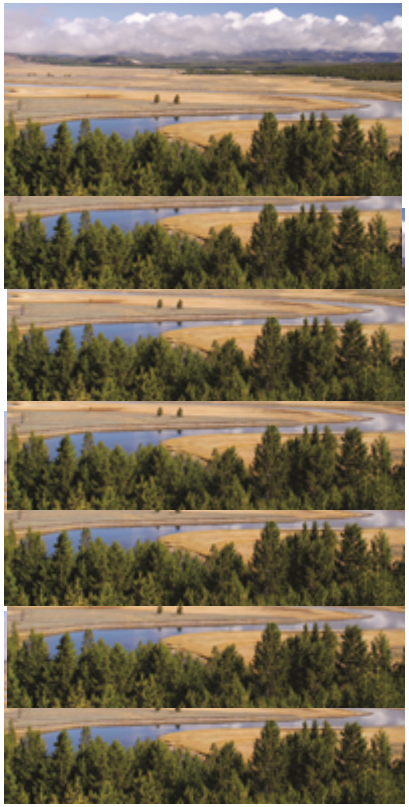
Propensity scores and causal inference using machine learning methods

Austin Nichols (Abt) &
Linden McBride (Cornell)

July 27, 2017

Stata Conference

Baltimore, MD

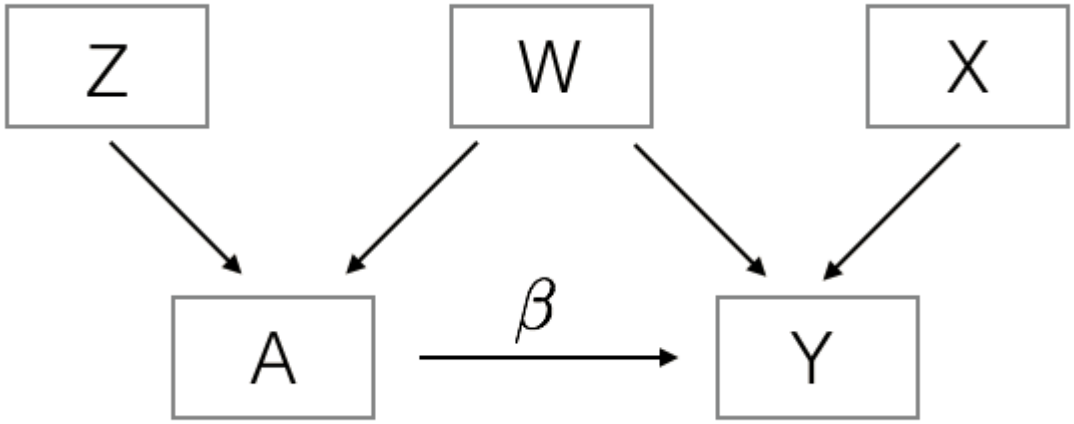


Overview

- Machine learning methods dominant for classification/prediction problems.
- Prediction is useful for causal inference if one is trying to predict propensity scores (probability of treatment conditional on observables);
- But sometimes better predictions lead to worse causal inference: greater bias or mean squared error (MSE).
- Simulation shows some machine learning methods are more robust to specification error when not all confounders are observed (the real world case).

Why propensity scores?

- With nonrandom selection of treatment status A, can estimate average treatment effect β by conditioning on all possible confounders W (if we observe all of them).
- Even if we observe all W, do we know the right functional form?
- Propensity score matching or weighting solves the **functional form** problem (not the incomplete observation problem).



Why propensity score weights?

- Where potential outcomes are conditionally independent of A given W, they are also conditionally independent given the conditional probability of A, E(A|W), a.k.a. the propensity score.
- We can condition on the propensity score to eliminate bias due to confounders (Rosenbaum and Rubin, 1983);
- We can improve efficiency by using **estimated** E(A|W) to reweight the data (Hirano, Imbens, & Ridder 2003), even if we know the true propensity score (i.e. we throw away that information).

How should we estimate E(A|W)?

