

# Comparing Covariate Prioritization via Matching to Machine Learning Methods for Causal Inference using Five Empirical Applications\*

Luke Keele<sup>†</sup>      Dylan Small<sup>‡</sup>

May 11, 2018

## Abstract

Matching methods have become one frequently used method for statistical adjustment under a selection on observables identification strategy. Matching methods typically focus on modeling the treatment assignment process rather than the outcome. Many of the recent advances in matching allow for various forms of covariate prioritization. This allows analysts to emphasize the adjustment of some covariates over others, typically based on subject matter expertise. While flexible machine learning methods have a long history of being used for statistical prediction, they have generally seen little use in causal modeling. However, recent work has developed flexible machine learning methods based on outcome models for the estimation of causal effects. These methods are designed to use little analyst input. All covariate prioritization is done by the learner. In this study, we replicate five published studies that used customized matching methods for covariate prioritization. In each of these studies, subsets of covariates were given priority in the match based on substantive expertise. We replicate these studies using three different machine learning methods that have been designed for causal modeling. We find that in almost every case matching and machine learning methods produce identical results. We conclude by discussing the implications for applied analysts.

## 1 Introduction

When attempting to use data for causal inference, randomized interventions are viewed as the “gold standard” due to the ability to remove measured and unmeasured confounding.

---

\*We thank Jake Bowers for useful feedback on the analysis plan.

<sup>†</sup>Associate Professor, Georgetown University, Email: lk681@georgetown.edu

<sup>‡</sup>Professor, University of Pennsylvania, E-mail: dsmall@wharton.upenn.edu

However, many interventions occur in settings where randomized experiments are difficult, impossible, or unethical. The primary alternative to an randomized trial is an observational study. Cochran (1965) defined an observational study as an empirical comparison of treated and control groups where the objective is to elucidate cause-and-effect relationships in contexts where it is not feasible to use controlled experimentation and subjects select their own treatment status. The primary weakness of observation studies is that when subjects select their own treatments, differing outcomes may reflect initial differences in treated and control groups rather than treatment effects (Cochran 1965; Rubin 1974). Pretreatment differences amongst subjects come in two forms: those that have been accurately measured, which are overt biases, and those that are unmeasured but are suspected to exist which are hidden biases. In an observational study, overt bias must be removed by investigators to render treated and control units comparable in terms of measured observed characteristics.

One approach method for removing overt bias is use a specific study design. For example, regression discontinuity designs remove overt bias through the use of a treatment assignment rule based on a cutoff on a continuous score (Lee and Lemieux 2010). More common is statistical adjustment where investigators apply a method designed to estimate treatment effects while removing overt bias. By far the most frequent statistical method used for adjustment is some type of regression model (Freedman 2005). To avoid bias from functional form misspecification, a wide array of alternatives had been proposed. One might categorize the alternatives as based on either matching, weighting, or machine learning methods. The extant literature contains many comparisons between some combination of these methods. See Dorie et al. (2017) for one particularly large scale comparison.

Among these methods, one particular difference has become particularly pronounced. Many of these methods –especially those based on matching – focus on modeling the treatment assignment process. Moreover, many new matching methods allow for some form of covariate prioritization so that subject matter expertise can be used to guide the removal of overt bias. In general, much of the recent matching literature has focused on adding function-

ality that allows more user input into the statistical adjustment process – see (Zubizarreta 2012) for the best example of this type of matching. Many methods based on machine learning take a very different approach. Here, the primary focus is on flexibly modeling the outcome. In general user input is considered unnecessary as an ensemble of statistical learners are applied. Thus the role of user input into statistical adjustment becomes particularly pertinent. That is, will investigators be better off with greater use of subject matter expertise, or should they instead rely flexible data algorithms that need no input?

Statistical simulation is the primary method used to compare and evaluate methods for statistical adjustment. Under this approach, data are generated where the truth is known and each method is applied repeatedly. The investigator can then measure comparative statistics on bias, coverage, or any other metric of interest. Comparisons based on simulation are widespread, since each method can be benchmarked against a known true quantity. Simulations, however, are ill-suited to understanding whether expertise should be used to guide methods of statistical adjustment. While simulations can be designed to closely mimic real data structures, it is difficult allow subject matter to guide a repeated simulation.

In this paper, we conduct a non-simulation based evaluation of customized matching and machine learning (ML) methods. Our goal is to understand whether using subject matter expertise to guide statistical adjustment leads to fundamentally different results from ML methods that do not rely on subject matter expertise. While little current evidence exists on this point, one study, using simulation evidence, finds that ML methods are decisively better Dorie et al. (2017).

Our evaluation is not simulation based. Instead, we identified five previously completed substantive applications where matching methods were customized to incorporate subject matter expertise. We then replicated the results of each study using machine learning methods that require no user input. We then compare the estimated treatment effects from these two different methods. We find that, in general, both methods produce similar results. We did find that ML methods were less able to model non-normal continuous outcomes. We also

find .... Next, we outline our notation and the causal assumptions needed by both methods.

## 2 Review: Causal Identification via Conditional Ignorability

Both customized forms of matching and ML adjustment methods are typically applied to observational data and invoke a set of causal assumptions that must hold for each method to produce consistent treatment effect estimates. Using the potential outcomes framework, we outline and review those assumptions. Let  $Y_i$  denote the outcome and  $Z_i$  denotes the treatment indicator such that we observe  $(Y_i, Z_i)$ . For each individual  $i$ , let  $Y_i(z)$  be the potential outcome given the treatment assignment value  $z \in \{0, 1\}$ . Observed outcomes are related to potential outcomes in the following way:  $Y_i = Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ . We denote observed covariates as  $\mathbf{x}_i$ . We control overt bias using  $\mathbf{x}_{ij}$ , but we may fail to control for  $u_i$  an unobserved covariate. Our notation implicitly assumes that there is no interference among units. This assumption is often referred to as one part of the stable unit treatment value assumption (SUTVA; Rubin 1986).

