Statistical Learning in Mediation Analysis

Chapter 2: Controlled direct effects

David Benkeser
Emory University

Iván Díaz
New York University

Marco Carone
University of Washington

MODULE 13

Summer Institute in Statistics for Clinical and Epidemiological Research

July 2023

Contents of this chapter

What is it and when is it identified?

Suppose that we are interested in quantifying the effect of a binary treatment A on outcome Y **through pathways not including the potential mediator** M.

We can imagine an intervention in which we set the level of A and M at will, resulting in the definition of the counterfactual outcome $Y(a, m)$. For an individual,

$$Y(1, m) - Y(0, m)$$

represents the (additive) effect of treatment when mediator level is fixed at m. Averaging over the target population leads us to

$$CDE(m) := E[Y(1, m) - Y(0, m)] \ ,$$

the **controlled direct effect of** A **on** Y **controlling for** M **at level** m.

Examples: What is the effect of...

...screen time on weight not through physical activity in children?
...SARS-CoV-2 infection on mortality not through modulation of IL-6 cytokine levels?
...a mRNA vaccine on risk of Covid not through anti-spike IgG antibody titer?

What is it and when is it identified?

We will focus on the case in which M is discrete.

To compute the controlled direct effect, it suffices to compute the counterfactual means $E[Y(1,m)]$ and $E[Y(0,m)]$ since

$$CDE(m) \;=\; E[Y(1,m) - Y(0,m)] \;=\; E[Y(1,m)] - E[Y(0,m)] \;.$$

The observed data consist of $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_0$, with $O_i := (W_i, A_i, M_i, Y_i)$ and

W_i = the vector of baseline patient characteristics (i.e., potential confounders);
A_i = the (binary) treatment/intervention received;
M_i = the mediator value experienced;
Y_i = the outcome of interest experienced.

Fundamental question:

When and how is $E[Y(a,m)]$ identifiable (i.e., estimable) from the observed data?

What is it and when is it identified?

First key condition: **conditional randomization** (or ignorability)

To learn causal effects from observational data, we typically need to have rich enough information recorded on patients to deconfound observed relationships.

This is formalized via the two-part conditional randomization condition:

1. $Y(a, m) \perp A \mid W$ (conditional treatment randomization)

   Within strata of W, the assignment of A gives no info about $Y(a, m)$.

   - Confounders of $A - Y$ relationship must have been recorded.
   - Can be enforced by design.
   - Similar to what is needed to estimate ATE of A on Y.

What is it and when is it identified?

First key condition: **conditional randomization** (or ignorability)

To learn causal effects from observational data, we typically need to have rich enough information recorded on patients to deconfound observed relationships.

This is formalized via the two-part conditional randomization condition:

2. $Y(a, m) \perp M \mid W, A = a$ (conditional mediator randomization given treatment)

   Within strata of W and $A = a$, the assignment of M gives no info about $Y(a, m)$.

   - Confounders of $M - Y$ relationship must have been recorded.
   - May or may not be enforceable by design, depending on context...

What is it and when is it identified?

First key condition: **conditional randomization** (or ignorability)

To learn causal effects from observational data, we typically need to have rich enough information recorded on patients to deconfound observed relationships.

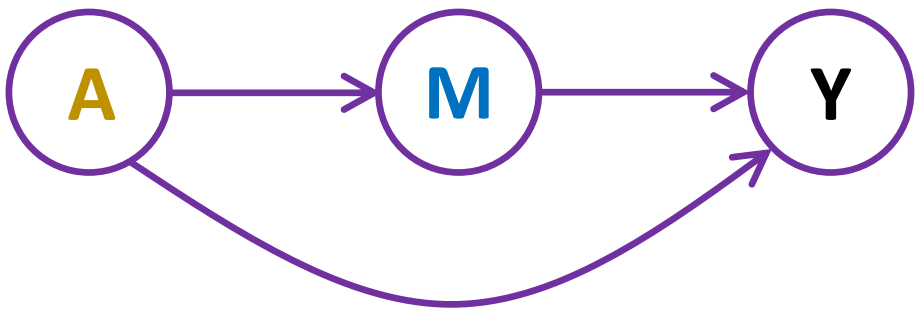
This is formalized via the two-part conditional randomization condition:

1. $Y(a, m) \perp A \mid W$ (conditional treatment randomization)
2. $Y(a, m) \perp M \mid W, A = a$ (conditional mediator randomization given treatment)

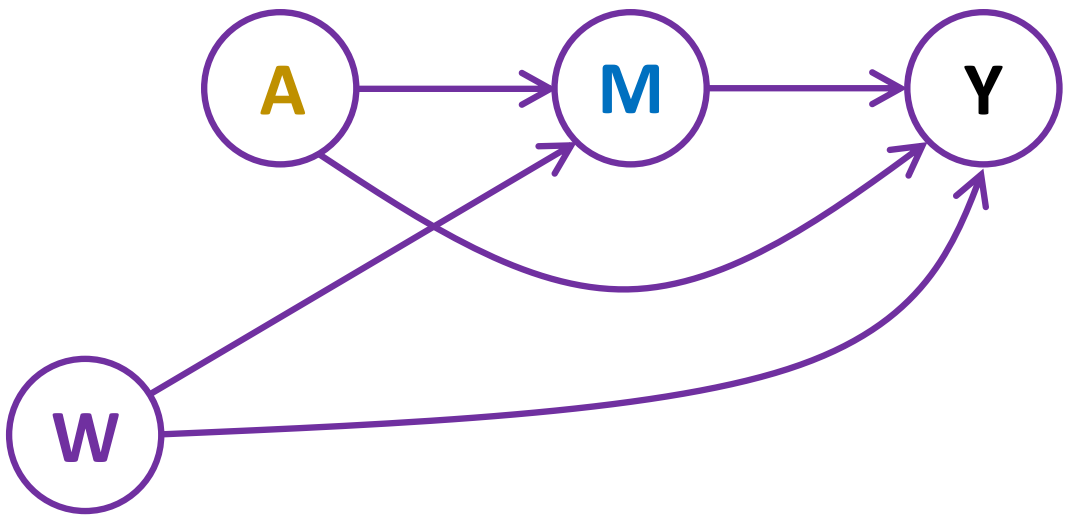
Without additional information, the conditional randomization condition is untestable (or empirically unverifiable) since it does not constrain the observed data distribution.

