Journal of the Operational Research Society
doi:10.1057/s41274-017-0258-8

AN EVOLUTIONARY ALGORITHM FOR SUBSET SELECTION IN CAUSAL INFERENCE MODELS

Wendy K. Tam Cho[1]

Abstract

Researchers in all disciplines desire to identify causal relationships. Randomized experimental designs isolate the treatment effect and thus permit causal inferences. However, experiments are often prohibitive because resources may be unavailable or the research question may not lend itself to an experimental design. In these cases, a researcher is relegated to analyzing observational data. To make causal inferences from observational data, one must adjust the data so that they resemble data that might have emerged from an experiment. The data adjustment can proceed through a subset selection procedure to identify treatment and control groups that are statistically indistinguishable. Identifying optimal subsets is a challenging problem but a powerful tool. An advance in an operations research solution that is more efficient and identifies empirically more optimal solutions than other proposed algorithms is presented. The computational framework does not replace existing matching algorithms (e.g. propensity score models) but rather further enables and augments the ability of all causal inference models to identify more putatively randomized groups.

Keywords: Causal Inference; Subset Selection; Optimization

[1]Wendy K. Tam Cho is Professor in the Department of Political Science and Department of Statistics and Senior Research Scientist at the National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign.

1 Problem Description

Experimental studies have enormous and unique potential. When an experiment is well-designed and flawlessly executed, the experimental framework isolates the treatment effect and allows one to examine causal effects (Neyman, 1923 [1990]; Fisher, 1935; Cochran and Cox, 1957). When a randomized experiment is not possible, we may hope to gain some traction on obviously important questions through observational data. When feasible, this route is appealing since observational data for a large number of phenomena often abound.

While it would be nice to be able to shift to the abundance of observational data for making causal inferences, this transition is far from trivial or simple. Applying standard statistical models to non-experimental or observational data generally allows the researcher to make associational inferences only. The thorny and complex problem is whether observational data may be adapted in such a way that they may be examined as experimental data that permit causal inferences to be made. If observational data can be successfully adjusted to mimic experimental data and all of the underlying assumptions are satisfied, then we can theoretically derive causal inferences from observational data (Holland, 1986; Rubin, 1974, 1978).

We present a computational model, embodying an advance in an optimization algorithm, for adjusting observational data so that they resemble a randomized experimental framework. We contribute to the potential outcomes literature by designing an efficient and effective evolutionary algorithm for implementing a computational model for making causal inferences. Our model framework was introduced in the statistics literature (Cho et al., 2013) as a way to incorporate the experimental counterpart of re-randomization. Whenever subjects are randomly chosen for a study, that study can be replicated by randomly choosing another set of treatment and control subjects. In any given observational study, the optimization framework presents an opportunity to exploit this replication link by embodying a design that builds replication directly into the statistical adjustment methodology.

To be sure, other causal inference methods could potentially also provide the ability to extract large numbers of solutions that satisfy a particular criteria, though no other method currently does. It would be fairly natural for Zubizarreta (2012) to extend his model in this way since it is also incorporates an optimization model. It would be more difficult and less natural to extract larger numbers of solutions from other methods, though it is possible, perhaps through chang-

ing particular parameters like the Mahalanobis weights, the propensity score model, the size of the calipers, etc. We are not proposing our algorithm as a replacement for all other matching methods—the idea of re-randomization within an optimization implementation can be applied to any causal inference model. The implementation, moreover, can encompass multiple matching models. Cho (2017) demonstrated the value of identifying large numbers of putatively randomized solutions. However, demonstrating the value and developing effective and efficient algorithms are separate research ventures. Our contribution here, the development of an algorithm that identifies a large number of putatively randomized groups in a more efficient and effective manner than any other existing algorithm, is in the operations research realm but advances the ability of statistical models to identify causal inferences.

1.1 Causal Inference via Observational Data

We begin by describing the causal inference problem and introducing notation. Experiments can be complex and multi-faceted, but let us assume, for simplicity, that a subject is either treated ($T = 1$) or not ($T = 0$). For subject i, $i = 1, \ldots, N$, the two potential outcomes are $Y_i(0)$ and $Y_i(1)$, where $Y_i(0)$ represents the dependent variable or outcome for subject i under the control condition and $Y_i(1)$ represents the dependent variable or outcome for subject i under the treatment condition. The causal effect of the treatment, as measured by the outcome or dependent variable, Y, on a particular subject i, is

$$Y_i(1) - Y_i(0). \tag{1}$$

The fundamental problem of causal inference is that it is impossible to observe the value of both $Y_i(1)$ and $Y_i(0)$ on the same subject because the subject has either been exposed to the treatment or has not. Since only one of the terms in (1), either outcome under treatment or outcome under control, is observable for a single unit, the expression cannot be evaluated (Holland, 1986).

The Rubin causal model reconceptualizes this framework so that *either* the outcome under treatment or under control, but not both, needs to be observed for each unit (Rubin, 1974, 1978). That is, one statistical solution to the fundamental problem of causal inference is to shift from an examination of individual treatment effects to an *average* treatment effect (ATE) over *all* of the subjects,

$$ATE = E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0)). \tag{2}$$

This alleviates the fundamental problem of causal inference because we no longer require two observations, $Y_i(0)$ and $Y_i(1)$, from one unit, i. Instead, we need only one of these observations.

If the outcome under control for unit i, $Y_i(0)$, is observed, we use that information to inform the average treatment effect under control, $Y(0)$. If the outcome under treatment for unit i, $Y_i(1)$, is observed, we use that information to inform the average treatment effect under treatment, $Y(1)$. We forego an ability to measure the treatment effect for a particular individual, $\tau_i = Y_i(1) - Y_i(0)$, in exchange for a measure of the average treatment effect across a range of individuals.

A critical issue for observational studies arises from the non-random nature of the subjects in the data set. One observes some set of subjects who have received a treatment, giving us $E(Y(1) \mid T = 1)$ rather than $E(Y(1))$. From this observed group, the average treatment effect *for the treated* (ATT) is

$$ATT = E(Y(1) - Y(0) \mid T = 1), \tag{3}$$

which quantifies the effect of the treatment on subjects that are treated. Because $E(Y(1)) \neq E(Y(1) \mid T = 1)$ and $E(Y(0)) \neq E(Y(0) \mid T = 1)$, the average treatment effect, $E(Y(1)) - E(Y(0))$, and the average treatment effect for the treated, $E(Y(1) \mid T = 1) - E(Y(0) \mid T = 1)$, are not generally interchangeable.

The ATE and the ATT would be interchangeable if the independence assumption—exposure to treatment is statistically independent of all other variables, including $Y(1)$ and $Y(0)$—holds because conditioning on treatment is then irrelevant. This allows us to compute the ATE as $E(Y(1) \mid T = 1) - E(Y(0) \mid T = 1)$, but we must still determine how to compute $E(Y(0) \mid T = 1)$, the untreated outcome for treated individuals. Notice here that if treatment is completely random, then a viable approach is to use the average outcome of similar subjects who were not exposed to treatment. We would then no longer require an observation of $Y_i(1)$ and $Y_i(0)$ from the *same* subject, but are able to use information from *different* subjects. If exposure to treatment satisfies the independence assumption, those who have been treated give us information about $E(Y(1))$, while those who have not been treated give us information on $E(Y(0))$ allowing the treatment effect can be calculated as

$$ATE = ATT = E(Y \mid T = 1) - E(Y \mid T = 0) = \frac{1}{N_t} \sum_{i \in \{T=1\}} Y_i(1) - \frac{1}{N_c} \sum_{i \in \{T=0\}} Y_i(0), \tag{4}$$

where N_t is the number of treated subjects, N_c is the number of control subjects, $\{T = 1\}$ denotes the set of treated subjects, and $\{T = 0\}$ denotes the set of control subjects.

With observational data, it would be very unusual for the independence assumption to hold; the treated group almost surely differs systematically from the non-treated group. Hence, if one wishes to make causal inferences from observational data, the task at hand is to adjust observational data so that exposure to treatment satisfies the independence assumption. If this difficulty can be satisfactorily accomplished, the adjusted data will resemble a randomized experiment, and one can then compute the treatment effect in a straightforward manner.

