Study Designs Appropriate for Comparative Effectiveness Research

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

Objectives of Session

• Motivation
• Incident vs. prevalent user cohorts
• Counterfactuals (aka comparators, controls)
• Understand definition of and threats to internal validity and external validity
  – Fundamental to study design selection
• Appreciate most rigorous quasi-experimental study designs
  – Quasi-experiments as “broken trials”

Motivation & Value of Quasi-Experimental Studies

• Quasi-experiment is 1st step to RCT
  – Big secondary data yield sample estimates of outcome variation and differences
  – May generate/support hypotheses of causes

• Quasi-experiment is only way to assess change in policy or clinical practice
  – Population effect in real world is parameter of interest
**Ideal Quasi-Experiment**

- Answers the research question
- Great internal validity
  - Treatment effect estimated is “true” effect
- External validity
  - Generalizable across settings, time

**Competing Demands in CER**

- As scientists, internal validity is #1 concern
  - Why RCTs considered gold standard
  - Why evidence from observational studies considered suspect or invalid
  - Historically, this is of greatest concern
- As policy/clinical experts, generalizability is of concern
  - Want results to be definitive across space (& time?)
  - In CER, this is more prominent and quasi-experimental studies can address this issue

**Four Design Choices**

- In absence of randomization, three study design choices determine internal validity of your study
  - Incident or prevalent users
  - Control group
  - Pre-period measurement
  - Post-period measurement
- The more thoughtfully these design elements are chosen a priori, the stronger the study’s internal validity and causal statements
Concept of Internal Validity

- Internal validity refers to ability to make causal inferences (e.g., Y is caused by X)
  - With successful randomization, X is only source of variation between groups
  - In quasi-experimental studies, X is no longer only source of variation b/c tx & control not balanced as in RCT
- Internal validity is not generalizability, reproducibility or reliability

Conditions for Causality

- Treatment precedes outcome
  - Temporal relationship
- Treatment and outcome covary
  - Significant correlation
- No other plausible competing hypotheses
  - Identification of “true” cause is process of eliminating all competing hypotheses but 1
  - If all other explanations for outcome variation are eliminated but treatment, treatment is cause

Incident User
& Prevalent User Designs
Three Cohort Choices

- Incident (aka new) users
  - Index date refers to first use with no history of prior use
  - Useful for medication and surgery studies
- Prevalent users
  - Index date refers to treatment use in reference to some other date (e.g., policy change)
  - Represents majority of medication users at point in time
- Both incident & prevalent users (‘wedge design’)

Tradeoffs of Incident User Cohort

- Advantages
  - Well suited to assess acute meds or initial risks of chronic meds
  - Captures events that occur early in therapy
  - Can control for risk factors before they are altered by exposure
- Disadvantages
  - Likely small sample
  - Limited precision/efficiency
  - Need to be careful about characterizing risk of adverse events if risk varies greatly over time in early period

Defining Pre-Period for Incident Cohort Design

- Two reasons to define pre-period carefully
  - Assess covariates before patient exposed to tx
  - Ensure no prior history of treatment exposure (wash out)
- Need to choose look back period carefully
  - 6 months: 32 of 1000 pts were eligible
  - 12 months: 27.5 of 1000
  - 24 months: 23.5 of 1000
  - 108 months: 17.2 of 1000
Tradeoffs of Prevalent User Cohort

- Advantages
  - Can characterize event risk because risk likely stable over time
  - Large sample
  - Improved precision/efficiency
  - Addresses experience of most current medication users

- Disadvantages
  - Under-ascertainment of events occurring early due to treatment
    - You miss folks who stopped meds due to adverse event
  - “Baseline” covariates may have been impacted by treatment exposure
  - Treatment duration unknown and variable
  - Patients who adhere to treatment over long term are not random sample