This condition must be justified by prior knowledge and scrutinized carefully.

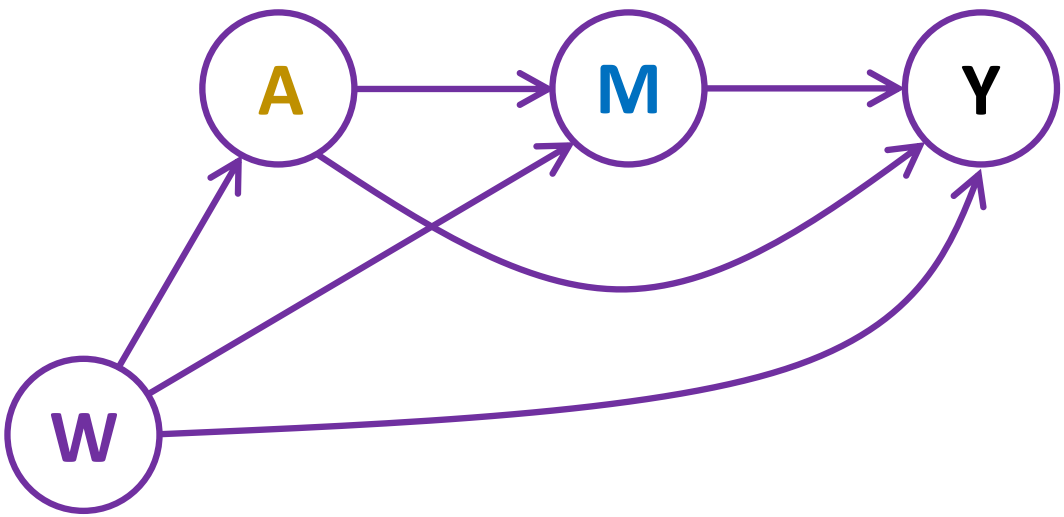
What is it and when is it identified?



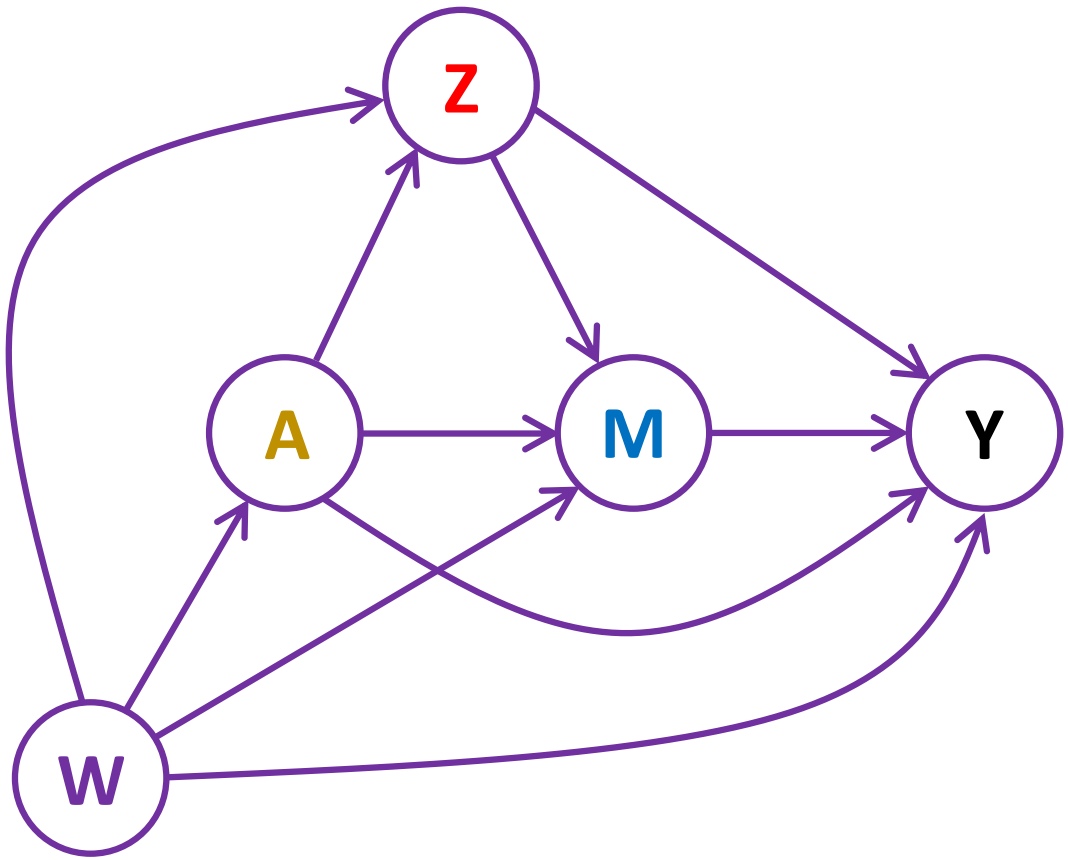
What is it and when is it identified?



