

Statistical Learning in Mediation Analysis

Chapter 5: Interventional mediation effects

David Benkeser
Emory University

Iván Díaz
New York University

Marco Carone
University of Washington

MODULE 14

**Summer Institute in Statistics for
Clinical and Epidemiological Research**

July 2024

Contents of this chapter

- 1 Natural vs. interventional effects with no exposure-induced confounding
- 2 Natural vs. interventional effects with exposure-induced confounding
- 3 Examples in R

Interventional direct and indirect effects

Recall that the **natural effects** involve interventions that, for example,

- set $A = 1$ and $M = M(1)$, the value M would naturally take under $A = 0$.

The challenge with identifying their effects is their **cross-world nature**.

- Intervening to set M equal to what your mediator would be under a^* .

Interventional effects consider a different form of interventions. For someone with covariates w , we

- set $A = a$ (as usual);
- set $M = M^*$, where M^* is a random draw from $M(a^*) \mid W = w$.

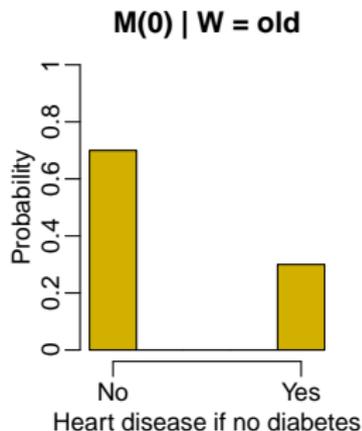
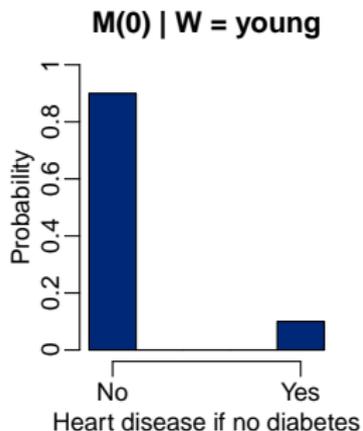
We are interested in the effect decomposition:

$$ATE = \underbrace{E[Y(1, M(1))] - E[Y(1, M^*)]}_{\text{interventional indirect effect}} + \underbrace{E[Y(1, M^*)] - E[Y(0, M(0))]}_{\text{interventional direct effect}} .$$

Interventional direct and indirect effects

Example: A = diabetes, M = heart disease, Y = all cause mortality, W = age at onset of diabetes. The **interventional indirect effect** compares:

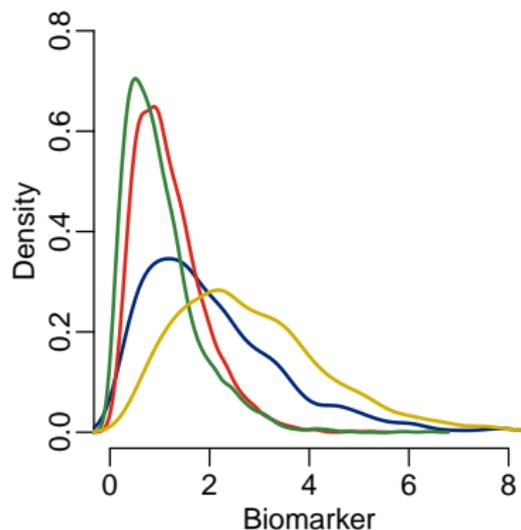
- set diabetes = 1 vs.
- set diabetes = 1 and set heart disease status to a random draw from the distribution of heart disease among non-diabetics of a similar age;
 - for young adults, draw from $M(0) | W = \text{young}$;
 - for older adults, draw from $M(0) | W = \text{old}$;



Interventional direct and indirect effects

Example: A = diabetes, M = biomarker, Y = all cause mortality, W_1 = age at onset of diabetes, W_2 = sex. The **interventional indirect effect** compares:

- set diabetes = 1 vs.
- set diabetes = 1 and set biomarker to a random draw from the distribution of biomarker among non-diabetics of a similar age and sex.



