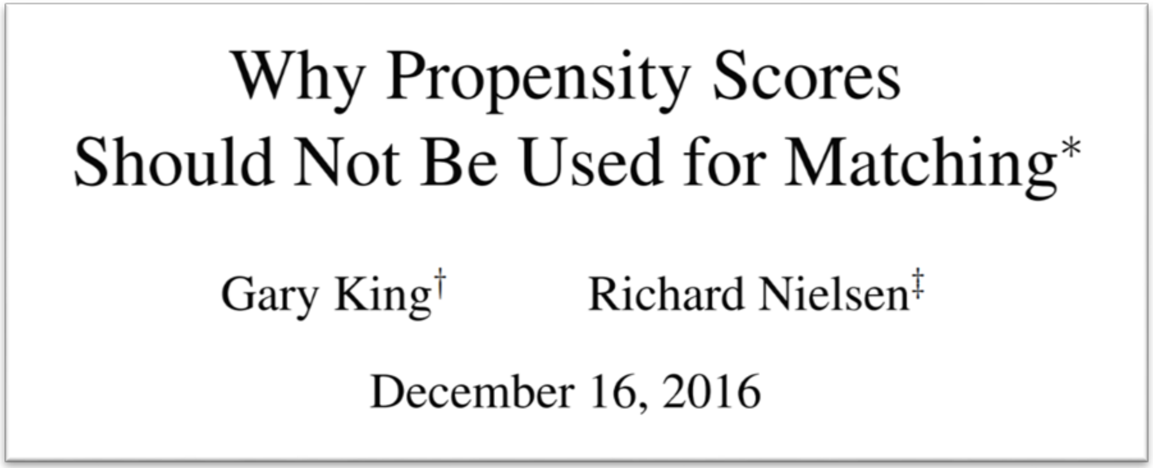
Weill Cornell
Medicine



Why Propensity Scores Should Not Be Used for Matching[*]

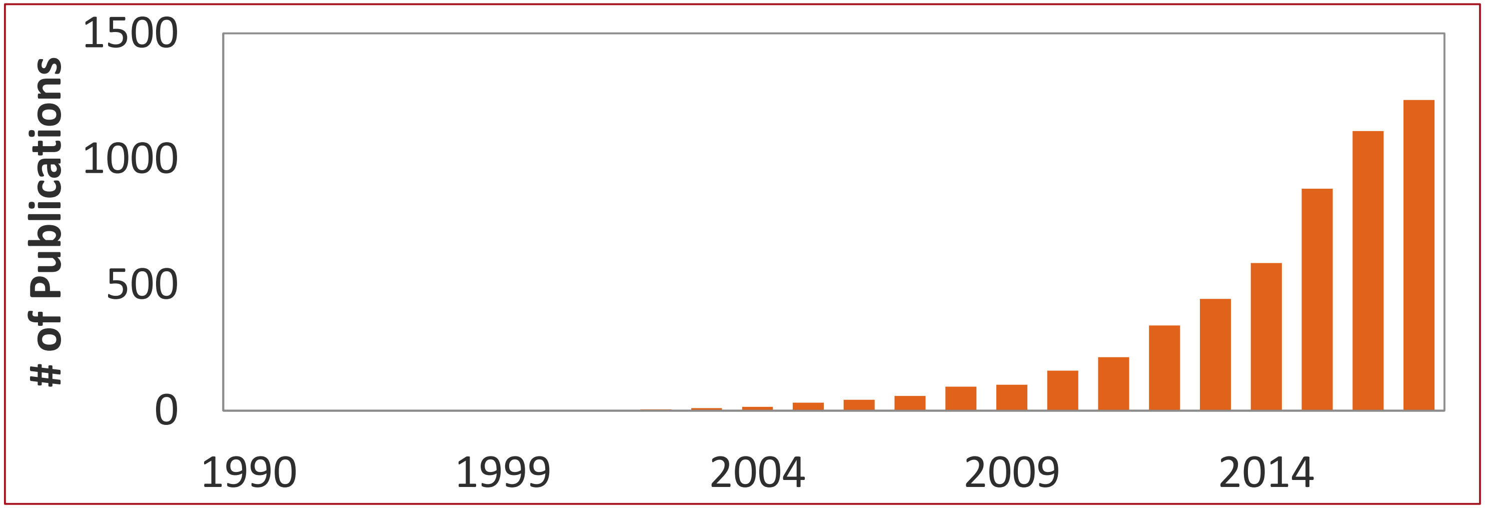
Gary King[†] Richard Nielsen[‡]

