

Selection Bias and Confounding in Observational Studies



Steven D. Pizer
June 4, 2007



Outline

- Overview of selection bias: how it arises in study designs and problems it causes.
- Methods
 - Propensity scores.
 - Instrumental variables.
 - Selection models.
- Conclusions.

Overview

- Selection bias is well known.
 - Randomized controlled trials eliminate it
- Why conduct observational studies?
 - Cost of data collection.
 - Ethical considerations.
 - Faster results.
- But selection into treatment often correlated with outcome.
 - For example . . .



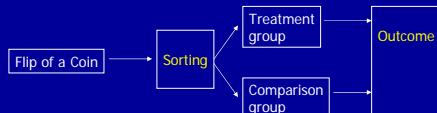
Study Suggests TV-watching Lowers Physical Activity 27 Aug 2006



A study of low-income housing residents has documented that the more television people say they watched, the less active they were, researchers from Dana-Farber Cancer Institute and colleagues report.

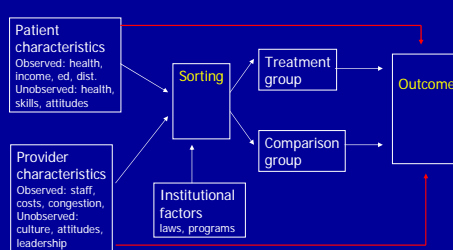
The findings of television's effects on physical activity are the first to be based on objective measurements using pedometers, rather than the study subjects' memories of their physical activity, say the researchers. The study will be published online by the *American Journal of Public Health* on July 27 and later in the journal's September 2006 issue.

Overview: Source of Bias in RCTs



- In RCTs, randomization ensures that
 - Observed (and unobserved) covariates are balanced between treatment and control groups
 - Only difference is treatment assignment
 - Thus, only cause of outcome difference is treatment
- **No bias** b/c coin flip is only driver of sorting and coin flip has no impact on outcomes

Overview: Source of Bias in Observational Studies

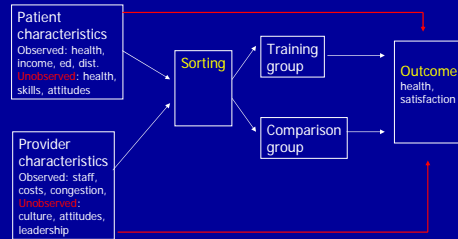


- In non-randomized studies, things get messy b/c there are many drivers of sorting that also impact outcomes.

Observational Study Scenarios

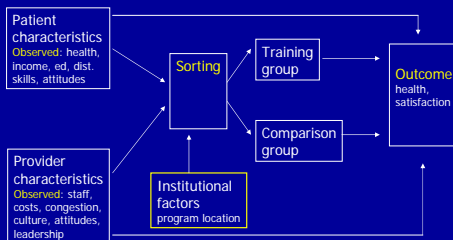
- Scenario A: Pilot clinical intervention. Self-care training for CHF. Self-reported health & satisfaction.
- Scenario B: Network-level study of guideline adherence. Annual eye and foot exams for diabetics. Amputations and retinopathy.
- Scenario C: National study of Medicaid HCBS. NH admission, mortality.

Scenario A: Pilot Study



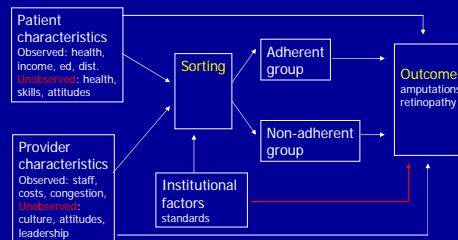
- Unobserved factors are important.
- No variables drive sorting without affecting outcome.

Scenario A: Pilot Study



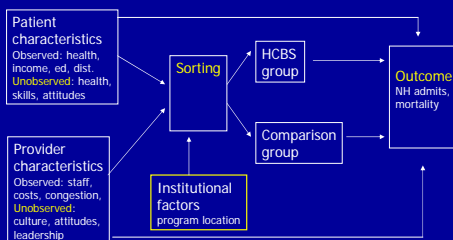
- All important factors are observed.
- Institutional factors drive sorting w/o affecting outcome.

Scenario B: Guidelines Study



- Unobserved factors are important.
- Institutional factors are related to outcome.

Scenario C: HCBS Study

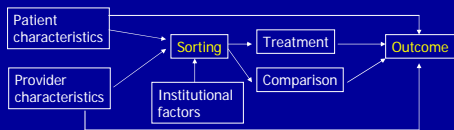


- Unobserved factors are important.
- Institutional factors drive sorting w/o affecting outcome.

Scenarios: Lessons

- A: Small study w/o unobservables. Propensity scores.
- B: Larger study w/ unobservables & w/o sorting variable uncorrelated w/ outcome. Fatally flawed.
- C: Large study w/ unobservables & uncorrelated sorting variable. Selection model or IV.