What is it and when is it identified?



What is it and when is it identified?



When is identification possible?

Second key condition: **positivity** (or experimental treatment assignment)

We must also be able to observe treatment level a and mediator value m in each relevant patient subpopulation.

This is formalized via the two-part positivity condition:

1. $P(A = a \mid W = w) > 0$ for every possible w (treatment positivity)
2. $P(M = m \mid A = a, W = w) > 0$ for every possible w (mediator positivity)

In a randomized trial, 1 is usually true by design, but 2 may fail unless M is set as part of the randomized intervention. In an observational study, both 1 and 2 can fail.

e.g.: patients with mild disease cannot be assigned to (risky) experimental treatment
immunosuppressed patients cannot generate high antibody levels after vaccine

The plausibility of this condition can usually be assessed empirically.

What is it and when is it identified?

Under the randomization and positivity conditions stated, $E[Y(a, m)]$ can generally be identified from the observed data.

It can be calculated as a summary of the distribution P_0 of the observed data unit O.

We will focus on the two most important identification formulas:

- the G-computation formula; (Robins, 1986)
- the inverse-probability-weighting (IPW) formula.
  (Horvitz & Thompson, 1952; Robins, Hernan & Brumback, 2000)

The G-computation identification formula

The **G-computation** identification formula gives an expression of $CDE(m)$ in terms of the observed data distribution.

Provided the randomization and positivity conditions hold for each a, it holds that

$$\begin{aligned} E[Y(a,m)] &= E\left[E(Y \mid A=a, M=m, W)\right] \\ &= \sum_w E(Y \mid A=a, M=m, W=w)P(W=w) \ , \end{aligned}$$

and so, the controlled direct effect $CDE(m)$ can be expressed as

$$\begin{aligned} CDE(m) &= E\left[E(Y \mid A=1, M=m, W) - E(Y \mid A=0, M=m, W)\right] \\ &= \sum_w \left[E(Y \mid A=1, M=m, W=w) - E(Y \mid A=0, M=m, W=w)\right] P(W=w) \ . \end{aligned}$$

Is it fair to compare $E(Y \mid A=1, M=m)$ and $E(Y \mid A=0, M=m)$?

What about comparisons within strata of W?

The G-computation identification formula

The G-computation formula can be derived as follows.

$$
\begin{aligned}
E[Y(a,m)] &= \sum_w E[Y(a,m) \mid W = w] P(W = w) && \text{(law of total expectation)} \\
&= \sum_w E[Y(a,m) \mid A = a, M = m, W = w] P(W = w) && \text{(randomization property)} \\
&= \sum_w E(Y \mid A = a, M = m, W = w) P(W = w) && \text{(consistency)}
\end{aligned}
$$

For $E(Y \mid A = a, M = m, W = w)$ to be defined, the positivity condition must hold.

The G-computation identification formula

Special case: partially linear outcome regression model

If the partially linear model

$$E(Y \mid M = m, A = a, W = w) = \beta_M m + \beta_A a + \beta_{MA} ma + f(w)$$

holds for some unspecified function f, then

$$CDE(m) = \beta_A + \beta_{MA} m$$

since $E(Y \mid M = m, A = 1, W = w) - E(Y \mid M = m, A = 0, W = w) = \beta_A + \beta_{MA} m$.

In the presence of interactions involving W, the form typically also depends on the distribution of W.

The IPW identification formula

The **inverse-probability-weighting (IPW)** identification formula gives an alternative means of expressing $CDE(m)$ in terms of the observed data distribution.

Provided the randomization and positivity conditions hold, then it holds that

$$E[Y(a,m)] \;=\; E\left[\left\{\frac{I(A=a,M=m)}{P(A=a,M=m \mid W)}\right\}Y\right].$$