We cannot estimate unit level causal effects, since we do not observe the potential outcomes. In general, investigators tend to focus on either the average treatment effect (ATE):

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)] \quad (1)$$

This quantity is often referred to as a causal estimand, since it based on a contrast of potential outcomes. Often the causal estimand is defined as averages over specific subpopulations. In particular, the average treatment effect is often defined for the subpopulation exposed to the treatment or the average treatment effect on the treated (ATT).

$$ATT = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = 1] \quad (2)$$

Investigators of causal effects often assume that the treatment assignment is strongly ignorable (Rosenbaum and Rubin 1983). Formally, this assumption states that the treatment assignment is unconfounded,

$$\Pr(Z_i = 1 | \{(Y_i(1), Y_i(0), \mathbf{x}_i, u_i), i = 1, \dots, n\}) = \Pr(Z_i = 1 | \{\mathbf{x}_i, i = 1, \dots, n\}),$$

and probabilistic

$$0 < \Pr(Z_i = 1 | \{(Y_i(1), Y_i(0), \mathbf{x}_i, u_i), i = 1, \dots, n\}) < 1,$$

for each unit  $i$  (Imbens and Rubin 2015, Section 3.4).

Under this set of assumptions and identification strategy, the analyst asserts that only overt bias is present. That is, selection into treatment regimes is solely on the basis of observed covariates – often referred to as “selection on observables” (Barnow et al. 1980). Thus, under this set of assumptions, investigators must remove overt bias – differences in covariates distributions across levels of  $Z_i$ . As we noted above, a widely variety of methods exist for this purpose. Critically, this assumption cannot be verified with observed data (Manski 2007), thus studies based on this assumption cannot ensure that some form of hidden bias may not be present. The promise and the peril of this approach is apparent in that it can at times recover experimental benchmarks (Dehejia and Wahba 1999; Sekhon and Diamond 2012) and also fail to recover such benchmarks (Arceneaux et al. 2006; Ramsahai et al. 2011).

### 3 Matching

One method for removing overt bias is matching. We provide a very brief review of the statistical literature on matching. Mostly, we focus on one recent strand of the matching literature, which has focused on covariate prioritization in the matching process.

### 3.1 Matching: A Review

While the concept of matching treated and control units with similar  $\mathbf{x}_i$  distributions to remove overt bias has a long history in the statistics literature (Rubin 1973b,a), the literature on matching is now well developed. Early matching methods used optimization methods to form matched pairs (Rosenbaum 1989). While pair matching is common, matched strata can take many forms depending on the study design (Ming and Rosenbaum 2001, 2000; Hansen and Klopfer 2006; Hansen 2004). In general, matching first requires the calculation of a distance matrix that contains a measure of similarity between each treated unit and all controls. Propensity score distances and Mahalanobis distance are frequently used to measure similarity between units. Matches are formed by using an algorithm to find treated to control assignments that minimize the total distances between two groups. See Diamond and Sekhon (2013); Hansen (2004); Iacus et al. (2011); Rosenbaum (1989) for examples. Once matches are formed, investigators typically check how successful the matching process was by comparing treated and control distributions. This diagnostics process is often referred to as balance checking.

One important feature of matching is that the removal of overt bias happens without references to outcomes. In fact, there are recommendations in the literature that all outcome data be withheld until matches are finalized (Rubin 2008). This occurs simply due to the mechanics of matching, but it is also consistent with a strand of the literature which emphasizes modeling the treatment assignment process rather than the outcome.

### 3.2 Matching for Covariate Prioritization

We now focus on the concept of covariate prioritization in matching. Covariate prioritization occurs when an analysts decides to improve balance on a single or set of covariates at the expense of the balance on other covariates. In general, prioritizing balance on one or a set of covariates comes at cost since that prioritization increases imbalances on other covariates. At times, prioritization comes at little cost, since some covariates might be

very well balanced. As such, increasing imbalance on already well balanced covariates can result in small imbalances across all covariates. Alternatively, the investigator may choose to ensure that one or a set of covariates are highly balanced, while other covariates are still imbalanced. Covariate prioritization is the primary means by which investigators can bring subject matter expertise to bear on statistical adjustment. That is, qualitative knowledge guides the identification of which covariate or set of covariates should be prioritized.

The simplest form of covariate prioritization in a multivariate match is exact matching. Here, treated and control units are matched exactly on a nominal covariate such as race. This serves as a form of covariate balance prioritization since all imbalance is removed from the nominal covariate, however, an exact match may increase the imbalance on other covariates. Caliper matching is another simple form of covariate balance prioritization when it is applied to specific covariates. Under a caliper match, two subjects are eligible to be matched if the covariate distance is less than a pre-specified tolerance (Cochran and Rubin 1973). Some methods of covariate prioritization such as fine balance are computationally less costly variations on exact matching (Rosenbaum et al. 2007; Yang et al. 2012).

However, more general forms of multivariate matching allow for more specific forms of prioritization. Ramsahai et al. (2011) demonstrate how genetic matching can be used for highly flexible types of covariate prioritization. Pimentel et al. (2015) develops general method of covariate prioritization for standard optimal match methods based on network flows. Matching via integer programming (Zubizarreta 2012) allows for the most flexible form of covariate prioritization. This form of matching allows analysts to not only select covariates for prioritization, but also apply a wide variety of balance constraints.

## 4 Black Box Methods

The suite of statistical methods referred to as “machine learning” have typically been used for statistical prediction problems. However, much recent work has adapted these methods to the estimation of treatment effects under selection on observables (McCaffrey et al. 2004;

Hill 2011; Hill et al. 2011; Sinisi et al. 2007; Wager and Athey 2017). As Dorie et al. (2017) note, many of these ML methods are designed to be black-box approaches that require no input from the investigator. As such, they stand in stark contrast to matching methods that are designed for the explicit incorporation of subject matter expertise via covariate prioritization. Next, we review two of these black box methods.

