

# 14. Causal Mechanisms

ISS5096 || ECI

Jaewon (“Jay-one”) Yoo

National Tsing Hua University

# 1/ Causal Mechanisms

# Theory and Causality

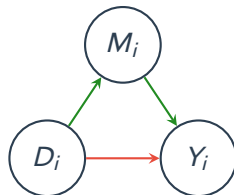
- Theory  $\implies$  (or  $\equiv$ ) causal effects
- But they also tell us **how** those causes should impact the outcomes.
  - Theory A: causal effect is “due to” path A
  - Theory B: causal effect is “due to” path B
- How to adjudicate between theories that predict the same ATE?
- Put differently: what **mechanism** drives a particular causal effect?

# What is a Causal Mechanism?

- A massive diversity of definitions
- But basically: **how** a treatment affects an outcome
- Cannot estimate a mechanism, only test for observable implications:
  - causal mediation (effect decomposition)
  - effects modification (null effect among a subgroup)
  - presence or absence of direct effects
  - placebo tests

# Notation

- DAG representation:
  - Treatment variable  $D_i \in \{0, 1\}$
  - Mediator,  $M_i \in \mathcal{M}$
  - Potential outcome variable  $Y_i(d, m)$



- Mediation goal: decompose total effect into **direct** and **indirect** effects.
- Moderator vs mediator:
  - **Moderator**: pretreatment variable correlated with the treatment effect.
  - **Mediator**: a posttreatment variable that changes the treatment effect.
  - Mediator has potential outcomes as well:  $M_i(d)$
- Consistency:  $M_i = M_i(D_i)$  and  $Y_i(D_i, M_i(D_i))$ .

# Interpreting the Potential Outcomes

- Example:  $D_i$  is exercise,  $M_i$  is diet, and  $Y_i$  is weight.
  - $D_i = 1$  is “run 10 km/day”,  $D_i = 0$  is don’t run
  - $M_i$  is the number of calories to eat.
- Some different possible potential outcomes:
  - $Y_i(1, 1500)$ : weight you would have if we forced you to run 10 km/day and eat 1500 kcals a day.
  - $Y_i(1, M_i(1))$ : weight if you run 10 km/day, but no intervention on diet.
  - $Y_i(0, M_i(1))$ : weight if you didn’t run, but ate like you did.
- Cross-world counterfactuals  $Y_i(0, M_i(1))$  logically impossible to observe.
  - Not just the fundamental problem of CI.

## 2/ Estimands

# Controlled Direct Effects (CDE)

- Definition for each  $m \in \mathcal{M}$ :

$$\text{Individual: } \xi_i(m) = Y_i(1, m) - Y_i(0, m)$$

$$\text{Average: } \bar{\xi}(m) = \mathbb{E} [Y_i(1, m) - Y_i(0, m)]$$

- Interpretation:
  - Effect of treatment when holding mediator fixed at  $m$ .
  - The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
  - Target of experiment manipulating  $D_i$  and  $M_i$ .
- If  $M_i$  fully mediates effect of  $D$ , then CDEs will be 0 for all  $m$ .
  - $\rightsquigarrow$  can be used to establish existence of unmediated path from  $D \rightarrow Y$ .
- Can capture **interactions** if  $\bar{\xi}_i(m) \neq \bar{\xi}_i(m')$



# Natural Indirect Effects (NIE)

- Definition of the **natural indirect effect** (NIE):

$$\text{Individual: } \delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

$$\text{Average: } \bar{\delta}(d) = \mathbb{E} [Y_i(d, M_i(1)) - Y_i(d, M_i(0))]$$

- Interpretation:
  - Effect of a change in the mediator induced by the effect of  $D_i$  on  $M_i$ .
  - Holding fixed the value of treatment.
- Also called the **causal mediation effect**
- If  $D_i$  doesn't affect  $M_i$ , so that  $M_i(1) = M_i(0)$ , then  $\delta_i = 0$ .

