

Causal Inference in Epidemiology

(Directed Acyclic Graphs)

Rino Bellocco^{1,2}, Sc.D.
&
Arvid Sjolander¹, Ph.D.

^{1,2} Department of Statistics and Quantitative Methods
University of Milano-Bicocca

&

²Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Causal Inference in Epidemiology
Sismec Working Group
Milan, December 12, 2016

Ideal randomized trials

- Exposed and unexposed are exchangeable:

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

- Association = causation:

$$RR = CRR$$

Observational studies

- Exchangeability is often implausible
- We may consider exchangeability more plausible if we stratify on some set of covariates

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

$$RR|L = CRR|L$$

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

Motivating Example

- Consider an observational study aimed to investigate whether smoking during pregnancy (exposure) causes malformations (outcome) in newborns
- For a large number of pregnancies, we collect data on both exposure and outcome
- Investigators recorded five additional covariates:
 - Mother's age at conception
 - Mother's socioeconomic status at conception
 - Mother's diet during pregnancy
 - Family history of birth defects
 - Status at birth (liveborn or stillborn)

Motivating example, cont'd

- We observe an unadjusted inverse association between smoking and malformations ($RR = 0.8$)
- 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 bias due to confounding we want to adjust for observed covariates

Covariate selection

- One strategy would be to adjust for all observed covariates
- This approach may
 - increase non-exchangeability, if the covariates are not confounders
 - increase statistical uncertainty (e.g. wider confidence intervals)
- Therefore, it is desirable to select a subset of covariates to adjust for
- What covariate selection strategy should we use?

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

Problem with traditional strategies

- They rely on statistical analyses of observed data, rather than *a priori* knowledge about causal structures
- Two important implications
 - they 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

Directed Acyclic Graphs

- Directed Acyclic Graphs (DAG's) can be used to overcome the problems with traditional covariate selection strategies
- A DAG is a graphical representation of underlying causal structures
- DAG's are a way to be explicit about your assumptions about the causal structure of the relationships between exposures, potential confounders and their consequences
- When there is uncertainty about the correct structure of the DAG, it is useful to conduct sensitivity analyses to examine how different assumptions impact the estimated causal effect

Directed Acyclic Graphs

- DAGs for covariate selection:
 - Encode our *a priori* causal knowledge/beliefs into a DAG
 - The proposed DAG may or may not correctly represent the true state of nature. The analysis that is dictated by a particular DAG estimates the causal effect of exposure only if the DAG is correct
 - apply simple graphical rules to determine what covariates to adjust for

Outline

Graph terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

Subtle points

Outline

Graph terminology

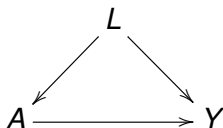
Covariate selection in DAGs

Motivating example, revisited

Potential problems

Subtle points

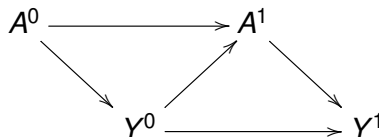
A simple DAG



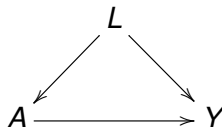
- Each arrow represents a causal influence
- The graph is
 - Directed: Time flows from left to right via directed, since each connection between two variables consists of an arrow
 - Acyclic, since the graph contains no directed cycles
- Formal connection to potential outcomes/counterfactuals through non-parametric structural equations
 - beyond the scope of this course

A note on acyclicity

- We impose acyclicity since a variable cannot 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

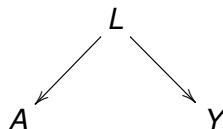
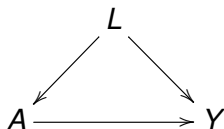


Underlying assumptions



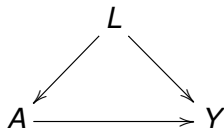
- Assumptions are encoded by the direction of arrows
 - The arrow from A to Y means that A may affect Y , but not the other way around

Underlying assumptions, cont'd



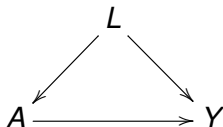
- Assumptions are encoded by the absence of arrows
 - The presence of an arrow from A to Y means that A may or may not affect Y
 - The absence of an arrow from A to Y means that A does not affect Y

