Introduction to Causal Inference and Causal Data Science

Day 3: Target Trial Emulation

Bas B.L. Penning de Vries
Dept. Data Science & Biostatistics
Julius Center, UMC Utrecht

UMC Utrecht

summerschool UTRECHT

Test-your-knowledge quiz

Question 1

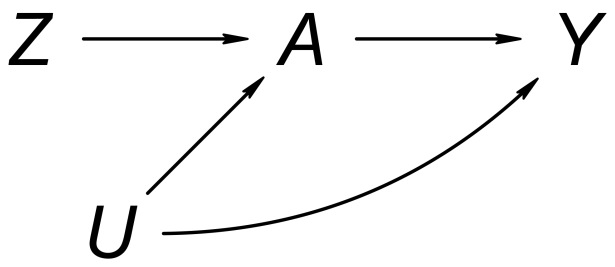
Which of the following options best describes the **meaning of a potential outcome** according to the counterfactual (or potential) outcomes framework?

a. A possible value of the outcome variable
b. The outcome of an individual that would be observed had treatment been set (by intervention) to a certain value
c. The best outcome an individual can achieve
d. A possible outcome of a study

Question 2

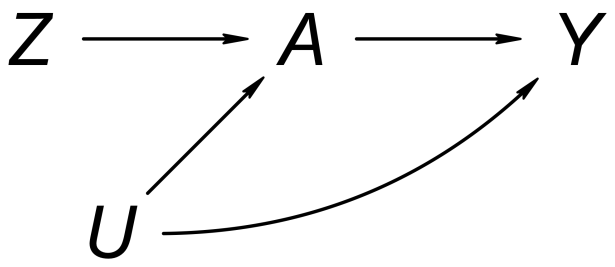
True or false? The **backdoor criterion** is fulfilled by a set of variables if (conditioning on) it closes all backdoor paths from treatment to outcome.

Question 3



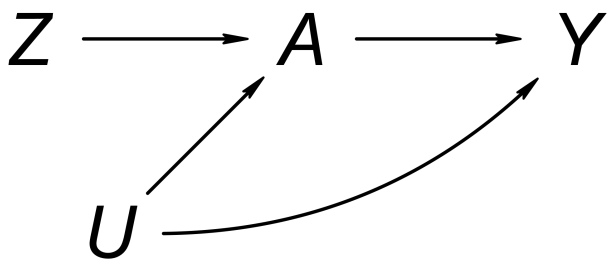
True or false? The **backdoor criterion** is satisfied (by the empty set) for treatment/exposure Z and outcome Y.

Question 4



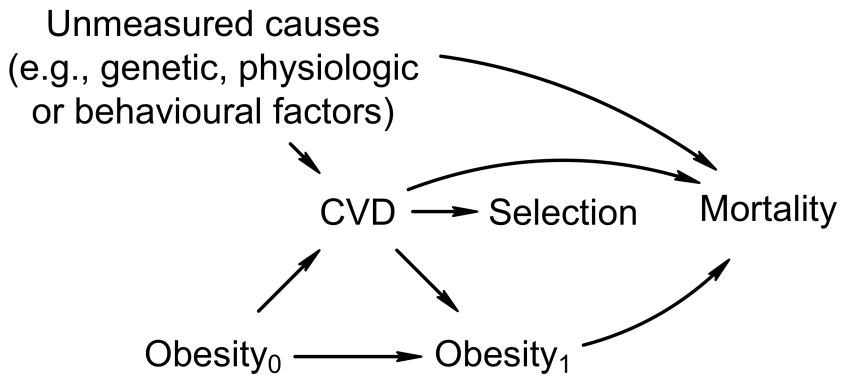
True or false? If the **backdoor criterion** is satisfied for *Z* and *Y*, then the exposure groups (defined by *Z*) are (marginally) **exchangeable** with respect to the outcome *Y*.

Question 5



True or false? The **backdoor criterion** is satisfied for treatment/exposure *A* and *Y*.

Question 6



True or false? The **backdoor criterion** is satisfied by (conditioning on) Selection for Obesity_1 (current obesity) and Mortality.

Question 7

True or false? Recent methodological developments allow epidemiologists to **falsify the presence of confounding using a statistical test** that does not rely on causal assumptions.

