Causal estimation and emulated trials

David Hajage

Sorbonne Université

9 June 2026

SANTÉ SORBONNE UNIVERSITÉ

Conflicts of interest

Private sector links: None

Public sector links: None

Causality in epidemiology

Pervasive questions in public health:

- Do fine particles **cause** cardiovascular disease?
- Does pesticide exposure **cause** lymphomas?
- Does an antihypertensive treatment **reduce** long-term mortality?

Public health decisions rely on causal reasoning.

But association ≠ causation

- Confounding bias
- Selection bias
- Measurement bias

Course outline

1. **Counterfactual reasoning and G-computation**
   - Potential outcomes, fundamental assumptions
   - Standardisation (G-computation)
2. **Propensity score and IPTW**
   - Propensity score, inverse probability of treatment weighting
3. **IPCW**
   - Informative censoring, per-protocol analysis
4. **Target trial emulation**

SANTÉ SORBONNE UNIVERSITÉ

1

Counterfactual reasoning

Everyday causality

When I drop my pen, it falls.

- **Necessary** association: the pen always falls when I drop it.
- **Specific** association: the pen only falls when I drop it.
- **Immediate** association: the pen falls immediately when I drop it.

⇒ A simple causal relationship to establish.

Causality in epidemiology

Now imagine:

- If I drop my pen, sometimes it falls, sometimes it doesn't. When it does fall, it only does so 5 years later.
- If I don't drop my pen, it falls anyway occasionally (but less often).

⇒ Dropping the pen changes the **probability** that the pen falls in 5 years.

Most causal relationships in epidemiology are neither necessary, nor specific, nor immediate.

Presentation of the study individuals

Name	A	Y
Alice	1	1
Brahim	1	0
Cécile	0	0
David	0	1
Emma	0	1
François	0	0
Giulia	1	0
Hugo	0	0
Inès	0	1
Julien	0	0
Karine	0	1
Léo	0	0
Marie	1	0
Nour	1	1

A: drug taken ($A = 1$) or not ($A = 0$)

Y: recovery ($Y = 1$) or not ($Y = 0$)

Adapted from an example by Hernán & Robins, "What If?"

Alice

Name	A	Y
Alice	1	1
Brahim	1	0
Cécile	0	0
David	0	1
Emma	0	1
François	0	0
Giulia	1	0
Hugo	0	0
Inès	0	1
Julien	0	0
Karine	0	1
Léo	0	0
Marie	1	0
Nour	1	1

- Alice took the drug and recovered.
- Did the drug **cause** the recovery?

Alice

Name	A	Y
Alice	1	1
Brahim	1	0
Cécile	0	0
David	0	1
Emma	0	1
François	0	0
Giulia	1	0
Hugo	0	0
Inès	0	1
Julien	0	0
Karine	0	1
Léo	0	0
Marie	1	0
Nour	1	1

- Alice took the drug and recovered.
- Did the drug **cause** the recovery?

We cannot know, because we would need to know what would have happened to Alice **without** the drug.

Potential outcomes and counterfactuals

- A: observed treatment (exposure); Y: observed outcome
- $Y^{A=a}$: **potential** outcome under the alternative $A = a$:
  - Y^0: outcome **if** the drug is not taken
  - Y^1: outcome **if** the drug is taken
- Only one of the two values is observed:
  - if $A = 0 \rightarrow$ we observe Y^0
  - if $A = 1 \rightarrow$ we observe Y^1
- The unobserved outcome is called the **counterfactual** outcome

Alice (continued)

Name	Y0	Y1
Alice	0	1
Brahim	0	0
Cécile	0	0
David	1	1
Emma	1	0
François	0	0
Giulia	1	0
Hugo	0	1
Inès	1	1
Julien	0	1
Karine	1	0
Léo	0	0
Marie	0	0
Nour	1	1

- Alice took the drug and recovered: $Y = Y^1 = 1$
- Had Alice **not** taken the drug, he/she would **not** have recovered: $Y^0 = 0$
- For Alice, the drug has a **causal effect**

What about the others?

Individual causal effect

Name	Y0	Y1	Causal effect
Alice	0	1	Yes, benefic.
Brahim	0	0	No
Cécile	0	0	No
David	1	1	No
Emma	1	0	Yes, detrim.
François	0	0	No
Giulia	1	0	Yes, detrim.
Hugo	0	1	Yes, benefic.
Inès	1	1	No
Julien	0	1	Yes, benefic.
Karine	1	0	Yes, detrim.
Léo	0	0	No
Marie	0	0	No
Nour	1	1	No

The individual causal effect is defined by $Y^1 - Y^0$.

It is not necessarily the same for all individuals.

In reality...

Name	Y0	Y1	A	Y
Alice	0	1	1	1
Brahim	0	0	1	0
Cécile	0	0	0	0
David	1	1	0	1
Emma	1	0	0	1
François	0	0	0	0
Giulia	1	0	1	0
Hugo	0	1	0	0
Inès	1	1	0	1
Julien	0	1	0	0
Karine	1	0	0	1
Léo	0	0	0	0
Marie	0	0	1	0
Nour	1	1	1	1

We never observe Y^0 **and** Y^1 simultaneously — only A and Y.

$\Rightarrow$ The individual causal effect is **unknown**

Let us be less ambitious.

Population-level causal effect

Name	Y0	Y1	Causal effect
Alice	0	1	Yes, benefic.
Brahim	0	0	No
Cécile	0	0	No
David	1	1	No
Emma	1	0	Yes, detrim.
François	0	0	No
Giulia	1	0	Yes, detrim.
Hugo	0	1	Yes, benefic.
Inès	1	1	No
Julien	0	1	Yes, benefic.
Karine	1	0	Yes, detrim.
Léo	0	0	No
Marie	0	0	No
Nour	1	1	No

A **population-level** causal effect (*Average Treatment Effect*):

$$ATE = E(Y^1) - E(Y^0)$$

For binary Y (risk difference):

$$\Pr(Y^1 = 1) = 6/14$$

$$\Pr(Y^0 = 1) = 6/14$$

$$\text{ATE} = 6/14 - 6/14 = 0$$

→ no causal effect

The fundamental problem of causal inference

- As with the individual effect, we cannot directly compute $Pr(Y^1 = 1) - Pr(Y^0 = 1)$
- This is the **fundamental problem of causal inference**

Causal inference

Doing causal inference means seeking to estimate, from observed data, quantities that are **reasonable substitutes** for the unobservable quantities $Pr(Y^1 = 1)$, $Pr(Y^0 = 1)$.

