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^{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?

An important milestone

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

Outline

Outline

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

- We say that mothers age is a ‘confounder’ that may induce non-causal associations between SDP and malformations
- To reduce the amount of confounding bias we may wish to adjust for mothers age in the analysis
  - e.g. by stratification, regression modelling or propensity scores

Measured covariates

- 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

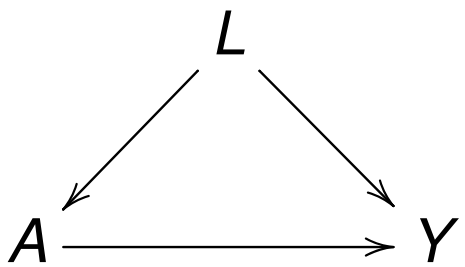
- They rely on statistical analyses of observed data, rather than *a priori* knowledge about causal structures
  - require that data is already collected, and cannot not be used at the design stage
- **They may select non-confounders, which may increase bias if adjusted for**

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

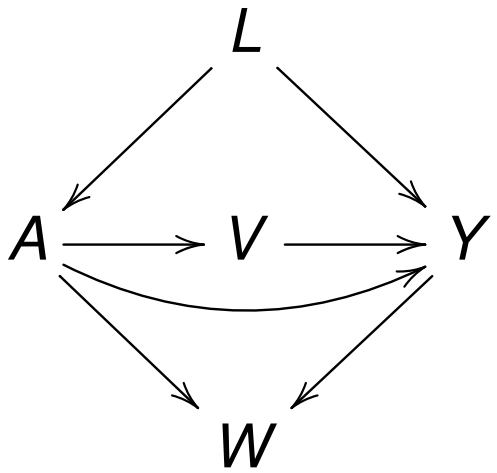
Outline

A simple DAG



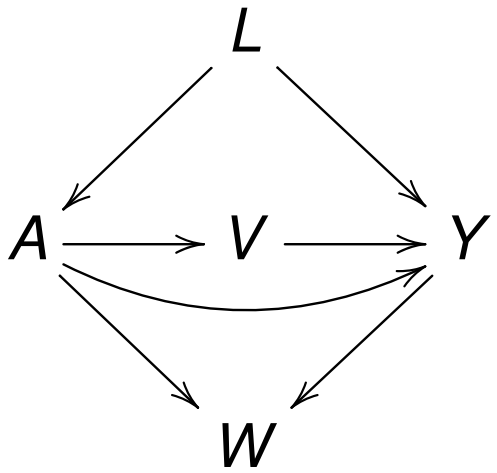
- Each arrow represents a causal influence
- The graph is
  - Directed, since each connection between two variables consists of an arrow
  - Acyclic, since the graph contains no directed cycles

Paths



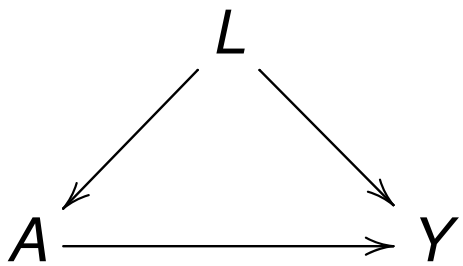
- A path is a route between two variables, not necessarily following the direction of arrows
- Four paths between A and Y:
  - $A \rightarrow Y$
  - $A \rightarrow V \rightarrow Y$
  - $A \leftarrow L \rightarrow Y$
  - $A \rightarrow W \leftarrow Y$

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
- 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'

  $$\longrightarrow L \longrightarrow$$

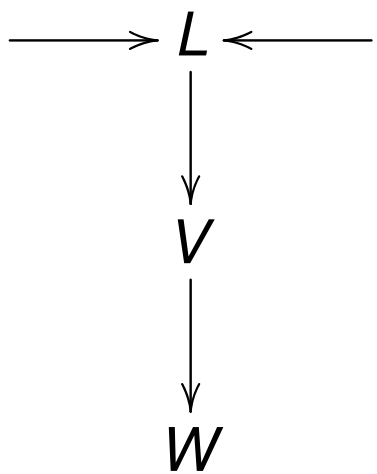
  or in a 'fork'

  $$\longleftarrow L \longrightarrow$$

  and we have adjusted for L

Rule 2

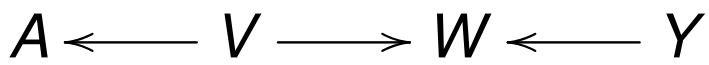
- A path is blocked if somewhere along the path there is a variable L that sits in an 'inverted fork'



