Uncovering Causal Mechanisms: Mediation Analysis and Surrogate Indices

Raj Chetty Kosuke Imai

Harvard University

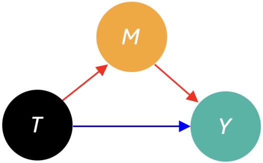
2025 NBER Methods Lecture

Mediation Analysis: Identifying Mechanisms Underlying Treatment Effects on Primary Outcomes

Part I. Introduction to Mediation

Causal Mechanism as Direct and Indirect Effects

- Directed Acyclic Graph (DAG; Pearl, 2000)
  - $T \in \mathcal{T} = \{0, 1\}$: treatment
  - $M \in \mathcal{M}$: mediator (mechanism variable)
  - $Y \in \mathcal{Y}$: observed outcome



- Direct effect: Effect of T on Y while holding M constant
- Indirect effect: Effect of T on Y through M

- DAG = Nonparametric Structural Equation Model (NPSEM)

$$Y = f_Y(M, T, \epsilon)$$
$$M = f_M(T, \eta)$$

where ϵ and η are i.i.d. and are usually omitted from DAG

Controlled Direct Effect (CDE)

- $Y(t, m) \in \mathcal{Y}$: potential outcome when $T = t$ and $M = m$
- Definition

$$\begin{aligned} \text{Individual:} \quad \text{CDE}_i(m) &:= Y_i(1, m) - Y_i(0, m) \\ \text{Average:} \quad \overline{\text{CDE}}(m) &:= \mathbb{E}[Y(1, m) - Y(0, m)] \end{aligned}$$

  for a given mediator value $m \in \mathcal{M}$

- Interpretation
  - direct effect of treatment while holding the mediator constant at m
  - effect of joint intervention on T and M

- If M fully captures treatment effect, CDEs will be zero for all m
- Potential interaction effects:

$$\text{CDE}_i(m) \neq \text{CDE}_i(m') \quad \text{for some } i \text{ and } m \neq m'$$

Natural Indirect Effect (NIE)

- Definition (Robins and Greenland, 1992; Pearl, 2001)

$$\text{Individual:} \quad \text{NIE}_i(t) \;:=\; Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$
$$\text{Average:} \quad \overline{\text{NIE}}(t) \;:=\; \mathbb{E}[Y(t, M(1)) - Y(t, M(0))]$$

- Interpretation
  - effect of change in M on Y induced by T
  - change M from $M(0)$ to $M(1)$ while holding T at $t = 0$ or $t = 1$
  - zero treatment effect on M implies zero NIE

- Represents the causal effect of T on Y through M
- Complete mediation $\rightsquigarrow$ $\text{NIE}_i = \text{TE}_i := Y_i(1, M_i(1)) - Y_i(0, M_i(0))$

Treatment Effect Decomposition

- Natural direct effect (NDE):

$$\begin{aligned} \text{Indiviual:} \quad \text{NDE}_i(t) &:= Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \\ \text{Average:} \quad \overline{\text{NDE}}(t) &:= \mathbb{E}[Y(1, M(t)) - Y(0, M(t))] \end{aligned}$$

  - change T from 0 to 1 while holding M constant at $M(t)$
  - causal effect of T on Y, holding M constant at its potential value that would be realized when $T = t$
- Represents all mechanisms other than through M
  - Complete mediation $\rightsquigarrow$ $\text{NDE}_i(t) = 0$
  - No mediation $\rightsquigarrow$ $\text{NDE}_i = \text{TE}_i$
- Effect decomposition:

$$\begin{aligned} \underbrace{Y_i(1, M_i(1)) - Y_i(0, M_i(0))}_{=\text{total effect } (\text{TE}_i)} &= \text{NIE}_i(t) + \text{NDE}_i(1-t) \\ &= \frac{1}{2}\sum_{t=0}^{1}\{\text{NIE}_i(t) + \text{NDE}_i(t)\} \end{aligned}$$

Gender Bias and Educational Attainment (Chen et al. 2019)

- Data on Taiwanese families
  - Y: educational attainment of the oldest child who is female
  - T: gender of the second oldest child
  - M: number of siblings

- Gender bias
  - Direct effect: having a brother takes away resources from a female child
  - Indirect effect: having a brother leads to a smaller number of siblings and hence more resources
  - Direct and indirect effects may have opposite signs

- Causal effects of interest
  - CDE: effect of having a brother while keeping sibling size constant at a fixed value, e.g., 2
  - NDE: effect of having a brother while keeping sibling size constant at a value that would result, e.g., if the second child were male
  - NIE: effect of having a brother through sibling size

Take-aways I

- Causal mechanism
  - how and why (not just whether) treatment affects outcome
  - understanding of causal structure (DAG $=$ NPSEM)

- Causal quantities of interest
  - Controlled direct effect (CDE)
  - Natural direct and indirect effects (NDE, NIE)
  - Effect decomposition: $\text{TE} = \text{NDE} + \text{NIE}$
  - No similar decomposition for CDE
  - Complete mediation: $\text{CDE} = \text{NDE} = 0$ and $\text{NIE} = \text{TE}$
  - No mediation: $\text{NIE} = 0$ and $\text{NDE} = \text{TE}$