Learning objectives

By the end of today, you'll be able to

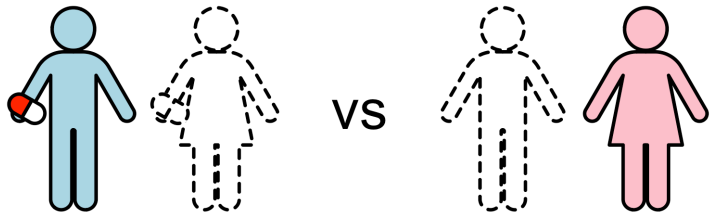
- Describe what is meant by a target trial and **target trial emulation**
- Identify **key components** of target trial emulation, including the determination of the start of follow-up (**time zero**)
- Recognise a **taxonomy of estimands** relevant to target trial emulation and distinguish between common targets such as intention-to-treat and per-protocol effects (*Friday!*)
- Describe the **relevance** of target trial emulation in causal inference from observational data
- Recognise **common deviations** from a target trial in observational studies
- Explain the basics of commonly used **methods** to address these deviations

"**Causal inference from observational data** can be viewed as an attempt to **emulate a hypothetical randomized trial**"

Hernán and Robins, 2020, *Causal Inference: What If*

Why trials?

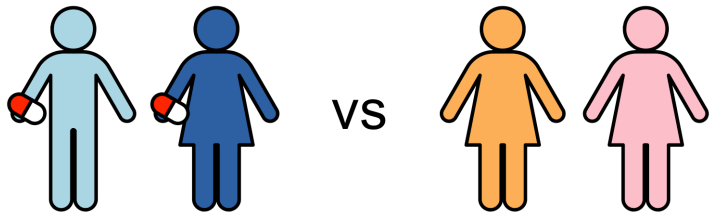
Causal inference is about speculating *what* would happen *if* ...



A **causal effect** is a contrast between the answers to what-if questions.

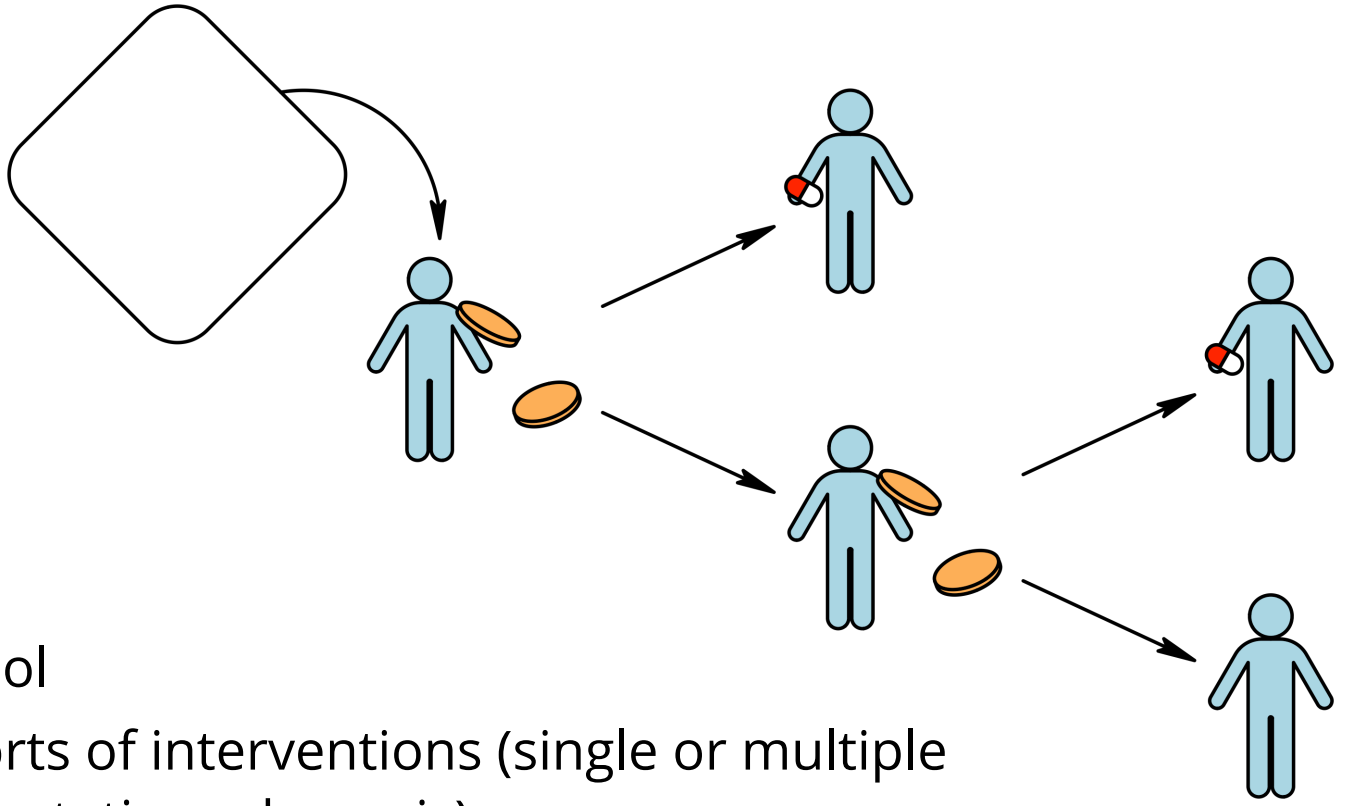
Fundamental obstacle: impossible to observe the consequences of ≥ 2 mutually exclusive actions (interventions, treatments, etc.)

