Using causal diagrams to understand problems of confounding and selection bias

Stijn Vansteelandt

Ghent University, Belgium
London School of Hygiene and Tropical Medicine, U.K.

MiRoR, Ghent

Often many explanations behind associations

'it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1.'

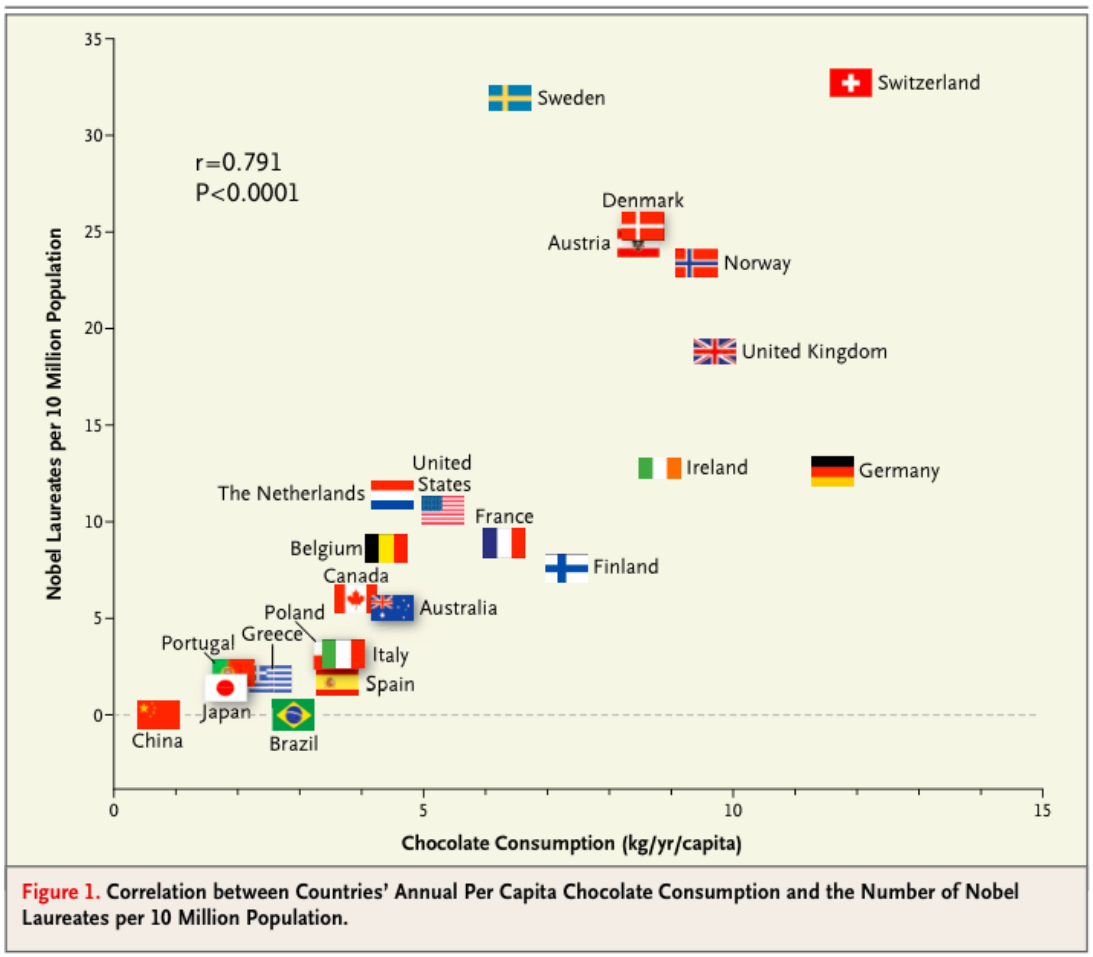


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

What might explain this?

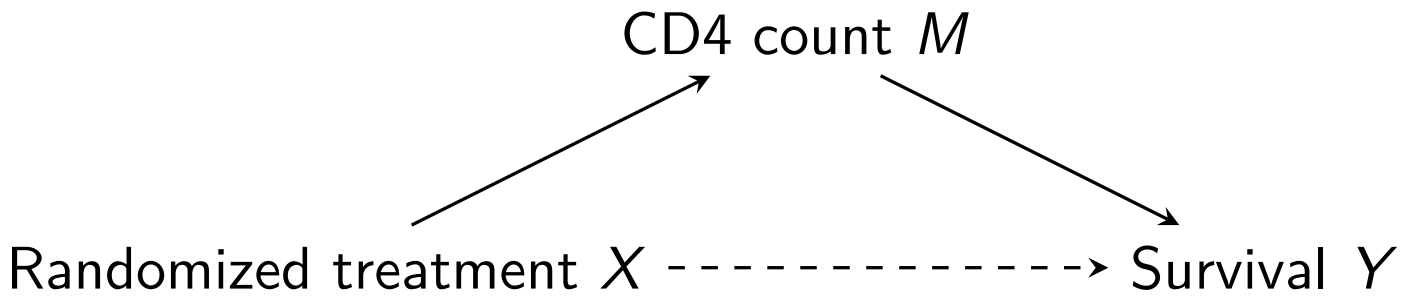
Causal diagrams

To gain insight into the origin of associations,
causal diagrams are becoming increasingly popular.

motivating example: search for biomarkers

- Pressure for accelerated evaluation of new AIDS therapies have led to CD4 and viral load as endpoints replacing time to clinical events.
- This raises the question whether an effect on the biomarker provides evidence for a clinical effect.

