Experimental Identification of Causal Mechanisms

Kosuke Imai\textsuperscript{1}  Dustin Tingley\textsuperscript{2}  Teppei Yamamoto\textsuperscript{3}

\textsuperscript{1}Princeton University
\textsuperscript{2}Harvard University
\textsuperscript{3}Massachusetts Institute of Technology

March 14, 2012
Royal Statistical Society, London
Causal inference is a central goal of most scientific research.

Experiments as gold standard for estimating causal effects

A major criticism of experimentation:

*it can only determine whether the treatment causes changes in the outcome, but not how and why*

Experiments merely provide a black box view of causality.

But, scientific theories are all about causal mechanisms.

Knowledge about causal mechanisms can also improve policies.

Key Challenge: How can we *design* and analyze experiments to identify causal mechanisms?
Overview of the Talk

- Show the limitation of a common approach
- Consider alternative experimental designs

- What is a minimum set of assumptions required for identification under each design?
- How much can we learn without the key identification assumptions under each design?

- Identification of causal mechanisms is possible but difficult
- Distinction between design and statistical assumptions
- Roles of creativity and technological developments

- Illustrate key ideas through recent social science research
What is a causal mechanism?
Cochran (1957)’s example: soil fumigants increase farm crops by reducing eel-worms
Political science example: incumbency advantage
Causal mediation analysis

Quantities of interest: Direct and indirect effects
Fast growing methodological literature
Alternative definition: causal components (Robins; VanderWeele)
Binary treatment: $T_i \in \{0, 1\}$

Mediator: $M_i \in \mathcal{M}$

Outcome: $Y_i \in \mathcal{Y}$

Observed pre-treatment covariates: $X_i \in \mathcal{X}$

Potential mediators: $M_i(t)$ where $M_i = M_i(T_i)$

Potential outcomes: $Y_i(t, m)$ where $Y_i = Y_i(T_i, M_i(T_i))$

Fundamental problem of causal inference (Rubin; Holland):

*Only one potential value is observed*

1. If $T_i = 1$, then $M_i(1)$ is observed but $M_i(0)$ is not
2. If $T_i = 0$ and $M_i(0) = 0$, then $Y_i(0, 0)$ is observed but $Y_i(1, 0)$, $Y_i(0, m)$, and $Y_i(1, m)$ are not when $m \neq 0$
Defining and Interpreting Indirect Effects

- Total causal effect:
  \[ \tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \]

- Indirect (causal mediation) effects (Robins and Greenland; Pearl):
  \[ \delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \]

- Change \( M_i(0) \) to \( M_i(1) \) while holding the treatment constant at \( t \)
  Effect of a change in \( M_i \) on \( Y_i \) that would be induced by treatment

- Fundamental problem of causal mechanisms:
  \[ \text{For each unit } i, \ Y_i(t, M_i(t)) \text{ is observable but} \]
  \[ \ Y_i(t, M_i(1 - t)) \text{ is not even observable} \]
Defining and Interpreting Direct Effects

- **Direct effects:**

  \[ \zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \]

- Change \( T_i \) from 0 to 1 while holding the mediator constant at \( M_i(t) \)

- Causal effect of \( T_i \) on \( Y_i \), holding mediator constant at its potential value that would be realized when \( T_i = t \)

- **Total effect = indirect effect + direct effect:**

  \[ \tau_i = \delta_i(t) + \zeta_i(1 - t) \]

  \[ = \delta_i + \zeta_i \]

  where the second equality assumes \( \delta_i(0) = \delta_i(1) \) and \( \zeta_i(0) = \zeta_i(1) \)
Mechanisms, Manipulations, and Interactions

Mechanisms
- Indirect effects:
  \[ \delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \]
- Counterfactuals about treatment-induced mediator values

Manipulations
- Controlled direct effects:
  \[ \xi_i(t, m, m') \equiv Y_i(t, m) - Y_i(t, m') \]
- Causal effect of directly manipulating the mediator under \( T_i = t \)

Interactions
- Interaction effects:
  \[ \xi(1, m, m') - \xi(0, m, m') \neq 0 \]
- Doesn’t imply the existence of a mechanism
Single Experiment Design

Assumption Satisfied

- Randomization of treatment

\[ \{ Y_i(t, m), M_i(t') \} \perp T_i, | X_i = x \]

Key Identifying Assumption

- Sequential Ignorability:

\[ Y_i(t', m) \perp M_i | T_i = t, X_i = x \]