Crude estimation

Name	Y0	Y1	A	Y
Alice	0	1	1	1
Brahim	0	0	1	0
Cécile	0	0	0	0
David	1	1	0	1
Emma	1	0	0	1
François	0	0	0	0
Giulia	1	0	1	0
Hugo	0	1	0	0
Inès	1	1	0	1
Julien	0	1	0	0
Karine	1	0	0	1
Léo	0	0	0	0
Marie	0	0	1	0
Nour	1	1	1	1

The most natural substitutes:

$\hat{Pr}(Y^1 = 1) \approx Pr(Y = 1 \mid A = 1) \quad = 2/5$

$\hat{Pr}(Y^0 = 1) \approx Pr(Y = 1 \mid A = 0) \quad = 4/9$

$\hat{ATE} \approx -2/45$

But $ATE = 0$ in reality!

What could explain this difference?

Crude estimation

Name	Y0	Y1	A	Y
Alice	0	1	1	1
Brahim	0	0	1	0
Cécile	0	0	0	0
David	1	1	0	1
Emma	1	0	0	1
François	0	0	0	0
Giulia	1	0	1	0
Hugo	0	1	0	0
Inès	1	1	0	1
Julien	0	1	0	0
Karine	1	0	0	1
Léo	0	0	0	0
Marie	0	0	1	0
Nour	1	1	1	1

The exposed and unexposed may **not have the same prognosis before treatment**

→ the observed association mixes the (possible) treatment effect with pre-existing differences between groups.

Lack of exchangeability: the groups are not interchangeable at baseline.

Confounding

Name	A	Y	L
Alice	1	1	1
Brahim	1	0	1
Cécile	0	0	0
David	0	1	0
Emma	0	1	1
François	0	0	1
Giulia	1	0	0
Hugo	0	0	1
Inès	0	1	0
Julien	0	0	0
Karine	0	1	0
Léo	0	0	0
Marie	1	0	1
Nour	1	1	0

L: initial severity ($L = 1$ if severe)

- Individuals with $L = 1$ take the drug **more often**: $Pr(A = 1 \mid L = 1) = 3/6$
- Individuals with $L = 1$ recover **less often** (even without the drug)

Confounding

Name	A	Y	L
Alice	1	1	1
Brahim	1	0	1
Cécile	0	0	0
David	0	1	0
Emma	0	1	1
François	0	0	1
Giulia	1	0	0
Hugo	0	0	1
Inès	0	1	0
Julien	0	0	0
Karine	0	1	0
Léo	0	0	0
Marie	1	0	1
Nour	1	1	0

L: initial severity ($L = 1$ if severe)

- Individuals with $L = 1$ take the drug **more often**: $Pr(A = 1 \mid L = 1) = 3/6$
- Individuals with $L = 1$ recover **less often** (even without the drug)

$\Rightarrow$ L is a **confounder**

Conditional exchangeability

Name	A	Y	L
Alice	1	1	1
Brahim	1	0	1
Cécile	0	0	0
David	0	1	0
Emma	0	1	1
François	0	0	1
Giulia	1	0	0
Hugo	0	0	1
Inès	0	1	0
Julien	0	0	0
Karine	0	1	0
Léo	0	0	0
Marie	1	0	1
Nour	1	1	0

Among $L = 1$:

$Pr(Y = 1 \mid A = 1, L = 1) = 1/3$

$Pr(Y = 1 \mid A = 0, L = 1) = 1/3$

$\rightarrow \hat{ATE} = 0$

Among $L = 0$:

$Pr(Y = 1 \mid A = 1, L = 0) = 1/2$

$Pr(Y = 1 \mid A = 0, L = 0) = 3/6$

$\rightarrow \hat{ATE} = 0$

No association within each stratum!

Conditional exchangeability

Name	A	Y	L
Alice	1	1	1
Brahim	1	0	1
Cécile	0	0	0
David	0	1	0
Emma	0	1	1
François	0	0	1
Giulia	1	0	0
Hugo	0	0	1
Inès	0	1	0
Julien	0	0	0
Karine	0	1	0
Léo	0	0	0
Marie	1	0	1
Nour	1	1	0

Within each stratum of L, the exposed and unexposed have the **same counterfactual prognosis**: they are exchangeable.

Formally: $Y^a \perp A \mid L, \quad \forall a$

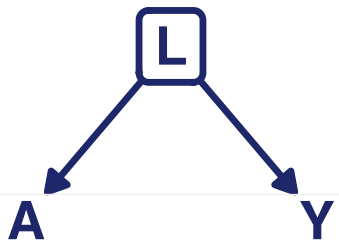
Conditional exchangeability

Within each stratum, exposed and unexposed individuals have the same baseline risk profile: they are "comparable". The only thing distinguishing them is the exposure → we can attribute the difference in outcomes to it.

SUMMARY: COUNTERFACTUAL REASONING

What we have learned:

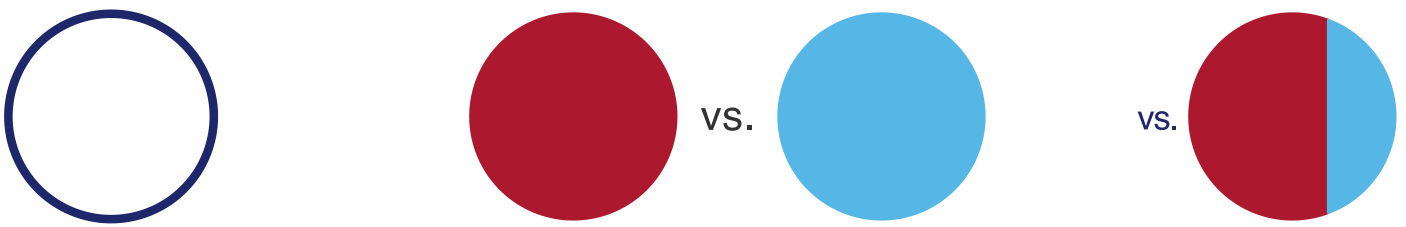
- The individual causal effect ($Y^1 - Y^0$) is unknown (fundamental problem)
- Outside a randomised trial, a crude association has no causal interpretation, due to lack of exchangeability
- The population-level causal effect ($ATE = E(Y^1) - E(Y^0)$) is estimable, but only under **certain assumptions**
- One of them is **conditional exchangeability**: stratifying the analysis on L allowed us to correct for confounding



The counterfactual intervention

Average Treatment Effect?

The ATE is the difference between two **hypothetical worlds**: *if everyone were treated* vs *if no one were treated*.



SANTÉ SORBONNE UNIVERSITÉ

2

The fundamental assumptions

The 4 fundamental assumptions

To give a causal interpretation to an association in an observational study, 4 assumptions are required:

1. **No interference**
2. **Consistency**
3. **Conditional exchangeability**
4. **Positivity**

No interference

Definition: The exposure of one individual does not affect the potential outcome of another individual.

Examples of violation:

- *Vaccination*: vaccinating one individual reduces the probability of infection in their contacts (herd immunity).
- *Air pollution*: reducing emissions in one area also affects air quality in neighbouring areas.

