Propensity Scores

Todd Wagner, PhD

July 2012

HERC
Health Economics Resource Center

HSR&D
HEALTH SERVICES RESEARCH & DEVELOPMENT

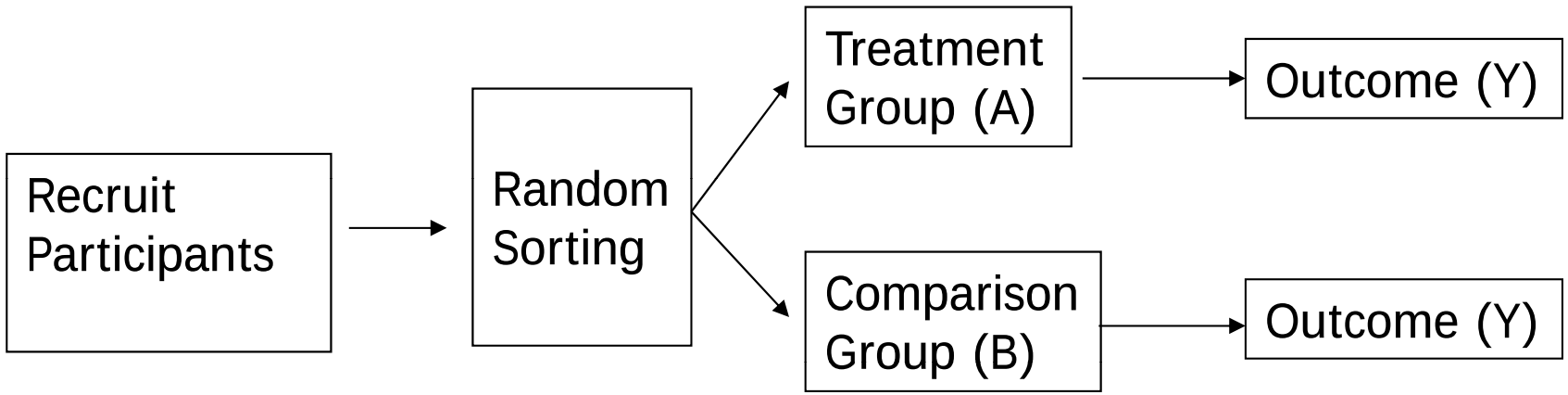
Outline

1. Background on assessing causation
   - Randomized trials
   - Observational studies
2. 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 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)