- for young females, draw from $M(0) \mid W_1 = \text{young}, W_2 = \text{female}$;
- for older females, draw from $M(0) \mid W_1 = \text{old}, W_2 = \text{female}$;
- for young males, draw from $M(0) \mid W_1 = \text{young}, W_2 = \text{male}$;
- for older males, draw from $M(0) \mid W_1 = \text{old}, W_2 = \text{male}$.

Interventional direct and indirect effects

Note that participant i could have $M_i(0) = 1$ but $M_i^* = 0$.

- The mediator value you get under our intervention might not be the same as the natural value your mediator would take under $A = 0$.

Example: Joe is an older adult who would only develop heart disease if he were diabetic ($M_{\text{Joe}}(0) = 0$); however, when implementing under our intervention, we happen to draw $M_{\text{Joe}}^* = 1$.

- In terms of natural indirect effects, Joe would contribute $Y_{\text{Joe}}(1, 0)$.
- In terms of interventional indirect effects, Joe would contribute $Y_{\text{Joe}}(1, 1)$.

However, in the population of people who are similar to Joe (i.e., older adults), the distribution of M^* = distribution of $M(0) \mid W = W_j$.

- Intervention is interesting **at a population level!**

Identification of interventional in/direct effects

With no exposure-induced confounders of M and Y , we require **three randomization assumptions** to identify $E[Y(a, M^*)]$.

- $Y(a, m) \perp A \mid W$
- $Y(a, m) \perp M \mid A = a, W$
- $M(a^*) \perp A \mid W$

Notably, **we do not require a cross-world assumption.**

We additionally require the **same positivity conditions** as for natural mediation effects:

- $P(A = a \mid W = w) > 0$ for all w ;
- $P(M = m \mid A = a^*, W = w) > 0$ implies $P(M = m \mid A = a^*, W = w) > 0$ for all m, w .

Identification of interventional in/direct effects

Under these assumptions, we have the following G-formula identification:

$$\begin{aligned} E[Y(a, M^*)] &= E(E[E(Y | A = a, W, M) | A = a^*, W]) \\ &= \sum_w \sum_m E(Y | A = a, W = w, M = m) P(M = m | A = a^*, W = w) P(W = w) . \end{aligned}$$

This is exactly the same as the G-formula for identification of natural effects!

Identification of interventional effects
=
Identification of natural effects.

In other words, the **same data analysis** may be **interpreted differently** depending on whether the **cross-world assumption** is believable.

- Is the tail wagging the dog?

Identification of interventional in/direct effects

Proof: For $a = 0, 1$ and M^* = a random draw from the distribution of $M(a^*) \mid W$,

$$\begin{aligned} & E[Y(a, M^*)] \\ &= \sum_w E[Y(a, M^*) \mid W = w]P(W = w) && \text{(tower rule)} \\ &= \sum_w \sum_m E[Y(a, m) \mid M^* = m, W = w]P(M^* = m \mid W = w)P(W = w) && \text{(tower rule)} \\ &= \sum_w \sum_m E[Y(a, m) \mid W = w]P(M(a^*) = m \mid W = w)P(W = w) && \text{(definition of } M^*) \\ &= \sum_w \sum_m E[Y(a, m) \mid A = a, W = w]P(M(a^*) = m \mid W = w)P(W = w) && \text{(randomization 1)} \\ &= \sum_w \sum_m (E[Y(a, m) \mid A = a, W = w] \\ &\quad \times P(M(a^*) = m \mid A = a^*, W = w)P(W = w)) && \text{(randomization 2)} \\ &= \sum_w \sum_m (E[Y(a, m) \mid A = a, M = m, W = w] \\ &\quad \times P(M(a^*) = m \mid A = a^*, W = w)P(W = w)) && \text{(randomization 3)} \\ &= \sum_w \sum_m E(Y \mid A = a, M = m, W = w)P(M = m \mid A = a^*, W = w)P(W = w) \\ &\hspace{15em} \text{(consistency)} \end{aligned}$$

Identification of interventional in/direct effects

The **key difference** between this identification proof and that for natural in/direct effects happens on **line 4**. Two things happen here:

1. We replace $P(M^* = m \mid W = w)$ with $P(M(a^*) = m \mid W = w)$.
 - M^* is a random draw from the distribution of $M(a^*) \mid W$.
2. We drop M^* from $E[Y(a, m) \mid M^* = m, W = w]$ and write $E[Y(a, m) \mid W = w]$.
 - M^* is a **random draw** from $M(a^*) \mid W$.
 - Once we know W , the **particular random value** that we draw tells us nothing about outcome. It's **drawn at random** – how could it?!

