

Propensity Scores

Todd Wagner, PhD

February 2011

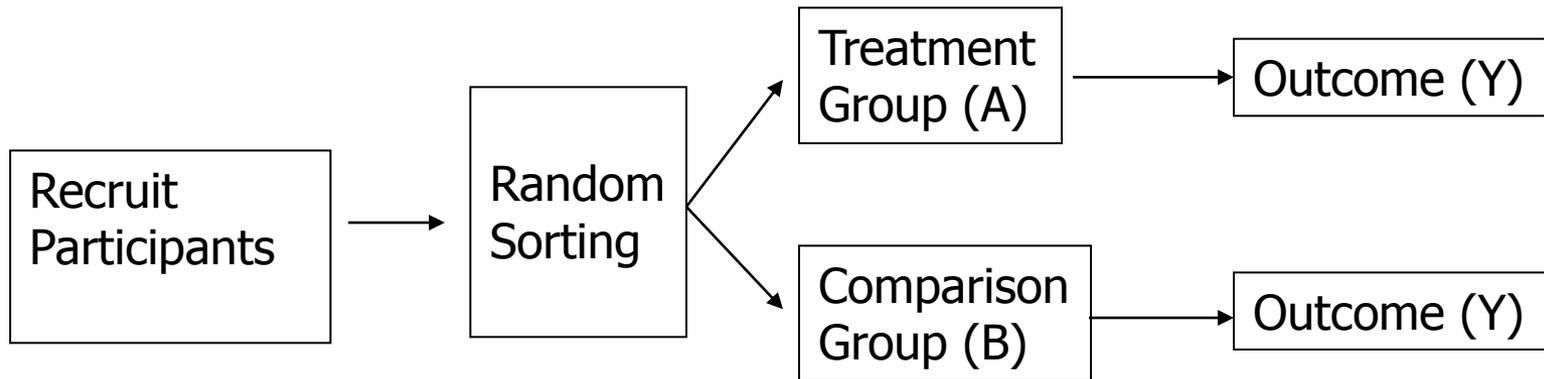
Outline

1. Background on assessing causation
 - Randomized trials
 - Observational studies
 2. Mechanics of calculating a propensity score
 3. Limitations
-

Causality

- Researchers are often interested in understanding causal relationships
 - Does drinking red wine affect health?
 - Does a new treatment improve mortality?
- Randomized trial provides a venue for understanding causation

Randomization



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

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 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

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

Where

- y is cognitive function
- 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)

Causation

- What two factors enable researchers to make statements about causation?