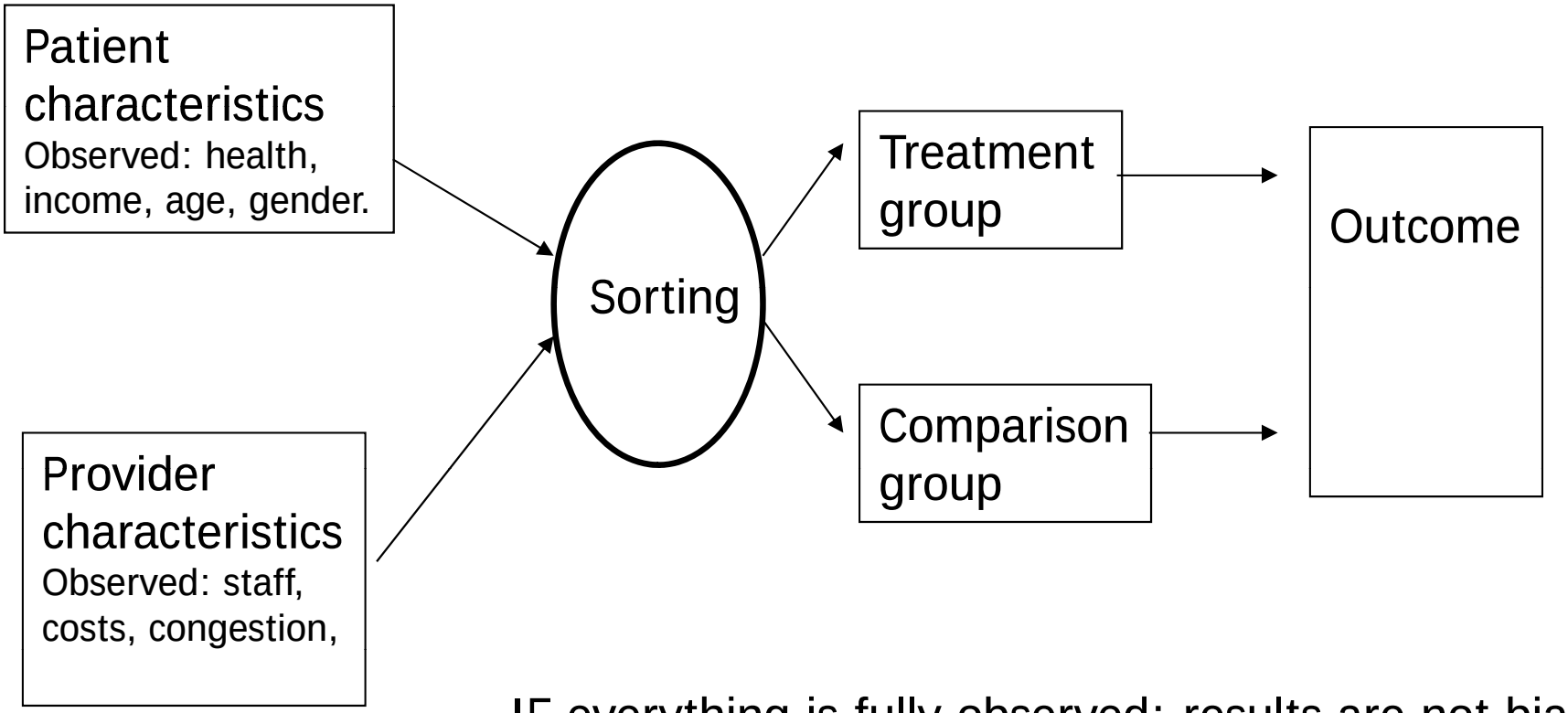
Assumptions

- 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

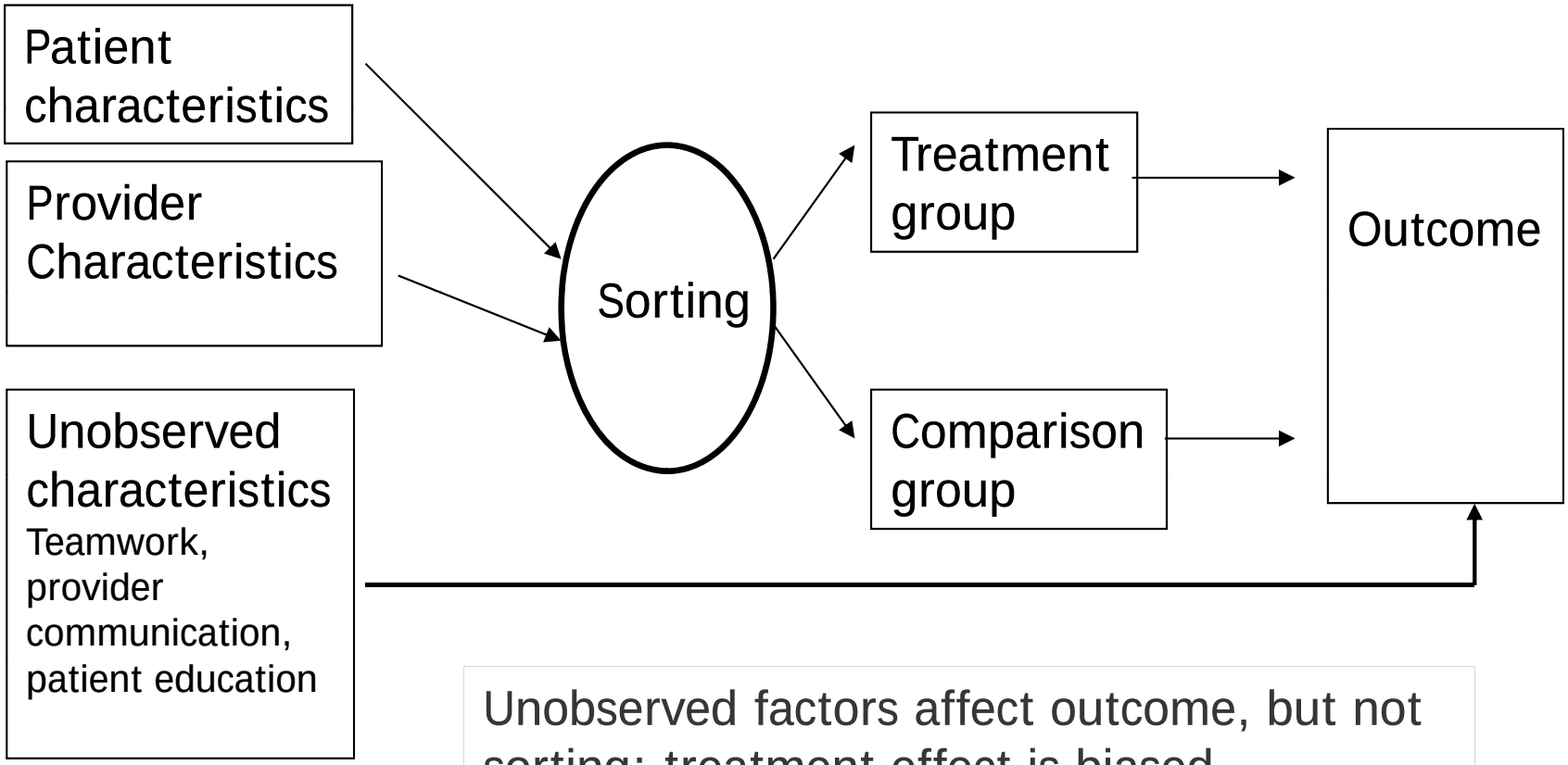
- Randomized trials may be
  - Unethical
  - Infeasible
  - Impractical
  - Not scientifically justified

