Understanding effect heterogeneity in observational and randomized studies of causality

Institute for Computational and Experimental Research in Mathematics (ICERM)

November 2023

Iván Díaz
Joint work with Thang Vo, Kara Rudolph, and Brian Gilbert
Division of Biostatistics
Department of Population Health
New York University Grossman School of Medicine

Motivating example[1]

- Moving to Opportunity study (MTO) was a large-scale housing voucher experiment conducted by the US Department of Housing and Urban Development in the 1990s-2000s in five cities.
- Families living in high-rise public housing developments in five US cities could sign up to be randomized to receive a Section 8 housing voucher.
- The vouchers can be used by families to move out of public housing into a rental on the private market.
- Participating MTO families were followed up for 10-15 years after randomization and economic, educational, and healthrelated outcomes were assessed.

[1] This research was conducted as a part of the U.S. Census Bureau's Evidence Building Project Series. Any opinions and conclusions expressed herein are those of the author and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project P-7504667: CBDRB-FY22-CES018-013, CBDRB-FY24-CES018-002)

Motivating example

- Multiple studies have contributed evidence of effect heterogeneity across MTO sites.
- It is natural to want to understand why: what components of the causal process are contributing to such heterogeneity?
    - Is this heterogeneity explained largely by differences in the distribution of **baseline compositional characteristics** (e.g., sociodemographic variables) between sites?
    - Or, are differences large due to differences in the **mediating mechanisms**?
    - Or, are they due to differences in the **outcome mechanisms** across sites?

This is a common problem

Randomized trials of the same therapies often produce different results:

- E.g., in treatments to reduce blood pressure in patients with cardiovascular disease.[2]
- E.g., in prevention of preterm delivery[3]
- E.g., in the treatment of sepsis[4]

[2] Basu, S., Sussman, J.B. and Hayward, R.A., 2017. Detecting heterogeneous treatment effects to guide personalized blood pressure treatment: a modeling study of randomized clinical trials. Annals of internal Medicine, 166(5), pp.354-360.

[3] Blackwell, Sean C., et al. "17-OHPC to prevent recurrent preterm birth in singleton gestations (PROLONG study): a multicenter, international, randomized double-blind trial." American journal of perinatology 37.02 (2020): 127-136.

[4] Kalil, Andre C., and Diana F. Florescu. "Severe sepsis: are PROWESS and PROWESS-SHOCK trials comparable? A clinical and statistical heterogeneity analysis." Critical Care 17.4 (2013): 167.

Causal model

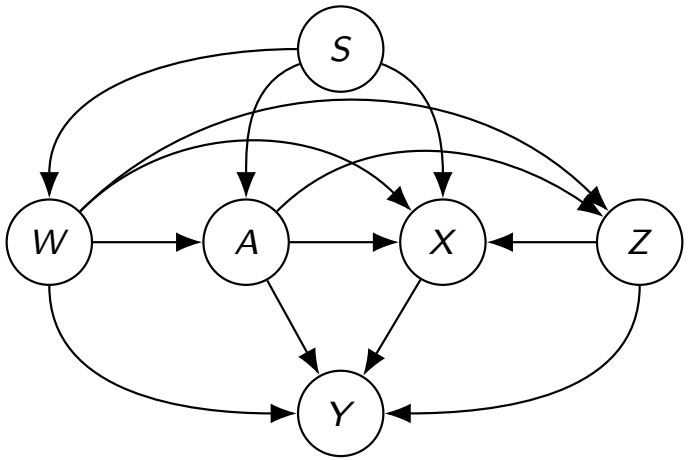

Figure 1: Causal selection diagram under mediation and intermediate confounding.

- For example:
  - A is randomization to receiving the voucher
  - X denotes actual use of the voucher to move
  - Z denotes confounders of treatment uptake and outcomes
  - Y denotes behavioral problems in adolescent children
- In what follows we will consider a bundled mediator $M = (X, Z)$

Counterfactual variables of interest

We define three counterfactual variables of interest:

- $Y(a)$ is the counterfactual outcome that would have been observed if, possibly contrary to fact, $A = a$ had been assigned w.p. 1
- $Y(a, m)$ is the counterfactual outcome that would have been observed if, possibly contrary to fact, $A = a$ and $M = m$ had been assigned w.p. 1
- $M(a, s)$ is the counterfactual outcome that would have been observed if, possibly contrary to fact, $A = a$ and $S = s$ had been assigned w.p. 1