In practice, adjusting observational data usually involves two population groups, treated (observed subjects who have a particular attribute) and control (observed subjects who do not have a particular attribute), and a set of pre-treatment covariates, $\mathbf{X}$. The objective is, given the observed treatment group, to identify a subset of the control pool so that the covariate distributions of the treated and chosen control group are statistically indistinguishable, creating the "appearance of randomization." If treatment is completely random, as measured by covariate randomization tests, we assume that the *unconfoundedness* or the *selection on observables* (SOO) assumption is satisfied. Formally, if

$$\textbf{Assumption 1: } P(T, Y(0) \mid \mathbf{X}) = P(T \mid \mathbf{X})\, P(Y(0) \mid \mathbf{X}), \tag{5}$$
$$P(T, Y(1) \mid \mathbf{X}) = P(T \mid \mathbf{X})\, P(Y(1) \mid \mathbf{X}), \text{and}$$
$$\textbf{Assumption 2: } 0 < P(T = 1 \mid \mathbf{X} = x) < 1, \tag{6}$$

hold, we claim evidence that the treatment assignment is "strongly ignorable" (Rosenbaum and Rubin, 1983). Assumption 1 states that the treatment is independent of the outcome, conditional on covariates, $\mathbf{X}$. The driving goal for adjusting observational data is to create data sets so that treatment assignment is strongly ignorable, i.e. adjust observational data to resemble randomized experimental data. Assumption 2 concerns the issue of overlap or common support, an important assumption in the development of causal inference methods. Later, we will discuss how this assumption lends an advantage to our model's framework over existing matching methods because we infuse greater flexibility in how we model the underlying experimental design.

To be sure, successfully adjusting the data is not simple and, furthermore, not always possible. One important obstruction is that we are limited to working with *observable* covariates. This selection on observables constraint presents a formidable problem in causal inferences models with observational data. Indeed, one cannot even be certain whether the assumption is satisfied in any particular analysis. When randomization is successful, all variables, observable and unobserv-

able, have the same distribution, in expectation, between the control and treatment group. When we attempt to mimic this "covariate balance" between the treatment and control groups in an observational data set, it is only possible to balance variables that are observed. Working solely with observed variables may result in balancing unobserved variables, but since there are unobserved variables, we can guess, but not know, whether the unobserved variables have fortuitously been balanced as well. This problem cannot be overcome by any causal inference method for observational data; and we do not purport to have solved this problem here. Neither statistical nor computational models can solve this type of data problem where unobserved covariates confound the analysis. Instead, we proceed as others methods proceed, with an acknowledgement of this pitfall and full awareness of the issues that then arise in interpreting the results of the model when this assumption is not satisfied. As with other models, the selection on observables assumption is presumed to be satisfied for our model to be useful. Causal inferences are only as valid as the assumptions upon which the model is built.

2 Matching versus Subset Selection Methods for Causal Inference

Traditionally, this adjustment of observational data so that they resemble experimental data by simulating statistical independence of treatment exposure and all other available variables is done via methods that fall under the term "matching" (Rubin, 1974, 1977, 1978). The term matching is used because the method attempts to match each treatment unit to a control unit that has attributes that are as similar as possible to that treatment unit. In this quest, the first step in matching is to calculate a distance or similarity measure for any two observations. The smaller the value of the distance metric between two observations, the more similar the two observations are to one another. The particular metric used by a matching method varies but some metric is essential in these methods. Once all pairwise distances have been computed, the next step is to match treatment units to control units. The matching may be done in a variety of ways, perhaps through greedy matching where each treated unit is sequentially paired with its closest non-matched control unit or a network flow optimization routine that seeks to minimize the total sum of pairwise differences (Rosenbaum, 1989). Once the treated observations are matched to control units, the chosen control units comprise the control group. At this point, some set of balance statistics are computed to provide a measure for how well the objective of adjusting the data to resemble an experiment has been achieved. The covariate distributions of the treated and control groups are

compared to one another and an assessment is made for how closely the treatment and control covariate distributions resemble one another. If the matching does not result in sufficient covariate balance, the entire procedure is repeated, perhaps with a different distance metric or another pairwise matching procedure with some parameter changed. The cycle is repeated until, ideally, sufficient balance is obtained so that the data may reasonably be statistically regarded as data that might have emerged from a randomized experiment.

Recently, another approach to achieving covariate balance, Balance Optimization Subset Selection (BOSS), was proposed (Cho et al., 2013). Specifically, instead of matching treatment units to control units, which mimics a pair randomized experiment, one could mimic the completely randomized experimental framework through a subset selection procedure. While matching methods have favored pair randomized designs, there is no particular reason to restrict one's search to pair randomization designs since there are many other valid experimental frameworks. The logic of the "matching" procedure is not well-aligned with a completely randomized experimental design since the completely randomized research design does not embody control units that are matched to a particular treatment unit. Instead, the randomization process results in control and treatment *groups* where the distributions of covariates for the two groups are similar. The subset selection procedure aims to identify two subsets/groups with similar covariate distributions, i.e. mimic the completely randomized experimental framework.

In either a pair randomized framework or a completely randomized framework, the resulting control and treatment *groups* have covariates values that are balanced. The intent of any causal inference method for observational data is to adjust observational data so that covariate balance between a treatment *group* and a control *group* is realized. Irrespective of the adjustment method or the framework that is being emulated, the end goal is the same: covariate balance. Since the end goal is unquestionably covariate balance, it makes sense that an adjustment procedure would optimize directly on covariate balance. With matching methods, the procedure is indirect; these methods match individual observations based on a distance metric and hope that this results in a control *group* with covariates that are balanced with those in the treatment group. This reasoning is logical and may result in covariate balance, but the distance metric in this framework is superfluous and unnecessarily constrains the solution space to pair randomized designs.

We also note that the subset selection design affords some advantages for the overlap assumption (6). In a pair matched design, if a treatment unit is unusual, there may not be a suitable or close

match for that unit. This might lead us to omit this unit from the analysis because it is outside the overlap region. When we transition to examining the covariate distributions for the entire group, however, we also create greater flexibility—a covariate distribution may be sufficiently similar to another covariate distribution even when there is not a paired match for each individual treatment unit. The impetus to change the parameters of the problem because of the overlap assumption, then, becomes less imperative. The subset selection framework is more pliable in this regard. In addition, one might also base a decision about imposing overlap on computational concerns. The BOSS framework is computationally intensive, and imposing overlap may significantly reduce the computational burden if a large number of observations in the control pool are substantially different from any of the treatment group units. For example, Dehejia and Wahba (1999) discarded 12,611 of the 15,992 observations in the control pool after imposing overlap based on the propensity score. Hence, another route is to impose overlap and then use the BOSS framework on the reamining observations that satisfy overlap. These are modeling decisions that should be made carefully and thoughtfully by the researcher, not imposed by the model.

3 Implementing a Computational Model

The theoretical framework of BOSS is fundamentally different from the rest of the matching literature. It seeks to adhere to a completely randomized experimental design. Because it foregoes the pair randomized experimental design, BOSS bypasses the need for a distance metric and is able to optimize directly on covariate balance. Nikolaev et al. (2013) provided a convincing proof of concept for the approach. At the same time, they emphasize that the concept has not yet realized its potential.

> Note that the main contribution of this paper is conceptual and theoretical. The goal . . . is to present the problem of causal inference in a new light, opening up a field where optimization tools developed within the operations research community can make an impact. By motivating and formalizing an alternative approach to a problem of great importance to multiple domains of modern science, this paper is intended as a seed for more applied, computational-oriented literature. Section 3 is not meant to be comprehensive; instead, it positions itself to illustrate that the proposed theory can shift the problem at hand into the computational realm. It is not intended to deliver comprehen-

sive numerical achievements, but rather supports the call for more intense, goal-driven computational research of BOSS.

This paper picks up where the Nikolaev et al. (2013) research left off, with the call for a more efficient optimization algorithm and implementation that is true to the underlying BOSS theory. We present an evolutionary algorithm that optimizes directly on covariate balance. Importantly, unlike the Nikolaev et al. (2013) implementation, our optimization algorithm does not rely on a reduction of the size of the solution space to make the optimization feasible. It also pays particular attention to the tradeoffs when the objective function includes many different independent but inter-related components.

To gain a sense of the computational complexity for implementing BOSS, consider the solution space in question. The idiosyncrasies of the solution space for a particular type of problem often point toward a specific search procedure that is likely to be more efficient. In the subset selection problem, the particular characteristics of the solution space are not difficult to identify. First, the solution space is enormous. If there are 10,000 units from which to choose a subset of 200, there are $\binom{10000}{200} \approx 1.71 \times 10^{424}$ possible subsets. Second, at a rough level of granularity, the solution space is rugged. At a finer level of granularity, the solution space is flat. That is, while the solution landscape is hilly in the sense that it has the usual peaks and valleys, these peaks and valleys are not a rapid succession of precipices, but instead, a series of vast plateaus, and hence, not rugged in the usual sense. These expansive plateaus manifest themselves throughout the landscape because many subsets have significant overlap. It is evident then that if one swaps out a single observation from a subset, the new subset is substantially similar and the covariate balance does not change significantly. For any subset, there is a slew of such minor modifications. A stochastic or random element should improve the solution search significantly and is necessary to move from one wide-ranging plateau in the solution space to another.

