Causal inference

Jon Michael Gran

Oslo Centre for Biostatistics and Epidemology,
Department of Biostatistics, UiO

j.m.gran@medisin.uio.no

Tromsø, April 24th, 2019

UiO : Institute of Basic Medical Sciences
University of Oslo

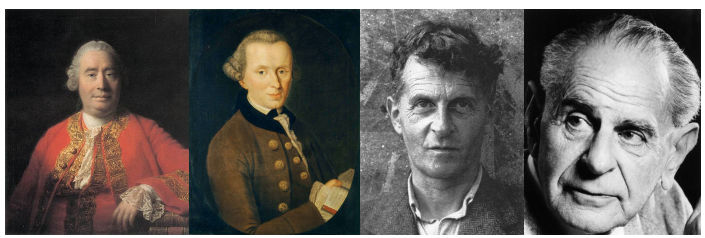
Outline

1 Introduction

Causality

- **Causality** is a topic of philosophy and metaphysics, intuitively about *cause and effect*, but hard to properly define – not logically possible to prove its existence (Hume, Kant 1700s) and even been dismissed as superstition (Wittgenstein 1900s)
- **Causal inference** on the other hand, if you are willing to do some assumptions, is about science and falsifiable hypothesis (Popper 1900s)

  For example; are the expected outcomes under treatment different than without treatment?



Interventions

- Want to assess **the effect of treatments**, or, more generally, any type of interventions
  - The effect on what outcomes?
  - Treatment compared to what?
  - For whom?
  - When treatment is given how?
  - What is the optimal treatment?
  - What is the mechanism behind?
- Causal inference can be seen as **a calculus for interventions**, built to formally address these questions

Traditional view on causal inference

- **Medicine**: *randomized controlled trials* (RCTs) as the gold standard
- **Epidemiology**: *informal* causal considerations, such as the Bradford Hill criterions
- **Statistics**: "association is *not* causation"

Prediction vs causal inference (exemplified by linear regression)

- **Prediction**:

$$\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

  - Collinearity between independent variables is unproblematic
  - Variable selection based on R^2 (stepwise regression, AIC etc)

- **Causal inference**:

$$Y = \beta_0 + \boldsymbol{\beta_1} X_1 + ... + \beta_k X_k$$

  - Collinearity between X_1 and other variables a major issue
  - Variable selection focus on removing bias and confounding; want "a causal model" (much more ambitious)

For example: NETFLIX versus 

History of causal inference

- Formal language for interventions in the context of randomized trials goes back to **Neyman (1923) and Fisher (1925)**; later introduced for observational studies by Rubin (1974)
- Roots in econometrics and structural equation modelling, going back to Nobel price winner **Trygve Haavelmo (1943)**
- The impact and change of paradigm caused by causal inference over the last decades is to a very large degree due the novel work of **Robins (1986) and Pearl (1986)**

Robin extended Rubins work and developed methods for time-dependent treatments and confounding, while Pearl formally connected (non-parametric) structural equation models with graphical models

Observational studies

- May have the same goal as RCTs, but differ in that the investigator **cannot control treatment assignment**
- In an ideal RCT association is causation because of the randomisation – in observational studies, generally, **association is not causation**
- If we want to mimick a RCT from observational data, we **need causal modelling**

Why analyse observational data?

- **RCT's can be difficult**: expensive, time consuming, unethical or to risky, e.g. in children or among pregnant women
- Even when RCT's exist, well-performed observational studies can **add to the overall evidence**; confirming or contradicting previous studies, allow for subgroup analyses, longer follow-up time, information on excluded patients, etc
- Most often **RCT's also have observational data in them** (post exposure measurements); can be used to understand or adjust for droput, non-complience, or the mechanisms behind how treatment works

Not about replacing RCT's, but **complimenting** them
(and analyzing them better)

Exchangeability

- In ideal RCTs, because individuals are randomly assigned to treatment groups, we say that the **groups are exchangeable**
- Exchangeability means that we **would have seen the same** effects if we exchanged the people in the two groups
- The corresponding term in the Rubin literature is **ignorability**

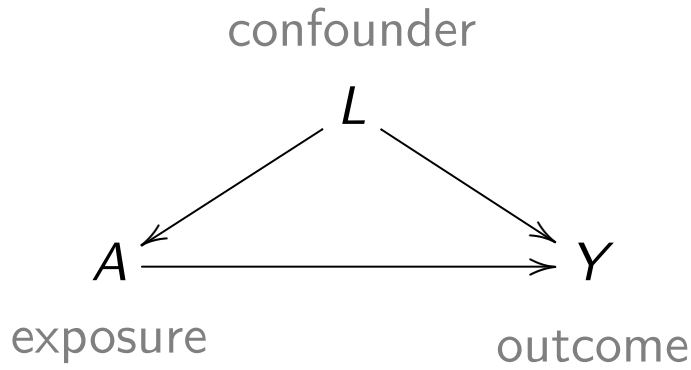
Conditional exchangeability

- Say that treatment assignment **depends on covariates**

  E.g.: Treatment is given to patients in critical condition with probability 0.75 and to patients in non-critical condition with probability 0.5
- This is a **conditional randomised trial**, as opposed to a *marginal* randomised trial – actually corresponds to *two* marginal RCTs; one in each patient group
- Treatment groups are subject to **conditional exchangeability**
- Treatment assignment can depend on one covariate or many – no conceptual difference between **observational studies** and conditional randomised trials

Confounding

- **Treatment and outcome share an common cause**, which can be represented grahically as



E.g.: Sicker people more likely to get treated and have adverse outcome

- Confounding in observational studies can be seen as a **lack of exchangeability**
- Conditional exchangeability (**no unmeasured confounding**) mean that we can adjust for L (typically *a set* of variables) using an appropriate statistical model and estimate the causal effect of A on Y

2 The counterfactual approach

Counterfactuals

- Consider a study where two treatments are being compared; each patient i have **two possible responses**:
  - Y_i^1 if they were treated
  - Y_i^0 if they were *not* treated
- Y^1 and Y^0 are called **counterfactual random variables**, or potential outcomes

  Alternative notation is $Y^{do(a=1)}$, just Y^1 or $Y(1)$

Causal estimands

- An *estimand* is the parameter we want to estimate; should be well-defined: **what causal effect do you want to identify?**
- For example, the **individual causal effect** for patient ***i***, θ_i, for a given outcome Y, is

$$\theta_i = Y_i^1 - Y_i^0$$

- **"The fundamental problem of causal inference"** (Holland 1986) is that both counterfactuals for person i can not be observed at the same time (only one of them is *factual*); identifying individual causal effects are generally not possible