Underlying assumptions, cont'd



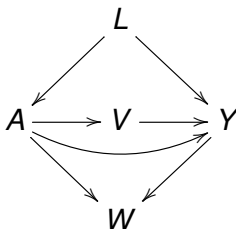
- Assumptions are encoded by the absence of common causes
 - The presence of L means that A and Y may or may not have common causes
 - The absence of L means that A and Y do not have any common causes

Ancestors and descendants



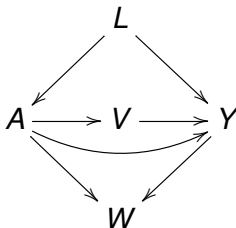
- The ancestors of a variable V are all other variables that affect V , either directly or indirectly
 - L is the single ancestor of A
- The descendants of a variable V are all other variables that are affected by V , either directly or indirectly
 - Y is the single descendant of A

Paths



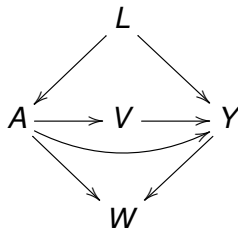
- A path is a route between two variables, not necessarily following the direction of arrows
- A path can be conceptualized as pipes or wires that carry the flow of association
- *Which are the paths between A and Y?*

Solution



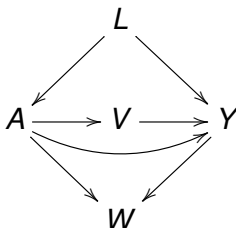
- Four paths between A and Y :
 - $A \rightarrow Y$
 - $A \rightarrow V \rightarrow Y$
 - $A \leftarrow L \rightarrow Y$
 - $A \rightarrow W \leftarrow Y$

Causal paths



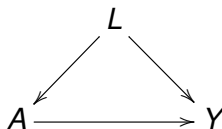
- A causal path is a route between two variables, **following the direction of arrows**
 - The causal paths from A to Y mediate the causal effect of A on Y , the non-causal paths do not
- *Which are the causal paths between A and Y ?*

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’

$$\longrightarrow L \longrightarrow$$

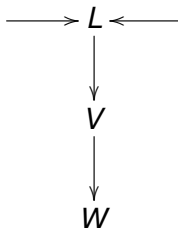
or in a ‘fork’

$$\longleftarrow L \longrightarrow$$

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

Graph terminology

Covariate selection in DAGs

Motivating example, revisited

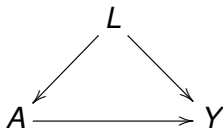
Potential problems

Subtle points

d-separation

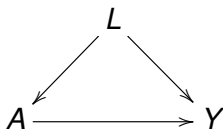
- If all paths between A and Y are blocked by adjusting for L , then A and Y are conditionally independent, given L
 - In graph terminology; A and Y are d-separated by L
- Conversely: if at least one path is open, then A and Y are (most likely) conditionally associated, given L

Example



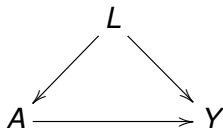
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
 - I.e. does the causal path $A \rightarrow Y$ exist?
- *Concrete example?*
- *Adjust or not adjust for L ?*

Heuristic argument



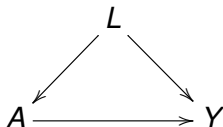
- A = smoking, Y = malformations, L = age
- Young mothers smoke more often, but their babies have smaller risk for malformations, than old mothers
- Hence, smokers are more likely to be young, and for this reason less likely to have babies with malformations, than old mothers
- Thus, by not adjusting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

Formal solution based on d-separation



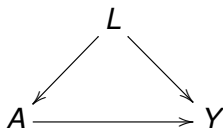
- Suppose that we don't adjust for L , and that we observe an association between A and Y
- There are two explanations for this association:
 - The causal path $A \rightarrow Y$
 - The open non-causal path $A \leftarrow L \rightarrow Y$ (Rule 1)
- Hence, an unadjusted association between A and Y does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

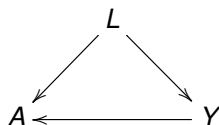
Conclusion



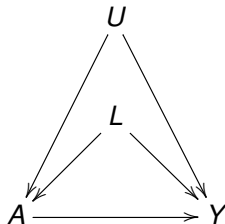
- If the aim is to test for a causal effect of A on Y , then we should adjust for L
- In terms of potential outcomes:
 - exposed and unexposed are not marginally exchangeable
 - exposed and unexposed are conditionally exchangeable, given L

