Causal diagrams for the design and analysis of epidemiological studies

Costanza Pizzi

Università degli Studi di Torino

Motivation

- The field of causal inference consists of three main parts:
  1. A **formal language** for unambiguously defining causal concepts.
  2. **Causal diagrams**: a tool for clearly displaying our causal assumption, useful for both design and analyses of epidemiological studies.
  3. **Statistical methods** to draw more reliable conclusions from the data at hand.
- In this lecture, we focus on 2.

Motivation

- Much work in epidemiology aims at identifying biological and behavioral **causes of diseases**
- From a public health perspective is also vital the assessment of **causal effects of interventions**, e.g. changing health policy, approving new drugs...

▷ **...so that optimal prevention strategies can be devised.**

Motivation

- Causal inference is the science of inferring the presence and magnitude of cause-effect relationships from data.
- **Association = causation** $\Longleftrightarrow$ if there are no source of bias.
- Thus **RCTs** represent the ideal study design to provide estimates that can be endowed with a causal interpretation
- However for ethical and practical reasons we often use **observational studies** to answer etiological questions $\Longrightarrow$ **confounding**

Motivation

- Thus the goal is to **identify a set of covariates that minimizes confounding**
- This requires background subjects-matter knowledge
- **Causal diagrams help us to organize this knowledge and identify whether or not confounding is present.**

Outline

Motivating example

- Consider an observational study 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
- We record information on four additional covariates:
  - mothers age at conception
  - mothers socioeconomic status at conception
  - family history of birth defects
  - indicator of whether the baby was liveborn or stillborn

Motivating example

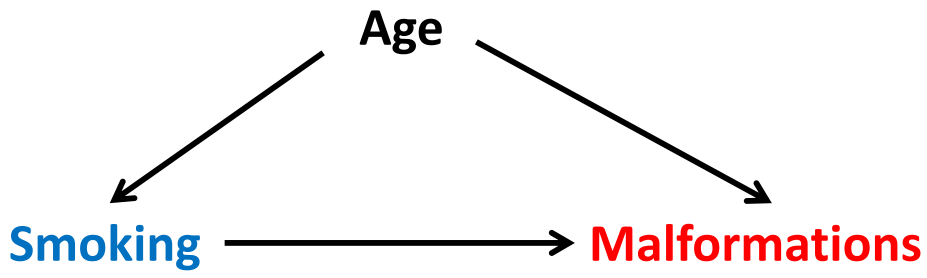
- We observe an unadjusted inverse association between smoking and malformations (RR=0.8)
- We suspect that this observed risk ratio cannot be given a causal interpretation
- We want to evaluate whether there is confounding and then adjust for a set of observed covariates to reduce confounding bias

How to construct a causal diagram (1)

Step 1

- Write down the **exposure** and the **outcome** of interest, with an arrow from the exposure to the outcome
- This arrow represents the **causal effect** we aim to estimate

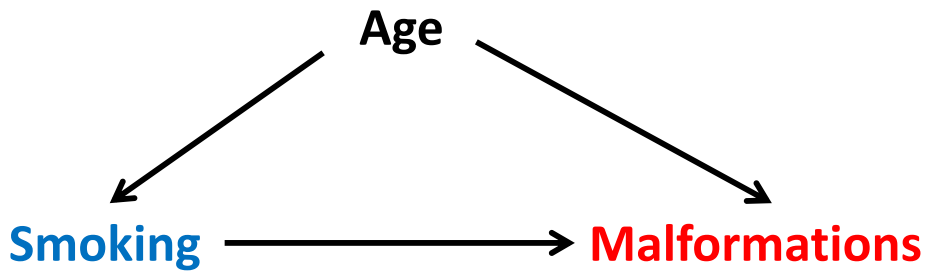
How to construct a causal diagram (2)



Step 2

- If there is any **common cause** of the exposure and the outcome we must write it in the diagram
- We must include this common cause irrespective of whether or not it has been measured in our study
- We continue in this way adding to the diagram any variable (observed or unobserved) which is common cause of two or more variables already included in the diagram

How to construct a causal diagram (2)



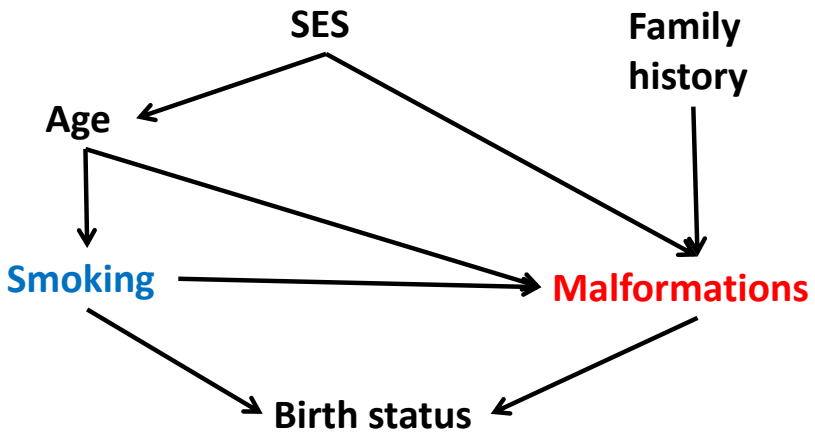
Step 2

- If there is any **common cause** of the exposure and the outcome we must write it in the diagram
- We must include this common cause irrespective of whether or not it has been measured in our study
- We continue in this way adding to the diagram any variable (observed or unobserved) which is common cause of two or more variables already included in the diagram

How to construct a causal diagram (3)

Step 3

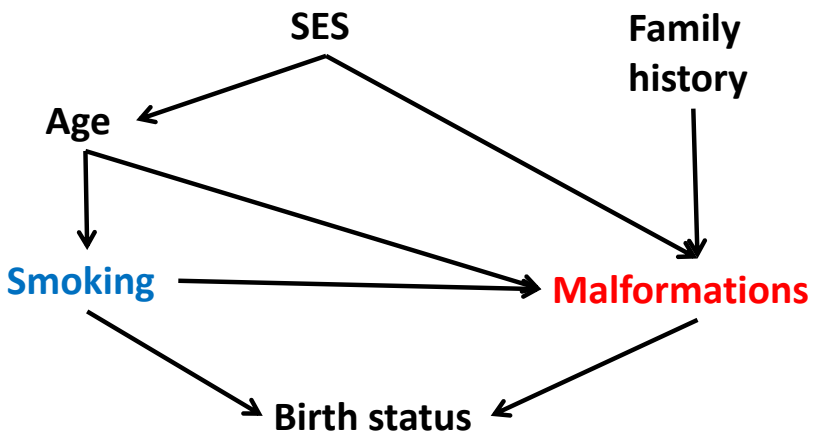
- We can choose to include variables that are not common cause of other variables in the diagrams
- For example birth status
- Suppose we finish at this point. The variables and arrows NOT in our diagram represent our causal assumptions



