To adjust or not adjust, that is the question

Brice Ozenne

December 7, 2017, NRU Christmas Symposium

Setting

In a medical studies, we often want to relate:

- an outcome Y — e.g. fMRI
- to an exposure variable E — e.g. SAD, season

We also know/suspect that other variables X (called covariates) may be related to Y or E or both. e.g. age, scanner type

Setting

In a medical studies, we often want to relate:

- an outcome Y e.g. fMRI
- to an exposure variable E e.g. SAD, season

We also know/suspect that other variables X (called covariates) may be related to Y or E or both. e.g. age, scanner type

What should we do with the covariates?

Setting

In a medical studies, we often want to relate:

- an outcome Y e.g. fMRI
- to an exposure variable E e.g. SAD, season

We also know/suspect that other variables X (called covariates) may be related to Y or E or both. e.g. age, scanner type

What should we do with the covariates?

- nothing
- stratification
- interaction
- ...

Why including covariates in the analysis?

Why including covariates in the analysis?

To avoid confounding bias:

- e.g. males got treatment A and females treatment B

Why including covariates in the analysis?

To avoid confounding bias:

- e.g. males got treatment A and females treatment B

To get more insight on the mechanisms of the outcome:

- genetic factors may explain failure/sucess of a treatment

Why including covariates in the analysis?

To avoid confounding bias:

- e.g. males got treatment A and females treatment B

To get more insight on the mechanisms of the outcome:

- genetic factors may explain failure/sucess of a treatment

To have more precise estimates/increase the power of the test:

- the PET signal varies across age groups

Why NOT including covariates in the analysis?

Why NOT including covariates in the analysis?

It makes the statistical analysis more complex:

- the interpretation of the result is more difficult
- less "objective": several types of analysis are possible

Why NOT including covariates in the analysis?

It makes the statistical analysis more complex:

- the interpretation of the result is more difficult
- less "objective": several types of analysis are possible

It can hurt the statistical analysis:

- when including useless covariates: loss of precision/power
- when including un-appropriate covariates: bias

Why NOT including covariates in the analysis?

It makes the statistical analysis more complex:

- the interpretation of the result is more difficult
- less "objective": several types of analysis are possible

It can hurt the statistical analysis:

- when including useless covariates: loss of precision/power
- when including un-appropriate covariates: bias

Parsimony principle

Including covariates, how to decide?

- solution 1:
- solution 2:
- solution 3:

Including covariates, how to decide?

- solution 1: intuition
- solution 2:
- solution 3:

Including covariates, how to decide?

- solution 1: intuition
- solution 2: causal inference & directed acyclic graphs
- solution 3:

Including covariates, how to decide?

- solution 1: intuition
- solution 2: causal inference & directed acyclic graphs
- solution 3: ask Santa

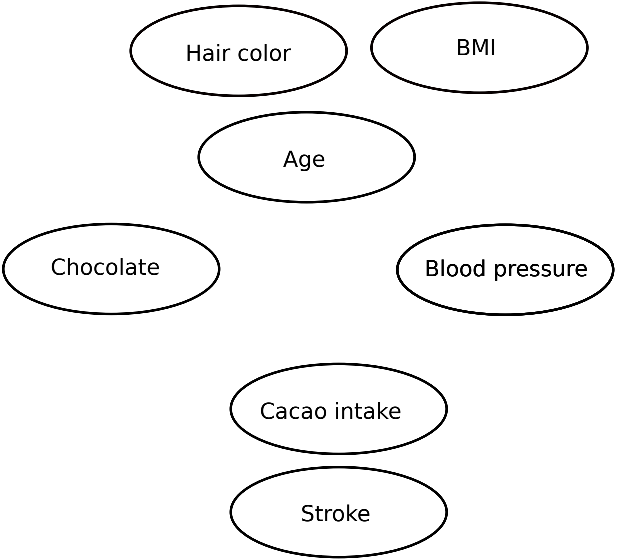
DAGs

Directed acyclic graphs (DAGs)