By its very construction $M^* \perp Y(a, m) \mid W$.

- The needed independence is moved from a **cross-world assumption** to the **definition of our causal quantity of interest**.

After this line, the proof continues **exactly as for natural effects**.

Interventional effects with exposure-induced confounding

When we have exposure-induced confounding of M and Y , we can define an **alternative effect decomposition** based on interventional effects.

- Let M^* be a random draw from $M(a^*) \mid W$.
- Let M° be a random draw from $M(a) \mid W$.

An **effect decomposition** based on interventional effects is

$$\underbrace{E[Y(a, M^\circ)] - E[Y(a, M^*)]}_{\text{total effect}} = \underbrace{E[Y(a, M^\circ)] - E[Y(a, M^*)]}_{\text{indirect effect}} + \underbrace{E[Y(a, M^*)] - E[Y(a^*, M^*)]}_{\text{direct effect}} .$$

Note that the **total effect** here is not the ATE.

- Under the intervention that defines the ATE, the mediator under intervention that, e.g., sets $A = a$ would have distribution $M(a) \mid Z(a), W$.
- Under the intervention that defines this total effect, the mediator under intervention that, e.g., sets $A = a$ has distribution $M(a) \mid W$.

For an effect decomposition of the ATE in terms of interventional effects see [Vansteelandt and Daniel \(2017\)](#).

Interventional effects with exposure-induced confounding

For this identification, we require three randomization conditions

- $Y(a, m) \perp A \mid W$
- $Y(a, m) \perp M \mid A = a, W, Z$
- $M(a^*) \perp A \mid W$

We need no unmeasured confounders of M and Y beyond W and Z .

- Unmeasured confounders of Z and Y are OK!

Interventional effects with exposure-induced confounding

We require two **positivity assumptions**:

1. $P(A = a \mid W = w) > 0$
2. For any z, w, m such that

$$P(Z = z \mid A = a, W = w) > 0 \text{ and } P(M = m \mid A = a^*, W = w) > 0 ,$$

we need $P(M = m \mid A = a, Z = z, W = w) > 0$.

Assumption 2 is a **stronger overlap condition** than before.

- Need overlap between $P(M \mid A = a^*, W = w)$ and $P(M = m \mid A = a, Z = z, W = w)$ **for every plausible value of z .**

Interventional effects with exposure-induced confounding

Example of violation of positivity condition 2:

A = vaccine, Z = asymptomatic infection, M = antibody level, Y = clinical disease.

Consider the following situation.

- Vaccine causes fever, effectively unblinding participants.
- Vaccinated people increase risk behaviors immediately and thus acquire asymptomatic infections prior to antibody measurements.
- Unvaccinated people are more conservative and acquire no asymptomatic infections prior to antibody measurements.
- There are some cases of vaccine failure, where no antibodies are generated.
- However, everyone produces antibodies in response to natural infection.