Sorting without randomization



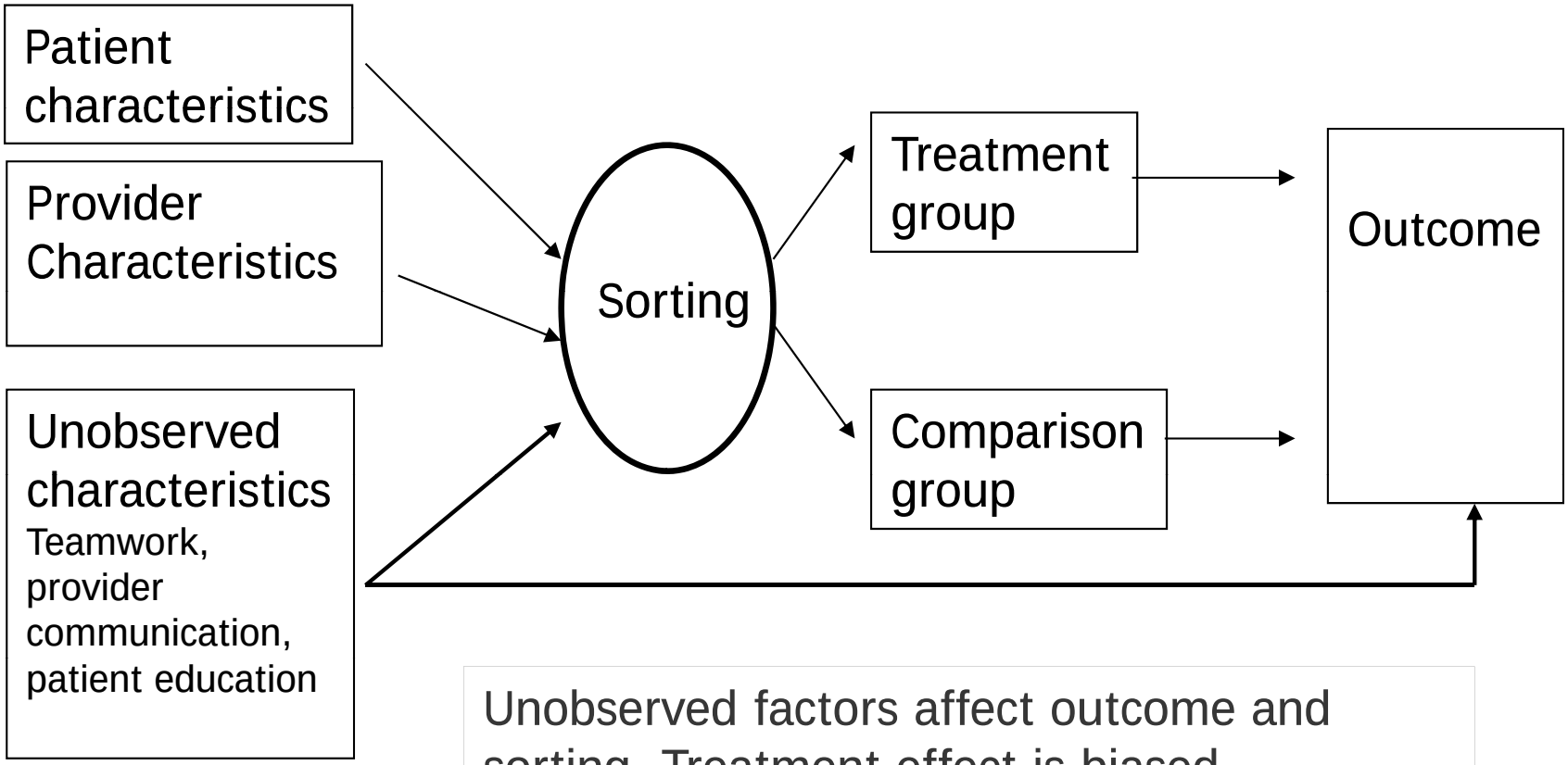
IF everything is fully observed; results are not biased.
Never happens in reality.

