Propensity Scores

Todd Wagner, PhD

January 25, 2023

Health Economics Resource Center

U.S. Department of Veterans Affairs
Veterans Health Administration
Health Services Research & Development Service

Learning Objectives

- We will:
  - Define a propensity score
  - Identify methods for implementing a propensity score
  - Highlight the assumptions needed to make causal claims with observational data

Outline

1. Background on assessing causation
2. Define propensity score (PS)
3. Calculate the PS
4. Use the PS
5. Limitations of the PS

Causality

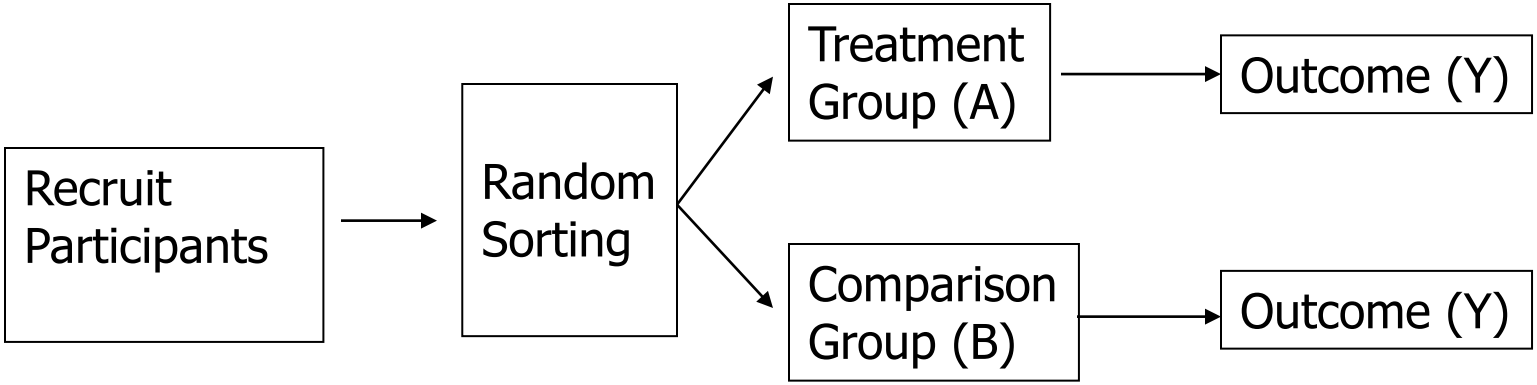
- Researchers are often interested in understanding causal relationships
  - Does treatment X reduce symptoms?
  - Does volume of work affect job burnout?
  - Does the Veterans Crisis Line reduce the likelihood of suicide?
  - Are there drugs that increase or decrease the risk of COVID-19?

Randomized Clinical Trial

- A RCT provides a methodological approach for understanding causation

- Understanding propensity score is assisted by understanding randomized trials.

Randomization



Note: random sorting can, by chance, lead to unbalanced groups. Most trials use checks and balances to preserve randomization

Just because a RCT can speak to causality, you must ask the question for whom– generalizability is often very limited

Trial analysis

- The expected effect of treatment is

  $E(Y)=E(Y^A)-E(Y^B)$

  Expected effect on group A minus expected effect on group B (i.e., mean difference).

Trial Analysis (II)

- $E(Y)=E(Y^A)-E(Y^B)$ can be analyzed using the following general model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Where

- y is the outcome
- α is the intercept
- x is the mean difference in the outcome between treatment A relative to treatment B
- ε is the error term
- i denotes the unit of analysis (person)

Trial Analysis (III)

- The model can be expanded to control for baseline characteristics (Z)

$$y_i = \alpha + \beta x_i + \delta Z_i + \varepsilon_i$$

Where

- y is outcome
- α is the intercept
- x is the added value of the treatment A relative to treatment B
- Z is a vector of baseline characteristics (predetermined prior to randomization)
- ε is the error term
- i denotes the unit of analysis (person)

Assumptions Needed for Causality

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- X, our right-hand side variable of interest, is measured without noise
  - Considered fixed in repeated samples
  - Noise, if it exists, is random, doesn't affect the mean, and biases towards the null

- There is no correlation between the X and the error term
  - In a RCT, this should happen by construction (coin flip) [$E(x_i\ \varepsilon_i)=0$]
  - Still must test balance of coin flip

- If these conditions hold, β on the treatment assignment is an unbiased estimate of the **causal** effect of X on the outcome

What if…

- The assumptions don’t hold in an RCT. Then what?

- You lose the unbiased estimate of causality.

Observational Studies

- Randomized trials may be
  - Unethical
  - Infeasible
  - Impractical
  - Not scientifically justified