Simulated annealing is one way to insert a random element into the search process and is the method implemented by Balance Optimization Subset Selection with Bins or BOSS-B (Nikolaev et al., 2013). BOSS-B begins by binning the data. Each unit is converted from a vector of covariate values $\{X_{1i}, X_{2i}, \ldots, X_{Ki}\}$ into a vector of bin numbers $\{X'_{1i}, X'_{2i}, \ldots, X'_{Ki}\}$ where $X'_{ki} = j$ if and only if $t^k_{j-1} \leq X_{ki} \leq t^k_j$ (i.e., unit i falls into bin j for covariate k). The bins are uniformly spaced across the covariate distributions. After binning the data, an initial solution is generated by randomly choosing n units from the control pool. At each iteration, the algorithm attempts a

1-exchange, replacing one unit in the control group with an unselected unit in the control pool. If the exchange improves the objective function, it is accepted unconditionally. If the exchange does not improve the objective function, then it is accepted with some probability according to a set of simulated annealing parameters. A random restart is invoked when little progress has been made after some number of iterations or a perfectly optimized control group is identified. The algorithm terminates after a pre-defined number of iterations.

Formally stated, the BOSS-B problem is formulated as follows.

Given: K covariates; a fixed integer N; set $\mathscr{S}^{\mathscr{T}}$, randomly selected from set $\mathscr{T}$ of units represented by vectors $\{X_{1u}, X_{2u}, \ldots, X_{Ku}\}$, $u \in \mathscr{T}$, with $|\mathscr{T}| = N$; set $\mathscr{C}$ of units represented by vectors $\{X_{1u}, X_{2u}, \ldots, X_{Ku}\}$, $u \in \mathscr{C}$, with $|\mathscr{C}| > N$; a set of covariate clusters $\mathbf{D}$; bins $\mathbf{B}^D$ for each $D \in \mathbf{D}$.

Objective: Find subset $\mathscr{S}^{\mathscr{C}} \subset \mathscr{C}$ of size N, that minimizes

$$\sum_{D \in \mathbf{D}} \sum_{b=1}^{|\mathbf{B}^D|} \frac{(|(\mathscr{S}^{\mathscr{C}}, B_b^D)| - |(\mathscr{S}^{\mathscr{T}}, B_b^D)|)^2}{\max(|(\mathscr{S}^{\mathscr{T}}, B_b^D)|, 1)} \tag{7}$$

where $|(\mathscr{S}^{\mathscr{X}}, B_b^D)|$ denotes the cardinality of the set of elements in group $\mathscr{X}$ that lie within bin B_b^D.

For the identified solutions, covariate balance is assessed from the actual (not binned) covariate values after the optimization routine has completed. Balance was assessed for identified solutions in Nikolaev et al. (2013) via statistical tests on first and second moments as well as the Kolomogorov-Smirnov distributional test.

3.1 BOSS-B versus Propensity Score Matching

The BOSS-B implementation performs favorably when compared with existing matching methods in the literature. A comparison with a propensity score model is provided by Nikolaev et al. (2013). They used the R Matching package to obtain the results for matching on a propensity score. A standard logistic regression model with first-order terms was used to estimate the propensity score. The data consisted of 3 covariates and a treatment group of size 500 and a control pool of size 10,000. The treatment group was chosen nonrandomly from a treatment pool of size 5,000. Individuals with covariate values in the tails of the distribution were chosen with higher probability than those with values closer to the mean of the distribution. There was no treatment.

Table 1: Comparison of single best solution from BOSS-B with Matching Methods

Objective Function	Obj Function Value	Treatment Effect	Kolmogorov-Smirnov Mean	Kolmogorov-Smirnov Maximum
BOSS-B: Obj Fcn from Equation (7), 32 bins	0.00000	−0.1142	0.025	0.026
BOSS-B: DOM	0.00002	−0.9877	0.093	0.118
BOSS-B: DOM and DOV	0.00038	0.0271	0.062	0.088
BOSS-B: DOM^2 and DOV	0.00027	0.1154	0.045	0.060
R Matching package: Propensity score model	NA	−1.3434	0.125	0.158

DOM = Difference of Means, DOV = Difference of Variances
Reproduced from Table 7 in Nikolaev et al. (2013)

The results from the experiment performed in Nikolaev et al. (2013) are shown in Table 1. As we can see, the estimate from the propensity score model was the furthest from the true treatment effect value of 0. BOSS-B, using four different objective functions all produced better results than the propensity score model. The improvement of BOSS-B over the propensity score model is likely due to the difficulty of properly specifying a propensity score model. A propensity score model that more closely estimates treatment propensity would perform better though, of course, in practice, it is not straightforward to identify a proper propensity score model. To be clear, we are not suggesting that propensity scores models are not valid causal inference models. We are only making the unremarkable claim that as with all statistical models, when the underlying assumptions hold, the model is most useful as it performs as expected. As stated in the literature, if the estimate of the propensity score is not good, then matching on a bad propensity score estimate will be problematic (Rosenbaum and Rubin, 1983). In a propensity score model, a propensity score must be obtained via some statistical model.

By encompassing the completely randomized experimental design, BOSS affords some advantages. Because BOSS focuses on the covariate distributions and optimizes directly on the entity of interest, covariate balance, it circumvents the necessity to match each treatment unit to a control unit, obviating the need to estimate a propensity score and thus removing a source of uncertainty as well as human bias in model specification. In addition, BOSS is implemented via an optimization framework, which allows it to seamlessly collect a large number of solutions satisfying some threshold. Propensity score models present a single matched solution, which makes it difficult for it to adopt a re-randomization frame.

Experiments may be replicated many times, each time producing a different realization of the ATE random variable. Any identified as-if-randomized group, whether from BOSS or a propen-

sity score model, is also only one realization of the ATE random variable. The question is not what method provides the single best solution, but how can we use multiple valid estimates of the same random variable. One could repeatedly estimate different propensity score models (or Mahalanobis metric matching, or another type of matching) in an ad hoc manner to provide multiple estimates or one can use a method like BOSS that is designed to automate and expand our capabilities in this regard. It would be highly unusual for an ad hoc procedure to be as effective or efficient as BOSS where the computational considerations are central to the implementation.

3.2 Improving Computational Efficiency

The BOSS framework embodies particular advantages but presents an *NP*-Hard optimization problem. Nikolaev et al. (2013) presented one implementation but clearly did not intend the BOSS-B framework with bins to be a final or complete implementation. Instead, it was created to provide a proof of concept for the BOSS theory and reasoning behind the subset selection procedure—using the subset selection framework, BOSS is able to identify as-if-randomized groups. The BOSS-B binning simplification made the solution search feasible. Their initial attempts to optimize directly on covariate balance measures were sufficiently inefficient that they were unsuccessful. They found that "close matches become difficult to find as the number of covariates increases" and that "using the KS score instead of [the] objective [function] caused the search process to stall and fail to make significant progress. . . suggest[ing] that a 1-exchange neighborhood is insufficient when used in conjunction with the KS score," pointing toward a clear need for research into increasing the efficiency of the optimization algorithm. Even coarsening the data with bins was not sufficient to overcome the computational/algorithmic inefficiency. Their empirical results demonstrate that four bins resulted in fast convergence but poor results while the algorithm using thirty-two bins was slower and failed to converge for some data sets.

Theoretically, the binning implementation is problematic because while it makes the computation feasible, the resulting estimates may be biased. To see this, assume that the treatment group and the control pool are random samples of their respective populations. We wish to estimate the difference in an outcome variable, $Y_t - Y_c$ for two groups, t and c. Assume that treatment is a function of only one continuous random variable x. Let $f_g(x)$ be the pdf of x, for units in group g. The mean of x in group t is

$$\mu_t = \int x f_t(x) dx. \tag{8}$$

The mean of x in group c is

$$\mu_c = \int x f_c(x) dx. \tag{9}$$

We may restrict the integrals (8) and (9) to points where $f_g(x) > 0$.