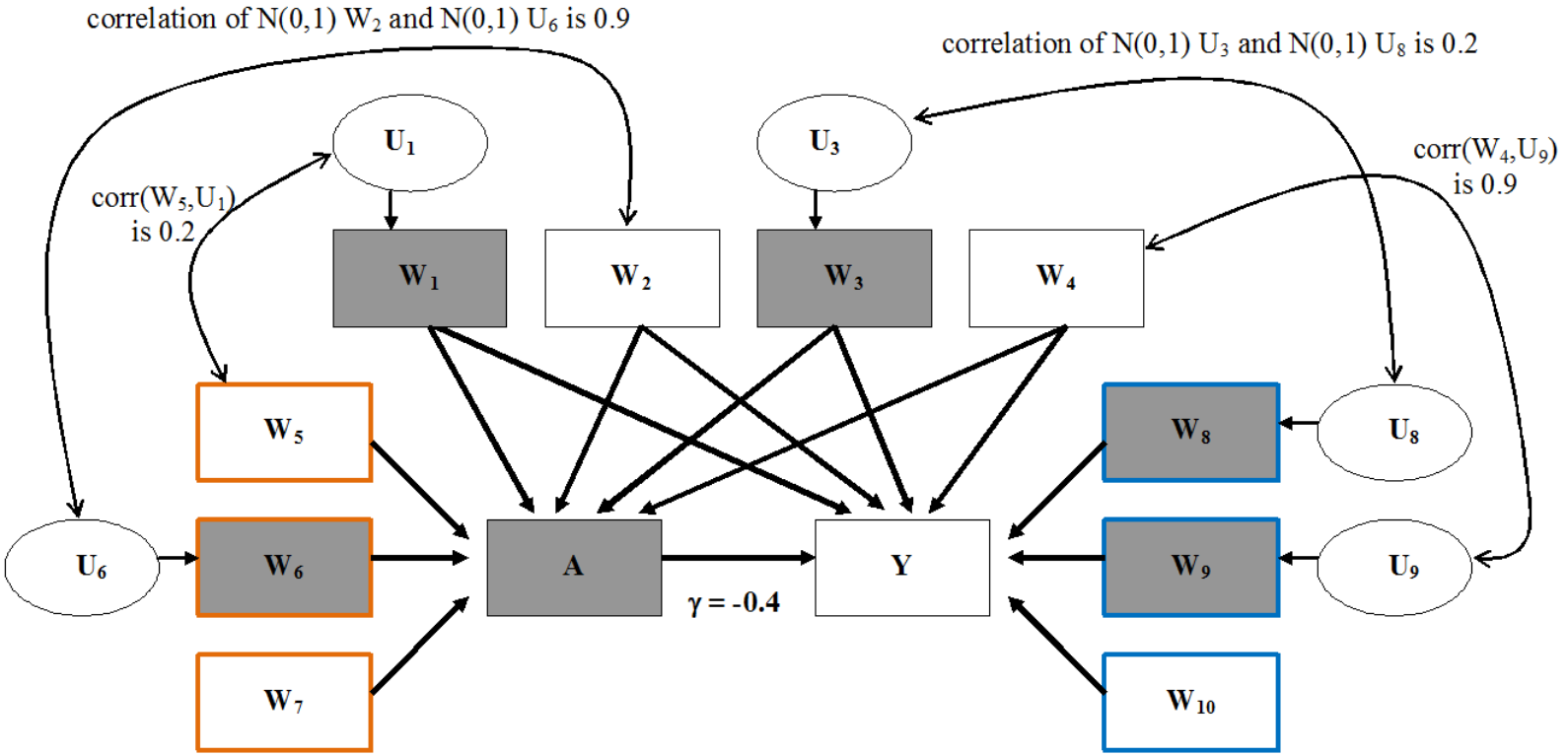
- This is a prediction problem, and one at which machine learning algorithms have been shown to excel
- A few papers explore the use of machine learning approaches, but with full set of confounders
  - Zador, Judkins, and Das (2001): MART for survey nonresponse adjustment
  - McCaffrey, Ridgeway & Morral (2004): Generalized boosted model for PSW
  - Setoguchi et al. (2008): Neural networks, CART for PSM
  - Lee, Lessler & Stuart (2009): CART, Pruned CART, Bagged CART, Random Forest, Boosted CART for PSW
  - Diamond & Sekhon (2013): Genetic Matching, Random Forest, Boosting for PSM

Simulations

- To ascertain finite-sample performance, we run a large set of simulations (each one 10,000 iterations)
- Each simulation imposes a causal diagram with binary treatment A and 10 potential confounders W, of which only 4 are actually confounders
- Vary the functional form for E(A|W): base case is logit with no interactions, but allow nonlinearity/interactions
- Estimation default as in **teffects ipwra** (outcome model is linear, treatment model is logit)

Simulation causal diagram

- Simulation structure follows Setoguchi et al. (2008) and Lee, Lessler, and Stuart (2009). Shaded boxes binary, orange excluded instruments (Z), blue controls (X).