Consistency

Definition: The potential outcome of an individual, if they received the exposure actually observed, equals the observed outcome:

$$Y^{A=a} = Y \text{ for individuals where } A = a$$

Examples of violation: When the exposure has multiple "versions"

- *Obesity*: $Y^{BMI=30} \neq Y^{BMI=40}$, but both correspond to obesity ($A = 1$)
- *Pesticides*: composition, season, application method = as many versions of the same "exposure"

Defining the exposure precisely enough reduces the risk of violation.

Conditional exchangeability

Exchangeability (in randomised trials):

$$Y^a \perp A \quad \forall a$$

Conditional exchangeability (in observational studies):

$$Y^a \perp A \mid L \quad \forall a$$

Within each stratum defined by L, exposed and unexposed individuals have the same average potential outcome.

This assumption is untestable from the data: it relies on domain knowledge.

In practice: L represents the set of measured confounders. Causal diagrams (DAGs) can help identify them.

Positivity

To compare exposed and unexposed individuals within each stratum of L, there must actually be both exposed **and** unexposed individuals in each stratum...

Definition: All individuals have a **non-zero** probability of being exposed and unexposed, regardless of their value of L:

$$Pr(A = a \mid L = l) > 0 \quad \forall a, \forall l$$

Examples of violation:

- No individual under 18 receives the treatment → the effect cannot be estimated in those under 18

POSITIVITY (CONT.)

- **Structural violation**: some strata *cannot* receive the exposure “by definition” (legal constraint, contraindication, ...) → the probability is truly 0 → it is fairly intuitive to exclude them!
- **Random violation**: the stratum is absent from the sample due to insufficient data, but exposure remains theoretically possible in the population → frequent situation when sample size is small

Towards G-computation

Stratification is a simple and pedagogically clear method for handling confounding, but it quickly becomes **impractical**:

- Many confounders $L_1, L_2, \ldots$
- Continuous or multi-category variables
- What if we want an **average effect** over the whole population, not a separate effect in each stratum?

How to generalise?

We will now see a method that directly extends this reasoning: **G-computation** (or regression standardisation).

SANTÉ
SORBONNE
UNIVERSITÉ

3

G-COMPUTATION

Identifiability of the ATE

In a randomised trial: exchangeability → the two groups represent the same population

$$ATE = E(Y \mid A = 1) - E(Y \mid A = 0)$$

Directly estimable by comparing the two groups.

In an observational study: conditional exchangeability on L → groups are comparable *within each stratum*

$$ATE_l = E(Y \mid A = 1, L = l) - E(Y \mid A = 0, L = l)$$

Causal effect estimable within each stratum... How to obtain a population average effect? (marginal population effect)

From stratification to regression

In a randomised trial: direct comparison $\leftrightarrow$ simple regression model

$$E(Y \mid A = a) = \alpha + \boldsymbol{\beta}\, a \quad \Rightarrow \quad \boldsymbol{\beta} = ATE$$

In an observational study: stratification on L $\leftrightarrow$ model adjusted for L with $A \times L$ interaction

$$E(Y \mid A, L) = \alpha + \gamma L + \boldsymbol{\beta}\, A + \boldsymbol{\delta}\, A \times L$$

- $\boldsymbol{\beta} + \boldsymbol{\delta}\, l = ATE_l$: effect **conditional** on $L = l$
- Generalisable: $E(Y \mid A, L_1, L_2, \ldots) = \alpha + \gamma_1 L_1 + \gamma_2 L_2 + \ldots + \boldsymbol{\beta}\, A + \boldsymbol{\delta_1}\, AL_1 + \boldsymbol{\delta_2}\, AL_2 + \ldots$

From stratification to regression (cont.)

- If no effect modification ($\delta = 0$): $ATE_l = ATE = \boldsymbol{\beta}$

In the absence of treatment effect heterogeneity, all stratum-specific conditional effects are equal to each other, and equal to the marginal population effect

- But what to do in the presence of such heterogeneity (or when we do not want to assume its absence)?

From regression to standardisation

The marginal effect is the **weighted average** of the conditional effects:

$$ATE = \sum_l ATE_l \times Pr(L = l)$$

The weights reflect the **distribution of L in the population**.

Estimation (standardisation):

$$\hat{ATE} = \sum_l (\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\delta}}_l) \times \hat{Pr}(L = l)$$

From regression to standardisation (cont.)

In practice, this can be computed individually for each observation:

$$\hat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{E}(Y \mid A = 1, L = l_i) - \hat{E}(Y \mid A = 0, L = l_i)\right]$$

→ This is the **G-computation** algorithm.

The G-computation algorithm (ATE)

In a sample of size n:

Step 1: Fit two regression models for Y, separately by exposure:

$$E(Y \mid A = 1) \sim f(L) \qquad E(Y \mid A = 0) \sim f(L)$$

(equivalent to a single model with $A \times L_k$ interactions)

Step 2: For **each individual** i, predict the two potential outcomes:

$$\hat{Y}_{1,i} = \hat{E}(Y \mid A = 1, L = l_i) \qquad \hat{Y}_{0,i} = \hat{E}(Y \mid A = 0, L = l_i)$$

The G-computation algorithm (ATE) (cont.)

Step 3: Average the predictions across all individuals:

$$\hat{E}(Y^1) = \frac{1}{n}\sum_i \hat{Y}_{1,i} \qquad \hat{E}(Y^0) = \frac{1}{n}\sum_i \hat{Y}_{0,i}$$

Step 4: Compute the desired measure of association:

$$\hat{DR} = \hat{E}(Y^1) - \hat{E}(Y^0) \qquad \hat{RR} = \frac{\hat{E}(Y^1)}{\hat{E}(Y^0)} \qquad \hat{OR} = \frac{\hat{E}(Y^1)/[1 - \hat{E}(Y^1)]}{\hat{E}(Y^0)/[1 - \hat{E}(Y^0)]}$$

Illustration on our example

Name	A	L	$\hat{Y}1$	$\hat{Y}0$
Alice	1	1	0.33	0.33
Brahim	1	1	0.33	0.33
Cécile	0	0	0.50	0.50
David	0	0	0.50	0.50
Emma	0	1	0.33	0.33
François	0	1	0.33	0.33
Giulia	1	0	0.50	0.50
Hugo	0	1	0.33	0.33
Inès	0	0	0.50	0.50
Julien	0	0	0.50	0.50
Karine	0	0	0.50	0.50
Léo	0	0	0.50	0.50
Marie	1	1	0.33	0.33
Nour	1	0	0.50	0.50

Steps 1 & 2: separate logistic models → counterfactual predictions (see table on the left)

Steps 3:

$\hat{E}(Y^1) = 0.427$

$\hat{E}(Y^0) = 0.427$

Step 4:

$\hat{ATE} = 0$

(True ATE = 0 by construction)

ATE vs ATT

Average Treatment Effect (ATE): effect if the **entire** population were exposed vs unexposed

$$ATE = E(Y^1) - E(Y^0)$$

Average Treatment Effect on the Treated (ATT): effect if the **exposed individuals** had not been exposed

$$ATT = E(Y^1 \mid A = 1) - E(Y^0 \mid A = 1)$$

ATE vs ATT (CONT.)