However, we do not have the exact values for the continuous variable, x. Instead, we have values for only a discrete variable that has been converted from the continuous variable by the following rule. Suppose $f_t(x)$ and $f_c(x)$ are partitioned into $b > 1$ equally sized partitions over the range $[x_{\min}, x_{\max}]$. The width of each of these partitions or "data bins" is then $w = \frac{x_{\max} - x_{\min}}{b}$, so the data distribution is partitioned at j_i, $i = 0, 1, \ldots, b$, where $j_0 = x_{\min}$, $j_b = x_{\max}$, and $j_i = j_0 + iw$. For any partition, the mean of the partition is

$$\overline{x}_j = \frac{\int_{j-1}^{j} x f_g(x) dx}{P_{gj}}, \qquad j = 1, 2, \ldots, b \tag{10}$$

where $P_{gj} = \int_{x_{j-1}}^{x_j} f_g(x) dx$. Any value, x_i, that falls in the range $(x_{j-1}, x_j]$ is assigned the value $\overline{x}_j$. If the data are placed into bins in this way, the difference or measurement error, $|\overline{x}_j - x_i|$, for each unit, x_i is bounded, $0 \leq |\overline{x}_j - x_i| \leq \frac{w}{2}$. As the number of bins, b, increases, the measurement error tends to zero.

$$\lim_{b \longrightarrow \infty} \frac{x_{\max} - x_{\min}}{2b} = 0. \tag{11}$$

When $b = \infty$, our bins are infinitesimally small which is equivalent to not binning or using the actual covariate values. Using binned data results in a (bounded) biased estimate of the ATE while using actual covariates values yields an unbiased estimate. BOSS-B used binned data to make the computational problem manageable. Optimizing using actual covariate value is clearly preferred, but also requires us to improve the efficiency of the optimization algorithm.

We note that there is no binning bias for categorical variables (e.g. gender), since we assume that units that are members of the same bin are identical with respect to that variable. In this case, as long as the number of categories for the variables is not too large, the BOSS-B algorithm is sufficient. When our data include continuous variables or ordered variables (e.g., income, age, socio-economic identifiers), however, if the units in a particular bin do not have identical underlying values, BOSS-B will yield a biased estimate.

4 Evolutionary Algorithm, E-BOSS

We pursue an implementation via an evolutionary algorithm (E-BOSS) that improves efficiency, allowing us to use the actual covariate values instead of binned values. E-BOSS bears similarities as well as dissimilarities with the BOSS-B implementation. Both algorithms are optimization models as the BOSS theory fundamentally embodies a computational/optimization model. They both seek to identify optimal control subsets. Our evolutionary algorithm is similar to simulated annealing in that we also employ probabilistic criteria to select and move between candidate solutions. BOSS-B employs a simulated annealing framework while E-BOSS embodies an evolutionary algorithm.

Some differences between the two approaches stem from the initial optimization framework choice. For instance, BOSS-B begins with one initial solution. E-BOSS begins with a large number of initial solutions (the initial population). E-BOSS explores many more solutions simultaneously. BOSS-B moves along the solution space via a 1-exchange mutation operator. E-BOSS includes a richer array of possible moves that allow us to search more of the solution space with more efficient operators. It includes the probabilistic mutation or 1-exchange operation, an n-exchange mutator, a crossover operator that enables larger movements to potential solutions, and an evolutionary framework for transforming the initial population to future generations. The increase in efficiency that results from these changes permits the use of actual covariates values in E-BOSS rather than the binned/rounded covariate values necessitated by the less efficient BOSS-B method.

4.1 Evolutionary Algorithm Implementation

In our subset selection problem, let the number of units in the control pool be N, and the number of treated units be n. In an actual randomized control trial, subjects are randomly chosen to participate, and then randomly assigned to either the treatment group or the control group. This results in roughly half the subjects being in one group and half in the other. Of course, there is no such guarantee as this 50–50 split occurs in expectation. In our setup, we incorporate this expectation by constraining the number of chosen control units to be the same size as the number of treated units. That is, we wish to choose n units from the N total control pool units so that the covariate balance between the n treated units and the n chosen control units is maximized. Obviously, the relative size of the control and treatment groups can be set in any way.

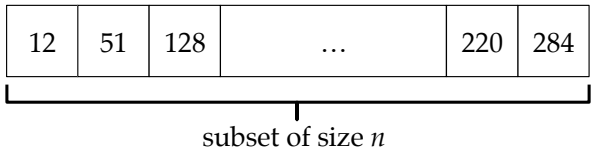

Figure 1: Chromosome encoding

Because the order in which subjects are chosen and placed into the treatment and control groups is inconsequential, we can begin by indexing the N total units arbitrarily and consecutively (from 1 to N). In the evolutionary algorithm, the chromosome is encoded as a list of n distinct integers (representing the indices) from the set $\{1, \ldots, N\}$. Figure 1 depicts our chromosome encoding, the basic data structure of our evolutionary algorithm. Each allele holds the index number of a chosen control unit.

Our evolutionary algorithm consists of three basic operators: reproduction, mutation, and crossover. Together, these operators define how the chromosomes will evolve from generation to generation. Each of these operators has a probabilistic nature, which can be modified and adapted for each particular application.

All of the chromosomes in the initial population may be randomly generated. Alternatively, some of the population of initial chromosomes may be seeded by chromosomes that are known to have good fitness values. The alleles of the chromosomes indicate which control units are chosen for the control group. The treatment group is fixed and consists of all treated units. The fitness of a chromosome represents *one* measure of balance between the treatment and control groups. One way to define the objective function, useful for many causal inference data sets, is

$$b = \sum_{i=1}^{C} w_i \left(KS_i + \mid t_i \mid + \left| \frac{\sigma_{ti}^2}{\sigma_{ci}^2} - \frac{\sigma_{ci}^2}{\sigma_{ti}^2} \right| \right) . \tag{12}$$

where i indexes the variable, C is the number of variables, w is a weight, KS_i is the Kolmogorov-Smirnov statistic, t is the t-statistic for the difference of means, and σ_t^2 and σ_c^2 indicate the variance of the treatment and control groups, respectively. The weights help guide the search routine to areas of the solution space that are less likely to be traversed with uniform weights. Larger weights lead the algorithm to work harder to balance substantively important covariates over less substantively important covariates. Some covariates may also be more difficult to balance, so the weights

also aid in distributing the balance more equally among all covariates. Adjusting the weights also helps the algorithm to expand the search to particular parts of the space that may otherwise be more difficult to locate.

4.2 Role of the Objective Function and Balance Function

Note that the particular objective function is flexible and does not define or limit the evolutionary algorithm. This specification is modular, incorporated into the objective function; different formulations are easily slotted in and out of the algorithm. Researchers are free to and should alter the objective function as they wish to match the particulars of their application. Indeed, there is no claim that Equation (12) represents a balance or objective function that should be used across all applications. It is simply one way to measure distributional similarity that happens to be effective for the application that we will present. We will see later that it results in a close fit between the control and treatment distributions for our particular data set and application. It may work well for other applications, but that decision should be made by the researcher. We do not intend to enter, nor is it necessary for us to enter, a debate on the proper balance function (Begg, 1990; Raab and Butcher, 2001; Imai, King and Stuart, 2008; Hansen and Bowers, 2008). As that literature unfolds, it is simple to extract its insights for particular applications and adapt them into our optimization framework.

It is also important to realize that the objective function and balance assessment are separate and distinct. The objective function guides the optimization routine in the search process and provides one means for identifying control units that ultimately will be well balanced with the treatment units. Balance, on the other hand, is assessed after solutions are identified. Balance was assessed for the BOSS-B solutions based on statistical tests using the actual covariate values, not via the objective function value. Control units were *identified* via the objective function specification but *assessed* via various balance measures. The intention of our algorithm is to identify solutions that perform well on any set of balance measures.

4.3 Process

Once the initial population is set and fitness values are computed, the evolutionary process is ready to begin. Each process involves three steps: reproduction, mutation, and crossover. Reproduction always occurs while mutation and crossover occur with some set probability. Mutation

usually occurs with low probability while crossover usually occurs with high probability. Both probabilities are set with tuning parameters that the researcher adapts to a particular application.

Reproduction. The purpose of the reproduction phase is to choose fit chromosomes to be parent chromosomes that may combine to produce a hopefully fitter child chromosome that will populate the next generation. Our reproduction scheme consists of three options which are user specified. The first option employs the notion of the roulette wheel where the slots of the wheel are proportional to the fitness values of the chromosomes. The roulette wheel is spun to identify parents that will enter the mating process. Since parents with higher fitness values occupy more area in the roulette wheel, this process results in highly fit chromosomes having higher numbers of offspring in succeeding generations. A second option involves a binary tournament. Here, two parents are chosen at random for the tournament. The one with the higher fitness wins. This process is repeated to choose the second parent. Lastly, parents chromosomes may be chosen randomly. Once parents are selected for the mating pool by one of these three methods, the algorithm moves to the crossover phase. Generally, the roulette wheel option works best for identifying promising regions in the solution space. The other options become more enticing as the population becomes more homogeneous since they are less likely to lead to premature convergence of the algorithm at that point.

Crossover. The crossover phase takes two pre-chosen parents and mates them, resulting in a child chromosome. In a standard crossover operation, a random point of division in the chromosome is chosen. If this occurs at position k, then two new strings are created by swapping all alleles between positions $k+1$ and n inclusively. These two new chromosomes are then part of the next generation. Note, however, that this procedure would not work with our encoding of the problem. Swapping all alleles after a certain position, k, may result in duplicate indicies within a single chromosome. Our formulation of the problem does not allow for duplicate alleles within a single chromosome. Each allele needs to be unique. The chromosomes represent subsets that need to be of a fixed length since our chosen control group solution is constrained to be of the same size as the treatment group. This requirement of a fixed-length subset creates a difficulty with a standard crossover because the procedure may not result in n unique alleles.