Effect heterogeneity decomposition for two studies

Let

$$\delta = \mathrm{E}[Y(1) - Y(0) \mid S = 1] - \mathrm{E}[Y(1) - Y(0) \mid S = 0],$$

denote the difference between the treatment effects in the two studies, and consider the decomposition

$$\delta = \delta_{\mathrm{EH}} + \delta_{\mathrm{CM}},$$

where

$$\delta_{\mathrm{EH}} = \int \Big\{ \mathrm{E}\left[Y(1) - Y(0) \mid W, S = 1\right] - \mathrm{E}\left[Y(1) - Y(0) \mid W, S = 0\right] \Big\} \mathrm{dP}(W \mid S = 0)$$

$$\delta_{\mathrm{CM}} = \int \mathrm{E}[Y(1) - Y(0) \mid W, S = 1] \Big\{ \mathrm{dP}(W \mid S = 1) - \mathrm{dP}(W \mid S = 0) \Big\}.$$

- δ_{EH} is a parameter that measures the extent of **differential effect heterogeneity** between the studies by comparing their conditional average treatment effects,
- δ_{CM} measures the extent to which the two studies represent different a **different case-mix** from the population.

Decomposing conditional average treatment effect heterogeneity

When δ_{EH} is non-zero, it is also of interest to understand whether the differential effect heterogeneity is due to

- between-study variations in effect modification by W, or
- between-study variation on the effect of treatment on the intermediate variables $M = (X, Z)$.

By definition, we have that $Y(a) = Y(a, M(a, S))$, so that

$$\begin{aligned}&\mathrm{E}\left[Y(1) - Y(0) \mid W, S = 1\right] - \mathrm{E}\left[Y(1) - Y(0) \mid W, S = 0\right] = \\ &\quad \mathrm{E}\left[Y(1, M(1,1)) - Y(0, M(0,1)) \mid W, S = 1\right] - \mathrm{E}\left[Y(1, M(1,0)) - Y(0, M(0,0)) \mid W, S = 0\right],\end{aligned}$$

The left hand side of this expression makes it clear that:

- The CATE may be different across studies due to different effect modification by W, or
- due to effects of A on the intermediate variables M that vary with S.

Decomposing conditional average treatment effect heterogeneity

To address the above, we propose the decomposition

$$\delta_{\text{EH}} = \delta_{\text{EM}} + \delta_{\text{MV}},$$

where

$$\begin{aligned}
\delta_{\text{EM}} = \int \Big\{ & \mathrm{E}\left[Y(1, M(1,1)) - Y(0, M(0,1)) \mid W, S = 1\right] \\
& - \mathrm{E}\left[Y(1, M(1,1)) - Y(0, M(0,1)) \mid W, S = 0\right] \Big\} \mathrm{dP}(W \mid S = 0) \\
\delta_{\text{MV}} = \int \Big\{ & \mathrm{E}\left[Y(1, M(1,1)) - Y(0, M(0,1)) \mid W, S = 0\right] \\
& - \mathrm{E}\left[Y(1, M(1,0)) - Y(0, M(0,0)) \mid W, S = 0\right] \Big\} \mathrm{dP}(W \mid S = 0),
\end{aligned}$$

- δ_{EM} denotes a parameter measuring between-study differences in **pure effect modification**, and
- δ_{MV} denotes a parameter that measures the between-study heterogeneity of **treatment effects on intermediate variables**.

Identification

Define

$$\theta(s_Y, s_M, s_W) = \int \mathrm{E}[Y(1, M(1, s_M)) - Y(0, M(0, s_M)) \mid W, S = s_Y] \mathrm{dP}(W \mid S = s_W),$$

Then

$$\begin{aligned}\delta_{\mathrm{CM}} &= \theta(1, 1, 1) - \theta(1, 1, 0) \\ \delta_{\mathrm{EH}} &= \theta(1, 1, 0) - \theta(0, 0, 0),\end{aligned}$$

and

$$\begin{aligned}\delta_{\mathrm{EM}} &= \theta(1, 1, 0) - \theta(0, 1, 0) \\ \delta_{\mathrm{MV}} &= \theta(0, 1, 0) - \theta(0, 0, 0).\end{aligned}$$

So that identifying $\theta(s_Y, s_M, s_W)$ will be sufficient

Identification assumptions

Assumption 1

Assume there exists $\epsilon > 0$ such that, for $s \in \{0, 1\}$ and with probability 1 over draws of W:

(i) $\epsilon < \mathrm{P}(A = 1 \mid S = s, W) < 1 - \epsilon$, and

(ii) $\epsilon < \mathrm{P}(S = 1 \mid W) < 1 - \epsilon$.

Assumption 2 (No unmeasured confounding in study $S = s$)

For all a and m, assume $Y(a, m) \perp\!\!\!\perp A \mid W, S = s$ and $M(a) \perp\!\!\!\perp A \mid W, S = s$.