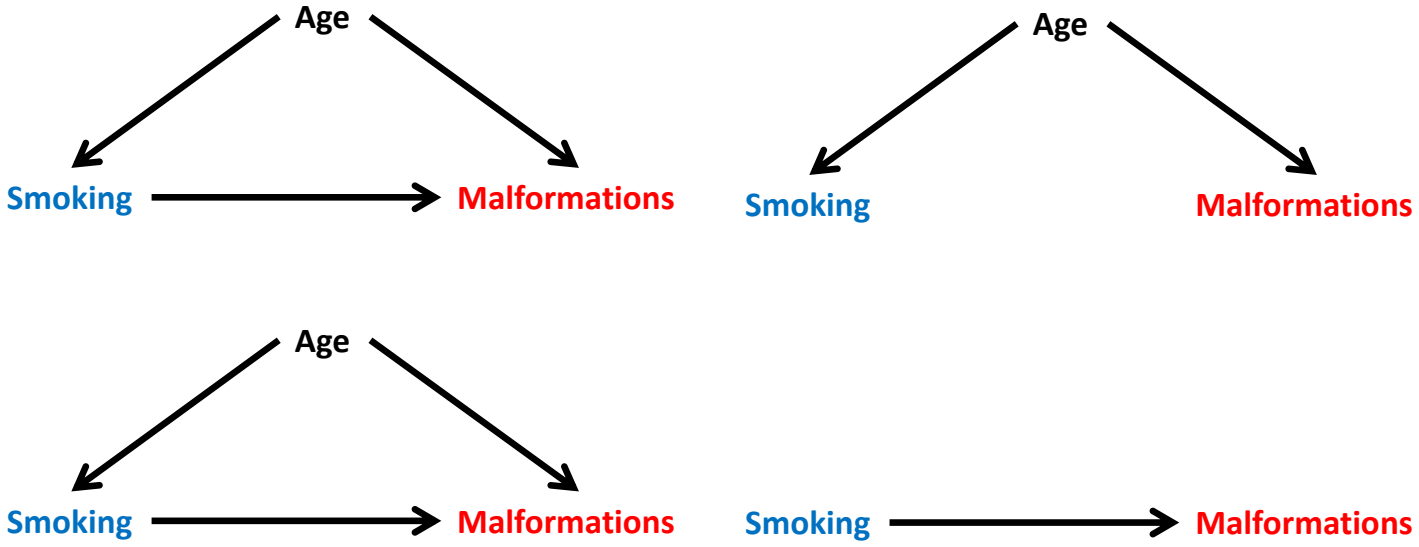
How to construct a causal diagram (3)

Step 3

- We can choose to include variables that are not common cause of other variables in the diagrams
- For example birth status
- Suppose we finish at this point. The variables and arrows NOT in our diagram represent our causal assumptions

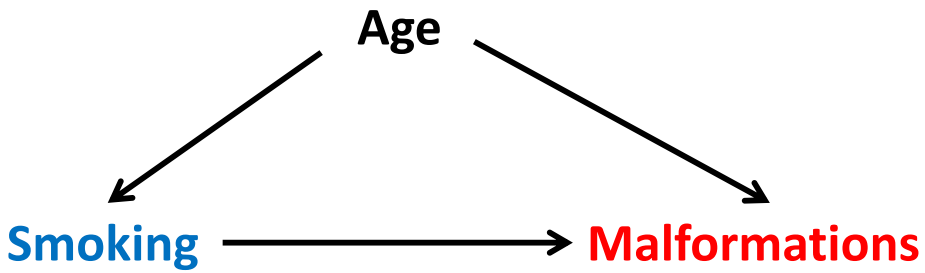


Underlying assumptions



- Assumption are encoded by:
  - the direction of arrows
  - the absence of arrows
  - the absence of common causes

Directed Acyclic Graphs

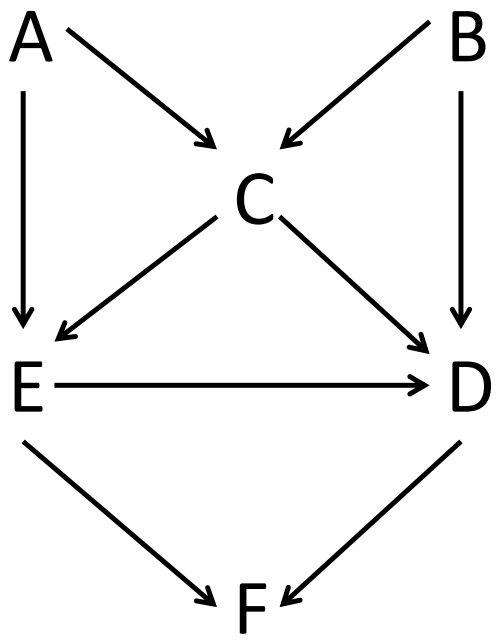


- Each arrow represents a causal influence
- The graph is
  - **Directed**, since each connection between two variables consists of an arrow;
  - **Acyclic**, since the graph contains no directed cycles. We impose this since a variable can't cause itself; however we can depict time varying processes adding one realization of each variable per time unit.

Some terminology

Children, descendants, colliders, paths

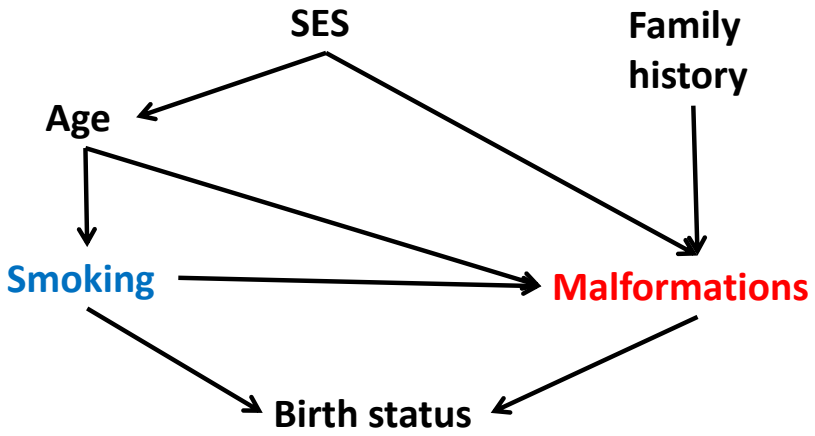
- E is a **child** of A.
- A is a **parent** of E
- F is a **descendant** of A
- A is an **ancestor** of F
- F is a **collider** along $E \rightarrow F \leftarrow D$
- $E \leftarrow A \rightarrow C \rightarrow D \rightarrow F$ is a **path** from E to F