White board exercise

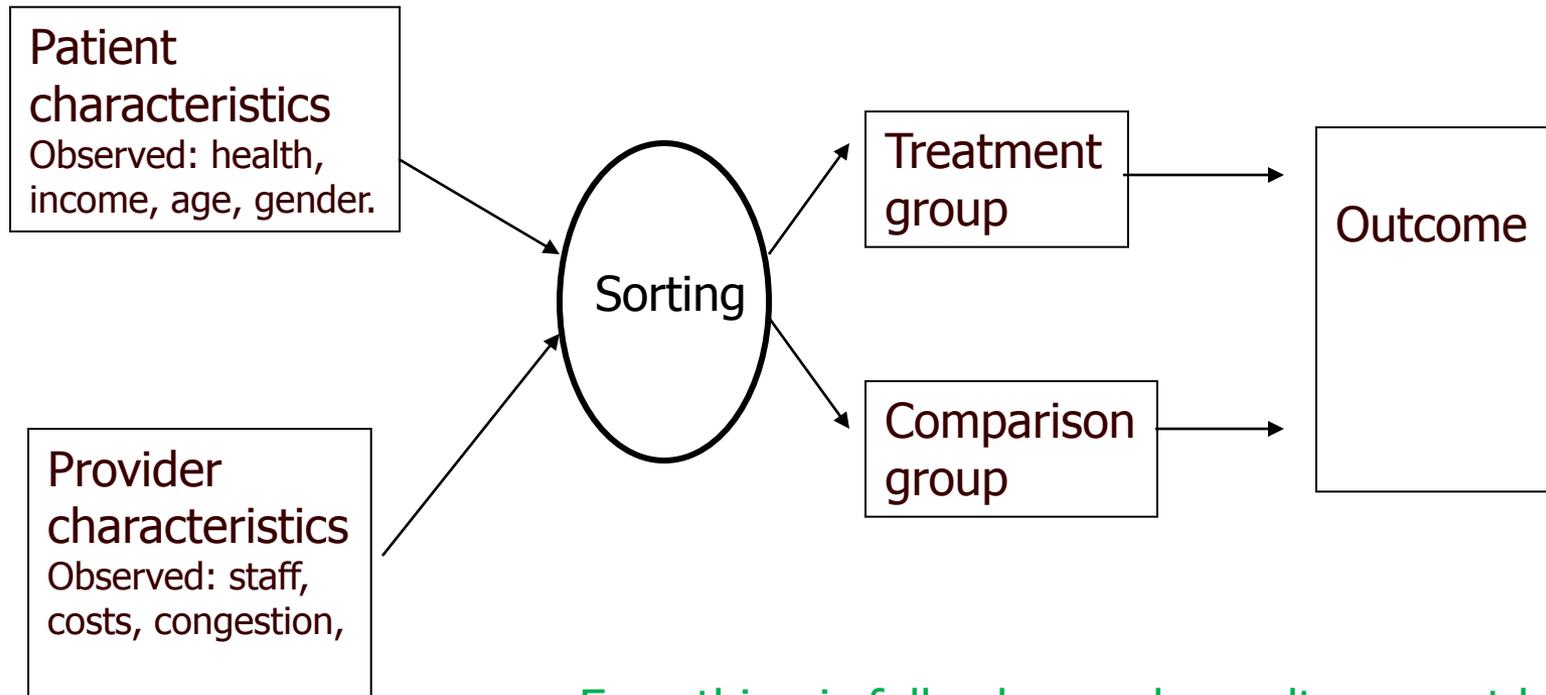
Assumptions

- Classic linear model (CLR) assumes that
 - Right hand side variables are measured without noise (i.e., considered fixed in repeated samples)
 - There is no correlation between the right hand side variables and the error term $E(x_i u_i) = 0$
- If these conditions hold, β is an unbiased estimate of the causal effect of the treatment on the outcome