Assumption 3 (Conditional exchangeability of study assignment)

For all a and s, assume $M(a, s) \perp\!\!\!\perp S \mid W$.

Assumption 4 (Counterfactual independences in study $S = s$)

For $s' \in \{0, 1\}$ and all a and m, assume $Y(a, m) \perp\!\!\!\perp M(a, s') \mid (W, S = s)$ and $Y(a, m) \perp\!\!\!\perp M \mid (A, W, S = s)$.

Identification assumptions

I personally find it difficult to reason about validity of assumptions using counterfactuals.

The following are sufficient conditions for to hold in terms of the errors U of a non-parametric structural causal model:

Proposition 1

The following statements are true:

(i) *Assume* $U_A \perp\!\!\!\perp (U_Y, U_M) \mid W, S = s$. *Then Ass.2 holds.*

(ii) *Assume* $U_S \perp\!\!\!\perp U_M \mid W$. *Then Ass.3 holds.*

(iii) *Assume* $U_Y \perp\!\!\!\perp U_M \mid W, S = s$. *Then Ass.4 holds.*

Note: Assumption (ii), for example, can be read as *W contains all common causes of S and M*.

Note 2: If we use X instead of $M = (X, Z)$, neither (iii) nor Ass.4 hold.

Identification

Under the above assumptions, we have

$$\theta(s_Y, s_M, s_W) = \int \big[\mathrm{E}(Y \mid A = 1, m, w, s_Y)\mathrm{dP}(m \mid A = 1, w, s_M) - \mathrm{E}(Y \mid A = 0, m, w, s_Y)\mathrm{dP}(m \mid A = 0, w, s_M)\big]\mathrm{dP}(w \mid s_W).$$

For estimation of this parameter, we want two main things:

- The ability to use flexible, data-adaptive regression methods in case M or W are continuous or multivariate
- The ability to study the sampling distribution of the estimator to construct frequentist valid uncertainty measures (confidence intervals, etc.)

Constructing estimators for θ

A crucial object to achieve this goals is a so-called von Mises expansion:

Definition 1 (First-order von Mises expansion)

A function $\varphi(O; \mathrm{P})$ that depends on the data and a distribution P satisfies a first-order von Mises expansion if the following holds

$$\theta(\mathrm{F}) - \theta(\mathrm{P}) = -\mathrm{E}_{\mathrm{P}}[\varphi(O; \mathrm{F})] + R(\mathrm{F}, \mathrm{P})$$

for any two distributions F and P, where $R(\mathrm{F}, \mathrm{P})$ is a second order term of the type $||\mathrm{F} - \mathrm{P}||^2$.

To see why this is so important, notice that:

- We can construct an estimate of the distribution $\hat{\mathrm{P}}$
- The von Mises expansion tells us that the first-order bias of a plug-in estimator is given by $-\mathrm{E}_{\mathrm{P}}[\varphi(O; \hat{\mathrm{P}})]$
- We can correct for this bias using one of two strategies:
  - Targeted minimum loss based estimation
  - One-step estimation

Extension to multiple studies

Recall the definition

$$\theta(s_Y, s_M, s_W) = \int \mathrm{E}[Y(1, M(1, s_M)) - Y(0, M(0, s_M)) \mid W, S = s_Y]\mathrm{dP}(W \mid S = s_W),$$

- Assume we now have K studies
- (s_Y, s_M, s_W) now take values in $\{1, \ldots, K\}^3$
- When $s_M = s_Y = s_W$, θ is the effect in the corresponding study.
- Let P_S denote a user-given joint distribution of S_Y, S_M, and S_W (e.g., independent, uniform).

We define the total between-study variability as

$$\tau^2 = \mathrm{Var}_S[\theta(S_Y.S_M, S_W)]$$

Extension to multiple studies

By using the law of total variance as

$$\tau^2 = \tau^2_{\text{EH}} + \tau^2_{\text{CM}},$$

where

$$\tau^2_{\text{CM}} = \text{E}_S\{\text{Var}_S[\theta(S_Y, S_M, S_W) \mid S_Y, S_M]\}$$
$$\tau^2_{\text{EH}} = \text{Var}_S\{\kappa(S_Y, S_M)\},$$

where $\kappa(S_Y, S_M) = \text{E}_S[\theta(S_Y, S_M, S_W) \mid S_Y, S_M]$.