- Observational data are limited by endogeneity

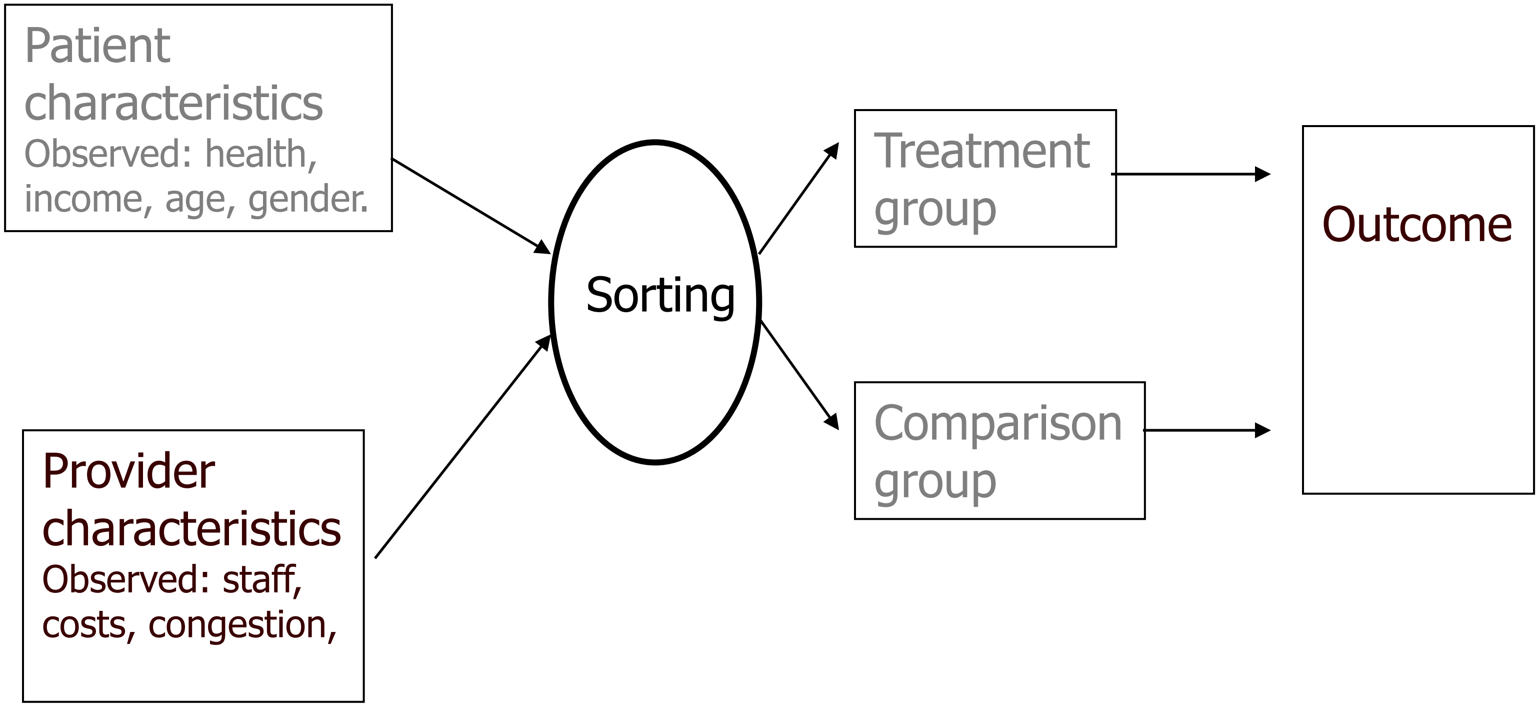
Endogenous

- Not attributable to any external factor.
- Example: Does smoking lead to cancer

$$cancer_i = \alpha + \beta smoking_i + \varepsilon_i$$

  - Smoking is correlated with income, education, parental exposure, etc.
  - We aren’t controlling for any of those factors, thus $E(smoking_i, \varepsilon_i) \neq 0$
  - Thus, smoking is endogenous

Sorting without randomization



Patient characteristics
Observed: health, income, age, gender.

Provider characteristics
Observed: staff, costs, congestion,

Sorting

Treatment group

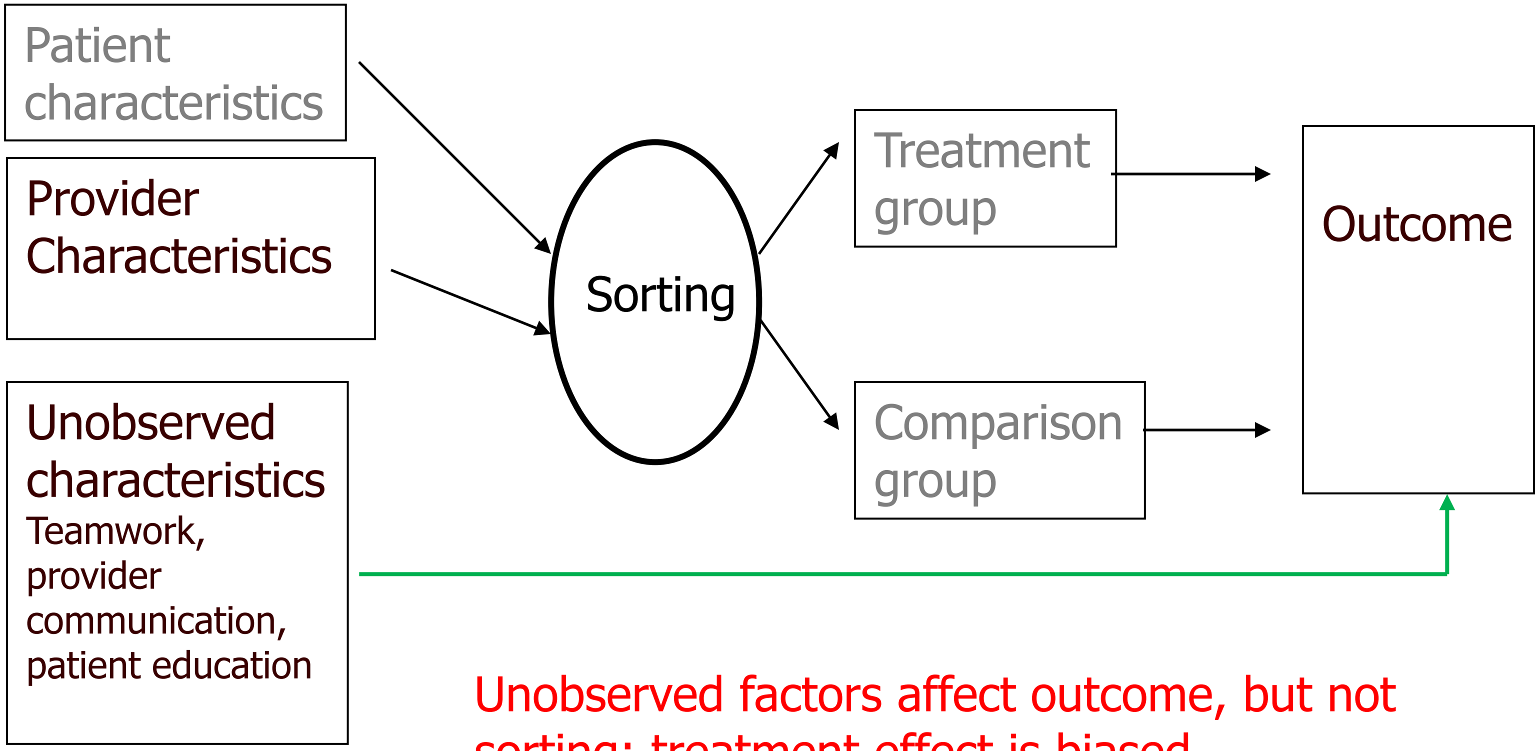
Comparison group

Outcome

If everything is fully observed and correctly specified; results are not biased. Never happens in reality.