- Selection on pre-treatment observables
- Unmeasured pre-treatment confounders
- Measured and unmeasured post-treatment confounders

1) Randomize treatment
2) Measure mediator
3) Measure outcome
Identification under the Single Experiment Design

- Sequential ignorability yields nonparametric identification

\[ \bar{\delta}(t) = \int \int \mathbb{E}(Y_i \mid M_i, T_i = t, X_i) \{dP(M_i \mid T_i = 1, X_i) - dP(M_i \mid T_i = 0, X_i)\} dP(X_i) \]

- Linear structural equation modeling (a.k.a. Baron-Kenny)
- Alternative assumptions: Robins, Pearl, Petersen et al., VanderWeele, and many others

- Sequential ignorability is an untestable assumption
- **Sensitivity analysis**: How large a departure from sequential ignorability must occur for the conclusions to no longer hold?

- But, sensitivity analysis does not solve the problem
A Typical Psychological Experiment

- Brader *et al.*: media framing experiment
- **Treatment**: Ethnicity (Latino vs. Caucasian) of an immigrant
- **Mediator**: anxiety
- **Outcome**: preferences over immigration policy

- Single experiment design with statistical mediation analysis
- **Emotion**: difficult to directly manipulate
- **Sequential ignorability assumption** is not credible
- **Possible confounding**
How much can we learn without sequential ignorability?

Sharp bounds on indirect effects (Sjölander):

\[
\max \left\{ \begin{array}{c}
-P_{001} - P_{011} \\
-P_{011} - P_{010} - P_{110} \\
-P_{000} - P_{001} - P_{100}
\end{array} \right\} \leq \hat{\delta}(1) \leq \min \left\{ \begin{array}{c}
P_{101} + P_{111} \\
P_{010} + P_{110} + P_{111} \\
P_{000} + P_{100} + P_{101}
\end{array} \right\}
\]

\[
\max \left\{ \begin{array}{c}
-P_{100} - P_{110} \\
-P_{011} - P_{111} - P_{110} \\
-P_{001} - P_{101} - P_{100}
\end{array} \right\} \leq \hat{\delta}(0) \leq \min \left\{ \begin{array}{c}
P_{000} + P_{010} \\
P_{011} + P_{111} + P_{010} \\
P_{000} + P_{001} + P_{101}
\end{array} \right\}
\]

where \( P_{ymt} = \Pr(Y_i = y, M_i = m \mid T_i = t) \)

The sign is not identified
Can we design experiments to better identify causal mechanisms?

Perfect manipulation of the mediator:
1. Parallel Design
2. Crossover Design

Imperfect manipulation of the mediator:
1. Parallel Encouragement Design
2. Crossover Encouragement Design

Implications for designing observational studies
The Parallel Design

- **No manipulation effect assumption**: The manipulation has no direct effect on outcome other than through the mediator value.

- Running two experiments in parallel:

  - **Experiment 1**
    1) Randomize treatment
    2) Measure mediator
    3) Measure outcome

  - **Experiment 2**
    1) Randomize treatment
    2) Randomize mediator
    3) Measure outcome
Identification under the Parallel Design

- Difference between manipulation and mechanism:

<table>
<thead>
<tr>
<th>Prop.</th>
<th>$M_i(1)$</th>
<th>$M_i(0)$</th>
<th>$Y_i(t, 1)$</th>
<th>$Y_i(t, 0)$</th>
<th>$\delta_i(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- $E(M_i(1) - M_i(0)) = E(Y_i(t, 1) - Y_i(t, 0)) = 0.2$, but $\bar{\delta}(t) = -0.2$

- Is the randomization of mediator sufficient? No
- The no interaction assumption (Robins) yields point identification

$$Y_i(1, m) - Y_i(1, m') = Y_i(0, m) - Y_i(0, m')$$

- Must hold at the unit level but indirect tests are possible
- Implication: analyze a group of homogeneous units
Identification under the Parallel Design

- Is the randomization of mediator sufficient? No!
- Sharp bounds: Binary mediator and outcome
- Use of linear programming (Balke and Pearl):
  - Objective function:
    \[
    \mathbb{E}\{Y_i(1, M_i(0))\} = \sum_{y=0}^{1} \sum_{m=0}^{1} (\pi_{ym1} + \pi_{y1m1})
    \]
    where \(\pi_{y1y0m1m0} = \Pr(Y_i(1, 1) = y_1, Y_i(1, 0) = y_0, M_i(1) = m_1, M_i(0) = m_0)\)
  - Constraints implied by \(\Pr(Y_i = y, M_i = m \mid T_i = t, D_i = 0)\), \(\Pr(Y_i = y \mid M_i = m, T_i = t, D_i = 1)\), and the summation constraint
- More informative than those under the single experiment design
- Can sometimes identify the sign of average direct/indirect effects
Why study brain?: Social scientists’ search for causal mechanisms underlying human behavior

- Psychologists, economists, and even political scientists

Question: What mechanism links low offers in an ultimatum game with “irrational” rejections?

