Directed Acyclic Graphs: a useful modern tool in epidemiology (DAGS)

Rino Bellocco, Sc.D.

Department of Statistics & Quantitative Methods
University of Milano-Bicocca
&
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

December 3, 2018

Ack Arvid Sjolander

Causal inference

- **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?

Causal inference

- 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.

Association vs Causation

- Many epidemiological research questions are centered around a particular **exposure** and a particular **outcome**
- Typically, we want to learn whether there is an **association** between the exposure and the outcome
- Often, the aim is more ambitious; we want to know whether the exposure has a **causal effect** on the outcome

Ideal randomized trials

- In ideal randomized trials exposed and unexposed are exchangeable:

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

- As a consequence, Association = Causation:

$$RR = CRR$$

Observational studies

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

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

$$RR|L = CRR|L$$

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

- Thus the goal is to identify a set of covariates such that conditional exchangeability holds given these (goal is to minimize confounding)
- This requires background subjects-matter knowledge
- Causal diagrams help us to organize this knowledge and identify whether or not conditional exchangeability holds.

Directed Acyclic Graphs

- UCLA computer scientist Judea Pearl developed Directed Acyclic Graphs (DAGs)
- Simplify interpretation and communication in causal inference
- We will motivate DAGs in the context of covariate selection

Outline

Outline

Motivating example

Graph terminology

Covariate selection in DAGs

Motivating example, revisited

Potential problems

Aim and data

- Suppose that we carry out 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 five additional covariates
  - mothers age at conception
  - mothers socioeconomic status/education level at conception
  - mothers diet during pregnancy
  - family history of birth defects
  - indicator of whether the baby was liveborn or stillborn

Confounding

- We observe an unadjusted 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 ('comparable'), and
  - the observed risk ratio cannot be given a causal interpretation
- To reduce bias due to confounding we want to adjust for a set of observed covariates

The need for covariate selection

- One strategy would be to control for all measured covariates
- This strategy may not be optimal, because
  - **some covariates may not be confounders, and may increase non-exchangeability if controlled 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 controlled for
- Therefore, it is often desirable to control for a subset of covariates

Traditional covariate selection strategies

- Control for covariates that are selected in a stepwise regression procedure
- Control for covariates that change the point estimate of interest with more than, say, 10%
- Control 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 controlled 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 control for

Directed Acyclic Graphs

- Directed Acyclic Graphs (DAGs) can be used to overcome the problems with the 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

The simplest DAG

$$X \longrightarrow Y$$

First Step

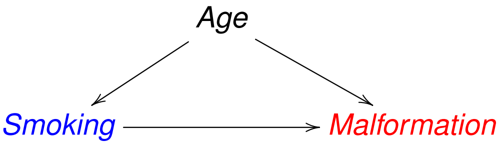
- We write the exposure and exposure of interest, with an arrow from the exposure to the outcome
- This arrow represents the causal effect we aim to estimate

How to draw a causal diagram - I

Smoking $\longrightarrow$ *Malformation*

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

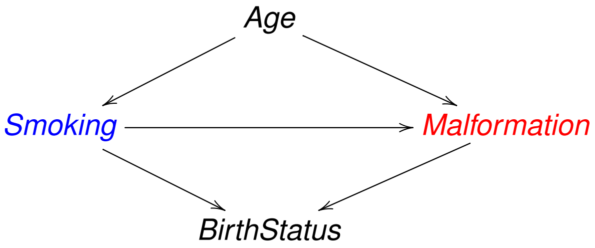
How to draw a causal diagram - II



- 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 draw a causal diagram - III

- 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



Directed Acyclic Graph

$$\begin{tikzcd}
 & Age(Z) \arrow[dl] \arrow[dr] & \\
Smoking(X) \arrow[rr] & & Malformation(Y)
\end{tikzcd}$$

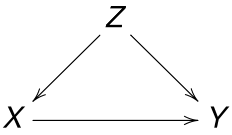
- Each arrow represents a causal influence
- The graph is
  - Directed, since each connection between two variables consists of an arrow
  - Acyclic, since the graph contains no directed cycles
- Formal connection to potential outcomes/counterfactuals through non-parametric structural equations
  - beyond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

A note on acyclicness

- We impose acyclicness 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 in DAGs be explicitly adding one ‘realization’ of each variable per time unit (more later)