Average causal effects

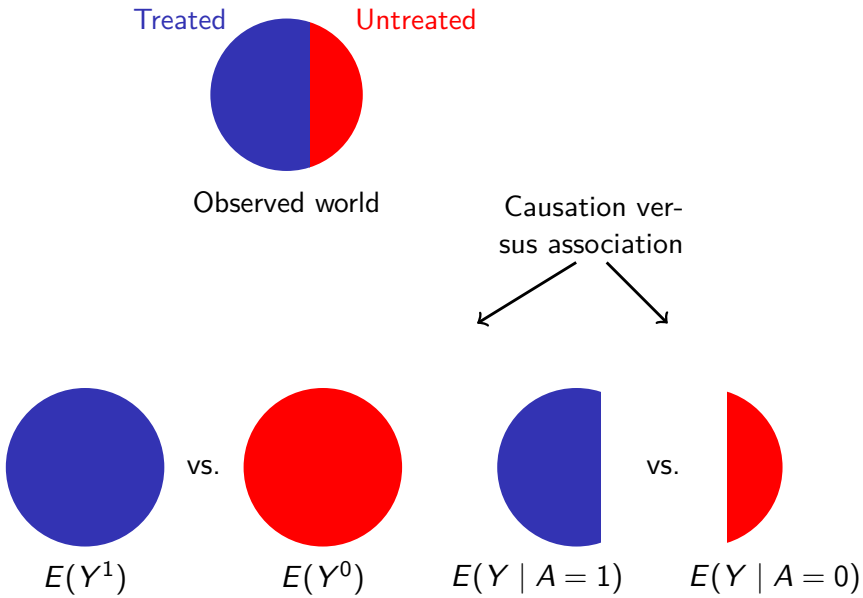
- Therefore we typically estimate **the average treatment effect** (ATE)

$$E(\theta) = E(Y^1) - E(Y^0),$$

  alternatively as a ratio $E(Y^1)/E(Y^0)$
- In other words, we seek to test the null hypothesis of no average causal effect; the so-called **causal null hypothesis**:

$$E(Y^1) - E(Y^0) = 0$$

Causation versus association



- $P(Y = 1|A = a)$ is a **conditional probability**
- $P(Y^a = 1)$ is a counterfactual **marginal probability**

Other causal estimands

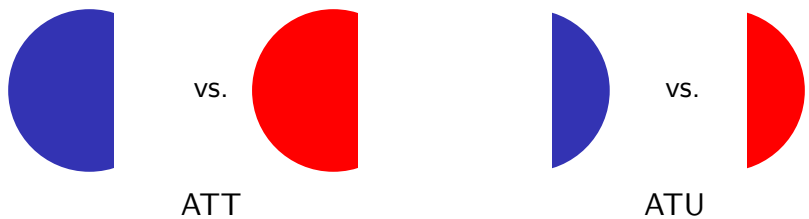
- Average causal effects can also be conditional, for example the **treatment effect on the treated**

$$ATT = E(Y^1|A = 1) - E(Y^0|A = 1),$$

or the treatment effect on the untreated

$$ATU = E(Y^1|A = 0) - E(Y^0|A = 0)$$

Note: two of the above counterfactuals are fully observed



Causal assumptions

- An observational study can be analysed as a conditional randomised trial under the following **three general assumptions**:
  1. Treatment is well-defined (no multiple versions): consistency
  2. No unmeasured confounding: conditional exchangeability
  3. The conditional probability of receiving every value of treatment is always greater than zero: positivity
- The first two are generally **not testable** and will rely on expert knowledge
- However, **sensitivity analysis** is useful – so is describing the ideal *target trial* you want to model

Estimation methods

- Given the three general causal assumptions there are **four types of methods** for analysing observational data as conditional randomised experiments:
  - Stratification (including multiple regression)
  - Matching
  - Standardisation
  - Inverse probability of treatment weighting (IPTW)
- The **first two are conditional methods** and the two latter based on marginalization – can identify different type of effects

  When can estimates differ? For *non-collapsable* models (e.g. for OR's, HR's)

Also: Under other assumptions, **instrumental variables** and the frontdoor formula (not much used) are other general alternatives

Inverse probability weighting

- A popular method for dealing with **missing data**, even though it for a long period gained little acceptance relative to more popular missing data approaches such as multiple imputation (Rubin, 1987)

  Changed with work following Robins, Rotnitzky, Zhao (1994)
- Will look briefly at the inverse probability of treatment weighting (IPTW) approach in a point treatment setting, as it is **a key method** in causal inference

The weight

- **Want to observe the full counterfactual data**, but obviously don't observe all (*How many do we observe?*)
- We want **weight** up the observed individuals in a smart way to account for the missing counterfactuals. Treated individuals are weighted with

$$w_i = \frac{1}{P(A = 1|L = l_i)},$$

and untreated individuals are weighted with

$$w_i = \frac{1}{P(A = 0|L = l_i)}$$

- The weighting gives us a counterfactual **pseudo dataset**, twice the size of the actual observed dataset
- The probabilities P(A=1|L=l) are also known as **propensity scores** (Rosenbaum and Rubin, 1983)

Toy example

- Say that you have a study with **three identical individuals** (with respect to covariates)

  Individual 1 is given treatment and individual 2 and 3 are not

- Inverse probability weighting is to **give individual 1 the weight**

$$w_1 = \frac{1}{1/3} = 3$$

  and individual 2 and 3 the weights

$$w_2 = w_3 = \frac{1}{2/3} = \frac{3}{2} = 1.5$$

- **Weighted data** corresponds to having 3 individuals in the treated group and $1.5 + 1.5 = 3$ individuals in the control group (a full counterfactual dataset)

Key points on IPTW

- Mathematically **equivalent with standardisation** for point treatments
- Typically use **stabilized weights** on the form

$$w_i = \frac{P(A = a)}{P(A = a|L = l_i)},$$

  for a equal to 0 or 1
- IPTW (and standardisation) was shown by Robins to generalize to **time-varying treatments** $\rightarrow$ marginal structural models

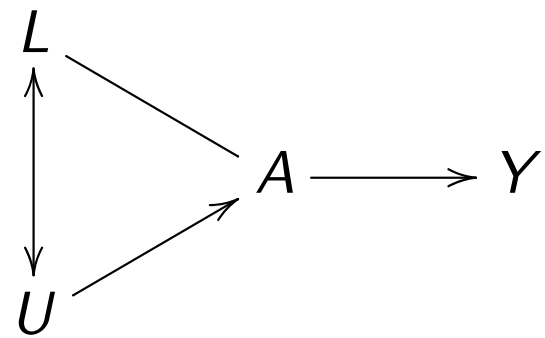
The counterfactual approach in summary

1. Think of your ideal **target trial**
2. Specify the **causal estimand**
3. List all the **(causal and statistical) assumptions** needed for the estimand to be identified
4. Perform the **analysis**
5. Assess the results using **sensitivity analysis**

3 Causal DAGs

Terminology

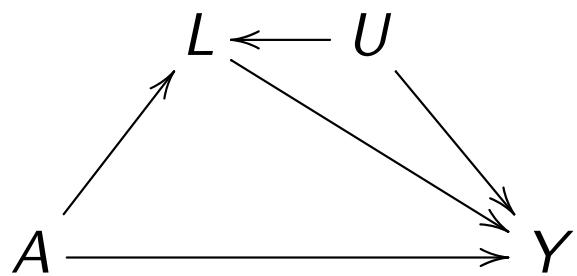
- **A graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$** is a set V of *vertices* or *nodes* and a set E of *edges*, which can be graphically illustrated, for example;



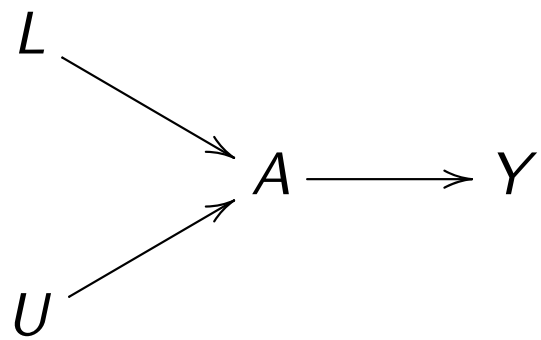
A graph with nodes L, U, A and Y

- **Nodes** typically represent random variables
- **Edges** (arrows) can be undirected, **directed** or bi-directed and typically indicate a certain relationship between nodes