December 16, 2016

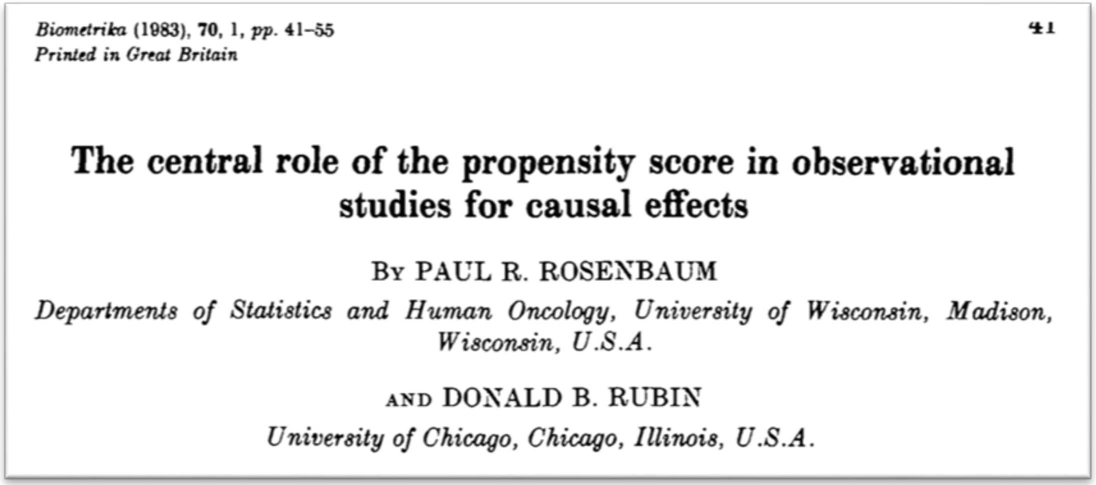
Hanhan Wang
August 29th, 2017

Introduction

- Matching is often used for preprocessing data for causal inference
- Matching reduces:
  - Imbalance
  - Model dependence
  - Inefficiency and bias
- Propensity score matching (PSM) is one of the most popular matching methods



Introduction



Biometrika (1983), **70**, 1, *pp.* 41–55
Printed in Great Britain

41

The central role of the propensity score in observational studies for causal effects

By PAUL R. ROSENBAUM

Departments of Statistics and Human Oncology, University of Wisconsin, Madison, Wisconsin, U.S.A.

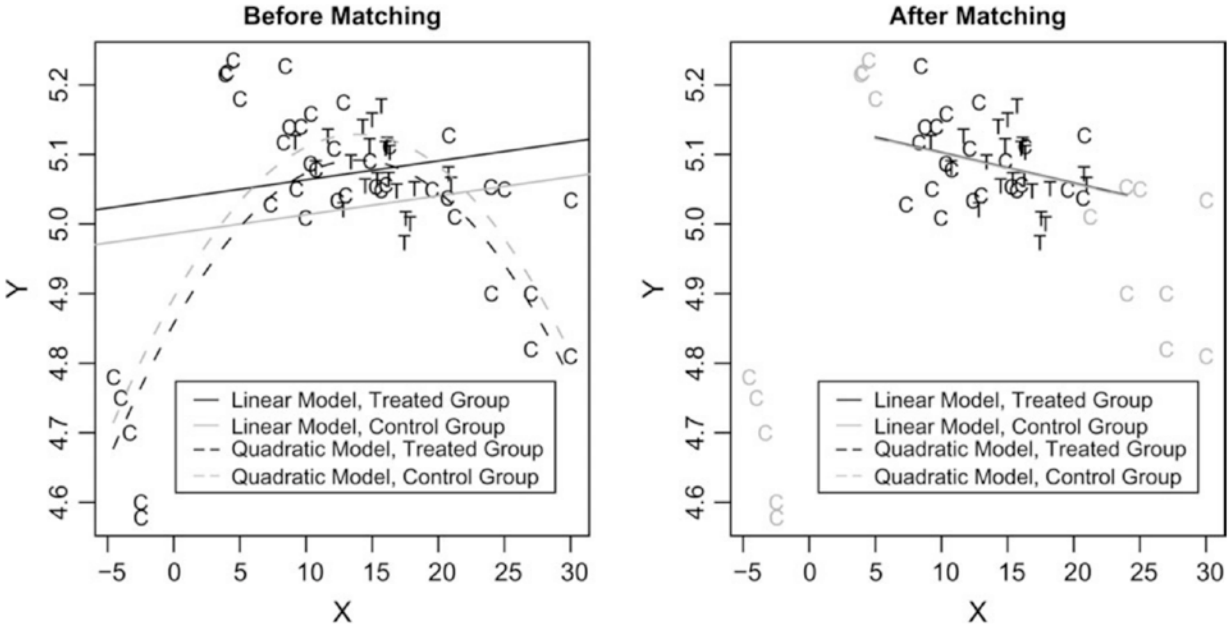
and DONALD B. RUBIN

University of Chicago, Chicago, Illinois, U.S.A.

- Other applications of propensity scores
  - Regression adjustment (Vansteelandt and Daniel, 2014)
  - Inverse weighting (Robins, Hernan and Brumback, 2000)
  - Stratification (Rosenbaum and Rubin, 1984)
  - Use of propensity score in other methods (e.g. Diamond and Sekhon, 2012; Imai and Ratkovic, 2014)

The Problem of Model Dependence in Causal Inference

- Estimating causal effect
  - o Ideal situation – experiment data
    - – Expensive
    - – Not feasible/ethical
  - o Reality – observational data



Matching

- Response variable: Y_i
- Treatment variable: $T_i \in \{0,1\}$ (1=treated, 0=control)
- Confounders: X_i

- Treatment Effect for treated observation i:

$$TE_i = Y_i - Y_i\,(0)$$

- Assumptions:
  - Overlap assumption: $0 < \Pr(\,T_i = 0|X) < 1$ for all i
  - Stable unit treatment value assumption: Y_i (0) does not change if T_i changes from 0 to 1
  - Unconfoundedness assumption: $[Y(0), Y(1)] \perp T|X$
- Quantities of Interest:
  - SATT: Sample Average Treatment effect on the Treated:

$$\tau = \text{mean}_{i \in \{i|T_i=1\}}(\text{TE}_i)$$

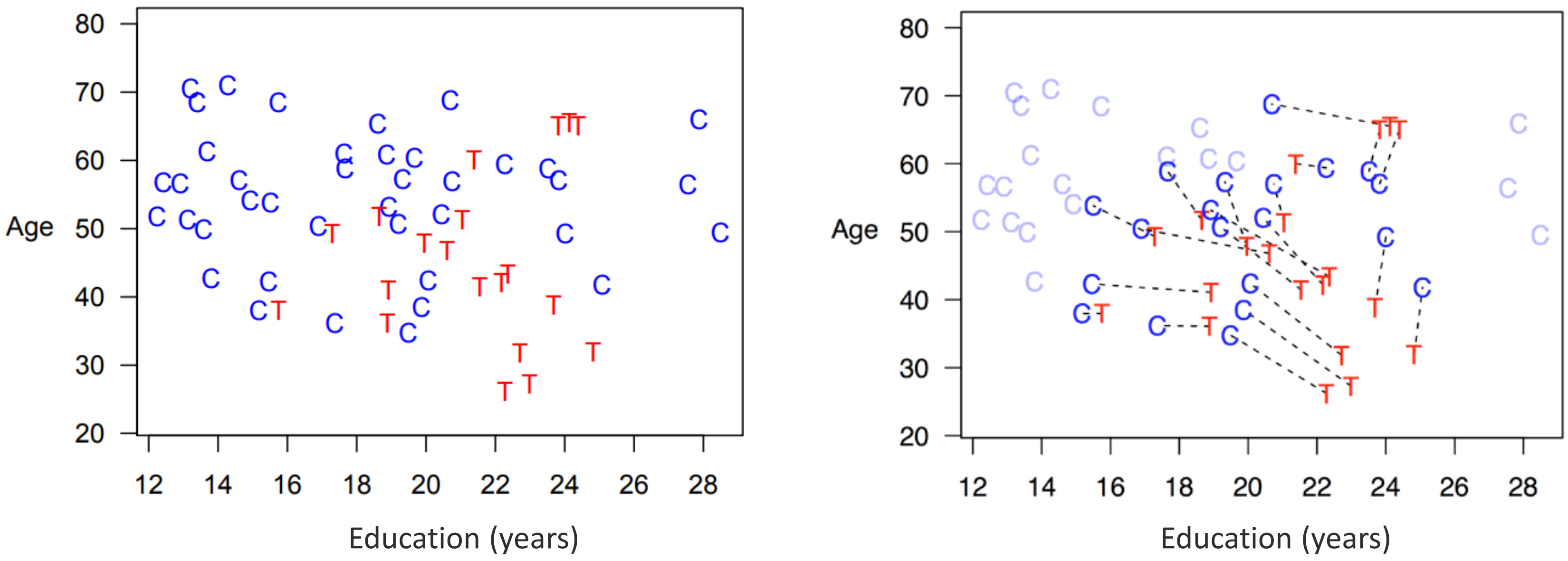
  - FSATT: Feasible SATT (prune badly matched treated subjects)

Matching

- Benefits:
  - Can simplify the analysis of causal effects
  - Reduces dependence of estimates on parametric models.
- Matching finds hidden randomized experiments within the observational data
  - PSM approximates complete randomization
  - Other matching methods approximates fully blocked randomized design
- Completely randomized experiments do not balance observed covariates exactly, thus less optimal

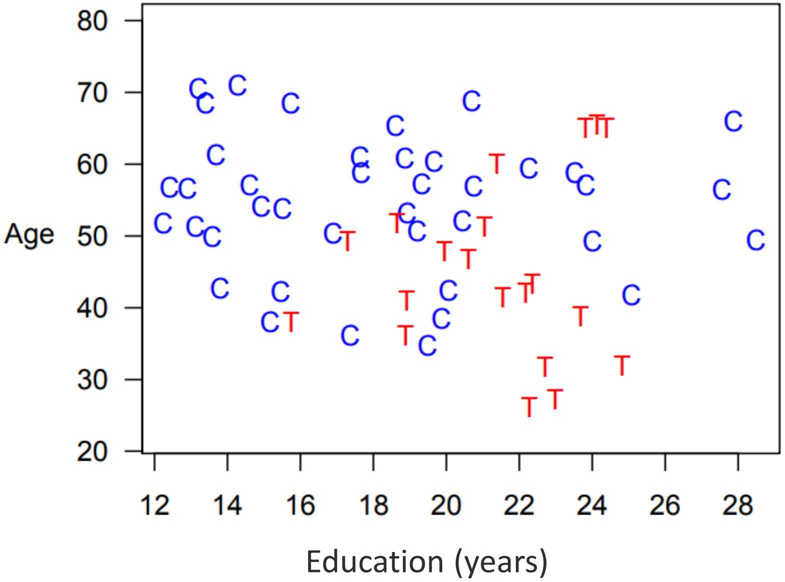
Mahalanobis Distance Matching (MDM)

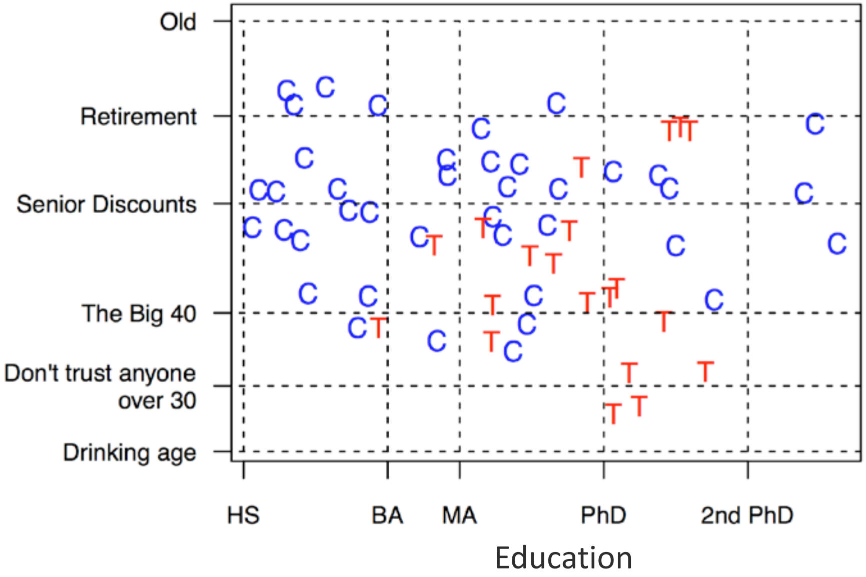
$$\mathbb{X}_{\text{MDM}} = M\left(X \,\middle|\, \sqrt{(X_i - X_j)S^{-1}(X_i - X_j)} < \delta\right)$$