Sorting without randomization



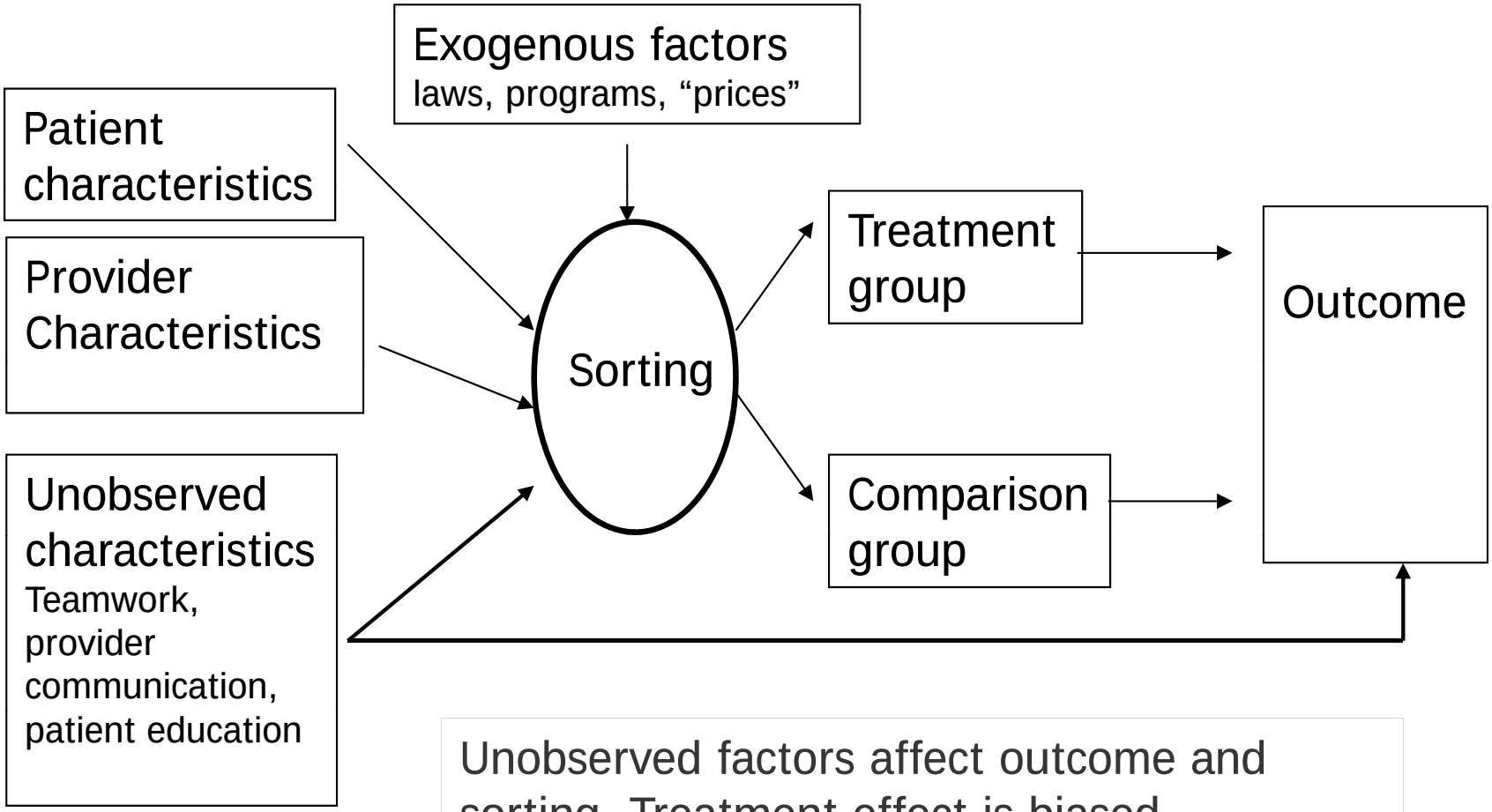
Unobserved factors affect outcome, but not sorting; treatment effect is biased.
Fixed effects would be potential fix.