Example: search for surrogate markers



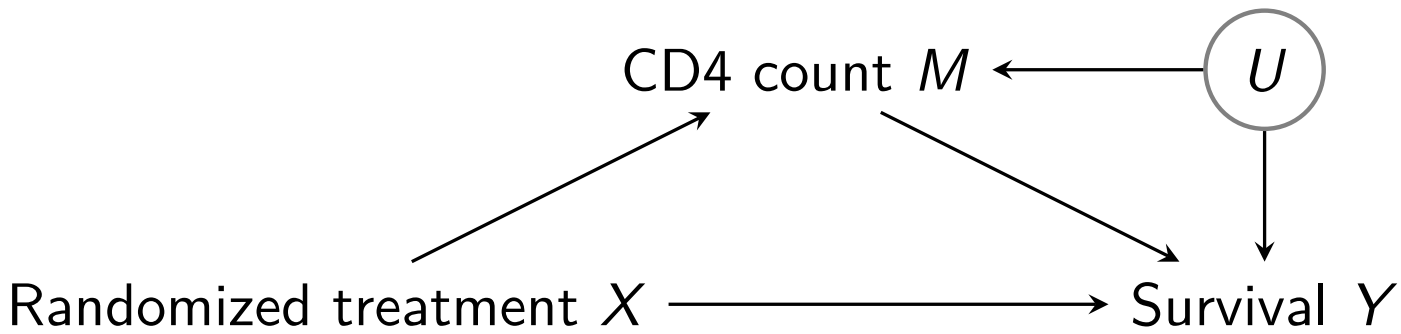
scientific question

- Is effect of treatment on clinical endpoint entirely mediated by its effect on the biomarker?
- Is there a direct effect of treatment on the clinical endpoint, not through the biomarker?

Causal diagrams

- To gain insight, we use causal graphs, causal diagrams, causal Directed Acyclic Graphs (DAG) or causal Bayesian networks.

(Pearl, 2000)

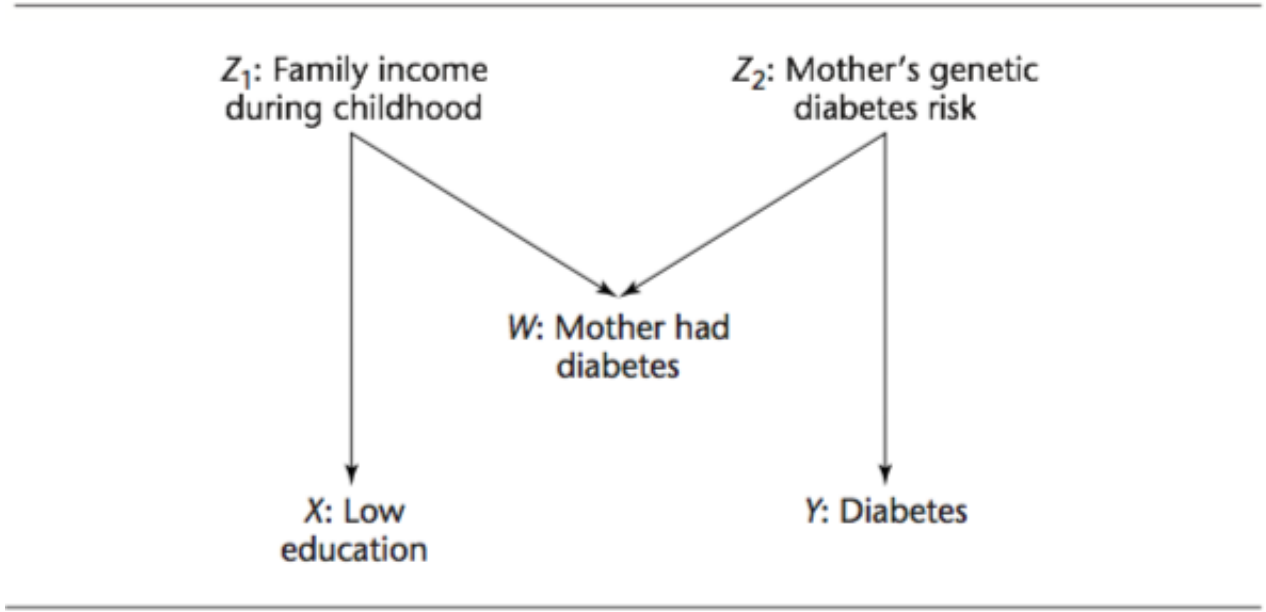


- Informally, these are graphical representations of the (causal) data-generating mechanism, for which we shall adopt the structure of a DAG.

Directed Acyclic Graph (DAG) or Bayesian network

a system of directed edges between variables, without cycles.

