Directed Acyclic Graphs

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

A short course on concepts and methods in Causal Inference

Ideal randomized trials

- In ideal randomized trials, exposed and unexposed are exchangeable:

$$(Y_0, Y_1) \perp\!\!\!\perp A$$

- As a consequence, association = causation

Observational studies

- In observational studies, exchangeability is often implausible
- We may achieve conditional exchangeability by adjusting for an appropriate set of covariates:

$$(Y_0, Y_1) \perp\!\!\!\perp A \mid L$$

- But selecting an appropriate set of covariates to adjust for is a non-trivial task

Motivating example

- Does smoking during pregnancy (exposure) causes malformations (outcome) in the offspring?
- For a large number of pregnancies, we collect data on both exposure and outcome
- We record five additional covariates:
  - the mothers age at conception
  - the mothers socioeconomic status at conception
  - the mothers diet during pregnancy
  - indicator of whether there is a family history of birth defects
  - indicator of whether the child was liveborn or stillborn

Motivating example, cont'd

- We observe an unadjusted inverse association between smoking and malformations; risk ratio = 0.8
- However, we suspect that there is confounding of the exposure and outcome
  - if so, exposed and unexposed are not exchangeable, and
  - the observed risk ratio cannot be given a causal interpretation
- To reduce confounding bias we want to adjust for observed covariates

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 non-exchangeability 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, say, 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 non-exchangeability if adjusted for

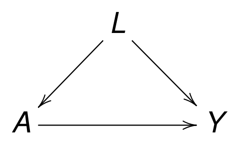
Covariate selection with DAGs

- 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

Outline

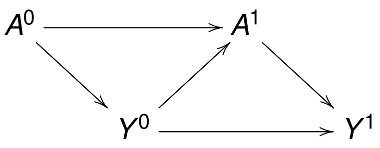
A simple DAG



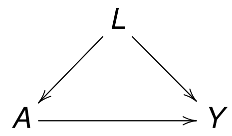
- Each arrow represents a causal influence
- The graph is
  - Directed, since each connection between two variables consists of an arrow
  - Acyclic, since the graph contains no directed cycles
- Formal connection to potential outcomes/counterfactuals through non-parametric structural equations
  - beyond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

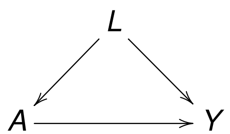


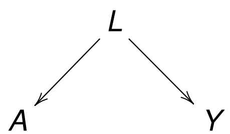
Underlying assumptions



- 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

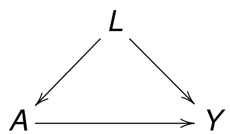
Underlying assumptions, cont'd

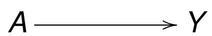




- Assumptions are encoded by the absence of arrows
  - the presence of an arrow from A to Y means that A may or may not affect Y
  - the absence of an arrow from A to Y means that A does not affect Y

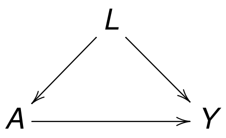
Underlying assumptions, cont'd





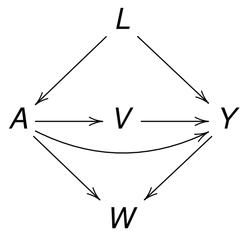
- Assumptions are encoded by the absence of common causes
  - the presence of L means that A and Y may or may not have common causes
  - the absence of L means that A and Y do not have any common causes

Ancestors and descendents



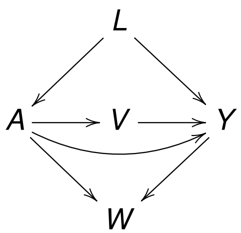
- The ancestors of a variable V are all other variables that affect V, either directly or indirectly
  - L is the single ancestor of A
- The descendents of a variable V are all other variables that are affected by V, either directly or indirectly
  - Y is the single descendent of A

Paths



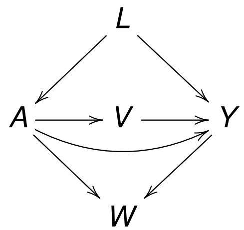
- A path is a route between two variables, not necessarily following the direction of arrows
- *Which are the paths between A and Y?*

Solution



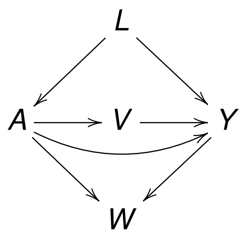
- Four paths between A and Y:
  - $A \rightarrow Y$
  - $A \rightarrow V \rightarrow Y$
  - $A \leftarrow L \rightarrow Y$
  - $A \rightarrow W \leftarrow Y$

Causal paths



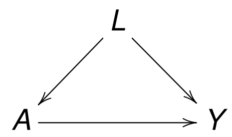
- A causal path is a route between two variables, **following the direction of arrows**
  - the causal paths from A to Y mediate the causal effect of A on Y, the non-causal paths do not
- *Which are the causal paths between A and Y?*

