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, Tartu 2023

Statistical associations vs causal effects in epidemiology

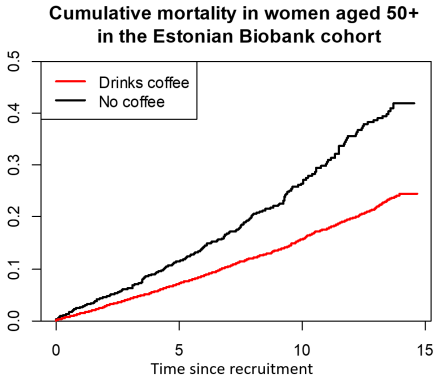
Does the exposure (smoking level, obesity, etc) have a causal effect on the outcome (blood pressure, cancer diagnosis, mortality, etc)?

is not the same question as

Is the exposure associated with the outcome?

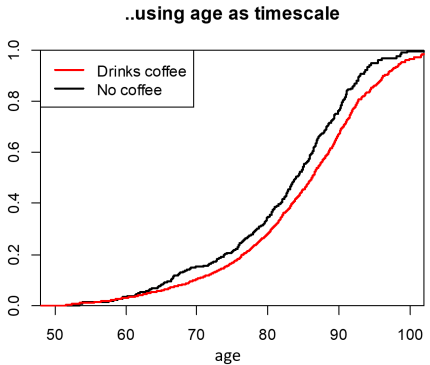
Conventional statistical analysis will answer the second one, but not necessarily the first.

Example



Does coffee-drinking prolong life?
(so drastically???)

Example (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

Causal effects and counterfactuals

- To define causal effects more properly, counterfactual (what-if) thinking is useful.
- Mathematically, the individual causal effect can be defined as the difference

$$Y^1 - Y^0,$$

  where $Y^1 = Y(X = 1)$ and $Y^0 = Y(X = 0)$ are defined as individual's potential (counterfactual) outcomes if this individual's exposure level X were set to 1 or 0, respectively.
- 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;
- For a particular individual, either Y^1 or Y^0 can be observed at any moment.

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 and Y–a disease outcome, one could formalize the effect of intervention on outcome by using $Y^{X,Z(X)}$

A causal model in terms of potential outcomes

- More generally Y^x is defined as the potential outcome following the exposure level $X = x$
- A linear causal model can be specified as

$$Y_i^x - Y_i^0 = x\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|x) = 0$$

- Note that the observed outcome $Y_i = Y_i^x$ for individuals with $X_i = x$.
- The model could be generalized to include nonlinear terms or interactions with other covariates, or as a generalized linear model (logistic regression, survival model).
- However, as we don't observe Y^0 and Y^x (with $x > 0$) for the same individuals at the same time, thus it is not straightforward to actually fit the model on data.

Statistical model vs causal model

- More generally Y^x is defined as the potential outcome following the exposure level $X = x$
- A linear causal model can be specified as

$$Y_i^x - Y_i^0 = x\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|x) = 0$$

- Note that the observed outcome $Y_i = Y_i^x$ for individuals with $X_i = x$.
- A classical linear regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|X_i) = 0$$

or

$$E(Y_i|X_i) = \beta_0 + X_i\beta_1.$$

- When are the two equivalent?

Statistical model vs causal model

- Rewrite the linear causal model as

$$Y_i^x = Y_i^0 + x\beta_1 + \varepsilon_i, \quad \text{with } E(\varepsilon_i|x) = 0$$

- Note that this would be equivalent with the classical linear model, if

$$E(Y_i^0 + \varepsilon_i|X_i) = \beta_0,$$

thus when the potential exposure-free outcome Y^0 is not associated with the exposure X

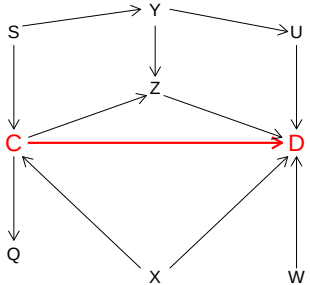
- For instance, this would mean that in the absence of smoking, the cancer risk for current smokers and current nonsmokers would be the same ($E(Y|X = 0) = E(Y^0)$).
- In other words, the two models are equivalent in the absence of confounding.