Observational Studies

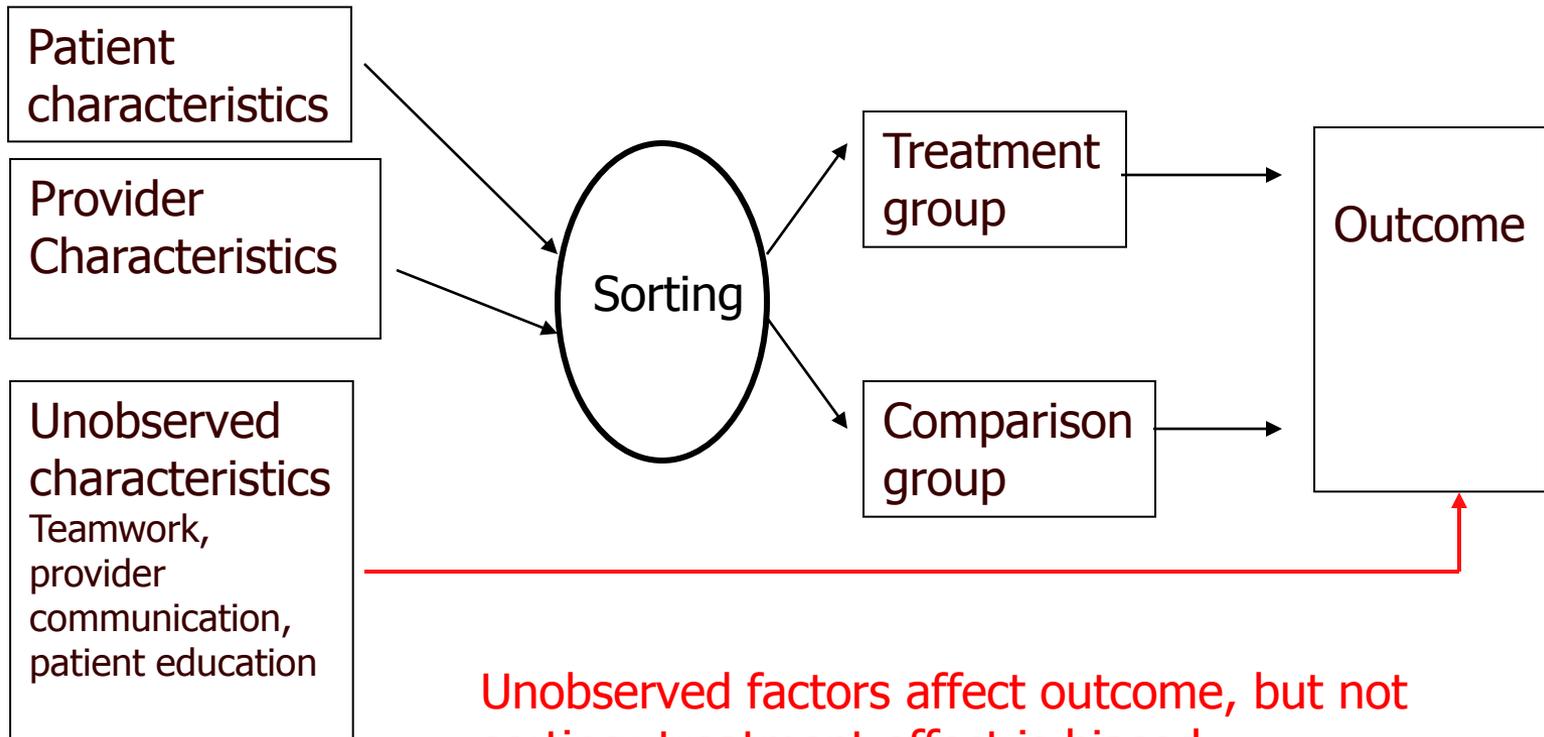
- Randomized trials may be
 - Unethical
 - Infeasible
 - Impractical
 - Not scientifically justified

