Innføring i DAGs, kausalitet og kausal inferens

HEL-8020 Analyse av registerdata i forskning
Universitetet i Tromsø
24. april 2017
Odd O. Aalen
Avdeling for biostatistikk, IMB,
Universitetet i Oslo

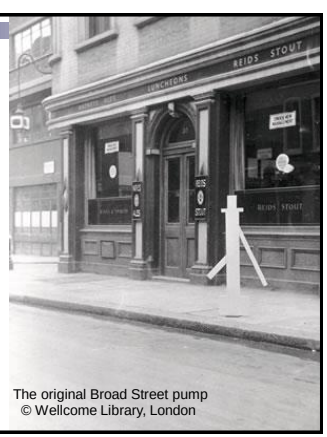
Plan

- Hvilken rolle spiller statistikken i å etablere kausalitet
- Metoder: grafiske og kontrafaktiske
- Kausal inferens: Marginal structural model
- Medieringsanalyse (Mediation)
- Liten smakebit fra et stort og voksende felt

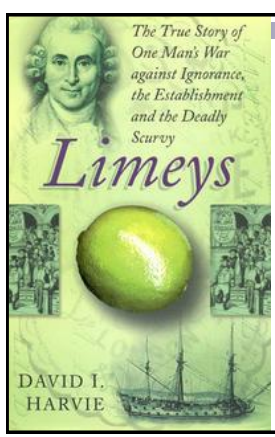
New England Journal of Medicine, Editorial, Jan. 6, 2000, p. 42-49

- *The eleven most important developments in medicine in the past millennium*
  - Elucidation of human anatomy and physiology
  - Discovery of cells and their substructures
  - Elucidation of the chemistry of life
  - *Application of statistics to medicine*
  - Development of anesthesia
  - Discovery of the relation of microbes to disease
  - Elucidation of inheritance and genetics
  - Knowledge of the immune system
  - Development of body imaging
  - Discovery of antimicrobial agents
  - Development of molecular pharmacotherapy

- From NEJM: The origin of modern epidemiology:
- 1854, when John Snow demonstrated the transmission of cholera from contaminated water
- The majority of people who got ill used the Broad Street Pump in London's Golden Square
- He removed the pump handle from the polluted well and the spread of the disease stopped.



The original Broad Street pump
© Wellcome Library, London

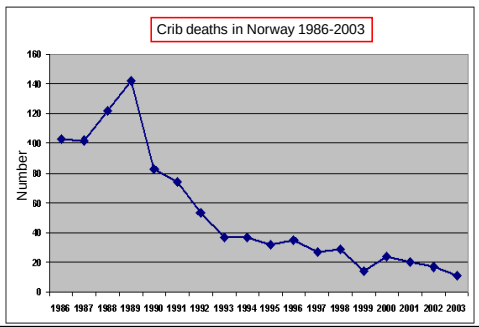


- From NEJM:
- Earliest clinical trial in 1747
- Scurvy (Serious disease: Magellan lost 80% of his men from scurvy)
- James Lind treated 12 scorbutic ship passengers on a British navy ship with cider, an elixir of vitriol, vinegar, sea water, oranges and lemon
- Those who got oranges and lemon did not get ill
- Supply of lemon juice eliminated scurvy from the navy

So, it is all about causality

- Statistics is important because it is *conceived as* contributing to a causal understanding which is needed in prevention and treatment of disease.
- Statistics can indicate causality even in the absence of a mechanistic understanding.
  - Treatment of scurvy far ahead of the knowledge of vitamin C
  - John Snow: 20 years ahead of Pasteur
- Going to modern times next: Causality and statistics – a happy couple?