Part II. Mediation Analysis Under Pretreatment Confounding

Linear Structural Equation Model (LSEM)

- Let's build some intuition with LSEM
- Homogeneous effects without interaction:

$$
\begin{aligned}
Y_i &= \alpha_Y + \beta_Y T_i + \gamma_Y M_i + \epsilon_i \\
M_i &= \alpha_M + \beta_M T_i + \eta_i
\end{aligned}
$$

  - $\overline{\text{CDE}}(m) = \overline{\text{NDE}}(t) = \beta_Y$ for any m and t
  - $\overline{\text{NIE}}(t) = \beta_M \times \gamma_Y$ for any t
  - CDE and NDE are identical

- Homogeneous effects with interaction:

$$
Y_i = \alpha_Y + \beta_Y T_i + \gamma_Y M_i + \delta_Y T_i M_i + \epsilon_i
$$

  - $\overline{\text{CDE}}(m) = \beta_Y + m\delta_Y$
  - $\overline{\text{NDE}}(t) = \beta_Y + \delta_Y(\alpha_M + t\beta_M)$
  - $\overline{\text{NIE}}(t) = \beta_M \times \gamma_Y + t\beta_M \times \delta_Y$
  - CDE is different from NDE

LSEM with Heterogeneous Effects and Interaction

- Model

$$Y_i = \alpha_Y + \beta_Y^{(i)} T_i + \gamma_Y^{(i)} M_i + \delta_Y^{(i)} T_i M_i + \epsilon_i$$
$$M_i = \alpha_M + \beta_M^{(i)} T_i + \eta_i$$

  - $\overline{\text{CDE}}(m) = \bar{\beta}_Y + m\bar{\delta}_Y$ where $\bar{\beta}_Y = \mathbb{E}[\beta_Y^{(i)}]$ and $\bar{\delta}_Y = \mathbb{E}[\delta_Y^{(i)}]$
  - $\overline{\text{NDE}}(t) = \bar{\beta}_Y + \alpha_M \times \bar{\delta}_Y + \mathbb{E}[\delta_Y^{(i)}(t\beta_M^{(i)} + \eta_i)]$
  - $\overline{\text{NIE}}(t) = \mathbb{E}[\beta_M^{(i)} \times (\gamma_Y^{(i)} + t\delta_Y^{(i)})]$

- Heterogeneous effects may be correlated with one another
  - For example, $\mathbb{E}[\beta_M^{(i)} \times \gamma_Y^{(i)}] \neq \bar{\beta}_M \times \bar{\gamma}_Y$
  - Possible to have $\bar{\beta}_M, \bar{\gamma}_Y > 0$ but $\mathbb{E}[\beta_M^{(i)} \times \gamma_Y^{(i)}] < 0$ or vice versa

- $\bar{\beta}_M$, $\bar{\gamma}_Y$, $\bar{\delta}_Y$, etc. are identifiable under exogeneity
- But, $\mathbb{E}[\beta_M^{(i)} \times \gamma_Y^{(i)}]$, $\mathbb{E}[\beta_M^{(i)} \times \delta_Y^{(i)}]$, etc. are unidentifiable
- This is essentially a problem of unobserved pre-treatment confounding

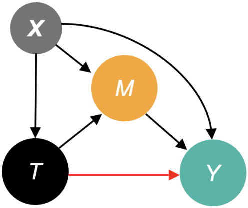
Identification of CDE with Pre-treatment Confounding

- Assumptions:
  1. Unconfoundedness

$$\{Y_i(t,m), M_i(t)\}_{t,m} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i = \boldsymbol{x}$$
$$\{Y_i(t,m)\}_m \perp\!\!\!\perp M_i \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}$$

  2. Overlap

$$P(T_i = t \mid \boldsymbol{X}_i = \boldsymbol{x}) > 0$$
$$P(M_i = m \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}) > 0$$



- Identification:

$$\overline{\text{CDE}}(m) = \sum_{\boldsymbol{X}} \left(\mathbb{E}[Y \mid T = 1, M = m, \boldsymbol{X}] - \mathbb{E}[Y \mid T = 0, M = m, \boldsymbol{X}]\right) P(\boldsymbol{X})$$

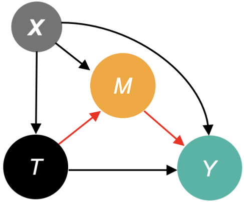
Identification of NDE/NIE with Pretreatment Confounding

- Replace the following assumption

$$\{Y_i(t,m)\}_m \perp\!\!\!\perp \underbrace{M_i}_{=M_i(t)} \mid T_i = t, \mathbf{X}_i = \mathbf{x}$$

with the cross-world independence

$$\{Y_i(t',m)\}_{t',m} \perp\!\!\!\perp M_i(t) \mid T_i = t, \mathbf{X}_i = \mathbf{x}$$