G-computation and ATT?

ATT algorithm: steps 1 and 2 are identical; at step 3, average **only over exposed individuals** ($A = 1$).

$$\hat{E}(Y^0 \mid A = 1) = \frac{1}{\sum A} \sum_{i:A_i=1} \hat{Y}_{0,i}$$

In a randomised controlled trial: $ATE = ATT = ATC$.

Variance of the estimate

The algorithm readily provides the ATE estimate; obtaining its variance is more involved.

Analytical method (delta method)

- Analytical propagation of uncertainty from the step-1 models
- Implemented in the R package `marginaleffects`

Bootstrap

- Repeat the 4 steps on B resamples
- Variance estimated from the spread of the B estimates
- Simple to implement, valid under general assumptions

Both approaches yield valid confidence intervals. Bootstrap is more general but more computationally intensive.

SUMMARY: G-COMPUTATION

Strengths:

- Intuitive: a direct extension of counterfactual reasoning
- Flexible: any regression model (linear, logistic, survival...)

Limitations:

- Relies on **correct specification** of the outcome model $E(Y \mid A, L)$
- Can be difficult when L is high-dimensional and Y is a rare event

Coming next:

An alternative method that models $E(A \mid L)$ rather than $E(Y \mid A, L)$: the **propensity score**.

SANTÉ SORBONNE UNIVERSITÉ

4

Propensity score and IPTW

A COMPLEMENTARY APPROACH

G-computation models $E(Y \mid A, L)$

- Requires correct specification of the outcome model
- Difficult when L is high-dimensional

Propensity score models $P(A = 1 \mid L)$

- Does **not** model Y at all
- Directly targets balance between groups
- Creates a pseudo-population where $A \perp L$

Both methods rest on the same fundamental assumptions and target the same marginal estimands.

The propensity score

Definition:

$$e = P(A = 1 \mid L = l)$$

Key property (Rosenbaum & Rubin, 1983):

If $Y^a \perp A \mid L \; \forall a$ (conditional exchangeability on L), then:

$$Y^a \perp A \mid e \; \forall a$$

Key point

e is a **balancing score**: conditioning on e is sufficient to restore exchangeability between treatment groups.

Estimating the propensity score

Logistic regression (most common):

$$P(A = 1 \mid L) = \text{logit}^{-1}(\beta_0 + \beta_1 L_1 + \ldots + \beta_k L_k)$$

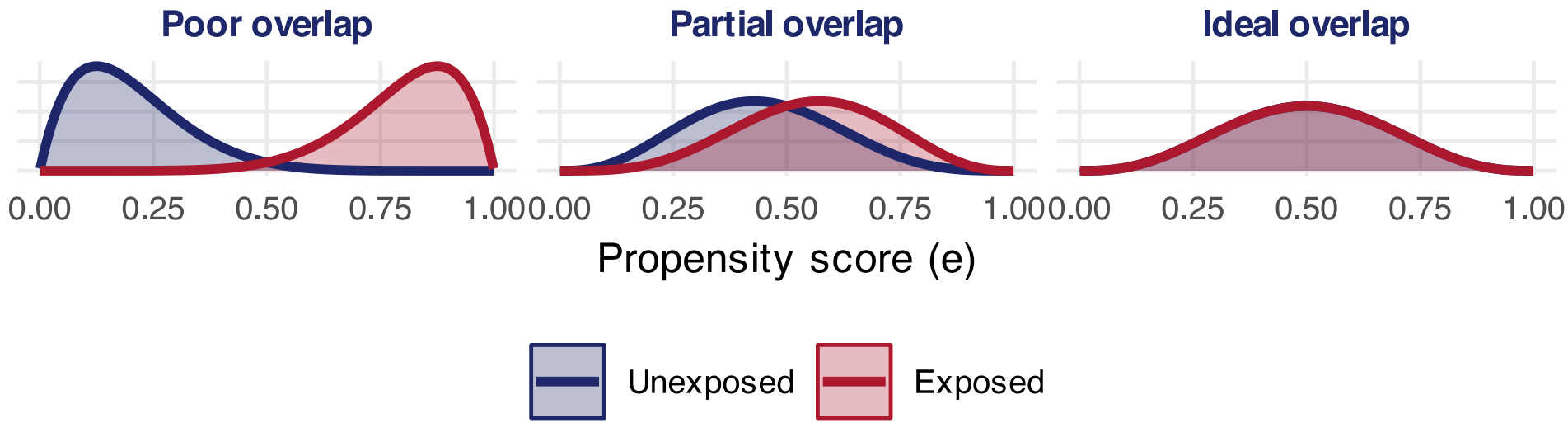
Which variables to include in the model?

- **Confounders** (related to A **and** Y) → **Yes**
- **Prognostic factors** (related to Y only) → **Yes** (improve precision)
- **Instrumental variables** (related to A only) → **No** (reduce precision, may increase bias)

These are therefore the same variables to account for as in G-computation!

Positivity and overlap of distributions

Plotting the distribution of the propensity score in exposed and unexposed individuals allows exploration of the positivity assumption.



Positivity ↔ overlap of the distributions

Methods for using the propensity score

Method	Principle
Adjustment	Regression $Y \sim A + e$
Stratification	Divide into quintiles of e; analyse within each stratum
Matching	Match each exposed individual to an unexposed individual with a similar score e
Weighting (Inverse Probability of Treatment Weighting)	Weight each individual i by $w_i = f(e_i)$

Methods for using the propensity score (cont.)

Why IPTW in this course?

- Directly provides a **marginal effect** (ATE), without additional standardisation
- Naturally generalises to **time-varying confounders** and longitudinal data
- Directly extends to **IPCW** to correct for informative censoring (Part 3)
- Naturally fits within the framework of **target trial emulation** (Part 4)

IPTW — WEIGHTS AND PSEUDO-POPULATION

Weights for the ATE:

$$w_i = \frac{1}{P(A_i \mid L_i)} = \begin{cases} 1/e_i & \text{if } A_i = 1 \\ 1/(1-e_i) & \text{if } A_i = 0 \end{cases}$$

Intuition:

- a treated individual with a small e is “rare” among the treated → high weight to compensate for this rarity
- an untreated individual with a small e is “common” among the untreated → small weight to compensate for this frequency

This process re-balances the L characteristics of the two groups

Stabilised IPTW weights

Stabilised weights:

$$w_i^s = \frac{P(A_i)}{P(A_i \mid L_i)} = \begin{cases} P(A=1)/e_i & \text{if } A_i = 1 \\ P(A=0)/(1-e_i) & \text{if } A_i = 0 \end{cases}$$

The numerator $P(A_i)$ is the **marginal** probability of exposure: simply the prevalence of exposure $P(A = 1)$ among the exposed, and that of non-exposure $P(A = 0)$ among the unexposed.