Modern breakthrough based on statistics:
Sleeping position influences risk of crib death

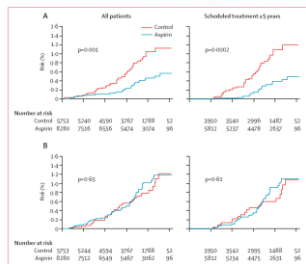


Sudden infant death syndrome (SIDS)

- The risk of SIDS is strongly increased (RR up to 13) when the infant is sleeping on its stomach compared to sleeping on its back.
- This is simple because
  - An intervention could be conceived and was easy to carry out in practice
  - The effect was immediate
  - The effect was very strong
- None of these conditions normally hold in epidemiology

Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials

Peter M Rothwell, Michelle Wilson, Carl-Eric Elwin, Bo Norrving, Ale Algra, Charles P Warlow, Tom W Meade



Lancet 2010; 376: 1741–50

Proximal colon

Distal colon

Figure 3: Pooled analysis of the effect of aspirin (75–1200 mg) versus control on incidence of colorectal cancer
Pooled analysis is by cancer site: proximal colon (A), distal colon (B), unspecified colon site (C), and rectum (D).
Statistical significance is given by the log-rank test.

Mechanistic understanding vs statistical documentation

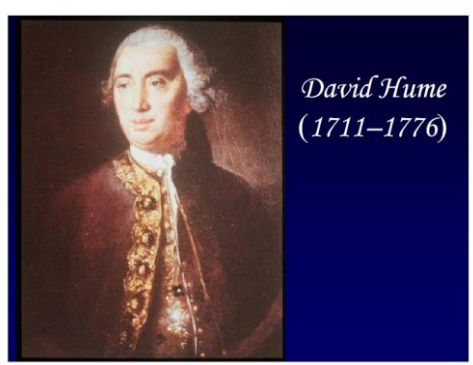
- Often an effect would be expected on the basis of mechanistic understanding, but does not show up in statistical studies
  - According to mechanistic understanding intake of antioxidants should be good for you. It prevents oxidative stress that might be damaging.
  - However, statistical studies show very little effect of antioxidants either in food or supplements (and existing effects are often negative)

What is causality

- Aristoteles – four concepts, we mention two:
- *Causa efficiens:* the efficient cause that produces change. This is the modern concept of science.
  - The father is the efficient cause of the child
- *Causa finalis:* the purpose, (formålet). Causa finalis is not an accepted idea of modern natural science.
  - The purpose (final cause) of taking walks is to improve your health

Aristoteles (Rafael) from Wikipedia

Taken from Pearl



Philosophical aspects of causality

- First discussed seriously by Hume.
- Stressing empirical view of causality: Causality is the "constant conjunction" between events
  - E.g. water "causes" fire to be extinguished
- Hume was strongly opposed to a mechanistic understanding of causality
- Hume 1748: "We may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words, *where, if the first object had not been, the second never had existed*."
- Second part points to *counterfactual* causality

13

Immanuel Kant's view of causality

- Hume inspired Kant
  - Causality is a *category* for experiencing reality, just like time and space
  - But: "Das ding an sich" is unknown!
  - Many major philosophers have thought that we cannot experience true reality

1724-1804 (Wikipedia)

14

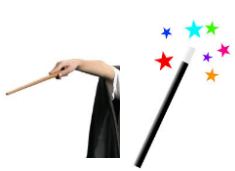
Questions

- Why is causality important in medicine?
- How can statistics say something about causality?
- Why have philosophers struggled with the causality concept?

15

Causal inference

- No magic wand:



- But, a way of thinking:

16

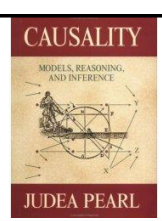
Going to modern times: Statistical approaches to causality

- Directed acyclic graphs – DAGs
  - Mathematical definition of causality.
  - Causality is defined by intervention
  - Notice the importance of interventions in the historical examples
  - Developed by Pearl and others. Inspired by Trygve Haavelmo (Norwegian recipient of Nobel prize in economics in 1989)
- Counterfactual causality
  - Distinguishing actual and counterfactual world.
  - Developed mainly at Harvard university by Robins, Rubin, Hernán and others. Nobel prize in economics (Heckman, 2000)
- Granger causality
  - Based on prediction. Developed in econometrics. Increasingly used in neuroscience
  - Nobel prizes to Granger (2003), Sims, Sargent (2011)

17

First approach: Causal DAGs

DAG: Directed acyclic graph



- Doing versus seeing: Pearl's do-calculus has become a major tool in epidemiology and other fields. Extensive mathematical theory for calculating causal effects.
- *Do*-operator:

  P(y | do(x), z)

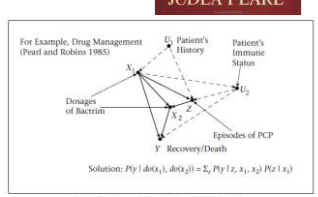


Figure 15. Learning to Act by Watching Other Actors (Treatment Management).