Paths

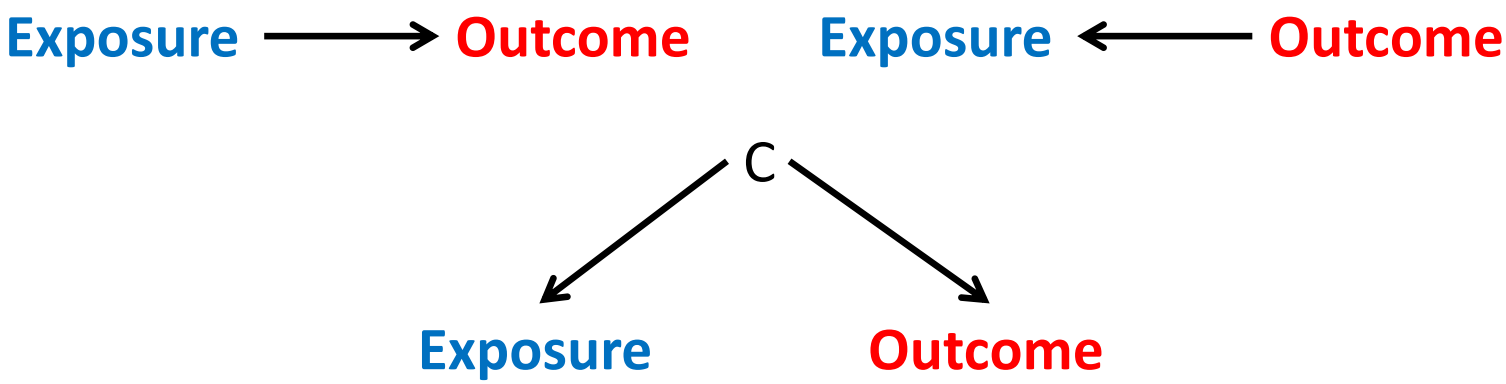
- A path is a route between two variables, not necessarily following the directions of arrows
- A **causal path** is a route between two variables, **following the directions of arrows**
- Paths are either open (association-transmitting) or blocked

Exercise



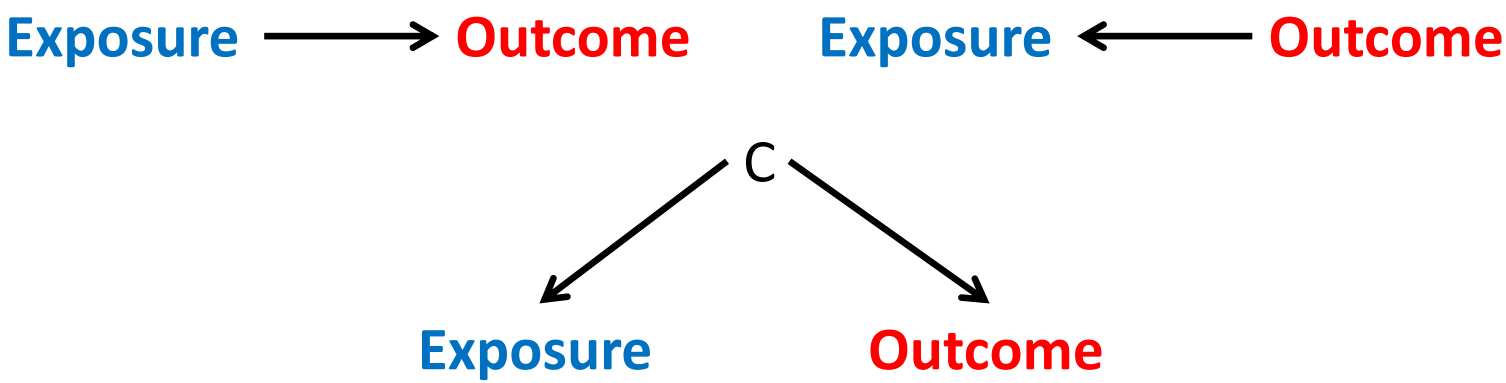
1. Identify a collider in the route between smoking and malformations
2. Identify an ancestor of smoking
3. Identify the non causal path between smoking and malformations
4. Which are the causal paths between smoking and malformations?

Association in the population



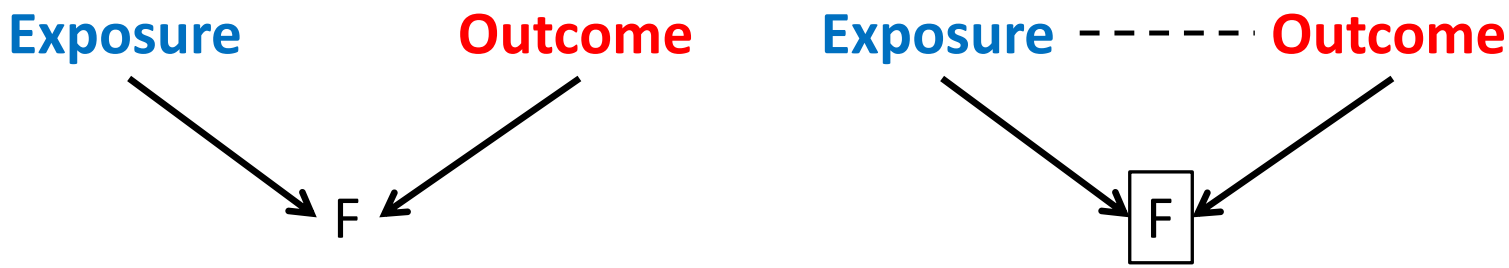
- If the exposure and the outcome are associated in the population (marginal association) then at least one of the above must be true
- Conditioning on C in the third example removes the association (block the Exposure-C-Outcome path)
  - → removes the confounding due to C

Association in the population



- If the exposure and the outcome are associated in the population (marginal association) then at least one of the above must be true
- **Conditioning on C in the third example removes the association (block the Exposure-C-Outcome path)**
  - → removes the confounding due to C