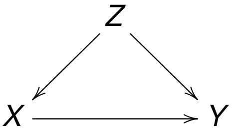
$$
\begin{tikzcd}
X^0 \arrow[rr] \arrow[dr] \arrow[drrr] & & X^1 \arrow[dr] & \\
& Y^0 \arrow[ur] \arrow[rr] & & Y^1
\end{tikzcd}
$$

Underlying assumptions



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

Underlying assumptions, cont'd



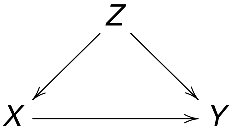
- Assumptions are encoded by the absence of arrows
  - the presence of an arrow from X to Y means that X may or may not affect Y
  - the absence of an arrow from X to Y means that X does not affect Y

Underlying assumptions, cont'd

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

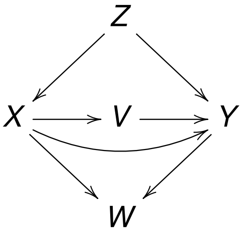
- Assumptions are encoded by the absence of common causes
  - the presence of Z means that X and Y may or may not have common causes
  - the absence of Z means that X and Y do not have any common causes

Ancestors and descendents



- The ancestors of a variable V are all other variables that affect V, either directly or indirectly
  - Z is the single ancestor of X
- 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 X

Paths



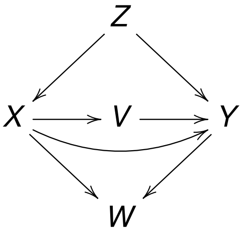
- A path is a route between two variables, not necessarily following the direction of arrows
- *Which are the paths between X and Y?*

Solution

$$
\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[r] \arrow[rr, bend right] \arrow[dr] & V \arrow[r] & Y \arrow[dl] \\
 & W &
\end{tikzcd}
$$

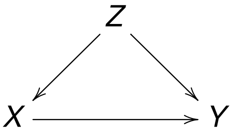
- Four paths between X and Y:
  - $X \rightarrow Y$
  - $X \rightarrow V \rightarrow Y$
  - $X \leftarrow Z \rightarrow Y$
  - $X \rightarrow W \leftarrow Y$

Causal paths



- A causal path is a route between two variables, **following the direction of arrows**
  - the causal paths from X to Y mediate the causal effect of X on Y, the non-causal paths do not
- *Which are the causal paths between X and 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 Z that sits in a 'chain'

$$\longrightarrow Z \longrightarrow$$

or in a 'fork'

$$\longleftarrow Z \longrightarrow$$

and we have controlled for Z

Rule 2

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

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

and we have **not** controlled for Z, or any of its descendents


Excuse me guys...
but what the heck happened Thur-Fri?

Once blocked stays blocked

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

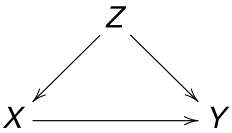
- Adjusting for V blocks the path from A to Y (rule 1)
- Adjusting for W leaves the path open (rule 2)
- Adjusting for both V and W blocks the path

Outline

Relation between 'blocking' and independence

- If all paths between X and Y are blocked, then X and Y are independent
- If at least one path is open between X and Y, then X and Y are generally associated

Example



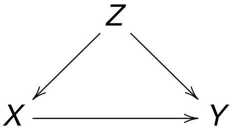
- Suppose that the DAG above depicts the true causal structure
- We want to test whether there is a causal effect of X on Y
  - i.e. does the causal path $X \rightarrow Y$ exist?
- *Control or not control for Z?*

Heuristic argument

$$
\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y
\end{tikzcd}
$$

- X = smoking, Y = malformations, Z = 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
- By not controlling 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 control for Z, and that we observe an association between X and Y
- There are two explanations for this association:
  - the causal path $X \to Y$
  - the open non-causal path $X \leftarrow Z \to Y$ (Rule 1)
- Hence, an association between X and Y, when not controlling for Z, does not prove that the causal path $X \to Y$ exists

Formal solution, cont'd

$$\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y
\end{tikzcd}$$

- Suppose that we control for Z
  - we block the non-causal path $X \leftarrow Z \rightarrow Y$ (Rule 1)
- Suppose that we then observe an association between X and Y
  - this can only be explained by the causal path $X \rightarrow Y$
- Hence, an association between X and Y, when controlling for Z, proves that there is a causal effect of X on Y

Conclusion

$$\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y
\end{tikzcd}$$

- If the aim is to test for a causal effect of X on Y, then we should control for Z
- We don't have unconditional exchangeability