Based on: Maciejewski and Pizer (2007) Propensity Scores and Selection Bias in Observational Studies. HERC Cyberseminar

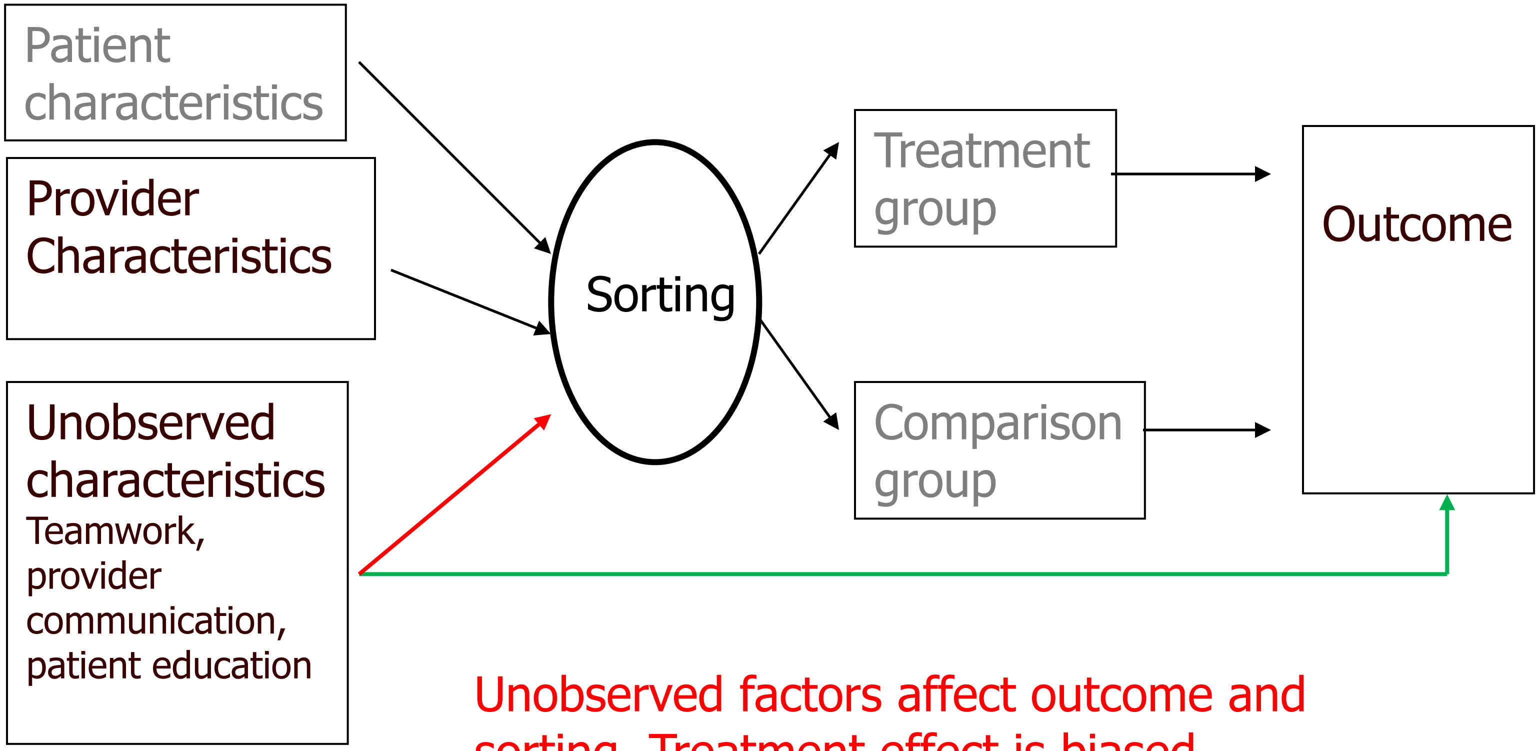
Sorting without randomization



Unobserved factors affect outcome, but not sorting; treatment effect is biased.

Fixed effects would be potential fix.

Sorting without randomization



Patient characteristics

Provider Characteristics

Unobserved characteristics
Teamwork, provider communication, patient education

Sorting

Treatment group

Comparison group

Outcome

Unobserved factors affect outcome and sorting. Treatment effect is biased. Causality isn't identified.

Example: Residential Treatment Programs

Fixed effect removes level effect. Still assumes exogeneity

FIGURE 1 Unadjusted Average Daily Costs for Inpatient Psychiatry (*N* = 141)

Note: RTP = rehabilitation treatment program.

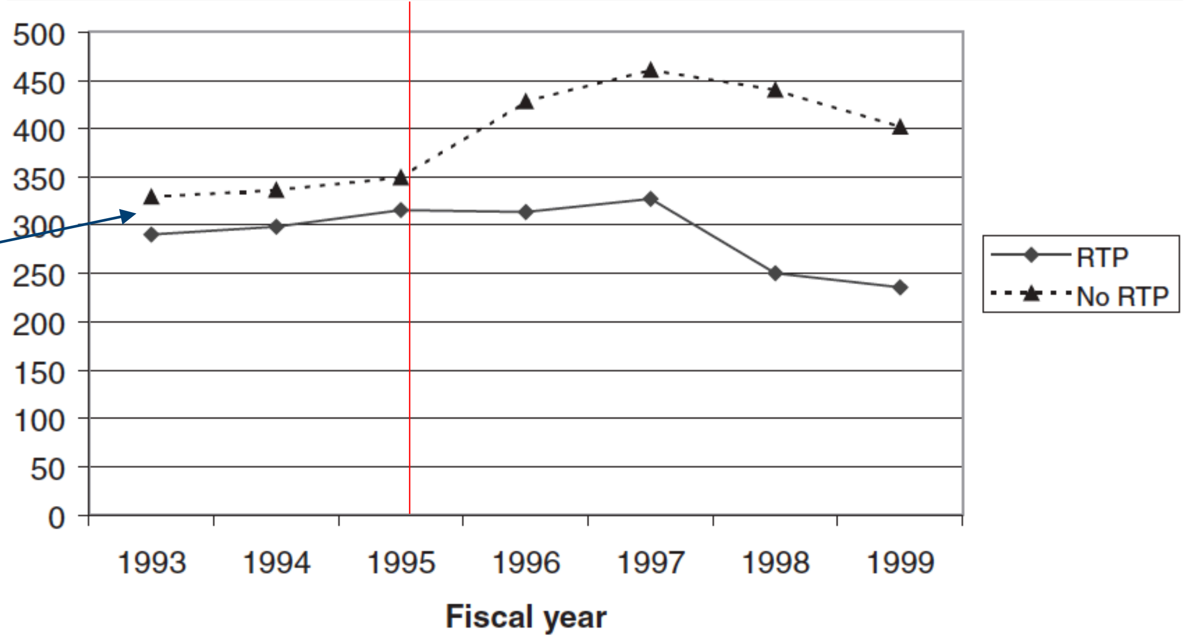


FIGURE 2 Unadjusted Average Daily Costs for Inpatient Substance Use (*N* = 134)

Note: RTP = rehabiliation treatment program.