Variation across 7 scenarios

- x x (base case) means no interactions in true model: non-additivity means e.g. coef on W_2W_4 is zero; nonlinearity means e.g. coef on W_2W_2 is zero.

Scenario	Non-additivity	Non-linearity
a	x	x
b	x	+
c	x	++
d	+	x
e	+	+
f	++	x
g	++	++

Simulations

- Can assume we know the true functional forms (unrealistic), or there is some specification error (i.e. unmodeled interactions in a logit).
- Can assume we observe all confounders W_1, W_2, W_3, W_4 (unrealistic) or some subset.
- Conditioning on a proper subset of confounders can decrease bias or amplify bias: useful figures in Steiner and Kim (2016).
- We compare mean-squared error (MSE) and bias reduction.

Not every confounder can be observed

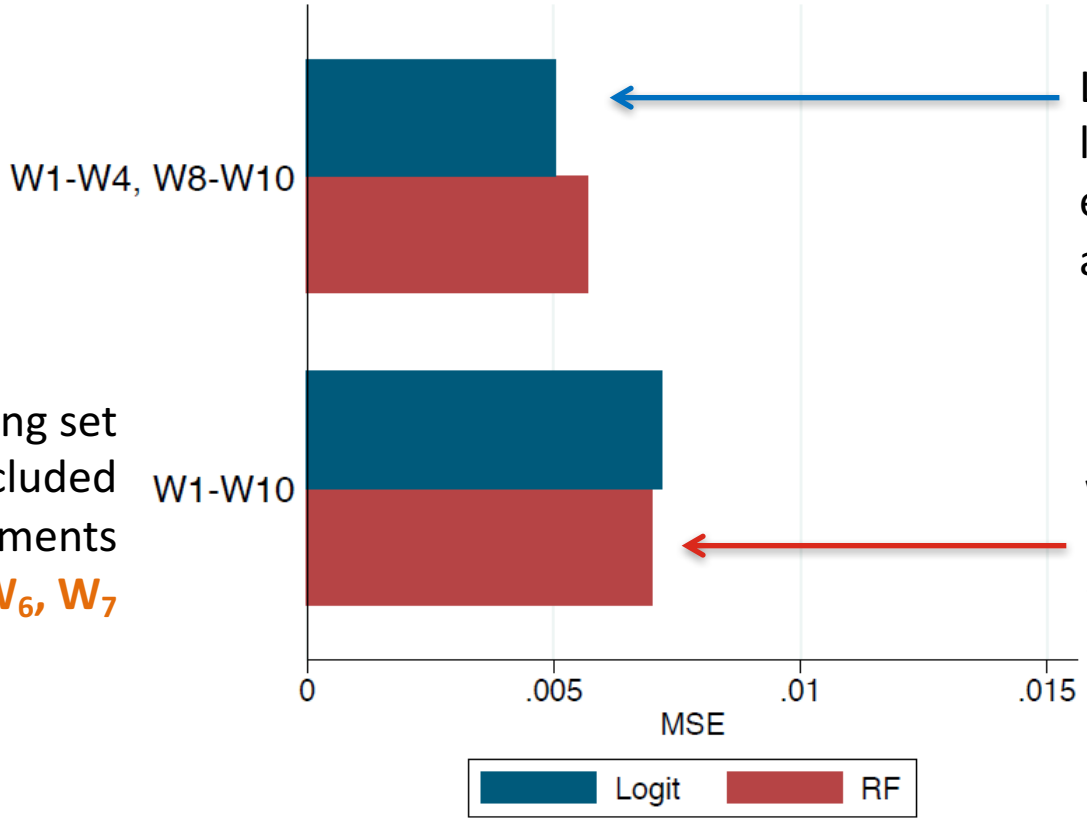
- This means we should not assume we can see W1, W2, W3, W4.
- Where there are omitted variables (i.e., absent full ignorability), including an excluded instrument in propensity score estimation increases the inconsistency (Heckman and Navarro-Lozano 2004, Battacharya and Vogt 2007) and bias (Pearl 2011, Wooldridge 2009) of the estimator.
- We report results for simulations with and without conditioning on the excluded instruments W5, W6, W7.