Solution



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

Blocking of paths



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

Rule 1

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

$$\longrightarrow L \longrightarrow$$

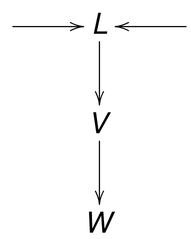
or in a 'fork'

$$\longleftarrow L \longrightarrow$$

and we have adjusted for L

Rule 2

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



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

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

Relation between ‘blocking’ and independence

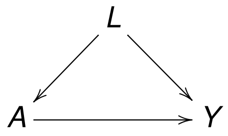
- If all paths between A and Y are blocked, then A and Y are independent
- Conversely: if there is an associaton between A and Y, then there is at least one open path between A and Y

Example

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

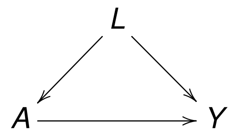
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
  - i.e. does the causal path $A \to Y$ exist?
- *Adjust or not adjust for L?*

Heuristic argument



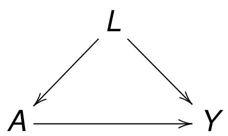
- A = smoking, Y = malformations, L = age
- Young mothers smoke more often, but their babies have smaller risk for malformations, than old mothers
- Hence, smokers are more likely to be young, and for this reason less likely to have babies with malformations, than non-smokers
- Thus, by not adjusting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

Formal solution



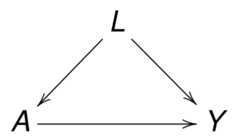
- Suppose that we don't adjust for L, and that we observe an association between A and Y
- There are two explanations for this association:
  - the causal path $A \rightarrow Y$
  - the open non-causal path $A \leftarrow L \rightarrow Y$ (Rule 1)
- Hence, an unadjusted association between A and Y does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

Formal solution, cont'd



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

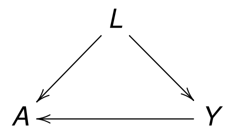
Conclusion



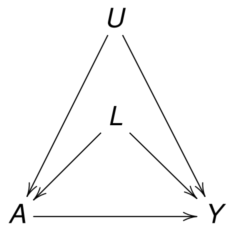
- If the aim is to test for a causal effect of A on Y, then we should adjust for L

Remark

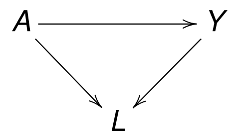
- Adjusting for L does not give a causal effect if the DAG is incorrect, e.g. if
  - Y causes A



  - there are additional common causes of A and Y

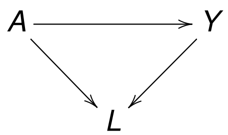


Example



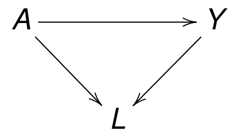
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
  - i.e. does the causal path $A \to Y$ exist?
- *Adjust or not adjust for L?*

Heuristic argument



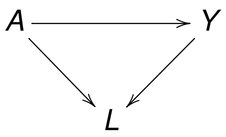
- A = smoking, Y = malformations, L = birth status (live/stillborn)
- Smoking and malformations increase the risk for stillbirth
- Consider the group of woman who has stillbirths: **what caused the stillbirths?**

Heuristic argument, cont'd



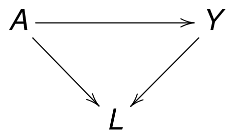
- For the non-smokers who had a stillbirth, smoking was obviously not the cause
  - perhaps malformations then?
- When smoking is ruled out as the cause of malformation, the likelihood of malformation increases
  - an inverse non-causal association between smoking and malformation!
- Thus, by adjusting 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



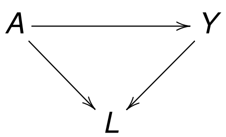
- Suppose that we adjust for L, and that we observe an association between A and Y
- There are two explanations for this association:
  - the causal path $A \rightarrow Y$
  - the open non-causal path $A \rightarrow L \leftarrow Y$ (Rule 2)
- Hence, an adjusted association between A and Y does not prove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

Relation between ‘blocking’ and exchangeability

- If all non-causal paths are blocked, then exposed and unexposed are typically exchangeable
- Thus, the observed association can typically be interpreted as a causal effect
  - e.g. the (conditional) risk ratio is equal to the (conditional) causal risk ratio

Examples revisited

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

- In the left DAG we have conditional exchangeability, given L:

$$(Y_0, Y_1) \amalg A \mid L$$

  so that the conditional risk ratio, given L, is equal to the conditional causal risk ratio, given L
- In the right DAG, we have marginal exchangeability:

$$(Y_0, Y_1) \amalg A$$

  so that the marginal risk ratio is equal to the marginal causal risk ratio

Technical note

- If all non-causal paths are blocked, then exposed and unexposed are typically exchangeable
- **But it is possible to construct counterexamples**

$$A \longrightarrow Y \longrightarrow L$$

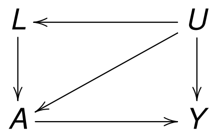
- If we adjust for L in the DAG above, then all non-causal paths between A and Y are blocked
  - there are no non-causal paths to start with
- Thus, a conditional association between A and Y proves that there is a causal effect of A on Y
- However, adjusting for L does not give exchangeability
  - e.g. the conditional risk ratio, given L, is not equal to the conditional causal risk ratio, given L
- Adjusting for L gives a valid test, but not a valid estimate