- Additional conditional independence between $Y_i(t',m)$ and $M_i(t)$
- Identification (Imai et al. 2010)

$$\overline{\text{NDE}}(t) = \sum_{M,\mathbf{X}} (\mathbb{E}[Y \mid M, T=1, \mathbf{X}] - \mathbb{E}[Y \mid M, T=0, \mathbf{X}]) \times P(M \mid T=t, \mathbf{X}) P(\mathbf{X})$$

$$\overline{\text{NIE}}(t) = \sum_{M,\mathbf{X}} \mathbb{E}[Y \mid M, T=t, \mathbf{X}] \times \{P(M \mid T=1, \mathbf{X}) - P(M \mid T=0, \mathbf{X})\} P(\mathbf{X})$$

Experimental Identification (Imai et al. 2013)

- Parallel design
  1. Randomize T and observe M and Y
  2. Randomize T and M and observe Y
- We can identify $P(M(t))$, $P(Y(t, M(t))$, and $P(Y(t, m))$
- CDE is identified
- NDE/NIE is still not identifiable:
  - randomization cannot break correlation between $Y(t', m)$ and $M(t)$
  - partial identification: sharp bounds contain zero
- Crossover design
  1. Randomize T and observe M and Y
  2. On the same sample, change T to the opposite condition while holding M at the same value and observe Y
- $Y(t, M(t))$, $M(t)$, and $Y(1 - t, M(t))$ are observable
- Additional assumption: no carryover effects
- NDE/NIE is identifiable

No Interaction Assumption

- No individual-level interaction

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

  - $\text{NDE}_i(t) = \text{CDE}_i(m) = \text{CDE}_i$
  - $\overline{\text{NDE}}(t) = \overline{\text{CDE}}(m) = \overline{\text{CDE}}$
  - $\overline{\text{NIE}}(t) = \text{ATE} - \overline{\text{NDE}}$

- Testable implication:

$$\mathbb{E}[Y_i(1, m) - Y_i(0, m) \mid \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[Y_i(1, m') - Y_i(0, m') \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

  for all $\boldsymbol{x}$

- NDE/NIE is identifiable so long as CDE can be identified
- Experimental identification, and identification with pretreatment and posttreatment confounding are all possible

Estimation of Natural Direct and Indirect Effects

- Recall the identification formula (NIE)

$$\overline{\text{NIE}}(t) = \sum_{M,\boldsymbol{X}} \mathbb{E}[Y \mid M, T = t, \boldsymbol{X}] \times \{P(M \mid T = 1, \boldsymbol{X}) - P(M \mid T = 0, \boldsymbol{X})\} P(\boldsymbol{X})$$

1. predict M given each treatment value: $\{M_i(1),\ M_i(0)\}$
2. predict Y by first setting $T_i = t$ and $M_i = M_i(0)$, and then $T_i = t$ and $M_i = M_i(1)$: $\{Y_i(t, M_i(0)), Y_i(t, M_i(1))\}$
3. compute the average difference between two predicted outcomes

- Estimation of NDE is similar

$$\overline{\text{NDE}}(t) = \sum_{M,\boldsymbol{X}} (\mathbb{E}[Y \mid M, T = 1, \boldsymbol{X}] - \mathbb{E}[Y \mid M, T = 0, \boldsymbol{X}]) \times P(M \mid T = t, \boldsymbol{X}) P(\boldsymbol{X})$$

- One can also do: $\overline{\text{NDE}}(t) = \text{ATE} - \overline{\text{NIE}}(1 - t)$

Weighting Methods for NDE and NIE

- Three weighting formulae:

$$
\begin{aligned}
\mathbb{E}[Y(t, M(t'))] &= \mathbb{E}\Bigg[\underbrace{\frac{1\{T = t'\}}{\Pr(T = t' \mid \boldsymbol{X})}}_{\text{weighting to get } P(M(t')|\boldsymbol{X})} \times \mathbb{E}[Y \mid M, T = t, \boldsymbol{X}] \Bigg] \\
&= \mathbb{E}\Bigg[\underbrace{\frac{1\{T = t\}}{\Pr(T = t \mid \boldsymbol{X}_i)}}_{\text{treatment weighting}} \times \underbrace{\frac{P(M \mid T = t', \boldsymbol{X})}{P(M \mid T_i = t, \boldsymbol{X}_i)}}_{\text{mediator weighting}} \times Y \Bigg] \\
&= \mathbb{E}\left[\frac{1\{T = t\}}{\Pr(T = t \mid M, \boldsymbol{X})} \times \frac{\Pr(T = t' \mid M, \boldsymbol{X})}{\Pr(T = t' \mid \boldsymbol{X})} \times Y \right]
\end{aligned}
$$

  - The third expression follows from Bayes rule
  - Useful when the mediator is high-dimensional

- Multiply-robust semiparametric estimator (Tchetgen Tchetgen and Shpitser, 2012); Double machine learning (Farbmacher et al. 2022)

Sensitivity Analysis

- Examine the robustness of empirical findings to the violation of untestable assumptions
- How large a departure from the key identification assumption must occur for the conclusions to no longer hold?
- Potential existence of unobserved pretreatment confounding (T is assumed to be unconfounded)

$$\{Y_i(t', m)\}_{t',m} \not\perp\!\!\!\perp M_i(t) \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}$$

- Recall LSEM (or more generally, additive semiparametric model)

$$Y_i = \alpha_Y + \beta_Y T_i + \gamma_Y M_i + \underbrace{\lambda_\epsilon U_i + \tilde{\epsilon}_i}_{=\epsilon_i}$$

$$M_i = \alpha_M + \beta_M T_i + \underbrace{\lambda_\eta U_i + \tilde{\eta}_i}_{=\eta_i}$$

- How much does U_i have to matter for the results to go away?

