Observational Studies 4 (2018) 97-110 Submitted 3/17; Published 3/18

Pre-analysis Plan for a Comparison of Matching and Black Box-based Covariate Adjustment

Luke Keele **lk681@georgetown.edu**
School of Public Policy
Georgetown University
Washington D.C.

Dylan S. Small **dsmall@wharton.upenn.edu**
Department of Statistics
University of Pennsylvania
Philadelphia, PA

Abstract

This article presents a pre-analysis plan for a comparison of methods for the statistical adjustment of observed confounders. In the planned analysis, we intend to replicate five existing studies that used customized form of matching and substantive input from subject matter experts. We will replicate the treatment effect estimates from these studies using machine learning methods that need little user input. In this article, we outline the five studies we will use for replication and discuss the methods we use for replication.

Keywords: Observational Studies, Matching, Machine Learning, Analysis Plan

1. Introduction

Many evaluations occur in settings where randomized experiments are difficult or impossible. When randomized interventions are not possible researchers often choose to conduct 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. 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 or selection biases 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 of treatment effects, analysts use pretreatment covariates and a statistical adjustment strategy to remove overt biases.

In this analysis plan, we outline future analyses that are designed to assess the relative merits of statistical methods that remove overt biases. While the extant literature has many such comparisons, the bulk of that evidence relies on simulated data. In this study, we propose a novel approach to such analytic comparisons. We begin with a brief overview of notation, study goals, and key assumptions.

2. Study Overview

2.1 Notation and Assumptions

We first define basic notation and outline the identification strategy under which we operate. There are I units, $i = 1, \ldots, I$, with observed covariates, $\mathbf{x}_i$, which are measured before assignment to treatment. Similarly, u_i represents an unobserved covariate for individual unit i. We denote whether a unit is treated such that $Z_i = 1$ if treated and $Z_i = 0$ if not. Each subject also has two potential responses, which we denote as y_{Ti} if $Z_i = 1$ or y_{Ci} if $Z_i = 0$. We do not observe the pair of potential outcomes, $(y_{Ti},\ y_{Ci})$, but we do observe the responses, $Y_i = Z_{ij}y_{Ti} + (1 - Z_i)y_{Ci}$. Note that 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).

On common approach to the identification of causal effects is to assume that treatment assignment is strongly ignorable (Rosenbaum and Rubin, 1983). Formally, this assumption states that treatment assignment is unconfounded

$$\Pr(Z_i = 1|\{(y_{Ti}, y_{Ci}, \mathbf{x}_i, u_i)\}) = \Pr(Z_i = 1|\{\mathbf{x}_i\}),$$

and probabilistic

$$0 < \Pr(Z_i = 1|\{(y_{Ti}, y_{Ci}, \mathbf{x}_i, u_i)\}) < 1,$$

for each unit i. Intuitively, this assumption implies that after adjusting for observed covariates there are no systematic, pretreatment differences in unobserved covariates between the treatment and control groups, and that every unit has a non-zero probability of receiving treatment. That is, we assume that all bias is overt and thus can be removed via statistical adjustments. This assumption is alternatively referred to as the "selection on observables" identification strategy, where the analyst asserts that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow et al., 1980). Critically, the selection on observables assumption is nonrefutable, insofar as it cannot be verified with observed data (Manski, 2007). The study we propose is a comparison of statistical methods that assume selection on observables holds.

2.2 Methods of Adjustment

Under a selection on observables identification strategy, analysts must employ a statistical method to adjust for observed confounders. A wide array of methods may

be used to adjust for observed confounders including regression modeling of various types, matching, and inverse probability weighting. We next review two approaches to such methods of statistical adjustment. Instead of focusing on technical details, we highlight a broader philosophical difference the between methods. First, we review methods for matching.

2.3 Customized Forms of Matching

Matching in its most basic form consists of forming matched pairs via minimization of a distance metric that measures multivariate distances between treated and control units. The matching process is then evaluated through balance tests, which consist of comparing summary statistics across the treated and control groups to evaluate whether the treated and control groups are indeed comparable on observed covariates. Currently a large number of matching algorithms are available that implement the basic process described above (Diamond and Sekhon, 2013; Hansen, 2004; Iacus et al., 2011; Rosenbaum, 1989). See Rosenbaum (2017, ch. 11) for a broad overview of different forms of matching that are used in applied research.