Defining Pre-Period for Prevalent Cohort

- Need to choose lookback period and history of treatment exposure carefully

<table>
<thead>
<tr>
<th>Study</th>
<th>Pre-Period Duration (months)</th>
<th>Number of Intervals</th>
<th>Number of Fills per Interval</th>
<th>Each Interval Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blais 2003</td>
<td>70</td>
<td>1</td>
<td>1+</td>
<td>70</td>
</tr>
<tr>
<td>Roblin 2005</td>
<td>12</td>
<td>2</td>
<td>4+, 1+</td>
<td>6, 6</td>
</tr>
<tr>
<td>Landsman 2005</td>
<td>24</td>
<td>2</td>
<td>1+</td>
<td>3, 24</td>
</tr>
<tr>
<td>Schneeweiss et al., 2007</td>
<td>12</td>
<td>2</td>
<td>0, 1+</td>
<td>6, 12</td>
</tr>
<tr>
<td>Yin, 2008</td>
<td>24</td>
<td>1</td>
<td>1+</td>
<td>24</td>
</tr>
<tr>
<td>Doshi 2009</td>
<td>24</td>
<td>2</td>
<td>1+</td>
<td>3, 24</td>
</tr>
<tr>
<td>Maciejewski 2010</td>
<td>12</td>
<td>2</td>
<td>1+</td>
<td>3, 9</td>
</tr>
</tbody>
</table>

The Value of a Control Group for Internal Validity
Importance of Control Group

- Terminology
  - Control group = counterfactual = comparator

- When observing treatment, how do we know what would have happened in the absence of treatment?
  - Counterfactual evidence is needed

- Control group is needed to compare
  - Outcomes under treatment
  - Outcomes NEVER under treatment

What is a Counterfactual?

- Counterfactual is measurement on subject that mirrors subject in every respect
  - Observed and unobserved

- Perfect counterfactual represents
  - What would have happened to treated patient had they not received treatment
  - What would have happened to control patient had they received treatment

- Ideal counterfactual: two universes
  - Same person in treatment & in control at same time

Ways of Estimating Counterfactuals

- Second best: Identical twins in RCT
- Third best: RCT
- Fourth best: Quasi-experiment w/ equivalent controls
  - Recommended by Shadish (2008)
- Fifth best: Quasi-experiment w/ non-equivalent controls
  - This is typical counterfactual in quasi-experiment
What is Effect of QI Program on Hospital Acquired Infections?

Harris, Maciejewski, in press Health Affairs

Counterfactuals in Quasi-Experiments

- Self-selection in quasi-experiments
  - Imbalance in observed covariates
  - Can assume imbalance in unobserved factors
  - Imbalance = nonequivalence (compared to RCT)

- Non-equivalence complicates interpretation of treatment effect b/c several explanations
  - Treatment? Selection bias? Both?
Threats to Internal Validity

Biggest Internal Validity Threats in Quasi-Experimental CER

• Ambiguous Temporal Precedence
• Selection Bias
• Regression
• History

Threats to Internal Validity

• Ambiguous Temporal Precedence
  – Major concern in cross-sectional studies
  – Eg., Volume-outcome relationship
• Selection
  – “confounding of treatment effects with population differences”
  – Observed (easier to fix) or Unobserved
  – May require knowledge of how people get treatment, not just knowledge of outcome
### Example of Selection Bias on Observables (Bariatric Surgery)

<table>
<thead>
<tr>
<th></th>
<th>Surgical Cases (N=850)</th>
<th>Non-Surg Controls (N=41,244)</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age (Mean, SD)</td>
<td>49.5 (8.3)</td>
<td>54.7 (10.2)</td>
</tr>
<tr>
<td></td>
<td>Male (%)</td>
<td>73.9</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>Caucasian (%)</td>
<td>77.9</td>
<td>67.8</td>
</tr>
<tr>
<td></td>
<td>Non-Caucasian (%)</td>
<td>16.0</td>
<td>19.3</td>
</tr>
<tr>
<td></td>
<td>Married (%)</td>
<td>52.1</td>
<td>57.7</td>
</tr>
<tr>
<td></td>
<td>Previously Married (%)</td>
<td>30.4</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>Never Married (%)</td>
<td>16.4</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>Unknown Marital (%)</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>BMI (Mean, SD)</td>
<td>47.4 (7.8)</td>
<td>42.0 (5.0)</td>
</tr>
<tr>
<td></td>
<td>Diagnostic Cost Group score (Mean, SD)</td>
<td>0.60 (0.92)</td>
<td>0.47 (0.75)</td>
</tr>
</tbody>
</table>