Sensitivity Parameters

- R^2 parameterization
  1. Proportion of previously unexplained variance explained by U_i

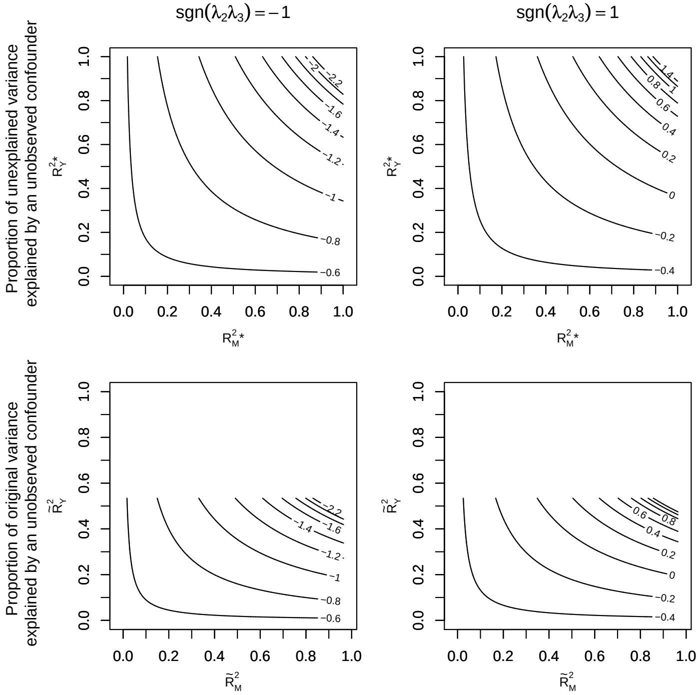
$$R_M^{2*} \equiv \frac{\mathbb{V}(\lambda_\eta U_i)}{\mathbb{V}(\eta_i)} \quad \text{and} \quad R_Y^{2*} \equiv \frac{\mathbb{V}(\lambda_\epsilon U_i)}{\mathbb{V}(\epsilon_i)}$$

  2. Proportion of original variance explained by U_i

$$\widetilde{R}_M^2 \equiv \frac{\mathbb{V}(\lambda_\eta U_i)}{\mathbb{V}(M_i)} \quad \text{and} \quad \widetilde{R}_Y^2 \equiv \frac{\mathbb{V}(\lambda_\epsilon U_i)}{\mathbb{V}(Y_i)}$$

- We also need to specify the direction of effects:

$$\text{sgn}(\lambda_\eta \lambda_\epsilon) = \begin{cases} 1 & \text{if same direction} \\ -1 & \text{if opposite directions} \end{cases}$$


sgn(λ2λ3) = −1
sgn(λ2λ3) = 1
Proportion of unexplained variance explained by an unobserved confounder
Proportion of original variance explained by an unobserved confounder
R_Y^2*
R_M^2*
$\tilde{R}_Y^2$
$\tilde{R}_M^2$
−0.6
−0.8
−1
−1.2
−1.4
−1.6
−2
−2.2
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.4

Gender Bias Application: Standard Mediation Analysis

- The original analysis fits LSEM with interaction

$$Y_i = \alpha_Y + \beta_Y T_i + \gamma_Y M_i + \delta_Y T_i M_i + \boldsymbol{\xi}_Y^\top \boldsymbol{X}_i + \epsilon_i$$
$$M_i = \alpha_M + \beta_M T_i + \boldsymbol{\xi}_M^\top \boldsymbol{X}_i + \eta_i$$

  - Y_i: university admission
  - T_i: the second child is male
  - M_i: sibling size is greater than two

- Estimates:

$\widehat{\text{ATE}}$	0.0020 (0.0013)
$\widehat{\overline{\text{CDE}}(\overline{M})}$	−0.0010 (0.0014)
$\widehat{\overline{\text{NDE}}(1)}$	−0.0001 (0.0014)
$\widehat{\overline{\text{NIE}}(0)}$	0.0022 (0.0005)

- Also, fits a random coefficient model to address heterogeneity
- Sensitivity analysis based on a semiparametric random coefficient model (Imai and Yamamoto, 2013)

Take-aways II

- Linear structural equation model
  - two key assumptions beyond exogeneity:
    1. homogeneous effects
    2. no interaction
  - CDE $=$ NDE under those assumptions
  - Relaxing these assumptions lead to different interpretations and identification issues

- Nonparametric identification analysis under pretreatment confounding
  - CDE is identifiable under standard exogeneity
  - NDE/NIE requires cross-world independence
  - alternatively, CDE $=$ NDE if we assume no individual-level interaction