Seeing vs doing (Pearl)

- Pearl makes a fundamental distinction between seeing and doing. Causality is about doing, while most statistical data is about seeing
- Seeing and doing may coincide in experiments because of the ability to control the setting. The "big" experiment in medicine is the randomized clinical trial where the effect of doing is apparent

19

Registries contain data on seeing only, and not doing

- We want to say something about the effect of intervention. BUT: the registry only contains a description of what has happened, there is no information about what could have happened if one acted differently. Therefore, you can't (directly) say anything about the effect of intervention
- This is the case for observational data in general
- Still, causal inference can help us if we collect enough data and the right type of data…

20

Intervention is fundamental in biostatistics

- The final aim of medical research is to intervene (either to treat or to prevent disease).
- When reading papers or listening to talks in medical research you should look for the interventions lurking behind.
- Important question: Most data are just "seeing". Can we deduce «doing» from «seeing»?
  - Well, sometimes by careful analysis

21

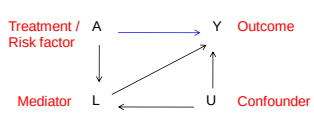
Graphical models

- Graphical models with arrows and boxes are common. However, Judea Pearl has lifted them to a new level
- A number of rules for evaluating graphs can be defined
- These are applicable in practice as shall be demonstrated

22

Directed acyclic graph – DAG

- Graph with arrows, where you never return to the same node

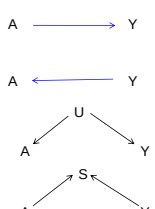


Collider: where two or more arrows meet

23

Statistical association

- If A and Y are associated, then this is compatible with four different types of causal relationship:
- Direct causation
- Reverse causation
- Confounding
- Collider effect (selection)



Examples:

Smoking/ Lung cancer

Heart disease / cancer

24

The following rules decide whether a path is open or closed

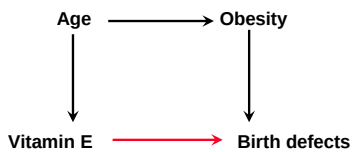
1. A path with colliding arrows is closed (→←). If there are no colliders the path is open.
2. To conditon on a non-collider closes the path.
3. To condition on a collider (or descendant of a collider) opens the path

What do we mean by "to condition on"

- We mean e.g. to include a variable in the regression.
  - To include a confounder is usually ok
  - To include a collider is dangerous
- However, a collider may not be avoided if it represents inherent selection in the data

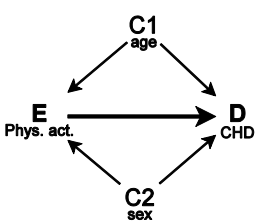
Keep causal paths open and non-causal paths closed

- Example (Hein Stigum): red arrow is causal, black path is not causal (backdoor path). Conditioning on age (or obesity) blocks the back-door path



Exercise (Hein Stigum)

- We want the causal effect of physical activity on CHD (coronary heart disease). What should we adjust for?



Birth defects. Adjustment for confounder?

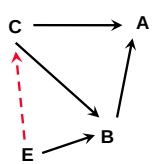
Source: Hernán et al, Amer. J. Epidem. 2002, **155**, 176-184

FIGURE 1. Low folate intake (*E*) may increase the risk of preterm delivery and infant low birth weight (*C*) (Am J Clin Nutr 2000;71(suppl):1295s–303s), and many birth defects (*D*) result in preterm deliveries and low birth weight infants (Am J Dis Child 1991;145:1313–18).