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

  - The descendents of L are all variables affected by L

Once blocked stays blocked



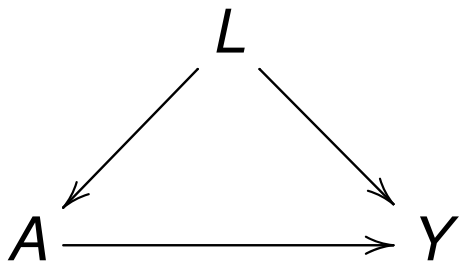
- Adjusting for V blocks the path from A to Y (rule 1)
- Adjusting for W leaves the path open (rule 2)
- Adjusting for both V and W blocks the path

Outline

Relation between 'blocking' and independence

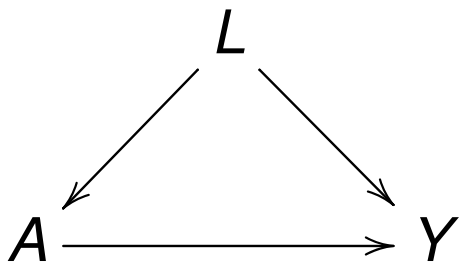
- If all paths between A and Y are blocked, then A and Y are independent
- Conversely: if there is an associaton between A and Y, then there is at least one open path between A and Y

Example



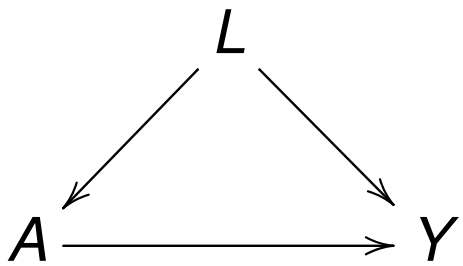
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
  - i.e. does the causal path $A \rightarrow Y$ exist?
- *Adjust or not adjust for L?*

Solution



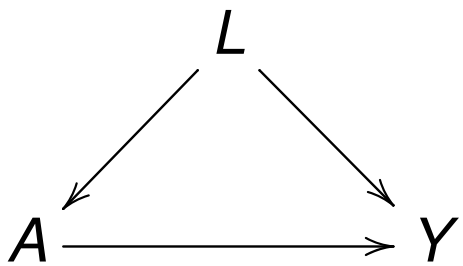
- Suppose that we don't adjust for L, and that we observe an association between A and Y
- There are two explanations for this association:
  - the causal path $A \rightarrow Y$
  - the open non-causal path $A \leftarrow L \rightarrow Y$ (Rule 1)
- Hence, an unadjusted association between A and Y does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

Solution, cont'd



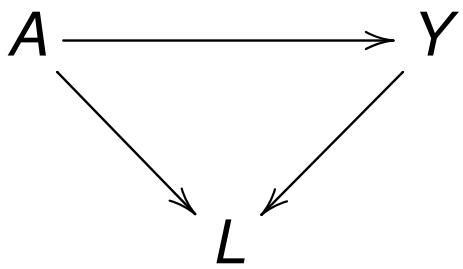
- Suppose that we adjust for L
  - we block the non-causal path $A \leftarrow L \rightarrow Y$ (Rule 1)
- Suppose that we observe an association between A and Y
  - this can only be explained by the causal path $A \rightarrow Y$
- Hence, an adjusted association between A and Y proves that there is a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

Conclusion



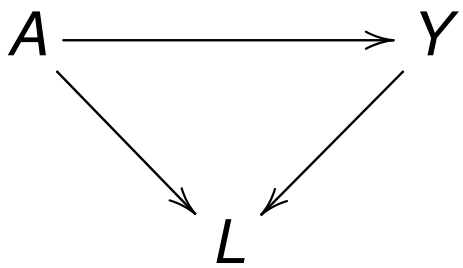
- If the aim is to test for a causal effect of A on Y, then we should adjust for L

Example



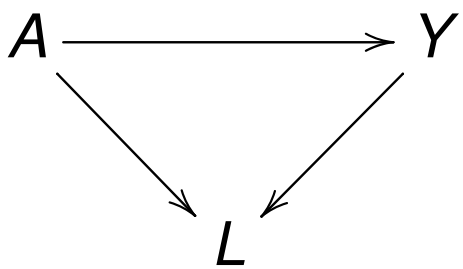
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of A on Y
  - i.e. does the causal path $A \rightarrow Y$ exist?
- *Adjust or not adjust for L?*