To bypass this difficulty, we conduct crossover via a variant of Random Assortment Recombination (RAR) (Radcliffe, 1993). In this procedure, the subjects or alleles are separated into different categories or sets. We identify four specific categories of alleles. The subset of alleles common to

both parents is denoted by w_2. Alleles that are in neither parent chromosome are in the subset $\overline{w}_2$. These are referred to as "barred alleles." If an allele is found in only one parent, it belongs to the set w_1. Lastly, those that are found in only one parent are also barred alleles from the other parents, so they also are represented in the $\overline{w}_1$ group. It should be evident that the set $\overline{w}_1$ and w_1 are the same: $\overline{w}_1$ consists of exactly the barred versions of w_1.

Once the sets of regular and barred alleles have been identified, we place them in a draw bag. Two copies of each allele in the w_2 and $\overline{w}_2$ sets are placed in the bag. One copy of each allele in the w_1 and $\overline{w}_1$ sets are placed in the bag. Alleles are then chosen randomly and without replacement from the draw bag. Because there are two copies of the w_2 and $\overline{w}_2$ alleles and only one copy of each w_1 and $\overline{w}_1$ allele, alleles that are present in both parents are twice as likely to be drawn as alleles that are present in only one parent. In addition, alleles that are in neither parents are twice as likely to not be drawn as those that are in one parent.

If a barred allele is chosen, that allele cannot appear in the offspring. When a regular/non-barred allele is drawn, it is placed into the offspring chromosome unless the matching barred allele has already been drawn. Note that after $N - n$ barred alleles have been drawn, the offspring is fully specified as the n remaining alleles. It is possible then that the offspring contain alleles found in neither parent (although the barred alleles are found in both parents). The introduction of barred alleles helps to diversify the population and aids in preventing premature convergence. Within crossover, there is thus some possibility of implicit mutation.

To prevent premature convergence, we incorporate two operations. The first is the implementation of a thresholding RAR (TRAR). It should be evident that the GA converges more quickly as increasing numbers of w_2 and $\overline{w}_2$ alleles are selected from the bag. These alleles are selected more often as the population converges since most alleles will fall into these two categories as convergence accelerates. One way to keep track of the rate of convergence is to compute a difference, D, for two chromosomes, C_1 and C_2, where

$$D = n - \mid C_1 \cup C_2 \mid . \tag{13}$$

This is a count of the number of alleles not shared by two chromosomes. When the average value of D in a generation drops below a certain threshold, the TRAR crossover procedure temporarily

sets the number of draw bag copies of w_2 and $\overline{w}_2$ alleles to 0 until the average D rises above the threshold, a parameter value set by a researcher to suit a particular problem.

We also incorporate a random restart procedure that is aimed to diversify the population to prevent premature convergence. In our implementation, we do not initiate a completely new random restart of the algorithm. The restart replaces almost the entire population. All members of the population except the elite chromosome (i.e. the one with the best fitness value) is replaced. The restart process then initializes the rest of the population randomly and invokes the roulette wheel selection to select members for the mating pool. This process retains information from the best current chromosome while infusing a significant amount of new evolutionary material.

Mutation. The goal of our mutation operator is to retain valuable evolutionary material that may possibly be lost inadvertently through the reproduction and crossover operations. We employ a simple mutation. In particular, we occasionally and with small probability randomly alter an allele. The specific probability value is a parameter that may be changed to suit the application. In general, mutation is a minor change while crossover carries the burden of making the more significant changes that define how one generation evolves from the previous generation.

Stopping Criteria. Each new generation proceeds through the same processes: reproduction, crossover, and mutation to form the next generation. This evolutionary process is repeated until particular stopping criteria are met. There are various stopping criteria including reaching a specified objective function value, evolving for a specific number of generations, evolving for a specified number of minutes, lack of generational improvement, and initiating an excessive number of random restarts without improvement. Each of these criteria may terminate the evolutionary process. The specific thresholds are set by the researcher to match the needs of a particular application.

5 LaLonde Data Case Study

To empirically examine the differences between propensity score models, BOSS-B, and E-BOSS, we turn to the classic LaLonde data set (LaLonde, 1986)—an observational data set that many matching methods have had notorious difficulty in identifying matches that yield good covariate balance. These data are from the National Supported Work Demonstration Program, a randomized job training experiment. An experimental benchmark was computed from the experiment, and then the data were augmented with survey data. LaLonde's intention in creating this data set

was to examine how well statistical methods would perform in trying to replicate the randomized experiment. In our analysis of these data, we used the Dehejia and Wahba (1999) subsample for the treatment group, which includes pretreatment income in 1974 as a covariate, and the Current Population Survey individuals for the control pool. The treatment group contains 185 individuals, and the control pool contains 15,992 individuals. There are eight covariates in this data set.

In this particular application, while there are eight covariates in the data set, simply examining means and variance ratios on these eight variables does not capture the distribution of these variables well enough. Even adding information from the Kolmogorov-Smirnov value proved to be insufficiently nuanced to capture the distributional shape of these variables since some of the variables, notably the real earnings variables, are bimodal or multi-modal in nature. One way to capture the covariate distributions better is to split the data up into quintiles (Zubizarreta, 2012). Accordingly, we created a number of new variables intended to provide additional data points that were more expressive of the idiosyncratic nature of the particular covariates in the data set. Real earnings for 1974, real earnings for 1975, and the age variable were broken down into quintiles. For the real earnings variables, we also created indicator variables for when the value fell at the minimum or the maximum values of the distribution. Education was broken down into three levels. These additional variables provided additional guidance for identifying subsets where the distributions of the control and treatment variables were more closely aligned.

In addition, we also adjusted the weights, w_i, to help the algorithm achieve sufficient balance for the set of variables. It is easier to obtain balance on some variables than others, so placing a higher weights on variables where it is more difficult to obtain balance encourages the algorithm to seek out these more rare traits, inducing greater diversity in solutions, and ultimately resulting in better overall balance. For the LaLonde data, we placed a weight of 2 on the first two terciles of education, and the first quintile of real earnings in 1974. For the covariate married, the first and third quintiles of age, and the zero values of real earnings in 1974, the weight was a 3. The covariate black and the zero values of real earnings in 1975 were weighted 4. The second quintile of age was weighted with a 4. All remaining covariates were weighted with a 1.

The best BOSS-B results are summarized in Table 2. We can see that while BOSS-B does not find it difficult to identify many solutions, it becomes more difficult to identify solutions that fall within increasingly better objective function ranges. The objective function ranges are shown in the first column of Tables 2. In the objective function value range of 40–45, the BOSS-B algorithm

Table 2: LaLonde data: BOSS-B, E-BOSS, and Propensity Score Model Solutions Sorted by Objective Function Value

	Observations	Treatment Effect			
Obj Fcn Range		Mean	SD	Minimum	Maximum
BOSS-B					
(40.0 , 45.0]	72	1534.03	259.11	930.84	2201.42
(45.0 , 50.0]	777	1481.05	256.78	695.39	2195.45
(50.0 , 55.0]	2,231	1415.61	269.84	572.07	2359.02
(55.0 , 60.0]	3,041	1318.83	296.90	397.96	2400.89
E-BOSS					
(16.0 , 17.0]	1,381	1762.32	80.03	1541.67	2026.59
(17.0 , 18.0]	4,137	1780.12	87.04	1489.97	2093.39
(18.0 , 19.0]	3,314	1827.52	104.75	1461.49	2115.63
Propensity Score Model					
(33.0 , 35.0]	742	1107.52	235.41	568.84	1899.65
(35.0 , 37.0]	93	779.24	269.47	420.78	1908.69
(37.0 , 39.0]	40	575.16	249.06	338.69	1353.00
(39.0 , 45.0]	6	1066.38	87.01	932.59	1141.55

Control and treatment group sizes are constrained to be equal. Control groups do not contain any duplicate observations (i.e. individuals are chosen "without replacement").

was able to identify 72 solutions. It was not able to find any solutions where the objective function value fell below 40.[2]

The propensity score model results (obtained via the *pairmatch()* function in the R optmatch package), and using the Dehejia and Wahba (2002) propensity model specification are also shown in Table 2. This specification has been shown to produce good balance for the LaLonde data. We obtained different results from the same propensity score model by changing the calipers for each run. The propensity score model produces many solutions with better objective value than BOSS-B, though its solutions exhibit the largest ranges as shown by the small minimum and maximum ATEs, highlighting the importance of obtaining multiple matched sets.