Coarsened Exact Matching (CEM)

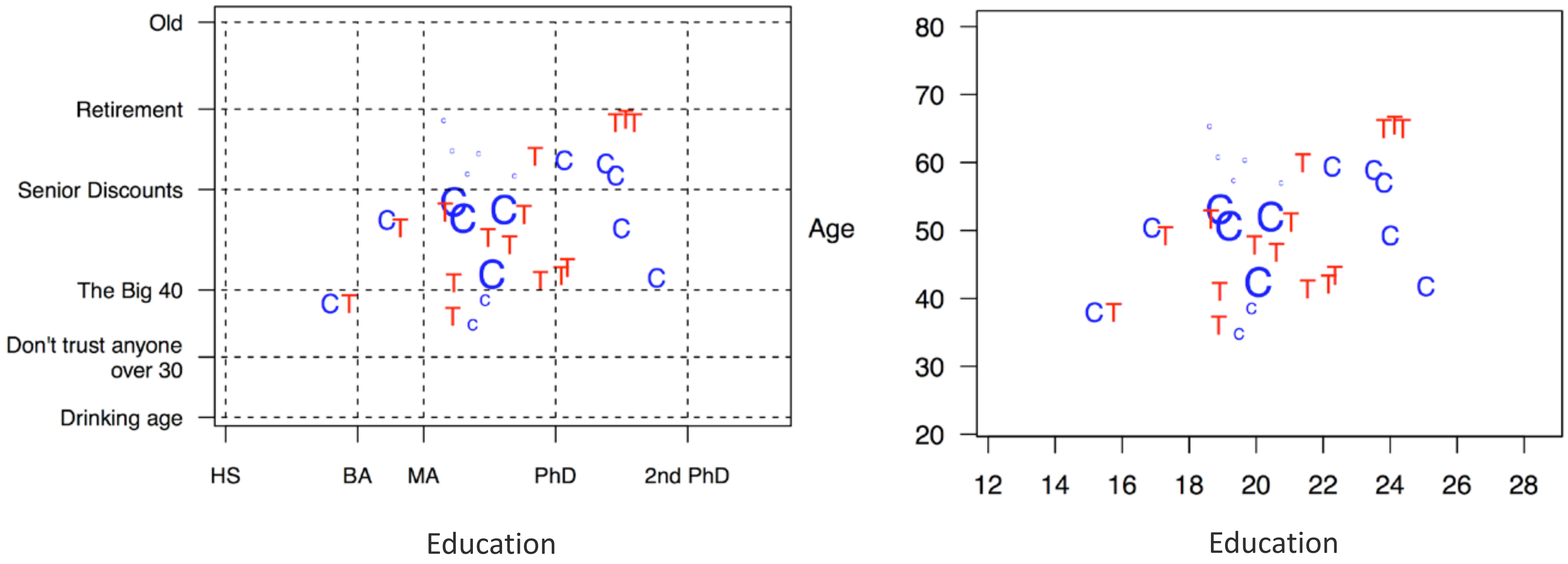
$$\mathbb{X}_{\text{CEM}} = M[X \mid C_\delta(X_i) = C_\delta(X_j)]$$





Coarsened Exact Matching (CEM)