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

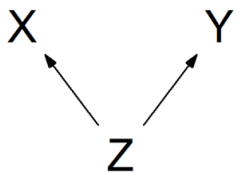
DAGs: directed acyclic graphs

- 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);
- Thus an arrow pointing from variable X to a variable Y on such graph represents a causal effect of X on Y.



“Classical” confounding

Third factors Z influence both, X and Y

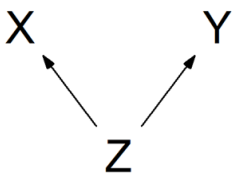


Also called as backdoor path between X and Y.
Implied statistical associations (Y is not independent of X in general, but it is independent of X, conditional on Z):

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

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

"Classical" confounding, mathematically



Assume:

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

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

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

If $b_{zx} \neq 0$, then also $r_{zx} \neq 0$ and so

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

. We see that:

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

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.

Example: COVID vaccination and Simpson's paradox

Suppose there are COVID infections in:

- 3000 unvaccinated individuals, 90 needing hospitalizations
- 1000 vaccinated individuals, 30 needing hospitalizations

No effect of vaccination?

More detailed data:

age	vaccination	total	hospitalized	% hospitalized
≥ 60	no	100	24	24%
	yes	300	24	8%
< 60	no	2900	66	2.3%
	yes	700	6	0.9%
all ages	no	3000	90	3%
	yes	1000	30	3%

Age is a confounder here!

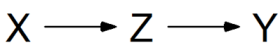
COVID vaccination and Simpson's paradox

Real data from Estonia (August 2021):

age	vaccination	total	hospitalized	% hospitalized
≥ 60	no	186	50	26.9%
	yes	202	16	7.9%
< 60	no	3075	57	1.9%
	yes	666	5	0.8%
all ages	no	3261	107	3.3%
	yes	868	21	2.4%

Causal chain (mediation, front-door path):

The effect of X on Y is mediated by Z:

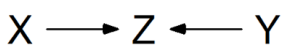


$$Y = \beta_0 + \beta_{xy}X + \beta_{zy}Z + \varepsilon,$$

- 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!

The case of a collider: adjustment is sometimes wrong!

X and Y have an effect on Z:



$Z = \beta_0 + \beta_{xz} X + \beta_{yz} Y + \varepsilon$, 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:
$Y = \beta_0^* + \beta_{xy} X + \beta_{zy} Z + \varepsilon^*$.

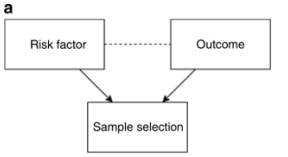
$$X \perp\!\!\!\perp Y \qquad X \not\perp\!\!\!\perp Y | Z$$

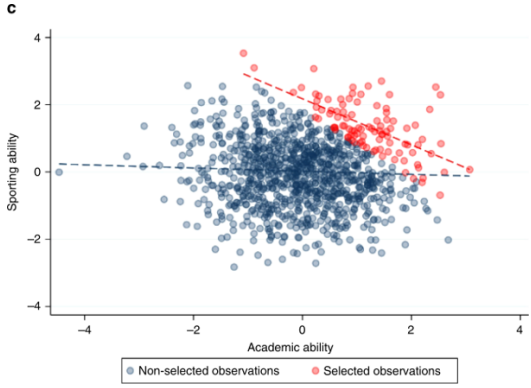
We see the association between X and Y only when the "effect" of Z has been taken into account.
But this is NOT a causal effect of X on Y.