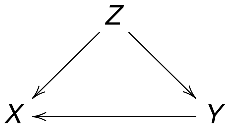
$$(Y_0, Y_1) \not\perp\!\!\!\perp X$$

but we have conditional exchangeability, given Z

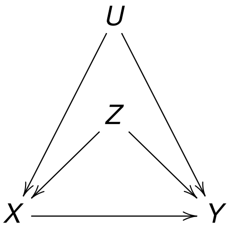
$$(Y_0, Y_1) \perp\!\!\!\perp X \mid Z$$

Remark

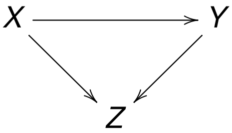
- Controlling for Z does not give a causal effect if the DAG is incorrect, e.g. if
  - Y causes X



  - there are additional common causes of X and Y



Example



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

Heuristic argument

$$\begin{tikzcd} X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd}$$

- X = smoking, Y = malformations, Z = 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

$$
\begin{tikzcd}
X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& Z &
\end{tikzcd}
$$

- 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!
- By controlling 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

$$\begin{tikzcd} X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd}$$

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

Formal solution

$$\begin{tikzcd} X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd}$$

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

Formal solution, cont'd

$$\begin{tikzcd} X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd}$$

- Suppose that we don't control for Z
  - we block the non-causal path $X \to Z \leftarrow Y$ (Rule 2)
- Suppose that we then observe an association between X and Y
  - this can only be explained by the causal path $X \to Y$
- Hence, an association between X and Y, when not controlling for Z, proves that there is a causal effect of X on Y

Conclusion

$$
\begin{tikzcd}
X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& Z &
\end{tikzcd}
$$

- If the aim is to test for a causal effect of X on Y, then we should not control for Z
- We don't have conditional exchangeability, given Z

$$(Y_0, Y_1) \not\perp\!\!\!\perp X \mid Z$$

but we have unconditional exchangeability

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

General strategy for covariate selection

- Control for covariates that block non-causal paths between the exposure and the outcome if controlled for
- Don't control for covariates that open non-causal paths between the exposure and the outcome if controlled for
- If we manage to block all non-causal paths, then any observed association must be due to a causal effect
  - we then have conditional exchangeability, given the covariates that we control for

$$(Y_0, Y_1) \amalg X \mid Z$$

Technical note: testing vs estimation

- If we manage to block all non-causal paths, then any observed association must be due to a causal effect
- We thus have a valid test for causation
- This typically, **but not necessarily**, means that we also have a valid estimate of the causal effect

Examples revisited

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

- In the left DAG, it can be shown that we have exchangeability:

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

  so that the risk ratio is equal to the causal risk ratio
  - not controlling for Z gives a valid estimate of the causal effect, as well as a valid test for causation
- In the right DAG, it can be shown that we have conditional exchangeability, given Z:

$$(Y_0, Y_1) \amalg X \mid Z$$

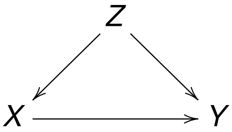
  so that the conditional risk ratio, given Z, is equal to the conditional causal risk ratio, given Z
  - controlling for Z gives a valid estimate of the causal effect, as well as a valid test for causation

Counterexample

$$X \longrightarrow Y \longrightarrow Z$$

- If we control for Z in the DAG above, then all non-causal paths between X and Y are blocked
  - there are no non-causal paths to start with
- Thus, a conditional association between X and Y, given Z, proves that there is a causal effect of X on Y
  - controlling for Z gives a valid test for causation
- However, it can be shown that controlling for Z does not give exchangeability
  - e.g. the conditional risk ratio, given Z, is not equal to the conditional causal risk ratio, given Z
  - controlling for Z does not give a valid estimate of the causal effect

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

$$\begin{tikzcd}
Z \arrow[d] & U \arrow[l] \arrow[dl] \arrow[d] \\
X \arrow[r] & Y
\end{tikzcd}$$

- A **confounder** is a variable that blocks a non-causal path between the exposure and the outcome, if controlled for
  - both Z 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, Z is not