Sorting without randomization



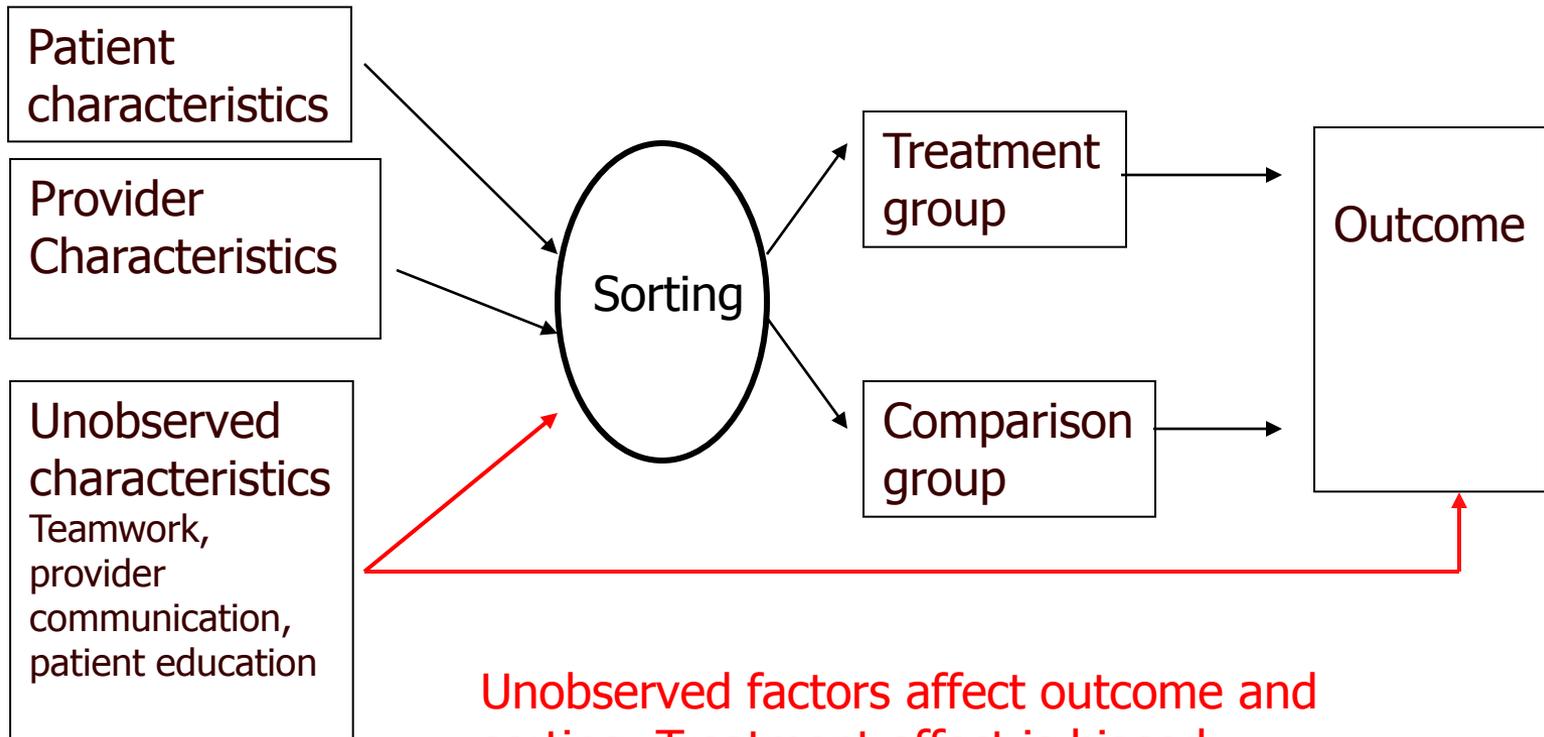
Everything is fully observed; results are not biased.
Never happens in reality.

