Introduction to causal inference

Krista Fischer

Institute of Mathematics and Statistics, University of Tartu
Institute of Genomics, University of Tartu
Estonian Academy of Sciences

Statistical Practice in Epidemiology, Lyon 2024

Statistical associations vs causal effects in epidemiology

Does the exposure (smoking level, obesity, etc) have a causal effect on the outcome (blood pressure, cancer diagnosis, mortality, etc)?

is not the same question as

Is the exposure associated with the outcome?

Conventional statistical analysis will answer the second one, but not necessarily the first.

Example 1

What is the effect of (white) bread consumption on body weight?
(Estonian Biobank)

```
> signif(summary(lm(weight~wbread+fruit,data=fe1))$coef,3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     75.10      0.174  431.00 0.00e+00
wbread           1.29      0.183    7.05 1.88e-12
fruit           -1.04      0.159   -6.51 7.56e-11
```

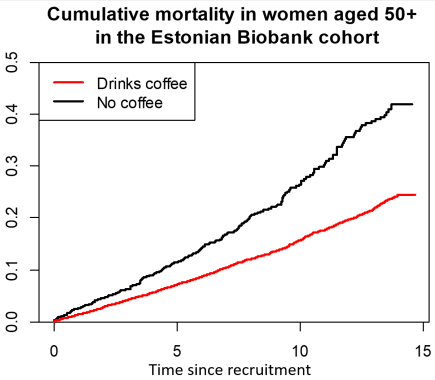
Adding sex effect to the model:

```
> signif(summary(lm(weight~ wbread+fruit+sex,data=fe1))$coef,3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    84.70      0.193  438.00 0.00e+00
wbread         -0.59      0.171   -3.45 5.54e-04
fruit           1.02      0.149    6.81 9.97e-12
sex           -13.40      0.148  -90.20 0.00e+00
```

Does white bread increase and fruit decrease body weight — or the other way around???

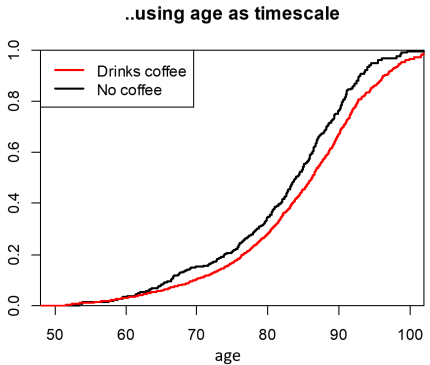
Example 2

Does coffee-drinking prolong life?



. . . and is the effect really that huge?

Example 2 (cont.)



Does coffee-drinking prolong life?
Or: do coffee-drinkers live longer (for several reasons)?

How to define causal effects (properly)?

- One can think of some basic guidelines (sometimes called as “criteria”) that must be satisfied for causal effect to be identifiable.
- Such principles may include temporality (cause preceding the outcome), consistency (reproducibility), monotonicity (dose-response), plausibility (e.g. biologically), etc. (Bradford Hill’s guidelines)
- However, although such general guidelines are useful, they are often not sufficient to establish causality

What is a causal effect?

Possible questions . . .

1. In case I stop eating white bread, would my weight drop? or . . .
2. In case white bread were eliminated from the world, would the average weight of all people decrease? or . . .
3. If, in a "parallel worlds" we could a) eliminate white bread (thus no-one gets any bread) and b) have everyone eating, say 200 grams of white bread per day, would the average weight (after a while) differ in the two worlds.

Actually, question 3 is what we are going to explore next, but first let's define the counterfactual quantities of interest.

Causal effects and counterfactuals

To define valid causal contrasts, we counterfactual (what-if) thinking is useful.