BART is a nonparametric regression tree method for fitting arbitrary functions using the sum of the fit from many small regression trees originally designed for statistical prediction Chipman et al. (2010). Hill (2011) adapted BART to the estimation of treatment effects under selection on observables. When BART is used for estimation treatment effects, it models the joint density of the treatment and covariates:  $f(z, \mathbf{x}_i)$ . This density is used to draw the posterior predictive distribution for  $Y(1) = f(\mathbf{x}_i, 1)$  and  $Y(0) = f(\mathbf{x}_i, 0)$ . The empirical posterior distribution for the average treatment effect is obtained by taking the differences between these quantities for each unit at each draw and averaging over the observations. In later work, Hill et al. (2013) develops trimming rules for BART to allow it to better deal with a lack of overlap in the treated and control covariate distributions. Hill et al. (2011) explicitly motivates BART as a black box method where user input is unnecessary and perhaps even undesirable. Moreover, BART focuses entirely on outcomes and does not model the treatment assignment mechanism.

Another prominent ML method designed for treatment effect estimation under selection on observables is based on Super Learner (SL) combined with Targeted Maximum Likelihood Estimation (TMLE) (Sinisi et al. 2007; Van der Laan and Rose 2011; Gruber et al. 2015). The SL approach is an ensemble algorithm, where the investigator can decide among a set of prediction methods—learners—that will be used in the ensemble. The set of learners selected by the investigator are used to make out-of-sample predictions through cross-validation. The predictions from each learner are combined according to weights that minimize the squared-error loss from predictions to observations. These weights are then used to combine the fitted values from the learner when fit to the complete data. Then TMLE correction is applied



to produce an estimate of the average treatment effect. Under SL, the learners are used to model both the assignment mechanism as well as the outcome.

## 5 Research Design

The review of statistical adjustment strategies we provided above highlights clear philosophical differences in treatment effect estimation under conditional ignorability. One approach, that typically uses matching envisions a role for subject matter expertise. Here, statistical adjustment is guided by qualitative input. Black box methods, in contrast, have no role for such information and simply allow flexible fits to determine the final model. It is worth noting that this contrast does not arise from the use of ML methods. One could easily design ML methods that allow for covariate prioritization.

This contrast begs the question of which approach is optimal. Should investigators seek to incorporate substantive information into their models or simply let black boxes do their thing? Unfortunately, the answer to this question is not readily available. Moreover, designing a study to answer this question is also difficult. Dorie et al. (2017) outline three specific challenges to designing studies focused on this question. First, they note that the number of methods compared is usually limited. Second, the comparisons may not be calibrated to closely mimic features of real data applications. That is, simulated data generating processes may be overly simplified. Finally, many comparisons are subject to the file drawer effect, where comparisons are never published or made public when the results do not match expectations.

To overcome these limitations, Dorie et al. (2017) designed a simulated data analysis competition. In the competition, investigators were given simulated data that was generated from real data. Submissions were registered to avoid any results being suppressed. In the competition, black box methods proved to be the best form of statistical adjustment. While the data analysis competition proved to be an innovative research design for comparing statistical adjustment strategies, some key weaknesses remained in terms of being able to

compare black box methods to customized forms of matching. Primarily, even though the simulated data was based on a real data set, there was little role for substantive knowledge, since the data were presented to the competitors without identifying features. As such, there was little room for competitors to rely on substantive insights about features of the data.

We devised the following research design to directly compare statistical adjustments that rely on substantive expertise to black box methods. First, we identified a set of published or completed observational studies where the statistical adjustments were guided by subject matter expertise. We use the treatment effect estimates from these studies as benchmark estimates. Since these studies are completed, we will not replicate the statistical adjustments in any way. It is our assumption that the process used by the authors is one designed to reflect the substantive expertise of the authors. For each the benchmark estimates from these studies, we will perform a re-analysis using a black box method designed for the estimation of causal effects. We will then document how often the two methods lead to substantively different conclusions. To avoid any appearance of bias in the selection of the studies, we published a pre-analysis plan which lists the five studies that were selected for replication using black box methods (Keele and Small 2018).

## 5.1 Selected Box Methods and Comparison Criteria

As outlined in Keele and Small (2018), we selected three different ML methods to replicate the published studies. Specifically, we selected BART, SL methods, and generalized random forests (GRF) (Athey et al. 2016). All three are machine learning methods designed to estimate treatment effects with minimal input from the user. We use a Superlearner as developed in Kennedy et al. (2015). This implementation has well defined asymptotic inferential properties. We will use an ensemble composed of three learners: the generalized linear model, random forests, and the LASSO. BART, like matching, allows for trimming of treated units based on overlap in the covariate distributions. In our re-analyses using BART, we will trim using the rules suggested in Hill et al. (2013). Finally, all methods three

have well defined approaches to variance estimation. The primary weakness of our proposed research design is that we do not have a known true quantity to benchmark against. If all every method produces a different result, we will have no way to know which, if any, of the methods is correct. However, our selection of ML methods was motivated by a strategy to help us identify why the treatment effect estimates may differ. We now outline this strategy.

There are four different possibilities from the replications that follow. First, the point estimates and confidence intervals are generally the same. This result will require little explanation. Second, the point estimates are highly similar, but there are clear differences in the confidence interval lengths. We expect this might occur since the ML methods use outcome information and matching does not. Third, we might find the point estimates have the same sign, but the confidence intervals do not overlap. Fourth, we might find that the sign of the point estimates differ. This will occur when point estimates have a different sign, and the confidence intervals do not overlap. Ideally, we would be able to trace exactly why the two final results happen. While formal characterization of why the methods produce different estimates is most likely not possible, we hope to provide insights into any differences via our choice of ML methods that we use for replicating the studies that used matching.

In fact, we selected the three ML methods to increase the possibility of identifying the source of any differences in the estimates. That is, each of the methods differ from each other in important respects. For example, while GRF is similar to BART in that it implements a flexible model for the outcome, GRF focuses on very local fits in the covariate space and BART fits a model that is global in the covariate space. GRF and BART both only model the outcome, while the Superlearner approach includes models for both the outcome and treatment assignment process. These differences may allow us to explain differences across the methods in the replications.

For example, if all the estimates based on blackbox methods differ from the matched estimate, this is probably due to the fact that the covariates selected for prioritization using subject matter expertise in the matches differ from the covariates given the most weight

by the ML methods. We can explore this possibility by re-running the matches without covariate prioritization and observe whether the estimates move closer to those based on ML. However, if one of the ML methods differ from the matched estimate, we will attribute the difference to the features of that ML method.