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
- If you read wikipedia, you will get the wrong impression about propensity scores

Strong Ignorability

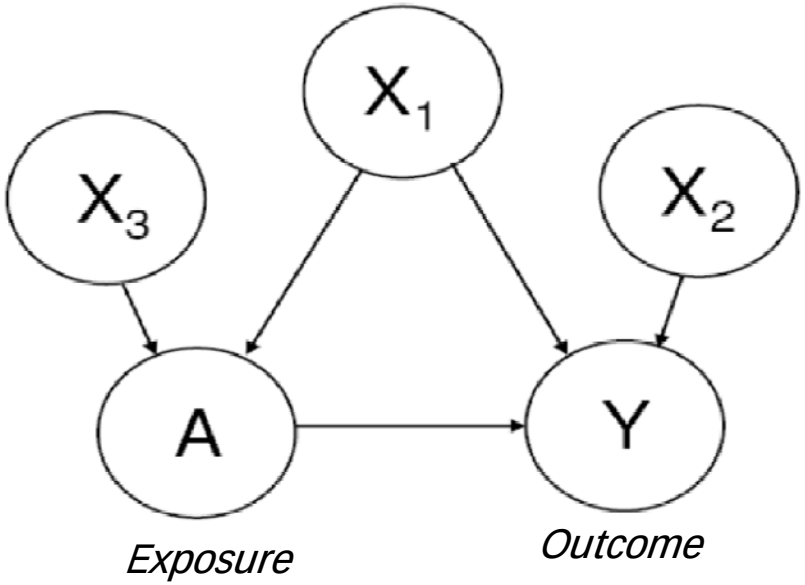
- Propensity scores were not developed to handle non-random sorting
- To make statements about causation, you would need to make an assumption that treatment assignment is strongly ignorable.
  - Similar to assumptions of missing at random
  - Equivalent to stating that all variables of interest are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

Calculating the Propensity Score

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

Variables to Include

- Include variables that are unrelated to the exposure but related to the outcome
- This will decrease the variance of an estimated exposure effect without increasing bias



Brookhart MA, et al Am J Epidemiol. 2006 Jun 15;163(12):1149-56. Variable selection for propensity score models.

Variables to Exclude

- Exclude variables that are related to the exposure but not to the outcome
- These variables will increase the variance

  decreasıng 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. Variable selection for propensity score models.

Example: Resident Surgery

- Do cardiac bypass patients have better / worse outcomes when their surgery is conducted by a resident?
- CSP 474
  - Randomized patients to radial artery or saphenous vein
  - Tracked primary surgeon

Is Resident Assignment Random?

- Assignment may depend on
  - Patient risk
  - Availability of resident
  - Resident skill
  - Local culture
- In CSP 474, 23% (167 / 725) of cases led by resident

Use of Resident Varies by Site

Site	Resident %
501	0%
506	81%
521	6%
523	0%
578	89%
580	0%
598	37%
618	61%
629	15%
652	0%
678	8%

Only supplies information on control group.

No variance within site. These cases are dropped if you use site fixed effects.

Resident Assignment in CSP 474

	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 strongly linked to site. Unclear why (culture, training patterns, supply of residents, etc.)

Bakaeen F, Sethi G, Wagner T, et al. Coronary Artery Bypass Graft Patency: Residents Versus Attending Surgeons. *Annals of Thoracic Surgery. in press*

Resident Assignment in CSP 474

	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

Bakaeen F, Sethi G, Wagner T, et al. Coronary Artery Bypass Graft Patency: Residents Versus Attending Surgeons. *Annals of Thoracic Surgery. in press*

Resident Assignment in CSP 474

	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 associated with angina symptoms and planned harvesting technique

Bakaeen F, Sethi G, Wagner T, et al. Coronary Artery Bypass Graft Patency: Residents Versus Attending Surgeons. *Annals of Thoracic Surgery. in press*