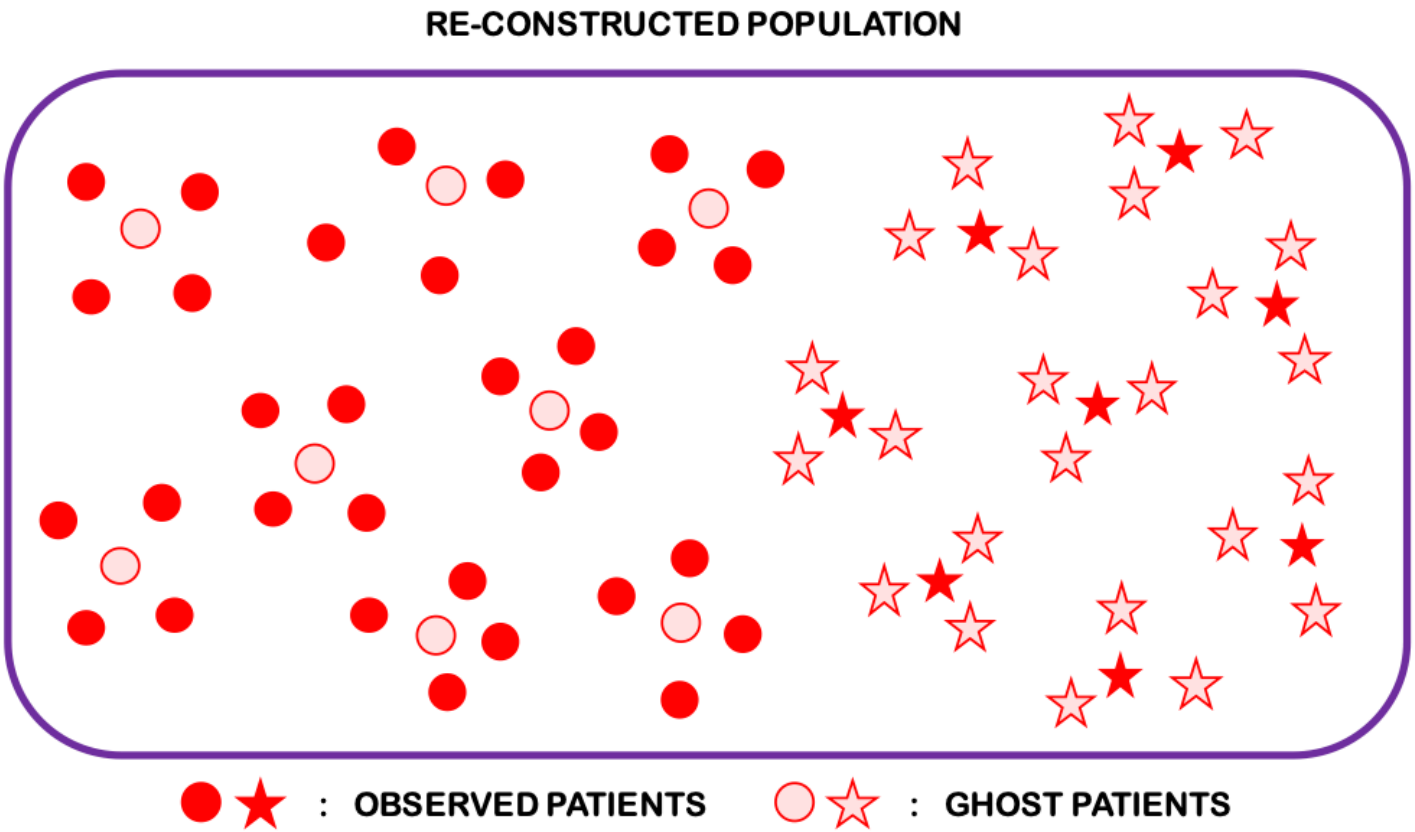
This is a weighted average of outcomes, with weight set according to the propensity of having had $(A,M)=(a,m)$ in the first place given $W=w$.

If $P(M=0,A=1 \mid W=w)=.25$, a patient with $W=w$ had a 25% chance of being treated and then experiencing mediator level $M=0$.

For each patient with $A=1$ and $M=0$, there are 3 similar patients with $A\neq 1$ or $M\neq 0$. Each such patient must stand in for the other 3, and so, have weight

$$\frac{1}{P(M=0,A=1 \mid W=w)} \;=\; \frac{1}{0.25} \;=\; 4\,.$$

The IPW identification formula



$$P(M = 0, A = 1 \mid W = \bigstar) = 0.25 \qquad P(M = 0, A = 1 \mid W = \bullet) = 0.80$$

The IPW identification formula

The IPW formula is equivalent to the G-computation formula.

By repeated use of the law of total expectation, we have that

$$
\begin{aligned}
E\left[\frac{I(A=a,M=m)Y}{P(A=a,M=m\mid W)}\right] &= E\left[E\left[\frac{I(A=a,M=m)Y}{P(A=a,M=m\mid W)}\middle|A,M,W\right]\right] \\
&= E\left[\frac{I(A=a,M=m)}{P(A=a,M=m\mid W)}E\left(Y\mid A,M,W\right)\right] \\
&= E\left[\frac{I(A=a,M=m)}{P(A=1,M=m\mid W)}E\left(Y\mid A=a,M=m,W\right)\right] \\
&= E\left[E\left[\frac{I(A=a,M=m)}{P(A=a,M=m\mid W)}E\left(Y\mid A=a,M=m,W\right)\middle|W\right]\right] \\
&= E\left[\frac{P(A=a,M=m\mid W)}{P(A=a,M=m\mid W)}E(Y\mid A=a,M=m,W)\right] \\
&= E\left[E(Y\mid A=a,M=m,W)\right] \ .
\end{aligned}
$$

Estimation based on the G-computation and IPW formulas

Via the identification formulas, we express quantities we care about in the counterfactual world as quantities defined in the observed data world.

This required certain **causal assumptions**.

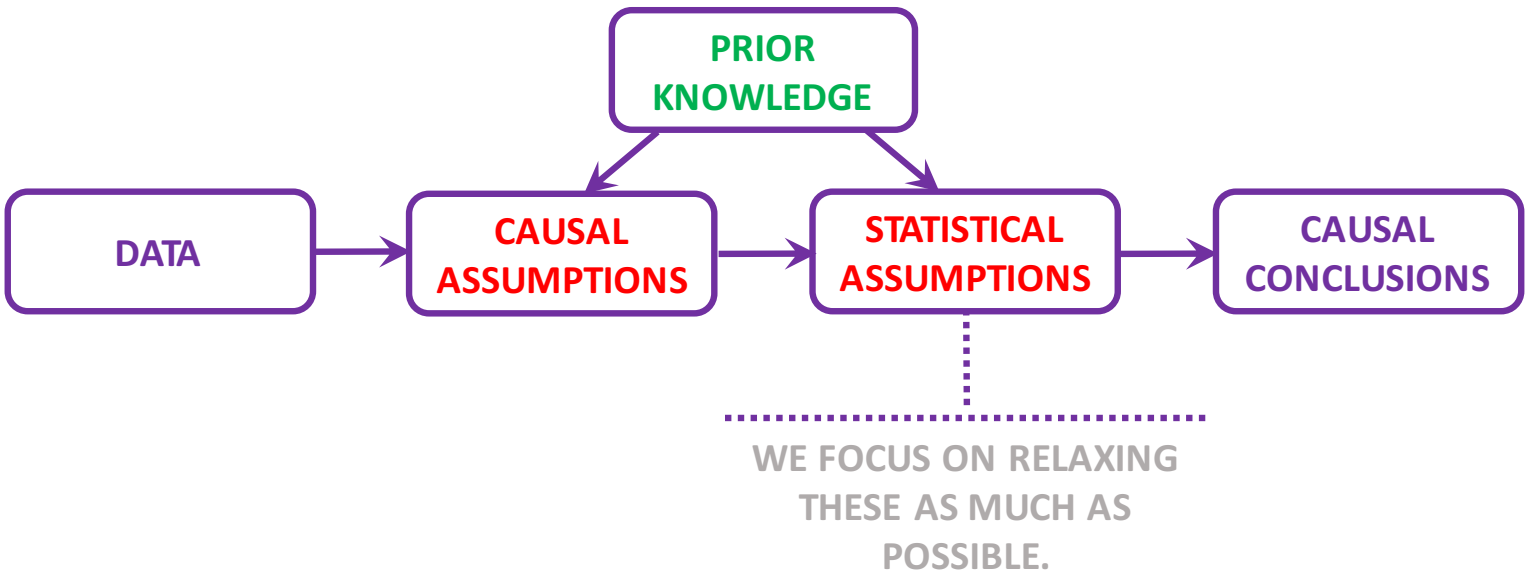
- Many of these are empirically unverifiable, and so cannot be relaxed for free.
- Alternative assumptions exist. Otherwise, partial identification is possible under weaker assumptions.

