Directed Acyclic Graphs

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Causal inference in epidemiology

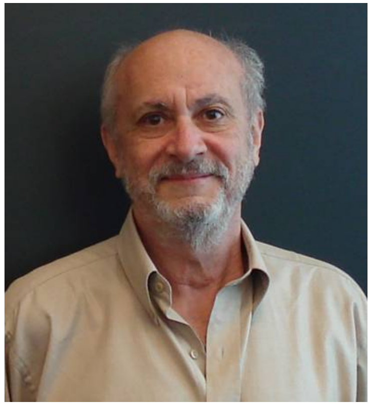
- Many epidemiological research questions are centered around a particular **exposure** and a particular **outcome**
- Typically, we want to learn whether the exposure has a **causal effect** on the outcome
  - E.g. does smoking during pregnancy (exposure) cause malformations (outcome) in newborns?

Causal inference in statistics

- Despite its relevance for epidemiological researchers, causality was largely ignored in the statistical field for most of the 20$^{\text{th}}$ century
  - ‘Statistics can only tell us about association, not causation’
- **Causal inference** is a rather new ($\sim$ 30 years) branch of statistics, specifically devoted to issues of causality
  - Under what conditions can we estimate causal effects?
  - Which statistical methods are most appropriate for causal effect estimation?

Brief history of causal inference, 70’s

- Donald Rubin developed a formal definition of causation
  - **potential outcomes**
  - **counterfactuals**



Brief history of causal inference, 80's

- James Robins discovered - and solved - some important problems with longitudinal studies, from a causal inference perspective
  - **Marginal Structural Models** (MSMs)
  - **Structural Nested Models** (SNMs)

Brief history of causal inference, 90's

- Judea Pearl developed **Directed Acyclic Graphs** (DAGs)
  - Simplify interpretation and communication in causal inference
  - Useful for covariate selection in observational studies

Outline

Motivating example

DAG terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

Outline

Motivating example

DAG terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

Statistical association

- Research question: does smoking during pregnancy (SDP) cause malformations in newborns?
- For a large number of pregnancies, we collect data on both exposure and outcome
- Suppose that we observe an inverse statistical association between SDP and malformations ($RR = 0.8$)
- *Can we then say that SDP protects against malformations?*

A possible non-causal explanation

- Young mothers smoke more often than old mothers
- Young mothers have smaller risk for malformations in their babies, than old mothers
- Hence, smokers are more likely to be young, and for this reason less likely to have babies with malformations, than non-smokers
  - Even in the absence of a causal effect

Confounding

- Mothers age is a potential ‘confounder’ that may induce non-causal associations between SDP and malformations
  - There are several definitions of the term ‘confounder’ in the literature - more later
- *How can we eliminate the influence of confounders?*

Randomization

- By randomizing the exposure, we guarantee that there are no systematic differences between exposed and unexposed
  - Same distribution of age, sex, ethnicity, genes etc
- In an ideal randomized controlled trial (RCT) there is no confounding
  - Observed statistical associations can be given causal interpretations
- *Any problems?*

Problems with randomized trials

- Unethical
- Expensive
- Unpractical
- Non-compliance
- Non-blinding
- etc

Adjusting for potential confounders in the analysis

- Suppose that we stratify the sample on age
  - Each stratum contains women of similar age
- Each stratum is analyzed separately
- Within-stratum associations can not be attributed to different age distribution between exposed and unexposed
- The confounding influence of age is eliminated
- *Any problems?*

Problems with confounder adjustment

- We can only adjust for confounders that
  - we are aware of
  - we have measured
- Often many potential confounders are unknown to the investigator, and/or difficult/expensive to measure
  - e.g. genetics, lifestyle factors

Technical note

- There are several methods for confounding adjustments, which are often combined
  - stratification
  - matching
  - standardization
  - propensity scores
  - regression modeling
  - inverse probability weighting
  - etc
- For realistic sample sizes, these methods have different pros and cons
- For HUGE samples, they are equivalent
- We use the term 'adjusting' generically, for any of the methods

Motivating example revisited

- Suppose that we have measured five covariates:
  - the mothers age at conception
  - the mothers socioeconomic status/education level at conception
  - the mothers diet during pregnancy
  - indicator of whether there is a family history of birth defects
  - indicator of whether the baby was liveborn or stillborn
- *Which of these are 'true' confounders, i.e. which should we adjust for?*