Sorting

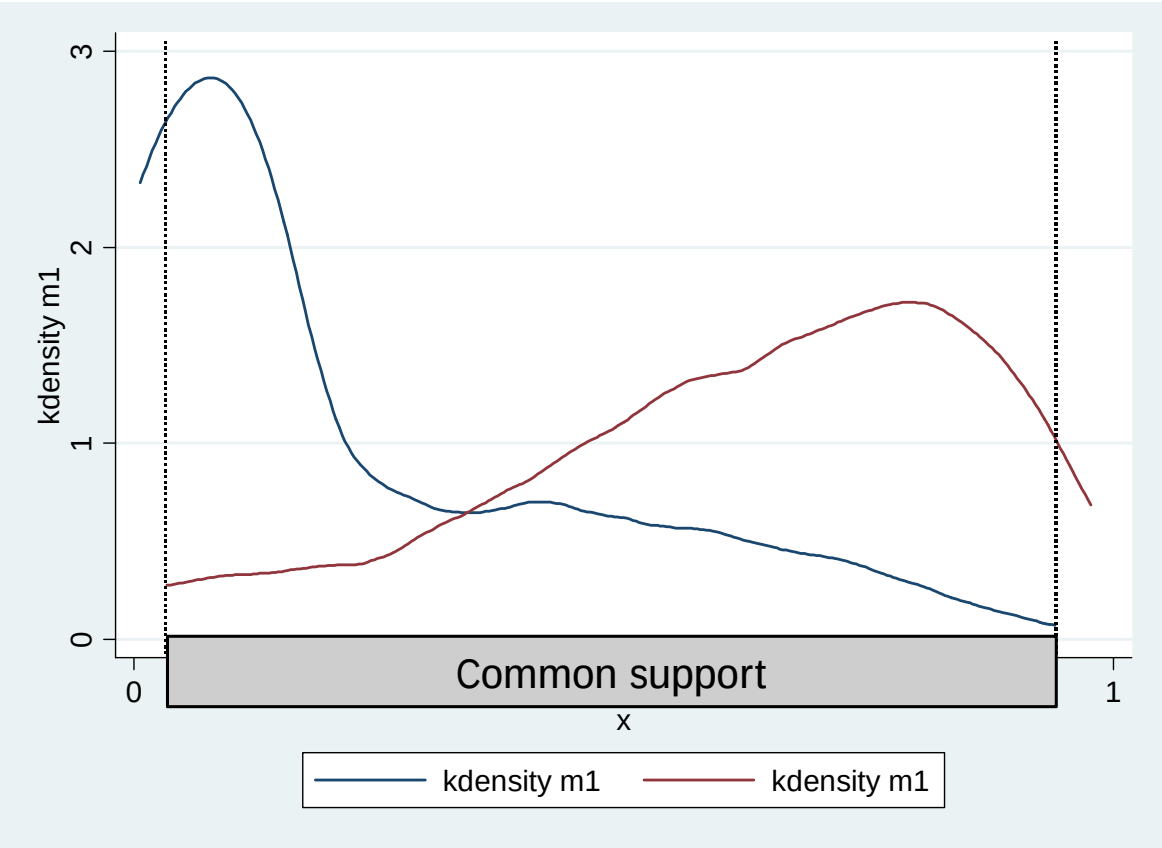
- Sorting is non-random
- If sorting is fully observed, we can estimate unbiased effect of resident surgeon effect
- Improbable that we fully observe the sorting process
  - Thus $E(x_i u_i) \neq 0$
  - Multivariate is biased and we need instrumental variables

Dimensionality

- The treatment and non-treatment groups may be different on many dimensions

- The propensity score reduces these to a single dimension

Common Support



These are the densities of having resident or non-resident surgery (m1 is propensity score)

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 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

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

Silk purse out of sow's ear?

- Propensity scores focus only on observed, not on unobserved.
- Improbable that we fully observe the sorting process
  - Thus $E(x_i u_i) \neq 0$
  - Multivariate (including propensity score) is biased and we need instrumental variables

Second Example

- CSP 474 was a randomized trial that enrolled patients in 11 sites
- Patients were randomized to two types of

- Is the sample generalizable?

  We compared enrollees to non-enrollees.