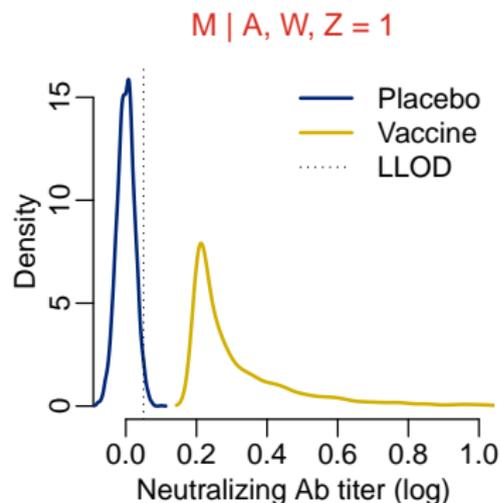
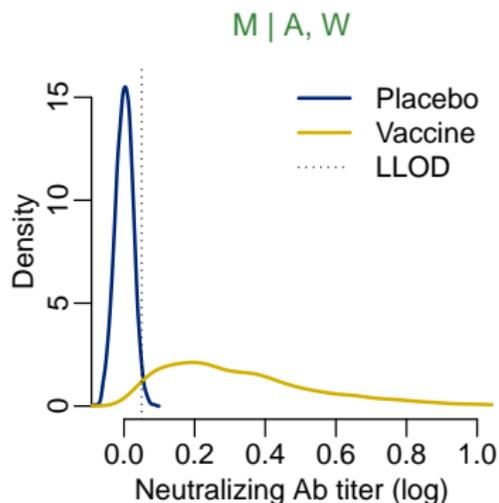
Interventional effects with exposure-induced confounding

In this case, we might have **overlap marginally**.

- Vaccine failures with no asymptomatic infection still have low antibodies.

But, we will have **no overlap conditional on Z**.

- Everyone with asymptomatic infection has positive antibodies.



Interventional effects with exposure-induced confounding

Under these assumptions, we can identify $E[Y(a, M^*)]$ where $M^* \sim M(a^*) \mid W$.

Let $\bar{Q}_a(z, m, w) = E(Y \mid A = a, Z = z, M = m, W = w)$. Then $E[Y(a, M^*)]$ equals

$$\sum_w \sum_m \sum_z \bar{Q}_a(z, m, w) P(Z = z \mid A = a, W = w) P(M = m \mid A = a^*, W = w) P(W = w) .$$

- First standardize the **outcome regression** with respect to $Z \mid A = a, W$.
- Next standardize with respect to the **mediator** using $M \mid A = a^*, W$.
- Finally, standardize with respect to the **covariates**.

Interventional effects with exposure-induced confounding

Proof: For $a = 0, 1$ and $M^* \sim M(a^*) \mid W$,

$$E[Y(a, M^*)] = \sum_w E[Y(a, M^*) \mid W = w]P(W = w) \quad (\text{tower rule})$$

$$\text{dog} = \sum_w \sum_m E[Y(a, m) \mid M^* = m, W = w]P(M^* = m \mid W = w)P(W = w) \quad (\text{tower rule})$$

$$= \sum_w \sum_m E[Y(a, m) \mid W = w]P[M(a^*) = m \mid W = w]P(W = w) \quad (\text{definition of } M^*)$$

$$= \sum_w \sum_m E[Y(a, m) \mid A = a, W = w]P[M(a^*) = m \mid W = w]P(W = w) \quad (\text{randomization 1})$$

$$= \sum_w \sum_m E[Y(a, m) \mid A = a, W = w]P[M(a^*) = m \mid A = a^*, W = w]P(W = w) \quad (\text{randomization 3})$$

$$= \sum_w \sum_m \sum_z (E[Y(a, m) \mid A = a, Z = z, W = w]P(Z = z \mid A = a, W = w) \times P[M(a^*) = m \mid A = a^*, W = w])P(W = w) \quad (\text{tower rule})$$

$$= \sum_w \sum_m \sum_z (E[Y(a, m) \mid A = a, Z = z, M = m, W = w]P(Z = z \mid A = a, W = w) \times P[M(a^*) = m \mid A = a^*, W = w])P(W = w) \quad (\text{randomization 2})$$

$$= \sum_w \sum_m \sum_z [E(Y \mid A = a, Z = z, M = m, W = w)P(Z = z \mid A = a, W = w) \times P(M = m \mid A = a^*, W = w)]P(W = w) \quad (\text{consistency})$$