Example



This diagram expresses that the data may have been obtained by a data-generating mechanism such as:

- First, generate Z_1 and Z_2 independently.
- Next, generate W in function of Z_1 and Z_2.

  e.g. W is binary $(0/1)$ with success probability $\text{expit}(2Z_1 - Z_2)$.
- Next, generate X in function of Z_1.

  e.g. X is binary $(0/1)$ with success probability $\text{expit}(-1 + 0.5Z_1)$.
- Finally, generate Y in function of Z_2.

Causal DAGs

We make the DAG causal
by letting each edge express the possibility of a direct causal effect.

Exclusion restriction

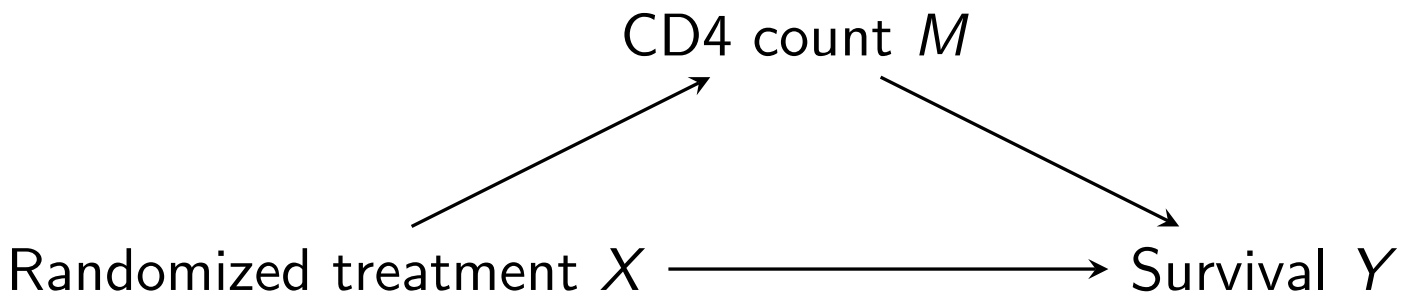
When there is no arrow from X directly into Y, manipulating X will not change Y once all parents of Y are manipulated.

For this interpretation to be justified,
one must adhere to the following principle.

no omitted confounders assumption

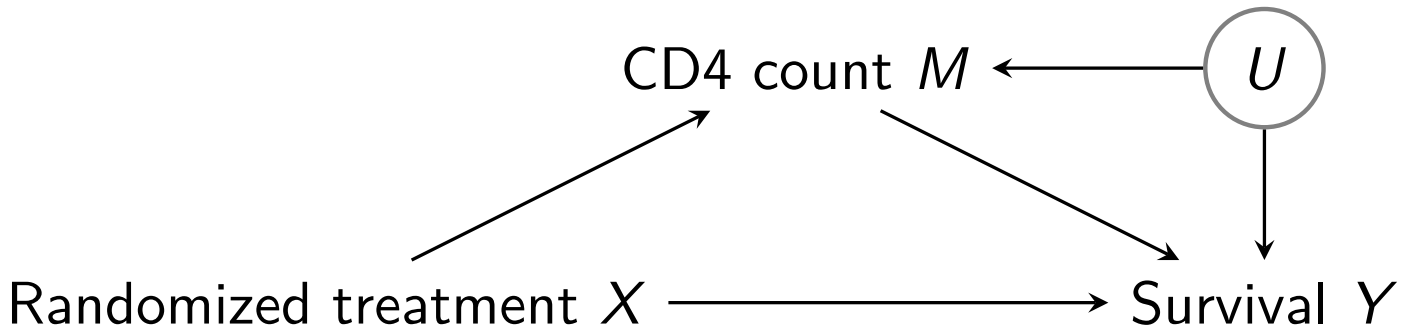
A causal DAG includes all common causes of any two variables.

Example: search for surrogate endpoints



- By randomization,
  no variables (measured or unmeasured) pointing to X.
- No omitted confounders, affecting X, must be added.
- This thus formally expresses the assumption
  that X is randomised!

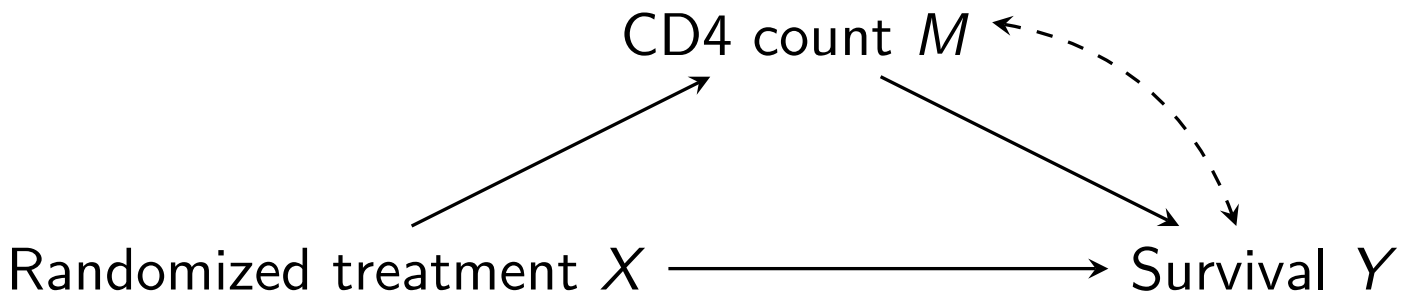
Example: search for surrogate endpoints



- There may be (unmeasured) health characteristics U jointly affecting CD4 count M and survival Y.
- Even if unmeasured, U must be added.

