Causal inference

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Definition of epidemiology

*Epidemiology is the science that studies the patterns, **causes, and effects** of health and disease conditions in defined populations.*

Wikipedia, 2017

Causality in epidemiology

- Does smoking cause lung cancer?
- Does red wine protect against cardiovascular disease?
- Does ADHD medication prevent traffic accidents?

Causality in other fields

- How much of recent climate changes are due to human greenhouse gas emission?
- Can we reduce criminality in society by employing more police and/or punish convicted criminals harder?
- Why have extreme right-wing parties recently gained popularity in many European countries?

Conclusion

- Most scientific research questions are about cause and effect
- In this sense, most research is ‘causal inference’

Definition of 'causal inference'

- A methodological branch of statistics, which aims to
  - establish a formal (mathematical) **language** for causal reasoning - done
  - use this language to develop appropriate **statistical methods** for making causal inference - ongoing

The language

- Potential outcomes
  - an algebraic tool to define causal parameters
- Direct Acyclic Graphs (DAGs)
  - a visual tool to derive appropriate analysis for estimating causal parameters

The statistical methods

- Instrumental variables
- Mediation analysis
- Interaction analysis
- Propensity scores
- Inverse probability weighting
- Marginal structural models
- Structural nested models
- ... and many others!

Outline

Outline

Due to ...

Donald Rubin (Harvard University)

James Robins (Harvard School of Public Health)

Example

- Research question: does smoking during pregnancy (X) cause malformations in the offspring (Y)?
- Data:

id	X	Y
1	1	0
2	1	1
3	0	1

- *Is there a statistical association between smoking and malformations?*

Solution

id	X	Y
1	1	0
2	1	1
3	0	1

- Malformations in offspring are more common among non-smokers than among smokers
- An inverse association!

$$RR = \frac{p(Y = 1|X = 1)}{p(Y = 1|X = 0)} = \frac{1/2}{1/1} = 0.5$$

Example

id	X	Y
1	1	0
2	1	1
3	0	1

- *Sampling variability aside, can we say that smoking protects against malformations?*

Solution

id	X	Y
1	1	0
2	1	1
3	0	1

- No!
- The smokers may be systematically different than the non-smokers
  - e.g. younger, more physically active, healthier diet etc
- 'Confounding'

What is the target parameter?

- Clearly, the associational risk ratio

$$RR = \frac{p(Y = 1|X = 1)}{p(Y = 1|X = 0)}$$

  is not the causal target parameter
- In fact, 'standard' statistical language cannot be used to define causal parameters
- **Without a proper definition of the target parameter, we can't be sure that we use an appropriate analysis**

Towards a causal target parameter

- The associational risk ratio

$$RR = \frac{p(Y = 1|X = 1)}{p(Y = 1|X = 0)}$$

  compares 'apples with pears'
  - the people in the numerator (smokers) are not the same people as those in the denominator (non-smokers)
- To avoid systematic differences, a causal parameter must compare 'apples with apples'
  - same people in numerator and denominator

Potential outcomes

- We think of each subject as having two **potential outcomes**
  - Y_0 = the outcome if the subject would hypothetically be unexposed (e.g. would not smoke)
  - Y_1 = the outcome if the subject would hypothetically be exposed (e.g. would smoke)

id	Y_0	Y_1
1	0	0
2	0	1
3	1	1

The causal risk ratio

id	Y_0	Y_1
1	0	0
2	0	1
3	1	1

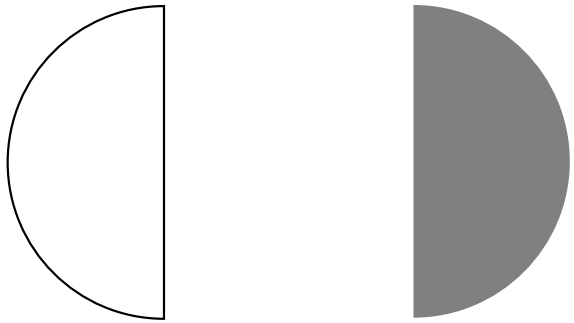
- We define the causal risk ratio as a comparison of two hypothetical scenarios
  - everybody unexposed, vs
  - everybody exposed

$$CRR = \frac{p(Y_1 = 1)}{p(Y_0 = 1)} = \frac{2/3}{1/3} = 2$$

Association vs causation

- Association:

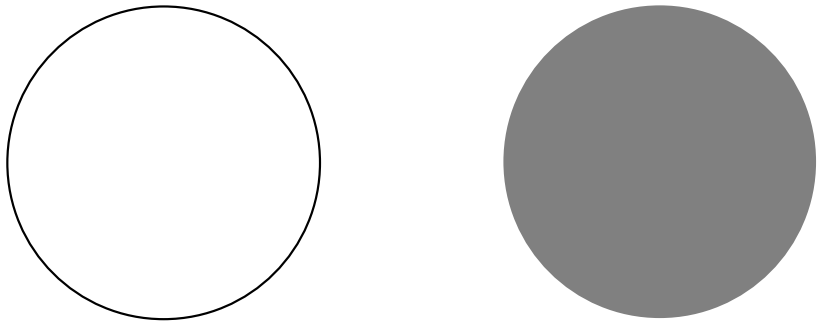
  Factually unexposed Factually exposed



$$p(Y = 1|X = 0) \text{ vs } p(Y = 1|X = 1)$$

- Causation:

  Everybody unexposed Everybody exposed



$$p(Y_0 = 1) \text{ vs } p(Y_1 = 1)$$

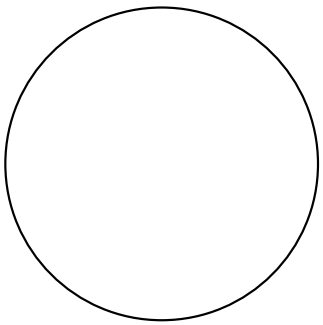
Ideal vs real data

- Ideally, we could observe both potential outcomes for any given subject
- In reality, we can only observe one of them - the one that corresponds to the factual exposure level for that subject
- The other is unobserved - or **counterfactual**

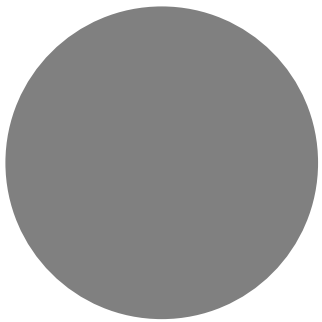
id	X	Y	Y_0	Y_1
1	1	0	? (0)	0
2	1	1	? (0)	1
3	0	1	1	? (1)

Want to do this...

Everybody unexposed



Everybody exposed

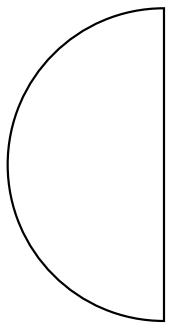


$$p(Y_0 = 1) \text{ vs } p(Y_1 = 1)$$

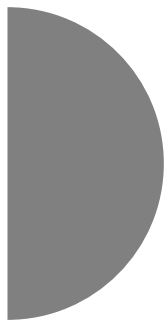
- No systematic differences

...but can only do this

Factually unexposed



Factually exposed



$$p(Y = 1|X = 0) \text{ vs } p(Y = 1|X = 1)$$

- Systematic differences

Solution

- Try to eliminate systematic differences between exposed and unexposed, so that association = causation
- By design: randomization
- By analysis: confounding control

Randomization

- Assign exposure levels by the flip of a coin
- Removes **all** systematic differences between exposed and unexposed: association = causation!
- Practical problems:
  - unethical
  - expensive
  - difficult

Confounding control

- Control for measured confounders in the statistical analysis
  - stratification
  - matching
  - regression modelling
  - propensity scores
  - inverse probability weighting
  - etc etc etc
- Only removes systematic differences due to confounders that we explicitly control for
- Systematic differences may remain, due to unmeasured confounders: association = causation?

What to control for?

- Often, we have measured a large set of variables, which we could potentially control for in the analysis
  - e.g. register-based research
- Which of these should we control for?
- Are there any variables we should **not** control for?
- Enter DAGs!

But really, what has been gained?

- We may define the causal effect, using potential outcomes, as

$$CRR = \frac{p(Y_1 = 1)}{p(Y_0 = 1)}$$

- But all we can ever observe is a statistical association

$$RR = \frac{p(Y = 1|X = 1)}{p(Y = 1|X = 0)}$$

- Even if potential outcomes may add conceptual clarity, one may question if they have any practical value

More complex scenarios

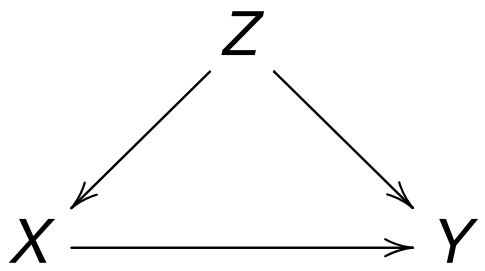
- Potential outcomes have proven extremely useful in more complex scenarios
  - Instrumental variable studies
  - Studies of mediation and interaction
  - Longitudinal studies with time-varying exposures and confounders
- In these scenarios, there is not one, but several possible causal target parameters
- **Without a proper definition of the target parameter, we can't be sure that we use an appropriate analysis**
- Largely overlooked in 'standard' statistical literature, not using potential outcomes

Outline

Due to ...