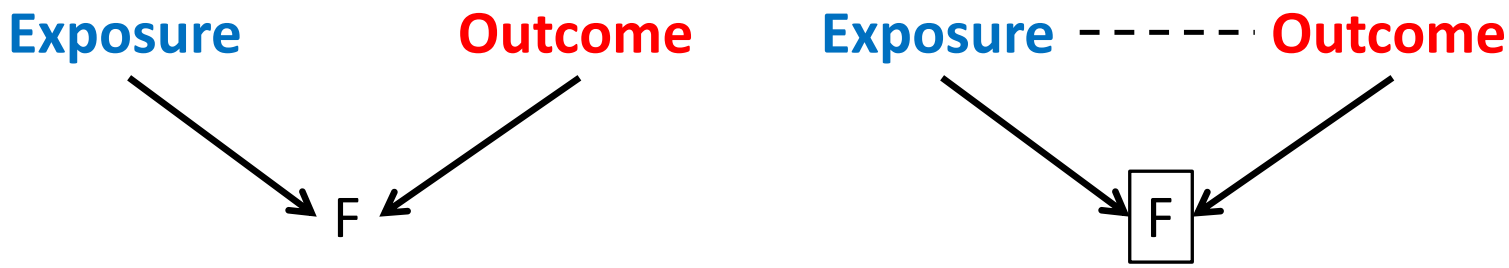
Association in a sub-population



- Even if the exposure and the outcome are independent in the population (marginally independent) the two variables will be associated within strata of the common effect F
- **Conditioning on F - denoted by the box around F in the second example - introduces a conditional association (spurious association) - denoted by the dashed line in the second example**

▷ ..we will come back to this later when discussing selection bias

Conditioning on a collider
Example

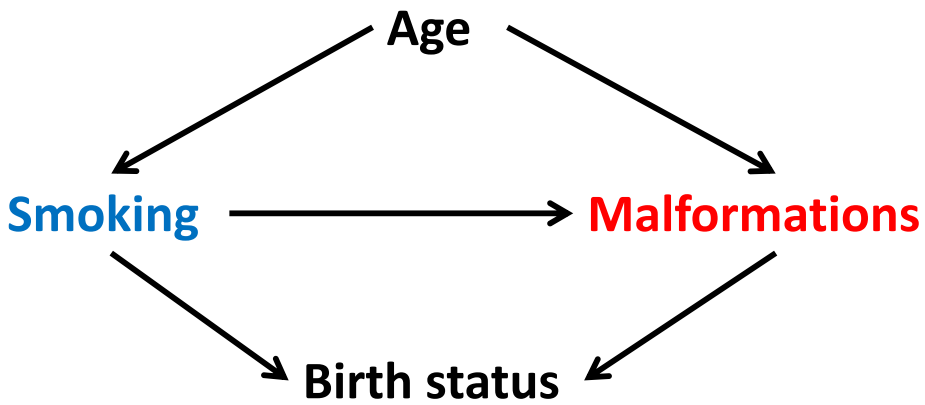


- Let F be studying at Harvard, Exposure being a basketball player and Outcome IQ score. Assume exposure and outcome are independent in the population.
- Acceptance to Harvard is positively influenced by both exposure and outcome; you're accepted either if your are a good at basketball or if you have a high IQ.
- Among Harvard students, if you have a low IQ you're likely to be good at basket $\rightarrow$ Exposure and Outcome become negatively associated.

Graphical rules to understand whether two variables are independent (d-separation)

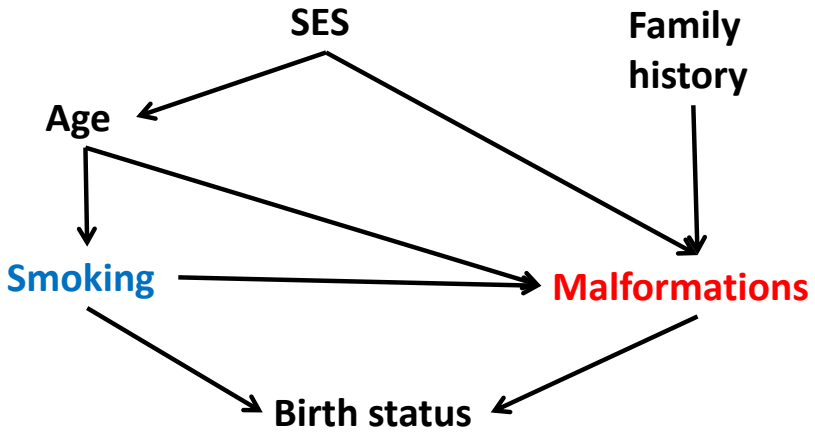
- Two variables are independent if all paths between the two variables are blocked.
  1. If there are no variables being conditioned on, a path is blocked if and only if it contains a collider: a variable F that sits in an inverted fork $\rightarrow F \leftarrow$
  2. If somewhere along the path there is a variable C (a non-collider) that sits in a chain $\rightarrow C \rightarrow$ or in a fork $\leftarrow C \rightarrow$ the path is blocked if we adjust for C

Exercise



1. Which are the paths between smoking and malformations?
2. Identify the open paths between smoking and malformations
3. Identify the blocked paths between smoking and malformations

Motivating example



- Suppose we agree that the causal structure for our example can be described by the DAG above
- We have observed an unadjusted inverse association between smoking and malformations (RR=0.8)
- We can now proceed to determine whether the smoking-malformation relationship is **confounded**
- This is done by using the **back-door criterion**

The essence of the backdoor criterion

- It looks to see whether exposure and outcome would be associated in the **absence** of a causal effect (that is presence of confounding)
- If so, it checks whether conditioning on a certain set of variables would remove the association (block all the non-causal paths) and create conditional exchangeability
- It does using the **building blocks**: (i) conditioning on a variable along an association-trasmitting path (open path) removes the association, (ii) conditioning on colliders,**or any of its descendents**, induces associations.
- **Removing some spurious associations may create others, so care is needed**

The backdoor criterion

Precisely

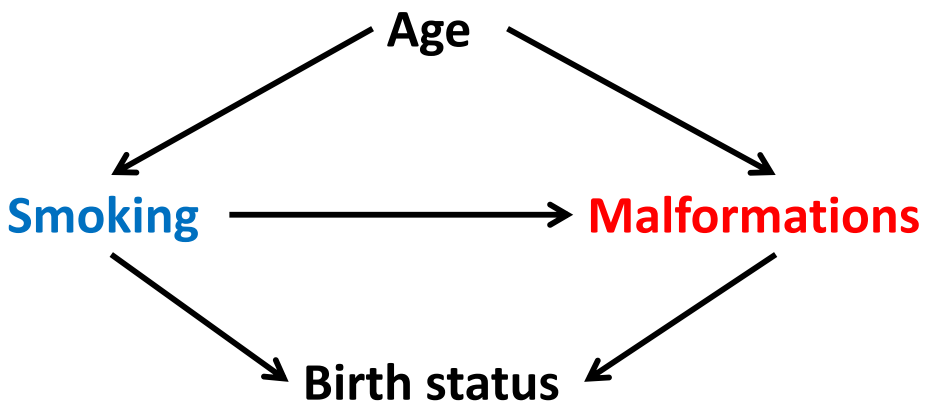
1. Choose a candidate set of variables $\Re$ **which does not contain any descendents of the exposure**
2. Remove all arrows emanating from the exposure
3. Join with a dotted line any two variables that share a child which is either itself in *Re* or has a descendant in $\Re$
4. Observe whether there is an open path (an open path does not contain colliders) from the exposure to the outcome that does not pass through a member of $\Re$
5. If NOT, then $\Re$ is **sufficient** to control for the confounding

In other words

- The backdoor criterion asks: after conditioning on $\Re$, and in absence of a causal effect of the exposure on the outcome, would we still see an association between the exposure and the outcome?