Methods to estimate E(A|W)

- **Logit**: Generalized linear model with log odds linear in parameters
- Regression tree (RT): Algorithm recursively partitions a feature space to minimize distance between mean and predicted outcomes within each partition; implemented with rpart in R
- Pruned regression tree (PRT): Prunes back RTs to prevent overfitting; rpart
- **Regression forest (RF)**: Produces many low bias RTs from random subsets of the data and then averages across those trees to reduce variance of predictor; randomForest
- Boosted regression tree (BRT): Builds out trees on residuals of prior bifurcations in the feature space; gbm
- Least absolute shrinkage and selection operator (LASSO): Penalized regression; glmnet

Scenario a

- Logit is the "right" model but underperforms RF when we improperly condition on excluded instruments:

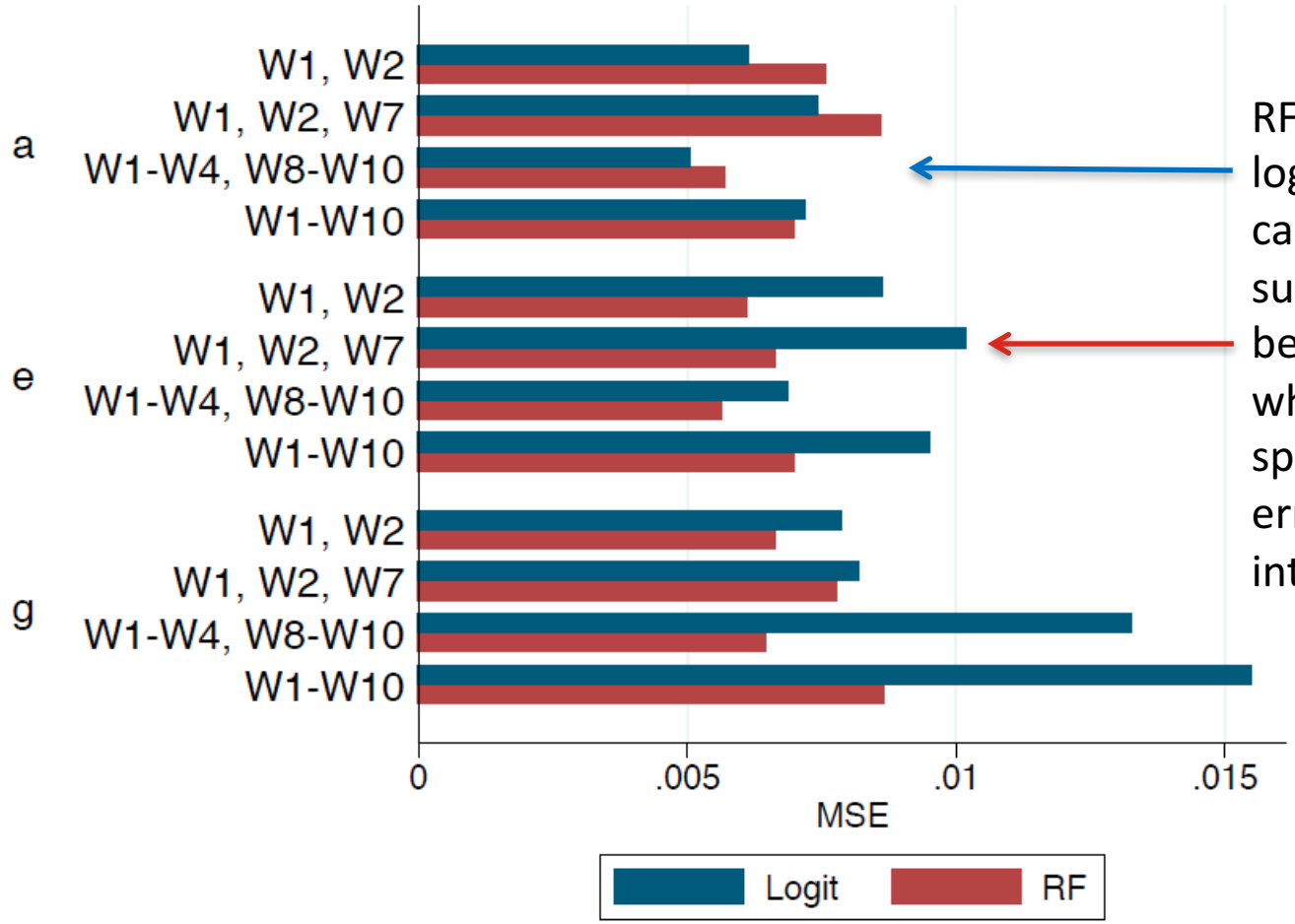


Lower MSE is better; logit wins when no excluded instruments are in treatment model

Conditioning set includes excluded instruments W_5, W_6, W_7

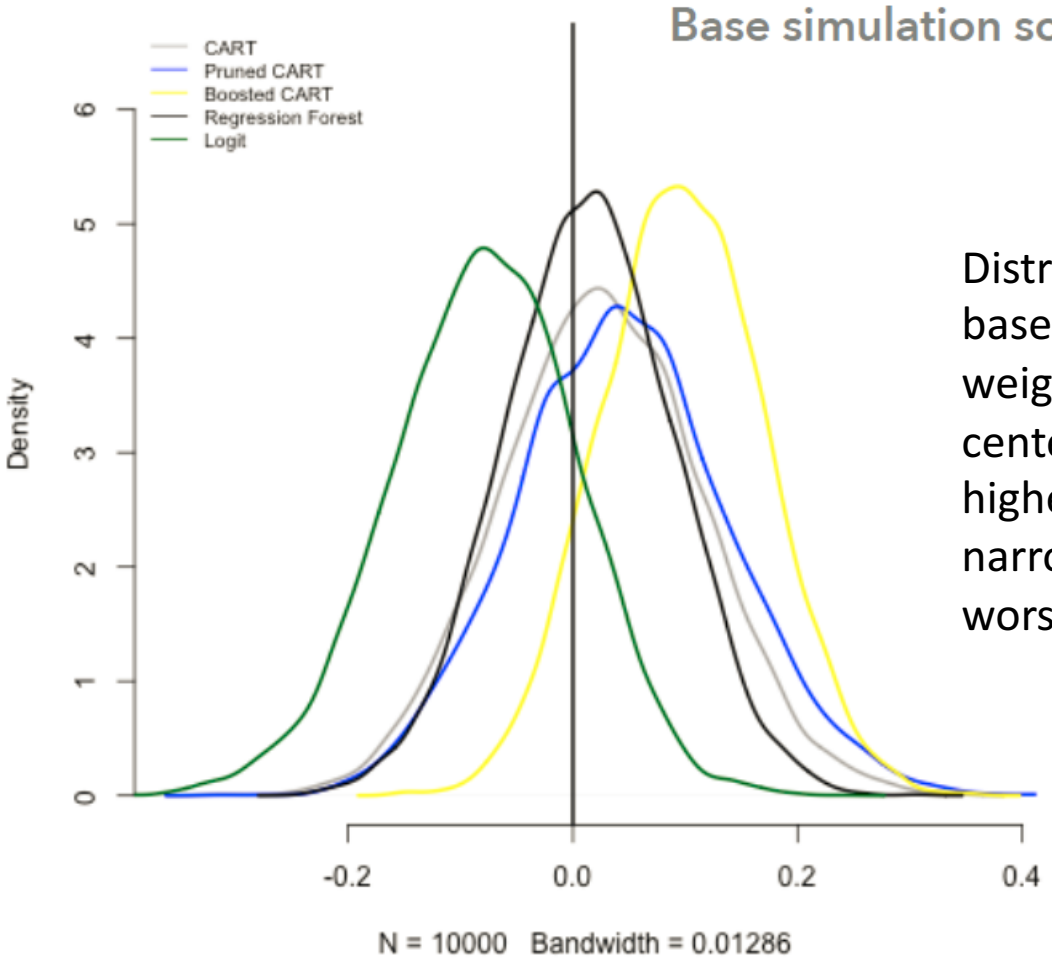
But RF wins, even when logit is the right model, with excluded instruments in tmodel

RF more robust to spec. error



RF similar to logit in easy cases, but substantially better than logit when some specification error is introduced

RF best in class on bias, MSE



Distribution of RF-based propensity-score-weighting estimator centered on truth, with higher peak and narrower spread, in the worst case scenario g.