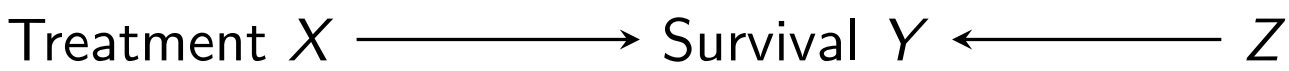
An alternative way to visualise common causes

We represent association between M and Y by means of an unmeasured common cause;
some authors use double-headed arrows.

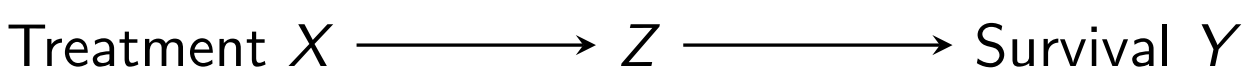


How to keep a causal DAG 'manageable' in practice?

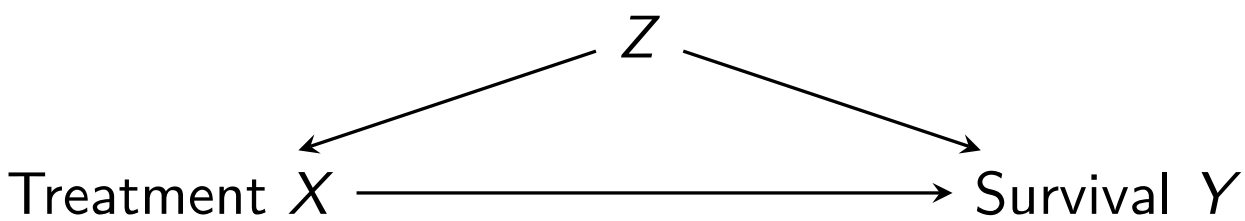
- A causal DAG need not include variables that are not of interest and not common causes of 2 variables in the DAG.



- A causal DAG need not include variables that lie on the causal path between an exposure and an outcome when there is no specific interest in them.



- Each node can represent a collection of (e.g. 50) variables.



- This has the advantage that no assumptions must be made about the causal relations between those components.

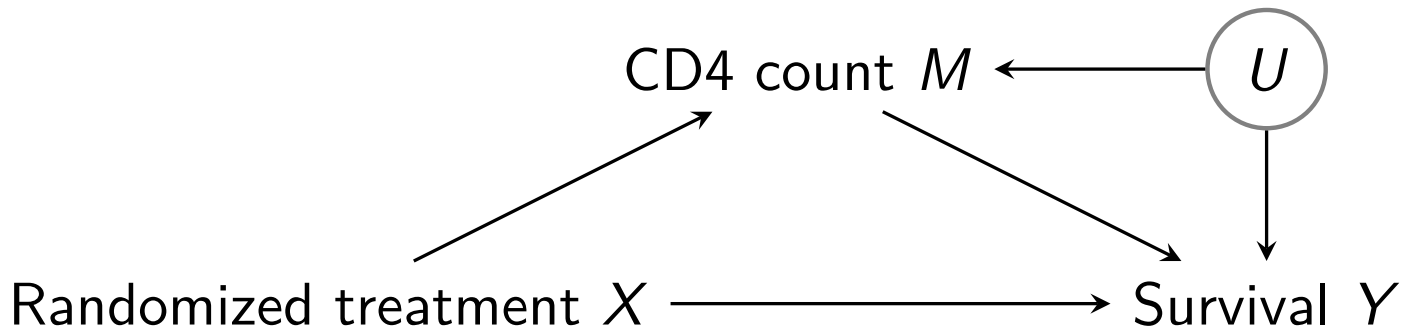
Causal diagrams versus path analysis

- In summary, a causal diagram forms a graphical, nonparametric representation, based on expert knowledge, of how the data were generated.
- It embodies causal assumptions, such as about:
  - the direction of causality;
  - the possible absence of causal effects between some measurements;
  - the possible absence of confounders;
  - the study design (e.g. ascertainment, missing data, ...)

  but no modelling assumptions.

How to use causal diagrams?

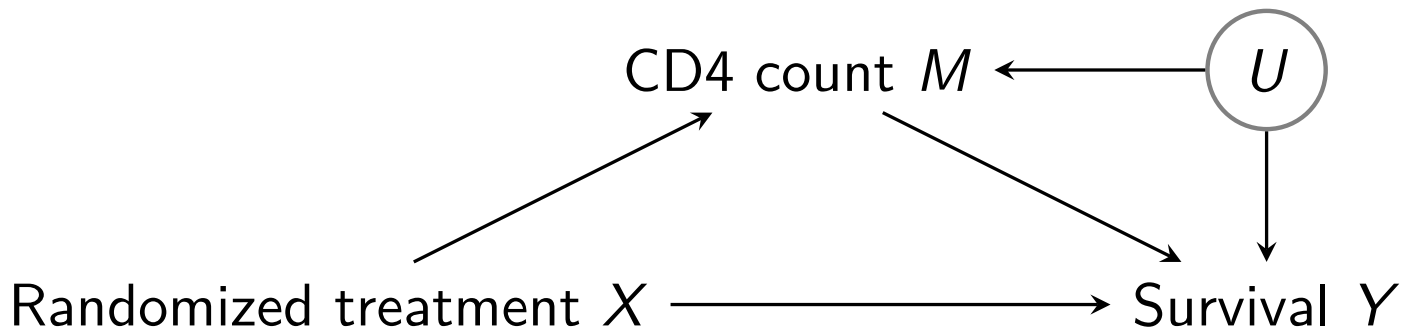
- On the causal diagram,
  we can assess how X may causally affect Y.
- A variable X in a causal diagram can only causally affect
  a variable Y when there is a directed path from X to Y.