Solution? Instead of comparing the *same individuals* between *different counterfactual ("what-if") situations*, ...



... compare *different individuals* who are actually *treated differently.*

Is this a **good solution?** Key concern: confounding



Randomisation

- Powerful (conceptual) tool
- Can accommodate all sorts of interventions (single or multiple time-point interventions, static or dynamic)

Why trials? Short answer: although the fundamental obstacle of causal inference means that causal effects cannot be observed directly — even in trials — they allow **causal effects** to be **readily identified** under **assumptions that are partly guaranteed by design**.

Example: identification average treatment effect (ATE) in a typical trial

$$\begin{aligned} \text{ATE} &= \mathrm{E}[Y^{A=1}] - \mathrm{E}[Y^{A=0}] && \text{(causal estimand)} \\ &= \mathrm{E}[Y^{A=1} \mid A=1] - \mathrm{E}[Y^{A=0} \mid A=0] && \text{(\textit{exchangeability} and \textit{positivity}, both guaranteed by design)} \\ &= \mathrm{E}[Y \mid A=1] - \mathrm{E}[Y \mid A=0] && \text{(\textit{consistency})} \end{aligned}$$

"Causation = correlation"!

Why not do trials?

- Expensive
- Unethical
- Impractical
- Untimely
- ...

But we can still target one ...

Target trial (emulation)

Target trial

A *hypothetical trial* that — if implemented — would readily allow us to answer our what-if question

- To help *communicate causal estimand* (because identification is “straightforward”)
- To *facilitate appraisal* of actual research designs (and *avoid methodological problems*)

Target trial emulation (TTE)

Explicit attempt to address deviations from a target trial, given the (observational) study data at hand

1. *Specify* target trial
2. *Emulate* it!

Step 1: specify target trial

	Target trial
Eligibility criteria	...
Treatment strategies	...
Outcome	...
Time zero and follow-up	...
Causal contrasts	...
Data analysis	...

Step 2: emulate

Compare and address departures from target trial (analytically)

	Target trial	Emulation study
Eligibility criteria	…	…
Treatment strategies	…	…
Outcome	…	…
Time zero and follow-up	…	…
Causal contrasts	…	…
Data analysis	…	…

Problem 1: Ill–defined or irrelevant interventions

Formulating a target trial helps to communicate the causal estimand and helps to avoid asking vague or “silly” questions (about ill-defined or irrelevant interventions).

- Eligibility defined by post-baseline events
- Causal effect of (a reduction/increase in) BMI?
- “Does water kill?” (Hernán, 2016)
- Unclear treatment strategies (e.g., stopping rules, dosage, etc.)

Treatment-variation (ir)relevance and well-definedness

- There may be many variations on an intervention and their impact on the outcome of interest need not be the same
- Interventions are sufficiently **well-defined** if there is no ambiguity about the variation or all possible variations equally affect the outcome variables of interest (i.e., there is treatment-variation irrelevance)
- Prerequisite of consistency

🔑 *Having to write a trial protocol forces you to be explicit and precise!*

Problem 2: Immortal-time bias

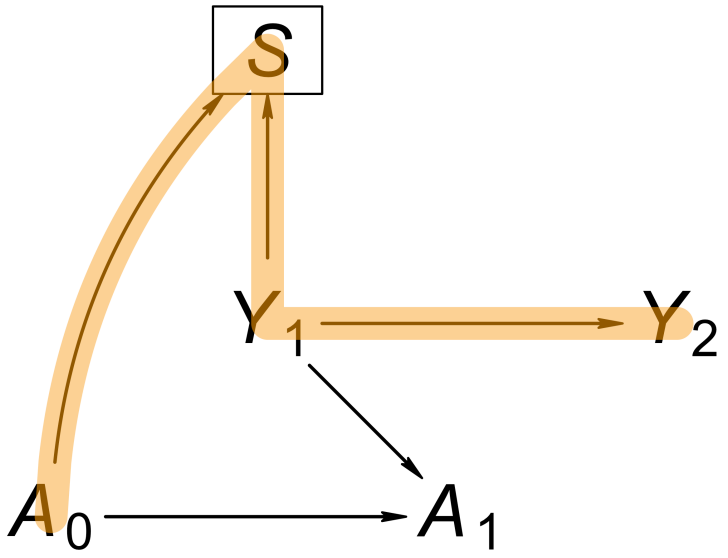
Example: do statins prevent cancer?