The need for covariate selection

- One strategy would be to adjust for all measured covariates
- This strategy may not be optimal, because
  - **some covariates may not be confounders, and may increase bias if adjusted for**
  - more covariates requires a bigger model, with a higher potential for bias due to model misspecification
  - some covariates may be prone to measurement errors, and may therefore lead to bias
  - some covariates may reduce statistical power/efficiency when adjusted for
- Therefore, it is often desirable to adjust for a subset of covariates

Traditional covariate selection strategies

- Adjust for covariates that are selected in a stepwise regression procedure
- Adjust for covariates that change the point estimate of interest with more than 10%
- Adjust for covariates that
  - are associated with the exposure, and
  - are conditionally associated with the outcome, given the exposure, and
  - are not in the causal pathway between exposure and outcome

Problems with traditional strategies

- They rely on statistical analyses of observed data, rather than *a priori* knowledge about causal structures
  - require that data is already collected, and cannot not be used at the design stage
- They may select non-confounders, which may increase bias if adjusted for

Directed Acyclic Graphs

- Directed Acyclic Graphs (DAGs) can be used to overcome the problems with traditional covariate selection strategies
- A DAG is a graphical representation of underlying causal structures
- DAGs for covariate selection:
  - encode our *a priori* causal knowledge/beliefs into a DAG
  - apply simple graphical rules to determine what covariates to adjust for

Outline

Motivating example

DAG terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

A simple DAG

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

- Each arrow represents a causal influence
- 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/counterfactuals through non-parametric structural equations
  - beyond the scope of this course

A note on acyclicness

- We impose acyclicness since a variable can't cause itself
  - e.g. my BMI today has no effect on my BMI today
- Observed variables are often snapshots of time varying processes
  - e.g. my BMI today certainly affects my BMI tomorrow
- Time varying processes can be depicted by explicitly adding one 'realization' of each variable per time unit
  - more later

$$\begin{tikzcd} A^0 \arrow[rr] \arrow[dr] & & A^1 \arrow[dr] & \\ & Y^0 \arrow[ur] \arrow[rr] & & Y^1 \end{tikzcd}$$

Underlying assumptions

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

- Assumptions are encoded by the direction of arrows
  - the arrow from A to Y means that A may affect Y, but not the other way around

Underlying assumptions, cont'd

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd} \qquad \begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A & & Y \end{tikzcd}$$

- Assumptions are encoded by the absence of arrows
  - the presence of an arrow from A to Y means that A may or may not affect Y
  - the absence of an arrow from A to Y means that A does not affect Y

Underlying assumptions, cont'd

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r] & Y \end{tikzcd}$$

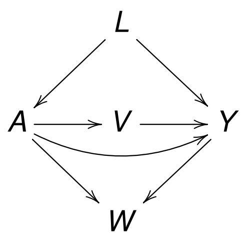
- Assumptions are encoded by the absence of common causes
  - the presence of L means that A and Y may or may not have common causes
  - the absence of L means that A and Y do not have any common causes

Ancestors and descendents

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

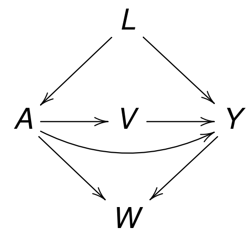
- The ancestors of a variable V are all other variables that affect V, either directly or indirectly
  - L is the single ancestor of A
- The descendents of a variable V are all other variables that are affected by V, either directly or indirectly
  - Y is the single descendent of A

Paths



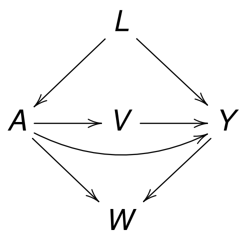
- A path is a route between two variables, not necessarily following the direction of arrows
- *Which are the paths between A and Y?*

Solution



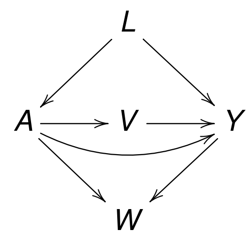
- Four paths between A and Y:
  - $A \rightarrow Y$
  - $A \rightarrow V \rightarrow Y$
  - $A \leftarrow L \rightarrow Y$
  - $A \rightarrow W \leftarrow Y$