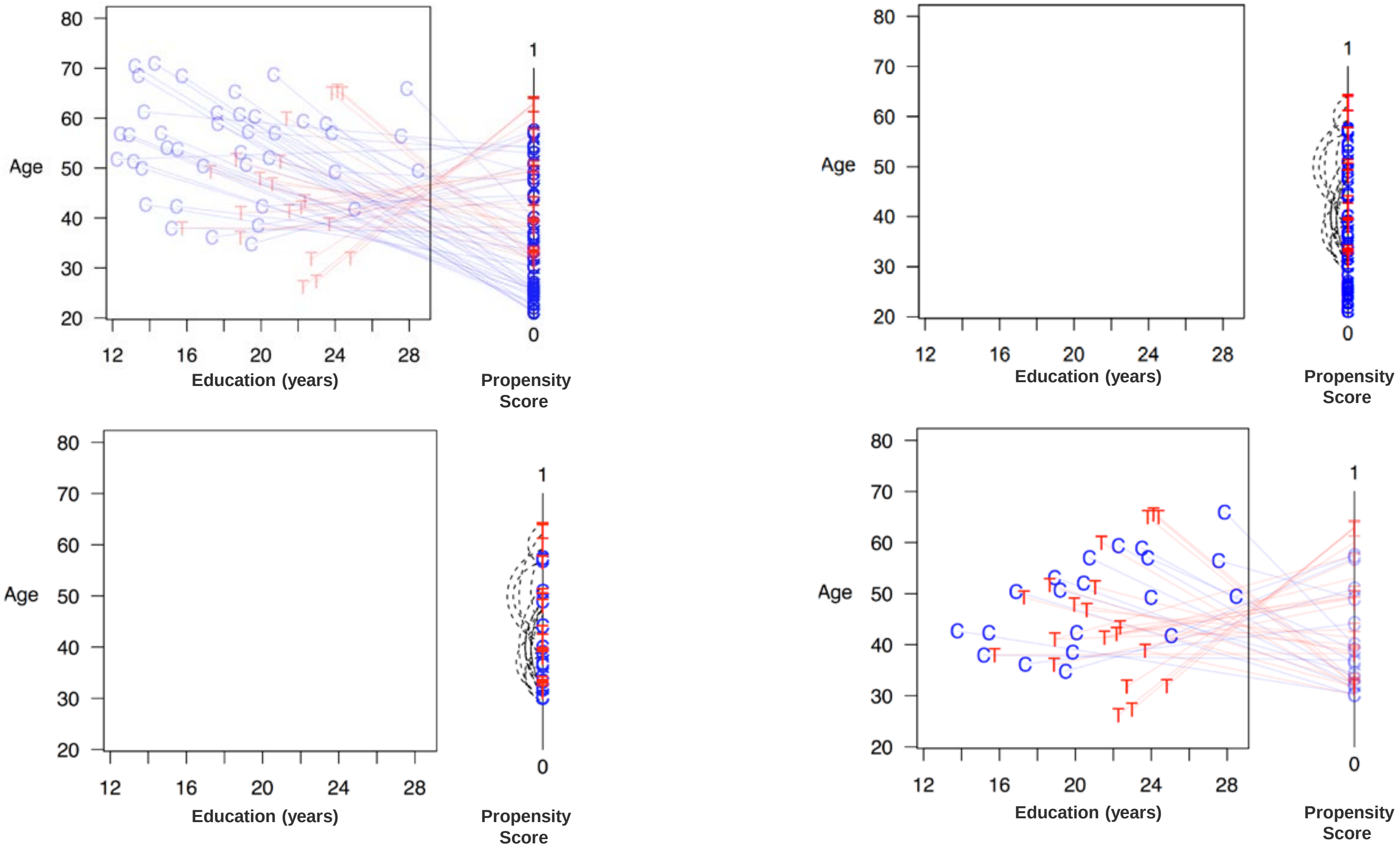
$$\mathbb{X}_{\text{CEM}} = M[X \mid C_\delta(X_i) = C_\delta(X_j)]$$



Propensity Score Matching (PSM)

$$\mathbb{X}_{\text{PSM}} = M\Big(X \Big| |\hat{\pi}_i - \hat{\pi}_j| < \delta\Big), \text{ where } \pi_i \equiv \Pr(T_i = 1|X_i)$$

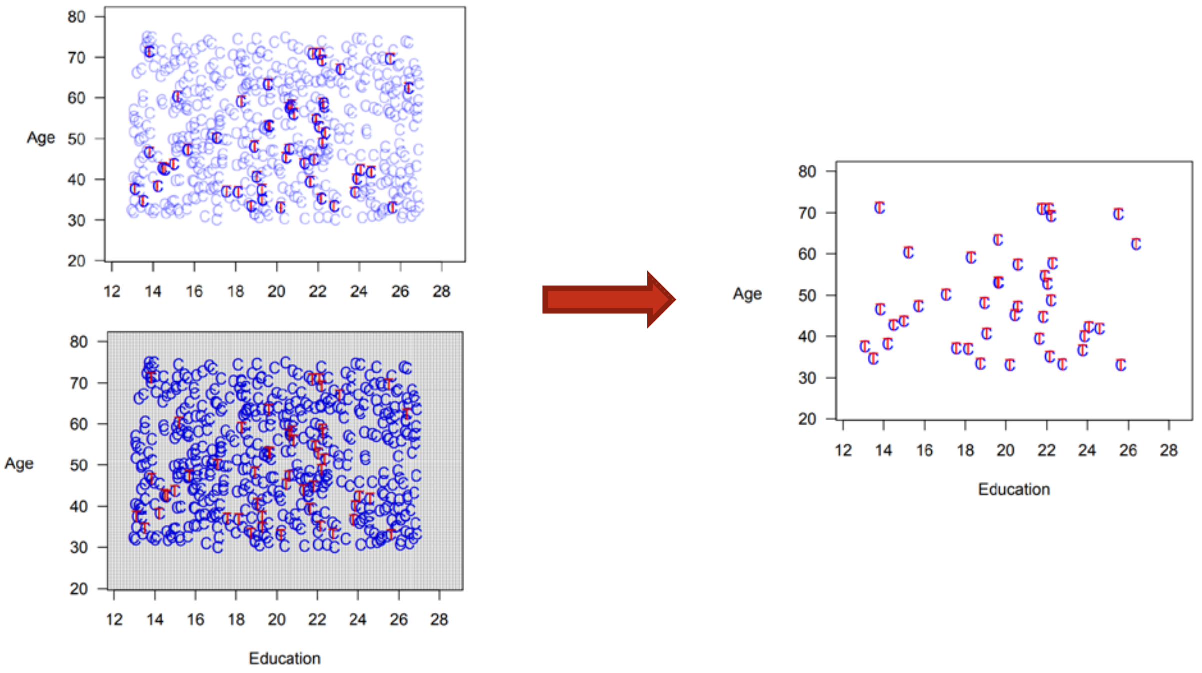
$$\hat{\pi}_i = (1 + e^{-X_i\hat{\beta}})^{-1}$$