Table 2 reports the E-BOSS results where the objective function value was lower than 19 on the measure (12). While the most optimal solution identified by BOSS-B had objective function value of 40.77, and the most optimal propensity score objective function was 33.2, E-BOSS was able to

[2]Note that the solutions considered are identical to those identified in Cho et al. (2013) but the objective function values computed in this table are from the formulation shown in Equation (12), which provides a more nuanced balance measure by examining additional points in the distributions. We recomputed the objective value of their solutions to provide 1) a uniform measure, allowing simple and transparent comparison with the solutions identified by E-BOSS, and 2) a measure that is based on actual covariate values, which they advocated as more ideal. The different objective functions primarily changed the scale not the nature of the measure. The correlation of the two objective function values for solutions identified by BOSS-B's binned measure and the formulation in Equation (12) was 0.89.

identify thousands of solutions with objective function values below 20. The best solution identified by E-BOSS has an objective function value of 16.48. Empirically, the E-BOSS implementation is more effective, clearly able to systematically identify more optimal solutions than either propensity score models or the BOSS-B implementation. It is possible that a different propensity score model might yield superior solutions, but in that framework, there is no systematic way to search for increasingly better balanced solutions. The method we used to obtain multiple propensity score estimates is ad hoc. Different ad hoc methods may be used, but these are also not systematic and not even designed to identify increasingly better balanced solutions. Instead, they are simply non-directed random searches. In addition, both BOSS-B and E-BOSS were able to identify thousands of solutions with objective function values in the 40s in less than 10 minutes while the propensity score model took 1 hour and 41 minutes to identify fewer than 1000 solutions. Hence, E-BOSS is not only more effective at finding solutions, it is also more efficient.

5.1 Assessing Balance via Multiple Means and Dimensions

The goal of the objective function is to provide a similarity measure for distributions of the treatment and control groups. Assessing balance is multi-faceted. To gain a sense for how well BOSS-B, a propensity score model, and E-BOSS perform in identifying control groups with similar covariate distributions to the treatment group, we can examine the distributions visually. Figures 2 and 3 present the treatment and control covariate distributions for the best solutions identified by the propensity score model, BOSS-B, and E-BOSS. The gray area shows the treatment covariate distribution while the large red dashes outline the distribution of the control covariates identified by E-BOSS; the short blue dashes indicate the outlines of the distribution of control covariates chosen by BOSS-B; and the dot-dash pattern shows the results from the propensity score model. The E-BOSS solution has identical treatment and control distributions for the black, Hispanic, and married covariates. The BOSS-B solution had fewer blacks, more Hispanics, and fewer married people than the treatment group while the propensity score model solution had more blacks and Hispanics and slightly fewer married in its control group than the treatment group. All three methods identified fewer people without degrees and more people with degrees than the treatment group, with E-BOSS faring the worse. All were also off for age and education. BOSS-B was closer on education while it is difficult to say which produced a better control group for the age covariate. E-BOSS performed significantly better in matching the real earnings's treatment distri-

Figure 2: Balance Plots for Age, Education, Black, and Hispanic Covariates: Gray area is treatment group distribution. Dotted lines show covariate distributions for the control groups identified by BOSS-B and E-BOSS.

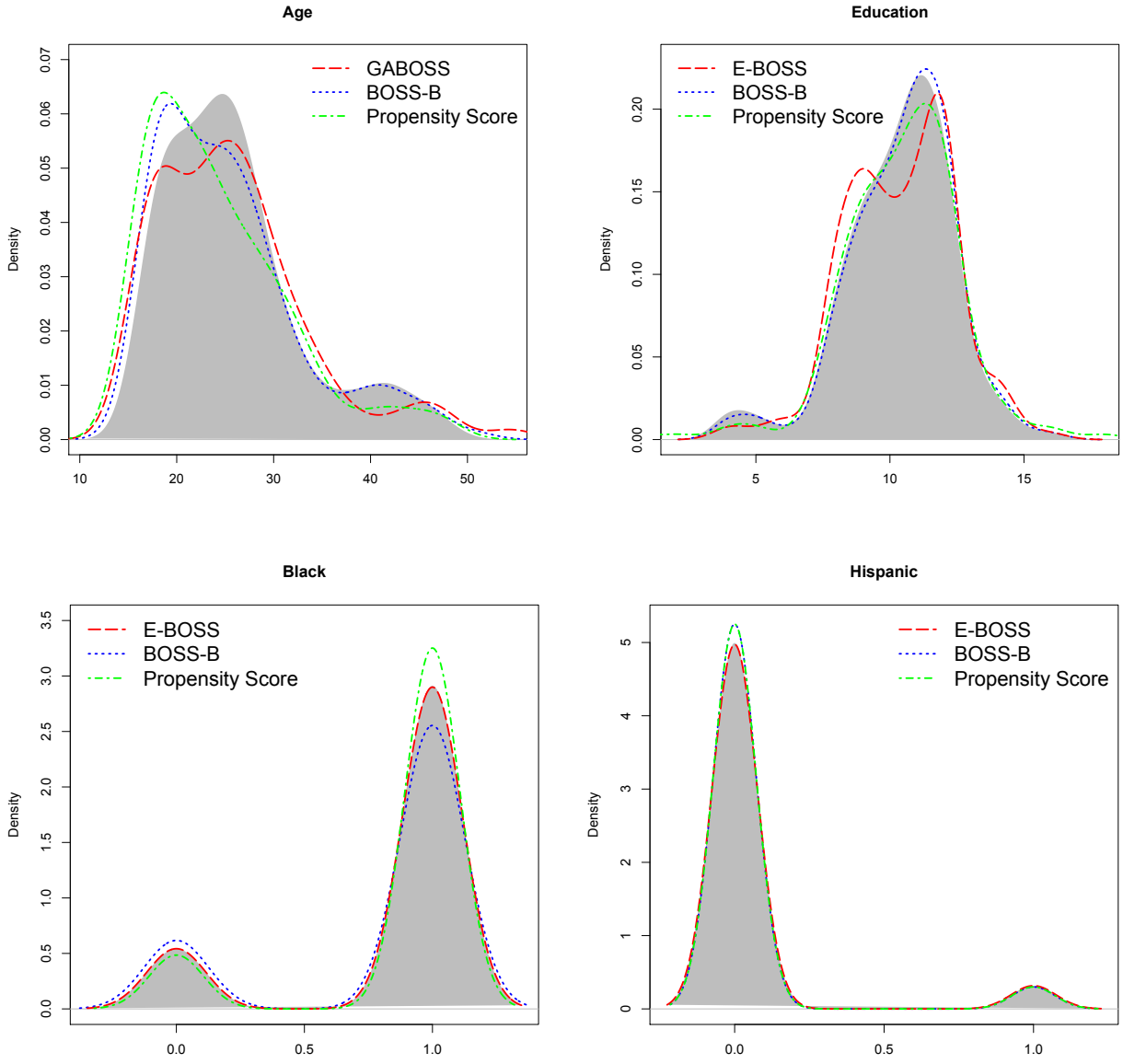

Figure 3: Balance Plots for Married, No Degree, Real Earnings in 1974, and Real Earnings in 1975 Covariates: Gray area is treatment group distribution. Dotted lines show covariate distributions for the control groups identified by BOSS-B and E-BOSS.

