Addressing Selection Bias In Observational Studies

AcademyHealth Methods Workshop
June 14, 2011
Outline of Workshop

• Overview of causation in observational studies
• How selection bias can arise
• Best practices for propensity score modeling
• Review of instrumental variables (IV) analysis
• Conditions under which quasi-experiments produce estimates similar to experiments
Speakers

- Paul Hebert, PhD
  - Seattle VA HSR&D and University of Washington

- Matt Maciejewski, PhD
  - Durham VA HSR&D and Duke University

- Steve Pizer, PhD
  - Boston VA HSR&D and Boston University
Conditions under which Quasi-Experiments generate estimates similar to Experiments

Matthew L. Maciejewski, PhD
Durham VA HSR&D and Duke University
RCT considered Gold Standard of Benefit Design for Several Reasons

• Create balance in observed covariates
  – Reduces number of competing hypotheses for variation in outcomes to one (treatment assignment)
  – Control group outcome is a valid counterfactual (unbiased estimate of outcome for treatment group had they not been randomized to treatment)

• Treatment effect generalizes to entire sample
Context for Perceived Inferiority of Quasi-Experiments

• Prior comparisons of RCTs and non-RCTs
  – Experimental results rarely replicated
  – Even when applying instrumental variables (IV) methods (LaLonde 1986)
  – Concordance of RCT & IV estimates (Stukel ‘07 JAMA)

• RCTs typically compared to non-identical samples and non-identical outcomes in different data
  – Conclusion has been that design (quasi-experiment) is the cause of difference, not sample or outcomes
  – Could outcomes be similar across designs if same sample & outcomes?
## LaLonde (1986) Job Training Results

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Wage Difference for Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted RCT</td>
<td>$886</td>
</tr>
<tr>
<td>Non-RCT estimates from PSID &amp; CPS-SSA</td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>Low=-$1637, High=$1714</td>
</tr>
<tr>
<td>Age adjusted</td>
<td>Low=-$1388, High=$195</td>
</tr>
<tr>
<td>Age, schooling, race &amp; pre-period wage</td>
<td>Low=-$1228, High=$1466</td>
</tr>
<tr>
<td>IV</td>
<td>Low=-$667, High=$889</td>
</tr>
</tbody>
</table>
### Stukel 2007 JAMA: Mortality Impact of Cardiac Cath

<table>
<thead>
<tr>
<th>Model</th>
<th>Risk Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted survival</td>
<td>0.36 (0.36, 0.37)</td>
</tr>
<tr>
<td>Multivariate adjustment</td>
<td>0.51 (0.50, 0.52)</td>
</tr>
<tr>
<td>Simple PS Adjustment: Deciles + Covariates</td>
<td>0.52 (0.51, 0.53)</td>
</tr>
<tr>
<td>Fancy PS Adjustment: Deciles + Covariates</td>
<td>0.52 (0.51, 0.53)</td>
</tr>
<tr>
<td>Instrumental Variables</td>
<td>0.84 (0.79, 0.90)</td>
</tr>
<tr>
<td>RCT Results</td>
<td>0.79-0.92</td>
</tr>
</tbody>
</table>
Re-appraising the Value of Quasi-experiments

• An under-used design that allows direct comparison of results from RCT & non-RCT
  – Within-study comparison study
• Four-arm study: 2-stage process
  – Randomize to randomized treatment or self-selected treatment
  – Same treatments. same controls, same outcomes
• Estimate two treatment effects
  – Difference between treatment & control in RCT “arm”
  – Difference btn treatment & control in non-RCT “arm”
Design of Within-Study Comparison by Shadish (2008)

Recruited Students

Pretests then Randomly Assigned

Randomized Experiment
- Mathematics Training
- Vocabulary Training

Nonrandomized Study
- Mathematics Training
- Vocabulary Training
Criteria of Good Within-Study Comparison Design

1. Variation in mode of assignment--random or not
2. No third variables correlated with both assignment and outcome
   • Careful measurement of selection process
3. Randomized experiment properly executed
4. Properly executed quasi-experiment
5. Both design types estimate the same causal entity
   • RCT typically estimates ITT, PS match estimates ATT
Details of Shadish (2008) Design