# Natural Direct Effects (NDEs)

- Definition of the **natural direct effect** (NDE) of the treatment:

$$\text{Individual: } \zeta_i(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$

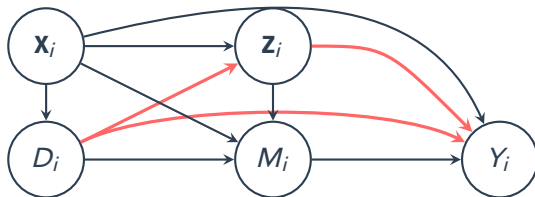
$$\text{Average: } \bar{\zeta}(d) = \mathbb{E} [Y_i(1, M_i(d)) - Y_i(0, M_i(d))]$$

- Interpretation:
  - Effect of treatment when mediator is at its natural value for  $D_i = d$ .
  - Effect of a redesigned treatment that doesn't affect the mediator
- Total effect decomposition:

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \underbrace{\delta_i(d)}_{\text{NIE}} + \underbrace{\zeta_i(1-d)}_{\text{NDE}}$$

## 3/ Identification

# Identification for CDEs



- Conditioning sets:
  - $\mathbf{X}_i$ : pre-treatment confounders
  - $\mathbf{Z}_i$ : post-treatment or intermediate confounders

- Sequential ignorability** (Robins):

$$\{Y_i(d', m), M_i(d')\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

$$Y_i(d, m) \perp\!\!\!\perp M_i \mid \mathbf{X}_i = \mathbf{x}, D_i = d, \mathbf{Z}_i = \mathbf{z}$$

- Interpretation: two “selection-on-observables” assumptions.
  - $D_i$  randomly assigned conditional on  $\mathbf{X}_i$ .
  - $M_i$  randomly assigned conditional on  $\mathbf{X}_i$ ,  $D_i$ , and  $\mathbf{Z}_i$ .

# Identifying the ACDE

- Post-treatment bias if we just condition on  $\mathbf{Z}_i$ :

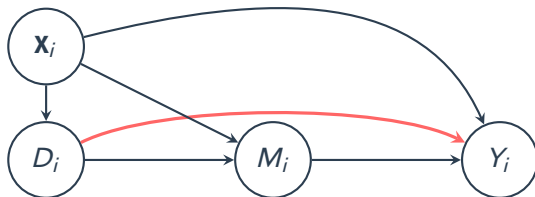
$$\begin{aligned}\bar{\xi}(m) \neq \sum_{\mathbf{x}, \mathbf{z}} \{ & \mathbb{E}[Y_i \mid D_i = 1, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] \\ & - \mathbb{E}[Y_i \mid D_i = 0, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] \} \mathbb{P}(\mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z})\end{aligned}$$

- Ignores that  $\mathbf{Z}_i$  depends on  $D_i$ !
- Nonparametric identification of the ACDE:

$$\begin{aligned}\bar{\xi}(m) = \sum_{\mathbf{x}, \mathbf{z}} \{ & \mathbb{E}[Y_i \mid D_i = 1, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] \mathbb{P}(\mathbf{Z}_i = \mathbf{z} \mid D_i = 1, \mathbf{X}_i = \mathbf{x}) \\ & - \mathbb{E}[Y_i \mid D_i = 0, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] \mathbb{P}(\mathbf{Z}_i = \mathbf{z} \mid D_i = 0, \mathbf{X}_i = \mathbf{x}) \} \\ & \times \mathbb{P}(\mathbf{X}_i = \mathbf{x})\end{aligned}$$

- **g-formula** (Robins) generalizes to any number of treatments

# Identification for Mediation



- **Sequential ignorability** (Imai et al. 2010):

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

$$Y_i(d, m) \perp\!\!\!\perp M_i \mid \mathbf{X}_i = \mathbf{x}, D_i = d$$

- $\rightsquigarrow$  ANIE and ANDE are identified.
- **No post-treatment confounders** (measured or unmeasured)
  - Assumes away post-treatment bias conditioning on  $M_i$

# Identifying (In)direct Effects

- ANIE under binary treatment/mediator:

$$\begin{aligned}\bar{\delta}(d) = & \sum_{\mathbf{x}} \left( \underbrace{\{\mathbb{P}[M_i = 1 \mid D_i = 1, \mathbf{x}_i = \mathbf{x}] - \mathbb{P}[M_i = 1 \mid D_i = 0, \mathbf{x}_i = \mathbf{x}]\}}_{\text{effect of } D_i \text{ on } M_i} \right. \\ & \times \underbrace{\{\mathbb{E}[Y_i \mid M_i = 1, D_i = d, \mathbf{x}_i = \mathbf{x}] - \mathbb{E}[Y_i \mid M_i = 0, D_i = d, \mathbf{x}_i = \mathbf{x}]\}}_{\text{effect of } M_i \text{ on } Y_i} \Big) \\ & \times \mathbb{P}(\mathbf{x}_i = \mathbf{x})\end{aligned}$$

- Multiply paths given  $\mathbf{X}_i$  and aggregate intuitive given DAG:



# (In)direct Effects with Non-binary Mediators

- Let's say that the mediator has  $J$  categories:

$$\begin{aligned}\bar{\delta}(d) = \sum_{\mathbf{x}} & \left( \sum_{m=0}^{J-1} \mathbb{E}[Y_i \mid M_i = m, D_i = d, \mathbf{x}_i = \mathbf{x}] \right. \\ & \times \{ \mathbb{P}[M_i = m \mid D_i = 1, \mathbf{x}_i = \mathbf{x}] - \mathbb{P}[M_i = m \mid D_i = 0, \mathbf{x}_i = \mathbf{x}] \} \Big) \\ & \times \mathbb{P}(\mathbf{x}_i = \mathbf{x})\end{aligned}$$

- The ANDE is the following:

$$\begin{aligned}\bar{\zeta}(d) = \sum_{\mathbf{x}} & \left( \sum_{m=0}^{J-1} \{ \mathbb{E}[Y_i \mid M_i = m, D_i = 1, \mathbf{x}_i = \mathbf{x}] - \mathbb{E}[Y_i \mid M_i = m, D_i = 0, \mathbf{x}_i = \mathbf{x}] \} \right. \\ & \times \mathbb{P}[M_i = m \mid D_i = d, \mathbf{x}_i = \mathbf{x}] \Big) \mathbb{P}(\mathbf{x}_i = \mathbf{x})\end{aligned}$$

- Effect of  $D_i$  for a fixed  $m$  averaged over the distribution of  $M_i$  when  $D_i = d$ .



# Alternative Identification

- Robins proposed a different identification strategy, based on a **no-interactions assumption**:

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

- The CDE does not depend on  $m$  for any unit  $i$ .
- $\rightsquigarrow$  ACDE = ANDE.
- Strong assumption because it has to hold at the individual level (like monotonicity for IV).

## 4/ Linear Structural Equation Models

# Estimation

- Let's say that we have a linear, structural model for all variables:

$$M_i(d) = \alpha_0 + \alpha_1 d + \eta_i$$

$$Y_i(d, m) = \beta_0 + \beta_1 d + \beta_2 m + \varepsilon_i$$

- Here the effect of treatment and mediator are constant across units.
- This is a huge simplification and may be incorrect.
- Allows us to “plug-in” and get potential outcomes:

$$\begin{aligned} Y_i(1, M_i(1)) &= \beta_0 + \beta_1 \times 1 + \beta_2 M_i(1) + \varepsilon_i \\ &= \beta_0 + \beta_1 \times 1 + \beta_2 (\alpha_0 + \alpha_1 \times 1 + \eta_i) + \varepsilon_i \end{aligned}$$

# Linear Models and Mediation

- It's clear that we can write the total effect of the treatment in the following way:

$$\begin{aligned}Y_i(1, M_i(1)) - Y_i(0, M_i(0)) &= \beta_0 + \beta_1 + \beta_2(\alpha_0 + \alpha_1 + \eta_i) + \varepsilon_i \\&\quad - \beta_0 - \beta_2(\alpha_0 + \eta_i) - \varepsilon_i \\&= \beta_1 + \beta_2 \cdot \alpha_1\end{aligned}$$

- What about the indirect effect:

$$\begin{aligned}Y_i(0, M_i(1)) - Y_i(0, M_i(0)) &= \beta_0 + \beta_2(\alpha_0 + \alpha_1 + \eta_i) + \varepsilon_i \\&\quad - \beta_0 - \beta_2(\alpha_0 + \eta_i) - \varepsilon_i \\&= \beta_2 \cdot \alpha_1\end{aligned}$$

# Estimation with LSEMs

- Estimate the total effect from a regression of  $Y_i$  on  $D_i$  and  $\mathbf{X}_i$
- Estimate the  $\hat{\beta}_1$  and  $\hat{\beta}_2$  from a regression of  $Y_i$  on  $D_i$ ,  $M_i$ , and  $\mathbf{X}_i$ .
- Estimate  $\hat{\alpha}_1$  from a regression of  $M_i$  on  $D_i$
- Direct effect is  $\widehat{\text{ANDE}} = \hat{\beta}_1$
- Indirect effect as the product:  $\widehat{\text{ANIE}} = \hat{\alpha}_1 \hat{\beta}_2$ .