Sorting without randomization



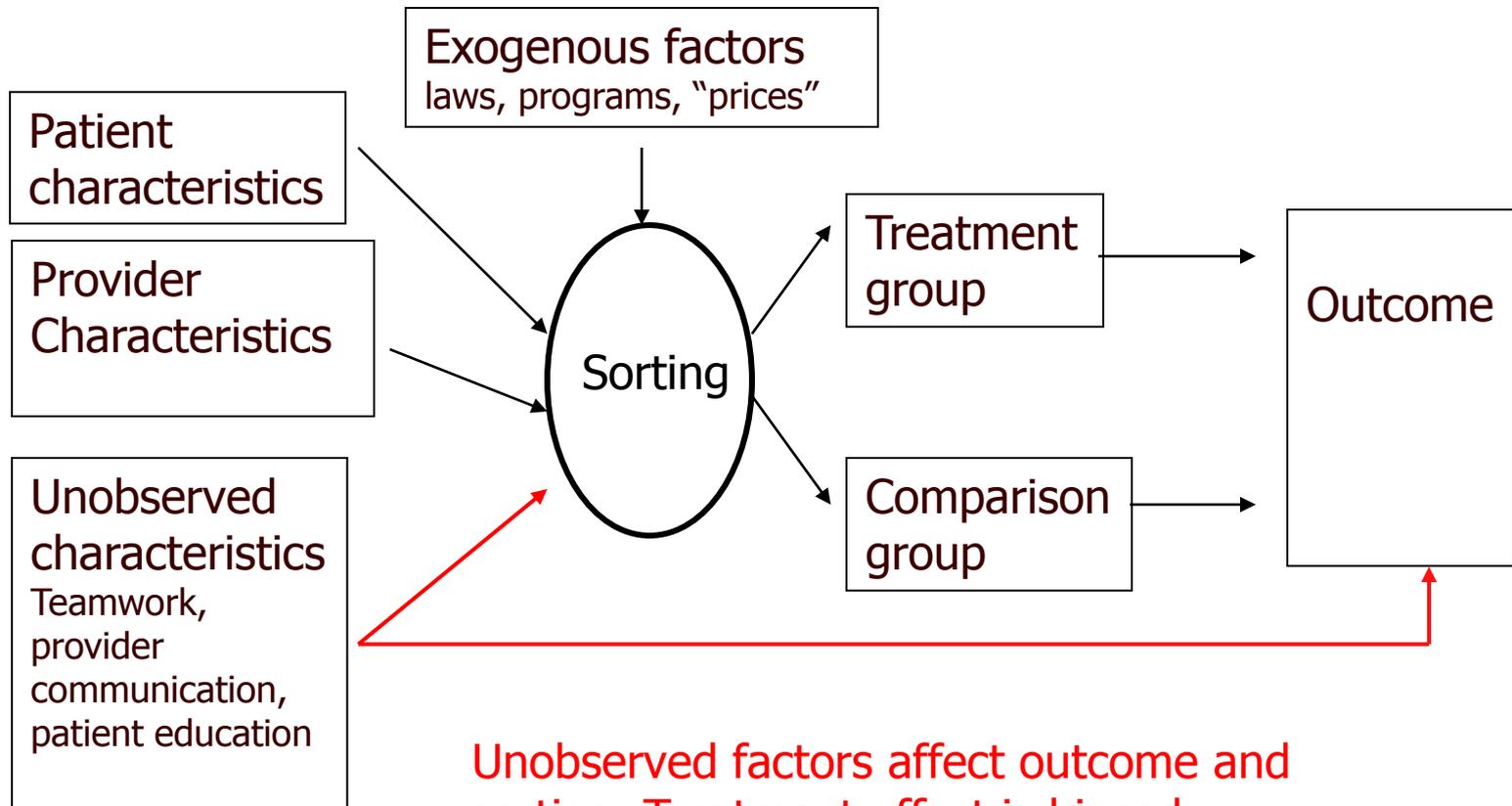
Unobserved factors affect outcome, but not sorting; treatment effect is biased.
Fixed effects would be potential fix.

Sorting without randomization



Unobserved factors affect outcome and sorting. Treatment effect is biased. Provides little or no information on causality
No fix.

Sorting without randomization



Unobserved factors affect outcome and sorting. Treatment effect is biased.
Instrumental variables is potential fix.

Propensity Scores

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

Strong Ignorability

- Propensity scores were not developed to handle non-random sorting
- To make statements about causation, you would need to make an assumption that treatment assignment is strongly ignorable.
 - Similar to assumptions of missing at random
 - Equivalent to stating that all variable of interest are observed

Calculating the Propensity

- One group receives treatment and another group doesn't.
- Use a logistic regression model to estimate the probability that a person received treatment.
- This probability is the propensity score.

Dimensionality

- The treatment and non-treatment groups may be different on many dimensions
- The propensity score reduces these to a single dimension

Using the Propensity Score

- Match individuals (perhaps most common approach)
- Include it as a covariate (quintiles of the PS) in the regression model
- Include it as a weight in a regression (i.e., place more weight on similar cases)
- Conduct subgroup analyses on similar groups (stratification)

Matched Analyses

- The idea is to select a control group to make them resemble the treatment group in all dimensions, except for treatment
 - Different metrics for choosing a match
 - Nearest neighbor, caliper
 - You can exclude cases and controls that don't match. If the groups are very different, this can reduce the sample size/power.
-

White board

- What would happen if you took a randomized trial and reran it with a propensity score?

Example

- CSP 474 was a randomized trial that enrolled patients in 11 sites
- Patients were randomized to two types of heart bypass
- Is the sample generalizable?
 - We compared enrollees to non-enrollees.