• Participants from one college
• Participants pretested on several covariates
• Chose math and vocabulary training because
  — Good analogue to educational interventions
  — Easy to control effect size with item difficulty
  — Math phobias cause plausible selection bias
• All participants treated together without knowledge of the different conditions
• All participants post-tested on both math and vocabulary outcomes
Unadjusted Results:
Vocabulary Training Effect on Vocabulary Outcome

<table>
<thead>
<tr>
<th></th>
<th>Vocab Training</th>
<th>Math Training</th>
<th>Mean Difference</th>
<th>Absolute Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted RCT</td>
<td>16.19</td>
<td>8.08</td>
<td>8.11</td>
<td></td>
</tr>
<tr>
<td>Unadjusted Quasi-exper.</td>
<td>16.75</td>
<td>7.75</td>
<td>9.00</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Conclusions
1. The effect of vocab training on vocab scores was larger (9 of 30 points) when participants could self-select into vocab training.
2. The 8.11 point effect (out of 30 possible points) in the randomized experiment was overestimated by 11% (.89 points) in the nonrandomized experiment
Propensity Score Modeling

• Based on a priori model of selection process that informed prospective pre-test assessments

• Extensive adjustment
  – Math & vocabulary pretest scores, ACT, GPA, prior exposure to math courses, math anxiety, demographic
  – “Big 5” personality traits (extraversion, emotional stability, agreeableness, intellect, & conscientiousness)

• Limited adjustment (comparable to claims)
  – Age, sex, race & marital status had reduced bias modestly (12-30%)
Results with Extensive Propensity Score Adjustments

![Graph showing percent bias remaining after adjustment for different methods of propensity score adjustment. The graph compares vocabulary outcome and mathematics outcome across four methods: unadjusted quasi-experiment, propensity score stratification, propensity score ANCOVA, and propensity score weighting. The graph indicates a decrease in percent bias after adjustment, with the largest decrease seen in the propensity score stratification method.]
Implications of Shadish (2006)

• Sampling design produced non-equivalent groups on observables
• Big overlap in baseline values in RCT & non-RCT groups due to 1st stage randomization made propensity scores more valid
• Extensive measurement of relatively simple selection process, though not homogeneous
  – Propensity score matching may not be effective if selection process is complex (as in job training)
• Propensity score results from extensive adjustment matched RCT results
Diaz & Handa (2006)

• Progresa: Anti-poverty program in Mexico
  – Multi-stage randomization: Village then family based on material resources
• Assignment into treatment is completely known because Mexican gov’t implemented
  – 506 rural localities in 7 states randomized
  – Other localities not randomized
• Outcomes
  – Food expenditures, teen school dropout, child labor
Diaz & Handa (2006)

• Two comparison groups
  – #1: People in eligible villages randomized to control
  – #2: Poorer people who were ineligible because they lived in richer villages

• Results (RCT Tx vs. #1) vs. (RCT Tx vs. #2)
  – Unadjusted results differed
  – Results similar when adjustment for covariates known to drive selection
Conditions Under Which Quasi-Experiments Match RCT Results

• Population matching on pre-period values
  – When geographically local, comparison groups may not differ on major observables b/c provider & site effects controlled
  – For example: Twins, cohorts in same organization

• Rigorous conceptualization and measurement of selection process to support effective matching if non-equivalent controls
  – Pre-period assessments on the outcome are particularly important
  – Adjustment using “off the shelf” vars not enough

• Large sample sizes
Reconsider Value of Quasi-Experiments for Causal Inference?

• Comparing good RCT to poor quasi-experiment confounds design type and the quality of its implementation
  – Logical fallacy
• This conclusion is ex post facto because we know RCT results in advance
• Quasi-experiments satisfying three conditions are more likely to match RCT results
Questions?
References


• Shadish, Clark & Steiner, 2008. *JASA*, 103(484): 1334-1356

• http://steinhardt.nyu.edu/scmsAdmin/uploads/002/477/Tom%20Cook-FINAL.pdf