Causal paths



- A causal path is a route between two variables, **following the direction of arrows**
  - the causal paths from A to Y mediate the causal effect of A on Y, the non-causal paths do not
- *Which are the causal paths between A and Y?*

Solution



- Two causal paths from A to Y:
  - $A \rightarrow Y$
  - $A \rightarrow V \rightarrow Y$

Blocking of paths

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

- Paths (both causal and non-causal) are either open or blocked, according to two rules

Motivating example DAG terminology Covariate selection in DAGs Motivating example, revisited Potential problems

Rule 1

- A path is blocked if somewhere along the path there is a variable L that sits in a ‘chain’

$$\longrightarrow L \longrightarrow$$

or in a ‘fork’

$$\longleftarrow L \longrightarrow$$

and we have adjusted for L

Motivating example DAG terminology Covariate selection in DAGs Motivating example, revisited Potential problems

Rule 2

- A path is blocked if somewhere along the path there is a variable L that sits in an ‘inverted fork’

$$\begin{tikzcd} \arrow[r] & L \arrow[d] & \arrow[l] \\ & V \arrow[d] & \\ & W & \end{tikzcd}$$

and we have **not** adjusted for L, or any of its descendents

Motivating example DAG terminology Covariate selection in DAGs Motivating example, revisited Potential problems

Once blocked stays blocked

$$A \longleftarrow V \longrightarrow W \longleftarrow Y$$

- Adjusting for V blocks the path from A to Y (rule 1)
- Adjusting for W leaves the path open (rule 2)
- Adjusting for both V and W blocks the path

Outline

Motivating example

DAG terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

Relation between ‘blocking’ and independence

- If all paths between A and Y are blocked, then A and Y are independent
- Conversely: if there is an associaton between A and Y, then there is at least one open path between A and Y

Example

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

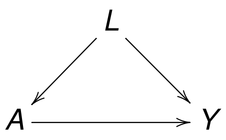
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
  - i.e. does the causal path $A \to Y$ exist?
- *Adjust or not adjust for L?*

Heuristic argument

$$\begin{tikzcd} & L \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

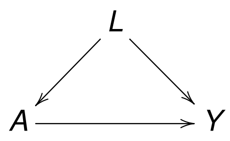
- A = smoking, Y = malformations, L = age
- Young mothers smoke more often, but their babies have smaller risk for malformations, than old mothers
- Hence, smokers are more likely to be young, and for this reason less likely to have babies with malformations, than non-smokers
- Thus, by not adjusting for age, we may observe an inverse association between smoking and malformations, even in the absence of a causal effect

Formal solution



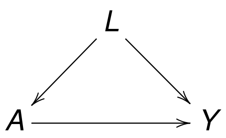
- Suppose that we don't adjust for L, and that we observe an association between A and Y
- There are two explanations for this association:
  - the causal path $A \to Y$
  - the open non-causal path $A \leftarrow L \to Y$ (Rule 1)
- Hence, an unadjusted association between A and Y does not prove that the causal path $A \to Y$ exists

Formal solution, cont'd



- Suppose that we adjust for L
  - we block the non-causal path $A \leftarrow L \to Y$ (Rule 1)
- Suppose that we observe an association between A and Y
  - this can only be explained by the causal path $A \to Y$
- Hence, an adjusted association between A and Y proves that there is a causal effect of A on Y

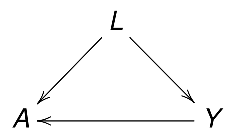
Conclusion



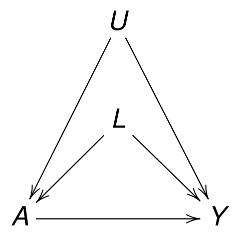
- If the aim is to test for a causal effect of A on Y, then we should adjust for L

Remark

- Adjusting for L does not give a causal effect if the DAG is incorrect, e.g. if
  - Y causes A



  - there are additional common causes of A and Y



Example

$$\begin{tikzcd} A \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & L & \end{tikzcd}$$

- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
  - i.e. does the causal path $A \rightarrow Y$ exist?
- *Adjust or not adjust for L?*

Heuristic argument

$$\begin{tikzcd} A \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & L & \end{tikzcd}$$

- A = smoking, Y = malformations, L = birth status (live/stillborn)
- Smoking and malformations increase the risk for stillbirth
- Consider the group of woman who has stillbirths: **what caused the stillbirths?**