This is certainly progress since we can estimate quantities in the observed data world!

Practitioners make **statistical assumptions** of varying degrees to tackle the resulting estimation/inference problem.

- Most of these are verifiable and thus unnecessary (except for convenience).
- The approach we advocate for uses modern statistical learning to reduce the risk of misleading conclusions due to inappropriate statistical assumptions.

Estimation based on the G-computation and IPW formulas



Estimation based on the G-computation and IPW formulas

Several quantities (defined in the observed data world) play a critical role in the methods we will describe:

$$\begin{aligned} \text{the outcome regression :} \quad & \bar{Q}(m,a,w) := E(Y \mid M=m, A=a, W=w) \\ \text{the propensity scores :} \quad & g_M(m \mid a,w) := P(M=m \mid A=a, W=w) \\ & g_A(a \mid w) := P(A=a \mid W=w) \quad . \end{aligned}$$

The various methods we will discuss explicitly require estimates of $\bar{Q}$, g_M and g_A.

In the following, we denote by $\bar{Q}_n$, $g_{M,n}$ and $g_{A,n}$ estimators of $\bar{Q}$, g_M and g_A, respectively.

Estimation based on the G-computation and IPW formulas

Plug-in estimation via the G-computation formula

$$\begin{aligned} CDE(m) &= E[\bar{Q}(m,1,W) - \bar{Q}(m,0,W)] \\ CDE_{n,G}(m) &:= \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n(m,1,W_i) - \bar{Q}_n(m,0,W_i)\right] \end{aligned}$$

Plug-in estimation via the IPW formula

$$\begin{aligned} CDE(m) &= E\left[\left\{\frac{I(A=1,M=m)}{g_M(m \mid 1,W)g_A(1,W)} - \frac{I(A=0,M=m)}{g_M(m \mid 0,W)g_A(0,W)}\right\}Y\right] \\ CDE_{n,IPW}(m) &:= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i=1,M_i=m)}{g_{M,n}(m \mid 1,W)g_{A,n}(1,W_i)} - \frac{I(A_i=0,M_i=m)}{g_{M,n}(m \mid 0,W_i)g_{A,n}(0,W_i)}\right\}Y_i \end{aligned}$$

Provided estimators $\bar{Q}_n$, $g_{M,n}$ and $g_{A,n}$ are obtained by fitting parametric models or using empirical moment estimators, the standard bootstrap can be used for inference.

Estimation based on the G-computation and IPW formulas

We illustrate the approaches using simulated data.

```
set.seed(1234)
n <- 5000
# confounder of A/Y
W1 <- rnorm(n)
# confounder of M/Y
W2 <- rnorm(n)
# treatment
A <- rbinom(n, 1, plogis(-1 + W1 / 2))
# binary mediator
M <- rbinom(n, 1, plogis(-2 + A / 2 + W2 / 3))
# binary outcome
Y <- rbinom(n, 1, plogis(-1 + A - M / 2 + W1 / 3 + W2 / 3))
full_data <- data.frame(W1 = W1, W2 = W2, A = A, M = M, Y = Y)
```

Here, we have that $CDE(0) = 0.223$ and $CDE(1) = 0.200$.

Estimation based on the G-computation and IPW formulas

Suppose we are interested in estimating $CDE(0)$.

```
# fit outcome regression
or_fit <- glm(Y ~ A + M + W1 + W2, family = binomial(), data = full_data)
# new data setting A and M
data_A1_M0 <- data_A0_M0 <- full_data
data_A1_M0$A <- 1; data_A1_M0$M <- 0
data_A0_M0$A <- 0; data_A0_M0$M <- 0
# predict on new data
Qbar_A1_M0 <- predict(or_fit, newdata = data_A1_M0, type = "response")
Qbar_A0_M0 <- predict(or_fit, newdata = data_A0_M0, type = "response")
# gcomp estimate of CDE(0)
mean(Qbar_A1_M0 - Qbar_A0_M0)
```

```
## [1] 0.2515527
```