Table 2—*Effect of Statin Therapy and Its Duration on the Odds for Lung Cancer*

Variables	Cases, No.	Controls, No.	Crude OR	95% CI	Adjusted OR*	95% CI*	p Value for Adjusted OR
Overall	7,280	476,453					
Not exposed to statins	5,286	314,785					
Exposed to statins before lung cancer diagnosis (statin use > 0 yr)	1,994	161,668	0.73	0.70–0.77	0.55	0.52–0.59	< 0.01
Duration of statin use, yr							
0–0.5	446	10,259	2.59	2.34–2.86	2.32	2.05–2.63	< 0.01
0.5–1.0	214	15,564	0.82	0.71–0.94	0.75	0.63–0.89	< 0.01
1.0–2.0	416	30,590	0.81	0.73–0.90	0.70	0.61–0.79	< 0.01
2.0–4.0	649	55,516	0.70	0.64–0.76	0.49	0.44–0.55	< 0.01
> 4.0	269	49,739	0.32	0.28–0.36	0.23	0.20–0.26	< 0.01

*Adjusted for effects of age, race, sex, BMI, smoking, alcohol use, and diabetes.

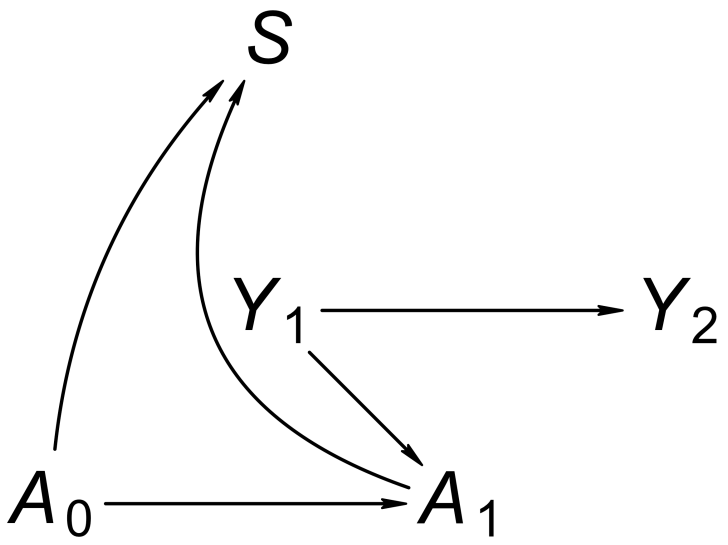
- **Immortal time** — a period of follow-up during which death or the study outcome cannot occur by design
- Arises from using *postbaseline information* to define …
  - inclusion/eligibility/selection, or
  - the exposure/contrast — *against trial principles!*
- May result in bias depending on how it is handled!
- Key to depicting this in a DAG is to include time-specific instances of variables



Immortal-time bias by selection

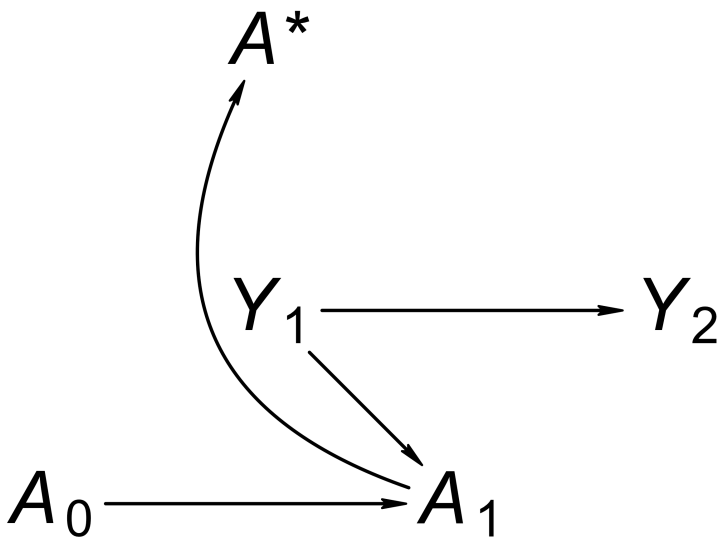
🧩 Are groups defined by A_0 exchangeable relative to outcome Y_2 conditional on $S = 1$?