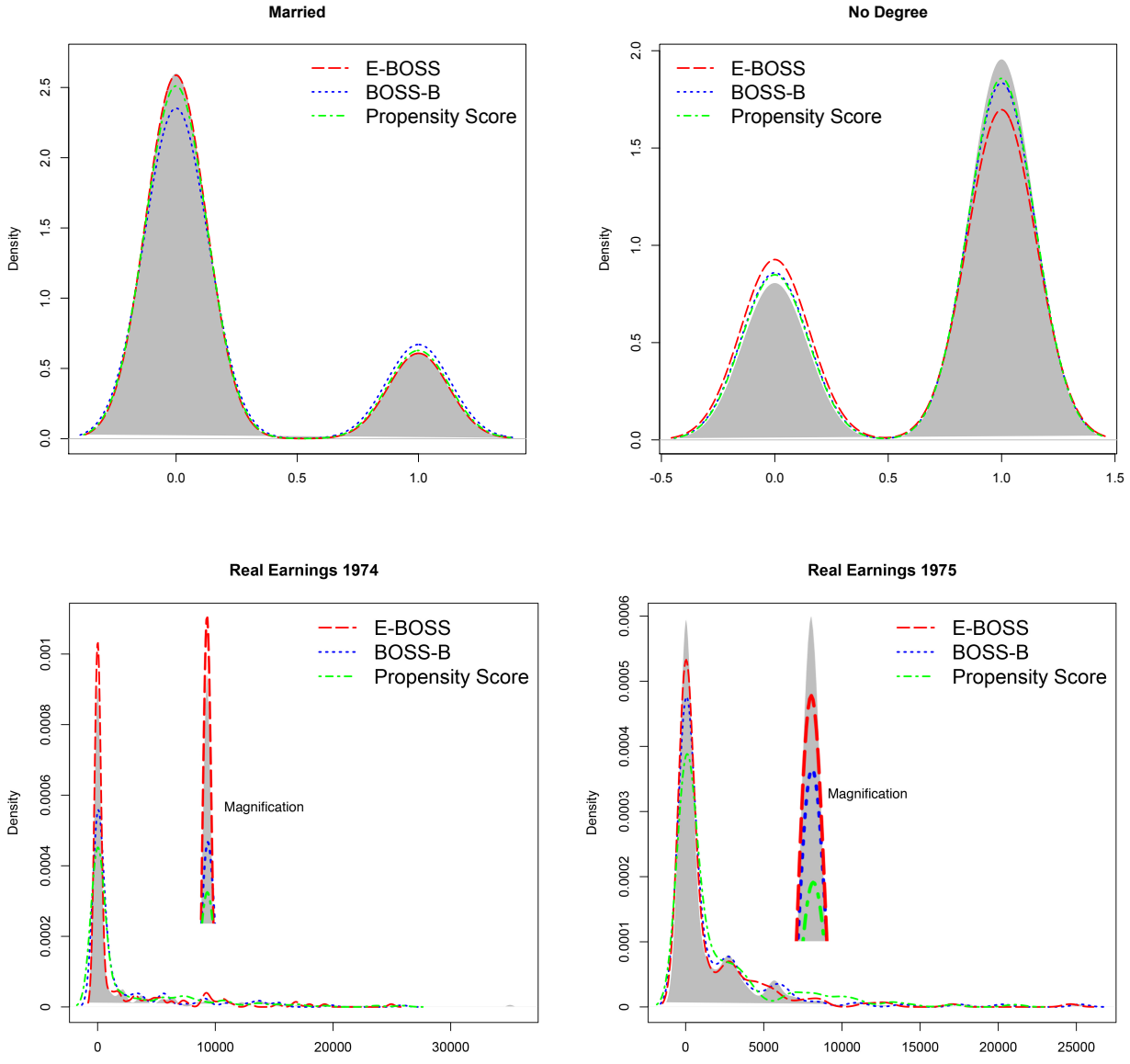

Table 3: Comparison of BOSS-B and E-BOSS solutions

		E-BOSS		BOSS-B		Propensity Score	
Variable	Treated Mean	Control Mean	p-value	Control Mean	p-value	Control Mean	p-value
Age	25.816	24.995	0.284	25.351	0.545	24.211	0.040
Education	10.346	10.427	0.693	10.454	0.602	10.341	0.980
Black	0.843	0.843	1.000	0.805	0.340	0.870	0.297
Hispanic	0.059	0.059	1.000	0.054	0.821	0.054	0.821
Married	0.189	0.189	1.000	0.222	0.441	0.200	0.773
No degree	0.708	0.659	0.315	0.681	0.573	0.687	0.646
Real Earnings 1974	2095.574	1968.862	0.791	2183.555	0.854	2484.594	0.343
Real Earnings 1975	1532.055	1526.942	0.988	1673.932	0.683	2006.768	0.091
d^2		3.29		3.76		9.70	
p-value		0.92		0.88		0.29	

butions. The BOSS-B solution had only about half as many units with $0 earnings in 1974 as the treatment group and the propensity score model was even worse on this score. Both BOSS-B and the propensity score model had the most trouble with fit for the real earnings in 1974 variable, though it exhibited a similar problem finding as many control units with $0 earnings in 1975 as the treatment group. On the other hand, E-BOSS had exactly the same number of $0 earners for 1974 and 1975, apparently trading off some fit on the no degree, age, and education variables. The Real Earnings plots have part of the plot magnified so that one is better able to see the fit at the left end of the distribution.

For these same solutions, we also present some statistics in Tables 3 and 4. These tables show the difference of means tests for the various covariates. The difference of means test is one way to measure similarity, though by itself, two distributions can clearly be quite different while having the same means. Examining only the first moment is plainly limited. By itself, the test is not robust, particularly with multi-modal distributions. However, since we have already visually observed the distribution fits in Figures 2 and 3, we can understand the difference of means tests in context. In Table 3, we display the test statistics for the base set of eight covariates while Table 4 expands the test to additional points in the covariate distributions. Age and real earnings are broken down into quintiles while education is broken down into terciles. Real earnings are also examined at their extremes since the distributions are heavily skewed with the majority of the mass at zero. At the bottom of the tables, we present the omnibus balance test, d^2, that combines the individual

Table 4: Comparison of BOSS-B and E-BOSS solutions

		E-BOSS		BOSS-B		Propensity Score	
Variable	Treated Mean	Control Mean	p-value	Control Mean	p-value	Control Mean	p-value
Age (quintile 1)	0.198	0.159	0.095	0.176	0.343	0.172	0.321
Age (quintile 2)	0.308	0.270	0.231	0.270	0.228	0.233	0.014
Age (quintile 3)	0.049	0.059	0.626	0.041	0.724	0.049	0.971
Age (quintile 4)	0.182	0.154	0.141	0.184	0.917	0.149	0.099
Age (quintile 5)	−0.032	−0.021	0.569	−0.029	0.877	−0.028	0.815
Education (tercile 1)	0.261	0.258	0.914	0.272	0.666	0.252	0.737
Education (tercile 2)	0.458	0.464	0.645	0.457	0.911	0.448	0.485
Education (tercile 3)	−0.191	−0.189	0.882	−0.186	0.748	−0.171	0.280
Real Earnings 1974 (quintile 1)	0.466	0.466	0.947	0.458	0.561	0.449	0.174
Real Earnings 1974 (quintile 2)	0.440	0.441	0.930	0.433	0.588	0.426	0.185
Real Earnings 1974 (quintile 3)	0.055	0.058	0.853	0.071	0.318	0.079	0.108
Real Earnings 1974 (quintile 4)	0.031	0.030	0.930	0.036	0.588	0.041	0.185
Real Earnings 1974 (quintile 5)	0.009	0.005	0.501	0.003	0.306	0.005	0.551
Real Earnings 1975 (quintile 1)	0.328	0.328	0.990	0.316	0.599	0.289	0.072
Real Earnings 1975 (quintile 2)	0.234	0.236	0.976	0.252	0.614	0.280	0.174
Real Earnings 1975 (quintile 3)	−0.055	−0.055	0.996	−0.047	0.722	−0.018	0.078
Real Earnings 1975 (quintile 4)	0.303	0.303	0.979	0.296	0.607	0.281	0.134
Real Earnings 1975 (quintile 5)	−0.140	−0.140	0.995	−0.133	0.563	−0.121	0.054
Real Earnings 1974 (lower extreme)	0.708	0.654	0.265	0.519	0.000	0.611	0.024
Real Earnings 1974 (upper extreme)	0.005	0.000	0.317	0.000	0.317	0.000	0.317
Real Earnings 1975 (lower extreme)	0.600	0.600	1.000	0.519	0.117	0.470	0.009
Real Earnings 1975 (upper extreme)	0.000	0.005	0.317	0.005	0.317	0.000	1.000
Black	0.843	0.843	1.000	0.805	0.340	0.870	0.297
Hispanic	0.059	0.059	1.000	0.054	0.822	0.054	0.819
Married	0.189	0.189	1.000	0.222	0.441	0.200	0.773
No Degree	0.708	0.659	0.315	0.681	0.573	0.686	0.646
d^2		25.51		37.83		48.00	
p-value		0.226		0.014		0.0004	

difference of means tests (Hansen and Bowers, 2008). In particular,

$$d^2(\mathbf{z}; \mathbf{x}_1, \ldots, \mathbf{x}_k) = [d(\mathbf{z}, \mathbf{x}_1), \ldots, d(\mathbf{z}, \mathbf{x}_k)] \left[\text{Cov} \begin{pmatrix} d(\mathbf{Z}, \mathbf{x}_1) \\ \vdots \\ d(\mathbf{Z}, \mathbf{x}_k) \end{pmatrix} \right]^+ \begin{bmatrix} d(\mathbf{z}, \mathbf{x}_1) \\ \vdots \\ d(\mathbf{z}, \mathbf{x}_k) \end{bmatrix}, \tag{14}$$

where

$$d(\mathbf{z}, \mathbf{x}) = \frac{\mathbf{z}'\mathbf{x}}{n_t} - \frac{(\mathbf{1} - \mathbf{z})'\mathbf{x}}{n_c}, \tag{15}$$

and $\mathbf{z}$ is a vector of zeros and ones to indicate assignment to the treated or control group. This test, related to Hotelling's (1931) T-test, considers balance on the base set of k covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, as well as on all linear combinations of the covariates.