- For instance, X may have a direct causal effect on Y,
  as well as an indirect causal effect which is mediated by M.
- X does not causally affect Y along the path $X - M - U - Y$!

How to assess association in causal DAGs?

- On the causal diagram,
  we can assess how X may be associated with Y.
- The association between 2 variables is driven by
  possible associations along all directed and undirected paths
  that connect these variables.



- To understand which paths explain the association,
  we use d-separation:
  a graphical rule to read off independencies implied by a DAG.

(Pearl, 1995, 2000).

d-separation

- To understand what causes Y and X to be associated, we think of a DAG as an electric net.

  - colliders C are inactive

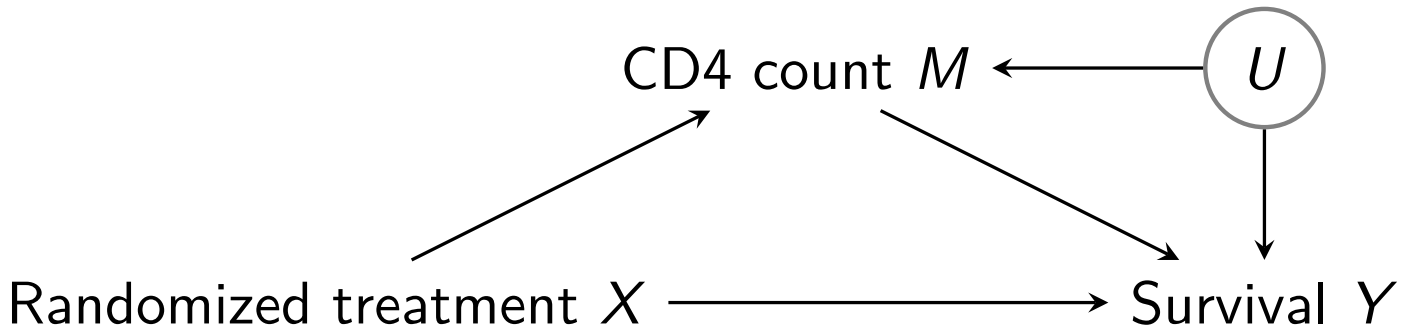
    $$X \rightarrow C \leftarrow Y$$

  - non-colliders C are active

    $$X \rightarrow C \rightarrow Y \text{ or } X \leftarrow C \rightarrow Y$$

- If there is no electric current between X and Y, then they are independent.
- There may be association along all active paths.

Example: search for surrogate endpoints



The association between X and Y is due to

- the direct causal effect,
- the indirect causal effect through M,
- but not due to a possible spurious association along the path $X - M - U - Y$.

We thus find that for the total effect, association = causation.

Adjusting or conditioning changes dependencies

- Suppose now that we 'adjust the analysis for C', either by restricting the analysis to subjects with the same value of C, or by including C in a regression model

$$E(Y|X, C) = \alpha + \beta X + \gamma C$$

- If there is no electric current between X and Y after adjusting for C, then X and Y are independent, conditional on C.
- There may be conditional association along all active paths.

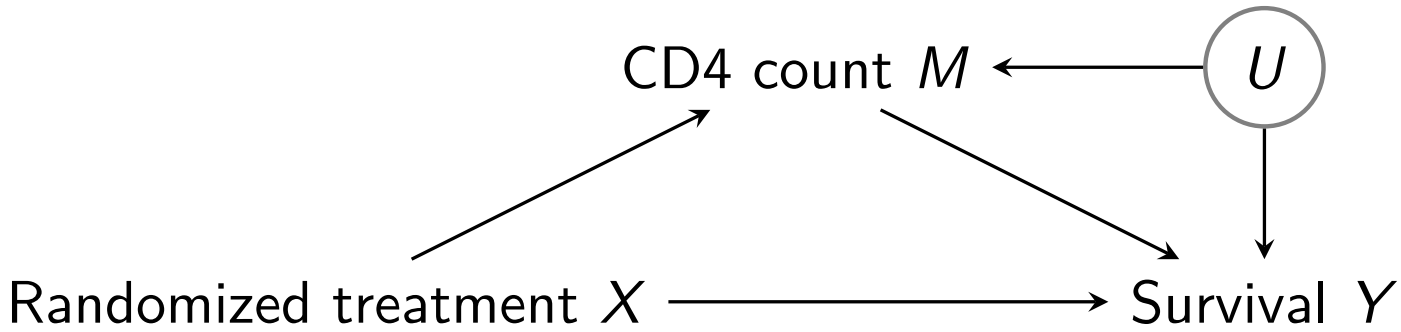
d-separation after conditioning

- Adjusting for a non-collider C changes
  active $\rightarrow$ inactive
- Adjusting for colliders or their descendants C changes
  inactive $\rightarrow$ active

The latter goes against intuition and is a source of much error.
It explains why e.g.