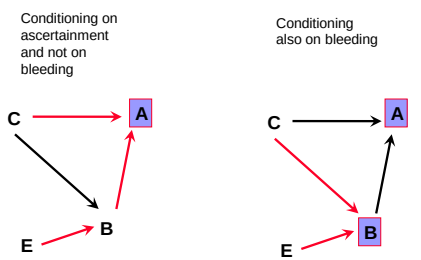
- When estimating the effect of E on D, shall you adjust for C?
- *No, one should not adjust for a collider.*
- Case-control study on folic acid supplementation and neural tube defects. Adjusted OR: 0.80 (0.62, 1.21), non-adjusted OR 0.65 (0.46,0.94)

Oestrogen and endometrial cancer

- Partly from Robins (2001).
- Does oestrogen supplements increase the risk of endometrial cancer?
- Use case control study
- Assume no effect of oestrogen on cancer
- Will statistical analysis still show an effect of oestrogen on cancer? That is: Will there be bias?



C – cancer
A – ascertainment
B – bleeding
E – oestrogen



31

Questions

- What is a causal path?
- When is a path non-causal?
- How do we close a non-causal path with a confounder?
- Should we adjust for a collider?

32

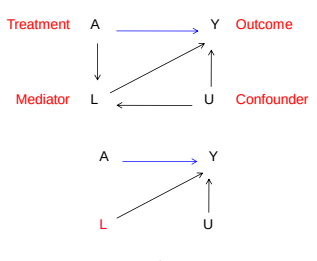
When is a DAG causal?
Two views

- Robins and Hernán: A DAG is causal when
  1. Lack of an arrow can be interpreted as lack of direct causal effect
  2. All common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph
  - Note: this requires a concept of direct cause
- Pearl: A DAG becomes causal if intervening on a node has the effect of removing all arrows into the node while the DAG is otherwise unchanged

33

Directed acyclic graph – DAG

- Graph with arrows, where you never return to the same node
- Intervention: all incoming arrows are removed



34

DAGs are useful

- DAGs are a useful way of formulating prior causal ideas and judging their consequences
- A warning: Causal ideas are usually rather vague and may not easily match the precision of the mathematical analysis of DAGs developed by Pearl and others.

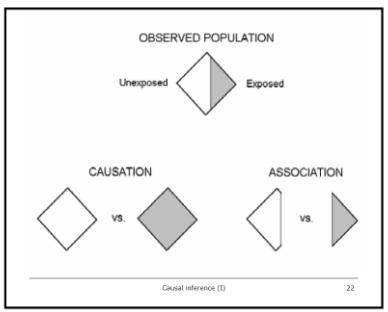
35

The second viewpoint:
Counterfactual worlds

- Increasing importance in epidemiology. (Rubin, Robins, Hernán at Harvard University)
- Example: Imagine one actual world where you do smoke and a counterfactual one where you don't and everything else is equal
- But you just observe one!
- The causal effect can be defined as the difference between the result in the actual and the counterfactual world.
- Normally this is not observable, but can be estimated from data given certain assumptions (like no unmeasured confounder)

36

Counterfactual causation vs. association (from Miguel Hernan)



How do we get from association to causation?

37

Basic problem of epidemiology

- To get from the observation a statistical association
- to a valid counterfactual statement

38

Defining causal effects

- Calculation on unobservable quantities (notice *unobservable*, not just *unobserved*) – Rubin, Robins
- First defining causal effects, and then seeing if they can be estimated (approximately)
- This cannot be covered here, but we shall look at an application

39

Questions

- What are the two definitions of causality that we focus on here?
- Why do we consider more than one version of the concept?

40

Randomised clinical trials

- The established solution to the confounder problem. We create both a factual and a counterfactual world
- One of the great pillars of medical research. An unrivalled source of reliable information. Thousands of clinical trials carried out every year.
- But limitations: very many exclusions (children e.g.), could be distant from clinical practice, extremely expensive (the development of a successful medication costs 1 billion dollars)
- Clinical trials become unethical once a secure effect has been established
- Increasingly data are collected in clinical registries, could they be used in addition? Or should all these data go to waste?

41

Can randomized trials be simulated from non-randomized data?