### Addressing Temporal Precedence and History Threats

- **Ambiguous Temporal Precedence**
  - **Design**: Choose sample to identify precedence
  - **Statistics**: Examine $Y = f(X)$ & $X = f(Y)$

- **Selection**
  - **Design**: Careful *a priori* choice of primary comparator
  - **Measurement**: Additional data collection
  - **Statistics**: Propensity score, selection methods

### History Threat to Internal Validity

- **History**
  - Co-occurring events (e.g., pt in multiple RCTs) that could impact outcome of interest
  - Attitudes about race, homosexuality change

- **Address by understanding, measuring and adjusting for co-occurring changes over time**
History Threat:
Near Simultaneous Copay Changes

December 6, 2001
Specialty Care up from $15 to $50
Primary Care copay ($15) introduced

2000 2001 2002 2003

February 4, 2002
Medication copay up from $2 to $7

Maciejewski, et al., 2010 AJMC

---

Attempt to Account for
History Threat (1 of 2)

Full Model includes lagged OP visits

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copay non-exempt</td>
<td>0.030</td>
<td>0.075</td>
</tr>
<tr>
<td>Short-term (ST) post</td>
<td>0.292</td>
<td>0.055</td>
</tr>
<tr>
<td>Long-term (LT) post</td>
<td>-0.078</td>
<td>0.076</td>
</tr>
<tr>
<td>Copay*ST post</td>
<td>-0.082</td>
<td>0.060</td>
</tr>
<tr>
<td>Copay * LT post</td>
<td>-0.150</td>
<td>0.082</td>
</tr>
</tbody>
</table>

Wang, Liu & Maciejewski, HSR in press

---

Attempt to Account for
History Threat (2 of 2)

Full Model includes lagged OP visits
Reduced model (excludes lagged OP visits)
Reduced model (includes current OP visits)

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>S.E.</th>
<th>β</th>
<th>S.E.</th>
<th>β</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copay non-exempt</td>
<td>0.030</td>
<td>0.075</td>
<td>0.030</td>
<td>0.075</td>
<td>0.030</td>
<td>0.075</td>
</tr>
<tr>
<td>Short-term (ST) post</td>
<td>0.292</td>
<td>0.055</td>
<td>0.288</td>
<td>0.055</td>
<td>0.288</td>
<td>0.055</td>
</tr>
<tr>
<td>Long-term (LT) post</td>
<td>-0.078</td>
<td>0.076</td>
<td>-0.078</td>
<td>0.076</td>
<td>-0.078</td>
<td>0.076</td>
</tr>
<tr>
<td>Copay*ST post</td>
<td>-0.082</td>
<td>0.060</td>
<td>-0.078</td>
<td>0.060</td>
<td>-0.078</td>
<td>0.060</td>
</tr>
<tr>
<td>Copay * LT post</td>
<td>-0.150</td>
<td>0.082</td>
<td>-0.145</td>
<td>0.082</td>
<td>-0.144</td>
<td>0.082</td>
</tr>
</tbody>
</table>

Wang, Liu & Maciejewski, HSR in press
Regression Threat to Internal Validity

• History
  – Co-occurring events (e.g., pt in multiple RCTs) that could impact outcome of interest
  – Attitudes about race, homosexuality change

• Address by understanding, measuring and adjusting for co-occurring changes over time

What is Effect of QI Program on Hospital Acquired Infections?

Harris, Maciejewski, in press Health Affairs
What is Effect of QI Program on Hospital Acquired Infections?

<table>
<thead>
<tr>
<th>Year</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yr -1</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>Yr 1</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Yr 2</td>
<td>20</td>
<td>5</td>
</tr>
</tbody>
</table>

Harris, Maciejewski, in press Health Affairs

Measurement Issues for Internal Validity
Pre-tests, Post-tests & Non-equivalent Dependent Variables

- Pre-test measurements
  - Can assess selection bias
- Post-test measurements
  - Can reduce/eliminate ambiguity about temporal precedence of cause & effect
  - The more the better to see pattern of association
- Non-equivalent outcomes (post-tests)
  - Assesses degree of change due to other (unmeasured) contextual factors

Natural Experiments

- Definition: A study that contrasts a naturally occurring event, such as an earthquake, with a comparable condition (Shadish, Cook & Campbell)
  - Intervention cannot be manipulated by researcher or research subject