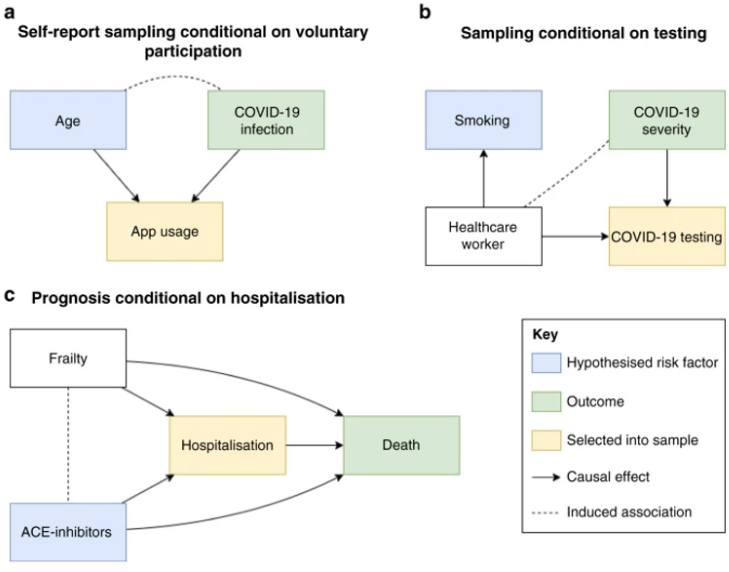
One should NOT adjust the analysis for Z!

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

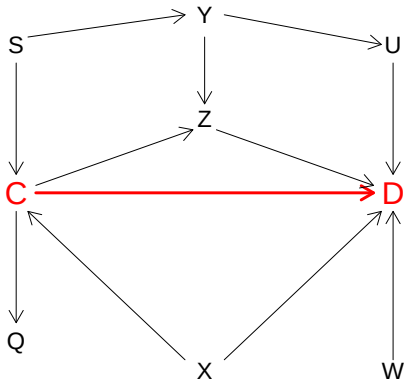
- All analysis are done conditional on the selected sample
- However, selection itself might be a collider (Griffith et al. 2020, https://www.nature.com/articles/s41467-020-19478-2)






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

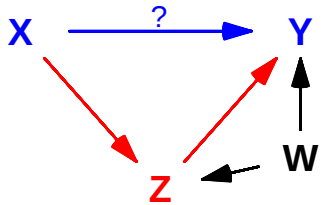
Q, S, U, W, X, Y, Z - other factors that influence cancer risks and/or smoking (genes, social background, nutrition, environment, personality, . . .)

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

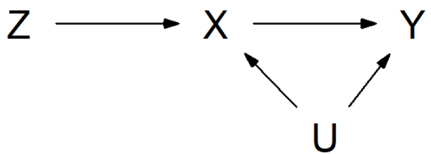


Paths: $X \to Z \to Y$ (open) and $X \to Z \leftarrow W \to Y$ (closed).

- 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:



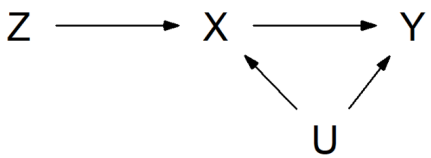
Assuming:
$Y = \alpha_y + \beta X + \gamma U + \epsilon,\ \mathrm{E}(\epsilon|X, U) = 0,$

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:

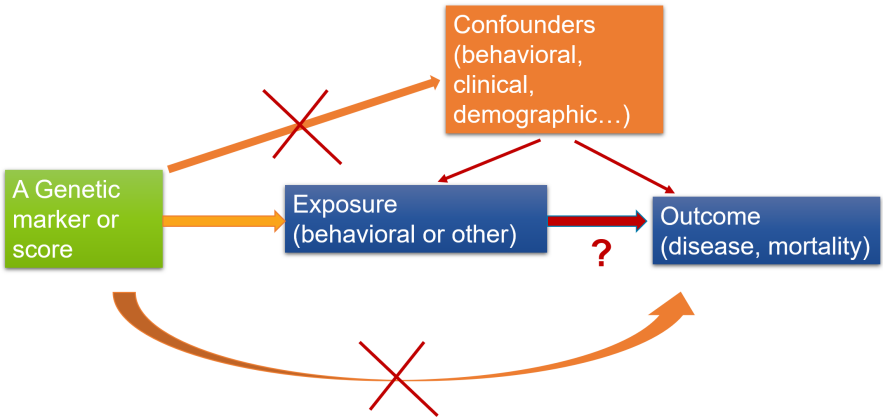
1. Z has a direct causal effect on X
2. Z does not have any direct or indirect causal effect on Y or the confounders U.

- 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”
- Mendelian randomization: Sheehan, N., Didelez, V., et al., Mendelian Randomization and Causal Inference in Observational Epidemiology, PLoS Med. 2008; papers by G.D. Smith, J. Bowden, S. Burgess and others.