Wagner TH, Chen S. An economic evaluation of inpatient residential treatment programs in the Department of Veterans Affairs. Medical care research and review. 2005 Apr;62(2):187-204.

Propensity Score Defined

- The PS uses observed information to calculate a single variable (the score)

- The score is the predicted propensity to get sorted into 1 of 2 groups (usually thought of as propensity to get treatment).

  Expected treatment effect: $E(Y)=E(Y^{A})-E(Y^{B})$

  Propensity Score is: $Pr(Y=A \mid X_{i})$

Propensity Scores

- What it is: Another way to correct for observable characteristics

- What it is not: A way to adjust for unobserved characteristics

- The only way to make causal claims is to make **huge** assumptions.

Strong Ignorability / Unconfounded

- To make statements about causation, you would need to assume that treatment assignment is strongly ignorable.
  - Similar to assumptions of missing at random
  - Equivalent to stating that all variables of interest are observed
- Growing interest in using propensity scores for prediction, which is a separate issue

Creating a Propensity Score

Calculating the Propensity Score

- You observe key covariate of interest
  $\text{cancer}_i = \alpha + \beta\text{smoking}_i + \varepsilon_i$

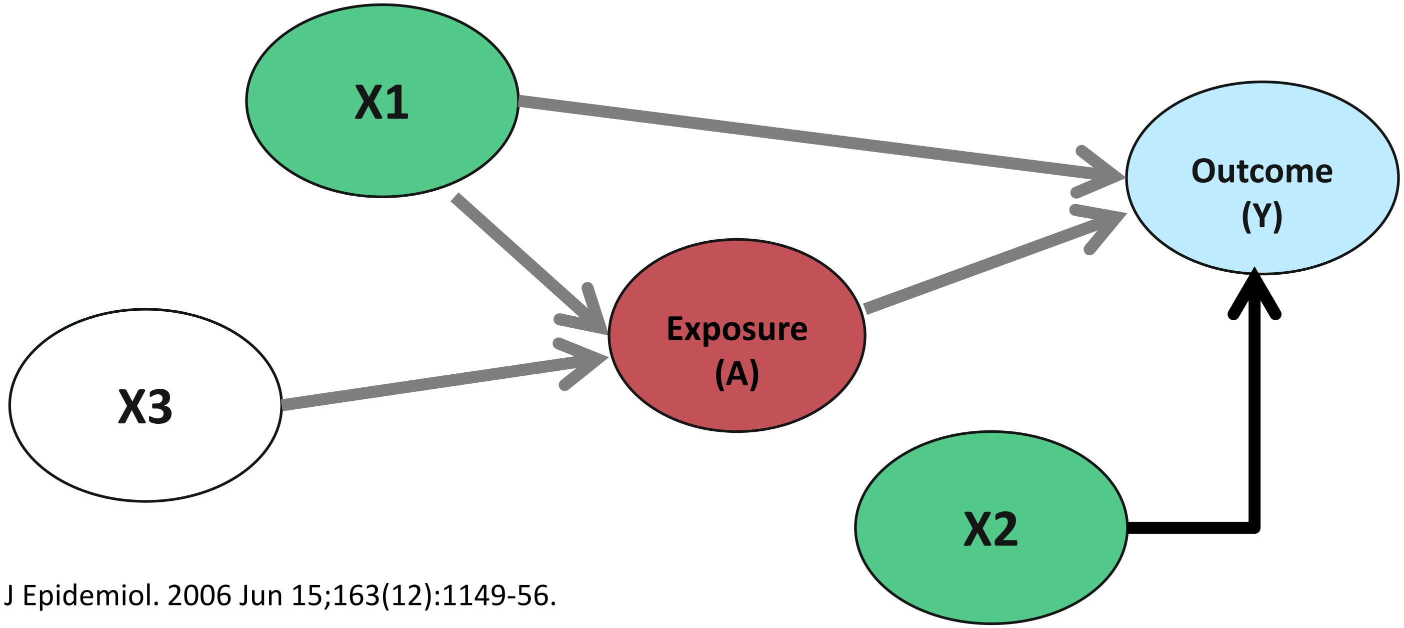
- Use multivariate logistic regression to estimate the probability that a person smoked

- The predicted probability from the logistic model is the propensity score

- PS models typically focus on sort into 2 groups; Melissa Garrido will be presenting later this year on 3-group PS models

Variables to Include

- Include variables that are related to the observed outcome
- This will decrease the variance of an estimated exposure effect without increasing bias
- Do not include variables affect only correlated with exposure



Brookhart MA, et al Am J Epidemiol. 2006 Jun 15;163(12):1149-56.

Variables to Exclude

- Exclude variables that are related to the exposure but not to the outcome
- These variables will increase the variance of the estimated exposure effect without decreasing bias
- Variable selection is particularly important in small studies (n<500)

Brookhart MA, et al Am J Epidemiol. 2006 Jun 15;163(12):1149-56.

Consider the Functional Form

- Age
  - Dummies (<45, 45-64, 65-74, >=75)
  - Linear (age)
  - Non-linear (age^2 or age^3)
- In regression, it is often recommended to **demean/ center** covariates so that the covariates have mean 0.
  - This makes it easier to interpret the intercept term
  - Age
  - Calendar year
- The functional form matters
  - Dummies create discontinuities in risk
  - Linear may not be accurate
  - Demeaned cubic polynomial

Example: Resident Surgery

- Are patient outcomes different when the surgery is conducted by a resident or an attending?