Example of a Natural Experiment

- Value-based insurance design (VBID) program implemented by BCBSNC in Jan 2008
  - Lowered generic copays for some, brand copays for all
  - Population based implementation, so tons of patients across multiple drug classes
  - Many employers opted out, so good controls
- Controls weren’t equivalent, but made so via propensity score matching
- Non-equivalent outcomes
  - ACEIs & CAIs only had brand Rx available

Maciejewski, Health Affairs Nov 2010
Design of VBID Evaluation

- Pre-tests
  - One year prior to VBID implementation
- Post-tests
  - Two years after VBID implementation
- Controls
  - Non-equivalent, but made so via propensity score matching
- Non-equivalent outcomes (ACEIs & CAIs)
  - Only brand Rx available

Maciejewski, Health Affairs Nov 2010

---

Unadjusted Trends in Diuretic Adherence (MPR)

<table>
<thead>
<tr>
<th>Year</th>
<th>Cases</th>
<th>Controls</th>
<th>Difference (3%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>79%</td>
<td>79%</td>
<td>6%</td>
</tr>
<tr>
<td>2008</td>
<td>79%</td>
<td>76%</td>
<td>3%</td>
</tr>
</tbody>
</table>

Adjusted Adherence Difference Associated with VBID Program

Note: GLM in unmatched sample, controlling for age, gender, Episode Risk Groups, number of unique medications filled, generic dispensing rate, total copay, and whether an enrollee filled one or more prescriptions with a 90-day supply; Maciejewski et al. Nov 2010 Health Affairs
Characteristics of Quasi-Experiments with Strong Internal Validity

- At least one pre-test measurement
- At least one post-test measurements
  - Outcomes of interest
  - Non-equivalent outcomes
- Control group
  - The more equivalent the better
- Choice of incident or prevalent user design depends on research question

External Validity

Concept of External Validity

- External validity = generalizability of results
  - Persons, settings, treatments outside of study

- Fundamental question: Is response to treatment is homogeneous or heterogeneous?
  - Assessment by subgroup (defined by pt characteristic, site, time) can inform this issue
Threat to External Validity due to Heterogeneity in Units

- Due to Heterogeneity in Units: Response may differ across subgroups
  - Age groups, genders, ethnicities
  - Income levels
  - Health systems (VBID example)
- Due to Treatment Variation
  - Interventionist variation, if 2+ interventionists
  - In multi-component trials, intervention delivered may vary if some pts got components #2-5 but others got components #6-9?

Example of External Validity in VBID Results

External Validity & Treatment Variation in DPP

<table>
<thead>
<tr>
<th></th>
<th>Original DPP Lifestyle Arm</th>
<th>DEPLOY</th>
<th>East Harlem DPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Design</td>
<td>RCT</td>
<td>RCT</td>
<td>RCT</td>
</tr>
<tr>
<td>Setting</td>
<td>AHCs</td>
<td>YMCA</td>
<td>Urban low income</td>
</tr>
<tr>
<td>Intervention</td>
<td>Individual lifestyle</td>
<td>16 group meetings at YMCA</td>
<td>??</td>
</tr>
<tr>
<td>Weight loss</td>
<td>6% greater</td>
<td>5.5% greater</td>
<td>1.9% greater</td>
</tr>
<tr>
<td>Intervention cost</td>
<td>$2,269 over 3 years</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Cost/QALY</td>
<td>$1,100</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>
External Validity and CER/Translation Research

- Fundamental issue in disseminating RCTs to communities (we are struggling with)
- If intervention we test in RCT is implemented in community clinic, would results differ?
  - Protocol: We shorten intervention intensity and/or change interventionist
  - Setting: Non-academic
  - Subjects: Lower severity & motivation
  - Funding: None (unlike RCT)

Take-Away Points

- Quasi-experimental study designs are widely used b/c secondary data analysis is common
  - Several threats to internal validity
  - Greatest external validity
- Have more design/statistical issues than RCTs
  - Require more data than RCTs
  - More assumptions about data (not observed)
  - Creativity in control group and statistics selection
- Carefully consider trade-offs in incident & prevalent user designs

References & Resources