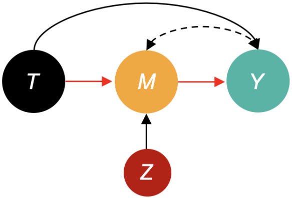
- Difficulty of identification
  - even when M is randomized, NIE/NDE are unidentifiable
  - sensitivity analysis plays an important role for assessing robustness

Part III. Coping with Identification Difficulties

Instrumenting the Mediator

- Instrument: Z_i
- Mediator: $M_i(t, z)$
- Exclusion restriction

$$Y_i(t, m, z) = Y_i(t, m)$$



- NPSEM:

$$\begin{aligned} Y &= f_Y(M, T, \epsilon) \\ M &= f_M(T, Z, \eta) \end{aligned} \quad \text{where} \quad \epsilon \not\perp\!\!\!\perp \eta$$

- If M and Z are continuous, we can use the control function approach (Imbens and Newey, 2009)
  1. Independence: $Z \perp\!\!\!\perp (\epsilon, \eta)$
  2. Monotonicity: η is a continuous scalar variable with its CDF and $f_M(\cdot, \cdot, \eta)$ being strictly monotonic in η
- Then, $(M, T) \perp\!\!\!\perp \epsilon \mid C$ where $C = F_{M \mid T, Z}(T, Z) = F_\eta(\eta)$
  - Recall the control function approach to 2SLS
  - Regress Y on M, T and the first stage residual $\hat{\eta}$
- Extension: an additional instrument for T (Florich and Huber, 2017)

Gender Bias Application: IV Analysis

- Instrument Z: twinning at the second birth

$$M_i = \alpha_M + \beta_M T_i + \zeta_M Z_i + \lambda_M T_i Z_i + \boldsymbol{\xi}_M^\top \boldsymbol{X}_i + \eta_i$$

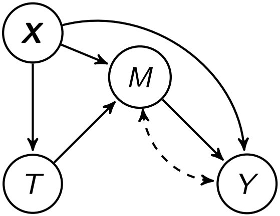
- Assumptions:
  - exogenous instrument: twinning is random conditional on $\boldsymbol{X}$
  - exclusion restriction: twinning affects Y only through M

- Findings:

	Standard analysis	**IV analysis**
$\widehat{\text{ATE}}$	0.0020 (0.0013)	0.0021 (0.0013)
$\widehat{\overline{\text{CDE}}(\overline{M})}$	−0.0010 (0.0014)	−0.0092 (0.0061)
$\widehat{\overline{\text{NDE}}(1)}$	−0.0001 (0.0014)	−0.0203 (0.0106)
$\overline{\text{NIE}}(0)$	0.0022 (0.0005)	0.0224 (0.0105)

Complete Mediation Analysis (Kwon and Roth 2024)

- Complete mediation: $Y_i(t, m) = Y_i(m)$
- Assumption: No unobserved confounding between T and M and between T and Y
- Possible unobserved confounding between M and Y



- Under monotonicity $M_i(1) \geq M_i(0)$ (in the binary mediator case), we can use the following test of instrumental validity

$$P(Y, M = 0 \mid T = 0, \boldsymbol{X}) \geq P(Y, M = 0 \mid T = 1, \boldsymbol{X})$$
$$P(Y, M = 1 \mid T = 1, \boldsymbol{X}) \geq P(Y, M = 1 \mid T = 0, \boldsymbol{X})$$

- Randomized experiment: test of complete mediation
- Observational study: unobserved confounding between T and Y can also lead to the rejection of the null hypothesis

Implicit Mediation

- What if we want to avoid the untestable assumptions at all costs?
- What can we infer from ATE_M and ATE_Y that are identifiable without such assumptions?

Table 1. Possible Implicit-Mediation Findings

Result	Inference	Rationale
X affects M and Y	M may be a mediator.	X appears to influence M, and this effect seems to coincide with a change in Y, as would be expected if M were a mediator.
X affects M but not Y	M appears not to be a mediator.	Although X affects M, this effect seems not to have any consequences for Y.
X affects Y but not M	Some variable other than M may be a mediator.	X appears to have no effect on M, which means that X's apparent effect on Y is not due to changes in M.
X affects neither M nor Y	There seem to be no indirect pathways from X to Y through M or other mediators.	X seems not to set in motion any causal effects.