Remark

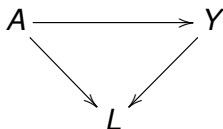
- The argument assumes that the DAG is correct, in particular that
 - Y does not cause A



- There is no additional common cause of A and Y

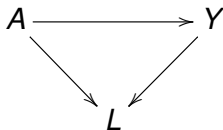


Example



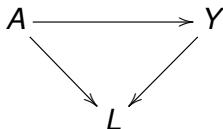
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
 - I.e. does the causal path $A \rightarrow Y$ exist?
- *Concrete example?*
- *Adjust or not adjust for L ?*

Heuristic argument



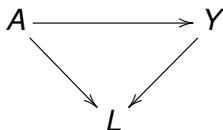
- A = smoking, Y = malformations, L = birth status (live/stillborn)
- Smoking and malformations increase the risk for stillbirth
- Consider the group of woman who has stillbirths, but do not smoke
 - There must be another reason for the stillbirth, presumably that the baby had malformations
- Thus, by 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 based on d-separation



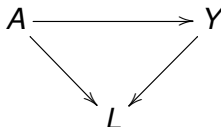
- Suppose that we adjust for L , and that we observe an association between A and Y
- There are two explanations for this association:
 - The causal path $A \rightarrow Y$
 - The open non-causal path $A \rightarrow L \leftarrow Y$ (Rule 2)
- Hence, an adjusted association between A and Y does not prove 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
- In terms of potential outcomes:
 - exposed and unexposed are marginally exchangeable
 - exposed and unexposed are not conditionally exchangeable, given L

QUESTION

True or False: In DAGs, a causal effect is represented by arrows connecting variables, and the direction of the arrows is ignored

- ☐ True
- ☐ False

QUESTION



In the DAG above,

- 1 Diet soda and cancer risk are unconditionally independent
- 2 Diet soda causes cancer
- 3 Diet soda and cancer risk are independent after conditioning on BMI
- 4 The effect of diet soda consumption on cancer risk is mediated by body mass index

QUESTION

AIR POLLUTION \longrightarrow OXIDATIVE STRESS \longrightarrow ASTHMA

In the DAG above,

- 1 Adjusting for oxidative stress is necessary to study the causal effect of air pollution on asthma risk
- 2 Adjusting for oxidative stress blocks the path between air pollution and asthma
- 3 There are several pathways linking air pollution to asthma risk
- 4 Adjusting for oxidative stress opens the path between air pollution and asthma and induces a bias