- short basketball players tend to be faster than tall ones;
- college students with poor math abilities tend to be good at sports;
- hospital patients without diabetes are more likely to have cholecystitis;
- ...

Example: search for surrogate endpoints

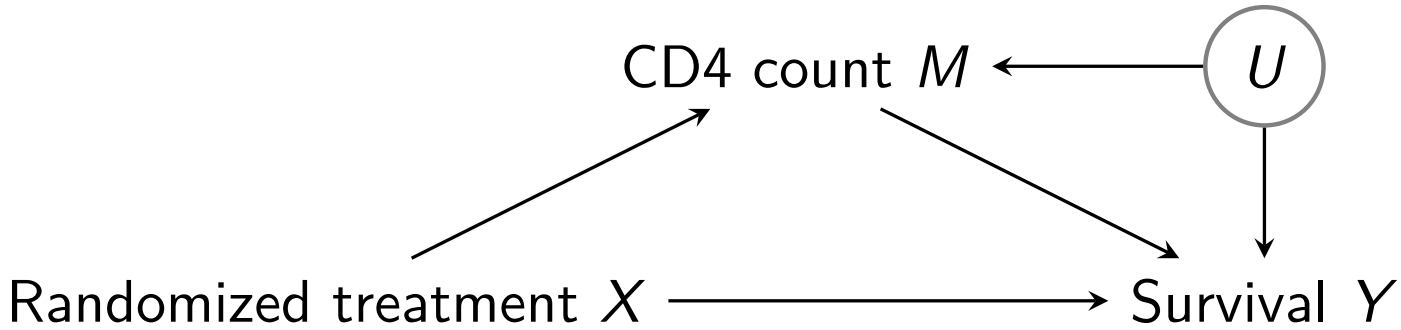


Conditional association between X and Y, given M is due to

- the direct causal effect,
- spurious association along the path $X - M - U - Y$,
- but not due to the indirect causal effect through M.

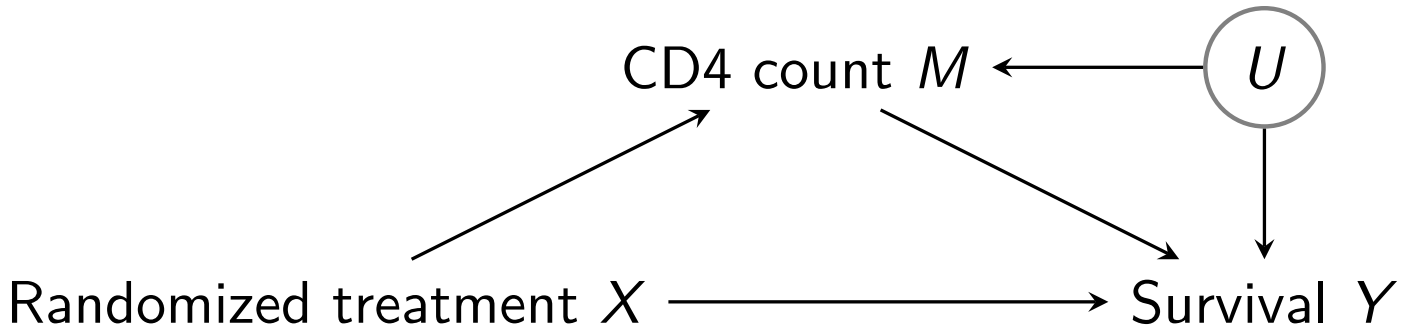
We thus find that for the direct effect, association $\neq$ causation.

Why does conditioning on a collider induce bias?



- Suppose that both treatment X and a low baseline level U of immunosuppression independently increase CD4 count.
- Then these attributes will be correlated among patients with high CD4 count.
- Indeed, untreated patients with high CD4 count likely have a low baseline level of immunosuppression, which explains their high CD4 count.

Example: search for surrogate endpoints



- Some criteria for validation of surrogate endpoints are based on testing whether $\beta = 0$ in model