- **Paths**: A trail of edges going from one node to another. For example, the following graph has two *directed* paths and one *undirected* path from A to Y:



- A **cyclic graph** have at least one path that can be followed through directed edges back to the original node
- An **acyclic graph** is a graph that contains no such cycles
- A **directed acyclic graph (DAG)** is a graph that is both directed and acyclic



A DAG with nodes L, U, A and Y

Additional terminology

- **Children and parents**: Nodes directly affected by and affecting other nodes respectively

$$P \longrightarrow C$$

- **Ancestors and descendants**: Nodes directly or indirectly affected by and affecting other nodes

$$A \longrightarrow P \longrightarrow C \longrightarrow D$$

- **Exogenous and endogenous nodes**: Nodes without and with parents respectively

$$X \longrightarrow N$$

DAGs and causal inference

- On their own, DAGs are **just dots and arrows**
- They are mathematical objects, part of a larger class of graphical models such as Bayesian networks or Markov networks – **used in various areas**
- To use them for causal inference, we need to put further *interpretation* on them and connect them to both data (probability distributions) and counterfactuals – say hello to **causal DAGs**

Causal DAGs

- **Causal DAGs**, also known as causal diagrams or causal bayesian networks, are, by the definition of Hernan, Robins (2019), DAGs where
  1. the lack of an arrow between two variables A and Y can be interpreted as an absence of a direct causal effect of A on Y
  2. where all common causes, even unmeasured, of any pair of variables in the graph are included
  3. any variable is a cause of its descendants
- Additional **assumptions** are needed to make these causal DAGs useful:
  1. The causal Markov assumption
  2. Faithfulness

  Markov is what links the causal DAG to conditional probabilities (e.g. data)

The causal Markov assumption

- Conditional on its parents, a variable X_j is independent of non-descendant (conditional independence)

  Mathematically equivalent to the statement that the joint distribution of the variables $\boldsymbol{X} = \{X_1, X_2, ..., X_n\}$ in a DAG can be factorized using the **Markov factorization**

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i | pa(X_i)),$$

  where $pa(X_i)$ denotes the parents of X_i

Note: This is **"just" the math behind** – if too much, jump to slide 38 and causal DAGs can still be of great practical use without detailed knowledge of this mathematical background

- It follows that the probability distribution generated by an intervention, for example $do(X_1 = x_1)$, is given by **truncated factorization** (Pearl)

$$P(X_2, ..., X_n | do(X_1 = x_1)) = \prod_{i \neq 1} P(X_i | pa(X_i))$$

  This is also known as the manipulation theorem (Spirtes), intervention formula (Lauritzen) or G-formula (Robins)

- Now say $\boldsymbol{X} = \{A, Y, X_3, ..., X_n\}$, the causal effect of A on Y can now be derived by **marginalizing** (summing) the truncated factorization formula over $\boldsymbol{X'} = \{X_3, ..., X_n\}$;

$$P(Y | do(A = a)) = \sum_{x'} P(Y, \boldsymbol{X'} | do(A = a))$$

  (note that these formulas are not restricted to interventions one a single variable)

- This allow us to calculate **causal effects**, such as

$$ATE = E(Y|do(A = a)) - E(Y|do(A = a'))$$

or

$$ATT = E(Y|A) - E(Y|do(A = a'))$$

The faithfulness assumption

- **No cancellation of effects**. For example: if half of a population were men where exposure A (e.g. heart transplant) had a beneficial effect and the other half were women where exposure A had a harmful effect, and these effects canceled out perfectly, there would be no association

$$A \longrightarrow Y$$

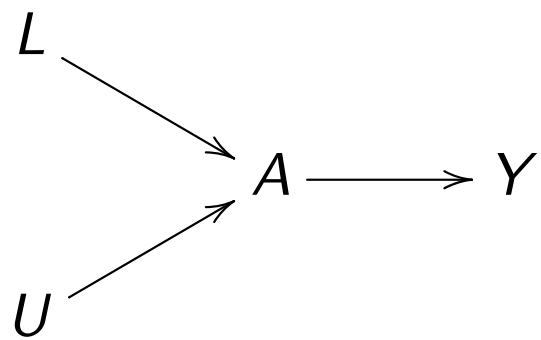
A causal DAG for the variables A and Y

The joint distribution of the data is said to be unfaithful to the causal DAG

- Perfect cancellation of effect is very rare, and **faithfulness is assumed to make independence and graphical independence equivalent concepts**

DAGs and independence

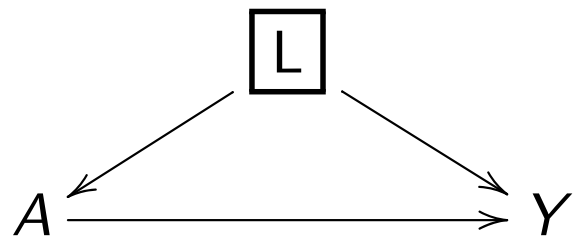
- Two variables are **marginally independent** if there are no arrow between them and they share no common cause
- If two variables are not marginally independent, they are **marginally dependent**



For example: L and U are marginally independent, L and A (or L and Y) are marginally dependent

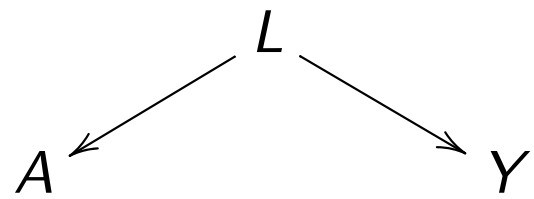
DAGs and conditional independence

- **Notation**: conditioning on a variable in a graph is typically denoted using a square box around it:

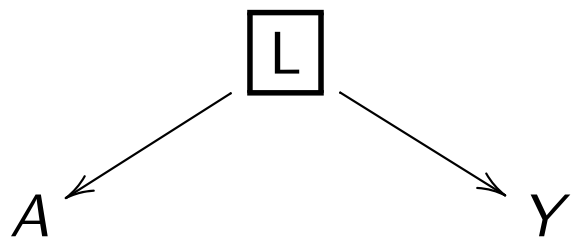


- Statements about **conditional dependencies and independencies** between variables can also be read of from a DAG

- If there are no causal relationship between A and Y, but they share a common cause L, ***A* and *Y* are associated**:

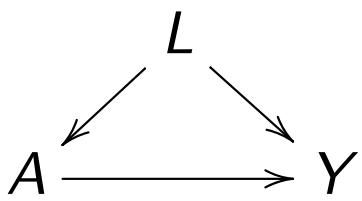


- However, conditioning on L, A and Y would not be associated, and ***A* and *Y* are conditional independent**:

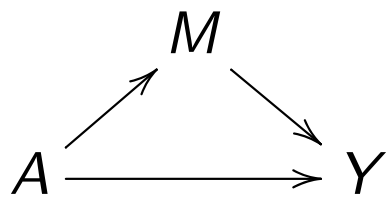


Causal structures

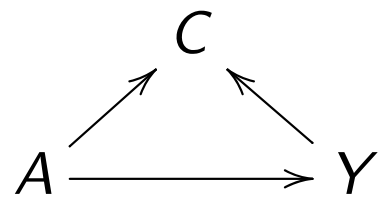
- There are **three basic causal structures** between a treatment variable A, an outcome Y and a third variable:



Confounder



Mediator



Collider

(e.g. selection bias)

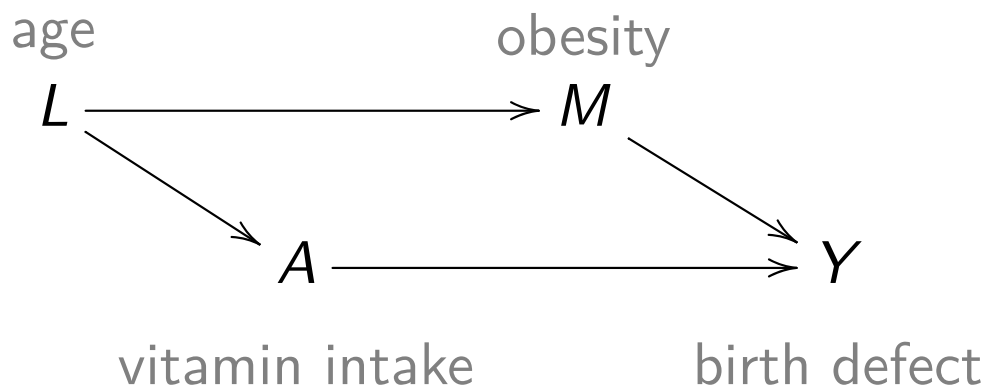
- **Three basic rules**:

  1. Adjust for confounders
  2. Adjusting for mediators do not identify total effects
  3. *Never* adjust for colliders

- Formally, in larger DAGs; we look at all *paths* from A to Y and use so-called **d-separation** and *back-door/adjustment criterions*

Example: Confounder

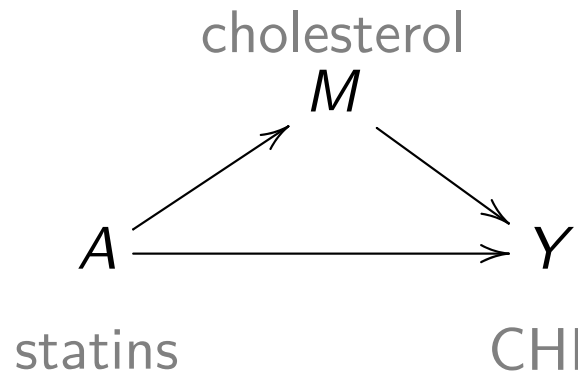
- Consider **the effect of maternal vitamin intake during pregnancy on birth defects**



- A confounder **does not need to be a *direct* cause** of exposure and outcome
- Confounding need to be **adjusted for** to estimate the causal effect of A on Y

Example: Mediator

- Consider **the effect of cholesterol lowering drugs on coronary heart disease**

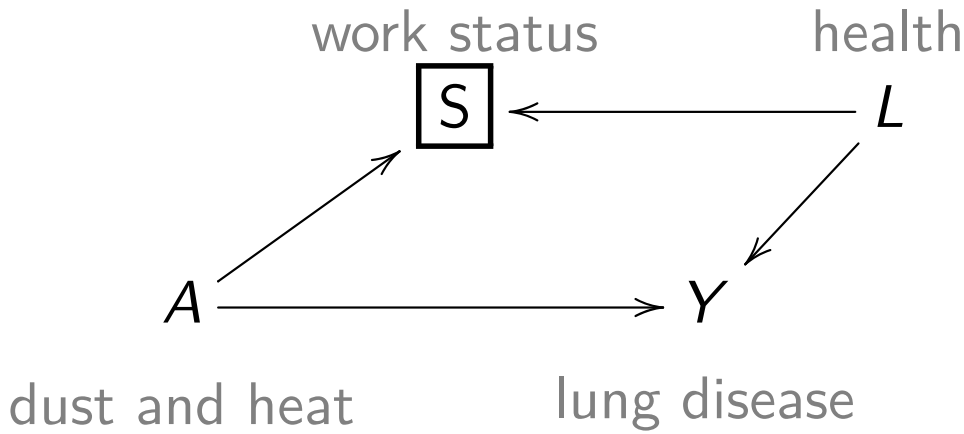


- Adjusting for M, which is on the causal pathway between A and Y, ***removes*** **part of the effect** that goes through M

  More on how to do formal mediation analysis later

Example: Collider

- Lets consider so-called **Healthy worker bias**, where only subjects that remain working is included in the study



  This can cause bias, that if large enough, may result in dust and heat looking protective against lung disease

- Again, a collider does not need to be a *direct* effect of exposure and outcome, for example in the case of **selection bias**

  Here, only one stratum of S is included in the study, implying that **we condition on S**

D-separation (open and closed paths)

- Say that a path with colliding arrows ($\rightarrow\leftarrow$) is closed and a path with no colliders is open – d-separation **with respect to a conditioning set of nodes** Z implies
  1. Conditioning on a non-collider closes/blocks/d-separates the path
  2. Conditioning on a collider, or a descendant of a collider, opens/d-connects the path

  Note: A path closed in at least one point is considered closed and Z might be a empty set $\{\}$
- D for "**directional**" (Pearl, 1988)

The adjustment criterion (Shpitser et al. 2010)

- The causal effect of A on Y can be identified when:

  all non-causal paths between A and Y are closed and all causal paths are open

- A **causal path** is a path where all arrows are pointing in the same direction ($\rightarrow\rightarrow$); otherwise the path is non-causal
- The adjustment criterion is **complete**; detects all sets Z that indentify the effect of A on Y by conditioning

The backdoor criterion (Pearl 1995)

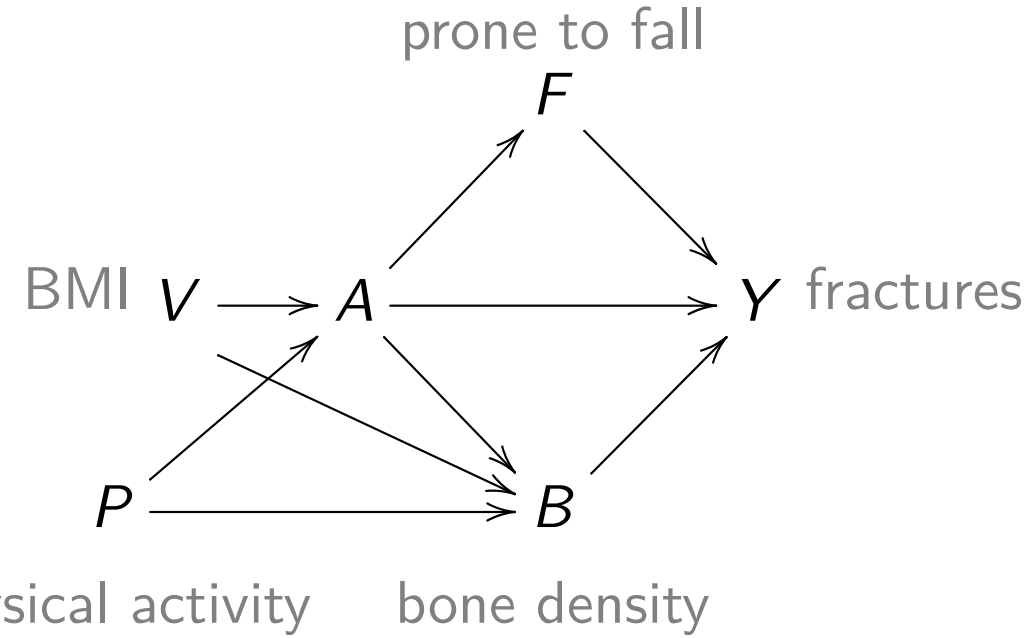
- Alternatively to the adjustment criteria, the causal effect of A on Y can be identified when adjusting for the set of variables Z that satisfy:

  1. No element of Z is a descendant of A
  2. Z blocks all backdoor paths from A to Y

  where a backdoor path from A to Y is any path starting with a arrow into A (on the form $A \leftarrow \cdots Y$)
- The backdoor criterion **omits some unnecessary conditioning sets** compared to the adjustment criterion

Exercise: Diabetes and fractures

- Want to estimate the total effect of **diabetes** $\boldsymbol{A}$ on fractures:

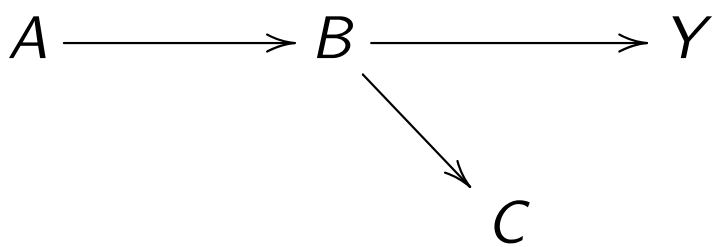


1. What are the relevant paths? How should they be treated?
2. Identify the mediators and confounders
3. Is B a collider?

	Path	Type	Status	
1	$A \rightarrow Y$	Causal	Open	**V**
2	$A \rightarrow F \rightarrow Y$	Causal	Open	**V**
3	$A \rightarrow B \rightarrow Y$	Causal	Open	**V**
4	$A \leftarrow \boxed{\text{V}} \rightarrow B \rightarrow Y$	Non-causal	Closed	**V**
5	$A \leftarrow \boxed{\text{P}} \rightarrow B \rightarrow Y$	Non-causal	Closed	**V**

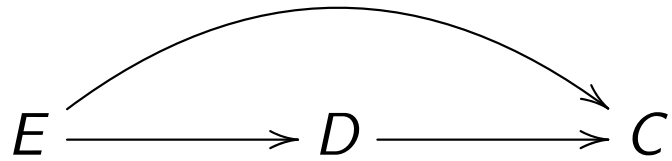
Exercise: Adjustment for C?

Why would conditioning on C ruin the identification of the total effect of A on Y?



Exercise: Adjustment for birth weight?

Consider an example on birth defects;

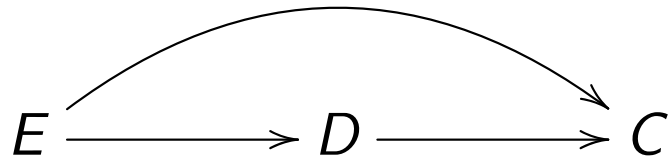


Low folate intake E may increase the risk of preterm delivery and infant low birth weight C, and many birth defects D result in preterm deliveries and low birth weight infants (Hernan, American Journal of Epidemiology 2002)

Should we adjust for C when estimating the effect of E on D?

Exercise: Adjustment for birth weight?

Consider an example on birth defects;



Low folate intake E may increase the risk of preterm delivery and infant low birth weight C, and many birth defects D result in preterm deliveries and low birth weight infants (Hernan, American Journal of Epidemiology 2002)

Should we adjust for C when estimating the effect of E on D?

No, it is 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)

Causal DAGs in summary

- Construct an assumed causal DAG based on **subject matter background knowledge**, typically involving unobserved quantities

  Remember that the absence of an arrow (or a path) between two variables is a stronger statement than the opposite
- Check if there is a set of observed variables that satisfy one of the mentioned **critions**
- If such a set exist, **adjust** for it using a suitable statistical method

4 Further topics

Time-dependent confounding

- A more complex confounding problem that could arise when you have **time-dependent treatments**
- The problem had no formal solution before Robins (1986) introduced his three **g-methods**; inverse probability of treatment weighting, g-computation and g-estimation

  These methods have lead to a revolution in how we analyse observational data

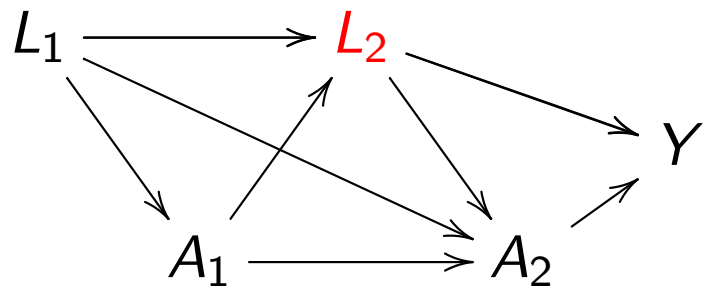
Time-varying treatments

- Individuals can **change from untreated to treated** (and possibly back and forth), other variabels may also vary with time and explain treatment and outcome
- Instead of differing between two values of treatments, we now have different **treatment regimes** $\bar{a} = \{a_1, a_2, a_3, ...\}$, for example $\bar{a} = \{0, 0, 1, 1, 1\}$

  As a result treatment effects such as the ATE are no longer uniquely defined

Definition of time-dependent confounding

- Present when **covariates L_2, affected by past exposure A_1, both affects future exposure A_2 and the outcome Y**



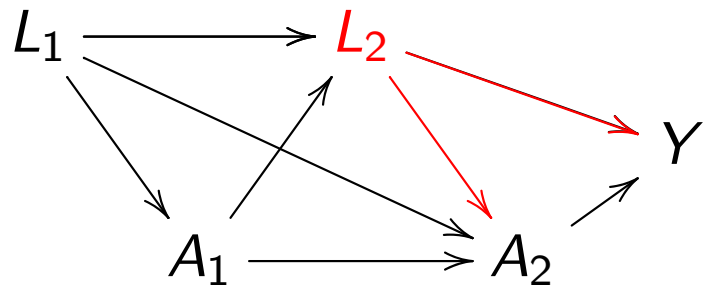
Simplified DAG with only two time points.

Classical example: HIV treatment A, CD4 cell count L and AIDS or death Y

- Can't both adjust and *not* adjust for L_2 using traditional methods – **need advanced methods**

Definition of time-dependent confounding

- Present when **covariates L_2, affected by past exposure A_1, both affects future exposure A_2 and the outcome Y**



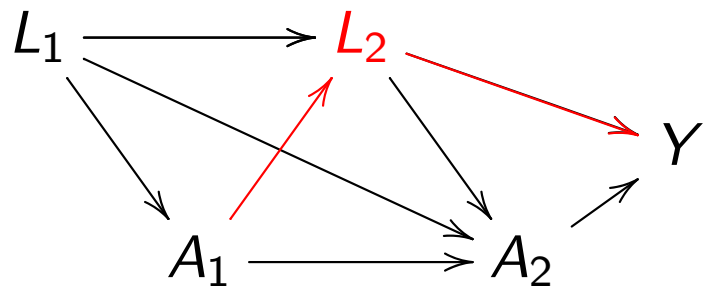
Simplified DAG with only two time points.