Solution



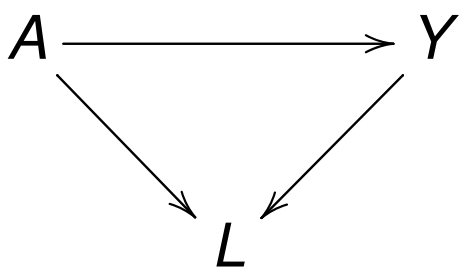
- Suppose that we adjust for L, and that we observe an association between A and Y
- There are two explanations for this association:
  - the causal path $A \rightarrow Y$
  - the open non-causal path $A \rightarrow L \leftarrow Y$ (Rule 2)
- Hence, an adjusted association between A and Y does not prove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

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

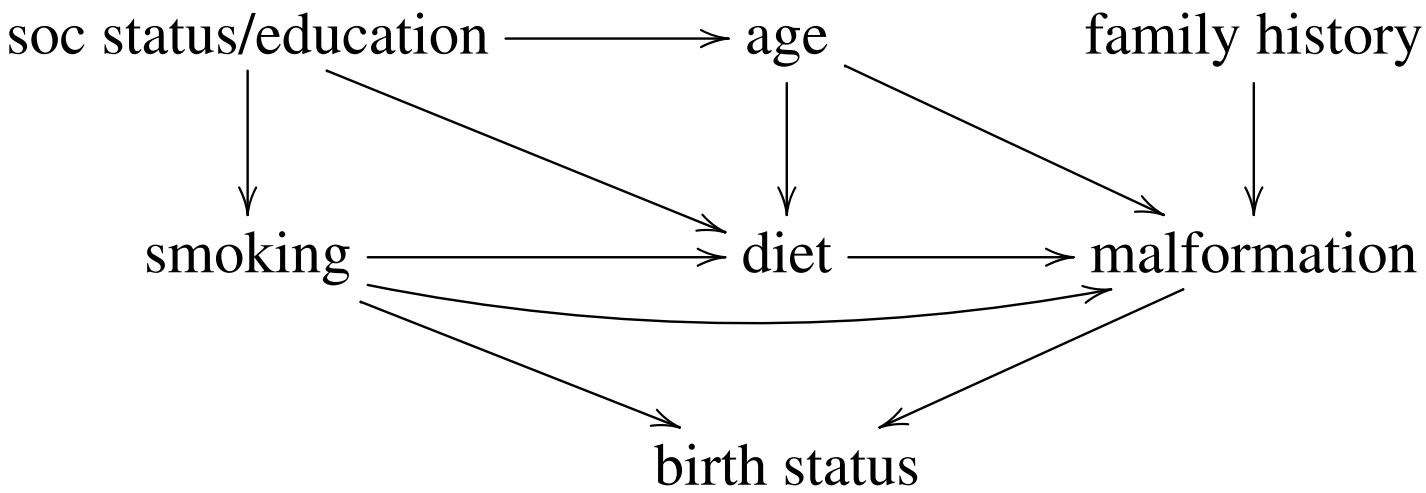
DAG strategy for covariate selection

- 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

Outline

A possible DAG for the motivating example

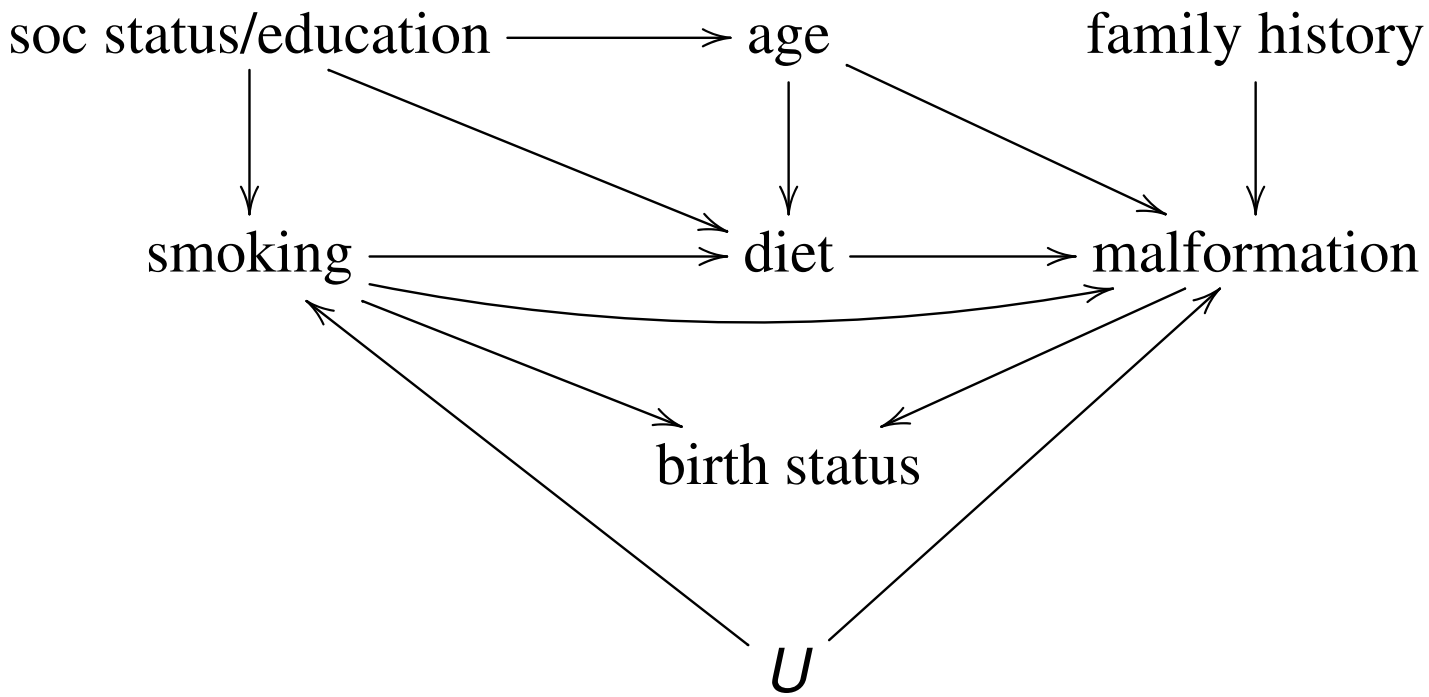
- Suppose we agree that the causal structures for our data can be described by the DAG below