- We had a dataset that tracked the primary surgeon for heart bypass

Uses

- Understanding sorting and balance
  - Sorting is multidimensional
  - The PS provides a simple way of reducing this dimensionality to understand the similarity of the treatment groups
- Adjusting for covariance

Example

- Are surgical outcomes worse when the surgeon is a resident?

- Resident assignment may depend on
  - Patient risk
  - Availability of resident
  - Resident skill
  - Local culture

Resident Assignment

	OR	P value
Age	1.00	0.79
Canadian Functional Class		
Class 2	1.93	0.15
Class 3	2.12	0.09
Class 4	4.25	0.02
Urgent priority	0.93	0.89
Artery condition at site		
Calcified	0.67	0.25
Sclerotic	2.63	0.00
site 2	62.89	0.00
site 3	0.67	0.60
site 5	138.16	0.00
site 7	11.66	0.00
site 8	19.85	0.00
site 9	1.76	0.43
endo vascular harvest	0.20	0.01
On pump surgery	1.20	0.75
1-2 grafts	1.70	0.16
4-5 grafts	0.79	0.46

Assignment not associated with age or number of grafts

Assignment associated with angina symptoms and planned harvesting technique

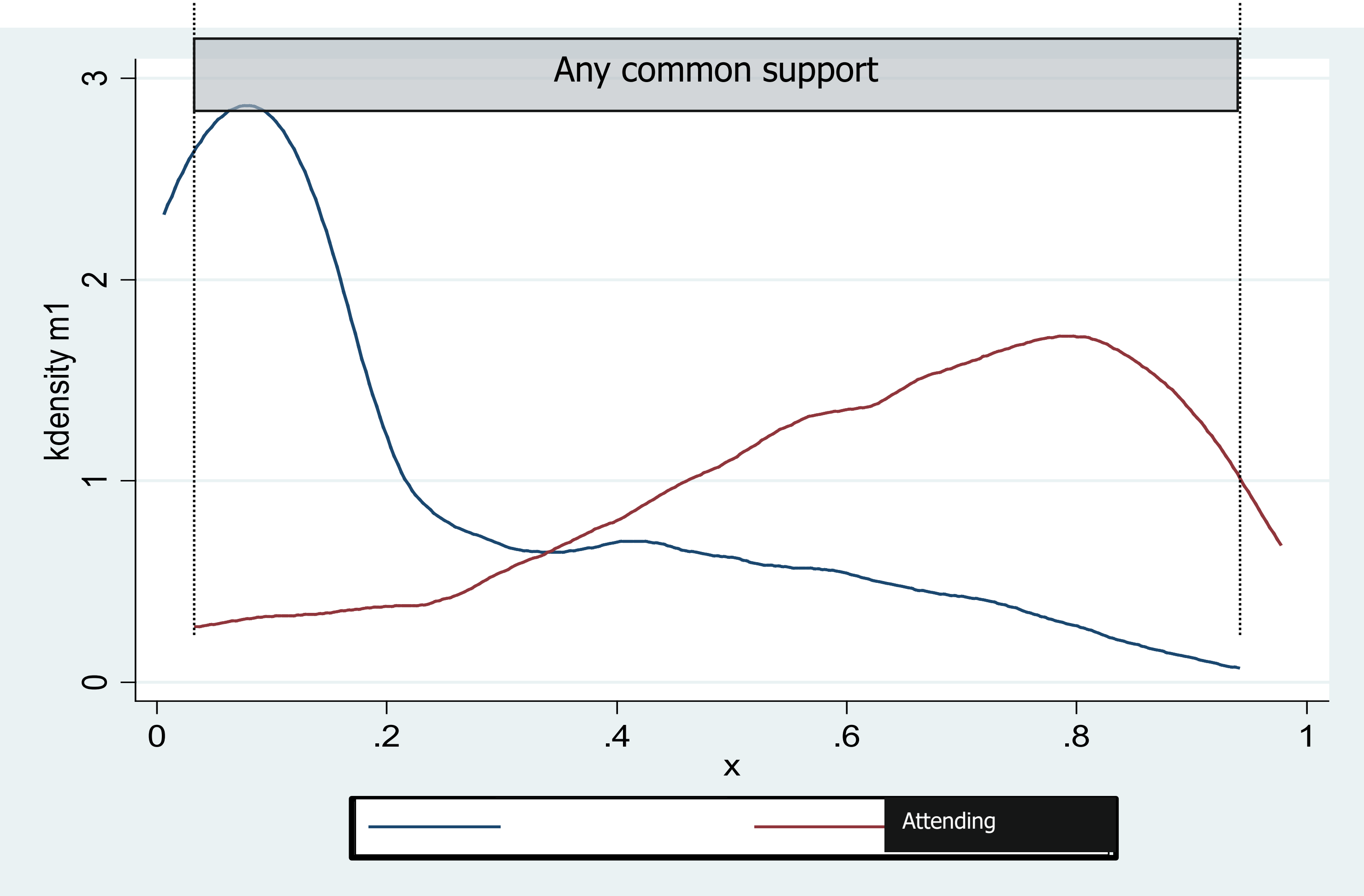
Bakaeen F et al. Coronary Artery Bypass Graft Patency: Residents Versus Attending Surgeons. *Annals of Thoracic Surgery.*

Shared / Common Support

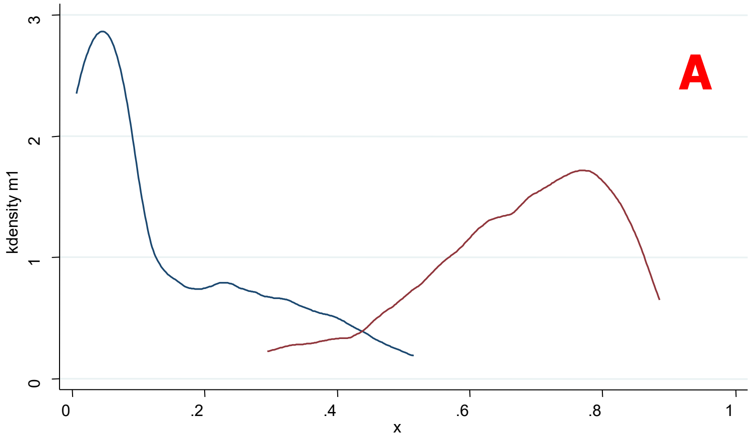
- Measures the similarity of people in both treatments
- Conditional on covariates, there exist people who choose both treatments.