Conclusions

- We offer guidance on selection of variables and methods for the estimation of propensity scores when measuring average treatment effects (ATE) or ATE on the treated (ATT) under different scenarios.
- Machine learning estimators, especially regression forest (RF), perform well where the treatment assignment mechanism is unknown and can offer better protection against improper conditioning on excluded instruments when not all confounders are included (the realistic case).

Conclusions

- No statistical test can distinguish confounders and excluded instruments.
- Theory and assumptions (i.e. a good causal diagram) play an outsized role in which variables to include and which estimation approaches to use in which settings;
- However, propensity score reweighting using regression forest dominates several alternatives in a realistic class of settings.
- MSE for IV is largest in all simulations.

Next steps

- Still more variations on this theme to be explored.
- But.
- Current generation of machine learning algorithms targets quality of predictions, not quality of causal inference using those predictions.
- In process: improved stochastic ensemble methods (along the lines of regression forests) as a Stata package.

References

Bhattacharya, J. and Vogt W. (2007.) "Do Instrumental Variables Belong in Propensity Scores?" NBER Technical Working Papers, 343. Cambridge, MA: National Bureau of Economic Research.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984.) *Classification and regression trees.* CRC press.

Breiman, L. (1996.) "Bagging predictors." Machine learning 24(2): 123-140.

Chernozhukov, V., C. Hansen, and M. Spindler. (2015.) "Post-Selection and Post-Regularization Inference in Linear Models with Very Many Controls and Instruments." American Economic Review.

Cole, Stephen R. and Elizabeth A. Stuart. (2010.) Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. American Journal of Epidemiology 172(1): 107-15.

Diamond, A., and J. Sekhon (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.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004). "Least angle regression." Annals of Statistics, 32(2): 407-499.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005.) "The elements of statistical learning: data mining, inference and prediction." The Mathematical Intelligencer, 27(2), 83-85.

Hirano, K., Imbens, G. W., Ridder, G. (2003.) "Efficient estimation of average treatment effects using the estimated propensity score." *Econometrica,* 71(4), 1161-1189.

Lee, Brian, Justin Lessler, and Elizabeth A. Stuart. (2010.) "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine,* 29:3, 337346.

Liaw, A. and M. Wiener (2002.) Classification and Regression by randomForest. R News 2(3),18-22.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral. (2004.) "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." Psychological Methods 9:4: 403-425.

Pearl, J. (2011). "Invited Commentary: Understanding Bias Amplification." *American Journal of Epidemiology,* 174(11): 1223-1227.

Ridgeway, G. with contributions from others (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. http://CRAN.R-project.org/package=gbm

Rosenbaum, Paul R., and Donald B. Rubin, The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika,* 70:1 (1983), 4155.

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., and Cook, E.F. (2008). "Evaluating uses of data mining techniques in propensity score estimation: a simulation study." *Pharmacoepidemiol Drug Saf,* 17(6): 546-555.

Steiner, Peter M., and Yongnam Kim. (2016). "The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases." *Journal of Causal Inference*, 4(2).

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society Series A 174 (2): 369-386.

Therneau, T., B. Atkinson, B. Ripley (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. http://CRAN.R-project.org/package=rpart

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." J. Royal. Statist. Soc B., 58(1): 267-288.

Wooldridge, J. (2009). "Should instrumental variables be used as matching variables?" East Lansing, MI: Michigan State University. [http://econ.msu.edu/faculty/wooldridge/docs/treat1r6.pdf]

Zador, Paul, David Judkins, and Barnali Das. (2001). Experiments with MART to Automate Model Building in Survey Research: Applications to the National Survey of Parents and Youth. *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*