Estimation based on the G-computation and IPW formulas

Here, we demonstrate the two-part estimation of the propensity scores.

```
# model for P(A = 1 | W)
ps_fit1 <- glm(A ~ W1 + W2, family = binomial(), data = full_data)
P_A1_W <- predict(ps_fit1, type = "response")
P_A0_W <- 1 - P_A1_W
# model for P(M = 0 | A, W)
ps_fit2 <- glm(M ~ A + W1 + W2, family = binomial(), data = full_data)
# P(M = 0 | A = 1, W)
data_A1 <- full_data; data_A1$A <- 1
P_M0_A1_W <- 1 - predict(ps_fit2, newdata = data_A1, type = "response")
# P(M = 0 | A = 0, W)
data_A0 <- full_data; data_A0$A <- 0
P_M0_A0_W <- 1 - predict(ps_fit2, newdata = data_A0, type = "response")
# ipw estimate of CDE(0)
mean( (A == 1) / P_A1_W * (M == 0) / P_M0_A1_W  * Y ) -
  mean( (A == 0) / P_A0_W * (M == 0) / P_M0_A0_W * Y )
```

```
## [1] 0.2553091
```

Estimation based on the G-computation and IPW formulas

In practice, which of these two approaches should we adopt?

If $\bar{Q}$ is easier to estimate well, G-computation seems like a good bet. If instead the propensity scores are easier to estimate well, the IPW approach is sensible. In reality, we can improve upon both estimators, as we present next.

In any case, we need to estimate at least one of $\bar{Q}$ or (g_M, g_A).

There are many approaches possible for estimating a regression function, ranging from very flexible (e.g., nonparametric methods) to rather rigid (e.g., parametric methods).

- (nonparametric) empirical moment, kernel regression, neural networks, random forests;
- (semiparametric) generalized additive models, partially linear additive models;
- (parametric) linear regression, logistic regression, spline regression.

It is often a good idea to do principled ensembling (e.g, Super Learning).

Important caveat:

valid inference is difficult to achieve when using flexible learning to build G-computation or IPW estimators.

Doubly-robust estimation

The **augmented IPW (AIPW) estimator** is a hybrid estimator combining both approaches seen so far. It is defined as

$$CDE_{n,AIPW}(m) := CDE_{n,G}(m) + B_n(\bar{Q}_n, g_{M,n}, g_{A,n}) \ ,$$

where we define the augmentation term

$$\begin{aligned} B_n(\bar{Q}_n, g_{M,n}, g_{A,n}) \ := \ & \frac{1}{n}\sum_{i=1}^{n} \frac{I(M_i = m, A_i = 1)}{g_{M,n}(m \mid 1, W_i) g_{A,n}(1 \mid W_i)} [Y_i - \bar{Q}_n(m, 1, W_i)] \\ & - \frac{1}{n}\sum_{i=1}^{n} \frac{I(M_i = m, A_i = 0)}{g_{M,n}(m \mid 0, W_i) g_{A,n}(0 \mid W_i)} [Y_i - \bar{Q}_n(m, 0, W_i)] \ . \end{aligned}$$

The AIPW estimator can also be seen as an augmentation of the IPW estimator.

Doubly-robust estimation

Properties of the AIPW estimator:

- Doubly-robust consistency

  $CDE_{n,AIPW}(m) \xrightarrow{P} CDE(m)$ provided $\bar{Q}_n \xrightarrow{P} \bar{Q}$ or $(g_{A,n}, g_{M,n}) \xrightarrow{P} (g_A, g_M)$.

- Asymptotic normality

  Under certain regularity conditions (allowing some flexible learning), we have that

  $$\sqrt{n}\left[CDE_{n,AIPW}(m) - CDE(m)\right] \xrightarrow{d} N(0, \sigma^2) \ ,$$

  where σ^2 can be estimated consistently by $\sigma_n^2 := \frac{1}{n}\sum_{i=1}^{n}(D_{i,n} - \bar{D}_n)^2$ and

  $$\begin{aligned} D_{i,n} := & \frac{I(M_i = m, A_i = 1)}{g_{M,n}(m \mid 1, W_i) g_{A,n}(1 \mid W_i)}[Y_i - \bar{Q}_n(m, 1, W_i)] \\ & - \frac{I(M_i = m, A_i = 0)}{g_{M,n}(m \mid 0, W_i) g_{A,n}(0 \mid W_i)}[Y_i - \bar{Q}_n(m, 0, W_i)] + \bar{Q}_n(m, 1, W_i) - \bar{Q}_n(m, 0, W_i) \ . \end{aligned}$$

Doubly-robust estimation

```
# aipw estimate of E[Y(1,0)]
aiptw_EY_A1_M0 <- mean(Qbar_A1_M0) +
  mean( (A == 1) / P_A1_W * (M == 0) / P_M0_A1_W  * (Y - Qbar_A1_M0) )

# aipw estimate of E[Y(0,0)]
aiptw_EY_A0_M0 <- mean(Qbar_A0_M0) +
  mean( (A == 0) / P_A0_W * (M == 0) / P_M0_A0_W  * (Y - Qbar_A0_M0) )

# aipw estimate of CDE(0)
aiptw_EY_A1_M0 - aiptw_EY_A0_M0
```

```
## [1] 0.2554265
```

Doubly-robust estimation

Can we find a good estimator $\bar{Q}_n^*$ of $\bar{Q}$ for which, given estimators $g_{A,n}$ and $g_{M,n}$,

$$CDE_{n,G}(m) = CDE_{n,AIPW}(m) \ ?$$

This would require, in particular, that $B_n(\bar{Q}_n^*, g_{M,n}, g_{A,n}) = 0$.

The **targeted minimum loss-based estimation (TMLE)** algorithm can be used to revise a given estimator $\bar{Q}_n$ of $\bar{Q}$ into an estimator $\bar{Q}_n^*$ that achieves precisely this goal.