In the literature on matching, researchers have developed a variety of refinements to more standard matching algorithms. Examples of such refinements include optimal subset matching (Rosenbaum, 2012), almost exact matching (Rosenbaum, 2010), methods for fine and near-fine balance (Rosenbaum et al., 2007; Yang et al., 2012), near-far matching (Lu et al., 2001), covariate balance prioritization (Zubizarreta, 2012; Kilcioglu and Zubizarreta, 2017; Ramsahai et al., 2011), cardinality matching (Zubizarreta et al., 2014), and refined covariate balance (Pimentel et al., 2015). In general, these refinements are designed to give applied analysts greater control over the matching process.

One motivation for introducing increased flexibility into the matching process is to allow the user to bring subject matter to bear on the design of the observational study. Under most regression modeling approaches, observed confounders are treated equally in the adjustment process. However, it is typically the case that while investigators match on many covariates, they often view some covariates as more important than others. The recent advances in matching noted above, allow analysts to inject a considerable amount of discretion into the matching process by targeting covariates deemed to be the most important. The implicit assumption is that scientific knowledge of the relevant subject matter should be an element of the design of an observational study and may play an important role in developing the "best" specification. For example, the investigator might choose to exact match on one covariate deemed to be of critical importance, finely balance other important nominal covariates, increase the comparability of higher moments for some covariates, while allowing other covariates to be less well balanced. In short, the match can be customized using subject matter expertise.

2.4 Black Box Methods

One alternative set of methods for covariate adjustment relies on insights and methods from machine learning (ML). The use of ML methods for covariate adjustment in observational studies is currently the new frontier in treatment effect estimation (Sinisi et al., 2007; Van der Laan and Rose, 2011; Gruber et al., 2015; Hill, 2011; Wager and Athey, 2017). In general, covariate adjustment via ML is often viewed as a black box process whereby a set of statistical learners are used to develop the "best" specification with little input from the investigator. In fact, the lack of investigator input if often considered an advantage of ML methods (Hill et al., 2013, 2011) since human judgement may be a poor substitute for sophisticated statistical algorithms. Thus one distinction between newer matching methods and black box methods based on ML is that the former are designed for customization based on subject matter knowledge, and the latter is designed to preclude such information. One obvious question is whether any general patterns can be inferred about which approach might be preferable in applied applications. Next, we outline a study design aimed at this question.

3. Study Design

Comparisons of statistical adjustment methods for observed confounders are common in the statistical literature. Generally such comparisons use simulated data. Simulations have the advantage that investigators know the full data-generating process and thus the true causal estimate. Simulations can be repeated many times to assess inferential properties as well. However, simulation based comparisons generally give little insight into the advantages of customization via matching, since it is difficult to design simulations to mimic a wide variety of applications. Moreover, it is generally not practical to both tailor the specification using subject matter knowledge and repeat the simulations many times. To that end, below we propose a study design to compare these methods fairly.

The first component of our study design is the selection of a set of published or completed observational studies where matching was used to customize the nature of the statistical adjustment using subject matter expertise. These studies will serve as a set of treatment effect benchmark estimates that use customized forms of matching. More specifically, the benchmarks in our study will be the treatment effect estimates derived after matching for this set of studies. Below, we outline the studies we have selected to serve as the benchmarks. Since these studies are published, we will not replicate the matching process in any way. It is our assumption that the matching process used by the authors is one designed to reflect the expertise of the authors. For each of the published benchmark studies, we will perform a re-analysis using a set of machine learning methods designed for the estimation of causal effects. In this re-analysis, we will use the identical set of covariates used for matching as control variables in the ML methods. In fact, the primary purpose of this analysis plan is to

pre-publish the set of studies we use of comparisons. This will preclude the possibility that we selected papers for replication to influence the results of the study. Below is an overview of the study design:

- Identify a set of published observational studies that use a customized form of matching.
- Identify covariates and outcome(s) used in each observational study based on matching.
- Use the same set of covariates to estimate causal effects via black box methods.
- Compare estimates and confidence intervals across black box methods and matching.