Counterfactual outcomes

- Suppose Y is the outcome of interest and X is a binary exposure (for simplicity), thus $X = 1$ for exposed and $X = 0$ for unexposed individuals.
- Define $Y^1 = Y^{X=1}$ and $Y^0 = Y^{X=0}$ as individual's potential (counterfactual) outcomes if this individual's exposure level X were set to 1 or 0, respectively.
- For a particular individual, either Y^1 or Y^0 can be observed at any moment.
- Consistency:

$$\text{if} \quad X_i = x, \quad \text{then} \quad Y_i^x = Y_i^X = Y_i$$

(X is a random variable for the actual treatment.)

Example: Y^1 individual's blood pressure, if he/she were a smoker; Y^0 individual's blood pressure, if he/she were a nonsmoker;

Individual and average causal effects (ACE)

The individual causal effect

The treatment X has a causal effect on an individual's outcome Y if

$$Y^1 \neq Y^0.$$

In epidemiology we are mostly interested in the averages.

The (population) Average Causal Effect (ACE)

- An average causal effect of exposure X on outcome Y is present if

$$E(Y^1) \neq E(Y^0).$$

- For binary Y, this would be:

$$P(Y^1 = 1) \neq P(Y^0 = 1)$$

Do we usually estimate causal effects?

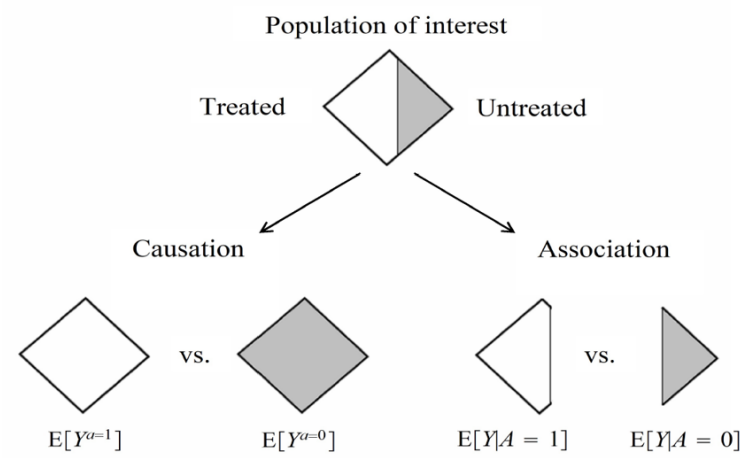
Causal effects are easily estimable, if: $Y^x \perp\!\!\!\perp X$,

as in this case $\mathrm{E}(Y|X=x) = E(Y^x|X=x) = E(Y^x)$.

This condition is, however, rarely satisfied.
If Y is weight and X the quantity of bread consumed (slices/day),then:

- $\mathrm{E}(Y^x)$ would be the average weight in population, in case everyone eats exactly x slices of bread a day;
- $\mathrm{E}(Y|X=x)$ would be the average weight in subset who actually eat x slices of bread a day;

The population can be divided into subsets with $X=0$, $X=1$, $X=2$, etc. $\mathrm{E}(Y|X=x)$ would be the average in one of such subsets, but $\mathrm{E}(Y^x)$ would be the average for the entire population for each possible value of x.



(by: M. Hernán and J. Robins)

The "naïve" association analysis

- With a binary exposure X, compare average outcomes in exposed and unexposed populations:

$$E(Y|X=1) - E(Y|X=0)$$

  Is cancer incidence different in smokers and nonsmokers?

- But mostly:

$$E(Y|X=1) \neq E(Y^1)$$

  Cancer risk in smokers is not the same as the potential cancer risk in the population if everyone were smoking

- Similarly:

$$E(Y|X=0) \neq E(Y^0)$$

- In most cases there is always some unobserved confounding present and therefore the naïve analysis does not provide causal effect estimates.