## 6 Results

For each of the five studies, we obtained the data. We then identified the exact set of covariates used in the published studies. Finally, we used the three ML methods to estimate the treatment effects from each study along with 95% confidence intervals. In our re-analyses using BART, we trimmed using the rules suggested in Hill et al. (2013). In general, we found these trimming rules did little to change the estimates. In one of the replications, estimation problems with one of the ML methods caused us to change the approach slightly. We detail those differences below. For each replication, we compare point estimates and 95% confidence intervals. Finally, several of the published benchmark studies published treatment effect estimates as median differences across the treated and control units. BART is designed to estimate average treatment effects. To ensure comparability, we re-estimated results from the benchmark studies that used medians using average effects and reported those results in the analysis plan (Keele and Small 2018).

### 6.1 Study 1: Right Heart Catheterization

In the first application we selected for replication, the goal was to study the effect of Right Heart Catheterisation (RHC) an invasive and controversial monitoring device that is widely used in the management of critically ill patients (Connors et al. 1996). We focus on a replication of the original study that applied covariate prioritization (Ramsahai et al. 2011). In the second study, the authors estimated the effect of RHC on mortality within 6 months, length of stay, and total medical costs. This study exemplifies the use of subject matter expertise to design the statistical adjustment for observed covariates. In the study, the investigators adjusted for the following set of covariates: sex, probability of 2-month survival estimated

at baseline, coma score, an indicator for do not resuscitate status, the APACHE III acute physiology score, education, an index of daily activities 2 weeks prior to admission, Duke Activity Status Index, physiological measurements, ethnicity, income, insurance class, primary disease category, admission diagnosis, an indicator for cancer,  $\text{PaO}_2/\text{FiO}_2$  ratio, creatinine,  $\text{PaCO}_2$ , albumin, number of comorbid illnesses, temperature, respiratory rate, heart rate, and white blood cell count.

However, they did not treat this set of covariates equally. First, the authors used exact matching. Second, they consulted a panel of clinical experts to identify a subset of covariates that should have higher priority in the matching process. Then they prioritized balance on the covariates identified as most important by the panel of experts. The clinicians identified eight covariates that were of overriding importance, which included age, coma score, and the APACHE III acute physiology score. They alter the loss function in genetic matching (Diamond and Sekhon 2013) to give this set of covariates higher priority in the match. After matching, they reported average differences by treatment status on the three outcomes.

We replicated the results of this study using the three ML methods. We applied the data trimming rules outlined in Hill et al. (2013), however, we found that these rules did not discard any data. For each of the ML methods, we did not attempt to perform any advanced tuning beyond the software defaults. Table 1 contains the estimates from the original article as well as the estimates using the three ML methods. For the mortality outcome, the point estimates from matching and the three ML methods are all very similar. The confidence intervals based on the ML methods are narrower relative to matching, but differ little across the three methods. For the length of stay outcome, we find some minor differences. First, both BART and matching produce very similar results in terms of both the point estimate (1.25 vs. 1.4) and both confidence intervals include zero. The estimates based on SL and GRF the point estimates are larger and the confidence intervals no longer include zero as they are shorter. This pattern repeats itself with the cost outcome. BART and matching have similar point estimates, while the estimates based on SL and GRF are larger. In this setting,

the confidence intervals based on matching are, however, much wider than those based on the ML methods. Here, we find some small differences for the two continuous outcomes, however, those differences are not large enough that we can draw any strong conclusions.

Table 1: Outcome Analysis Comparison: Study 1 Right Heart Catherization

		Mortality	Length of Stay	Total Costs
Match	Point Estimate	0.046	1.25	9927
	95% Confidence Interval	[ 0.00 , 0.09 ]	[ -1.16 , 3.66 ]	[ 3197 , 16,656 ]
BART	Point Estimate	0.046	1.4	10496
	95% Confidence Interval	[ 0.02 , 0.072 ]	[ -0.53 , 3.32 ]	[ 7495 , 13496 ]
SL	Point Estimate	0.047	2.01	12985
	95% Confidence Interval	[ 0.025 , 0.068 ]	[ 0.6 , 3.41 ]	[ 10025 , 15944 ]
GRF	Point Estimate	0.039	2.93	14986
	95% Confidence Interval	[ 0.014 , 0.064 ]	[ 1.42 , 4.44 ]	[ 11714 , 18259 ]

## 6.2 Study 2: Minority Candidates and Co-Racial Turnout

One area of study in political science focuses on identifying whether voter turnout is higher among minority voting populations when a co-racial candidates runs for office (Barreto et al. 2004; Brace et al. 1995; Gay 2001; Griffin and Keane 2006; Tate 1991, 2003; Voss and Lublin 2001; Washington 2006; Whitby 2007; Fraga 2016a,b; Henderson et al. 2016; Keele and White 2018). Keele et al. (2014) study this question in the context of Louisiana mayoral elections. In their study, the treatment is an African-American candidate on the ballot in Louisiana mayoral elections from 1988 to 2011. In their research design, they matched municipalities with African-American mayoral candidates to municipalities without any African-American candidates for a given election. They matched cities and towns on population, the percentage of African-American residents, the percentage of residents with college degree, the percentage of residents with a high school degree, the percentage of residents unemployed, median income, the percentage of residents below the poverty line, an indicator for home rule municipal charter, and election year.

The investigators used subject matter expertise to inform the statistical adjustments in

a number of ways. First, they identified the percentage of African–American residents in the municipality and year of the election as two key covariates. For this covariate, they set a balance constraint such that municipalities could differ by no more than a tenth of a percentage on the first of these two covariates. Moreover, they enforced balance not only on central moment of this covariate but on higher moments as well. For the election year covariate, they used almost exact matching and allowed elections to differ by no more than two years. They performed three different matches: (1) for general elections, (2) for runoff elections, and (3) for a subset of runoff elections where it was thought that the threat of unobserved confounding was lessened. Overlap between treated and control units was poor, as such the analysts applied optimal subset matching to find the largest set of observations that met the balance constraints. Optimal subsetting removed a significant number of treated observations. The outcome in the study was turnout among African–Americans in the municipality measured as a percentage.