Translating Diagram Into Equations



Eq 1: $\text{Outcome} = \text{Treatment} + X_{\text{patient}} + X_{\text{provider}} + u_1$

Eq 2: $\text{Treatment} = X_{\text{patient}} + X_{\text{provider}} + X_{\text{institutions}} + u_2$

Selection bias occurs in Eq 1 when u_1 is correlated with u_2 , and therefore with Treatment.

Methods

- Propensity scores.
- Instrumental variables.
- Selection models.

Intuition Behind Propensity Scores

- In most observational studies, observed covariates not equal between treatment and comparison groups.
- But subgroups in the comparison group may look like subgroups in the treatment group.
- Identify people/groups with similarity in propensity to choose treatment.
 - People/groups with similar propensity are likely to have similar covariates.

Propensity Scores: How It's Done

- With parameter estimates (betas) from logistic regression on treatment, create predicted probability of treatment
 - This is the propensity score
- Use propensity score to match, stratify, or weight observations.
- Check that covariates are balanced afterwards.

Propensity Scores: Issues

- Successful balancing more likely if treatment regression is powerful.
- Results are unbiased if ALL variables that affect treatment and outcome are included.
 - This is the strong ignorability assumption.
- If unobservables affect both treatment and outcome, results will be BIASED.

Intuition Behind IV and Selection Models

- Selection bias means naively estimated effect of Treatment on Outcome reflects influence of unobservables correlated with Treatment variable.
- So we have to either remove effect of correlated unobservable (IV) or control for it (selection model).

Intuition: IV

- Instrumental variables (IV) uses variables that affect sorting but are not related to patient or provider unobservables.
- These are often institutional factors.
 - Example: Residence in county with HCBS waiver program. HCBS recipients vs. others => bias. Waiver county residents vs. others => no bias.

IV: How It's Done

Eq 1: Outcome = Treatment + X_{patient} + X_{provider} + u_1

Eq 2: Treatment = X_{patient} + X_{provider} + $X_{\text{institutions}}$ + u_2

- Estimate Eq 2 (like propensity score estimation).
- Construct predicted probability of Treatment (not a function of u_2).
- Substitute Pr(Treat) for Treatment.

IV: Issues

Eq 1: Outcome = Treatment + X_{patient} + X_{provider} + u_1

Eq 2: Treatment = X_{patient} + X_{provider} + $X_{\text{institutions}}$ + u_2

- Must have identifying instrument(s): $X_{\text{institutions}}$ in this case.
- Identifying instrument(s) must be strongly correlated with Treatment.
- Estimate only applies to those affected by instrument(s).
- Outcome equation must be linear.

Intuition: Selection Model

- Instead of eliminating correlated unobservables (like IV), try to control for them.
- Estimated Treatment effect wouldn't be biased by correlated unobservables if we could approximate and include them as control variables.

Selection Model: How It's Done

Eq 1: Outcome = Treatment + X_{patient} + X_{provider} + u_1

Eq 2: Treatment = X_{patient} + X_{provider} + $X_{\text{institutions}}$ + u_2

- Estimate Eq 2 and construct predicted residual ($E\{u_2|\text{Treatment}\}$).
- Add $E\{u_2|\text{Treatment}\}$ as another regressor in Eq 1.
- Adjust standard errors to reflect 2-step procedure.

Selection Model: Issues

Eq 1: Outcome = Treatment + X_{patient} + X_{provider} + u_1

Eq 2: Treatment = X_{patient} + X_{provider} + $X_{\text{institutions}}$ + u_2

- Must have identifying instrument(s): $X_{\text{institutions}}$ in this case.
- Identifying instrument(s) must be strongly correlated with Treatment.
- Need a probability distribution for u_2 to construct $E\{u_2|\text{Treatment}\}$.

Conclusions: Observational Studies

- In observational studies where you have a limited set of variables, selection bias is probably an issue.
- The methods you use can strongly impact the results you get.
- Qualitative understanding (QU) of sorting is as important as QU of outcome.
- All methods have assumptions that must be understood and satisfied to work well.

Conclusions: Propensity Scores

- Using Propensity Scores doesn't turn an observational study into an RCT.
- Propensity scores are effective at balancing observed covariates, but deal with selection bias if and only if:
 - You know exactly what drives selection.
 - you measure those drivers.
- Otherwise, propensity scores unlikely to have dealt with selection bias.

Conclusions: IV and Selection Models

- IV and selection models are effective at dealing with unobserved covariates if you have suitable instruments.
- These methods require technical expertise to implement and make presentation of results more difficult.
- Both approaches depend critically on the availability of variables that drive sorting but not the outcome.

The Last Word

- If unobservables are important, sorting is related to outcome, and you don't have unrelated variables that explain sorting . . .
- **DANGER!** Results likely biased. Conclusions will be hard to defend.
- Recommendation: Find different variables or approach.

Selection Bias and Confounding in Observational Studies



Steven D. Pizer
June 4, 2007