Potential outcomes (counterfactuals) in different settings

- Randomized trials: probably the easiest setting to imagine Y^X for different X.
- "Actionable" exposures: smoking level, vegetable consumption, ... – potential interventions may alter exposure levels in future.
- Non-actionable exposures: e.g genotypes. It is difficult to ask *"What if I had different genes?"*. Still useful concept to formalize genetic effects (heritability, attributable risk).
- Combinations: With X– a behavioral intervention level, Z– smoking level after intervention and Y–a disease outcome, one could formalize the effect of intervention on outcome by using $Y^{X,Z(X)}$

Estimability of causal effects in an ideal randomized trial

Causal effect is estimable in a randomized trial (RCT), where the subjects are randomly assigned to two arms: $R = 1$ and $R = 0$. In an ideal (but often not realistic) RCT:

$$\mathrm{E}(Y^0) = \mathrm{E}(Y^0|R = 0) \equiv \mathrm{E}(Y|R = 0),$$

because when $R = 0$ then also $X = 0$ and we observe $Y = Y^0$ (consistency!).
Similarly:

$$\mathrm{E}(Y^1) = \mathrm{E}(Y^1|R = 1) \equiv \mathrm{E}(Y|R = 1)$$

and thus we can estimate the causal effect of treatment in such study.

In epidemiology, however, exposures are not randomized and therefore the estimation of causal effects is complicated (and not always possible).

Does adjustment for covariates help?

Adjustment for confounders is helpful, if:

$$Y^x \not\perp\!\!\!\perp X, \quad \text{but} \quad Y^x \perp\!\!\!\perp X|Z,$$

(conditional exchangeability), implying

$$E(Y^x) \neq E(Y|X = x), \quad \text{but} \quad E(Y^x|Z) = E(Y|X = x, Z = z),$$

for a set of covariates Z.
(Within levels of Z, X is randomly "assigned".)

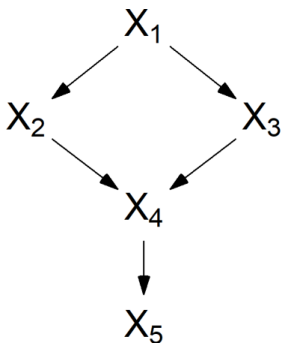
Under conditional exchangeability, one can test for the causal effect by simple regression adjustment. For estimation of the average causal effect, a few "tricks" can be used (Esa's lecture on Thursday).

Classical/generalized regression estimates vs causal effects?

- In the presence of confounding, regression analysis provides a biased estimate for the true causal effect
- To reduce such bias, one needs to collect data on most important confounders and adjust for them
- However, too much adjustment may actually introduce more biases
- Causal graphs (Directed Acyclic Graphs, DAGs) may be extremly helpful in identifying the optimal set of adjustment variables

DAGs: directed acyclic graphs

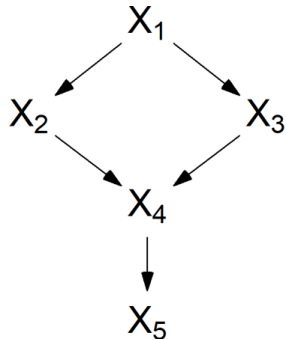
- A Directed Acyclic Graph (DAG) is a graphical representation of the causal association structure in the data, where variables are presented as nodes (points) and the associations are presented as edges (lines, arrows);



Terminology:

The parents of a node V_m are nodes with a direct arrow into V_m, whereas V_m is a child of its parents.

V_m is a descendant of V_j (and V_j is an ancestor of V_m) if following the direction indicated by the arrows, one can reach V_m by starting at V_j.



X_1 is a parent of X_2 and X_3; X_4 is a child of X_2 and X_3. X_1, X_2, X_3 and X_4 are ancestors of X_5, etc

A causal DAG:

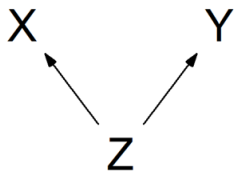
1. the lack of an arrow from node V_j to V_m can be interpreted as the absence of a direct causal effect of V_j on V_m
2. all (measured and unmeasured) common causes of any pair of variables on the graph are themselves on the graph
3. any variable is a cause of its descendants.