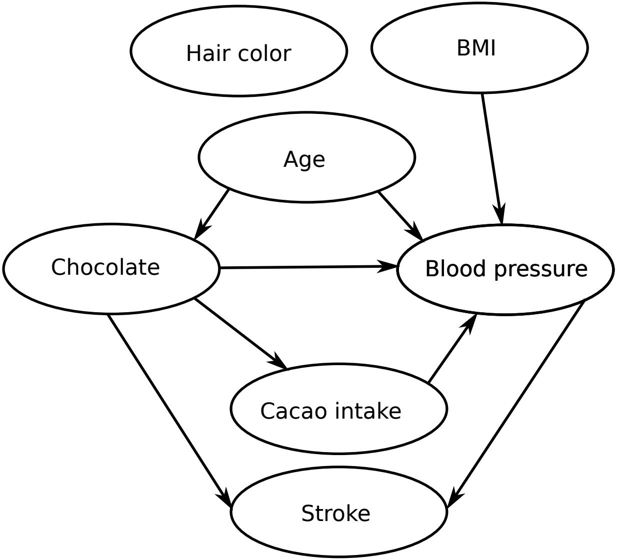
Graphical representation:

- of the variables that are being studied
- and their (causal) relationship
- in an ideal world where we could measure everything

Example of DAG with seven variables



Example of DAG with seven variables

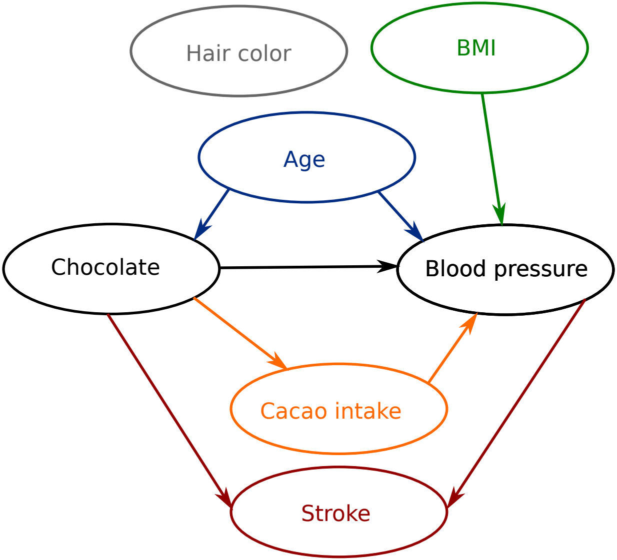


Causal DAGs

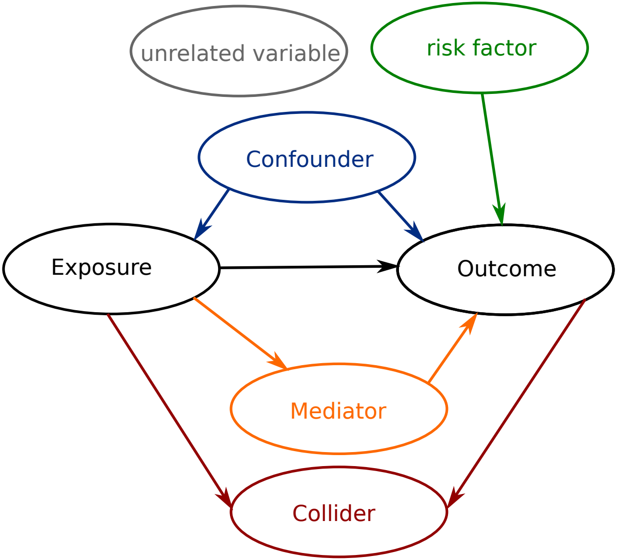
A causal DAG satisfies:

- lack of an arrow $\implies$ absence of direct causal effect
- any variable is a cause of its descendents
- all common causes (even unmeasured) are on the graph

Covariates as a structure in a DAG



Covariates as a structure in a DAG



Including covariates, how to decide?