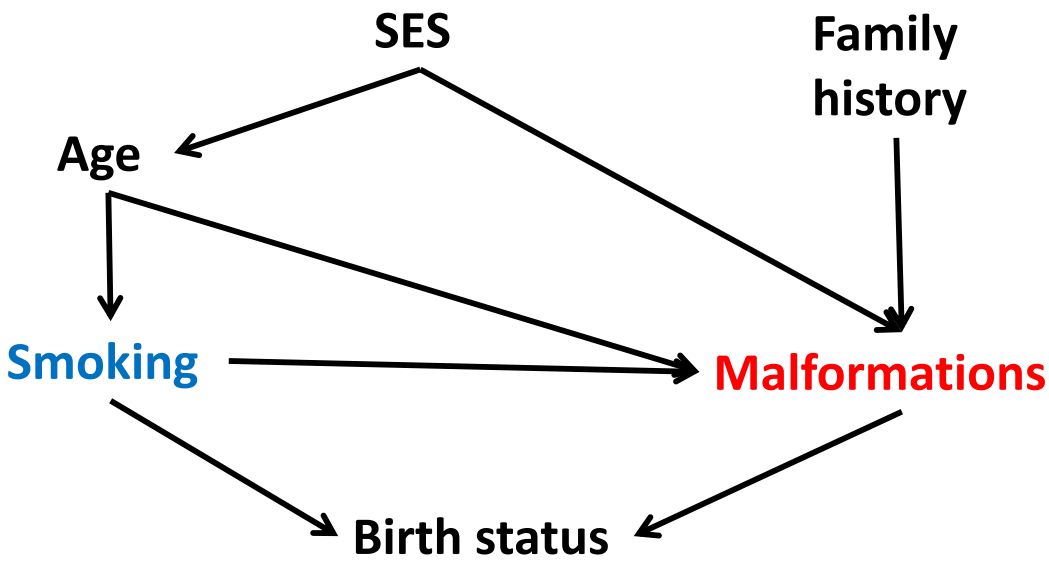
  If YES $\Re$ is not sufficient and there is still confounding

Example



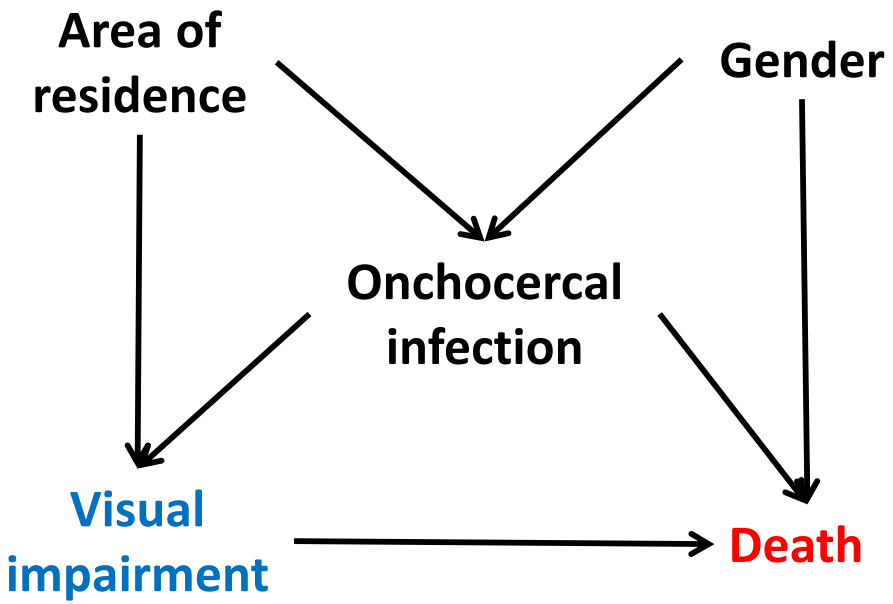
- To estimate the causal effect of smoking on malformation, which variable should we control for?

Exercise



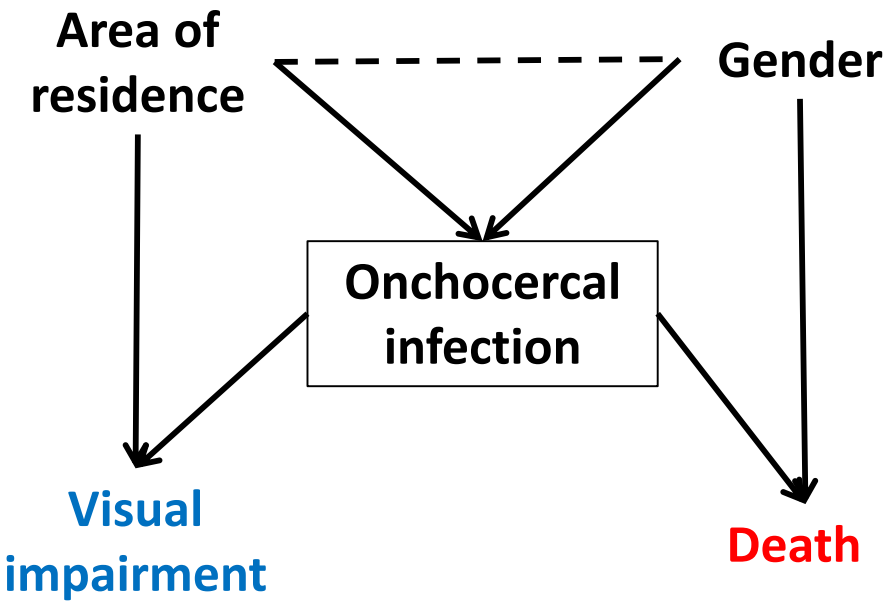
- To estimate the causal effect of smoking on malformation, which variable should we control for?