- Mean close to **1** → easy to verify
- Reduced variance, extreme values less frequent
- Target the same **ATE** as unstabilised weights

In practice, stabilised weights are preferred.

Illustration on our example

Name	L	A	e	w
Alice	1	1	0.50	2.00
Brahim	1	1	0.50	2.00
Cécile	0	0	0.25	1.33
David	0	0	0.25	1.33
Emma	1	0	0.50	2.00
François	1	0	0.50	2.00
Giulia	0	1	0.25	4.00
Hugo	1	0	0.50	2.00
Inès	0	0	0.25	1.33
Julien	0	0	0.25	1.33
Karine	0	0	0.25	1.33
Léo	0	0	0.25	1.33
Marie	1	1	0.50	2.00
Nour	0	1	0.25	4.00

Estimated propensity score:

$\hat{e}(L = 0) = P(A = 1 \mid L = 0) = 0.25$

$\hat{e}(L = 1) = P(A = 1 \mid L = 1) = 0.5$

Balance on L before/after weighting:

	$A = 1$	$A = 0$
Crude $P(L = 1)$	0.6	0.33
Weighted $P_w(L = 1)$	0.43	0.43

After weighting, both groups have the same prevalence of L.

A ⊥ L after weighting → exchangeability restored!

Checking balance: the love plot

Standardised mean difference (SMD) for each L_k, before and after weighting:

Without weighting

$$SMD_k = \frac{\bar{L}_{k,A=1} - \bar{L}_{k,A=0}}{\sqrt{(\hat{\sigma}^2_{k,A=1} + \hat{\sigma}^2_{k,A=0})/2}}$$

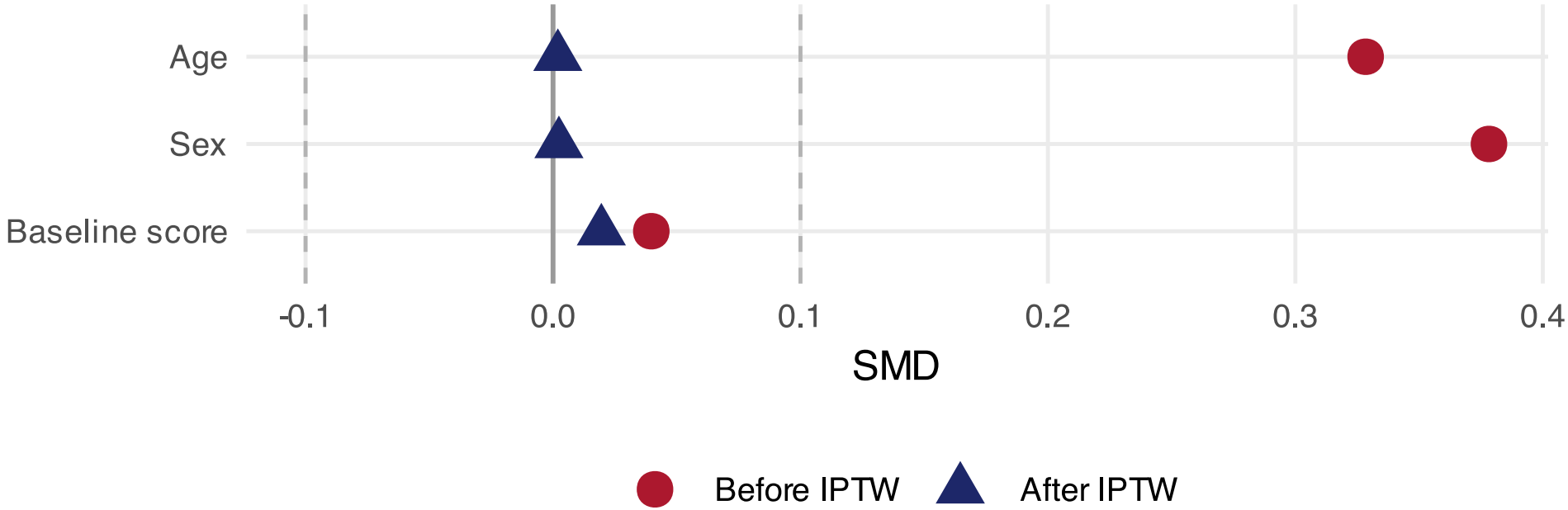
With weighting

$$SMD^w_k = \frac{\bar{L}^w_{k,A=1} - \bar{L}^w_{k,A=0}}{\sqrt{(\hat{\sigma}^2_{k,A=1} + \hat{\sigma}^2_{k,A=0})/2}}$$

where $\bar{L}^w_{k,A=a} = \frac{\sum_{i:A_i=a} w_i L_{ki}}{\sum_{i:A_i=a} w_i}$

The denominator uses **unweighted** variances in both cases. We aim for $|SMD^w| < 0.1$.

Checking balance: the love plot (cont.)



No statistical tests (p-values) needed to assess balance!

Analysis and variance

In the weighted pseudo-population, we estimate weighted expectations:

$$\hat{E}(Y^1) = \frac{\sum_{i:A_i=1} w_i Y_i}{\sum_{i:A_i=1} w_i} \qquad \hat{E}(Y^0) = \frac{\sum_{i:A_i=0} w_i Y_i}{\sum_{i:A_i=0} w_i}$$

then compute the desired measure of association (*DR*, *RR*, *OR*...), exactly as in G-computation.

No additional adjustment needed

Once balance is restored, **no need to adjust for** *L*: covariates are already balanced by the weights.

Analysis and variance (cont.)

Variance: the “ordinary” variance may underestimate uncertainty (weights are estimated).

- Analytical estimators (rather complex)
- **Bootstrap**: re-estimating the propensity score model in each resample

Estimand and type of weighting

The weights presented ($w_i = A_i/e_i + (1 - A_i)/(1 - e_i)$) target the **ATE**.

To obtain the ATT, only the weights change:

$$w_i^{ATT} = \begin{cases} 1 & \text{if } A_i = 1 \\ e_i/(1 - e_i) & \text{if } A_i = 0 \end{cases}$$

The rest of the analysis is identical: compute weighted expectations then the chosen measure of association.

Other weightings exist depending on the target population (ATC, ATO, ...) but the principle is always the same; only the weights differ.