Heuristic argument, cont'd

$$\begin{tikzcd} A \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & L & \end{tikzcd}$$

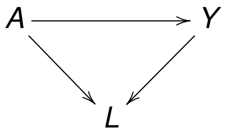
- For the non-smokers who had a stillbirth, smoking was obviously not the cause
  - perhaps malformations then?
- When smoking is ruled out as the cause of malformation, the likelihood of malformation increases
  - an inverse non-causal association between smoking and malformation!
- Thus, by adjusting for (e.g. stratifying on) birth status, we may observe an inverse association between smoking and malformations, even in the absence of a causal effect

Formal solution

$$\begin{tikzcd} A \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & L & \end{tikzcd}$$

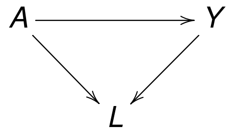
- Suppose that we adjust for L, and that we observe an association between A and Y
- There are two explanations for this association:
  - the causal path $A \rightarrow Y$
  - the open non-causal path $A \rightarrow L \leftarrow Y$ (Rule 2)
- Hence, an adjusted association between A and Y does not prove that the causal path $A \rightarrow Y$ exists

Formal solution, cont'd



- Suppose that we don't adjust for L
  - we block the non-causal path $A \rightarrow L \leftarrow Y$ (Rule 2)
- Suppose that we observe an association between A and Y
  - this can only be explained by the causal path $A \rightarrow Y$
- Hence, an unadjusted association between A and Y proves that there is a causal effect of A on Y

Conclusion

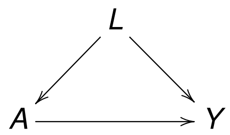


- If the aim is to test for a causal effect of A on Y, then we should not adjust for L

General strategy for covariate selection

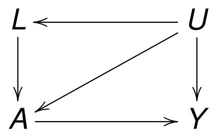
- We should adjust for those covariates that block non-causal paths between the exposure and the outcome
- We should not adjust for those covariates that open non-causal paths between the exposure and the outcome
- If we manage to block all non-causal paths, then any observed association must be due to a causal effect
- Thus, if all non-causal paths are blocked, then we have a valid test for causation

Confounding



- Common causes of the exposure and the outcome lead to non-causal paths
- We say that there is **confounding** if the exposure and the outcome have common causes

Confounder



- A **confounder** is a variable that blocks a non-causal path between the exposure and the outcome, if adjusted for
  - both L and U are confounders in the DAG above
- A (set of) variable(s) is **sufficient for confounding control** if the variable(s) blocks all non-causal paths
  - U is sufficient for confounding control, L is not

Outline

Motivating example

DAG terminology

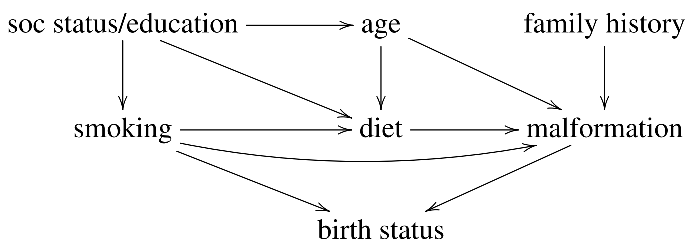
Covariate selection in DAGs

Motivating example, revisited

Potential problems

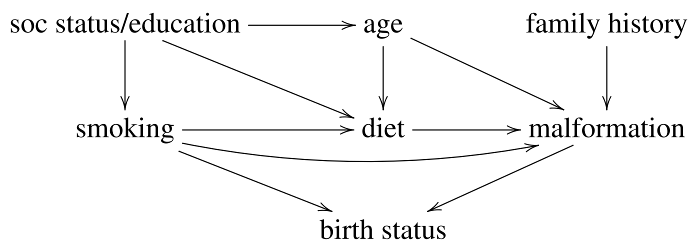
A possible DAG for the motivating example

- Suppose we agree that the causal structures for our data can be described by the DAG below



- *Which assumptions are encoded in this DAG?*
- *Can these assumptions be tested?*

Covariate selection



- *Given the DAG, which covariates should we adjust for?*
- *Which covariates would be selected by the traditional strategies?*