- unrelated variable: do not adjust (would decrease precision)
- risk factor: adjust (will increase precision)
- confounder: adjust (will reduce bias)
- mediator: it depends in what we are interested in
  (direct or total effect)
- collider: do not adjust (would increase bias)

A more general criteria (bias)

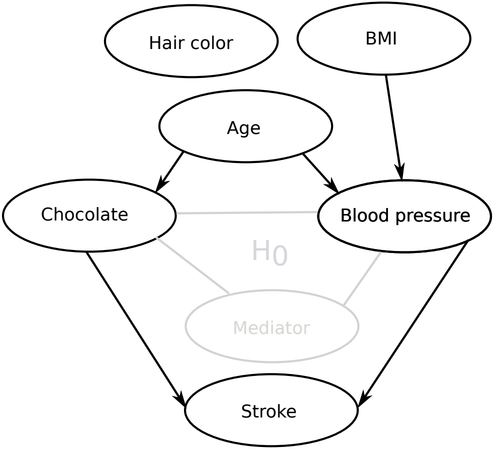
d-separation:

Two nodes Y and E are d-separated conditional on the node(s) X if every path between Y and E is blocked.

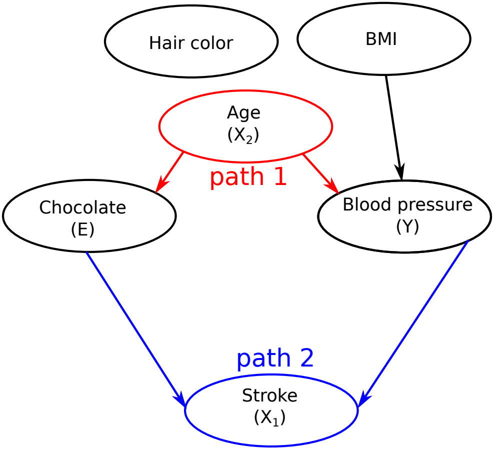
A path can be block if:

- it is a "colliding" path and does not intersect X
- it is not a "colliding" path and X it intersect X

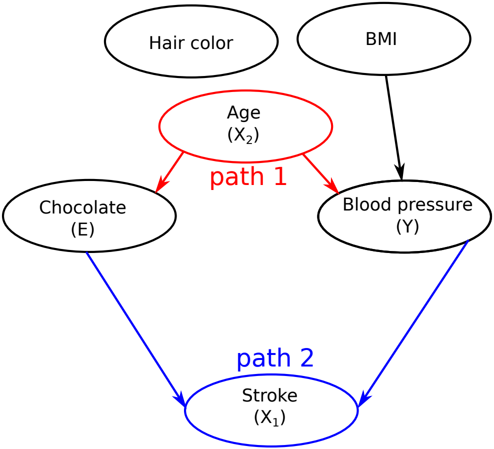
Example - d-separation



Example - d-separation



Example - d-separation



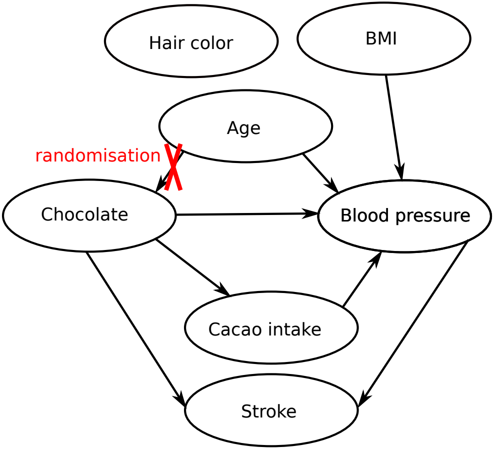
So we should adjust on Age and not on Stroke

Applications: Random assignment

Applications: Random assignment

Remove the link with the parents of a node.

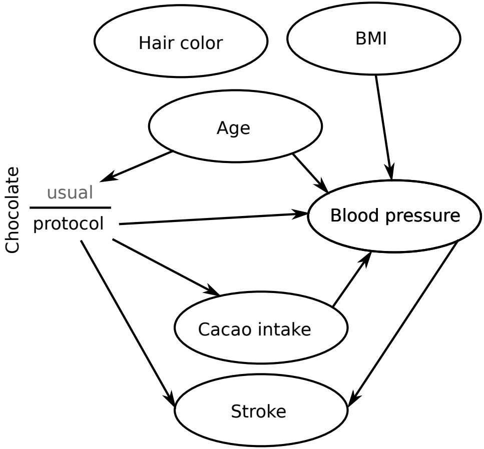
- no confounding possible
- unbiased if we don't condition on any variables



Applications: Random assignment

Remove the link with the parents of a node.

