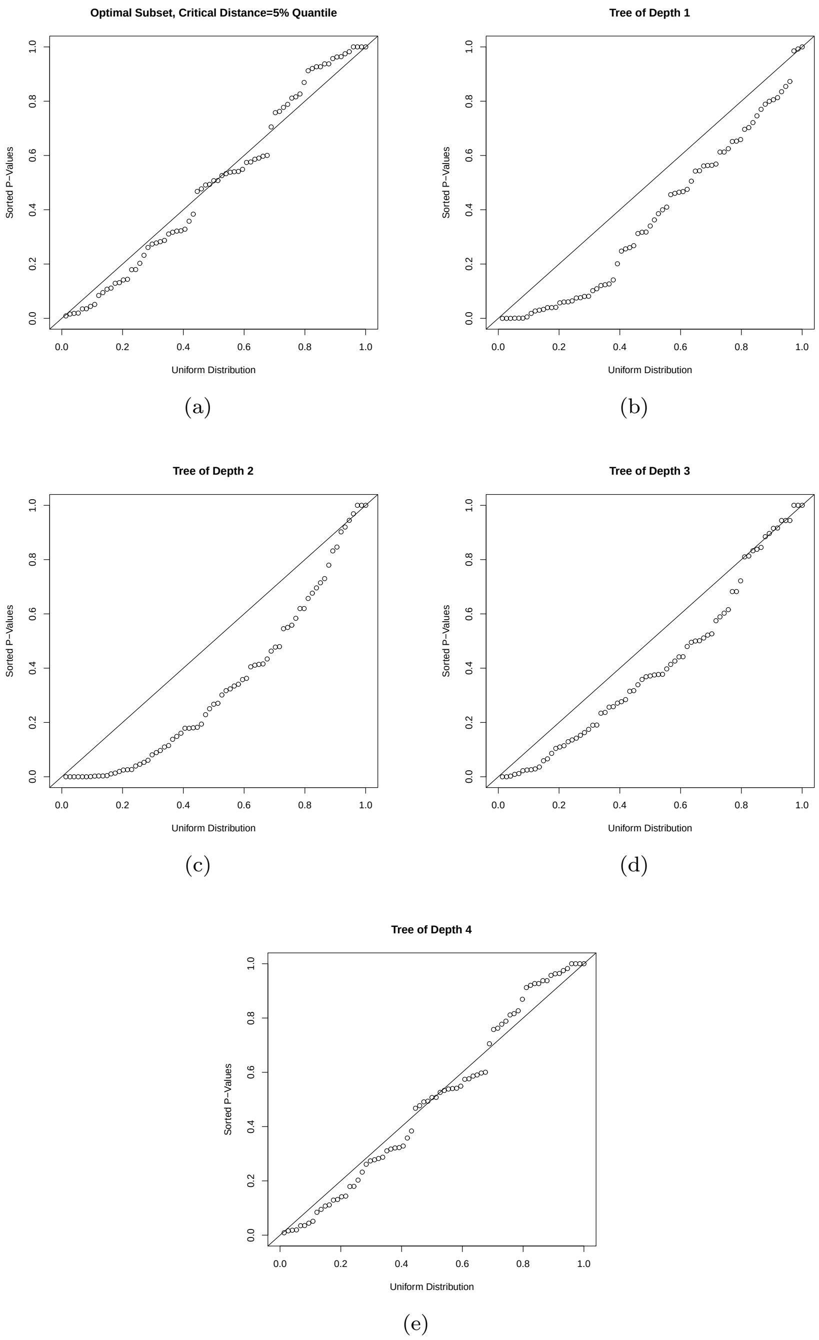


Fig. 8 For the RHC data, estimated ratio of variance of best trees of various depths to that of the estimated optimal subpopulation by Crump et al.'s [8] criterion: $\hat{\mathbb{A}} = \{\mathbf{X} \in \mathbb{X} : 0.1026 \le \hat{e}(\mathbf{X}) \le 0.8974\}$

Fig. 9 For the right heart catheterization study, absolute standardized difference in means between treated and matched control groups for 73 covariates, including the propensity score, for optimal subset of treated subjects using a critical distance of the 5% quantile of distances in $\boldsymbol{\Delta}$ and trees of various maximum depths fit to the classification of treated subjects as in or out of this optimal subset.





(a) (b)

(c) (d)

(e)

Fig. 10 For the right heart catheterization study, quantile-quantile plots comparing the two-sample p-values for 73 covariates, including the propensity score, to the uniform distribution in five matched comparisons

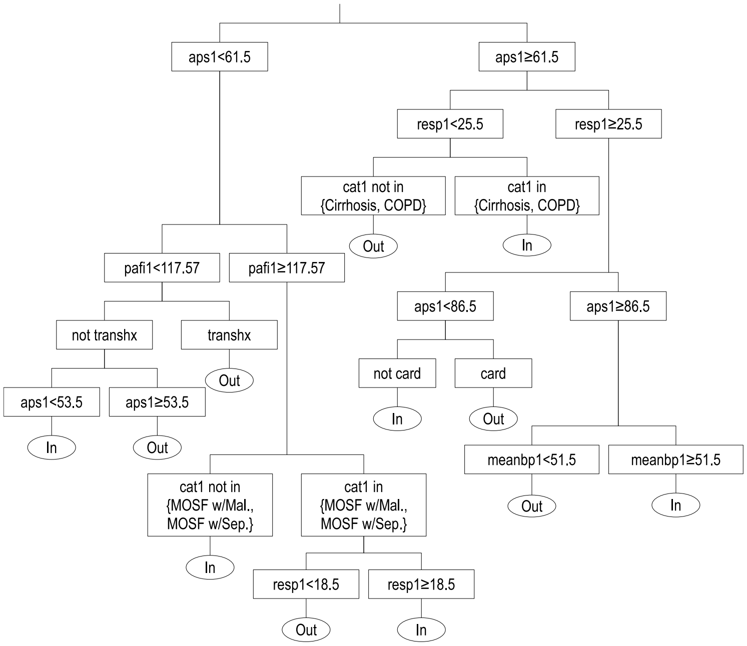


Fig. 11 Best tree of depth 4 for the balance for treatment on treated effect criterion for the right heart catheterization study.