All three solutions pass this omnibus d^2 test for balance on the eight baseline covariates (though the propensity score model just misses 0.05-level significance for the age variable). For the balance

test at more points in the distributions, the E-BOSS solution produces covariates means that were statistically indistinguishable from the treated means for all 26 variables tested. As well, the omnibus test for all 26 covariates combined also produced a statistically insignificant value at the 0.20-level. BOSS-B performed well on all but one variable, real earnings in 1974 at \$0. We had already noticed the significant underfit on the real earnings in 1974 variable at the lower extreme in Figure 2. The p-value here was less than 0.001. At the same time, BOSS-B fails the omnibus d^2 test of randomization. The propensity score model performs worse than BOSS-B, especially for the second quintile of age and the lower extreme of Real Earnings in 1974. Several other variables are close to statistically significant differences at the 0.05 level. The propensity score model also fails the d^2 test of randomization.

Note that our objective function (12) is one way assess balance, though its primary function is to guide the optimization search. The BOSS-B objective function, based on binned values, provides another balance measure. The graphical displays in Figures 2–3 provide a different view of how close the control and distributions are. The d^2 statistic (14) provides yet another balance measure that is nuanced in a different way. Indeed, this set of measures do not comprise the only ways to measure balance. Balance statistics can encompass more or less *and* different information and should be understood in tandem and individually for the virtues that they bring to our ability to measure the similarity of covariate distributions. Because the BOSS-B balance statistic uses binned values, its fit is coarse. If we look more closely at the distributions, we see, as expected, that measuring fit with this statistic is less than satisfying within the bin ranges. The d^2 statistic offers a way to combine the fit for the individual covariates into a single omnibus value that considers balance on each covariate as well as a linear combination of all covariates. The E-BOSS solutions are impressive, not simply for faring well with the E-BOSS objective function but for its impressive fit on a variety of balance measures that tap different aspects of distributional similarity.

Quite clearly, other balance measures could also be assessed. Explicitly multivariate balance measures could be utilized. In any particular application, the researcher should apply as many balance tests and measures as he deems necessary to convince himself that the control and treatment groups are similiar enough to be deemed as as-if-randomized groups. Randomization tests, such as the omnibus d^2 test, should be employed after these groups are identified to verify that their similarity falls within the bounds of a randomization rubric. Verification is an important and necessary step whether the vehicle for identifying the groups is via E-BOSS or some other

algorithm. Once verified, another important observation is that there are many as-if-randomized solutions. In this sense, the search for the optimally balanced solution is not an end to itself. In the process, a large number of other as-if-randomized solutions are identified, and these other solutions (while not optimal) are critical in gaining an understanding of the substantive problem.

The best solution as identified by the lowest objective function is always interesting in gauging the effectiveness and efficiency of an optimization algorithm. In the case of this causal inference computational model, however, the group of interesting solutions in more expansive. Any identified solution that yields a control group that is statistically indistinguishable from the treatment group is of interest because any of these control groups would pass a test of randomization.

It is clear that this process of choosing and assigning units can be repeated, and that if it were repeated, the next randomization process would yield empirically different units and a potentially different estimate of the treatment effect. Each new estimate of the treatment effect helps us contextualize and understand both the current treatment effect estimate as well as all previous estimates of the treatment effect. That variation is embodied within the many potential experiments is plain—it is induced by the randomization process itself.

It is only by conducting many experiments that we are able to contextualize any one experiment. Our computational model captures this realization. This is more difficult to embody in a purely statistical model. The statistical models that exist identify only a single solution and hope that that solution is not unusual and close to the mean of the underlying true distribution. Our thousands of solutions that satisfy a randomization test help us contextualize each single solution, giving us a far richer picture of the possible underlying phenomena.

6 Discussion

Substantive research that might benefit from the ability to make causal inferences abounds. Causal research questions transcend all scientific fields. As previously highlighted in the BOSS theory for obtaining causal inferences, BOSS embodies a new paradigm for developing an analytical toolbox that is based in operations research. The causal inference literature has been heavily ensconced in the statistical literature. While the statistical foundations are integral, advances in operations research algorithms may help forge new systematic paths that allows us to bypass some of the assumptions that necessarily underlie the statistical models. Guessing the form of a propensity score model or specifying a particular distance function is no longer necessary under the BOSS

framework, eliminating a potential source of human error. This human bias is replaced by the complexity of a non-trivial NP-Hard optimization problem. In this sense, we are simply replacing one difficult problem with another difficult problem. Available computational power and heuristic algorithmic development are non-trivial replacements. On the other hand, the nature of the problem shifts from uncertainty in model specification to the need for more computational power and algorithm development. Auspiciously computational power is perpetually rising and the algorithm development is straightforward, albeit not simple. Statistical modeling and computational modeling can and ideally do work in tandem with one informing the other and each capitalizing on the strengths of the other.

There are many research paths for causal inference models. Identifying an optimal subset is clearly interesting. However, it would be a greater boon to identify a host of *independent* subsets that satisfy a randomization test and could thus be regarded as a subset that might have emerged from a randomized experimental design. Together, all of these subsets represent the distribution for the realizations of the treatment effect. Identifying the distribution of treatment effects (under a set of assumptions) would be a significant step forward for causal inference models, and one potentially obtainable via computational modeling. The solution search for independent subsets is not the standard search for a single global optimum, but the search principles and strategies are the same. The goal of the optimization process, then, can be seen as two-fold. It is important to be able to search the entire space well and identify optimal solutions. It is also important for the search process to identify diverse solutions that can be regarded as independent subsets. This is difficult for optimization routines because the nature of these algorithms tends to identify an area of the solution space where an optimum may lie and then scour that neighbor space for the optimal solution. For causal inference, there are potentially many subsets that would be statistically indistinguishable from the treatment group. An experiment, after all, may be repeated with a different but equally valid randomized group. Indeed, an experiment can be replicated a large number of times, each subsequent iteration offering new information, and each one as valid as any other. We had instituted random restarts, but another tabu strategy to encourage diversification might be to eliminate members of the control pool once they have been chosen to be a member of an optimal control subset.

The causal inference problem, akin to all interesting problems, embodies its own set of peculiarities. Insights are gained when the right tools are employed but arriving at good solutions

also commands substantial substantive knowledge. Domain knowledge and expertise is integral and cannot be replaced with a default algorithm. Weighting the variables appears to be important in this problem to guide the algorithm either to areas of the solution space that are not being traversed or to indicate which variables have particular substantive importance. Incorporating operations research tools and the insights from computational modeling provide a promising avenue for pursuing advances in models for causal inference.

References

Begg, C. B. 1990. "Significance Tests of Covariate Balance in Clinical Trials." *Controlled Clinical Trials* 11:223–225.

Cho, Wendy K. Tam. 2017. "Causal Inference via Many Experiments." *Journal of Applied Statistics* .

Cho, Wendy K. Tam, Jason J. Sauppe, Alexander G. Nikolaev, Sheldon H. Jacobson and Edward C. Sewell. 2013. "An Optimization Approach for Making Causal Inferences." *Statistica Neerlandica* 67(2):211–226.

Cochran, William G. and G.M. Cox. 1957. *Experimental Designs*. London: Chapman & Hall.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448):1053–1062.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84(1):151–161.

Fisher, Ronald A. 1935. *Design of Experiments*. New York: Hafner.

Hansen, Ben B. and Jake Bowers. 2008. "Covariate balance in simple, stratified and clustered comparative studies." *Statistical Science* 23(2):219–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171(2):481–502.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76:604–20.

Neyman, Jerzy. 1923 [1990]. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923)." *Statistical Science* 5(4):465–472. reprint. Transl. by Dabrowska and Speed.

Nikolaev, Alexander G., Sheldon H. Jacobson, Wendy K. Tam Cho, Jason J. Sauppe and Edward C. Sewell. 2013. "Balance Optimization Subset Selection (BOSS): An Alternative Approach for Causal Inference with Observational Data." *Operations Research* 61:398–412.

Raab, G.M. and I. Butcher. 2001. "Balance in Cluster Randomized Trials." *Statistics in Medicine* 20:351–365.

Radcliffe, Nicholas J. 1993. Genetic Set Recombination. In *Foundations of Genetic Algorithms 2*, ed. L. Darrell Whitley. San Mateo, CA: Morgan Kaufmann Publishers.

Rosenbaum, Paul R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84(408):1024–1032.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.

Rubin, Donald B. 1977. "Assignment to a Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2(1):1–26.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6(1):34–58.

Zubizarreta, J.R. 2012. "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery." *Journal of the American Statistical Association* 107:1360–1371.