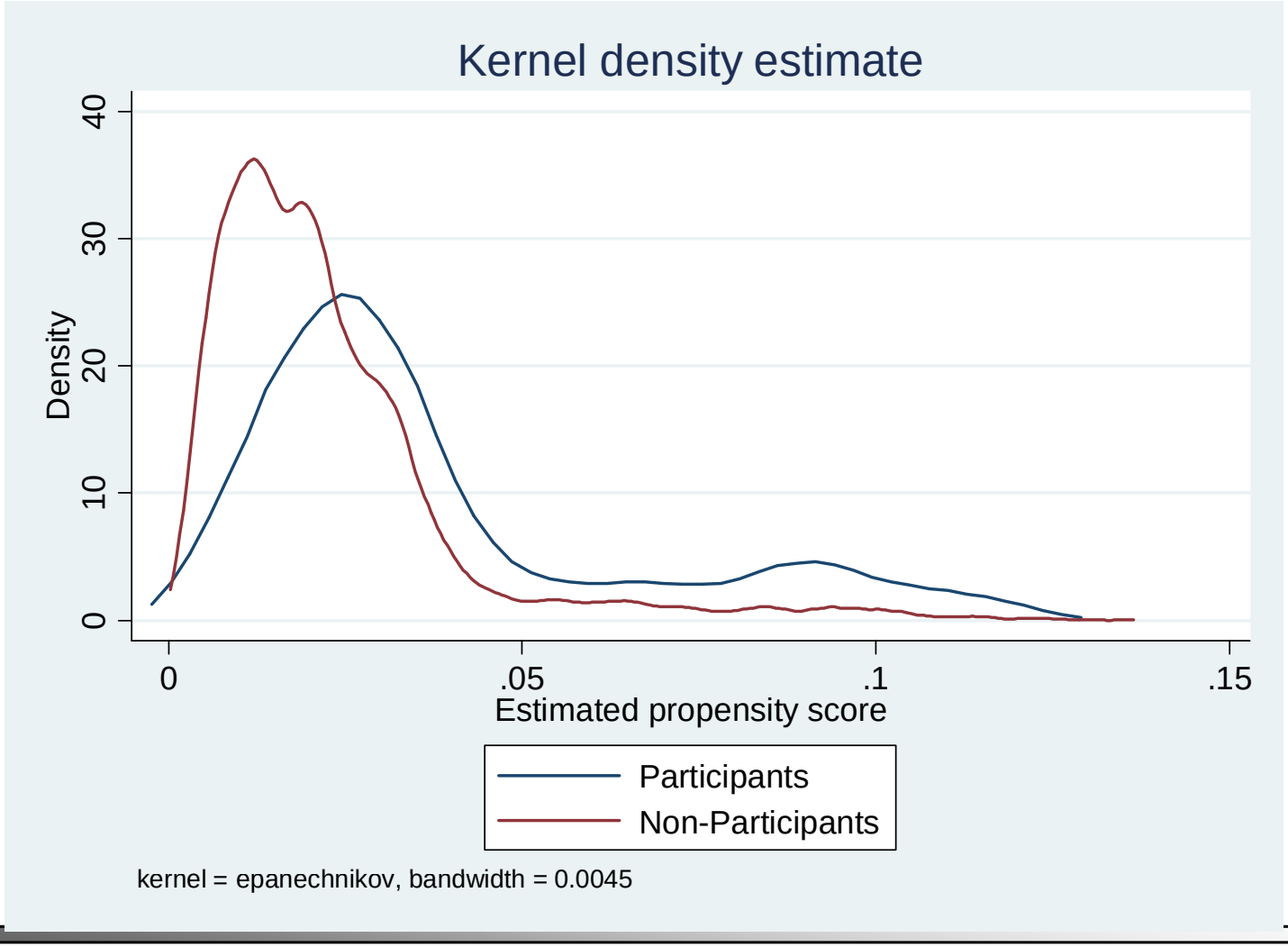
Methods

- 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 Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test 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 tended 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

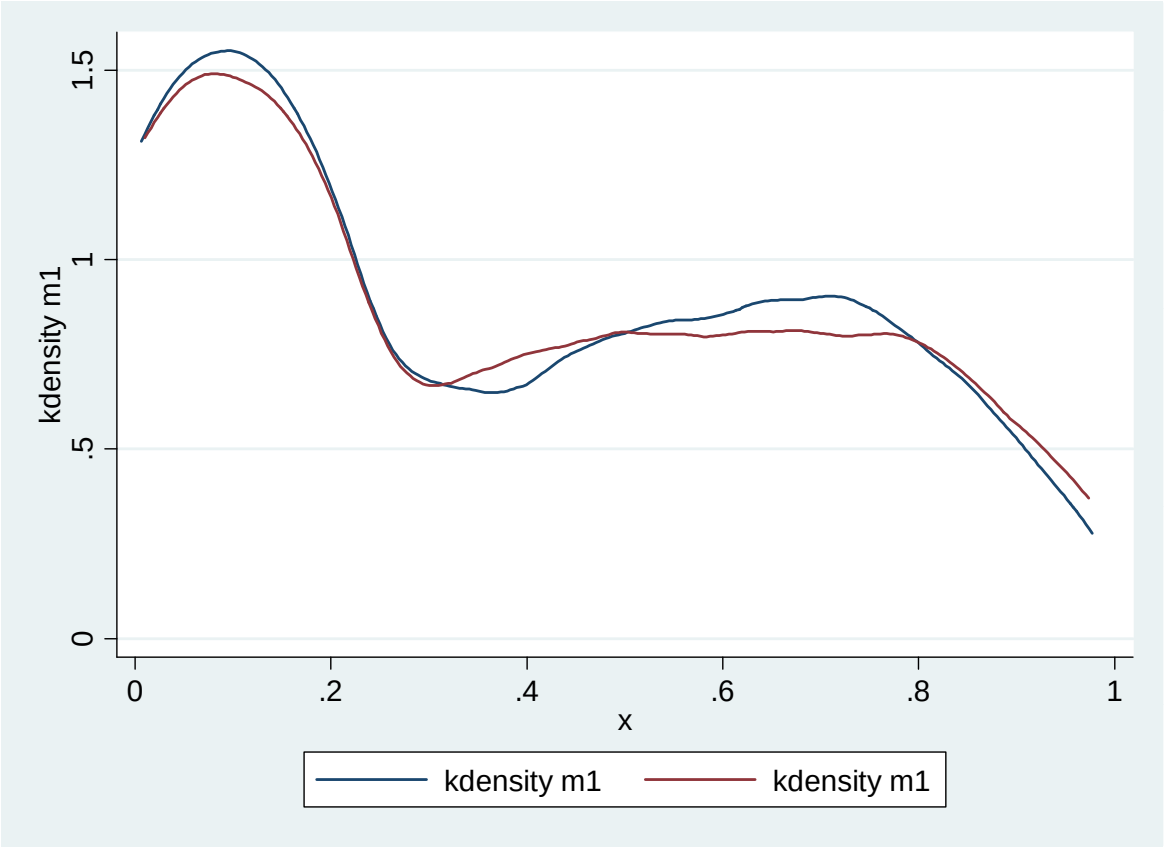
- To create generalizable estimates from the RCT, you can weight the analysis with the propensity score.