Outline

Motivating example

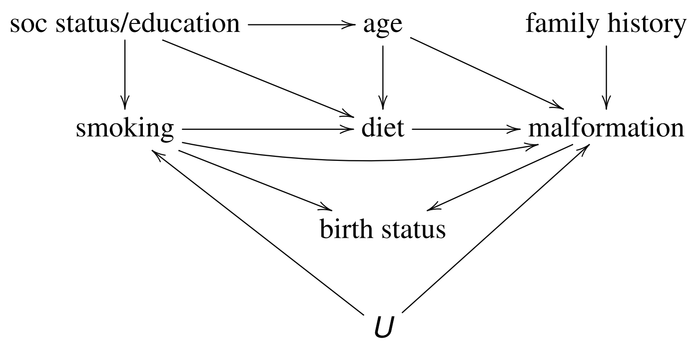
DAG terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

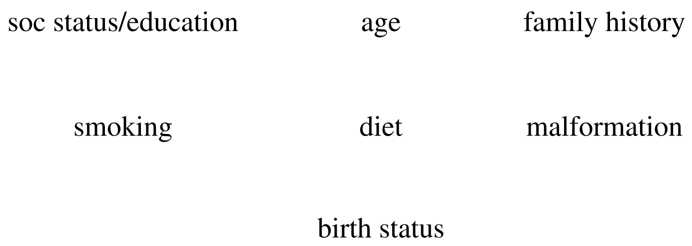
Unmeasured confounding



- Not a problem with DAGs, but with observational studies
- Try to reduce confounding bias as much as possible
  - i.e. block as many non-causal paths as possible

No *a priori* knowledge

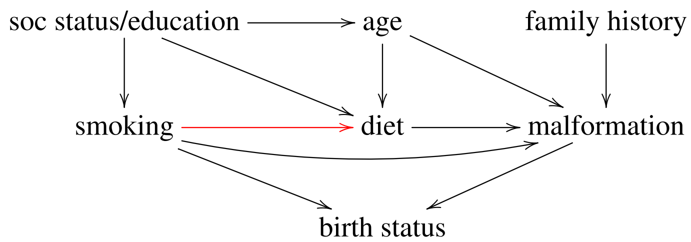
- Cannot construct a plausible DAG

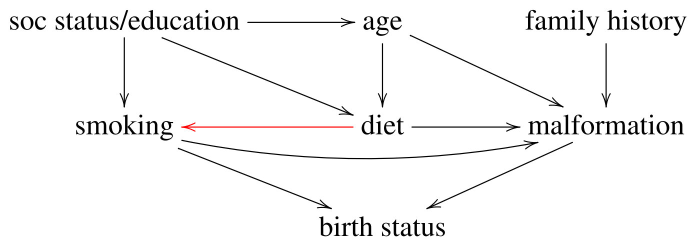


- DAG-based covariate selection cannot be used, and we have to resort to traditional strategies
  - but be aware of the pitfalls

Weak *a priori* knowledge

- Cannot settle with **one** plausible DAG

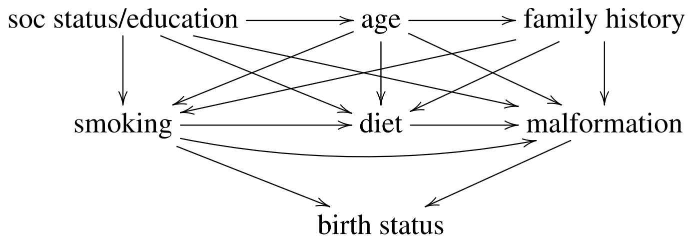




- Present all plausible DAGs, and the implied analyses

A complicated DAG

- No/little covariate reduction



- But remember that
  - more covariates requires a bigger model, with a higher potential for bias due to model misspecification
  - some covariates may be prone to measurement errors, and may therefore lead to bias
  - some covariates may reduce statistical power/efficiency when adjusted for
- It may sometimes be reasonable to exclude covariates with a weak 'confounding effect'

Summary

- Traditional covariate selection strategies
  - are difficult to apply at the design stage
  - may select non-confounders, which may increase bias if adjusted for
- DAGs can be used for covariate selection
  - encode our *a priori* causal knowledge/beliefs into a DAG
  - adjust for those covariates that block non-causal paths between the exposure and the outcome
- DAGs are not only tools for covariate selection
  - generally speaking, they are used to facilitate interpretation and communication in causal inference