$$E(Y|X, M) = \alpha + \beta X + \gamma M$$

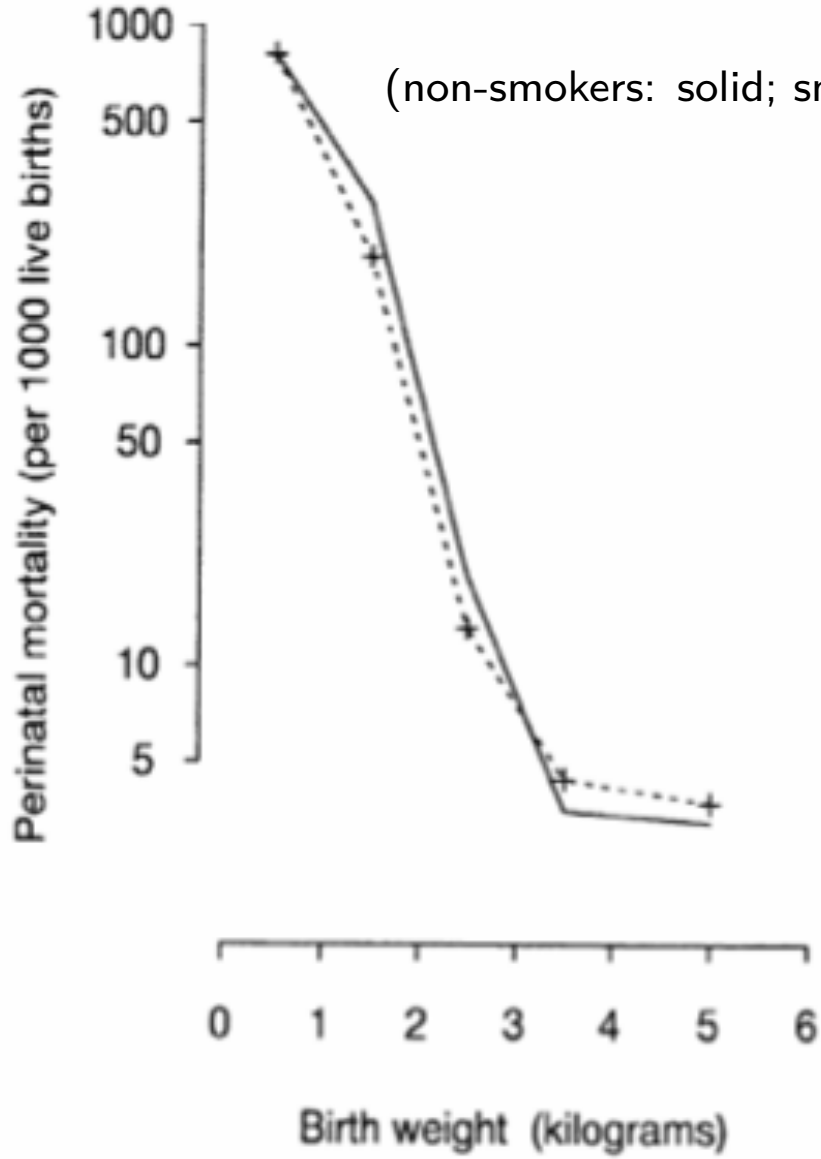
- These approaches are invalid
  in the presence of unmeasured confounders U.

Does it really matter?

- Birth weight is strong predictor of infant mortality.
- Investigators have therefore frequently stratified on birth weight when evaluating the effect of maternal smoking on infant mortality.

  (Yerushalmy, 1971; Wilcox, 1993)

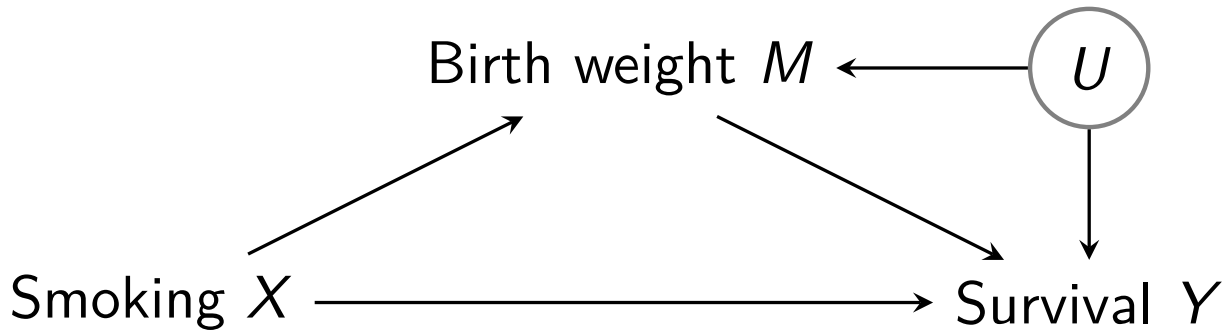
Kaiser Foundation Health Plan, SF, 1960-67



Does it really matter?

- Survey of 1991 U.S. births reveals that infant mortality rate ratio for exposed infants versus nonexposed infants is 0.79 (95% CI: 0.76, 0.82) among LBW infants.
- Birth weight paradox has been a controversy for decades.
- One suggestion is that the effect of maternal smoking is modified by birth weight
  in such a way that smoking is beneficial for LBW babies.

Does it really matter?



- Although birth weight is a strong predictor of infant mortality and adjustment is therefore common,
  it is inappropriate for answering this research question.
- The unadjusted rate ratio 1.55 (95% CI: 1.50, 1.59)
  expresses the causal effect (provided no further confounders).