The resulting estimator $CDE_{n,TMLE}(m)$ has the same large-sample properties as $CDE_{n,AIPW}(m)$ but can have better small-sample performance.

Inferential approaches used for $CDE_{n,AIPW}(m)$ remain valid for $CDE_{n,TMLE}(m)$.

A tip on implementation:

use software for estimating ATE contrasting "treatment" levels $A^* = 1$ versus $A^* = 0$, where we define $A^* := I(M = m, A = 1) - I(M \neq m)$.

What about continuous mediators?

The methods discussed so far are most relevant when M can only take a few values, so that a sufficient number of individuals are observed with $M = m$.

In many cases though, M can take many (even infinitely many) values. What then?

The G-computation (but not IPW) formula still provides a good way forward.

Some strategies for dealing with continuous mediators:

1. Consider a discretization $M_\sharp$ of M.
   - No need to change inferential procedure (except if discretization is data-driven).
   - If they hold with M, identifying conditions will also hold with M replaced by $M_\sharp$.
   - Definition of resulting CDE then depends on sampling distribution of M.
2. Use prior knowledge about the shape of the CDE curve.
   - Shape constraints (e.g., monotonicity) can be leveraged to perform valid inference (though only at irregular rates). (Westling et al., 2020)
   - Smoothness constraints can also be leveraged but inference is then more challenging.
3. Focus instead on a summary of the CDE curve.
   - Coefficients of projection onto a parametric model is a regression-inspired approach.

What about continuous mediators?

Example: **assessing the role of antibodies in preventing COVID**

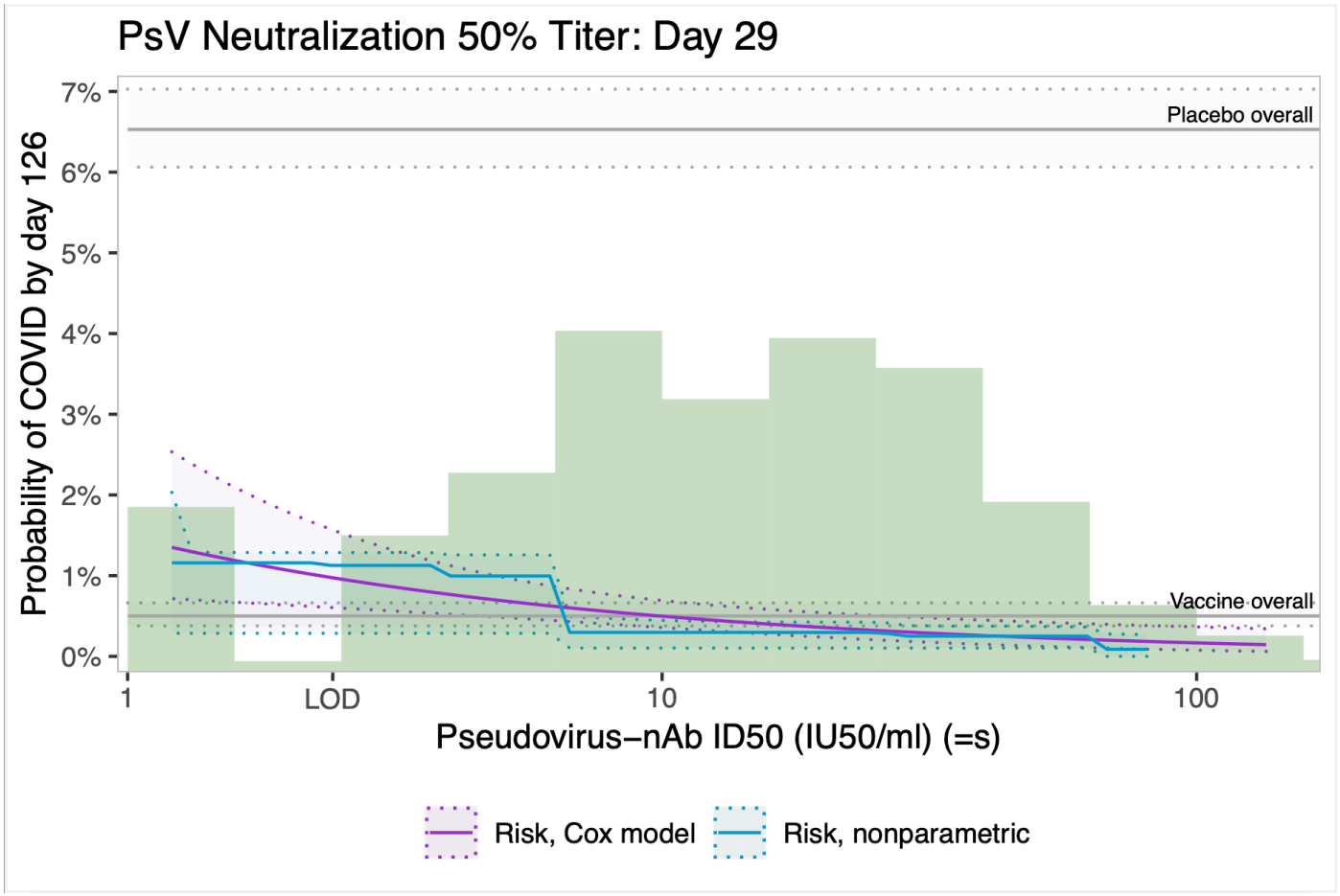
$$
\begin{aligned}
Y &= \text{COVID disease by day 126} \\
A &= \text{indicator of vaccination status} \\
M &= \text{day 29 pseudovirus ID50 neutralizing titer}
\end{aligned}
$$