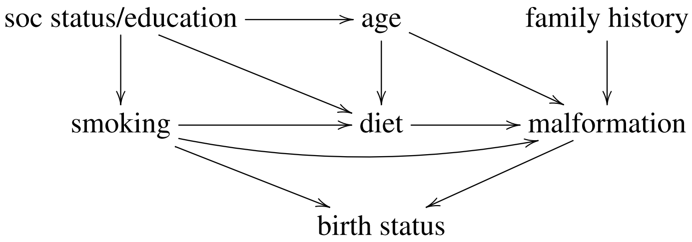
$$(Y_0, Y_1) \perp\!\!\!\perp X \mid U$$

$$(Y_0, Y_1) \not\perp\!\!\!\perp X \mid Z$$

Outline

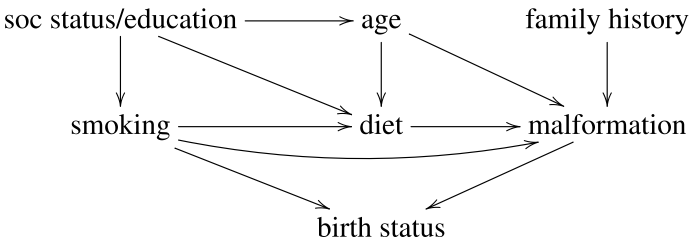
A possible DAG for the motivating example

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



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

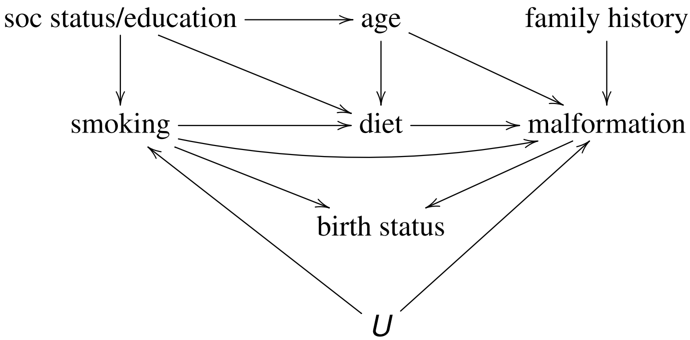
Covariate selection



- *Given the DAG, which covariates should we control 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

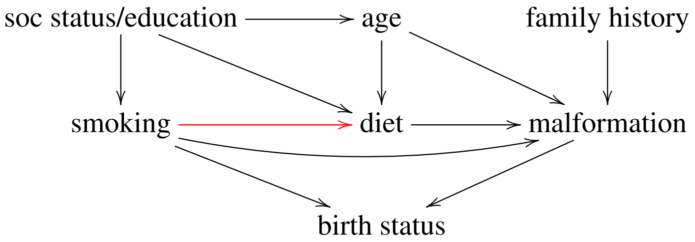
smoking | diet | malformation

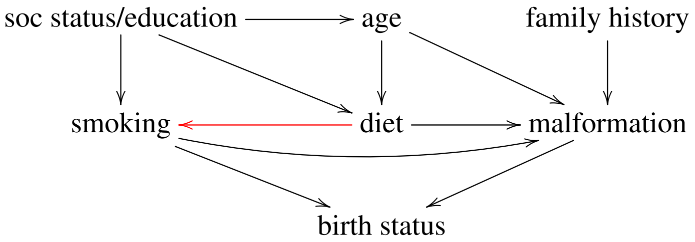
birth status

- 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

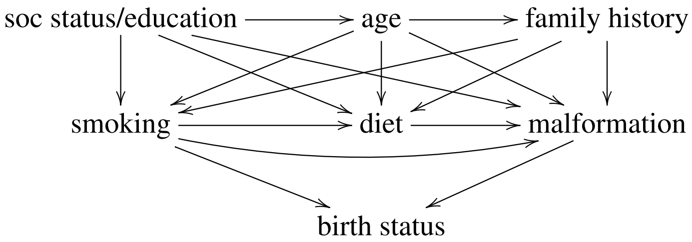




- Present all plausible DAGs, and the implied analyses

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 controlled 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
  - control for covariates that block non-causal paths between the exposure and the outcome if controlled for
- DAGs are not only tools for covariate selection
  - generally speaking, they are used to facilitate interpretation and communication in causal inference

Some References

- Causal Inference in Epidemiology (Sismec Working group) (http://www.causal.altervista.org)
- Harvard Causal Inference Group (http://www.hsph.harvard.edu/causal)
- Judea Pearl's: (http://bayes.cs.ucla.edu/jp_home.html)
- www.dagitty.com
- Hernan,M.A. A definition of causal effect for epidemiologic research, Journal of Epidemiology and Community Health (2004).
- Greenland,S, Pearl ,J, Robins ,JM. Causal diagrams for epidemiologic research. Epidemiology (1999).
- Hernan ,MA, Hernandez-Diaz, S, Werler ,MM, Mitchell, AA Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology (2002).