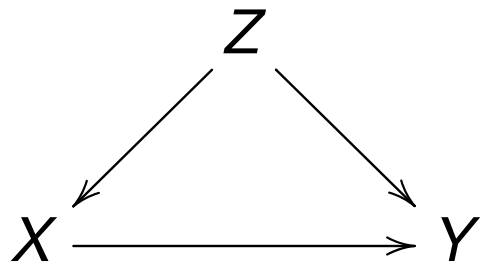
Judea Pearl (UCLA)

A simple DAG



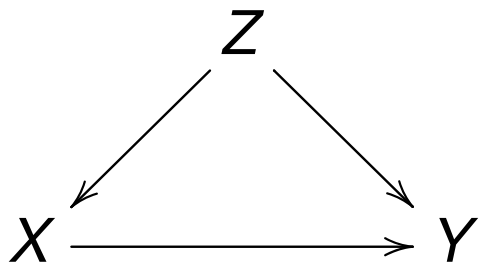
- Each arrow represents a causal effect
- The graph is
  - Directed, since each connection between two variables consists of an arrow
  - Acyclic, since the graph contains no directed cycles
- Formal connection to potential outcomes through non-parametric structural equations
  - beyond this seminar

Causal and non-causal paths



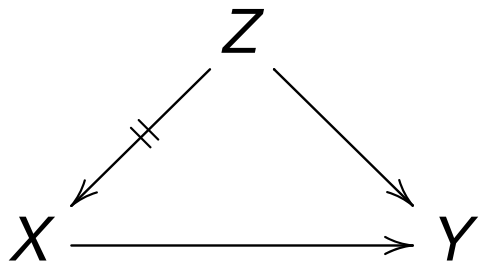
- There are two paths between X and Y:
  - $X \rightarrow Y$
  - $X \leftarrow Z \rightarrow Y$
- Only the first path is causal
  - if we remove the arrow from X to Y, then X has no causal effect on Y

Confounding in DAGs



- The variable Z is a common cause of the exposure X and the outcome Y - a confounder
- The non-causal path $X \leftarrow Z \rightarrow Y$ induces a statistical association between X and Y
  - even in the absence of the causal effect $X \rightarrow Y$