Classical example: HIV treatment A, CD4 cell count L and AIDS or death Y

- Can't both adjust and *not* adjust for L_2 using traditional methods – **need advanced methods**

Definition of time-dependent confounding

- Present when **covariates L_2, affected by past exposure A_1, both affects future exposure A_2 and the outcome Y**



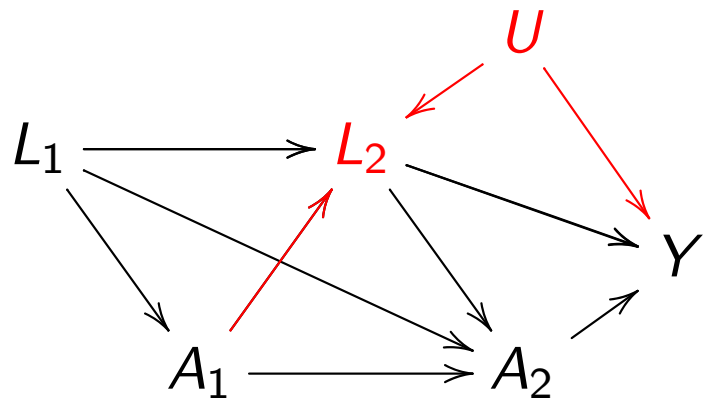
Simplified DAG with only two time points.

Classical example: HIV treatment A, CD4 cell count L and AIDS or death Y

- Can't both adjust and *not* adjust for L_2 using traditional methods – **need advanced methods**

Definition of time-dependent confounding

- Present when **covariates L_2, affected by past exposure A_1, both affects future exposure A_2 and the outcome Y**

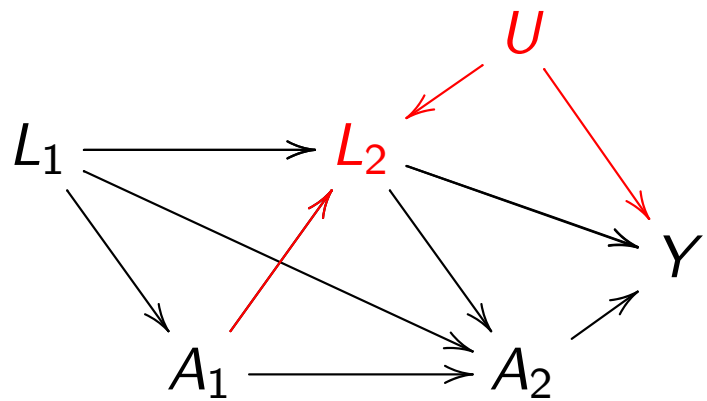


Simplified DAG with only two time points.

Classical example: HIV treatment A, CD4 cell count L and AIDS or death Y

- Can't both adjust and *not* adjust for L_2 using traditional methods – **need advanced methods**

- **Note** that even with just two time-points, the figure



  is *over-simplified*; we typically would have arrows $L_1 \to Y$ and $A_1 \to Y$
- If the outcome Y is **time-to-event**, we would instead have Y_1 and Y_2, and typically censoring C_1 and C_2

Example: Data with time-varying treatments

patient ID	time	treatment	CD4	death	censored
1	1	no	high	no	no
1	2	no	low	no	no
1	3	yes	high	no	no
1	4	yes	high	no	no
1	5	yes	low	yes	no
2	1	no	high	no	no
2	2	yes	high	no	no
2	3	yes	low	no	no
2	4	yes	low	no	yes

New target trial and causal assumptions

- The new target trial is **the sequential randomized trial**: A trial in which treatment is randomly assigned to individuals at each time point

  Not frequently used in practice, but a helpful concept to understand key conditions for estimation of causal effects of time-varying treatments
- Modified general causal assumptions: correspondingly we now need **sequential conditional exhangeability** to identify causal effects, together with sequential versions of positivity and consistency (See Hernan, Robins 2019)

Marginal structural models (MSMs)

- Marginal structural models were introduced by Robins (1998) and are formally models for the marginal comparison of the **ATE of given treatment regimes** $\bar{a} = \{a_1, a_2, a_3, ...\}$, for example $\{1, 1, 1, 1\}$ vs $\{0, 0, 0, 0\}$, or $\{0, 0, 1, 1\}$ vs $\{1, 1, 0, 0\}$
- Note that the marginal models for point treatments can also, strictly speaking, be labeled MSMs, but the term is typically used for **methods adjusting for time-dependent confounding**
- The most common way of fitting MSMs are through **inverse probability of treatment weighting** (IPTW), but g-computation is another alternative

  Same principal as before, but now every individual observed at time t is weighted by the inverse probability of the *treatment history* up to time t

- Again it can be showed that **the stabilized weights**

$$\boldsymbol{sw_i(t) = \prod_{k=1}^{t} \frac{P(\text{treatment at } k | \text{baseline covariates})}{P(\text{treatment at } k | \text{covariates up to } k)}}$$

  are *less* variable and still produces unbiased estimates
- The probabilities in the formula for the weights are typically estimated using **pooled logistic regression**

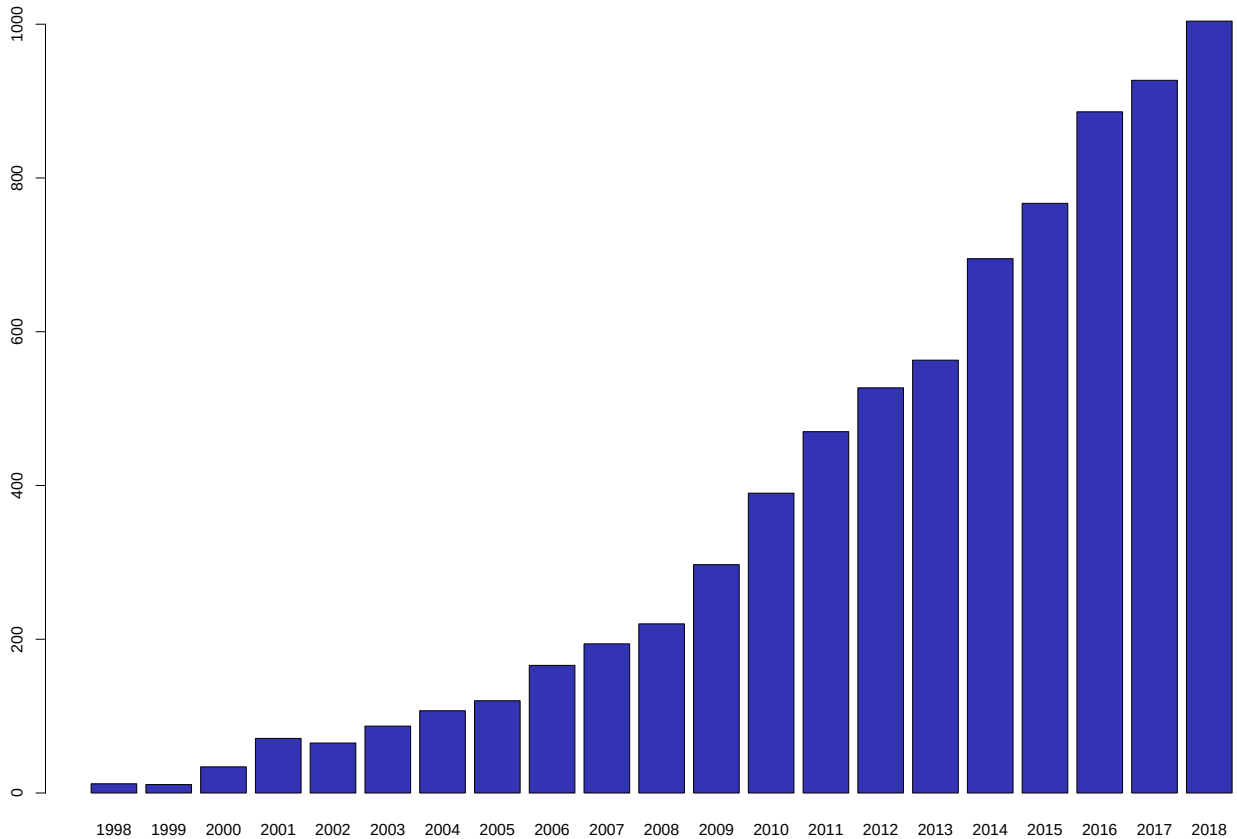
Intuition behind MSM by IPTW

- The weights create a world where **every individual is observed under any possible treatment regime**