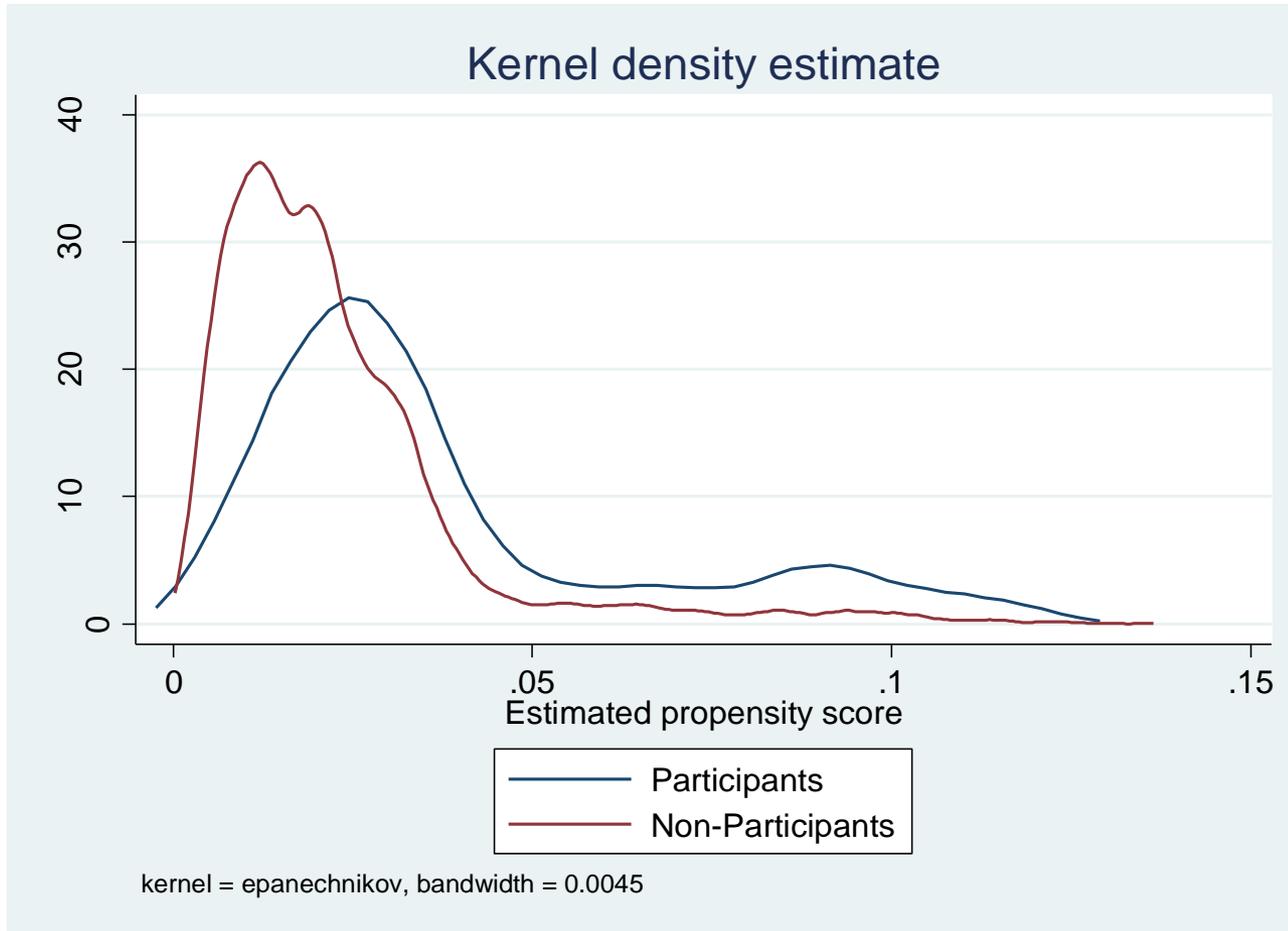
Methods

- We identified eligible bypass patients across VA (2003-2008)
 - We compared:
 - participants and nonparticipants within participating sites
 - participating sites and non-participating sites
 - participants and all non-participants
-

Propensity Scores

- A reviewer suggested that we should use a propensity score to identify degree of overlap
- Estimated a logistic regression for participation (pscore and pstest command in Stata)