- *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

soc status/education age family history

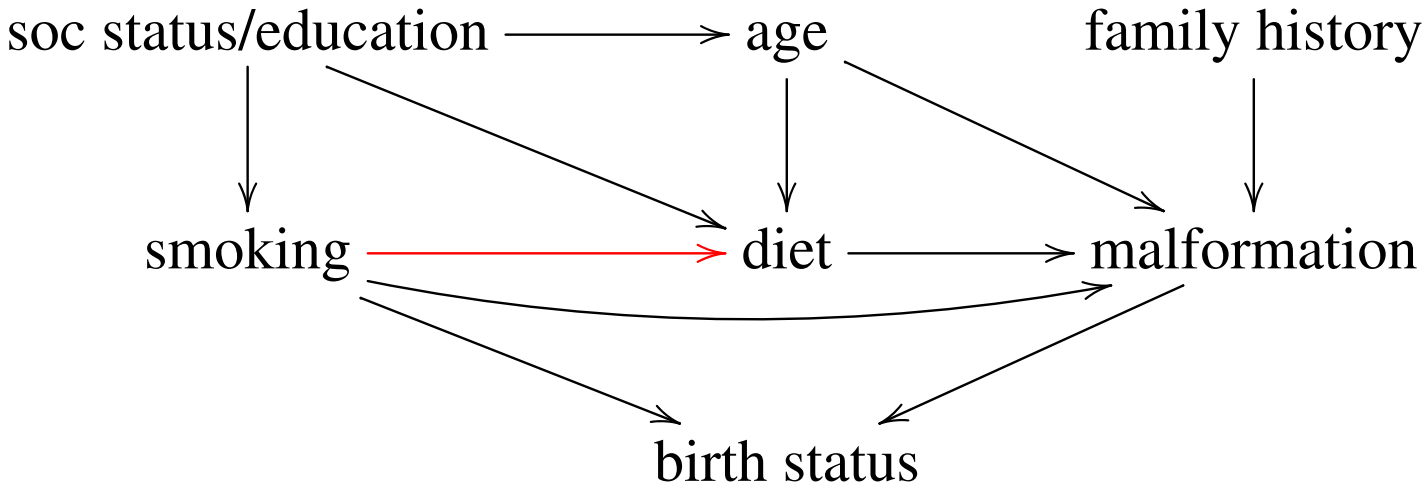
smoking diet malformation

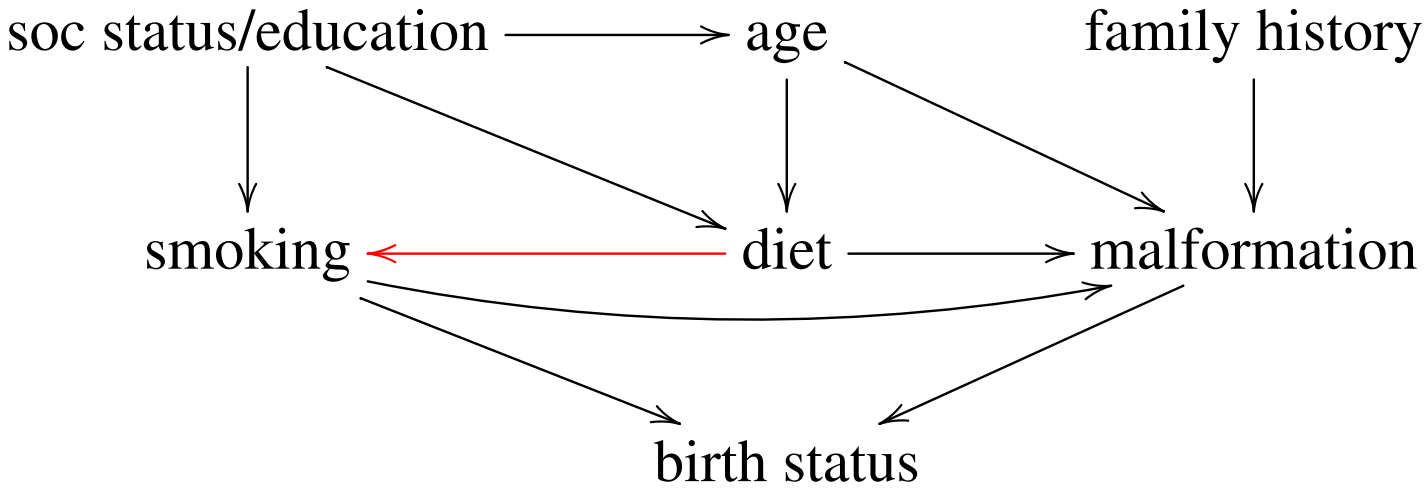
birth status