Summary: G-computation vs IPTW

	G-computation	**IPTW**
What is modelled	$E(Y \mid A, L)$	$P(A = 1 \mid L)$
Key assumption	Correct specification of $E(Y \mid A, L)$	Correct specification of $P(A = 1 \mid L)$
Advantage	More robust to positivity violations	Positivity and balance verifiable

SANTÉ SORBONNE UNIVERSITÉ

5

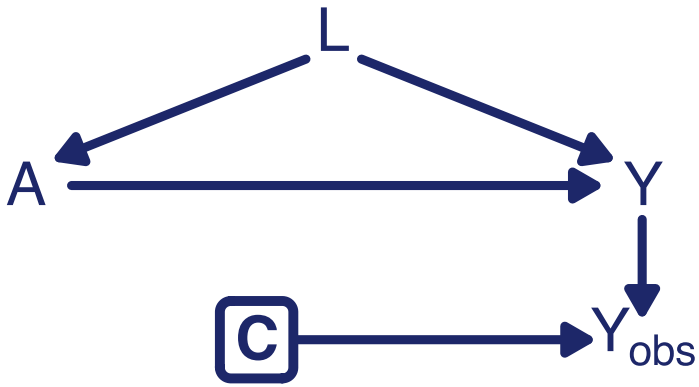
IPCW

Censoring in longitudinal studies

In a longitudinal study, **censoring** occurs when the outcome is not observed:

- Lost to follow-up, end of study, competing events

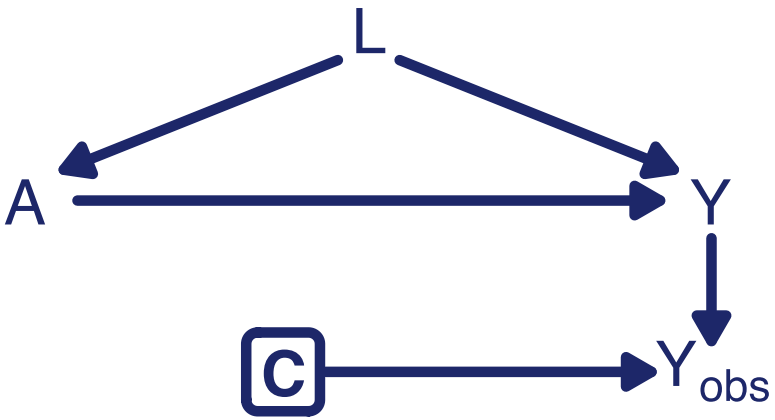
Counterfactual framework: we would like to know Y for **all** individuals, but censoring C **masks** it → we only observe Y_{obs}.



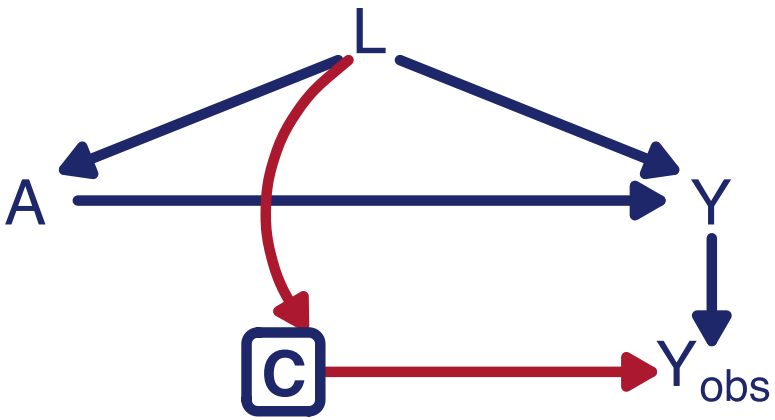
C masks Y: we only observe Y_{obs}.

Non-informative vs informative censoring

Non-informative: C independent of L → no bias



Informative: L predicts both Y and C → selection bias



Consequence and solution: IPCW

L predicts both Y **and** C: uncensored individuals ($C = 0$) **are not representative** of the entire population.

Consequence

The analysis restricted to $C = 0$ is biased.

Solution

Weight the uncensored individuals so that they also represent the censored individuals who resemble them → **IPCW** (Inverse Probability of Censoring Weighting).

Counterfactual intervention on censoring

Censoring can be treated as an **intervention**: the same reasoning as for exposure!

Key Point

What would happen if **the entire population remained uncensored**?

$$E(Y^{C=0})$$

Key assumption: conditionally on covariates L, censoring is independent of the potential outcome

$$Y^{C=0} \perp C \mid L$$

(*missing at random* (MAR) assumption, applied to censoring)

The IPCW principle — Basic case

Weight each **uncensored** individual to also represent the censored individuals who resemble them.

Basic case: censoring depends only on baseline covariates L.

Unstabilised weight:

$$wc_i = \frac{1}{Pr(C_i = 0 \mid L = l_i)}$$

Stabilised weight (recommended):

$$wc_i^s = \frac{Pr(C_i = 0)}{Pr(C_i = 0 \mid L = l_i)}$$

An uncensored individual with a **high risk of censoring** → high weight to compensate for similar censored individuals.

In practice: estimate $\hat{Pr}(C = 1 \mid L)$ by logistic regression, then $wc_i = 1/(1 - \hat{p}_{C,i})$.

In the weighted pseudo-population, uncensored individuals are representative of the entire population.

IPCW WEIGHTS — TIME-VARYING CONFOUNDERS

When censoring also depends on covariates **measured during follow-up** $\bar{L}_k$, the probability of remaining uncensored is computed **at each period** k:

$$wc_i^s = \frac{\prod_{k=0}^{K} Pr(C_k = 0 \mid A_0, C_{k-1} = 0)}{\prod_{k=0}^{K} Pr(C_k = 0 \mid \bar{A}_{k-1}, \bar{L}_k, C_{k-1} = 0)}$$

- **Numerator**: marginal probabilities (baseline characteristics) → serves to stabilise the weights
- **Denominator**: conditional probabilities on history $\bar{L}_k$ → corrects bias at each step

The individual weight wc_i is the **product** of $K + 1$ terms, one per follow-up period.

IPCW WEIGHTS — LONGITUDINAL OUTCOME

When the outcome is longitudinal (e.g., a censored outcome such as survival), the weight is **updated at each time** t:

$$wc_{it}^{s} = \frac{\prod_{k=0}^{t} Pr(C_k = 0 \mid A_0, C_{k-1} = 0)}{\prod_{k=0}^{t} Pr(C_k = 0 \mid \bar{A}_{k-1}, \bar{L}_k, C_{k-1} = 0)}$$

Same structure as the previous case: only the horizon changes (t instead of K).

- At each period, wc_{it} is recomputed for all still-observed individuals
- Individuals with a high risk of censoring accumulate increasing weights over follow-up

This is the usual case in practice: survival analyses, target trial emulation.