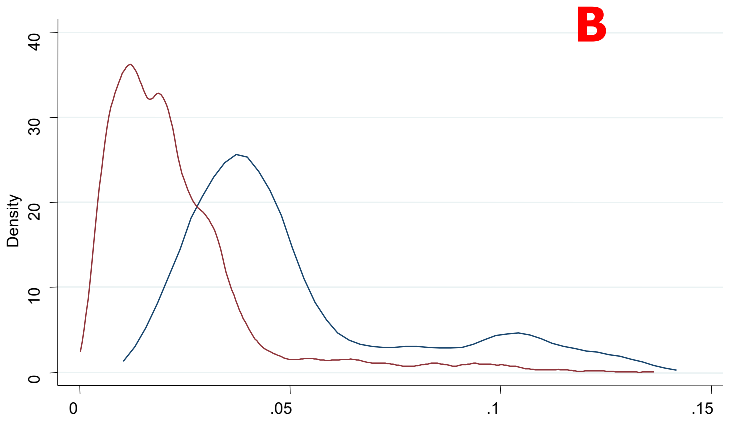
- Examining shared support offers insights not in multivariate models ✓

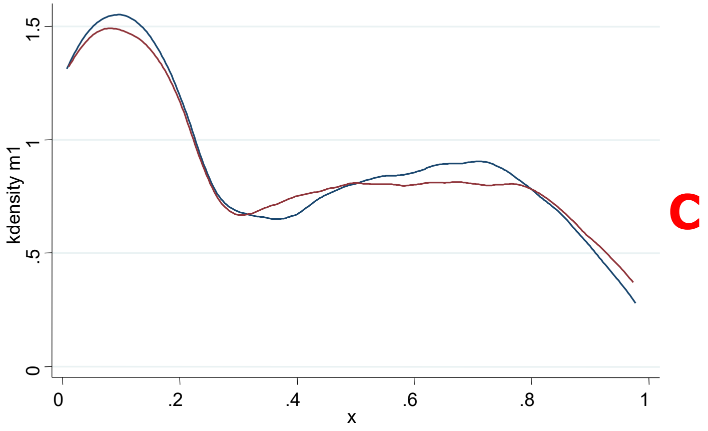
Propensity Score for Resident vs Attending Surgeon



Compare Three Diagrams



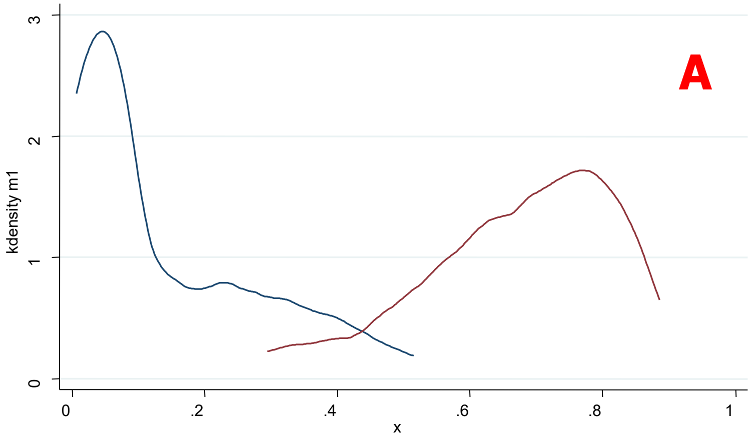


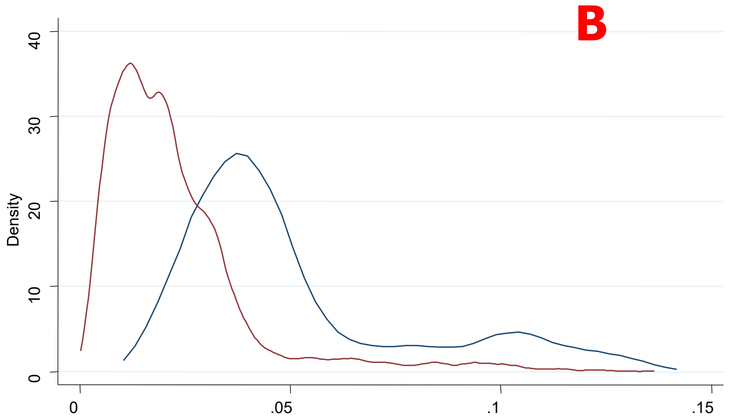


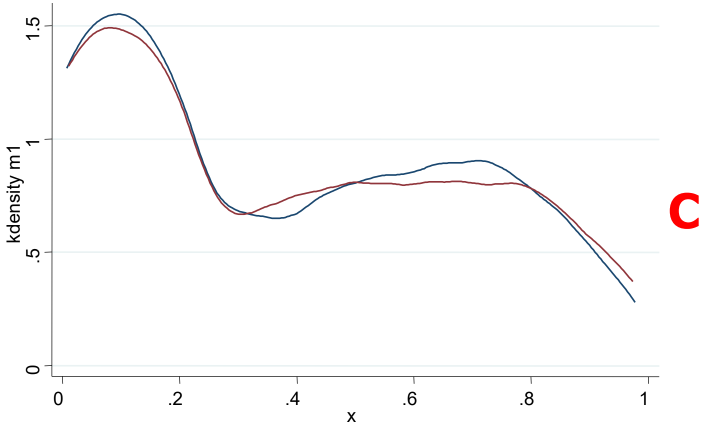
Poll

- Which graph is the most concerning? Choose one
  - A
  - B
  - C
  - All of them
  - None of them