  This corresponds to a sequentially randomized treatment
- In this new counterfactual world, treatment is no longer dependent of any covariate

  Analyze the weighted data, adjusting only for treatment (and baseline covariates in the case of stabilized weights)

Google scholar hits 1998–2018



Number of **hits by publication year** for *"marginal structural model" OR "marginal structural models"* per March 20th, 2019

For comparison; "Cox regression" from 1920 to 24300 hits, and "logistic regression" from 18100 to 51300, in the same period

Example: Swiss HIV Cohort Study (Stern et al, Lancet, 2005)

- An ongoing multi-center research project **following up HIV infected adults** aged 16 or older
- The data we use goes from 1996, when **HAART treatment** became available in Switzerland, to September 2003
- The data are organised in monthly intervals, with measures of for example **CD4 and HIV-1 RNA levels**, together with other variables describing sickness and treatment history
- Time between visits vary, but scheduled clinical follow up is every 6th month, with laboratory **measures taken every 3rd month**
- In total 2161 individuals, where observation time varies from 1 to 92 months. 202 people progressed to **'AIDS or death'** and 717 where ever treated with HAART

Statistical analysis

We used weighted Cox proportional hazards models to estimate hazard ratios for progression to AIDS or death, controlling for time-dependent confounding. These models estimate the parameters of marginal structural models.[13] The weights are based on the inverse of each patient's probability of the treatment history they actually had, given their covariate history. The weighted analysis creates a statistical population in which the probability of being treated at each time is unrelated to the measured prognostic factors (the time-dependent confounders). Because these confounders are controlled by the weights rather than by inclusion as covariates in the Cox models, this approach avoids the problem that such confounders could also be intermediate on the causal pathway from HAART to the outcome of AIDS or death. Follow up ended when AIDS or death occurred,

We estimated the probability of treatment with HAART using a pooled logistic regression in which the outcome was treatment with HAART (of patients not already on such treatment). The covariates were CD4 count, concentration of HIV-1 RNA, haemoglobin, and CDC stage B events, together with lagged and baseline values of these variables, time since January 1996, baseline age, sex, and presumed transmission group.

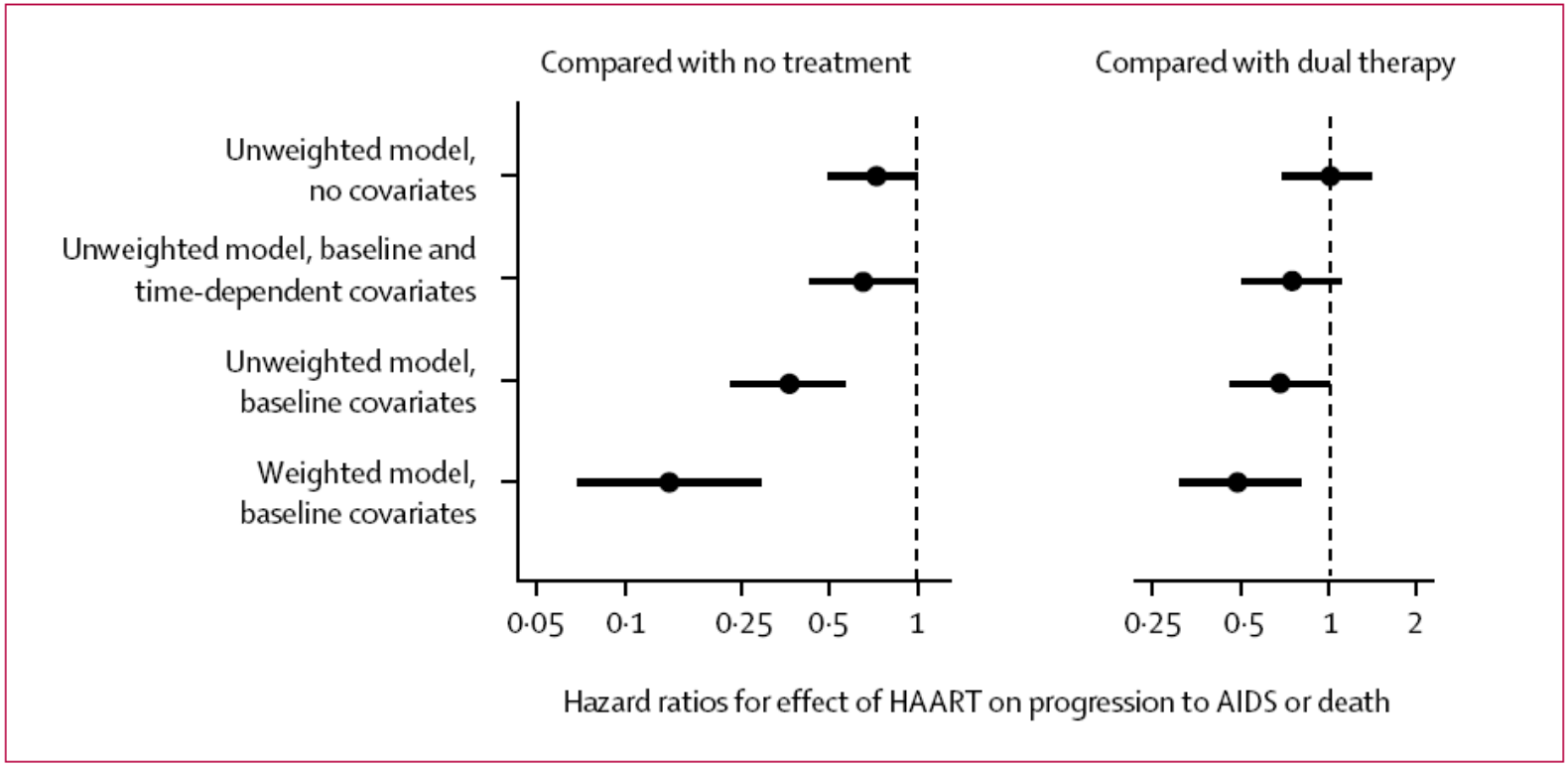


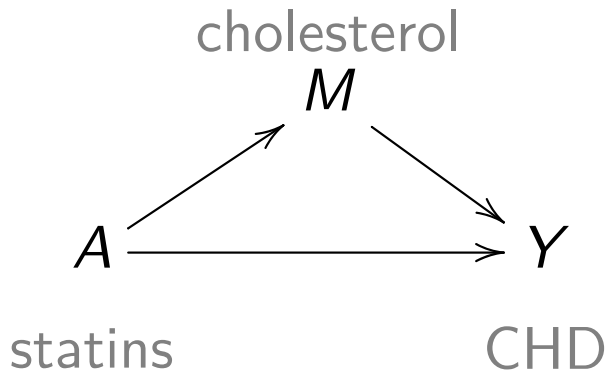
Figure: **Estimated effect of HAART from unweighted (standard) and weighted Cox models**
Weighted model with baseline covariates estimates parameters of marginal structural model. Weights adjust for confounding due to measured time-dependent covariates.