- no confounding possible
- unbiased if we don't condition on any variables



Applications: Selection bias

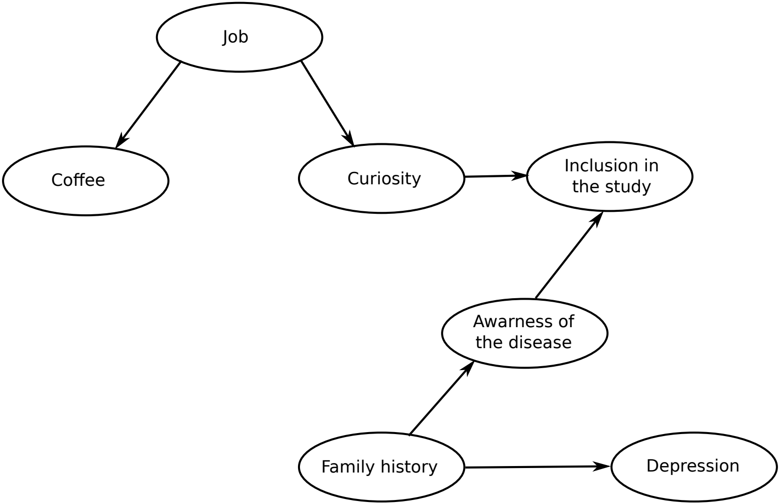
Study: relationship between coffee (E) and depression (Y)

- recrutement: volonteers

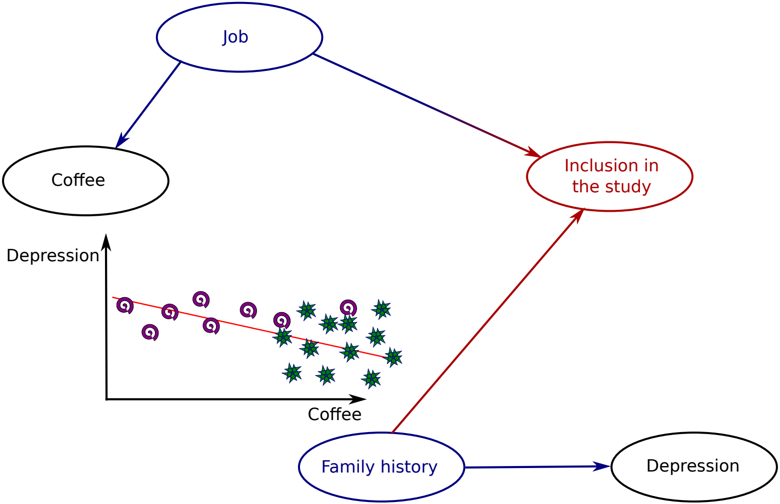
Fictitious world:

- being a researcher makes you drink more coffee an be more curious compared to other job (not the other way around)
- no relationship being researcher and depression
- your relatives influence your likelihood to be depressed and your interest in depression
- no relationship being coffee and depression
- main reasons for joining the study are curiosity and interested in depression

Applications: Selection bias



Applications: Selection bias



Summary: Christmas gift for your statistician

Summary: Christmas gift for your statistician

Offer him the DAG corresponding to your study:

- easy to draw: only potatoes and arrows, no math!
- use your expert knowledge to decide nodes/arrows