- Medical treatments: Large HIV cohorts in the US, UK and Switzerland have been used as a testing ground for new methodology. Harvard researchers at the forefront. We cooperate closely with the Swiss HIV cohort
  - The HIV cohorts are models for data registries that can be used for drawing causal conclusions. Data are collected at fixed times, and not only when clinical events occur. Model for quality registries?
- Epidemiology: Hernán et al, *Epidemiology 2008;19: 766–779,* analyzed the effect of Postmenopausal Hormone Therapy on Coronary Heart Disease. There has been a discrepancy between clinical trials and epidemiological studies. This disappears when the epidemiological studies are analyzed by mimicking the design of a randomized trial
- Conclusion on treatment effects from non-randomized studies may be feasible

42

Swiss HIV cohort data

- An ongoing multi-center research project following up HIV infected adults aged 16 or older.
- Data goes from 1996, when the highly active anti-retroviral treatment (HAART) became available in Switzerland, to September 2003. The data are organized in monthly intervals, with measures of CD4 count, viral load (HIV-1 RNA) and other blood values, together with variables describing sickness and treatment history.
- The end point of interest is AIDS or death
- 77 838 person-months of observation, 2161 individuals, observed over minimum 1 and maximum 92 months
- This dataset has already been analyzed using MSMs [Sterne *et al*. 2005]

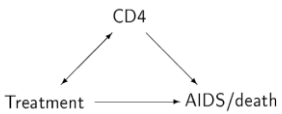
43

Time-dependent confounding

44

Time dependent confounding: HIV example

- When analyzing treatment effects on HIV, variables such as CD4 count (a measure of immune status) are time dependent confounders
- Such confounding could be present when a covariate, affected by past exposure, is both a predictor of the future exposure and the outcome



Notice the left arrow goes in both directions. This is not a DAG, but a local dependence graph. The feedback cannot be understood without considering time. Local dependence tells us how processes influence one another

45

Solutions to time-dependent confounding

- One solution is given by the *marginal structural model (MSM)* proposed by James M. Robins
  - The confounding is handled by inverse probability of treatment weighting (IPTW) and inverse probability of censoring weighting (IPCW)
  - Example: If there are fewer men than women in a study we can weight up the men to get a fair comparison
  - The MSM uses a sophisticated version of this
- An alternative, called sequential Cox regression is developed by Gran and coauthors.

46

The sequential Cox approach

Gran et al, 2009

- *Mimic a sequence of randomized clinical trials (RCTs)* based on each time period (month) of treatment start
- *Average over all* mimicked RCTs to find an overall effect estimated by composite likelihood
- Consider only individuals starting treatment in a certain time interval as the *treatment group.* Analysis start at the starting point of this interval
- Individuals still not on treatment in this interval serve as the *control group* – if they start treatment on a later stage they get censored (artificial censoring - Hernán)
- Dependent censoring? Solution: inverse probability of *censoring* weighting
- Adjust for covariates at baseline and at the start of the mimicked RCT (using for instance a Cox model)

47

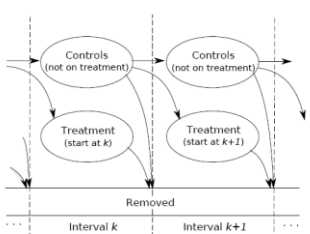


Figure 2: Illustration showing the movement between groups for individuals going from one sub analysis (in interval k) to another (in interval $k+1$). For each interval individuals already on treatment are removed, together with indivials being censored, dying or developing AIDS without ever starting treatment, while the individuals still not on treatment are compared with the individuals starting treatment in that interval.

48

Results from overall analysis

	HR	95% CI
Unweighted Cox model, baseline and time dependent covariates	0.647	0.430-0.973
Unweighted Cox model, baseline covariates only	0.334	0.232-0.483
MSM	0.140	0.066-0.299
Sequential Cox	0.176	0.105-0.296
Censor weighted sequential Cox	0.165	0.079-0.343

Estimated hazard ratio of HAART vs. no treatment. The three first rows correspond to results from Sterne et al (2005)

What is the issue here?

http://dallaslouis.net/?p=609

- Can we analyze treatment effects from non-randomized data?
  - Yes, we can (*if we have the right data*)
- Time-dependent confounding will be an issue
- This will be increasingly important when data from hospitals and medical practices become more available
- It is called comparative effectiveness research
- It is likely to be one of the major statistical challenges in medical research

Assumptions

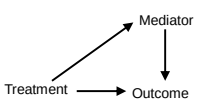
- Positivity
- No unmeasured confounders
- Design aspects: systematic follow-up
  - Must not only register those with events leading to contact with the hospital. Must know the development for the others as well

Questions

- What is the difference between time-dependent confounding and ordinary confounding?
- Why is time-dependent confounding more difficult?
- Which are the two approaches we mentioned for analyzing it?