3.1 Comparison Criteria

Finally, we outline our analytic plan for comparing estimates across estimation methods. One weakness of our study design compared to a simulation study is that we do not have a known true quantity to benchmark against. Thus we developed the following guidelines for comparing results from the different methods. We envision four different possibilities after the replications are complete. Under the first possibility, the point estimates and confidence intervals are generally the same. This outcome will require little explanation. Second, the point estimates are nearly identical, but confidence interval length differs. We might expect this to occur due to an ML method that relies on modeling the outcome or due to a form of matching that doesn't use the full data. The third possibility is that the point estimates differ but the sign of those estimates is identical. This result arises when the point estimates have the same sign, but the confidence intervals do not overlap. The fourth possibility is that the sign of the point estimates differ. This differences will result from point estimates of a different sign and non-overlapping confidence intervals.

The final two possibilities will merit further investigation. That is, ideally, we would identify why point estimates differ across methods. 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. Next, we outline the ML methods we will use for blackbox estimates, and how we will use these methods to possibly identify why results may differ across methods.

We use three different ML methods. The first method we will use is Bayesian Additive Regression Trees (BART). We selected BART as a method of comparison, since it has been proposed as a feasible alternative to methods like matching that requires little user input Hill et al. (2013). 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). The second

method we will use is generalized random forests (GRF) (Athey et al., 2016). GRF is an implementation of random forests designed to estimate treatment effects. The third method we use is a Superlearner (SL), another ML method designed specifically for estimating treatment effects Van der Laan and Rose (2011). The Superlearner method also allows us implement the ML methods as a true ensemble. In the SL framework, analysts can select from a variety of ML methods, and the results are based on an ensemble of these learners. 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. All three of these methods can be readily implemented using available software `R`.

We selected these three methods to increase the possibility of identifying the source of any differences in the estimates. That is, each of the three ML methods differ from each other in important respects. While GRF is similar to BART in that it implements a flexible model for the outcome, it differs from BART in that it focuses on very local fits in the covariate space. BART fits a model to the outcome that is global in the covariate space. However, RF and BART both focus on the outcome only, and the Superlearner allows us to include models for both the outcome and treatment assignment process. Such causal estimation methods are often referred to as "doubly robust" (Robins et al., 1994; Bang and Robins, 2005).

Differences in the ML methods may allow us to trace any differences we find in the estimates. If all the ML estimates differ from the matched estimate, this is probably a result of the fact that the covariates selected for prioritization using subject matter expertise were not the covariates that should have been prioritized, and the ML methods placed greater weight on other observed confounders. We can trace this possibility by re-running the matches without covariate prioritization and observe whether the estimates move closer to those based on ML. However, if only one of the ML methods differ from the matched estimate that will allow us to attribute the difference to the features of that ML method. For example, if only BART differs from the matched estimate, we can attribute the difference to the global nature of BART estimates, and its use of outcome information.

4. Study List

The primary goal of this analysis plan is to identify in advance the set of studies that will serve as published benchmarks. Below, we summarize the set of existing studies based on customized matching that we will use as the benchmarks. For each study, we review the intervention of interest, the baseline covariates, the customized form of matching, and the outcomes. In our comparison to ML methods, we focus on the average difference in the outcomes across the treated and control arms. For some of the studies, the original analysis did not report average differences in the outcomes after matching. Below, we note when this the case. For those studies, we

replicated the outcome analysis using average differences and included those results in the appendix. We do this to ensure that the outcome estimates from the matching that form are benchmarks are known before we use black-box methods.

4.1 Study 1: Right Heart Catherisation

Ramsahai et al. (2011) use a genetic matching algorithm to prioritize balance on subsets of the covariates included in the match. In the study, they consulted a panel of experts to identify the subset of covariates that should have higher priority in the matching process. In their match, they used exact matching, but prioritized balance on the covariates identified as most important by the panel of experts. The treatment of interest was the use of Right Heart Catherisation (RHC) an invasive and controversial monitoring device that is widely used in the management of critically ill patients. The covariates used in the match are are 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, PaO_2/FiO_2 ratio, creatinine, $PaCO_2$, albumin, number of comorbid illnesses, temperature, respiratory rate, heart rate, and white blood cell count. After matching, the primary outcome was mortality within 6 months, with secondary outcomes of length of stay and cost. Outcome estimates were reported as mean differences and included 95% confidence intervals. These mean differences and confidence intervals will be used as the points of comparison with the results using ML methods.