The causal Markov assumption:

Conditional on its direct causes, any variable V_j is independent of any variable for which it is not a cause (non-descendants).

"Classical" confounding

There are common causes (confounders) Z that have an effect on both, X and Y.

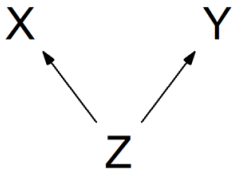


Also called as backdoor path between X and Y.
Implied statistical associations (Y is not independent of X in general, but it is independent of X, conditional on Z):

$$X \not\perp\!\!\!\perp Y \qquad X \perp\!\!\!\perp Y | Z$$

X and Y are independent, conditional on Z, but marginally dependent.

"Classical" confounding, mathematically



Assuming linear associations:

$$X = b_{0x} + b_{zx}Z + \varepsilon_x,\ \mathrm{E}(\varepsilon_x|Z) = 0$$

$$Y = b_{0y} + b_{zy}Z + \varepsilon_y,\ \mathrm{E}(\varepsilon_y|Z,X) = 0.$$

If $b_{zx} \neq 0$, then the correlation $\rho(Z,X) \neq 0$, and also:

$$\mathrm{E}(Z|X) = b_{0z} + b_{xz}X, \text{ where } b_{xz} \neq 0$$

We see that Y is associated with X, as now:

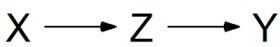
$$\mathrm{E}(Y|X) = b_{0y} + b_{zy}\mathrm{E}(Z|X) = b^*_{0y} + b_{zy}b_{xz}X,$$

with $b_{xz}b_{zy} = b_{xy} \neq 0$.

One should adjust the analysis for Z, by fitting a regression model for Y with covariates X and Z. There is a causal effect between X and Y, if the effect of X is present in such model.

Causal chain (mediation, front-door path):

The effect of X on Y is mediated by Z:



$$Y = \beta_0 + \beta_{xy}X + \beta_{zy}Z + \varepsilon,$$

- Don't adjust for Z, if you are interested in the total effect of X on Y
- Do adjust for Z, if you are interested in the direct effect of X on Y
- Adjusted analysis is valid only when the Z-Y association is unconfounded!

Collider

$$X \perp\!\!\!\perp Y, \quad \text{but} \quad X \not\perp\!\!\!\perp Y|Z$$

- Here the paths collide at Z and Z is called as the collider.
- In contrary to the situations seen before, adjusting for Z is opening the path between X and Y and so artificially creating the association between X and Y!
- Here you SHOULD NOT adjust the analysis for Z;
- If you still adjust for Z, you will create a collider bias.

Adjusting for a collider is wrong!

$Z = \beta_0 + \beta_{xz}X + \beta_{yz}Y + \varepsilon$, with $\beta_{xz} \neq 0$ and $\beta_{yz} \neq 0$

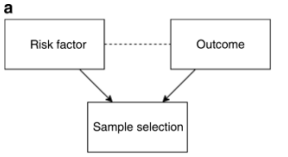
hence, there exist parameters $\beta_{xy} \neq 0$ and $\beta_{zy} \neq 0$, so that:
$Y = \beta_0^* + \beta_{xy}X + \beta_{zy}Z + \varepsilon^*$.

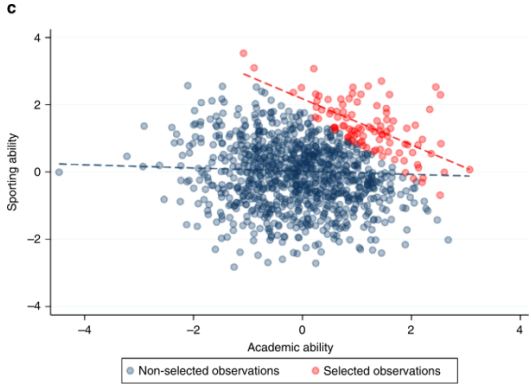
We see the association between X and Y only when the "effect" of Z has been taken into account.
But this is NOT a causal effect of X on Y.