Randomization in DAGs

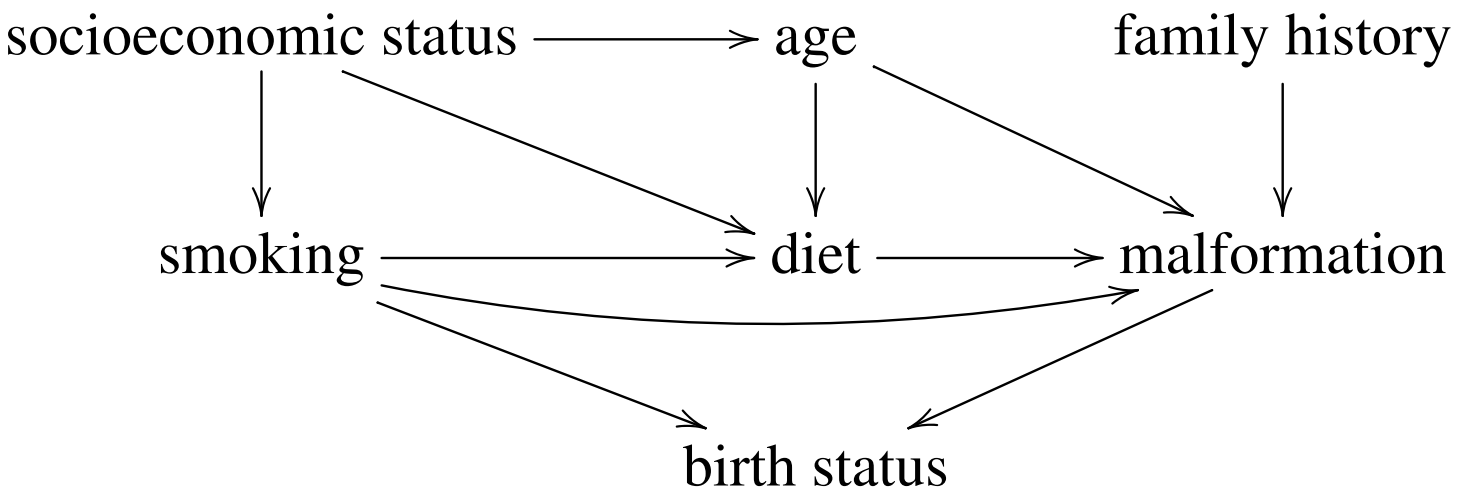


- Randomization breaks the influence of Z on X
- Thus, the non-causal path $X \leftarrow Z \rightarrow Y$ no longer exists
  - ... and no other non-causal paths either
- Association = causation

DAGs can be used for confounder selection

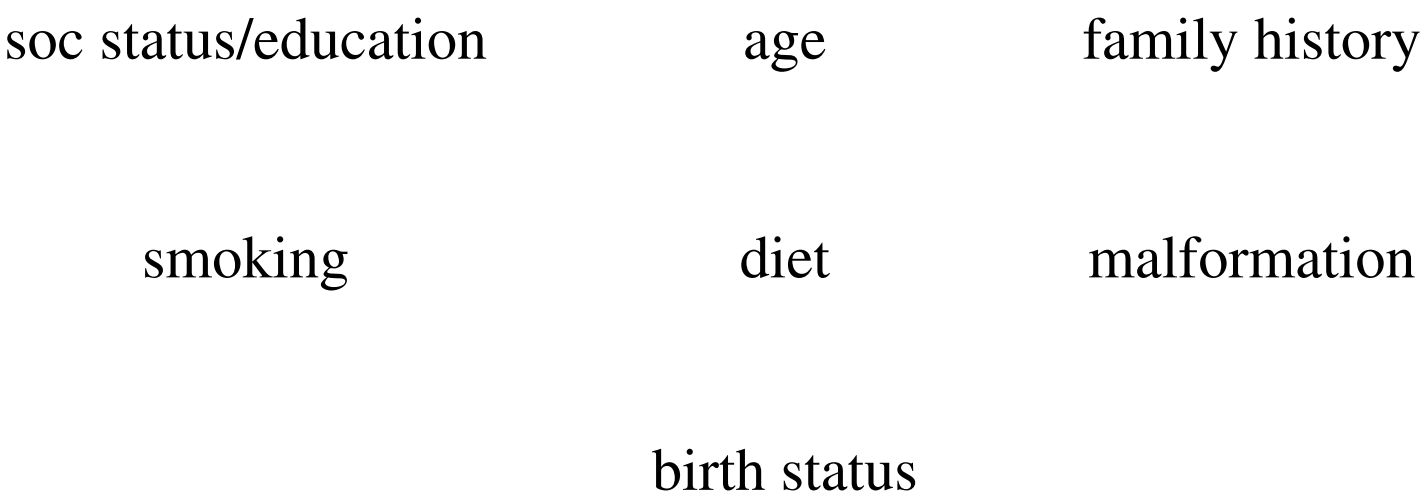
- 1. Use subject matter knowledge to draw the DAG (by no means trivial!)
- 2. Use simple graphical rules to determine what to control for
  - attempt to 'block' non-causal paths between the exposure and the outcome
  - if all non-causal paths are blocked, then association = causation