- A brain region known to be related to fairness becomes more active when unfair offer received (single experiment design)

Design solution: manipulate mechanisms with TMS

- Knoch et al. use TMS to manipulate — turn off — one of these regions, and then observes choices (parallel design)
The Parallel Encouragement Design

- Direct manipulation of mediator is often difficult
- Even if possible, the violation of no manipulation effect can occur
- Need for indirect and subtle manipulation

- Randomly encourage units to take a certain value of the mediator
- Instrumental variables assumptions (Angrist et al.):
  1. Encouragement does not discourage anyone
  2. Encouragement does not directly affect the outcome

- Not as informative as the parallel design
- Sharp bounds on the average “complier” indirect effects can be informative
A Numerical Example

Based on the marginal distribution of a real experiment
The Crossover Design

Basic Idea

- Want to observe $Y_i(1 - t, M_i(t))$
- Figure out $M_i(t)$ and then switch $T_i$ while holding the mediator at this value
- Subtract direct effect from total effect

Key Identifying Assumptions

- No Manipulation Effect
- No Carryover Effect: For $t = 0, 1$, 
  \[ \mathbb{E}\{Y_{i1}(t, M_i(t))\} = \mathbb{E}\{Y_{i2}(t, m)\} \]
  if $m = M_i(t)$
- Not testable, longer “wash-out” period
Example from Labor Economics


- Treatment: Black vs. White names on CVs
- Mediator: Perceived qualifications of applicants
- Outcome: Callback from employers

- Estimand: Direct effects of (perceived) race $\rightarrow$ overt racism
- Would Jamal get a callback if his name were Greg but his qualifications stayed the same?

- Round 1: Send Jamal’s actual CV and record the outcome
- Round 2: Send his CV as Greg and record the outcome

- Assumptions:
  1. No manipulation: potential employers are unaware
  2. Carryover effect: send resumes to different (randomly matched) employers at the same time
The Crossover Encouragement Design

**Key Identifying Assumptions**
- Encouragement doesn’t discourage anyone
- No Manipulation Effect
- No Carryover Effect

**Identification Analysis**
- Identify indirect effects for “compliers”
- No carryover effect assumption is indirectly testable (unlike the crossover design)
Comparing Alternative Designs

- No manipulation
  - Single experiment: sequential ignorability

- Direct manipulation
  - Parallel: no manipulation effect, no interaction effect
  - Crossover: no manipulation effect, no carryover effect

- Indirect manipulation
  - Encouragement: no manipulation effect, monotonicity, no interaction effect
  - Crossover encouragement: no manipulation effect, monotonicity, no carryover effect
Implications for the Design of Observational Studies

- Use of “natural experiments” in the social sciences
- Attempts to “replicate” experiments in observational studies
- Political science literature on **incumbency advantage**
- During 70s and 80s, the focus is on estimation of causal effects
- Positive effects, growing over time
- Last 20 years, search for causal mechanisms

- How large is the “**scare-off/quality effect**”?
- Use of cross-over design (Levitt and Wolfram)
  1. 1st Round: two non-incumbents in an open seat
  2. 2nd Round: same candidates with one being an incumbent

- Assumptions
  1. Challenger quality (mediator) stays the same
  2. First election does not affect the second election
Another Incumbency Advantage Example

- Redistricting as natural experiments (Ansolabehere et al.)
  1. 1st Round: incumbent in the old part of the district
  2. 2nd Round: incumbent in the new part of the district
- Assumption: No interference between the old and new parts of the district
Concluding Remarks

- Identification of causal mechanisms is difficult but is possible
- Additional assumptions are required

Five strategies:
1. Single experiment design
2. Parallel design
3. Crossover design
4. Parallel encouragement design
5. Crossover encouragement design

Statistical assumptions: sequential ignorability, no interaction
Design assumptions: no manipulation, no carryover effect

Experimenters’ creativity and technological development to improve the validity of these design assumptions