Bullock and Green (2021)

Identification Analysis of Implicit Mediation

- Questions:
  1. Does $\text{ATE}_M = 0$ imply $\overline{\text{NIE}} = 0$ and/or $\overline{\text{NDE}} \neq 0$?
  2. Does $\text{ATE}_M > 0$ and $\text{ATE}_Y > 0$ imply $\overline{\text{NIE}} > 0$?
- No! Recall even the no-assumption bounds from the parallel experiment design always contain zero
- The decomposition under a binary mediator:

$$\overline{\text{NIE}}(t) = \underbrace{\mathbb{E}[Y_i(t,1) - Y_i(t,0) \mid M(1) = 1, M(0) = 0]}_{\text{ATE of } M \text{ on } Y \text{ for compliers}} \cdot p_{10} - \underbrace{\mathbb{E}[Y_i(t,1) - Y_i(t,0) \mid M(1) = 0, M(0) = 1]}_{\text{ATE of } M \text{ on } Y \text{ for defiers}} \cdot p_{01}$$

  where $p_{m_1 m_0} = \Pr(M(1) = m_1, M(0) = m_0)$
- Cross-world assumption or homogeneity assumption leads to the usual product estimator

$$\overline{\text{NIE}}(t) = \underbrace{\mathbb{E}[Y_i(t,1) - Y_i(t,0)]}_{=\text{ATE of } M \text{ on } Y} \times \underbrace{(p_{10} - p_{01})}_{=\text{ATE}_M}$$

Identification under Monotonicity

(Blackwell et al. 2024; Kwon and Roth 2024)

- Monotonicity assumption (no defier) yields:

$$\overline{\text{NIE}}(t) = \mathbb{E}[Y_i(t,1) - Y_i(t,0) \mid M(1) = 1, M(0) = 0] \cdot p_{10}$$

- Sharp bounds

$$\max\{-\text{ATE}_M, -q_{1-t,t|t}\} \leq \overline{\text{NIE}}(t) \leq \min\{\text{ATE}_M, q_{tt|t}\}$$

where $q_{ym|t} = \Pr(Y = y, M = m \mid T = t)$

- Two fundamental difficulties remain:
  1. effect heterogeneity
  2. endogeneity of mediator
- Even under an additional assumption of $\mathbb{E}[Y(t,1) - Y(t,0)] > 0$, the sharp bounds still contain zero

Take-aways III

- Instrumental variable approach
  - addressing the endogeneity problem
  - the instrument must be exogeneous
  - exclusion restriction needs to be satisfied
  - nonparametric estimation is possible

- Complete mediation
  - hypothesis testing approach
  - no need to assume the exogeneity of mediator
  - no unobserved confounding between T and Y (satisfied in RCT)

- Implicit mediation
  - an attempt to sidestep assumptions
  - not informative even about the signs of NIE/NDE
  - monotonicity is not sufficient

Part IV. Mediation Analysis under Posttreatment Confounding

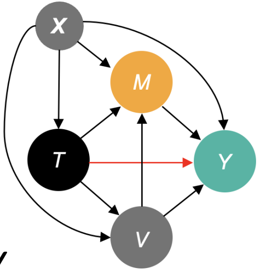
Identification of CDE with Posttreatment Confounding

- Replace the following assumption

$$\{Y_i(t,m)\}_m \perp\!\!\!\perp M_i \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x},$$

with

$$\{Y_i(t,m)\}_m \perp\!\!\!\perp M_i \mid \boldsymbol{V}_i = \boldsymbol{v}, T_i = t, \boldsymbol{X}_i = \boldsymbol{x}$$



- Post-treatment bias: cannot simply control for $\boldsymbol{V}$

$$\overline{\text{CDE}}(m) \neq \sum_{\boldsymbol{X},\boldsymbol{V}} (\mathbb{E}[Y \mid T=1, M=m, \boldsymbol{X}, \boldsymbol{V}] - \mathbb{E}[Y \mid T=0, M=m, \boldsymbol{X}, \boldsymbol{V}])P(\boldsymbol{X}, \boldsymbol{V})$$

- Identification: model $\boldsymbol{V}$ given T and $\boldsymbol{X}$

$$\overline{\text{CDE}}(m) = \sum_{\boldsymbol{X},\boldsymbol{V}} \{\mathbb{E}[Y \mid T=1, M=m, \boldsymbol{X}, \boldsymbol{V}]P(\boldsymbol{V} \mid T=1, \boldsymbol{X}) - \mathbb{E}[Y \mid T=0, M=m, \boldsymbol{X}, \boldsymbol{V}]P(\boldsymbol{V} \mid T=0, \boldsymbol{X})\}\, P(\boldsymbol{X})$$

Estimation of Controlled Direct Effects

1. Directly use the identification formula

$$\bar{\xi}(m) = \sum_{\boldsymbol{X},\boldsymbol{V}} \{\mathbb{E}[Y \mid T=1, M=m, \boldsymbol{X}, \boldsymbol{V}] P(\boldsymbol{V} \mid T=1, \boldsymbol{X}) - \mathbb{E}(Y \mid T=0, M=m, \boldsymbol{X}, \boldsymbol{V}) P(\boldsymbol{V} \mid T=0, \boldsymbol{X})\} P(\boldsymbol{X})$$

   - regression of Y on $T, M, \boldsymbol{X}, \boldsymbol{V}$
   - model $\boldsymbol{V}$ given T and $\boldsymbol{X}$ $\rightsquigarrow$ difficult if $\boldsymbol{V}$ is high-dimensional

2. Marginal structural models (Robins et al. 2000)

$$\mathbb{E}[Y(t,m)] = \mathbb{E}\left[\underbrace{\frac{1\{T=t, M=m\}}{\Pr(T=t \mid \boldsymbol{X})}}_{\text{IPW for treatment}} \cdot \underbrace{\frac{1}{\Pr(M=m \mid T=t, \boldsymbol{X}, \boldsymbol{V})}}_{\text{IPW for mediator given treatment}} \times Y\right]$$