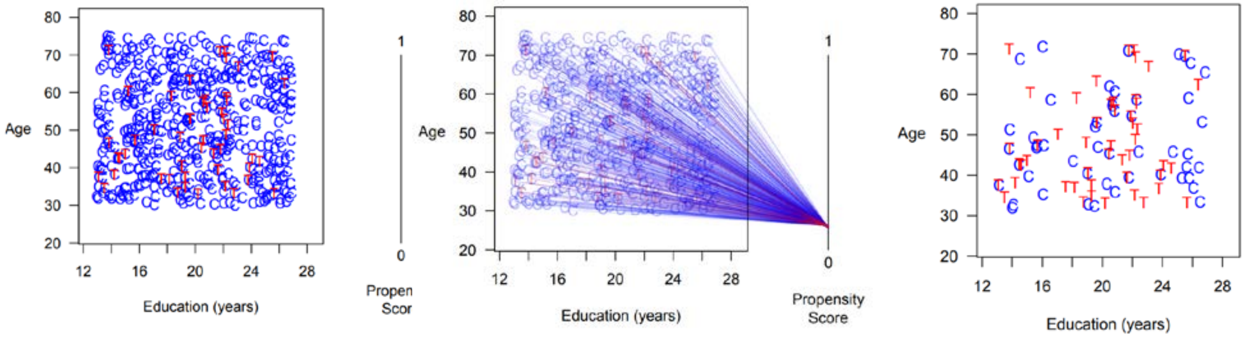
Best Case

MDM

CEM

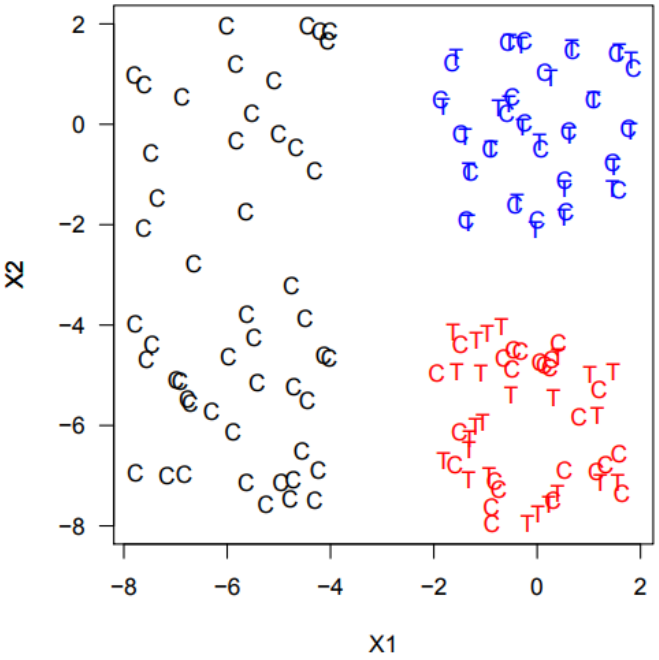


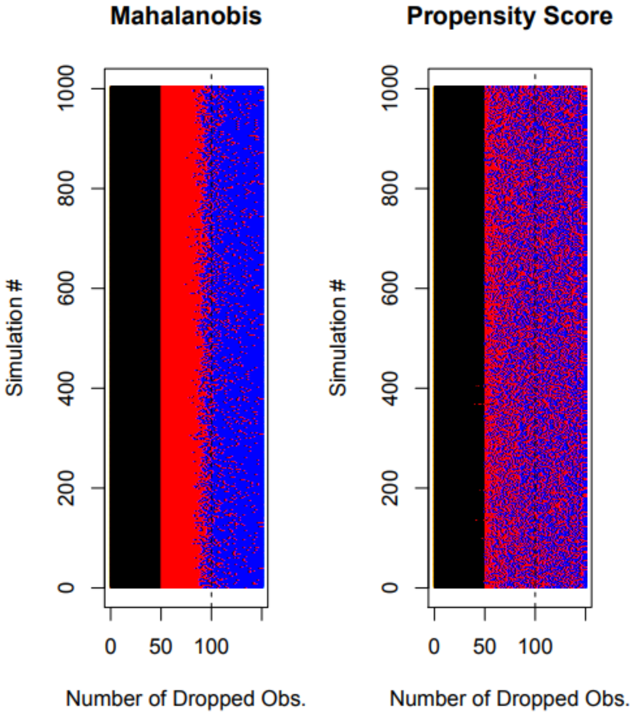
PSM



Information Ignored by Propensity Scores

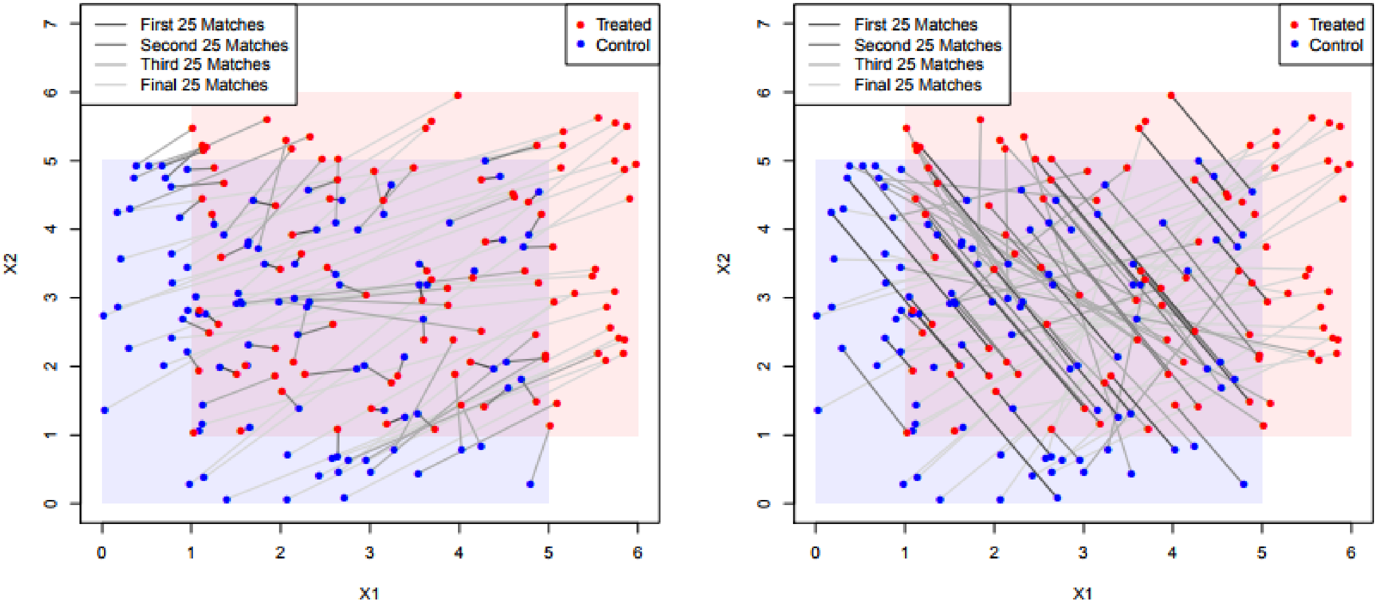
- **1000 simulated data set:**
  - **Matched pair randomized experiment**
  - **Completely randomized experiment**
  - **Control units from an imbalanced observational data set**