(Hint: use the *backdoor criterion!*)



Immortal-time bias by selection (2)

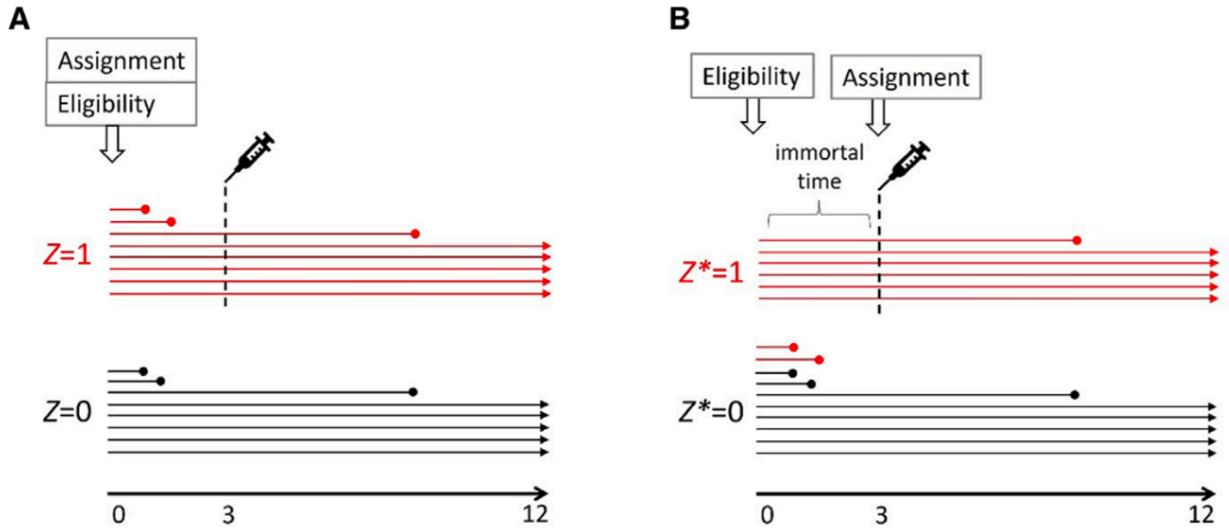
- S is a descendant of A_0 — violation of backdoor criterion!
- A_0 and Y_2 are marginally independent but not necessarily conditional on $S = 1$!



Immortal-time bias by misclassification

- Eg: compare ‘wait-time + surgery’ vs ‘no surgery’
- Can arise when we include everyone (no selection!) but make the *wrong contrast*

Example: ‘wait-time + surgery’ vs ‘no surgery’ (Hernán, 2025)



Problem 3: Including prevalent users

- Would you consider *initiating* a treatment regime now (at baseline) for a patient who is *already* on treatment (prevalent user)?
- Prevalent users are not part of the target population!
- Inclusion might result in (selection) bias!

Misalignment of **eligibility**, **treatment assignment**, and the **start of follow-up** can result in time-related bias such as immortal time and selection of prevalent users.

Matthews et al, 2023

Too few incident users at any given time?

- In studies on the effect of statin use and cancer incidence, there are few incident users at specified time baseline t_0 (or in short period starting at t_0)
- Consider a *sequence of trials* that are identical except for their baseline times
- To gain efficiency, we could emulate multiple such trials and *analyse simultaneously* (possibly according to flexible modelling assumptions to reflect heterogeneity across trials) — **sequential target trial emulation**
- NB: because individuals can be eligible for randomisation in multiple trials, need to respect *clustering* in estimating standard errors and constructing confidence intervals!

Statin-cancer example revisited

- Dickerman et al. (2019):
  - When *reanalysing* the observational data using the same approach as in earlier analyses, they found effect estimates *similar* to those found in earlier observational studies.
  - When applying trial principles to analyse observational data (*emulating a trial*), they found effect *estimates close to null*.
- Discrepancies between trials and observational studies are often attributable largely to sources of bias other than residual confounding!

Summary

Target trial emulation (TTE)

Explicit attempt to address deviations from a target trial, given the (observational) study data at hand

1. *Specify* target trial
2. *Emulate* it!

The target trial framework provides an **organizing principle** for the design of observational studies that leads to **clinically interpretable results** and analytic approaches that can **prevent common biases**.

Matthews et al, 2023