We replicated this study, using the three ML methods. To emulate the almost exact match on election year, we included year fixed effects in the ML specifications. Table 2 contains the treatment effect estimates for both the original study and those from BART. Despite the fact that we applied trimming rules to BART, no observations were removed. For the effect in general elections, all three ML methods produce very similar results that are somewhat smaller than the estimate based on matching. Matching and SL produce confidence intervals that include zero, while BART and GRF have wider intervals that include zero. Interestingly, BART uses all the original 1,006 data points, while the match is only based on 394 observations due to trimming. For the other two outcomes, there is little agreement across the four methods. In both cases, BART produces a result that largely agrees with matching. However, the estimate based on SL is negative, while the GRF estimates are close to zero. Unfortunately, all the confidence intervals are quite wide making stronger conclusions difficult. Moreover the matching estimates are based on 54 and 20 observations while the ML methods use all 187 and 96 observations. Clearly small

samples increase the difficulty of reliably producing estimates.

Table 2: Outcome Analysis Replication for Study 2: Minority Candidates and Co-Racial Turnout

		General Elections	Runoff Elections	Runoff Election Subset
Match	Point Estimate	3.41	5.11	2.97
	95% Confidence Interval	[0.72, 6.1]	[-3.5, 14]	[-11, 17]
BART	Point Estimate	2.69	4.59	5.35
	95% Confidence Interval	[ -0.42 , 5.8 ]	[ -0.5 , 9.69 ]	[ -0.51 , 11.2 ]
	N	1006	187	96
SL	Point Estimate	2.49	-2.21	-9.67
	95% Confidence Interval	[ 1.07 , 3.9 ]	[ -7.93 , 3.5 ]	[ -32.39 , 13.04 ]
GRF	Point Estimate	2.27	-0.23	0.76
	95% Confidence Interval	[ -2.07 , 6.61 ]	[ -6.77 , 6.32 ]	[ -6.53 , 8.06 ]

Note: Point estimates are differences in turnout rates expressed as percentages.

### 6.3 Study 3: Antibiotic Initiation in Critically Ill Children

Study 3 is a study of the the effect of Procalcitonin (PCT)-guided antimicrobial stewardship protocols on antibiotic usage for patients admitted to a pediatric intensive care unit between 2011 and 2014 (Ross et al. 2017). As such infants on the PCT protocol were considered treated, infants that did not use the PCT protocol were controls. In the study, the investigators matched on age, African American (yes/no), PRISM-III score, reason for PICU diagnosis, an indicator for chronic ventilator-dependent respiratory failure, indicator for oncologic comorbidity, an indicator for new mechanical ventilation within the first hour of the PICU admission, source of PICU admission (7 categories), an indicator for surgery preceding PICU admission, and an indicator for trauma preceding PICU admission.

The original study used two forms of covariate prioritization. First, they included KS test balance constraints for age and PRISM-III score. This results in close matches along the entire distribution of these covariates instead of just for the mean. Second, patients were exactly matched on 26 diagnoses. The exact match on diagnosis ensured other aspects of patient care were similar. The original study included two outcomes: 1) a binary indicator



for receiving less than 72 hours of therapy among patients for whom antibiotics were initiated and 2) a binary indicator for initiation of oral or parental antibiotic therapy in the 24 hours prior to or 48 hours after PICU admission. In the data, 505 infants were given PCT protocol and 4,539 controls were available for matching.

Table 3 contains the original results along with the estimates from the three ML methods. For the first outcome, all four methods produce very similar point estimates. Moreover, the confidence intervals are also quite similar. Here, the confidence intervals based on matching are not longer. For the second outcome, matching, SL and GRF all return similar point estimates (0.20, 0.22, 0.22). The estimate for BART is somewhat higher, and the confidence interval only narrowly overlaps with the other methods. Perhaps, BART places more emphasis on some global feature related to the outcome. We suspect that the underlying data generating process is mostly linear and the sample sizes are larger and as such all the methods generally agree.

Table 3: Outcome Analysis Replication for Study 3: Antibiotics in Critically Ill Children

		Antibiotics >72 hours (0/1)	Antibiotic Initiation (0/1)
Match	Point Estimate	0.112	0.202
	95% Confidence Interval	[0.049, 0.17]	[0.15, 0.25]
BART	Point Estimate	0.09	0.29
	95% Confidence Interval	[0.04 , 0.15]	[0.23 , 0.34]
	Point Estimate	0.11	0.22
	95% Confidence Interval	[ 0.06 , 0.15 ]	[ 0.184 , 0.248 ]
GRF	Point Estimate	0.10	0.215
	95% Confidence Interval	[ 0.05 , 0.15 ]	[ 0.178 , 0.252 ]

Note: Point estimates are differences in proportions.

## 6.4 Study 4: The 2010 Chilean Earthquake and Posttraumatic Stress

The next replication uses the results from an investigation of the effect of the Chilean earthquake in 2010 on mental health (Zubizarreta et al. 2013). The authors exploited the fact

that a survey panel study that had been in the field before the earthquake, collected further data after the earthquake. In the study, Chilean residents that lived in areas directly affected by the earthquake were designed as treated units, and residents that lived far enough to be unaffected were designated controls. Residents that directly experienced the earthquake were paired on 46 different covariates including age, gender, member of an indigenous ethnic group, household size, years of education, indicators for employment status, income, measures of housing status and measures of health status prior to the earthquake. The study design employed several form of covariate prioritization. Specifically, the authors exactly matched on sex, whether a resident was a member of an indigenous ethnic group and age categories. They also added fine balance constraints to scales of health and housing quality. Finally, KS test balance constraints were applied to income to balance higher moments of the distribution. The study included a single outcome: the Davidson Trauma Scale (DTS). The DTS is a 17-item self-report measure that assesses individuals for symptoms of post-traumatic stress disorder.