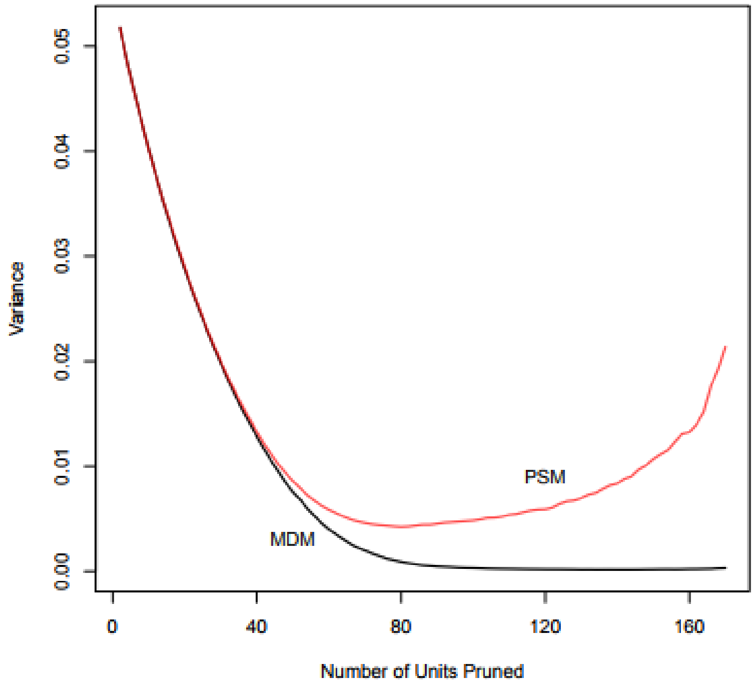
The Propensity Score Paradox

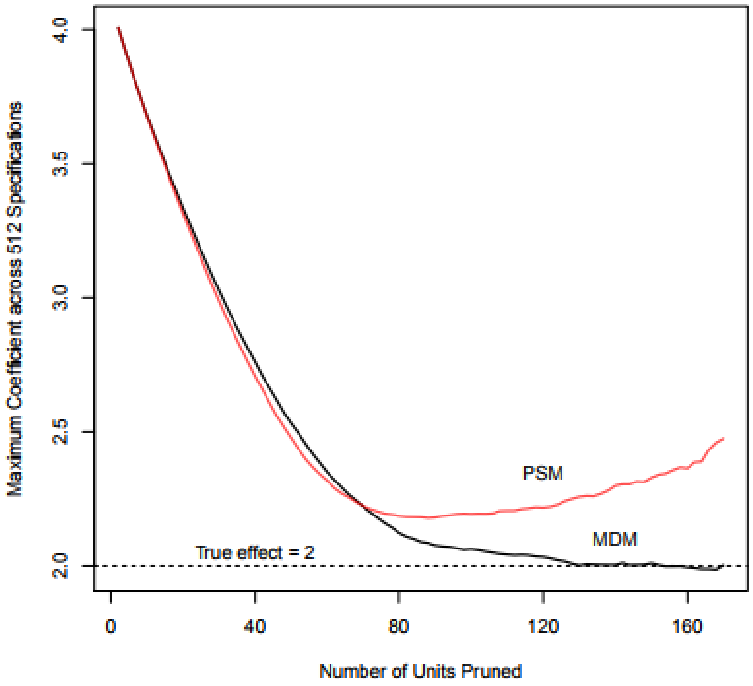
- When complete randomization has been approximated, more pruning using PSM will increase imbalance
- Generate 100 simulated dataset
  - o Randomly draw 100 control subjects from $Unif(0,5)$ and 100 treatment subjects from $Unif(1,6)$ for X_1 and X_2
  - o Generate $Y_i = 2T_i + X_{i1} + X_{i2} + \epsilon_i$, where $\epsilon \sim N(0,1)$