Interventional effects with exposure-induced confounding

Here, we consider estimation in the simplest setting, where M and Z are binary.

```
set.seed(123)
# simulate some data
n <- 5000
# treatment/outcome confounder
W1 <- rnorm(n)
# treatment/mediator confounder
W2 <- rbinom(n, 1, 0.5)
# mediator/outcome confounder
W3 <- runif(n)
A <- rbinom(n, 1, plogis(-1 + W1 / 3 + W2 / 4))
Z <- rbinom(n, 1, plogis(-2 + A / 2))
M <- rbinom(n, 1, plogis(-2 + A / 2 - Z / 2 + W2 / 4))
Y <- rbinom(n, 1, plogis(-1 + M / 2 + A / 4 - Z / 2 + W1 / 4 - W3 / 4))
full_data <- data.frame(W1 = W1, W2 = W2, W3 = W3, A = A, Z = Z, M = M, Y = Y)
```

The true values are:

- total effect = $E[Y(1, M^1)] - E[Y(0, M^0)] = 0.050$
- indirect effect = $E[Y(1, M^1)] - E[Y(1, M^0)] = 0.008$
- direct effect = $E[Y(1, M^0)] - E[Y(0, M^0)] = 0.042$

Interventional effects with exposure-induced confounding

To estimate the effects of interest, we need to fit three regression models:

- outcome regression = $E(Y | A, Z, M, W)$
- mediator regression = $P(M = 1 | A, W)$
- confounder regression = $P(Z = 1 | A, W)$

```
# outcome regression
or_fit <- glm(Y ~ A + Z + M + W1 + W2 + W3,
             family = binomial(), data = full_data)
# mediator regression
med_fit <- glm(M ~ A + W1 + W2 + W3, family = binomial(), data = full_data)
# confounder regression
z_fit <- glm(Z ~ A + W1 + W2 + W3, family = binomial(), data = full_data)
```

Interventional effects with exposure-induced confounding

To compute all effects of interest, we need estimates

- $\bar{Q}_{n,1}(z, m, W_i)$ for $z = 0, 1$ and $m = 0, 1$ and $i = 1, \dots, n$.
- $\bar{Q}_{n,0}(z, m, W_i)$ for $z = 0, 1$ and $m = 0, 1$ and $i = 1, \dots, n$.

```
Qbar_na_zm <- function(or_fit, a, z, m, full_data){
  pred_data <- full_data
  pred_data$A <- a; pred_data$M <- m; pred_data$Z <- z
  pred <- predict(or_fit, type = "response", newdata = pred_data)
  return(pred)
}
# a = 1
Qbar_n1_z1m1 <- Qbar_na_zm(or_fit, a = 1, z = 1, m = 1, full_data)
Qbar_n1_z1m0 <- Qbar_na_zm(or_fit, a = 1, z = 1, m = 0, full_data)
Qbar_n1_z0m1 <- Qbar_na_zm(or_fit, a = 1, z = 0, m = 1, full_data)
Qbar_n1_z0m0 <- Qbar_na_zm(or_fit, a = 1, z = 0, m = 0, full_data)
# a = 0
Qbar_n0_z1m1 <- Qbar_na_zm(or_fit, a = 0, z = 1, m = 1, full_data)
Qbar_n0_z1m0 <- Qbar_na_zm(or_fit, a = 0, z = 1, m = 0, full_data)
Qbar_n0_z0m1 <- Qbar_na_zm(or_fit, a = 0, z = 0, m = 1, full_data)
Qbar_n0_z0m0 <- Qbar_na_zm(or_fit, a = 0, z = 0, m = 0, full_data)
```

Interventional effects with exposure-induced confounding

We also need estimates

- $\hat{P}_n(Z = z | A = 1, W = W_i)$ for $i = 1, \dots, n$.
- $\hat{P}_n(Z = z | A = 0, W = W_i)$ for $i = 1, \dots, n$.

```
Phat_n_Z1_a <- function(z_fit, a, full_data){  
  pred_data <- full_data  
  pred_data$A <- a  
  pred <- predict(z_fit, type = "response", newdata = pred_data)  
  return(pred)  
}  
# A = 1  
Phat_n_Z1_a1 <- Phat_n_Z1_a(z_fit, a = 1, full_data)  
Phat_n_Z0_a1 <- 1 - Phat_n_Z1_a1  
# A = 0  
Phat_n_Z1_a0 <- Phat_n_Z1_a(z_fit, a = 0, full_data)  
Phat_n_Z0_a0 <- 1 - Phat_n_Z1_a0
```