Group Comparison before PS



Variable	Sample	Mean		%bias	%reduct bias	t-test	
		Treated	Control			t	p>t
ms_1	Unmatched	.09729	.10659	-3.1		-0.75	0.455
	Matched	.09729	.0986	-0.4	85.9	-0.22	0.827
ms_3	Unmatched	.35407	.36275	-1.8		-0.45	0.655
	Matched	.35407	.35769	-0.8	58.3	-0.37	0.710
male	Unmatched	.99043	.99069	-0.3		-0.07	0.946
	Matched	.99043	.99049	-0.1	76.6	-0.03	0.975
aa2	Unmatched	.12919	.09003	12.6		3.37	0.001
	Matched	.12919	.11989	3.0	76.3	1.36	0.173
aa3	Unmatched	.27113	.22301	11.2		2.86	0.004
	Matched	.27113	.26578	1.2	88.9	0.59	0.554
aa4	Unmatched	.27751	.22921	11.1		2.84	0.005
	Matched	.27751	.26658	2.5	77.4	1.20	0.230
aa5	Unmatched	.10367	.1388	-10.8		-2.52	0.012
	Matched	.10367	.11048	-2.1	80.6	-1.10	0.272
aa6	Unmatched	.09569	.13058	-11.0		-2.57	0.010
	Matched	.09569	.10471	-2.8	74.2	-1.51	0.132
aa7	Unmatched	.05104	.10121	-19.0		-4.14	0.000
	Matched	.05104	.05918	-3.1	83.8	-1.82	0.069
aa8	Unmatched	.01754	.05057	-18.3		-3.76	0.000
	Matched	.01754	.0204	-1.6	91.4	-1.07	0.285

Only partial
listing
shown

Standardized difference >10% indicated imbalance and >20% severe imbalance

Summary of the distribution of the abs(bias)

BEFORE MATCHING

	Percentiles	Smallest		
1%	.0995122	.0995122		
5%	.2723117	.2723117		
10%	1.809271	1.061849	Obs	38
25%	3.781491	1.809271	Sum of Wgt.	38
50%	10.78253		Mean	10.59569
		Largest	Std. Dev.	9.032606
75%	15.58392	18.99818		
90%	18.99818	19.16975	Variance	81.58797
95%	29.75125	29.75125	Skewness	1.848105
99%	46.80021	46.80021	Kurtosis	8.090743

AFTER MATCHING

	Percentiles	Smallest		
1%	.0321066	.0321066		
5%	.0638531	.0638531		
10%	.4347224	.332049	Obs	38
25%	.7044271	.4347224	Sum of Wgt.	38
50%	1.156818		Mean	1.416819
		Largest	Std. Dev.	1.215813
75%	1.743236	2.848478		
90%	2.848478	2.97902	Variance	1.4782
95%	3.083525	3.083525	Skewness	2.524339
99%	6.859031	6.859031	Kurtosis	11.61461

Results

- Participants tending to be slightly healthier and younger, but
- Sites that enrolled participants were different in provider and patient characteristics than non-participating site

PS Results

- 38 covariates in the PS model
 - 20 variables showed an imbalance
 - 1 showed severe imbalance (quantity of CABG operations performed at site)
 - Balance could be achieved using the propensity score
 - After matching, participants and controls were similar
-

Generalizability

- To create generalizable estimates from the RCT, you can weight the analysis with the propensity score.

Li F, Zaslavsky A, Landrum M. Propensity score analysis with hierarchical data. Boston MA: Harvard University; 2007.

Weaknesses

- Propensity scores are often misunderstood
- While they can help create balance on observables, they do not control for unobservables or selection bias

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

PS or Multivariate Regression?

- 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. Winkelmeier. Nephrol. Dial. Transplant 2004; 19(7): 1671-1673.
2. Cepeda et al. Am J Epidemiol 2003; 158: 280–287

Further Reading

- Imbens and Wooldridge (2007)
www.nber.org/WNE/lect_1_match_fig.pdf
- Guo and Fraser (2010) Propensity Score Analysis. Sage.