Table 4: Outcome Analysis Replication for Study 4: Chilean Earthquake

		Davidson Trauma Scale
Match	Point Estimate	18.4
	95% Confidence Interval	[17, 19]
BART	Point Estimate	28.6
	95% Confidence Interval	[ 26.5 , 30.7 ]
SL	Point Estimate	25.5
	95% Confidence Interval	[ 20.5 , 30.5 ]
GRF	Point Estimate	28.3
	95% Confidence Interval	[ 25.9 , 30.6 ]

Table 4 contains the results from the matched analysis as well as the 3 ML methods. For the first time, we observe a clear difference between methods. The matched estimate is 18.4 while the 3 ML methods have estimates that range from 25.5 to 28.6. The confidence interval does not overlap. Given the agreement between the ML methods, this would suggest that the black boxes placed greater weight on covariates other than those prioritized in the

match. To explore this possibility, we re-ran the match in the original data without the any covariate prioritization. For this match, we applied an optimal matching method with a robust version of the Mahalanobis distance and a propensity score caliper. For this match, we did not apply any specific balance constraints. The goal of this new match is to observe whether the balance constraints applied in the original match is the source of the discrepancy between the results.

Table 5: Outcome Estimates After Updated Match for Study 4: Chilean Earthquake

Davidson Trauma Scale	
Point Estimate	29.3
95% Confidence Interval	[27, 32]

Table 5 contains the outcome estimates from this new match. The treatment effect estimate from the new match is nearly identical to the estimates based on black box methods, and the confidence intervals overlap. Thus, if no covariate prioritization is employed, the estimates from matching agree with those based on ML methods. While we have traced the source of the discrepancy between the two methods, we do not know which estimate is actually closest to the truth.

## 6.5 Study 5: The Effectiveness of Emergency General Surgery in Elderly Patients

We now turn to the final replication. In the original study, the authors sought to estimate the effectiveness of emergency general surgery (EGS) in adults over the age of 65 (Sharoky et al. 2017). The authors were primarily interested in whether the presence of a dementia diagnosis acted as an effect modifier for the effect of surgery. In the replication, we only focus on the main effect of surgery.

To control for confounders, the authors matched on patient demographics, number of comorbidities, an indicator for sepsis, an indicator for a pre-operative disability, a series of indicators for dementia type, indicators for 31 comorbidities, 9 indicators for broad medical

condition types, 51 indicators for specific medical conditions, and two indicators for hospital admission source. The authors applied covariate prioritization by exactly matching on hospital, the 9 categories of surgical conditions, and the indicators for type of dementia. They also applied near fine balance to the 51 more specific sub-condition indicators. In the study, they estimated treatment effects for the following set of outcomes: mortality (0/1), prolonged length of stay (0/1), discharge to a higher level of care (0/1), discharge to a hospice (0/1), and length of stay (in days). In our replication, we only focus on mortality, prolonged length of stay, and length of stay outcomes. The other two outcomes have patterns of missingness that cause additional complications. As in other replications, we include hospital fixed effects in the ML methods to mimic the exact matching on hospitals.

For this study, we were unable to apply BART on the original data set used for matching. Specifically, we found that there were two aspects of the data that caused problems. First, we could not obtain BART estimates with original sample size of 1,221,303 observations. BART required more memory than was available on the secure computer where this data is housed. Note this is the same computer used for the matching in the original paper and has 16 GB of RAM. As we noted above, dementia status was thought to be a possible effect modifier, and the investigators exactly matched on dementia status. To that end, we then applied BART just within the dementia subgroup, which had a much smaller sample size of 96,473. Again BART could not allocate enough memory. We found that this was due to the large number of hospital fixed effects (more than 450).

To allow for a comparison across all the methods, we took the following approach. First, we re-ran the matching within the dementia subgroup and removed the hospital exact match. A match of this form reduces the sample size considerably and removes the large number of fixed effects. Table 6 contains the treatment effect estimates for the original match and for the smaller match. The outcome estimates based on the new match are nearly identical to the results from the original match. We then applied the 3 ML methods to the dementia subgroup and did not include hospital fixed effects. Table 6 also contains the estimates from

the 3 ML methods.

Table 6: Outcome Analysis Replication for Study 5: Emergency General Surgery in Elderly Patients

		Prolonged LOS	Mortality	LOS
Original Match	Point Estimate	0.093	-0.001	3.53
	95% Confidence Interval	[0.09, 0.095]	[-0.003, 0.0002]	[3.5, 3.6]
Modified Match	Point Estimate	0.098	-0.007	3.7
	95% Confidence Interval	[0.088, 0.11]	[-0.013, -0.0015]	[3.5, 3.9]
BART	Point Estimate	0.084	-0.01	3.35
	95% Confidence Interval	[0.076 , 0.093]	[-0.014 , -0.005]	[3.24 , 3.46]
SL	Point Estimate	0.08	-0.023	3.55
	95% Confidence Interval	[0.073 , 0.087]	[-0.031 , -0.016]	[3.37 , 3.73]
GRF	Point Estimate	0.082	-0.018	3.64
	95% Confidence Interval	[0.075 , 0.09]	[-0.026 , -0.009]	[-0.19 , 0.16]

Note: Point estimates are differences in turnout rates expressed as percentages.

In general, we find some minor differences between the ML treatment effect estimates and those based on matching. First, for the prolonged length of stay outcome, the ML estimates are all slightly lower than the estimates from matching. While the confidence intervals based on the Superlearner and matching do not quite overlap, the confidence interval from matching and the other two methods do overlap. In general, despite the slight differences, the methods provide highly similar results. For the mortality outcome, the estimates and confidence intervals based on BART and matching are very similar (-0.007 vs -0.01). The estimate based on the Superlearner, however, is larger and the confidence interval just narrowly fails to overlap with the confidence intervals for BART and matching. While the GRF estimate is closer to the estimate from SL, the GRF confidence interval is wider such that it overlaps with all the other methods. Thus, we observe some differences in the estimates for the mortality outcome. However, the differences are not large and do not vary by estimation method in an obvious way. Finally for the length of stay outcome, all the estimates are highly similar. We do not discern any clear pattern between methods and estimates. It is also worth noting that in this application use of outcome information under the ML methods

typically provides no advantage in terms of narrower confidence intervals. For example, the length of the confidence interval for the prolonged length of stay outcome under matching is 0.022, while for BART the confidence interval length is 0.017.