Gilbert et al. (2021) introduced **controlled risk** and **controlled vaccine efficacy** curves:

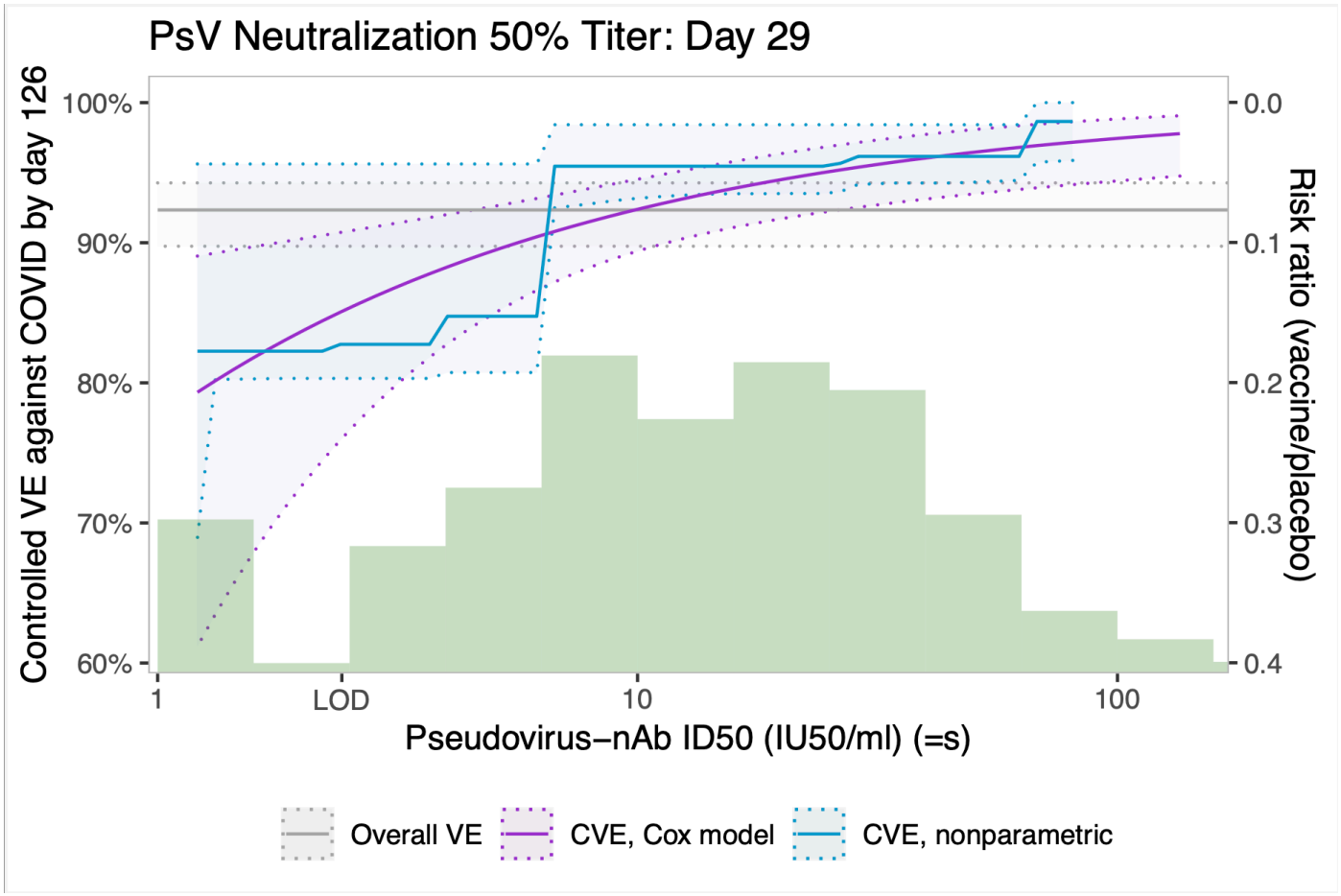
$$
\begin{aligned}
\text{CR}(m) &:= E[Y(1,m)] \\
&= \text{risk of COVID by day 126 in participants given vaccine and set to have antibody level } m \\
\text{CVE}(m) &:= 1 - \frac{E[Y(1,m)]}{E[Y(0)]} \\
&= \text{relative reduction in risk of COVID by day 126 comparing participants given vaccine and set to have antibody level } m \text{ vs participants given placebo.}
\end{aligned}
$$

In a randomized vaccine trial, the denominator $E[Y(0)]$ is easy to estimate, but the numerator $E[Y(1,m)]$ remains challenging...

What about continuous mediators?



What about continuous mediators?



References and additional reading

References:

Horvitz, D, Thompson, D (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260)663–685. doi: 10.1080/01621459.1952.10483446.

Robins, JM (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12)1393-1512. doi: 10.1016/0270-0255(86)90088-6.

Robins, JM, Hernan, MA, Brumback, B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5)550-560. doi: 10.1097/00001648-200009000-00011.

Westling, T, Gilbert, PB, Carone, M (2020). Causal isotonic regression. *Journal of the Royal Statistical Society: Series B*, 82(3):719-747. doi: 10.1111/rssb.12372.

Gilbert, PB, Fong, Y, Carone, M (2021). A controlled effects approach to assessing immune correlates of protection. *Biostatistics*, kxac024. doi: 10.1093/biostatistics/kxac24.

Additional reading:

Snowden, JM, Rose, S, Mortimer, KM (2011). Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7)731-738. doi: 10.1093/aje/kwq472.

VanderWeele, TJ (2015). Explanation in causal inference. *Oxford*.