# Interactions

- **Implicit assumption:** no interactions

$$\text{ANIE}(1) = \text{ANIE}(0)$$

- We could incorporate an interaction into the model here to allow for the indirect effect to vary.

$$Y_i(d, m) = \beta_0 + \beta_1 d + \beta_2 m + \beta_3 dm + \varepsilon_i$$

# Variance Estimates

- The variance of the total effect and the direct effect are straightforward.
  - Just the SE of the estimated coefficients.
- The indirect effect is more complicated because it is a function of multiple parameters.
- Using the delta method, the variance of  $\widehat{ANIE} = \widehat{\alpha}_1 \widehat{\beta}_2$  can be written:

$$\mathbb{V}[\widehat{ANIE}] \approx \widehat{\alpha}_1^2 \mathbb{V}[\widehat{\beta}_2] + \widehat{\beta}_2^2 \mathbb{V}[\widehat{\alpha}_1]$$

- We can use this formula to estimate standard errors for the indirect effects.

## 5/ Nonparametric Estimation



# Nonparametric Estimation

- LSEs require strong modeling assumptions  $\rightsquigarrow$  what about nonparametrics?
- If the number of categories in  $M_i$ ,  $D_i$ , and  $\mathbf{X}_i$  are small, use **plug-in estimator** for the CEF of  $Y_i$ :

$$\widehat{\mathbb{E}}[Y_i \mid M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}] = \frac{\sum_i Y_i \mathbb{1}\{M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}\}}{\sum_i \mathbb{1}\{M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}\}}$$

- Same for  $M_i$ :

$$\widehat{\mathbb{P}}[M_i = m \mid D_i = d, \mathbf{X}_i = \mathbf{x}] = \frac{\sum_i \mathbb{1}\{M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}\}}{\sum_i \mathbb{1}\{D_i = d, \mathbf{X}_i = \mathbf{x}\}}$$

# What About More Complicated Scenarios?

- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator.

$$\mu_{dm}(\mathbf{x}) = \mathbb{E}[Y_i \mid M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}]$$

- Flexibly estimate  $\mu_{dm}(\mathbf{x})$  as a function of  $\mathbf{x}$  using splines of  $\mathbf{x}$ .
- To get the standard errors, we can use bootstrapping.
- Need to be careful with the curse of dimensionality in  $\mathbf{X}_i$ . Use good nonparametric strategies (cross-validation, etc)

# Continuous Mediator, Nonparametric

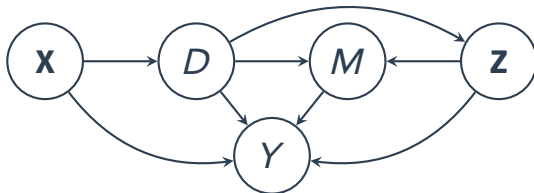
- What if the mediator is continuous? Things get tricky.
- Need to integrate over the distribution of the mediators to get the ANIE:

$$\bar{\delta}(d) = \int \int \mathbb{E}[Y_i \mid M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}] \\ \{dF_{M_i|D_i=1, \mathbf{x}_i=\mathbf{x}}(m) - dF_{M_i|D_i=0, \mathbf{x}_i=\mathbf{x}}(m)\} dF_{\mathbf{x}_i}(\mathbf{x})$$

- Obviously, this is a much harder problem. In this case, we actually can use Monte Carlo simulation to take the integral.
- Modeling  $M_i$  probably appropriate here.

## 6/ Estimating Controlled Direct Effects

# Sequential G-Estimation

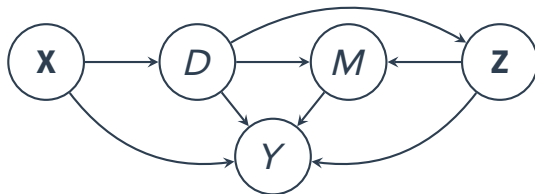


- **Sequential g-estimation** is one of many approaches in these settings.
  - Other approaches include weighting.
  - Linear version of a broader class called **structural nested mean models**
- Run the “long” regression:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + \mathbf{X}'_i \gamma_3 + \mathbf{Z}'_i \gamma_4 + \varepsilon_i$$

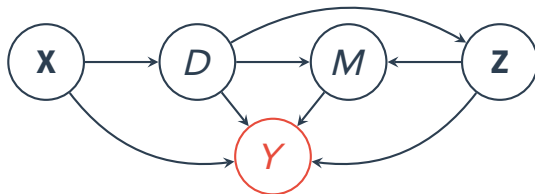
- $\gamma_1$  is not the CDE (posttreatment bias)
- $\gamma_2$  **is** the effect of  $M_i$  on  $Y_i$  (if model is correct)