## 7 Discussion

Here, we outlined two different approaches to the analysis of observational data under a selection on observables identification strategy. Under one approach, black-box machine learning methods are applied with little guidance from the investigator. Here, it is assumed that flexible fits are better than substantive knowledge. Under the second approach, based on matching, substantive expertise can be incorporated in the statistical adjustment process. This approach seeks scientific expertise from subject matter experts. We then compared whether these two approaches produce similar results across five diverse studies. We found that despite the underlying philosophical differences between these two approaches, they mostly produce nearly identical results. In the one case with the clearest difference, we found that a more traditional match agreed with the black box estimates. We did find that the black box methods often produce shorter confidence intervals since outcome information is utilized. However, for larger sample sizes the difference in confidence interval length is greatly reduced.

Our replication process did reveal one clear computational advantage of matching. Under matching, whether covariate prioritization is applied or not, once the match is completed relatively simple models can be fit for each outcome. For each of the black box methods, the entire method needs to be re-run for each outcome. For some of the larger data sets we worked with there was a clear time-savings advantage to being able to run the match once and then quickly estimate treatment effects for each outcome. Moreover, for the EGS replication, we were unable to produce BART estimates using a standard desktop computer. However, we were able to complete the original match with the full sample size in less than 24 hours.

Finally, we consider whether there are any larger lessons to be learned for the conduct of observational studies. At first blush one would be tempted to say no. We employed methods of statistical adjustment that appear to differ in fairly significant ways to a wide variety of data sets, and typically the treatment effect estimates were nearly identical. Thus one basic conclusion could be that so long as the method of adjustment is sufficiently flexible, analysts should use whatever method they are most comfortable using. One might liken the situation to a setting where on average two medical procedures for a condition have similar average performance. A doctor might be best off using the procedure he or she is most comfortable with because the doctor will use it in a more skilled way.

However, we would argue that there is one reason to be less agnostic. That is one might imagine a research design where covariate prioritization and black box methods are used in conjunction. As a first step, subject matter experts would be consulted to implement statistical adjustment with covariate prioritization. In the second step, black box methods would be applied to the same data. If the results from the two methods agree, little else need be done. However, if the results differ, then investigators could explore what aspects of covariate prioritization produced different estimates. This would serve as a useful way to explore whether conventional wisdom about covariate priority agrees with a flexible approach.

Finally, we think the dearth of differences serves to emphasize that in observational studies the form of statistical adjustment will rarely be what makes the evidence from the study compelling. Instead we suspect that careful study design is generally more important. For example, we would argue that the skillful use of negative controls or placebo tests (Lipsitch et al. 2010; Rosenbaum 2005) is far more likely to make a result conclusive than whether one used matching or a black box method. Rosenbaum (2010, ch. 5) offers one useful overview of devices that can make observational evidence more compelling. Alternatively, identification of a plausible natural experiment often provides a way to bolster evidence from observational data (Angrist and Pischke 2010).

## References

- Angrist, J. D. and Pischke, J.-S. (2010), “The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics,” *Journal of Economic Perspectives*, 24, 3–30.
- Arceneaux, K., Gerber, A. S., and Green, D. P. (2006), “Comparing Experimental and Matching Methods Using A Large-Scale Voter Mobilization Study,” *Political Analysis*, 14, 37–62.
- Athey, S., Tibshirani, J., and Wager, S. (2016), “Generalized Random Forests,” *ArXiv e-prints*.
- Barnow, B., Cain, G., and Goldberger, A. (1980), “Issues in the Analysis of Selectivity Bias,” in *Evaluation Studies*, eds. Stromsdorfer, E. and Farkas, G., San Francisco, CA: Sage, vol. 5, pp. 43–59.
- Barreto, M. A., Segura, G., and Woods, N. (2004), “The Effects of Overlapping Majority-Minority Districts on Latino Turnout,” *American Political Science Review*, 98, 65–75.
- Brace, K., Handley, L., Niemi, R. G., and Stanley, H. W. (1995), “Minority Turnout and the Creation of Majority-Minority Districts,” *American Politics Quarterly*, 23, 190–203.
- Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010), “BART: Bayesian additive regression trees,” *The Annals of Applied Statistics*, 4, 266–298.
- Cochran, W. G. (1965), “The Planning of Observational Studies of Human Populations,” *Journal of Royal Statistical Society, Series A*, 128, 234–265.
- Cochran, W. G. and Rubin, D. B. (1973), “Controlling Bias in Observational Studies,” *Sankhya-Indian Journal of Statistics, Series A*, 35, 417–446.
- Connors, A. F., Speroff, T., Dawson, N. V., Harrell, C. T. F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A., and Califf, R. M. (1996), “The effectiveness of right heart catheterization in the initial care of critically III patients,” *Jama*, 276, 889–897.
- Dehejia, R. and Wahba, S. (1999), “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.
- Diamond, A. and Sekhon, J. S. (2013), “Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies,” *Review of Economics and Statistics*, 95, 932–945.
- Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2017), “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition,” *arXiv preprint arXiv:1707.02641*.
- Fraga, B. (2016a), “Candidates or Districts? Re-evaluating the Role of Race in Voter Turnout,” *American Journal of Political Science*, 60, 97–122.