4.2 Study 2: Minority Candidates and Co-Racial Turnout

Keele et al. (2014) examine whether voter turnout is higher among African–American voters when a co-racial candidate is on the ballot in Louisiana mayoral elections. As such, the treatment is the presence of an African–American candidate in Louisiana mayoral elections from 1988 to 2011. In the study, they matched on municipal population, percentage of African–American residents, percentage of residents with college degree, percentage of residents with a high school degree, percentage of residents unemployed, median income, percentage of residents below the poverty line, an indicator for home rule municipal charter, and election year. They customized the match in multiple ways. First, they prioritized balance on the percentage of African–American residents in the municipality and almost exact matching was enforced on election year. Moreover, they enforced balance not only on central moments of the distributions but on higher moments as well. Finally, optimal subset matching was used to find the largest set of observations that met the balance constraints. The outcome is turnout among African–Americans in the municipality measure as a percentage. The authors reported three sets of results. One for general elections, one for runoff

elections, and then a subset of runoff elections where it was thought that the threat of unobserved confounding was lessened.

In the original study, the authors used randomization inference methods based on ranks to estimate the treatment effect and associated 95% confidence intervals. These methods estimate a point estimate closer to a median rather than mean difference across treated and control groups. Treatment effect estimates from ML methods such as BART are average treatment effect estimates. To that end, in the appendix we report revised point estimates and 95% confidence intervals based on mean differences.

4.3 Study 3: Antibiotic Initiation in Critically Ill Children

Ross et al. (2017) investigate the effect of Procalcitonin (PCT)-guided antimicrobial stewardship protocols on antibiotic usage for patients admitted to a single pediatric intensive care unit between February 1, 2011 and February 28, 2014. In the study, they 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. In the match, KS test balance constraints were applied to age and PRISM-III score and patients were exactly matched on reason for PICU diagnosis. The following two outcomes were examined: 1) a binary indicator for initiation of oral or parental antibiotic therapy in the 24 hours prior to or 48 hours after PICU admission and 2) a binary indicator for receiving less than 72 hours of therapy among patients for whom antibiotics were initiated. The original study only reported confidence intervals for risk ratios. In the appendix, we report confidence intervals for average treatment effects.

4.4 Study 4: The 2010 Chilean Earthquake and Posttraumatic Stress

The next study included is an investigation of the effect of the Chilean earthquake in 2010 on mental health (Zubizarreta et al., 2013). In this study, the authors compare Chileans that lived near the earthquake to those who lived quite far away. The outcome is the Davidson Trauma Scale. Chilean residents were matched on the following 46 different covariates: age, gender, member of an indigenous ethnic group, household size, indicators for married, divorced or single, years of education, three indicators for work status, personal income, total household income, 14 measures of housing status prior to the earthquake, 15 measures of health status prior to the earthquake, whether one lived in a rural area, and the estimated propensity score. The authors exactly matched on sex, member of an indigenous ethnic group and age categories, fine balance was applied to ratings of health and housing quality. KS test balance constraints were applied to income. Outcomes were analyzed using

rank tests, so we report the average treatment effect estimate and associated 95% confidence interval in the appendix.

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

The final study that will be replicated is an observational study of the effectiveness of emergency general surgery (EGS) in adults over the age of 65 (Sharoky et al., 2017). In the study the authors, compare patients that received EGS to those that received non-operative care for 59 medical conditions where clinical guidelines do not offer clear a recommendation on whether operative care is superior. Covariates included are patient demographics, number of comorbidities, an indicator for sepsis, an indicator for a pre-operative disability, a series of indicators 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 exactly matched on hospital, the 9 indicators for medical conditions, and the indicators for type of dementia. They also applied near fine balance to the 51 specific sub-condition variables. Outcomes analyzed were mortality, prolonged length of stay, discharge to a higher level of care, discharge to a hospice, and length of stay. We will restrict out comparison to the mortality, prolonged length of stay, and length of stay outcomes. Only these three outcomes will be replicated with black box methods. The other two outcomes have patterns of missingness that were fixed by design in the match. This will be difficult to replicate with the ML methods. In the original study, The measures for length of stay were analyzed using m-estimates. As such, the appendix contains average treatment effect estimates and associated 95% confidence intervals.