Exercise

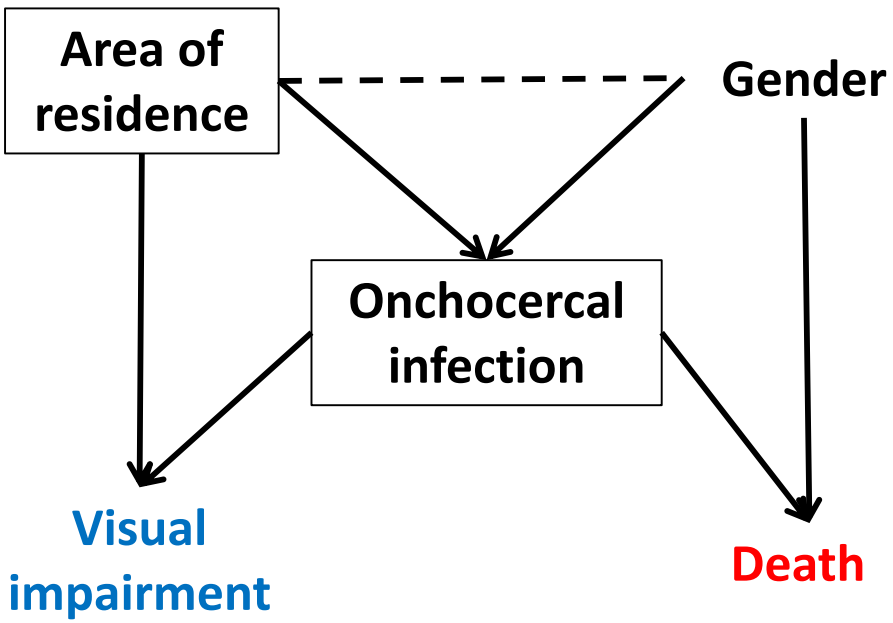


- To estimate the causal effect of visual impairment on death, which variable should we control for?

Exercise

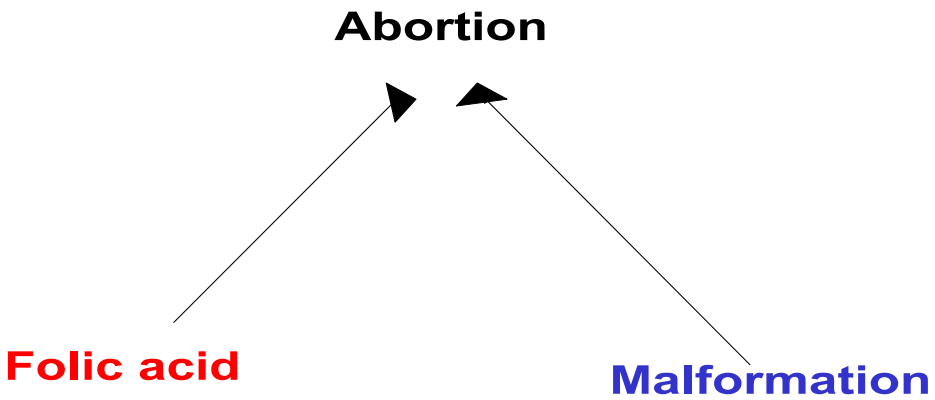

Area of residence
Gender
Onchocercal infection
Visual impairment
Death

Exercise


Area of residence
Gender
Onchocercal infection
Visual impairment
Death

An application - Folic acid and neural tube defect

Case-control study on intake of folic acid during pregnancy and risk of neural tube defects in the offspring.



- Is therapeutic abortion a confounder?

Hernan et al, Am J Epidemiol 2002;155:176-84

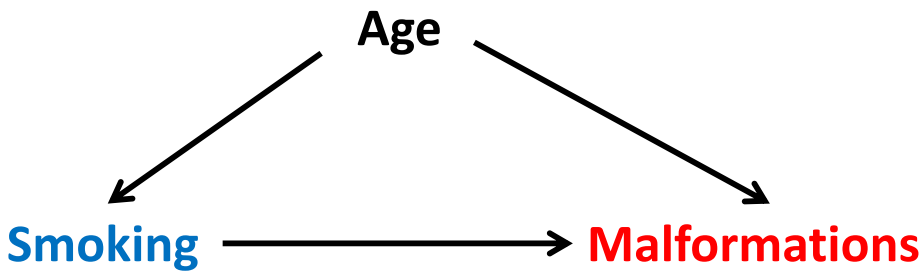
Traditional definition of confounding

- A variable that, when adjusted for, change the point estimate of interest with more than, say, 10%
- It's a variable which
  1. Independently associated with the outcome
  2. Is associated with the exposure
  3. Not on the causal pathway from exposure to outcome

Problem with traditional strategies

- They rely on statistical analyses of observed data, rather than *a priori* knowledge about causal structures.
  - Cannot be used at the design stage
  - May lead to select non confounders, which may increase bias if adjusted for

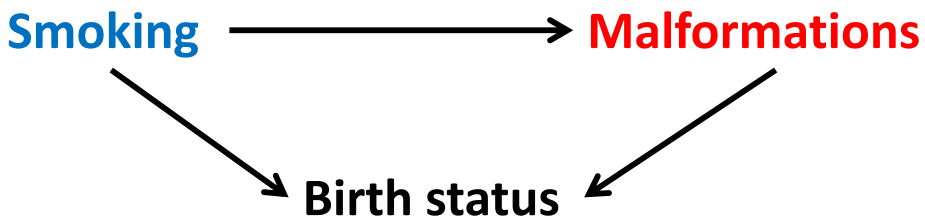
Example (1)



Some simple examples

- Age is a confounder according to both the traditional and causal diagram views

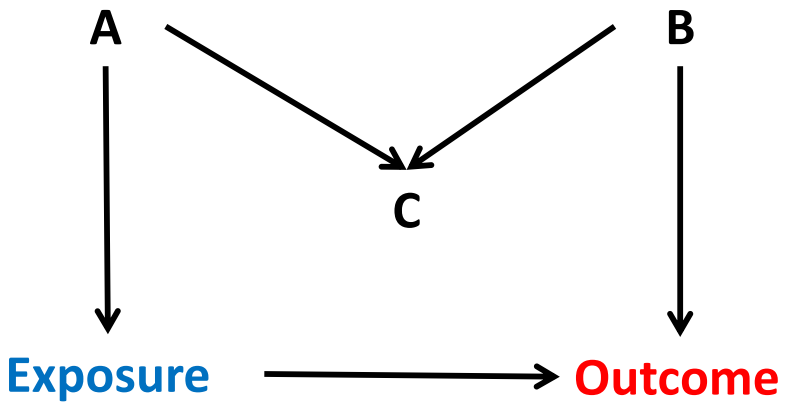
Example (2)



Some simple examples

- Birth status is NOT a confounder according to the causal diagram views (because it is a descendent of the exposure) - controlling for it create bias
- Birth status is a confounder according to the traditional view (it is not on the causal pathway). In practice, would epidemiologist control for it?