   - no need to model $\boldsymbol{V}$
   - covariate balancing methods are also available (Imai and Ratkovic, 2015)

Identification of NDE/NIE with Posttreatment Confounding

- Identification is impossible with *observed* posttreatment confounding
- Consider the following NPSEM

$$\begin{aligned} Y &= f_Y(M, \boldsymbol{V}, T, \epsilon) \\ M &= f_M(\boldsymbol{V}, T, \eta) \\ \boldsymbol{V} &= f_{\boldsymbol{V}}(T, \xi) \end{aligned}$$

- Cross-world independence cannot hold

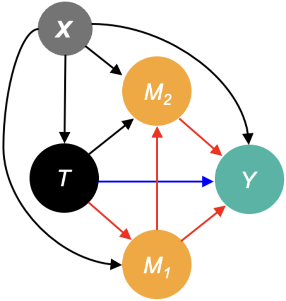
$$\underbrace{\boldsymbol{V}(1)}_{=f_{\boldsymbol{V}}(1,\xi)} \not\perp\!\!\!\perp \underbrace{\boldsymbol{V}(0)}_{=f_{\boldsymbol{V}}(0,\xi)} \quad \Longrightarrow \quad Y(t', m, \boldsymbol{V}(t'), \epsilon) \not\perp\!\!\!\perp M(t, \boldsymbol{V}(t), \eta)$$

- Conditioning on T and $\boldsymbol{V}$ does not solve this problem

Multiple Causally Related Mediators

- Same as the posttreatment confounding setting
- Path specific effects
  1. $T \to Y$
  2. $T \to M_1 \to Y$
  3. $T \to M_2 \to Y$
  4. $T \to M_1 \to M_2 \to Y$
- Combined effect:

$$
\begin{aligned}
& T \to M_1 \rightsquigarrow Y \\
= & (T \to M_1 \to Y) + (T \to M_1 \to M_2 \to Y)
\end{aligned}
$$



- Generalized cross-world independence assumptions:
  1. $\{M_{1i}(t), M_{2i}(t, m_1), Y_i(t, m_1, m_2)\}_{t, \mathbf{m_1}, \mathbf{m_2}} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i = \boldsymbol{x}$
  2. $\{M_{2i}(t', m_1), Y_i(t', m_1, m_2)\}_{t', \mathbf{m_1}, \mathbf{m_2}} \perp\!\!\!\perp M_{1i}(t) \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}$
  3. $\{Y_i(t', m_1, m_2)\}_{t', \mathbf{m_2}} \perp\!\!\!\perp M_{2i}(t, m_1) \mid M_{1i} = m_1, T_i = t, \boldsymbol{X}_i = \boldsymbol{x}$
- Identifiable decomposition:

$$
\text{ATE} = (T \to Y) + (T \to M_2 \to Y) + (T \to M_1 \rightsquigarrow Y)
$$

Interventional Direct and Indirect Effects (IDE and IIE)

- $\mathcal{P}_{M(t)}$: interventional distribution that independently generates $M(t)$
- Definition (Geneletti, 2007; Lok, 2016)

$$\text{Individual: } \begin{cases} \text{IIE}_i(t) & = & Y_i(t, \mathcal{P}_{M(1)}) - Y_i(t, \mathcal{P}_{M(0)}) \\ \text{IDE}_i(t) & = & Y_i(1, \mathcal{P}_{M(t)}) - Y_i(0, \mathcal{P}_{M(t)}) \end{cases}$$

$$\text{Average: } \begin{cases} \overline{\text{IIE}}(t) & = & \mathbb{E}[Y(t, \mathcal{P}_{M(1)}) - Y(t, \mathcal{P}_{M(0)})] \\ \overline{\text{IDE}}(t) & = & \mathbb{E}[Y(t, \mathcal{P}_{M(1)}) - Y(t, \mathcal{P}_{M(0)})] \end{cases}$$

- Interpretation
  - similar to NIE and NDE
  - IDE is a function of CDE:

$$\text{IDE}_i(t) = \sum_m \text{CDE}_i(m) \times P(M(t) = m)$$

  - no mediation: zero treatment effect on M implies zero IIE
- Effect decomposition

$$\underbrace{Y_i(1, \mathcal{P}_{M(1)}) - Y_i(0, \mathcal{P}_{M(0)})}_{\text{Interventional Total Effect (ITE)} \neq \text{TE}} \quad = \quad \text{IIE}_i(t) + \text{IDE}_i(1-t)$$

Identification of IDE and IIE

- Once CDE is identified, we can identify IDE:

$$\overline{\mathsf{IDE}}(t) = \sum_m \overline{\mathsf{CDE}}(m) P(M(t) = m)$$

- IIE is also identifiable:

$$\overline{\mathsf{IIE}}(t) = \sum_m \mathbb{E}[Y(t,m)] \left\{ P(M(1) = m) - P(M(0) = m) \right\}$$