- Fraga, B. L. (2016b), “Redistricting and the Causal Impact of Race on Voter Turnout,” *Journal of Politics*, Forthcoming.
- Freedman, D. (2005), “Linear statistical models for causation: A critical review,” *Encyclopedia of statistics in behavioral science*.
- Gay, C. (2001), “The Effect of Black Congressional Representation on Political Participation,” *American Political Science Review*, 95, 589–602.
- Griffin, J. D. and Keane, M. (2006), “Descriptive Representation and the Composition of African American Turnout,” *American Journal of Political Science*, 50, 998–1012.
- Gruber, S., Logan, R. W., Jarrín, I., Monge, S., and Hernán, M. A. (2015), “Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets,” *Statistics in medicine*, 34, 106–117.
- Hansen, B. B. (2004), “Full matching in an observational study of coaching for the SAT,” *Journal of the American Statistical Association*, 99, 609–618.
- Hansen, B. B. and Klopfer, S. O. (2006), “Optimal full matching and related designs via network flows,” *Journal of Computational and Graphical Statistics*, 15.
- Henderson, J. A., Sekhon, J. S., and Titiunik, R. (2016), “Cause or effect? Turnout in Hispanic majority-minority districts,” *Political Analysis*, 24, 404–412.
- Hill, J., Su, Y.-S., et al. (2013), “Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children’s cognitive outcomes,” *The Annals of Applied Statistics*, 7, 1386–1420.
- Hill, J., Weiss, C., and Zhai, F. (2011), “Challenges with propensity score strategies in a high-dimensional setting and a potential alternative,” *Multivariate Behavioral Research*, 46, 477–513.
- Hill, J. L. (2011), “Bayesian nonparametric modeling for causal inference,” *Journal of Computational and Graphical Statistics*, 20.
- Iacus, S. M., King, G., and Porro, G. (2011), “Causal inference without balance checking: Coarsened exact matching,” *Political Analysis*, 20, 1–24.
- Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference For Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge University Press.
- Keele, L. J., Shah, P. R., White, I. K., and Kay, K. (2014), “Black Candidates and Black Turnout: A Study of Viability in Louisiana Mayoral Elections,” *Journal of Politics*, Forthcoming.
- Keele, L. J. and Small, D. S. (2018), “Pre-analysis Plan for a Comparison of Matching and Black Box-based Covariate Adjustment,” *Observational Studies*, in press.

- Keele, L. J. and White, I. K. (2018), “African American Turnout in Majority-Minority Districts,” *Political Science Research and Methods*, Forthcoming.
- Kennedy, E., Sjölander, A., and Small, D. (2015), “Semiparametric causal inference in matched cohort studies,” *Biometrika*, 102, 739–746.
- Lee, D. S. and Lemieux, T. (2010), “Regression Discontinuity Designs in Economics,” *Journal of Economic Literature*, 48, 281–355.
- Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010), “Negative controls: a tool for detecting confounding and bias in observational studies,” *Epidemiology (Cambridge, Mass.)*, 21, 383.
- Manski, C. F. (2007), *Identification For Prediction And Decision*, Cambridge, Mass: Harvard University Press.
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), “Propensity score estimation with boosted regression for evaluating causal effects in observational studies,” *Psychological methods*, 9, 403–425.
- Ming, K. and Rosenbaum, P. R. (2000), “Substantial gains in bias reduction from matching with a variable number of controls,” *Biometrics*, 56, 118–124.
- (2001), “A note on optimal matching with variable controls using the assignment algorithm,” *Journal of Computational and Graphical Statistics*, 10, 455–463.
- Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), “Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons,” *Journal of the American Statistical Association*, 110, 515–527.
- Ramsahai, R. R., Grieve, R., and Sekhon, J. S. (2011), “Extending iterative matching methods: an approach to improving covariate balance that allows prioritisation,” *Health Services and Outcomes Research Methodology*, 11, 95–114.
- Rosenbaum, P. R. (1989), “Optimal Matching for Observational Studies,” *Journal of the American Statistical Association*, 84, 1024–1032.
- (2005), “Observational Study,” in *Encyclopedia of Statistics in Behavioral Science*, eds. Everitt, B. S. and Howell, D. C., John Wiley and Sons, vol. 3, pp. 1451 – 1462.
- (2010), *Design of Observational Studies*, New York: Springer-Verlag.
- Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), “Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer,” *Journal of the American Statistical Association*, 102, 75–83.
- Rosenbaum, P. R. and Rubin, D. B. (1983), “The Central Role of Propensity Scores in Observational Studies for Causal Effects,” *Biometrika*, 76, 41–55.

- Ross, R. K., Keele, L. J., Kubis, S., Lautz, A. J., Dziorny, A. C., Denson, A. R., O'Connor, K. A., Chilutti, M. R., Weiss, S. L., and Gerber, J. S. (2017), "Impact of Procalcitonin Availability on Antibiotic Utilization in Critically Ill Children," *Journal of the Pediatric Infectious Diseases Society*, In press.
- Rubin, D. B. (1973a), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183.
- (1973b), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203.
- (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.
- (1986), "Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961–962.
- (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840.
- Sekhon, J. S. and Diamond, A. (2012), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," *Review of Economics & Statistics*, Forthcoming.
- Sharoky, C. E., Sellers, M. M., Keele, L. J., Wirtalla, C. J., Martin, N. X., Braslow, B. X., Holena, D. N., Neuman, M., and Kelz, R. R. (2017), "Evidence to Guide Acute Surgical Decision-Making in Older Adults with Alzheimer's Disease and Related Dementias," Unpublished Manuscript.
- Sinisi, S. E., Polley, E. C., Petersen, M. L., Rhee, S.-Y., and van der Laan, M. J. (2007), "Super learning: an application to the prediction of HIV-1 drug resistance," *Statistical applications in genetics and molecular biology*, 6.
- Tate, K. (1991), "Black Political Participation in the 1984 and 1988 Presidential Elections," *American Political Science Review*, 85, 1159–1176.
- (2003), *Black Faces in the Mirror: African American and Their Representatives in the U.S. Congress*, Princeton, NJ: Princeton University Press.
- Van der Laan, M. J. and Rose, S. (2011), *Targeted learning: causal inference for observational and experimental data*, Springer Science & Business Media.
- Voss, S. and Lublin, D. (2001), "The Effects of Incumbency in U.S. Congressional Elections, 1950-1988," *American Politics Research*, 29, 141–182.
- Wager, S. and Athey, S. (2017), "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*.
- Washington, E. (2006), "How Black Candidates Affect Voter Turnout," *Quarterly Journal of Economics*, 121, 973–998.

- Whitby, K. J. (2007), “The Effect of Black Descriptive Representation on Black Electoral Turnout in the 2004 Elections,” *Social Science Quarterly*, 88, 1010–1023.
- Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), “Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes,” *Biometrics*, 68, 628–636.
- 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.
- Zubizarreta, J. R., Cerdá, M., and Rosenbaum, P. R. (2013), “Designing an observational study to be less sensitive to unmeasured biases: Effect of the 2010 Chilean earthquake on posttraumatic stress,” *Epidemiology*, 24, 79–87.