Time-dependent confounding in summary

- MSMs fitted using IPTW is the most common approach and it identifies the ATE between given **treatment regimes**, however; there are also methods for estimating ATT under such treatments (see e.g. Gran et al. 2018)
- Have not talked about all **alternatives**; g-estimation, doubly robust methods or stratification based methods (e.g. the sequential Cox approach; see Gran et al. 2010)
- The major benefit of the IPTW approach is that it naturally extend traditional regression methods; most **software** packages today easily let you include weights in your linear/logistic/Cox regression etc – various packages also tutorials also exist

Mediation analysis

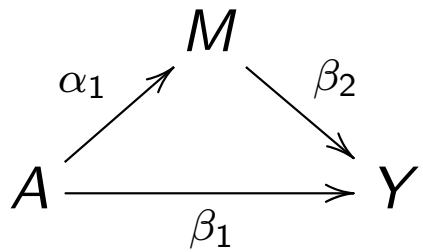
- Typically motivated by a wish to understand the mechanisms of how treatment works



- **How much of the effect go through *M*** and how much does *not* go through M?

Traditional approach

- From **Wright 1921, 1934 via Baron and Kenny 1986**: Using linear models, a common approach (the product method) is to estimate coefficients in



by fitting the two models:

$$\begin{aligned} M &= \alpha_0 + \alpha_1 X \\ Y &= \beta_0 + \beta_1 X + \beta_2 M \end{aligned}$$

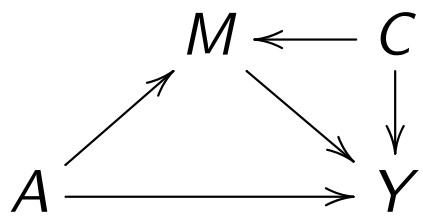
Then, β_1 is labeled a direct effect and $\alpha_1\beta_2$ a indirect effect

Can be extended to arbitrarily complex diagrams, as long as all models are simple linear models

- **Massively applied** in psychology, social science, epidemiology etc. (Baron and Kenny have 80 000 Google Scholar citations)

Critique against traditional approach

1. **Mediator-outcome confounding**; even if A is randomized, there might still be confounding between M and Y;



   was not mentioned in Baron and Kenny (1986), and as a result mostly ignored; even claimed to *not* be a problem in randomised trials – can lead to heavy bias

2. **Exposure-mediator interactions** can also lead to bias if neglected
3. Indirect and direct **effect definitions are model based**, which do not give an intuitive interpretation

Causal mediation analysis

- Robins and Greenland (1992) is often mentions as the starting point of causal mediation analysis; have contributed with a better understanding of the **strong assumptions needed** to identify direct and indirect effects with causal interpretation
- **Start with a well-defined causal estimand**, using counterfactuals, and indentify the causal and statistical assumptions needed for its identification
- If possible, perform **sensitivity analysis** for untestable assumptions

Causal estimands for mediation

- **Natural direct and indirect effects** are the most common (Robins and Greenland 1992);

$$\text{NDE} = E[Y^{1,M^0}] - E[Y^{0,M^0}]$$

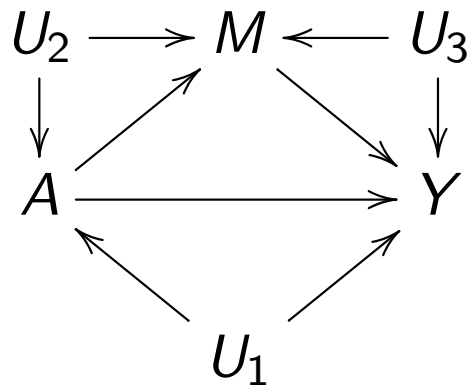
$$\text{NIE} = \text{ATE} - \text{NDE} = E[Y^{1,M^1}] - E[Y^{1,M^0}]$$

  where Y^{1,M^0} is a so-called nested counterfactual, a purely hypothetical quantity used to create model-free definitions of direct and indirect effects

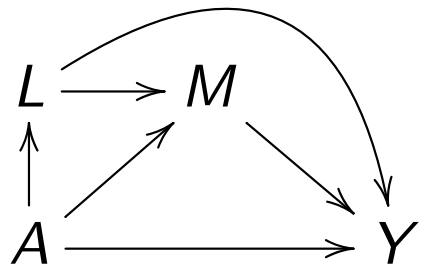
- Identification possible by **extending the causal assumptions** from earlier

Key causal assumptions for NDEs and NIEs

- **Conditional exchangeability**, which now means, as expected, that there are noe *unobserved* confounders U_1, U_2 or U_3:



- **And**; **the cross-world independence assumption**, which implies no intermediate confounders L:



Note that there exist an alternative assumption that allow for L (Petersen et al. 2006), but this assumption is also very strong

Mediation analysis in summary

- Given the assumptions, identification of NDEs and NIEs breaks down to a **mediation formula** with only observable quantities – typically need to model the main outcome *and* mediator using regression models
- Depending interactions, types of outcomes/mediators etc, estimation of NDEs and NIEs breaks down to either analytic **formulas or a simulation approach**
- Other measures exist, such as *controlled direct effects*, and are sometimes needed; **vast methodological development**, see e.g. the book by VanderWeele (2015)
- Well-developed **software** for causal mediation analysis can be found for R and Stata (e.g. packages `mediation`), with built in tools for sensitivity analysis

5 Take-home messages

- **Counterfactual worlds** are the actual gold standard; RCTs are just the next best thing
- Formalized causal inference start with a **target trial**; define causal estimand, identify needed assumptions – causal and statistical, analyze and do sensitivity analysis
- **Causal DAGs**, together with selection criterions, are great tools for identifying variables to adjust for (and not)
- **New methods for time-dependent confounding and mediation** analysis have showcased the importance of formal methods for causal inference and changed how we analyse observational data

- Causal inference can be theoretical and notational heavy, *but* one **can separate modelling and interpretation** by asking

  "what is the hypothetical randomised trial?"

- Causal inference is often misunderstood as to ambitious for observational data, but it is really about **being *more* explicit**, and thus transparent about the limitations

*The calculus of causation consists of two languages: causal diagrams, to express **what we know**, and a symbolic language, resembling algebra, to express **what we want to know***

The Book of Why, Pearl and Mackenzie (2018)

6 Further reading

- Hernan MA, Robins JM (2019). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming (book).
- Pearl J, Mackenzie D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books (book).
- Pearl J, Glymour M, Jewell NP. (2016). Causal Inference in Statistics: A Primer. John Wiley & Sons (book).
- VanderWeele T (2015). Explanation in Causal Inference: Methods For Mediation and Interaction. Oxford University Press (book).
- VanderWeele T (2019). Principles of Confounder Selection. European journal of epidemiology.
- Hernan MA (2018) The C-word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. American Journal of Public Health.
- Hernan MA (2010). The Hazards of Hazard Ratios. Epidemiology.