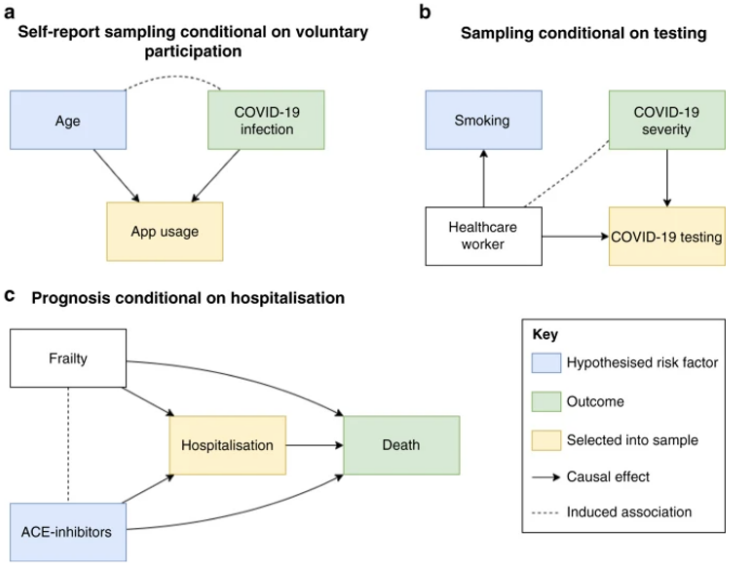
One should NOT adjust the analysis for Z!

Selection bias: a special (but common) case of collider bias

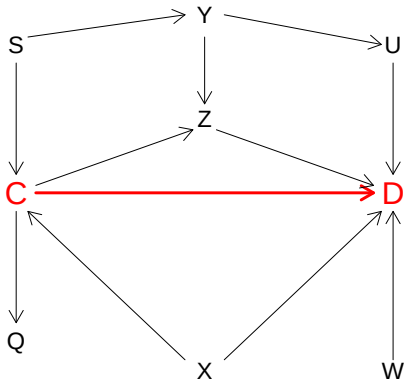
- All analysis are done conditional on the selected sample
- However, selection itself might be a collider (Griffith et al. 2020, https://www.nature.com/articles/s41467-020-19478-2)






a
Self-report sampling conditional on voluntary participation
Age
COVID-19 infection
App usage
b
Sampling conditional on testing
Smoking
COVID-19 severity
Healthcare worker
COVID-19 testing
c
Prognosis conditional on hospitalisation
Frailty
Hospitalisation
Death
ACE-inhibitors
Key
Hypothesised risk factor
Outcome
Selected into sample
Causal effect
Induced association

Actually there might be a complicated system of causal effects:



C-smoking; D-cancer

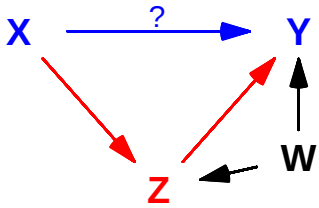
Q, S, U, W, X, Y, Z - other factors that influence cancer risks and/or smoking (genes, social background, nutrition, environment, personality, . . .)

What to do in complicated cases?

1. Sketch a causal graph
2. Identify all paths between the exposure and outcome (ways to go from X to Y regardless of the direction of the arrows).
3. Identify the closed paths that include colliders and open paths that don't.
4. You need to select adjustment variables that block all open paths.
5. Don't adjust for colliders (as they would open the closed paths)!
6. If you are looking for the total effects, you don't need to block the directed paths (that follow the directions of the arrows).
7. Often, there are unobserved confounders!

R package *dagitty* is useful for such tasks.