# Blip down



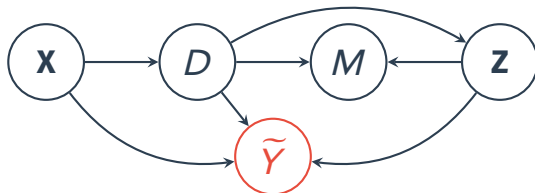
- Create a blipped down (or demediated) outcome:  $\tilde{Y}_i = Y_i - \hat{\gamma}_2 M_i$

# Blip down



- Create a blipped down (or demediated) outcome:  $\tilde{Y}_i = Y_i - \hat{\gamma}_2 M_i$
- The **blip-down** removes the effect of  $M_i$  on  $Y_i$  from the outcome.

# Blip down

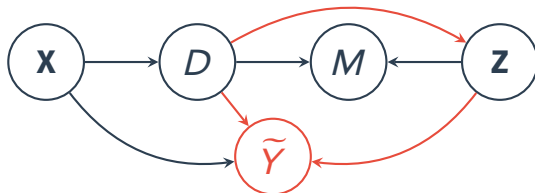


- Create a blipped down (or demediated) outcome:  $\tilde{Y}_i = Y_i - \hat{\gamma}_2 M_i$
- The **blip-down** removes the effect of  $M_i$  on  $Y_i$  from the outcome.
- Any remaining effect of  $D_i$  on  $Y_i$  is just the CDE:

$$\mathbb{E}[\tilde{Y}_i \mid D_i = d, \mathbf{x}_i] = \mathbb{E}[Y_i(d, 0) \mid \mathbf{x}_i]$$



# Blip down



- Create a blipped down (or demediated) outcome:  $\tilde{Y}_i = Y_i - \hat{\gamma}_2 M_i$
- The **blip-down** removes the effect of  $M_i$  on  $Y_i$  from the outcome.
- Any remaining effect of  $D_i$  on  $Y_i$  is just the CDE:

$$\mathbb{E}[\tilde{Y}_i \mid D_i = d, \mathbf{x}_i] = \mathbb{E}[Y_i(d, 0) \mid \mathbf{x}_i]$$

- Relies on correct modeling of the outcome!

# Sequential G-Estimation

1. Run a regression of  $Y_i$  on  $M_i, \mathbf{z}_i, D_i, \mathbf{x}_i$ .

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + \mathbf{x}_i' \gamma_3 + \mathbf{z}_i' \gamma_4 + \varepsilon_i$$

2. Subtract off the effect of  $M_i$  on  $Y_i$ :

$$\tilde{Y}_i = Y_i - \hat{\gamma}_2 M_i$$

3. Regress blipped-down outcome on  $D_i$  and  $\mathbf{x}_i$ :

$$\tilde{Y}_i = \beta_0 + \beta_1 D_i + \mathbf{x}_i' \beta_2 + \eta_i$$

$$\text{CDE}(0) = \mathbb{E}[Y_i(1, 0) - Y_i(0, 0)] = \beta_1$$

4. Bootstrap or complicated variance estimator for SEs
  - Second regression ignores the first regression.

# Notes on Sequential G-Estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.
- We can weaken this, but requires us to model the distribution of  $\mathbf{Z}_i$  which might be very high dimensional:

$$\begin{aligned} & \int_{\mathbf{x}} \int_{\mathbf{z}} \mathbb{E}[Y_i \mid \mathbf{x}, d = 1, \mathbf{z}, m] dF_{\mathbf{Z} \mid D, \mathbf{x}}(\mathbf{z} \mid d = 1, \mathbf{x}) dF_{\mathbf{x}}(\mathbf{x}) \\ & - \int_{\mathbf{x}} \int_{\mathbf{z}} \mathbb{E}[Y_i \mid \mathbf{x}, d = 0, \mathbf{z}, m] dF_{\mathbf{Z} \mid D, \mathbf{x}}(\mathbf{z} \mid d = 0, \mathbf{x}) dF_{\mathbf{x}}(\mathbf{x}) \end{aligned}$$

- Typical selection on observables: need correct model for covariates in both steps.
- ATE - ACDE  $\neq$  an indirect effect, but still can tell us something about mechanisms.

# Wrap-Up

- Mechanisms are hard.

# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.

# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.

# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.

# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation



# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation
  - Controlled direct effects

# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation
  - Controlled direct effects
  - Effect modification

# Wrap-Up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation
  - Controlled direct effects
  - Effect modification
  - Placebo tests