The Propensity Score Paradox

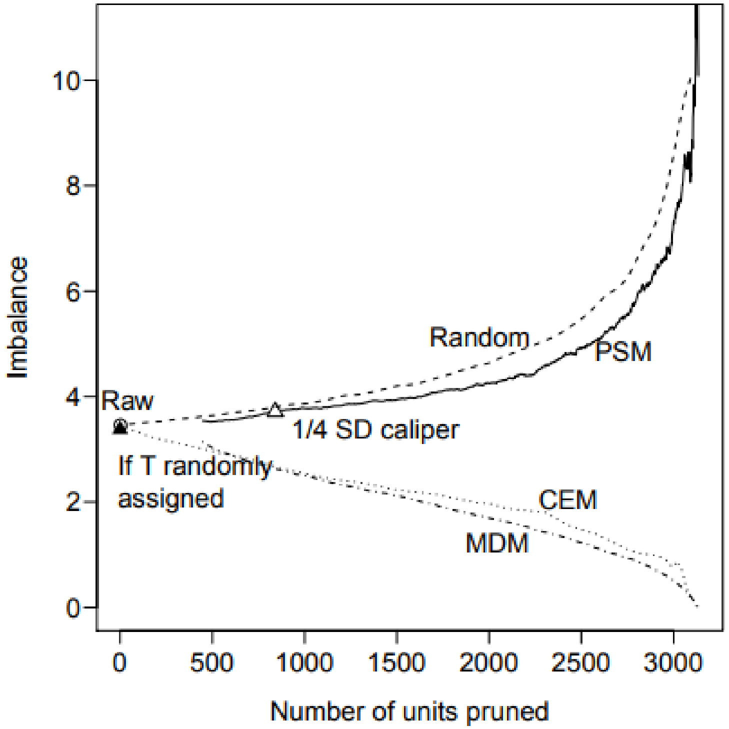
- Average variance in the causal effect estimate over 100 data sets across 512 models
- Maximum estimated causal effect from 512 models applied to 100 data sets

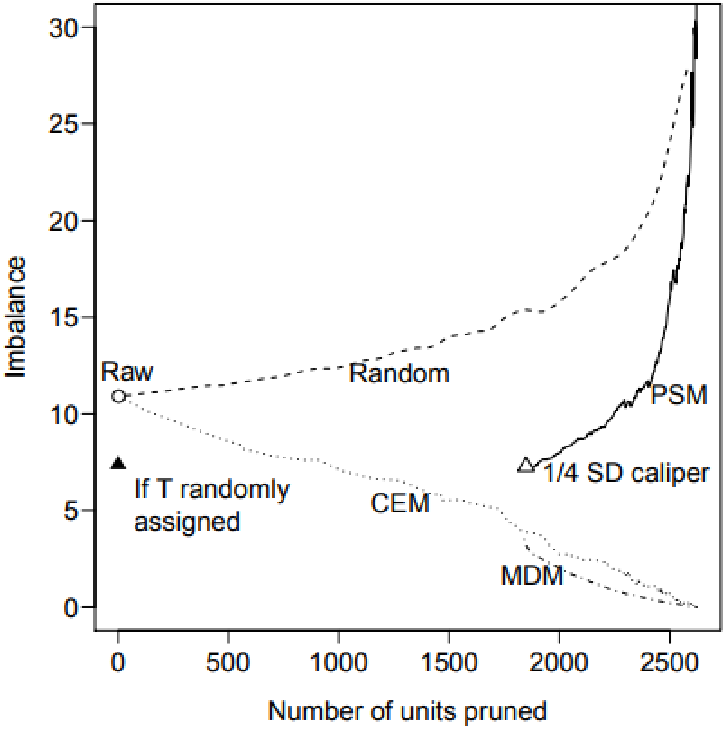




Damage Caused in Real Data

- Imbalance measured by “Mahalanobis Discrepancy” (Abadie and Imbens, 2006)





Advice for Users

- Propensity score matching:
  - Scale variables to represent their importance
  - Report techniques used to avoid problems (imbalance, model dependence, bias, etc.)
  - Be aware that PSM can help the most in data where valid causal inferences are least likely (i.e., with high levels of imbalance) and may do the most damage in data that are well suited to making causal inferences (i.e., with low levels of imbalance).
  - Understand what happens when combining PSM with other matching methods
- Other matching methods
  - Any matching method that prune independent of the covariates can increase imbalance
  - More data, more information
  - Choose a method that can match on all X

Conclusions

- PSM approximates completely randomized experiment
- However a fully blocked randomized experiment can do better
- The PSM paradox
- Doubts on PSM related practices and recommendations:
  - Perform PSM on data from completely randomized experiments
  - ¼ caliper on propensity score
  - Include all available covariates

References

- King, Gary, and Richard Nielsen. "Why propensity scores should not be used for matching." Copy at http://j. mp/1sexgVw Download Citation BibTex Tagged XML Download Paper 378 (2016).

- Stuart, Elizabeth A. “Matching Methods for Causal Inference: A Review and a Look Forward.” Statistical science : a review journal of the Institute of Mathematical Statistics 25.1 (2010): 1–21. PMC. Web. 29 Aug. 2017.

- Ho, Daniel E., et al. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." Political analysis 15.3 (2007): 199-236.

- Austin, Peter C. “A critical appraisa l of propensity-score matching in the medical literature between 1996 and 2003." Statistics in medicine 27.12 (2008): 2037-2049.

Discussion

Thank you!