Summing up

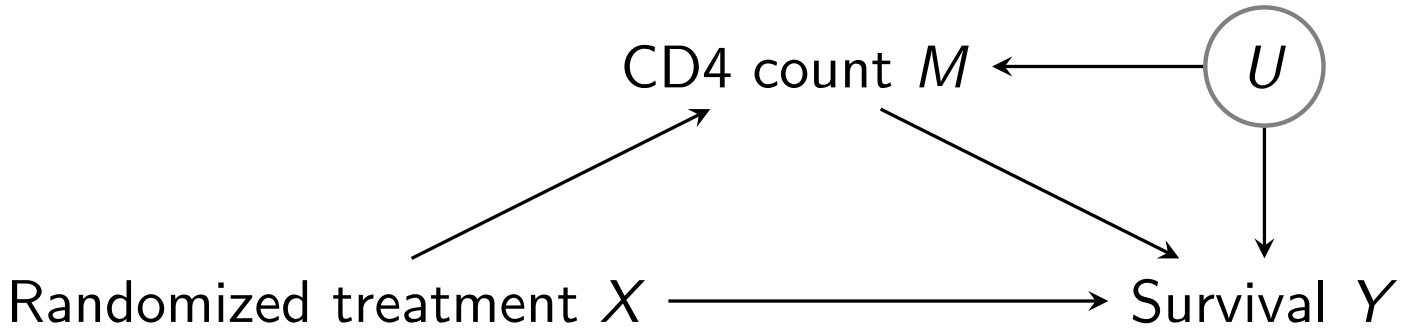
- The reason why standard approaches may fail, is because they try to uncover causation from statistical associations, but association $\neq$ causation.
- For instance, the decision to adjust for birth weight is based on birth weight having a strong association with infant mortality, but this has nothing to do with causal arguments.
- The only way to learn about the effect of some exposure on some outcome, is to express background knowledge about 'what may have a causal effect on what'.
- We can do this via causal diagrams.

Summing up

- Using d-separation, we can infer for which confounders C we need to adjust when estimating the effect of X on Y.
- Such adjustment may happen via standard regression

$$E(Y|X, C) = \alpha + \beta X + \gamma C$$

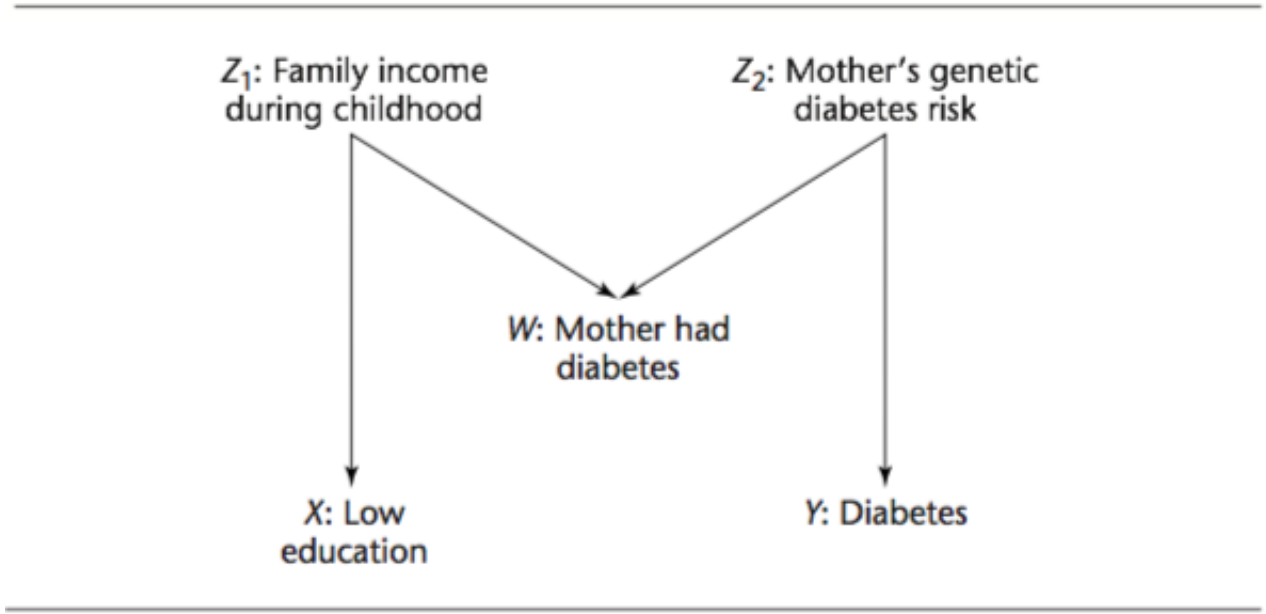
Summing up



Take home message 1: Mediation analyses demand confounding adjustment, even in randomized experiments

- They demand adjustment
  for confounding of the mediator - outcome association.
- The fact that the exposure is randomly assigned,
  does not prevent such confounding.

Summing up



Take home message 2: Standard criteria for covariate selection can be very misleading

They demand adjustment for strong correlates of the outcome,

regardless of whether the end result retains a meaningful interpretation.

References

Cole S and Hernan MA. Fallibility in estimating direct effects. *Int J Epidem* 2002; **31**:163-165.

Glymour MM. *Using causal diagrams to understand common problems in social epidemiology.* In Methods in Social Epidemiology, Oakes M, and Kaufman J, eds. Jossey-Bass. 2006.

Greenland S, Pearl J and Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**:37-48.

Pearl J. Causal Diagrams for Empirical Research (with discussion). *Biometrika* 1995; **82**:669-710.

Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001; **12**:313-320.

Methods in Research on Research

A presentation delivered at the

first MiRoR training event
October 19-21, 2016
Ghent, Belgium

This project has received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement #676207