Outline

Graph terminology

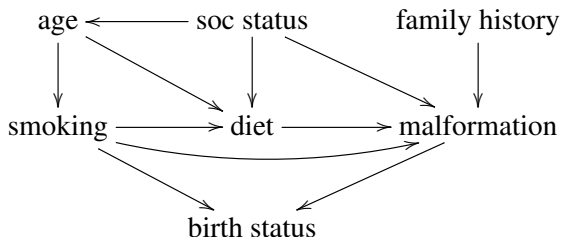
Covariate selection in DAGs

Motivating example, revisited

Potential problems

Subtle points

Covariate selection



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

Outline

Graph terminology

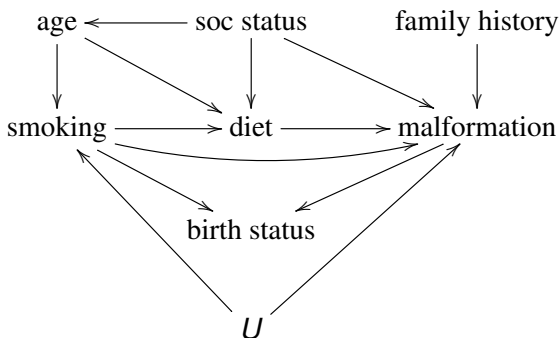
Covariate selection in DAGs

Motivating example, revisited

Potential problems

Subtle points

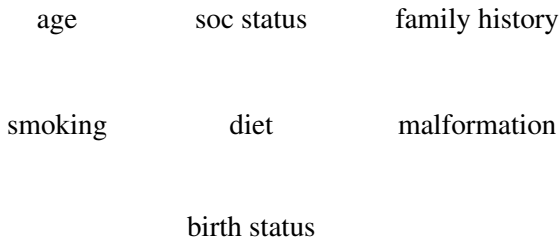
Unmeasured confounding



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

No *a priori* knowledge

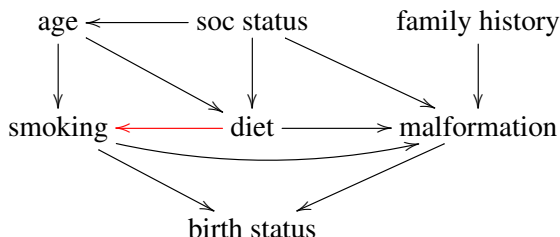
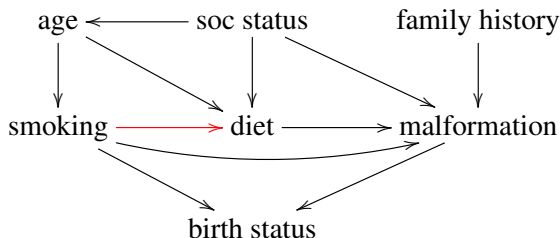
- Cannot construct a plausible DAG



- 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

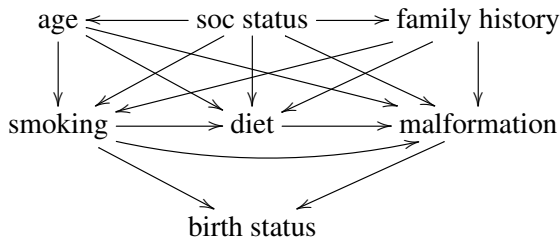
- Cannot settle with **one** plausible DAG



- Present all plausible DAGs, and the implied analyses

A complicated DAG

- No/little covariate reduction



- For every confounder that we adjust for, we may
 - decrease the bias due to confounding, but
 - increase the statistical uncertainty (e.g. wider confidence intervals)
- A reasonable trade off: exclude covariates with a relatively weak ‘confounding effect’

Outline

Graph terminology

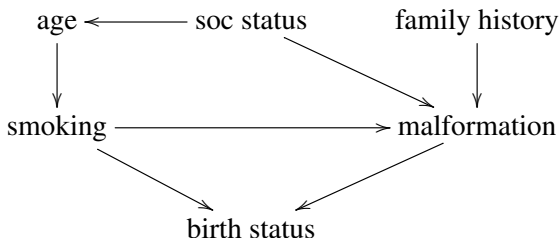
Covariate selection in DAGs

Motivating example, revisited

Potential problems

Subtle points

Confounding vs confounder

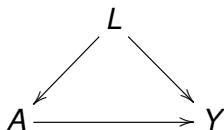


- We have **confounding** due to the non-causal path $\text{smoking} \leftarrow \text{age} \leftarrow \text{soc status} \rightarrow \text{malformation}$
 - **Confounding** means that the exposure and the outcome have common causes
- The path can be blocked by adjusting for either soc status (the common cause) or age
 - A **confounder** is any variable that can be used to block a non-causal path, not necessarily the common cause

Testing vs estimation

- We have learned how to use DAGs for covariates selection, given that we want to **test** for a causal effect
 - i.e. ‘is there an effect, or is there no effect?’
- Often, our aim is on **estimation**
 - i.e. ‘what is the magnitude of the effect?’

Examples revisited

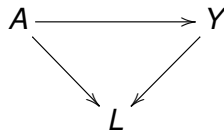


- In the left DAG, the conditional (given L) $RD/RR/OR$ equals the **causal** conditional $RD/RR/OR$

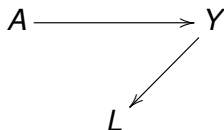
$$RD|L = CRD|L, \quad RR|L = CRR|L, \quad OR|L = COR|L$$

- In the right DAG, the marginal $RD/RR/OR$ equals the **causal** marginal $RD/RR/OR$

$$RD = CRD, \quad RR = CRR, \quad OR = COR$$



A valid test does not imply a valid estimate



- In this DAG, adjusting for L is not necessary
- Adjusting for L does not ‘invalidate’ the test
 - Only one path, $A \rightarrow Y$, which is causal
- Adjusting for L ‘invalidates’ the RD and the RR

$$RD|L \neq CRD|L, \quad RR|L \neq CRR|L$$

but not the OR

$$OR|L = COR|L$$

Summary

- Traditional covariate selection strategies do not use *a priori* knowledge about causal structures
 - 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
 - apply simple graphical rules to determine what covariates to adjust for
- DAGs are not only tools for covariate selection
 - Generally speaking, they are used to facilitate interpretation and communication in causal inference