Combining IPTW + IPCW

In the presence of both **baseline confounding** and **informative censoring**: the weights can be combined

$$\tilde{w}_i = w_{IPTW,i} \times wc_{IPCW,i}$$

SANTÉ SORBONNE UNIVERSITÉ

6

Application: per-protocol analysis

Scenario

Observational cohort study of a treatment **taken long-term**:

- At baseline ($t = 0$): some patients initiate treatment ($A = 1$), others do not ($A = 0$)
- L: confounders measured at baseline and during follow-up
- Follow-up until the event or end of study

ITT-analogue

What is the effect of **initiating** treatment?

→ **IPTW**

PP-analogue

What is the effect of **initiating AND maintaining** treatment?

→ **IPTW × IPCW**

ITT-ANALOGUE: ESTIMATING THE EFFECT OF INITIATION

We compare the two groups **as defined at baseline**, regardless of their subsequent trajectory.

- $A = 1$ (initiators) vs $A = 0$ (non-initiators), from start to end of follow-up
- **Baseline confounding** corrected by IPTW

What we estimate

The effect of the **decision to initiate** treatment at the start of follow-up: analogue of the intention-to-treat (ITT) analysis.

PP-ANALOGUE: ESTIMATING THE EFFECT OF ADHERENCE TO TREATMENT

Some treatments must be **maintained** to be effective: antiplatelets, antihypertensives, statins...

But in practice:

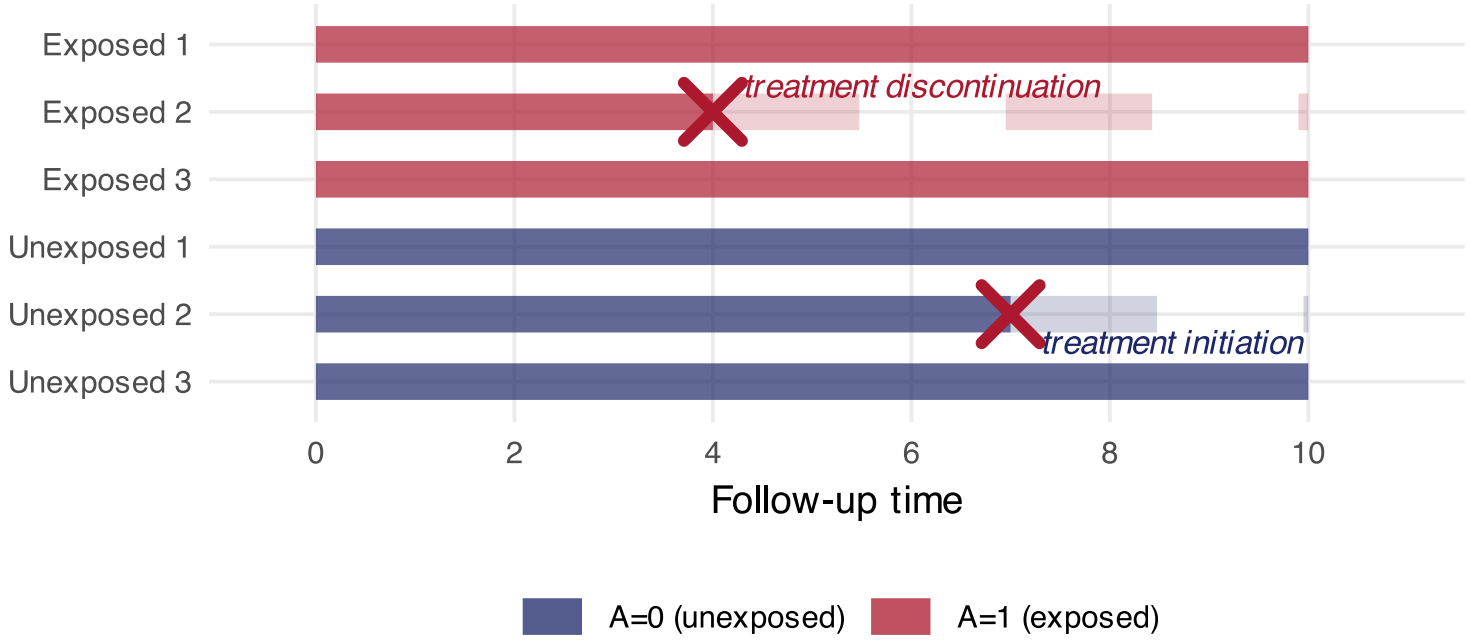
- Some $A = 1$ patients **stop** their treatment during follow-up
- Some $A = 0$ patients **start** treatment later

Per-protocol effect

What would the effect be if each individual had **followed their initial strategy** throughout follow-up?

→ **Censor** deviations + correct this informative censoring with **IPCW**.

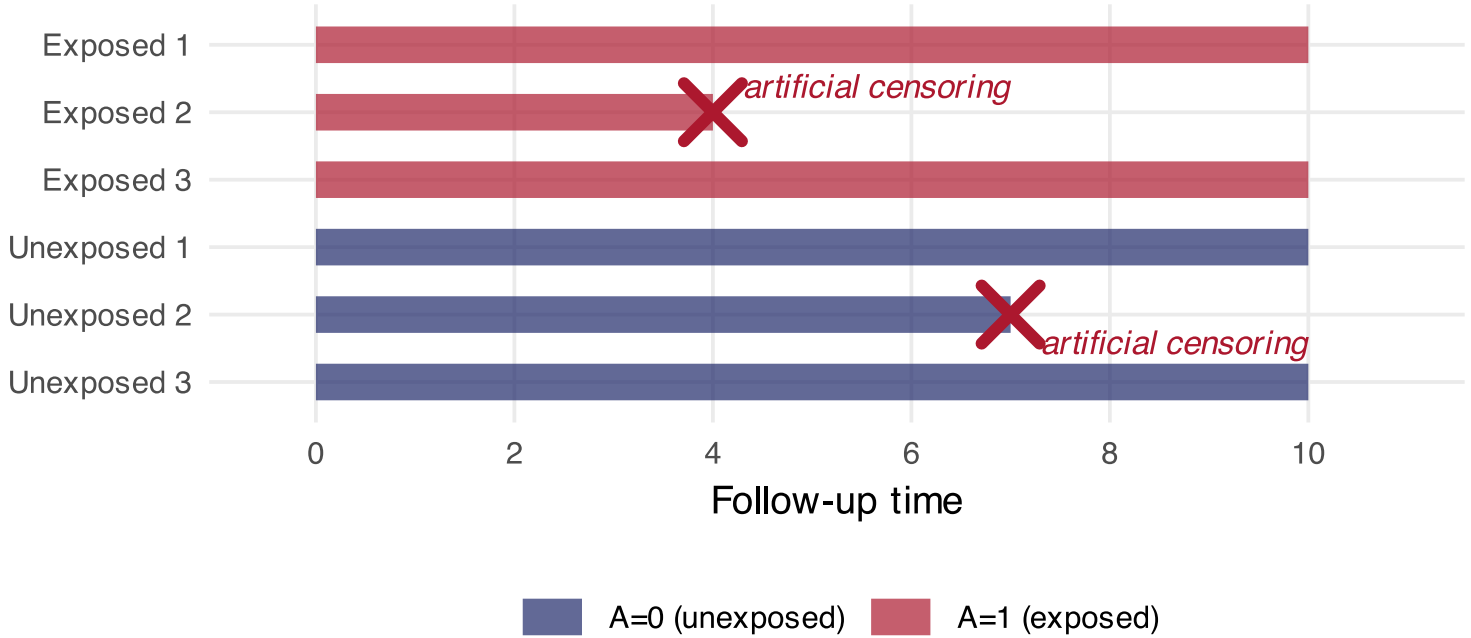
Two groups... and deviations



Transparent portion: the individual continues to be followed, but outside their initial strategy.