Three Scores



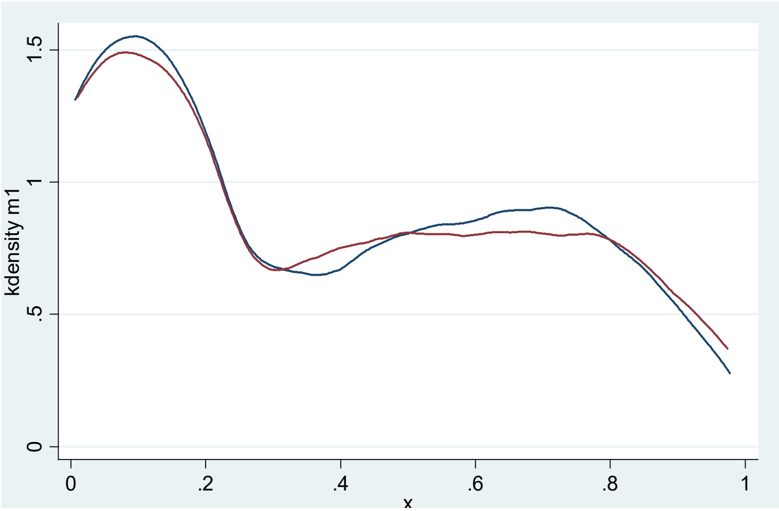


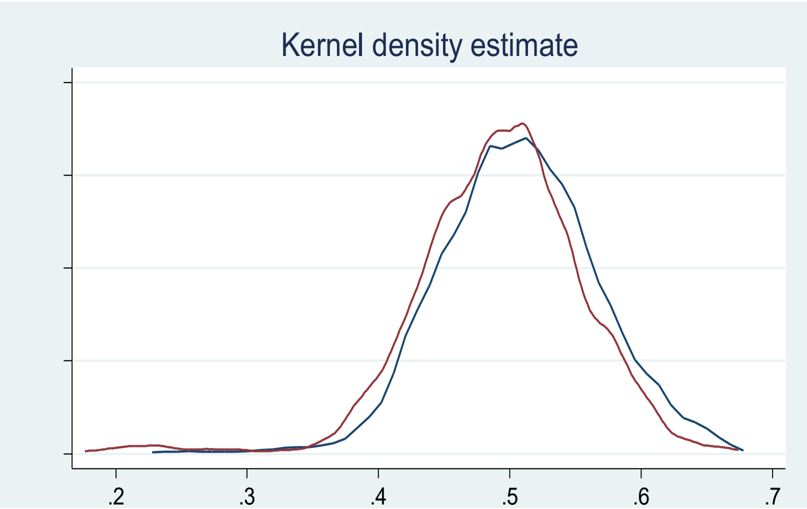


RCTs and Propensity Scores

- What would happen if you used a propensity score with data from a RCT?

Shared Common Support

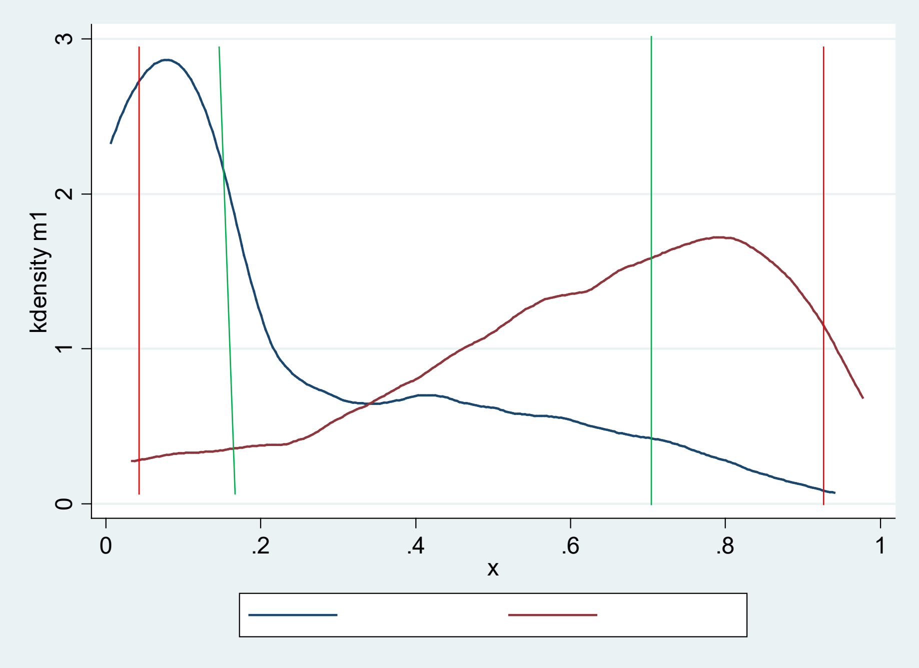




Don’t worry about the shape. Focus on the overlap

Common Support

- Understanding the shared support is critical
  - What do you do with observations that don't share support?
  - Where do you draw the line?
  - Trimming is arbitrary; extreme weighting is one possible solution.[1]



[1] Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. American journal of epidemiology. 2019 Jan 1;188(1):250-7.

Using the Propensity Score

Using the Propensity Score

1. Compare individuals based on similar PS scores (a matched analysis)
2. Conduct subgroup analyses on similar groups (stratification)
3. Include it as a covariate (quintiles of the PS) in the regression model
4. Use it to weight the regression (i.e., place more weight on similar cases)
5. Use both 3 and 4 together (doubly robust)

PS as a Covariate

- There seems to be little advantage to using PS over multivariate analyses in most cases.[1]
- PS provides flexibility in the functional form
- Propensity scores may be preferable if the sample size is small and the outcome of interest is rare.[2]

[1] Winkelmeyer. Nephrol. Dial. Transplant 2004; 19(7): 1671-1673.

[2] Cepeda et al. Am J Epidemiol 2003; 158: 280–287

Matched Analyses

- The idea is to select controls that resemble the treatment group in all dimensions, except for treatment
- You can exclude cases and controls that don’t match, which can reduce the sample size/power.
- Different matching methods

Matching Methods

- Nearest Neighbor: rank the propensity score and choose control that is closest to case.
- Caliper: choose your common support and from within randomly draw controls