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

RCTs and Propensity Scores

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

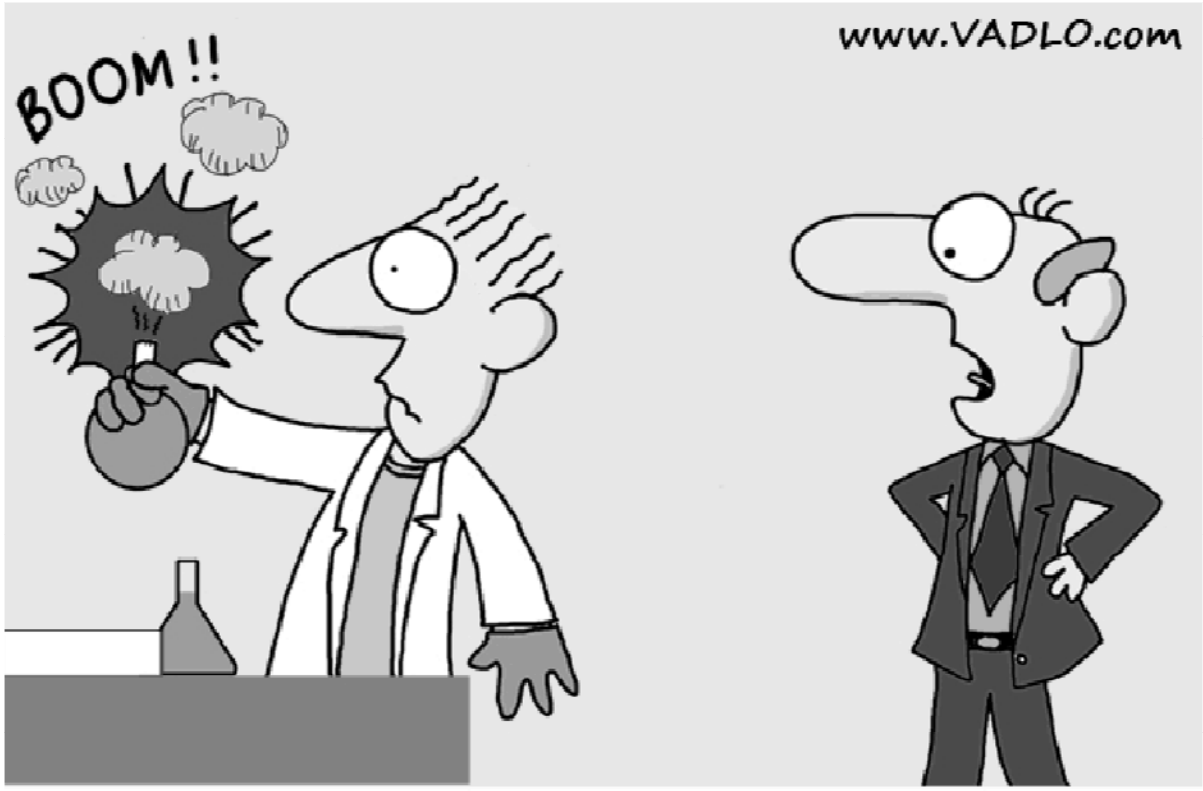
Share Common Support



Summary

- Propensity scores offer another way to adjust for confounding based on observables
- Reducing the multidimensional nature of confounding can be helpful
- Propensity scores do not attempt to adjust for unobserved.

Unrealistic Expectations



"I asked you not to mix Science with Religion."

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

of the estimators. [1]

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

Further Reading

- Imbens and Wooldridge (2007) www.nber.org/WNE/lect_1_match_fig.pdf
- Guo and Fraser (2010) Propensity Score Analysis. Sage.
- Brookhart MA, et al Am J Epidemiol. 2006 Jun 15;163(12):1149-56. Variable selection for propensity score models.