Summary: Christmas gift for your statistician

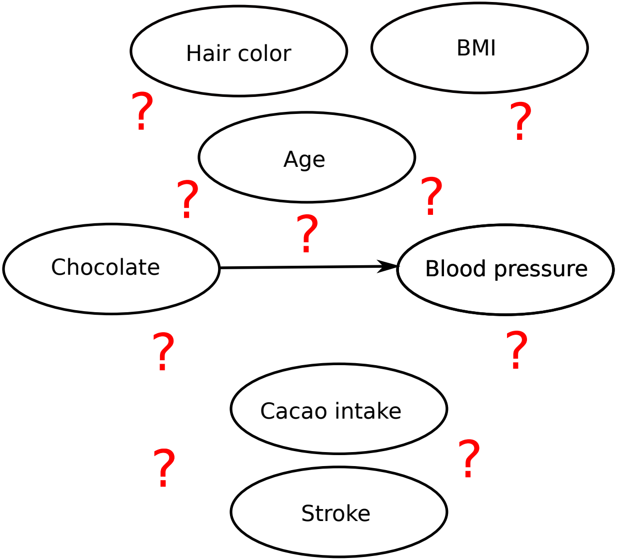
Offer him the DAG corresponding to your study:

- easy to draw: only potatoes and arrows, no math!
- use your expert knowledge to decide nodes/arrows

Limitations of DAGs:

- not well suited for displaying interactions
- difficult to do by hand when the number of variables is large
- require prior knowledge

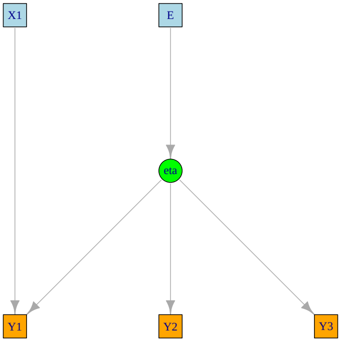
DAGs in practice (pessimistic view?)



DAGs and latent variable models

LVM as a DAGs

```
library(lava)
m <- lvm(c(Y1,Y2,Y3)~ eta, Y1 ~ X1, eta ~ E)
latent(m) <- ~eta
plot(m, plot.engine = "igraph")
```



Latent variable models

Latent variable models can be describe using path diagrams

- similar to DAGs, but can include covariance links

Can help you to decide on the presence/absence of an arrow

- but not on its direction
- and only if you enough power for the corresponding test

i.e. don't expect to identify the graph with n=10

Data-driven definition of the graph

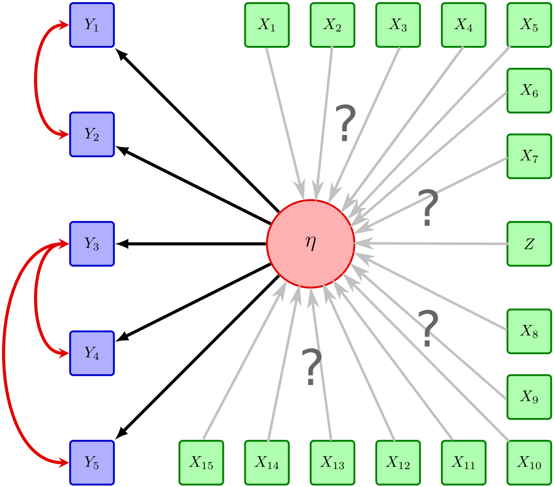
Functions `modelsearch` in lava under assumptions:

- linearity of the association
- Gaussian distribution

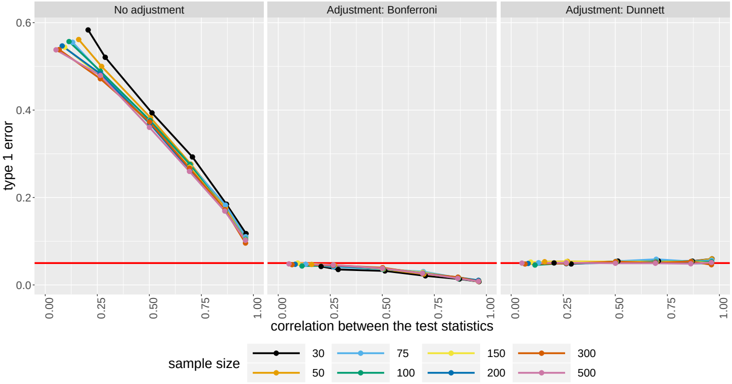
Testing several arrows requires ajustment for multiple comparisions:

- function `modelsearch2` in lava
- paper ready for submission!

(in short) Results: setting



(in short) Results: type 1 error



(in short) Results: power