Can Quasi-Experiments Yield Causal Inferences?

Matthew L. Maciejewski, PhD
Durham VA HSR&D and Duke University

Sample

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Race</th>
<th>SES</th>
<th>Health status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Intervention

<table>
<thead>
<tr>
<th>Intervention list</th>
<th># of interactions</th>
<th>Duration of each interaction</th>
<th>Single topic or multiple topics</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
- Statistical result is causal effect of treatment on outcome

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)
- 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?

Differences in Samples for RCTs and Quasi-Experiments

- RCTs
  - Conducted on highly selected populations
  - Rarely pregnant women, highest risk people, oldest
- Quasi-experiments
  - Conducted on general populations
- Differences not necessarily due to randomization
  - Could be entirely due to different samples included
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 Catheterization

<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>
</tbody>
</table>

Conclusion
1) Adjustment for covariates important in non-RCT
2) Multivariate & PS regressions are same

Stukel 2007 JAMA: Mortality Impact of Cardiac Catheterization

What to conclude?
1) Regression & PS results are both right?
2) Results are both wrong?
Stukel 2007 JAMA: Mortality Impact of Cardiac Catheterization

<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>

What to conclude?
1) Regression & PS results are both right?
2) Results are both wrong? **This is it.**

Re-appraising the Value of Quasi-experiments

- An under-used design 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, controls, outcomes, timing
- Can compare two treatment effects!
  - Difference btn treatment & control in RCT “arm”
  - Difference btn treatment & control in non-RCT “arm”

Design of Within-Study Comparison by Shadish (JASA 2008)

- Recruited Students
  - Pretests then Randomly Assigned
    - Randomized Experiment
    - Nonrandomized Study
      - Mathematics Training
      - Vocabulary Training
      - Mathematics Training
      - Vocabulary Training
Details of Shadish (2008) Design

- Participants from one college
- Participants pretested on several covariates
- Chose math and vocabulary training because
  - Easy to induce effect with item difficulty
  - Math phobias cause plausible selection bias
- All participants treated together (in same class) without knowledge of different conditions
  - People randomized to math in same training class as people self-selecting math
- Everyone post-test on math & vocab 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-experiment</td>
<td>16.75</td>
<td>7.75</td>
<td>9.00</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Conclusions
1. Effect of vocab training on vocab scores was larger (9 of 30 points) when participants could self-select into vocabulary training
2. The 8.11 point effect in RCT was overestimated by 11% (0.89 points) in the quasi-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)
  - Extensive adjustment reduced bias a lot (59-96%)
- Limited adjustment (comparable to claims)
  - Age, sex, race & marital status had reduced bias modestly (12-30%)
Bias Reduction Fairly Similar Across Different Propensity Score Methods

<table>
<thead>
<tr>
<th>Propensity Score Adjustment Method</th>
<th>Unadjusted Quasi-Experiment</th>
<th>Propensity Score Stratification</th>
<th>Propensity Score ANCOVA</th>
<th>Propensity Score Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Bias Remaining After Adjustment</td>
<td>Vocabulary Outcome</td>
<td>Mathematics Outcome</td>
<td>60%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Implications of Shadish (2008)

- 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)
- Bottom line: Propensity score results from extensive adjustment matched RCT results

Limitations of Shadish (2008)

- Short duration (15 minutes)
  - Not costly to conduct
  - Little incentive for non-compliance
- Absence of non-compliance with treatment assignment
- Short time between pretest & post-test, and short time between treatment & posttest
  - Change attributable to few things besides treatment
- Not generalizable to complex medical settings
  - Longer duration, have significant non-compliance and delay between treatment and outcomes assessment
Conditions Under Which Quasi-Experiments Match RCT Results

- Similarity between groups in pre-period values
  - When geographically local, comparison groups may not differ on major observables b/c provider & site effects controlled (e.g., pts in same clinic)
  - ACEI example (Hebert & Maciejewski)
- Rigorous conceptualization and measurement of selection process to support effective matching
  - Pre-period outcomes are particularly important
  - Adjustment using "off the shelf" vars not enough
- Regression discontinuity

Descriptive Statistics of Unmatched CA ACEI and Non-CA ACEI Cohorts

<table>
<thead>
<tr>
<th></th>
<th>CA ACEI Cohort</th>
<th>Non-CA ACEI Cohort</th>
<th>Standardized Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>76.1</td>
<td>75.9</td>
<td>7.17</td>
</tr>
<tr>
<td>Female (%)</td>
<td>65%</td>
<td>64%</td>
<td>2.09</td>
</tr>
<tr>
<td>White Race (%)</td>
<td>76%</td>
<td>83%</td>
<td>17.41</td>
</tr>
<tr>
<td>Black Race (%)</td>
<td>9%</td>
<td>7%</td>
<td>7.38</td>
</tr>
<tr>
<td>Baseline AMI (%)</td>
<td>6.7%</td>
<td>4.1%</td>
<td>11.52</td>
</tr>
<tr>
<td>Elixhauser Score</td>
<td>5.69 (7.79)</td>
<td>4.66 (6.99)</td>
<td>44.54</td>
</tr>
<tr>
<td>Baseline Expenditures</td>
<td>$8081 (15210)</td>
<td>$6180 (12798)</td>
<td>16.06</td>
</tr>
<tr>
<td>Baseline # Meds</td>
<td>6.7 (3.8)</td>
<td>6.4 (3.6)</td>
<td>26.03</td>
</tr>
<tr>
<td>Office visits</td>
<td>8.6 (8.3)</td>
<td>8.5 (9.0)</td>
<td>4.42</td>
</tr>
</tbody>
</table>

Careful Consideration of Selection Process

- Bias can be significantly reduced if three steps of confounder adjustment are done
  - Identification of all relevant confounders from literature, theory, and experts
  - Error-free measurement
  - Proper modeling
- Use of variables of convenience fails 1st step, so unlikely to reduce bias fully
  - Especially true in claims data?
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
  - Rarely true; more often have to infer a la Stukel
- Quasi-experiments satisfying three conditions more likely to generate valid causal estimates

Questions?

References

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