Likewise, we decompose τ^2_{EH} into a pure effect modification parameter and a mediator variability parameter as

$$\tau^2_{\text{EH}} = \tau^2_{\text{EM}} + \tau^2_{\text{MV}},$$

where

$$\tau^2_{\text{EM}} = \text{E}_S\{\text{Var}_S[\kappa(S_Y, S_M) \mid S_Y]\}$$
$$\tau^2_{\text{MV}} = \text{Var}_S\{\text{E}[\kappa(S_Y, S_M) \mid S_Y]\}\}.$$

Application to MTO

- Baseline variables:
  - Child: race/ethnicity, age, history of behavioral problems, enrollment in programs for gifted and talented students.
  - Adult and neighborhood: education, married, was under 18 when child was born, currently working, currently receiving welfare, neighborhood safety perception, satisfaction with neighborhood, disability, size of household, poverty rate of neighborhood, etc.

- Intermediate variable: did the child ever attend a school that was not high-poverty during followup?

Parameter	S.D.	S.E. of S.D.	%	S.E. of %
τ	0.023	0.018	100.000	
τ_{CM}	0.016	0.014	49.620	11.160
τ_{EH}	0.016	0.013	50.380	11.160
τ_{EM}	0.005	0.006	5.911	5.713
τ_{MV}	0.015	0.013	44.470	12.960

Table 1: Results of our analyses[5]

[5] All results were approved for release by the U.S. Census Bureau, authorization number CBDRB-FY22-CES018-013, CBDRB-FY24-CES018-002

Thank you

Estimation details

We denote $\eta = (q_Y, q_M, e, e_M, g, g_M, h)$ and

$$
\begin{aligned}
q_Y(m, a, w, s_Y) &= \mathrm{E}(Y \mid a, m, w, s_Y) \\
q_M(a, w, s_Y, s_M) &= \mathrm{E}[q_Y(M, a, W, s_Y) \mid a, w, s_M] \\
e(s \mid w) &= \mathrm{P}(s \mid w) \\
e_M(s \mid m, w) &= \mathrm{P}(s \mid mw) \\
g(a \mid w, s) &= \mathrm{P}(a \mid w, s) \\
g_M(a \mid m, w, s) &= \mathrm{P}(a \mid m, w, s) \\
h(s) &= \mathrm{P}(s).
\end{aligned}
$$

Estimation details

Lemma 2 (von-Mises expansion)

Let $\theta(\mathsf{F})$ denote the parameter evaluated at an arbitrary distribution F. Let η_{F} denote the parameters corresponding to distribution F. Then we have

$$\theta(\mathsf{F}) - \theta(\mathsf{P}) = -\mathsf{E}_{\mathsf{P}}[\varphi(O; \eta_{\mathsf{F}})] + R(\eta_{\mathsf{P}}, \eta_{\mathsf{F}}),$$

where we add the index P to the expectation for clarity, and where R is a second-order term of the form

$$R(\eta_{\mathsf{P}}, \eta_{\mathsf{F}}) = \int c_1(\mathsf{P}, \mathsf{F})\{q_{Y,\mathsf{P}} - q_{Y,\mathsf{F}}\}\big[\{e_{\mathsf{P}} - e_{\mathsf{F}}\} + \{g_{\mathsf{P}} - g_{\mathsf{F}}\} + \{h_{\mathsf{F}} - h_{\mathsf{P}}\}\big]\mathrm{d}\mathsf{P}$$
$$+ \int c_2(\mathsf{P}, \mathsf{F})\{q_{M,\mathsf{P}} - q_{M,\mathsf{F}}\}\big[\{e_{M,\mathsf{P}} - e_{M,\mathsf{F}}\} + \{g_{M,\mathsf{P}} - g_{M,\mathsf{F}}\}\big]\mathrm{d}\mathsf{P},$$

where c_1 and c_2 are some transformations of F and P, and

$$\varphi(O; \mathsf{P}) = \frac{g_M(A \mid M, L, s_M)}{g_M(A \mid M, L, s_Y)} \frac{e_M(s_M \mid M, W)}{e_M(s_Y \mid M, W)} \frac{e(s_W \mid W)}{e(s_M \mid W)} \frac{(2A-1)I(S = s_Y)}{g(A \mid W, s_M)h(s_W)}\{Y - q_Y(M, A, W, s_Y)\}$$
$$+ \frac{e(s_W \mid W)}{e(s_M \mid W)} \frac{(2A-1)I(S = s_M)}{g(A \mid W, s_M)h(s_W)}\{q_Y(M, A, W, S) - q_M(A, W, s_Y, s_M)\}$$
$$+ \frac{I(S = s_W)}{h(s_W)}\{q_M(1, W, s_Y, s_M) - q_M(0, W, s_Y, s_M) - \theta(s_Y, s_M, s_W)\}$$