- Choice of matching estimator important

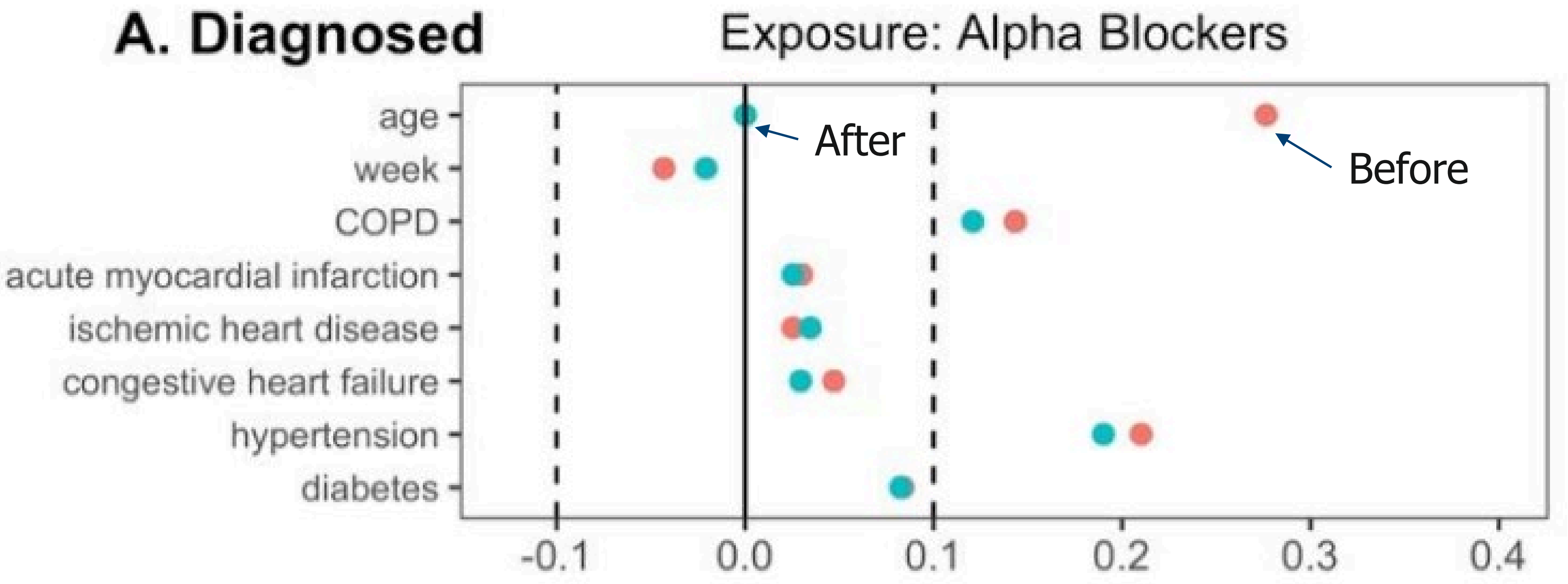
Next Step

- Choose your method
- Graph the overlap
- Compare the balance (Love plots)
  - Standardized difference of less than 10% is a common rule of thumb

Attributed to Thomas E Love (Heart failure, chronic diuretic use, and increase in mortality and hospitalization. European Heart Journal. 2006).

Love Plots

The Association Between Alpha-1 Adrenergic Receptor Antagonists and In-Hospital Mortality from COVID-19



After

Before

https://www.medrxiv.org/content/10.1101/2020.12.18.20248346v1

Recent Areas of Research

- Economics: choice of matching estimators
  - Busso M et al. New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *Review of Economics and Statistics*, 96.5 (2014): 885-897
  - Athey S, Imbens GW. The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives. 2017 May;31(2):3-2.
- Political Science
  - King G, Nielsen R. Why propensity scores should not be used for matching. Copy at http://j.mp/1sexgVw. 2016 Dec 16;378.
- Biostatistics: high dimensional propensity scores using big data
  - Schneeweiss, Sebastian, et al. "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." *Epidemiology* 20.4 (2009): 512.

Limitations

Do the Unobservables Matter?

- Propensity scores focus only on observed characteristics, not on unobserved.
- Improbable that we fully observe the sorting process
  - Thus $E(x_i \varepsilon_i) \neq 0$
  - Multivariate (including propensity score) is biased and we need another method, such as instrumental variables, fixed effects or RCT

Does Using PS Exacerbate Imbalance of Unobservables

- PS is based on observables.
- Brooks and Ohsfeldt, using simulated data, showed that PS models can create **greater** imbalance among unobserved variables.
- King G, Nielsen R. Why propensity scores should not be used for matching. https://dspace.mit.edu/handle/1721.1/128459

Brooks and Ohsfeldt (2013): Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health Services Research*.

Summary

A Propensity Score:

- Offers another way to adjust for confound by observables
- Reduce the multidimensional nature of confounding can be helpful
- Has many forms. There are many ways to implement propensity scores and a growing interest in matching estimators

Strengths

- Allow one to check for balance between control and treatment

- Without balance, average treatment effects can be very sensitive to the choice of the estimators.[1]

[1] Imbens and Wooldridge 2007 http://www.nber.org/WNE/lect_1_match_fig.pdf

Challenges

- Propensity scores are often misunderstood
- Not enough attention is placed on the PS model, itself
- Not enough attention is placed on robustness checks
- While a PS can help create balance on observables, PS models do not control for unobservables or selection bias

Further Reading

- Rosenbaum, P. R., D. B. Rubin. “The central role of the propensity score in observational studies for causal effects.” *Biometrika 70* (1983)*: 41–55*
- Imbens and Wooldridge (2007) www.nber.org/WNE/lect_1_match_fig.pdf
- Imbens, Guido W. "The role of the propensity score in estimating dose-response functions." *Biometrika* 87.3 (2000): 706-710.
- Imbens, Guido W. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and Statistics* 86.1 (2004): 4-29.
- Guo and Fraser (2010) Propensity Score Analysis. Sage.
- King G, Nielsen R. Why propensity scores should not be used for matching. Copy at http://j.mp/1sexgVw. 2016 Dec 16;378.
- Brooks, John M., and Robert L. Ohsfeldt. "Squeezing the balloon: propensity scores and unmeasured covariate balance." *Health Services Research* 48.4 (2013): 1487-1507.
- Garrido, Melissa M., et al. "Methods for constructing and assessing propensity scores." *Health Services Research* 49.5 (2014): 1701-1720.
- Busso M et al. “New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators.” *Review of Economics and Statistics*, 96.5 (2014): 885-897
- Imai, Kosuke, and Marc Ratkovic. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*76.1 (2014): 243-263.
- Reiffel JA. Propensity score matching: The ‘Devil is in the details’ where more may be hidden than you know. The American journal of medicine. 2020 Feb 1;133(2):178-81.

Questions?

- HERC@VA.gov

- @herc_va
- @toddhwagner

- **Next class: Instrumental Variables**

  Kritee Gijral, Ph.D. Feb 1.