Artificial censoring at the deviation

To estimate the effect of the *sustained* strategy, we **censor** at the point of deviation.



Is this censoring informative?

Treatment is not stopped (or started) at random:

- **Stopping**: spontaneous patient decision (fatigue, cost), prescription not renewed, change of physician without continuity of care...
- **Late start**: deterioration in health status, hospitalisation...

Patients who deviate have a different prognosis from those who remain in their strategy → artificial censoring is informative!

Solution: IPCW

Weight the uncensored individuals by the inverse of their probability of remaining in the strategy, conditional on their prognostic characteristics.

SANTÉ SORBONNE UNIVERSITÉ

7

Target trial emulation

HRT AND CORONARY HEART DISEASE — A CONTRADICTION

Hormone replacement therapy (HRT) and risk of coronary heart disease in postmenopausal women:

Observational study *(Nurses' Health Study)*

Comparison: **prevalent** users vs never-users

→ **–30% coronary risk**

Randomised trial *(Women's Health Initiative)*

Comparison: new **initiators** vs non-initiators

→ **+20% coronary risk**

Same apparent question, similar data. Opposite results. Why?

HRT AND CORONARY HEART DISEASE — NOT A CONFOUNDING PROBLEM

The re-analysis of the *same data* from the Nurses' Health Study, this time comparing **initiators** to non-users, yields results **consistent with the trial** *(Hernán et al., Epidemiology 2008)*.

The real problem: the design

The two studies answered **different questions**:

- **Cohort**: women *already on HRT* for years at the start of follow-up (vs. non-users)
- **WHI**: women who *initiate* HRT at the start of follow-up (vs. non-initiators)

It is not confounding that explains the discordance, but designs that do not answer the same question.

Target trial emulation (Hernán & Robins, 2016)

Idea: emulate, with real-world data, the trial one would have wanted to conduct

Step 1: Define

Write the complete protocol of the **target trial**: who to include, which strategies to compare, which outcome, which time zero...

As if one were actually going to conduct the trial.

Step 2: Emulate

Reproduce each component of this protocol with the **available observational data**, explicitly identifying the necessary approximations.

Components of the target trial

Component	Randomised trial	Observational emulation
Eligibility	Defined a priori	Same (depending on available data)
Compared strategies	Defined a priori and **imposed**	Defined a priori but **observed** in the data
Assignment	**Randomisation** → exchangeability guaranteed	**Not randomised** → adjustment (G-computation, IPTW, ...)
Alignment at start of follow-up (time zero T_0)	**Natural**: E = A = T_0 by construction	**To be constructed**: align E, A, and T_0 explicitly
Outcome	Defined a priori	Same (depending on available data)
Analysis	ITT / per-protocol	ITT-analogue / PP-analogue

Making the causal question explicit forces identification of potential biases before analysis.

Aligning E, A, and T_0

Every emulation protocol must align three landmarks: **E** (eligibility), **A** (assignment), $\mathbf{T_0}$ (start of follow-up).

A misalignment is the source of many biases (long known and discussed!)

Misalignment of E, A, and T_0

Depletion of susceptibles

T_0 after A = E: follow-up begins after patients have initiated treatment. At the start of follow-up, the most fragile have already died or stopped treatment: only selected survivors are observed.

Example: women who have been on HRT for 10 years are not comparable to non-initiators.

→ Solution: exclude prevalent users (**new-user design**)

Misalignment of E, A, and T_0 (cont.)

Immortal time bias

A after T_0 = E: follow-up starts at eligibility, but treatment is defined or assigned later. Patients classified as "treated" have necessarily survived until treatment → the interval $[T_0, A]$ is immortal.

Example: include patients operated on within 30 days with T_0 = admission; the "operated" patients have necessarily survived until the operation.

→ Solution: **align A and T_0**

G-COMPUTATION, IPTW, IPCW, ..., IN EMULATION

Target trial emulation defines the **design**; the methods seen today are its **analytical tools**:

Emulation is not an analysis method!

Emulation means

1) defining a causal question
2) defining a design that allows answering it from observational data

Thinking about the analytical methods to reach an estimate comes only after that.

A COMMON SITUATION: THE GRACE PERIOD

Situation: compare patients operated on within 30 days of admission to unoperated patients, for 10-year survival.

Naïve analysis: T_0 = admission, A = "operated within 30 days" → defined by looking at what happens *after* T_0.

→ A is after T_0: patients classified as "operated" have all survived until surgery: **immortal time bias**.

How to correctly align E, A, and T_0?

The grace period

Target trial: at T_0 = admission, define two strategies:

- Strategy 1: be operated on within 30 days
- Strategy 2: not be operated on within 30 days

These 30 days represent what is called a **grace period**

Clone → censor → weight

→ **Clone** each patient: assign each patient alternately **to each of the two strategies**

- E = A = T_0 aligned for all clones.

→ **Censor** each clone as soon as they deviate from the strategy assigned to them:

- "Operated" clone not operated by day 30 → censored at day 30
- "Unoperated" clone operated before day 30 → censored at the date of operation

→ These artificial censoring events are **informative** (the most severe cases tend to be operated on more quickly) → **IPCW**

Conclusion

Each tool addresses a specific threat to the validity of causal estimation:

Problem	**Tool**
Baseline confounding	G-computation, IPTW
Informative censoring / missing data	IPCW
Non-compliance, grace period	Artificial censoring + IPCW
Design bias (time zero, prevalent users)	Target trial emulation

These tools are complementary: choosing them means first understanding the **source of bias** one seeks to correct.

Practical session: eusebe.github.io

SANTÉ SORBONNE UNIVERSITÉ

References

- Hernán & Robins — *What If? (open access: hsph.harvard.edu/miguel-hernan/causal-inference-book)*
- Hernán & Robins, *AJE* 2016 — *Using Big Data to Emulate a Target Trial*
- Hernán et al., *J Clin Epidemiol* 2016 — *Specifying a target trial prevents immortal time bias*
- Lodi et al., *AJE* 2019 — *Comparing Apples With Apples*