Mediation

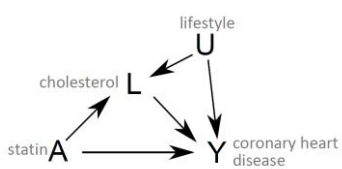
- Can we understand mechanisms by using statistics?
- Path analysis (Wright, 1921) introduced the idea of direct, indirect and total effects and presented a simple calculus for these effects based on linear regression models.
- A lot of recent and sophisticated development of these ideas in the causal inference literature



How much of the effect passes directly from treatment to outcome (direct effect) and how much passes through the mediator (indirect effect)?

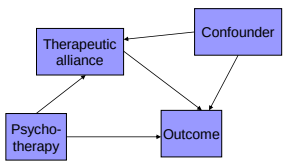
Mediation: Cholesterol treatment

Mediation and confounding. Can we estimate the direct effect of statin on coronary heart disease? Confounders between mediator an outcome may give a false impression of an increased indirect effect



Example

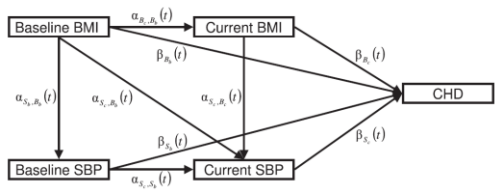
- Psychotherapy: It is well documented that therapeutic alliance appears to be a mediator. However, there is an obvious possibility of confounding effects
- Which type of effects could that be?



Dynamic path analysis for survival data

- First, assume a basic *causal structure* between the variables
- Carry out a set of *linear* (or additive) regression analyses for each node in the graph at each time point where an event occurs, conditioning on parents and baseline covariates
- Find the estimated direct and indirect effects by multiplying the estimated coefficients belonging to the arrows along each path
- Direct and indirect effects as *functions of time*
- Assume "no unmeasured confounders"

Direct and indirect effects for survival data

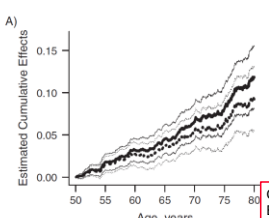


Dynamic Path Analysis in Life-Course Epidemiology, Michael Gamborg*, *Gorm Boje Jensen, Thorkild I. A. Sørensen, and Per Kragh Andersen*, **American Journal of Epidemiology,** DOI: 10.1093/aje/kwq502

Dynamic path analysis with time effects
(Generalizing standard path analysis)
Conclusion in paper:
Baseline BMI has a strong direct effect on CHD, and just very slight indirect effects via blood pressure

Direct and total effects of baseline BMI

- To which extent is the effect of baseline BMI on CHD mediated through later systolic blood pressure? Solid thick line indicates total effect, while thick dotted line shows the direct effect



Conclusion in paper:
Baseline BMI has a strong direct effect on CHD, and just very slight indirect effects via blood pressure

Questions

- Why are we interested in mediation?
- What kind of confounding might present a difficulty when analyzing mediation?

Summary

- Causal inference is a large and complex area
- It is no magic tool, but still has a lot of promise
- Causal inference more and more becomes the norm of analysis and presentation in an international epidemiological setting
- Whether we can do causal inference depends on how the data are collected. The HIV cohorts are a good example

Some references

- Lange, T. and Hansen, J. V, (2011). Direct and Indirect Effects in a Survival Context. *Epidemiology*, 22(4):575-581. The electronic supplement contains some information on software
- Røysland, K., Jon Michael Gran et al. (2011). Analyzing direct and indirect effects of treatment using dynamic path analysis applied to data from the Swiss HIV Cohort Study. *Statistics in Medicine*. DOI: 10.1002/sim.4324
- Pearl J. *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge University Press: Cambridge, 2009.
- Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**:360–372.
- Gran JM et al. A sequential Cox approach for estimating the causal effect of treatment in the presence of time dependent confounding. *Statistics in Medicine* 2010; **29**:2757–2768
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.