Example (3)



The M-structures

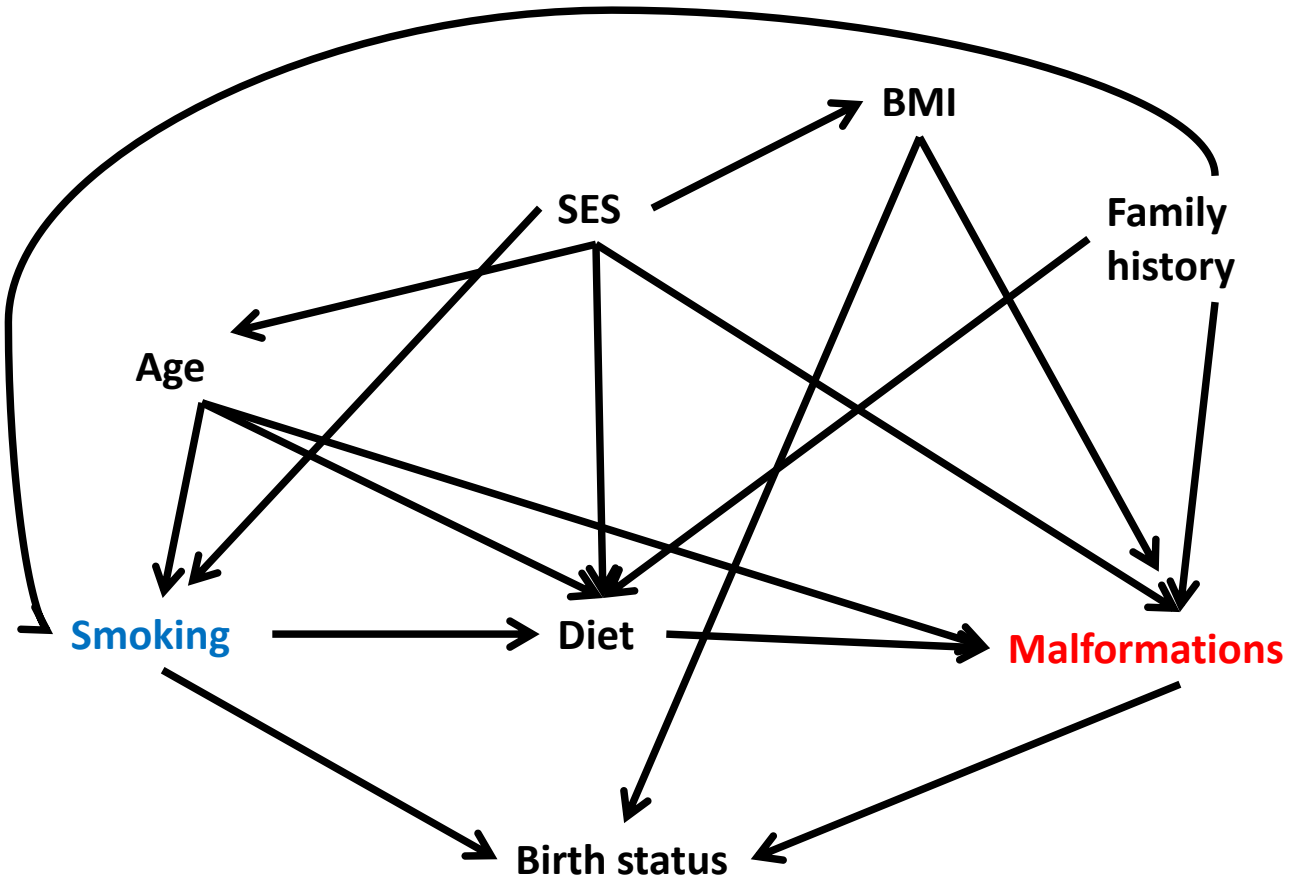
- C is NOT a confounder according to the causal diagram views (controlling for it create bias)
- C is a confounder according to the traditional view. Most epidemiologist would probably control for it.

Relationship with traditional view

- In summary, with the exception of the so-called ‘M’-structure, and related structures, the traditional and causal diagram views agree in most situations in which one confounder is being considered.

A complicated DAG

But in reality, life is more complicated!
The traditional view would not take us very far in this example..



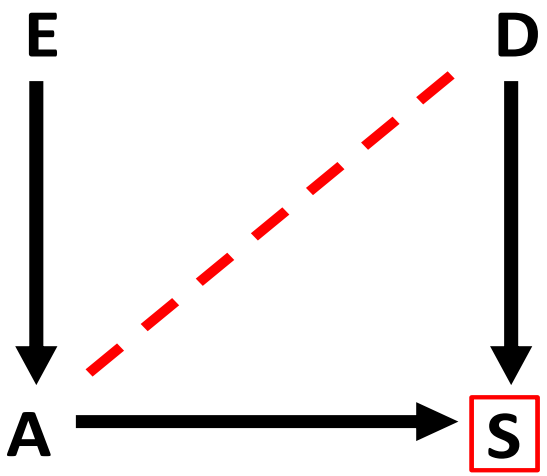
DAGs and Bias

An association between an exposure (**E**) and an outcome (**D**) can be produced by 3 causal structure (Hernan et al, Epidemiology 2004; 15;615-25) :

1. **Common causes**: E and D share a common cause $\rightarrow$ **Confounding**
2. **Common effects**: E and D share a child $\rightarrow$ **Selection Bias**
3. **Cause and effect**: E causes D or D causes E? If the latter $\rightarrow$ **Information Bias**

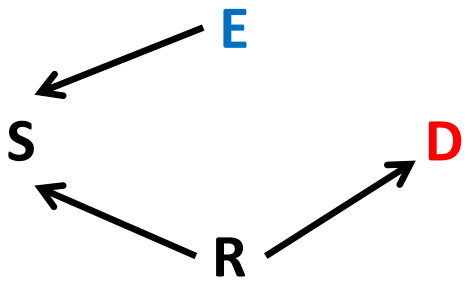
Selection bias
Case-control Study

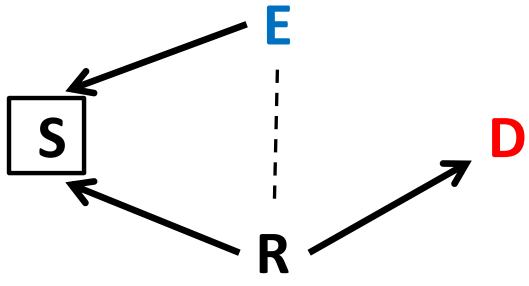
Inappropriate selection of controls in Case-Control Study