- Effect decomposition

$$\underbrace{\mathbb{E}[Y(1, \mathcal{P}_{M(1)}) - Y(0, \mathcal{P}_{M(0)})]}_{\neq \mathbb{E}[Y(1,M(1)) - Y(0,M(0))]} = \overline{\mathsf{IDE}}(t) + \overline{\mathsf{IIE}}(1-t)$$

- Complete mediation: $\overline{\mathsf{IDE}} = 0$
- Identification is possible with observed pretreatment and posttreatment confounding
- Experimental identification via parallel design is also possible

Take-aways IV

- Posttreatment confounding
  - CDE can be identified under exogeneity
  - estimation of CDE requires marginalizing posttreatment confounders
  - NIE/NDE are not identifiable under exogeneity
  - Different decomposition is identifiable under cross-world independence

- Alternative estimands
  - interventional direct and indirect effects (IDE/IIE)
  - interventional distribution on M
  - enables decomposition of alternative total effect
  - identification of CDE implies that of IDE/IIE

Conclusion, Resources, and References

Concluding Remarks on Causal Mechanisms

- Study of causal mechanisms is essential but challenging
- Triangulation of evidence is necessary
  - causal quantities
    - CDE
    - NDE/NIE, path specific effects
    - IDE/IIE
  - causal identification strategies
    - selection on observables
    - instrumental variables
    - experimental designs
    - partial identification
  - statistical methodologies
    - weighting and regression
    - sensitivity analysis
    - nonparametric modeling and machine learning

Resources

- Statistical software:
  - mediation (R and Stata)
  - Valeri and VanderWeele macros (SPSS, SAS, Stata)

- Review article by an economist:

Huber, Martin (2020). "Mediation Analysis". Handbook of Labor, Human Resources and Population Economics. Ed. by Klaus F. Zimmermann. Cham: Springer.

- Monographs:

VanderWeele, Tyler J. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Press.

Wodtke, Geoffrey T. and Xiang Zhou (Forthcoming). Causal Mediation Analysis. Cambridge University Press.

Works Cited I

Blackwell, Matthew, Ruofan Ma, and Aleksei Opacic (2024). "Assumption Smuggling in Intermediate Outcome Tests of Causal Mechanisms". arXiv preprint arXiv:2407.07072.

Bullock, John G. and Donald P. Green (2021). "The Failings of Conventional Mediation Analysis and a Design-Based Alternative". Advances in Methods and Practices in Psychological Science 4.4, pp. 1–14.

Chen, Stacey H., YenâChien Chen, and JinâTan Liu (2019). "The Impact of Family Composition on Educational Achievement". Journal of Human Resources 54.1, pp. 122–170.

Farbmacher, Helmut et al. (Jan. 2022). "Causal mediation analysis with double machine learning". The Econometrics Journal 25.2, pp. 277–300.

Works Cited II

Frölich, Markus and Martin Huber (2017). "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables". Journal of the Royal Statistical Society Series B: Statistical Methodology 79.5, pp. 1645–1666.

Geneletti, Sara (2007). "Identifying direct and indirect effects in a non-counterfactual framework". Journal of the Royal Statistical Society, Series B (Statistical Methodology) 69.2, pp. 199–215.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010). "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects". Statistical Science 25.1, pp. 51–71.

Works Cited III

Imai, Kosuke and Marc Ratkovic (2015). "Robust Estimation of Inverse Probability Weights for Marginal Structural Models". Journal of the American Statistical Association 110.511, pp. 1013–1023.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto (2013). "Experimental Designs for Identifying Causal Mechanisms (with discussions)". Journal of the Royal Statistical Society, Series A (Statistics in Society) 176.1, pp. 5–51.

Imai, Kosuke and Teppei Yamamoto (Spring 2013). "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments". Political Analysis 21.2, pp. 141–171.

Imbens, Guido W. and Whitney K. Newey (2009). "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity". Econometrica 77.5, pp. 1481–1512.

Works Cited IV

Kwon, Soonwoo and Jonathan Roth (Apr. 2024). "Testing Mechanisms". arXiv preprint arXiv:2404.11739.

Lok, Judith J. (2016). "Defining and Estimating Causal Direct and Indirect Effects When Setting the Mediator to Specific Values Is Not Feasible". Statistics in Medicine 35.22, pp. 4008–4020.

Pearl, Judea (2000). Causality: Models, Reasoning, and Inference. New York: Cambridge University Press.

— (2001). "Direct and Indirect Effects". Proc. of the 17th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, pp. 411–420.

Robins, James M. and Sander Greenland (Mar. 1992). "Identifiability and exchangeability for direct and indirect effects". Epidemiology 3.2, pp. 143–155.

Works Cited V

Robins, James M., Miguel Ángel Hernán, and Babette Brumback (2000). "Marginal Structural Models and Causal Inference in Epidemiology". Epidemiology 11.5, pp. 550–560.

Tchetgen, Eric J. and Ilya Shpitser (2012). "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness, and Sensitivity Analysis". Annals of Statistics 40.3, pp. 1816–1845.