5. Discussion

This paper proposes a study design to compare methods of adjustment for confounders in observational studies. The primary goal is to understand whether ML methods with little user input produce different treatment effect estimates compared to studies where matching was used in conjunction with substantive knowledge to customize the form of adjustment for observed confounders. In this analysis plan, we have outlined the five existing studies that will be replicated using ML methods. This will preclude the possibility that we only replicate published studies that either agree or disagree with the benchmark estimates.

Acknowledgments

We thank all the authors who agreed to allow us to replicate their studies. We also thank Jake Bowers and Jas Sekhon for useful advice and comments.

Appendix A.

We replicated the outcome analysis for the four studies that did not report average treatment effect estimates. We use the average treatment effect to ensure direct comparability between the estimates after matching and those from ML methods.

Table 1: Outcome Analysis Replication for Study 2: Minority Candidates and Co-Racial Turnout

	General Elections	Runoff Elections	Runoff Election Subset
Point Estimate	3.41	5.11	2.97
95% Confidence Interval	[0.72, 6.1]	[-3.5, 14]	[-11, 17]
N	394	54	30

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

Table 2: Outcome Analysis Replication for Study 3: Antibiotics in Critically Ill Children

	Antibiotics >72 hours (0/1)	Antibiotic Initiation (0/1)
Point Estimate	0.112	0.202
95% Confidence Interval	[0.049, 0.17]	[0.15, 0.25]
N	842	988

Note: Point estimates are differences in proportions.

Table 3: Outcome Analysis Replication for Study 4: Chilean Earthquake

	Davidson Trauma Scale
Point Estimate	18.4
95% Confidence Interval	[17, 19]
N	5040

Table 4: Outcome Analysis Replication for Study 5: EGS

	pLOS (0/1)
Point Estimate	0.093
95% Confidence Interval	[0.09, 0.095]
N	351850
	Mortality (0/1)
Point Estimate	-0.0013
95% Confidence Interval	[-0.003, 0.0002]
N	351850
	LOS (Days)
Point Estimate	3.53
95% Confidence Interval	[3.5, 3.6]
N	351850

References

Athey, S., Tibshirani, J., and Wager, S. (2016). Generalized Random Forests. *ArXiv e-prints*.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Barnow, B., Cain, G., and Goldberger, A. (1980). Issues in the analysis of selectivity bias. In Stromsdorfer, E. and Farkas, G., editors, *Evaluation Studies*, volume 5, pages 43–59. Sage, San Francisco, CA.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of Royal Statistical Society, Series A*, 128(2):234–265.

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(3):932–945.

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(1):106–117.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618.

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(3):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(3):477–513.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1).

Iacus, S. M., King, G., and Porro, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.

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.

Kennedy, E., Sjölander, A., and Small, D. (2015). Semiparametric causal inference in matched cohort studies. *Biometrika*, 102(3):739–746.

Kilcioglu, C. and Zubizarreta, J. R. (2017). Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings. *Annals of Applied Statistics*, In press.

Lu, B., Zutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253.

Manski, C. F. (2007). *Identification For Prediction And Decision*. Harvard University Press, Cambridge, Mass.

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(510):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(3-4):95–114.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(4):1024–1032.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer-Verlag, New York.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21(1):57–71.

Rosenbaum, P. R. (2017). *Observation and experiment: an introduction to causal inference*. Harvard University Press, Cambridge, MA.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007). Mimimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of propensity scores in observational studies for causal effects. *Biometrika*, 76(1):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. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 6(5):688–701.

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

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(1).

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

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(2):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(500):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(1):79–87.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1):204–231.