D $\rightarrow$ Myocardial Infarction
E $\rightarrow$ Postmenopausal estrogens
A $\rightarrow$ Hip fracture
S $\rightarrow$ Indicator of selection into the study

Sample selection in Cohort Studies





D $\rightarrow$ Outcome
E $\rightarrow$ Exposure of interest
R $\rightarrow$ Risk factor for the outcome
S $\rightarrow$ Indicator of selection into the sample

$\rightarrow$ **Conditioning on S induce a spurious association between E and R**

Sample selection in Cohort Studies

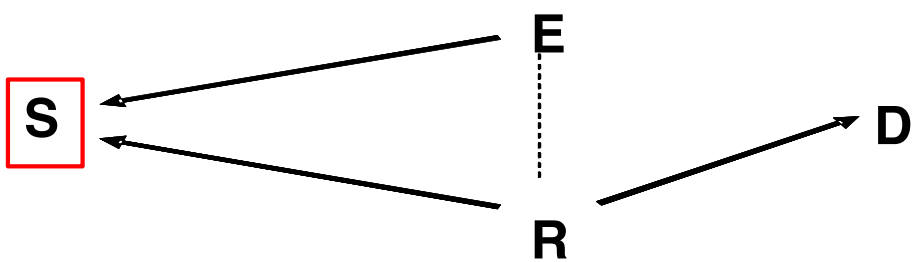
Consequence of conditioning

- If both E and R are associated with the selection, and **R unknown or unmeasured** $\rightarrow$ the backdoor path **E-R-D is opened** and the E-D association estimated in the restricted cohort may be biased
- But exposure is almost akways associated with some disease risk factors in the general population
- Thus bias depends on the net results of two components: the **induced E-R** association and the **true R-D** association
- **The confounding pattern in the restricted cohort will differ from that of the corresponding general population**

Selection bias
Sample selection in Occupational Cohort Studies

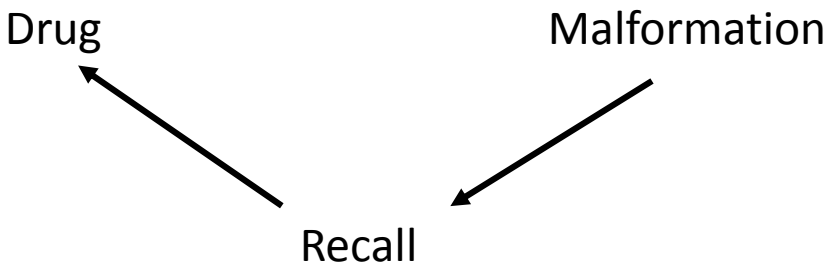
Healthy worker effect

D $\rightarrow$ Mortality
E $\rightarrow$ Exposure to Diesel exhaust
R $\rightarrow$ Health status
S $\rightarrow$ Being an active worker



Information bias
Recall Bias

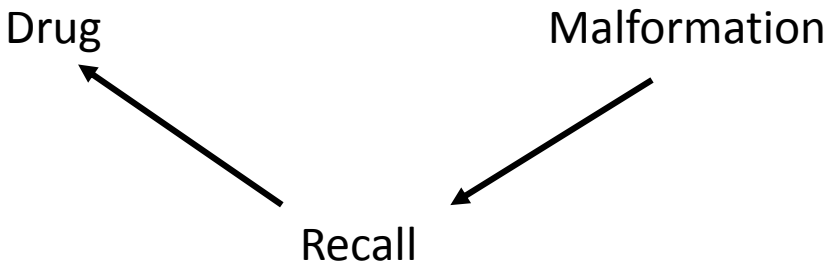
Case control study of malformation and drug use during pregnancy:



$\rightarrow$ Subjects usually report the exposure information from interviews after learning of their diagnosis, and **diagnosis may affect memory**.

Information bias
Recall Bias

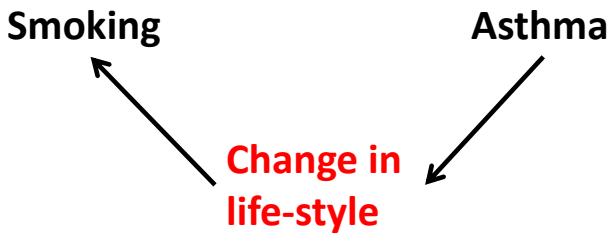
Case control study of malformation and drug use during pregnancy:



$\rightarrow$ Subjects usually report the exposure information from interviews after learning of their diagnosis, and **diagnosis may affect memory**.

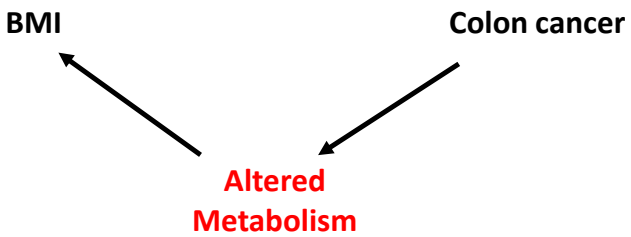
Information bias
Reverse Causality

1. Cross-sectional study of smoking status and asthma:



$\rightarrow$ Exposure and disease status are measured at the same time. It is likely that subjects who have already experienced an asthma attack quit smoking.

2. Cohort study of BMI and colon cancer risk:



$\rightarrow$ Prevalent cases are enrolled in the cohort.

Limits

- This approach **does not take into account**
  - **Sampling variation**
  - **Problem of model complexity**
- Causal relationships between variables should be specified $\rightarrow$ **Different DAGs can lead to different models**
- The magnitude and the form of the associations are not considered $\rightarrow$ **Qualitative non parametric approach**
- It is difficult to specify effect modifications

Summary

- Causal inference from observational data is challenging but important!
- Causal diagrams allow us to make our assumption explicit, and help identify an analysis that will more likely lead to causally interpretable results
- They should be used when designing the study too, so that anticipated confounders are measured
- But our causal inferences are only as valid as the causal diagram on which they rely.

Main references

- Greenland S, Pearl J, Robins J. Causal Diagrams for epidemiological research. Epidemiology 1999. 10: 37-84.
- Hernan MA, Hand Robins JM. Causal Inference. Draft chapters can be downloaded (for free) from www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- Pearl J. Causality: Models, Reasoning, and Inference. Cambridge Un Press, 2000.
- Hernan MA, Hernandez-Diaz S, Werler MM and Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. AJE 2002. 155:176-184.