Example: mediation with confounding



Example

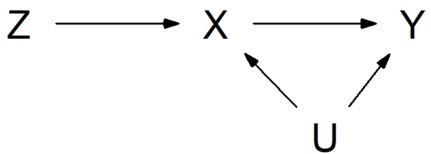
X: a genetic factor (SNP or a polygenic score)
Y: Type 2 Diabetes
W: Body Mass index (BMI)

Paths: $X \to Z \to Y$ (open) and
$X \to Z \leftarrow W \to Y$ (closed).

- The total effect of X on Y is estimable without any adjustment.
- For direct effect you need to adjust for Z, but that would open the closed path – to block that, you also need to adjust for W.
- If W is an unobserved confounder, direct effect of X on Y cannot be estimated.

Instrumental variables estimation: the idea

A DAG with the exposure X, outcome Y, confounder U and an instrument Z:



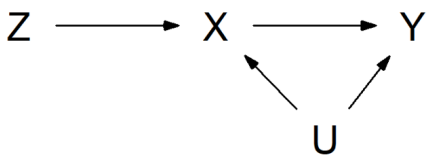
Assuming:
$Y = \alpha_y + \beta X + \gamma U + \epsilon,\ \mathrm{E}(\epsilon|X, U) = 0,$

simple regression will estimate:

$\mathrm{E}(Y|X) = \alpha_y + \beta X + \gamma \mathrm{E}(U|X).$

Thus the coefficient of X will be a biased estimate of β (as it also depends on γ).

Instrumental variables estimation: the idea



A variable Z is an instrument for the path $X \to Y$, if:

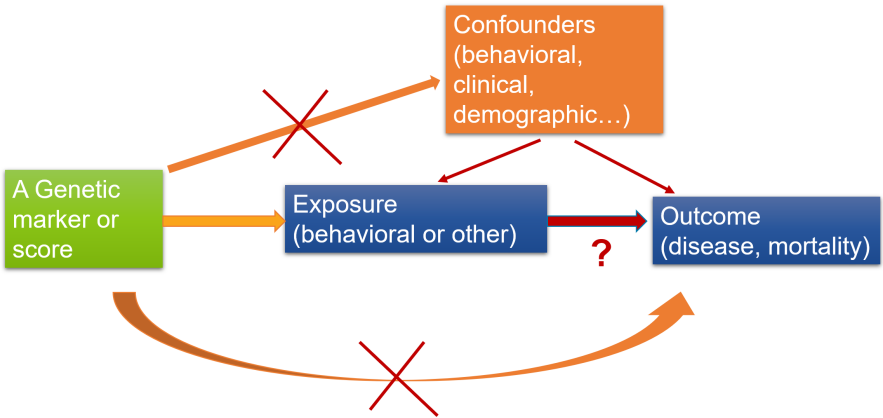
1. Z has a direct causal effect on X
2. Z does not have any direct or indirect causal effect on Y or the confounders U.

- It can be shown that the causal effect of X on Y equals:

$$\beta = \frac{cov(Z, Y)}{cov(Z, X)} = \frac{\beta_{ZY}}{\beta_{ZX}},$$

  where β_{ZY} and β_{ZX} are the coefficients of Z in a simple linear regression models for Y and X (with covariate Z).
- Replacing β_{ZY} and β_{ZX} by their estimates, we get the instrumental variables (IV) estimate of β.

Example
Mendelian randomisation



Summary

- There is no unique definition of “the causal effect”
- The validity of any causal effect estimates depends on the validity of the underlying assumptions.
- Adjustment for other available variables may remove (some) confounding, but it may also create more confounding. Do not adjust for variables that may themselves be affected by the outcome.
- Instrumental variables approaches can be helpful, but beware of assumptions!

Some references

- A webpage and a free online book by Miguel Hernan and Jamie Robins: http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- Judea Pearl, “The Book of Why”