Confounding

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

- 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

DAG terminology

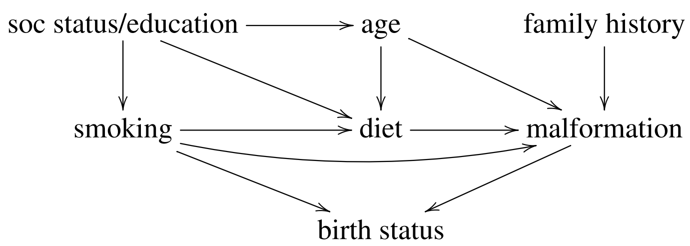
Covariate selection in DAGs

Motivating example, revisited

Potential problems

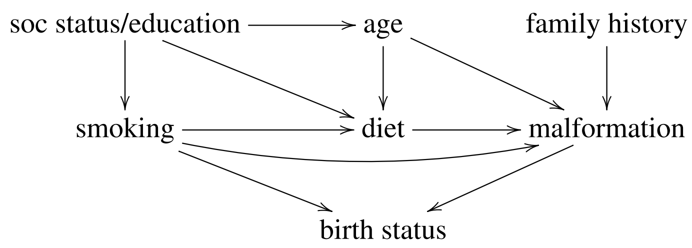
A possible DAG for the motivating example

- 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?*

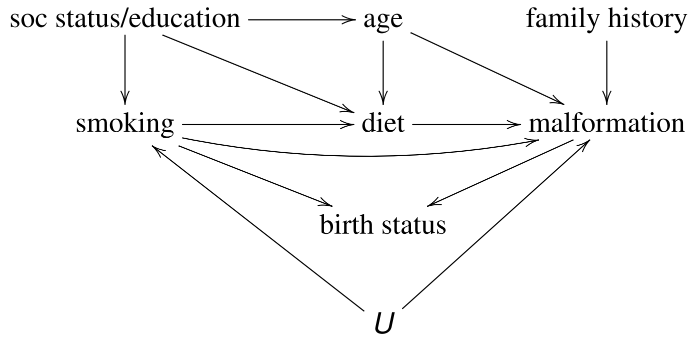
Covariate selection



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

Outline

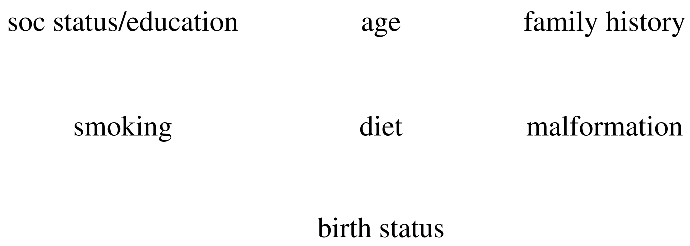
Unmeasured confounding



- 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

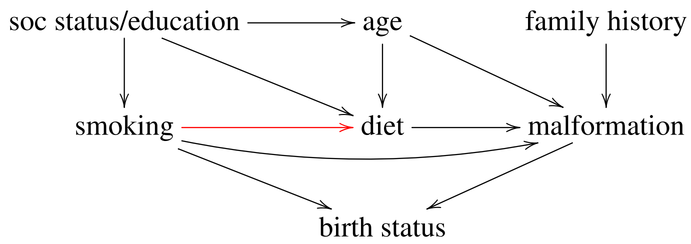
- Cannot construct a plausible DAG

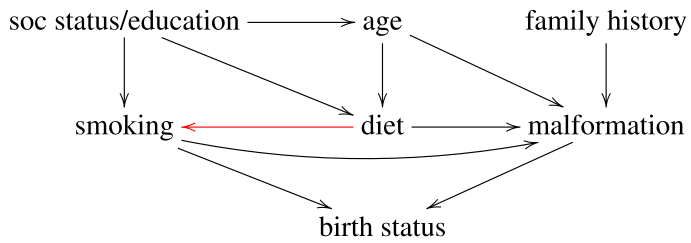


- DAG-based covariate selection cannot be used, and we have to resort to traditional strategies
  - but be aware of the pitfalls

Weak *a priori* knowledge

- Cannot settle with **one** plausible DAG

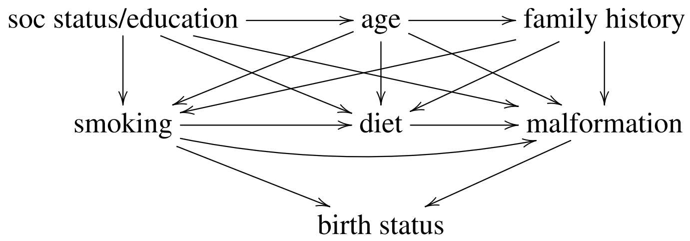




- Present all plausible DAGs, and the implied analyses

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 non-exchangeability
- 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