Interventional effects with exposure-induced confounding

Finally, we need estimates of the mediator distribution

- $\hat{P}_n(M = m \mid A = 1, W = W_i)$ for $m = 0, 1$ and $i = 1, \dots, n$
- $\hat{P}_n(M = m \mid A = 0, W = W_i)$ for $m = 0, 1$ and $i = 1, \dots, n$

```
Phat_n_M1_a <- function(med_fit, a, full_data){  
  pred_data <- full_data  
  pred_data$A <- a  
  pred <- predict(med_fit, type = "response", newdata = pred_data)  
  return(pred)  
}  
# A = 1  
Phat_n_M1_a1 <- Phat_n_M1_a(med_fit, a = 1, full_data)  
Phat_n_M0_a1 <- 1 - Phat_n_M1_a1  
# A = 0  
Phat_n_M1_a0 <- Phat_n_M1_a(med_fit, a = 0, full_data)  
Phat_n_M0_a0 <- 1 - Phat_n_M1_a0
```

Interventional effects with exposure-induced confounding

Now we can compute estimates of the components of the effects of interest.

```
# E[Y(1, M^1)]
EY1M1 <- mean(
  # terms in sum for z = 0, m = 0
  Qbar_n1_z0m0 * Phat_n_Z0_a1 * Phat_n_M0_a1 +
  # terms in sum for z = 1, m = 0
  Qbar_n1_z1m0 * Phat_n_Z1_a1 * Phat_n_M0_a1 +
  # terms in sum for z = 0, m = 1
  Qbar_n1_z0m1 * Phat_n_Z0_a1 * Phat_n_M1_a1 +
  # terms in sum for z = 1, m = 1
  Qbar_n1_z1m1 * Phat_n_Z1_a1 * Phat_n_M1_a1
)

EY1M1

## [1] 0.2777025
```

Interventional effects with exposure-induced confounding

Now we can compute estimates of the components of the effects of interest.

```
# E[Y(1, M^0)]
EY1M0 <- mean(
  # terms in sum for z = 0, m = 0
  Qbar_n1_z0m0 * Phat_n_Z0_a1 * Phat_n_M0_a0 +
  # terms in sum for z = 1, m = 0
  Qbar_n1_z1m0 * Phat_n_Z1_a1 * Phat_n_M0_a0 +
  # terms in sum for z = 0, m = 1
  Qbar_n1_z0m1 * Phat_n_Z0_a1 * Phat_n_M1_a0 +
  # terms in sum for z = 1, m = 1
  Qbar_n1_z1m1 * Phat_n_Z1_a1 * Phat_n_M1_a0
)

EY1M0

## [1] 0.2702332
```

Interventional effects with exposure-induced confounding

Now we can compute estimates of the components of the effects of interest.

```
# E[Y(1, M^0)]
EYOM0 <- mean(
  # terms in sum for z = 0, m = 0
  Qbar_n0_z0m0 * Phat_n_Z0_a0 * Phat_n_M0_a0 +
  # terms in sum for z = 1, m = 0
  Qbar_n0_z1m0 * Phat_n_Z1_a0 * Phat_n_M0_a0 +
  # terms in sum for z = 0, m = 1
  Qbar_n0_z0m1 * Phat_n_Z0_a0 * Phat_n_M1_a0 +
  # terms in sum for z = 1, m = 1
  Qbar_n0_z1m1 * Phat_n_Z1_a0 * Phat_n_M1_a0
)

EYOM0

## [1] 0.2421124
```

Interventional effects with exposure-induced confounding

Finally, we can compute the effects of interest.

```
# total effect
```

```
EY1M1 - EY0M0
```

```
## [1] 0.03559014
```

```
# indirect effect
```

```
EY1M1 - EY1M0
```

```
## [1] 0.007469372
```

```
# direct effect
```

```
EY1M0 - EY0M0
```

```
## [1] 0.02812076
```

References and additional reading

References:

Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*. PMC: [PMC5289540](#).

Additional reading:

Díaz I, Hejazi NS, Rudolph KE, van der Laan MJ. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*. doi: [10.1093/biomet/asaa085](#).