Example


socioeconomic status
age
family history
smoking
diet
malformation
birth status

No *a priori* knowledge

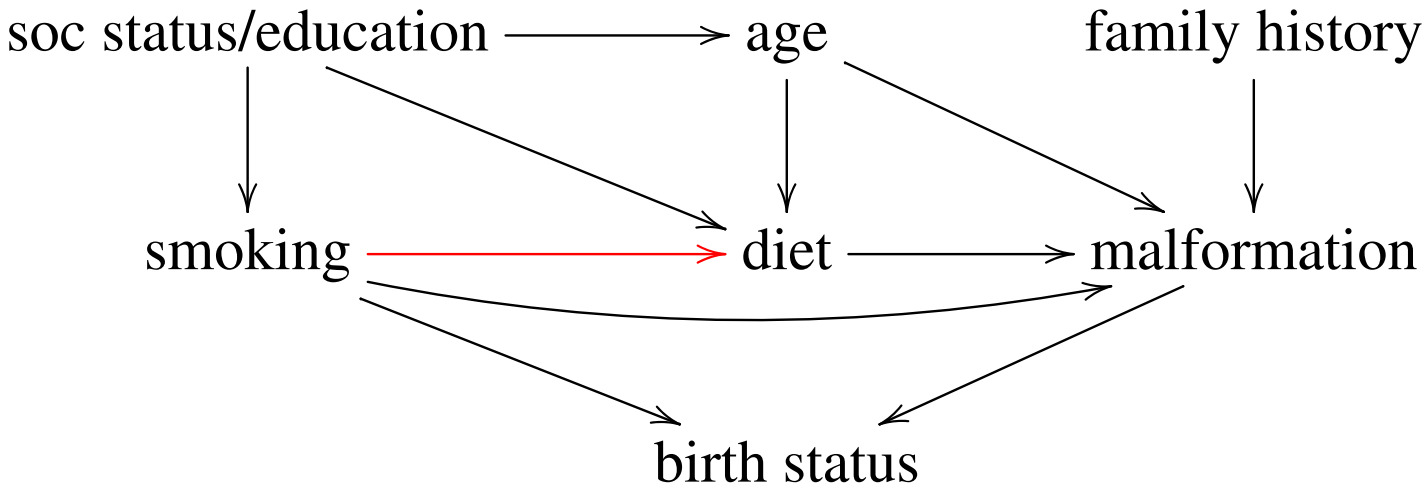
- Cannot construct a plausible DAG

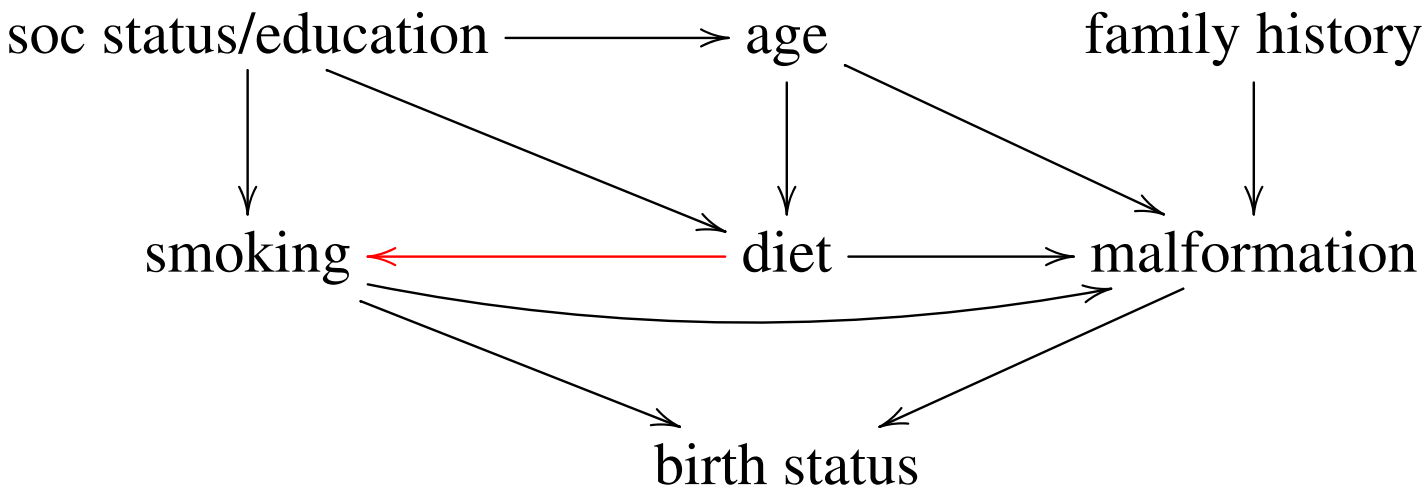


- Ok... but are you really the right person to do this study?

Weak *a priori* knowledge

- Cannot settle with **one** plausible DAG





- Present all plausible DAGs, and the implied analyses

Summary

- Causal inference has been an intense research field the last $\sim$30 years
- It has generated many new methods and countless papers
- Much of this success can be attributed to the development of a formal causal language
  - enables proper definitions of causal parameters
  - can be used to derive appropriate analyses for estimating causal parameters
- The key elements in this language are potential outcomes and DAGs

Read more

- Pearl, J (2009). *Causality*. Cambridge University Press (2nd edition).
- Judea Pearl's home page (search for 'introduction')
- Hernan MA, Robins JM (2018). *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.