- 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

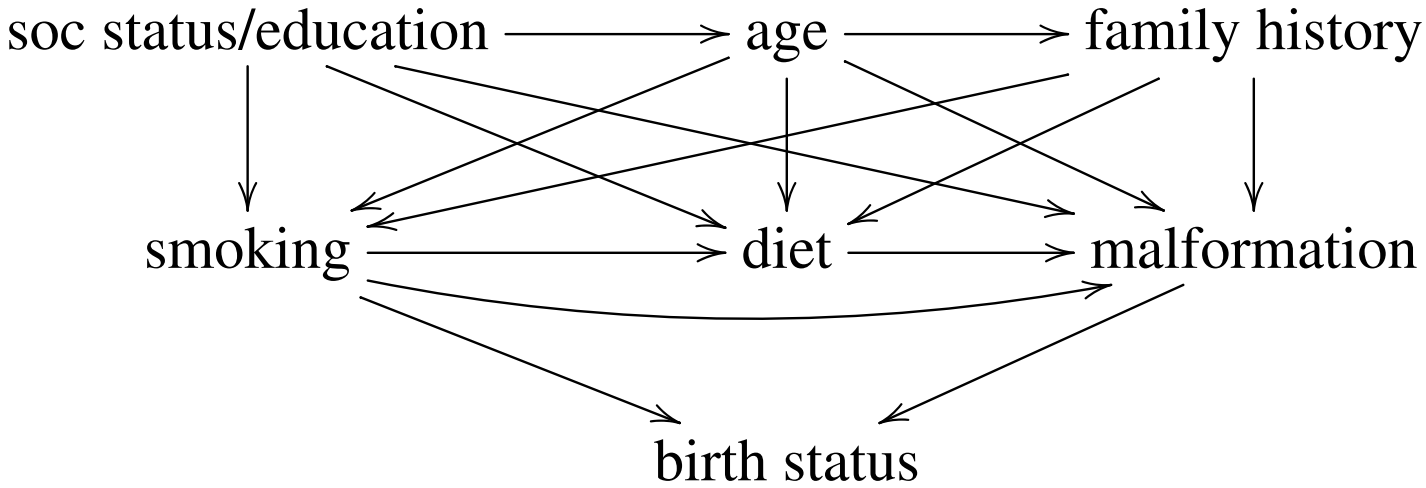




- Present all plausible DAGs, and the implied analyses

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