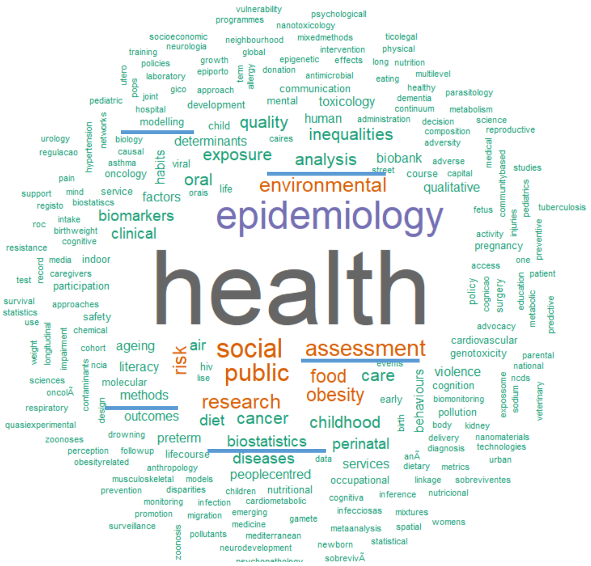
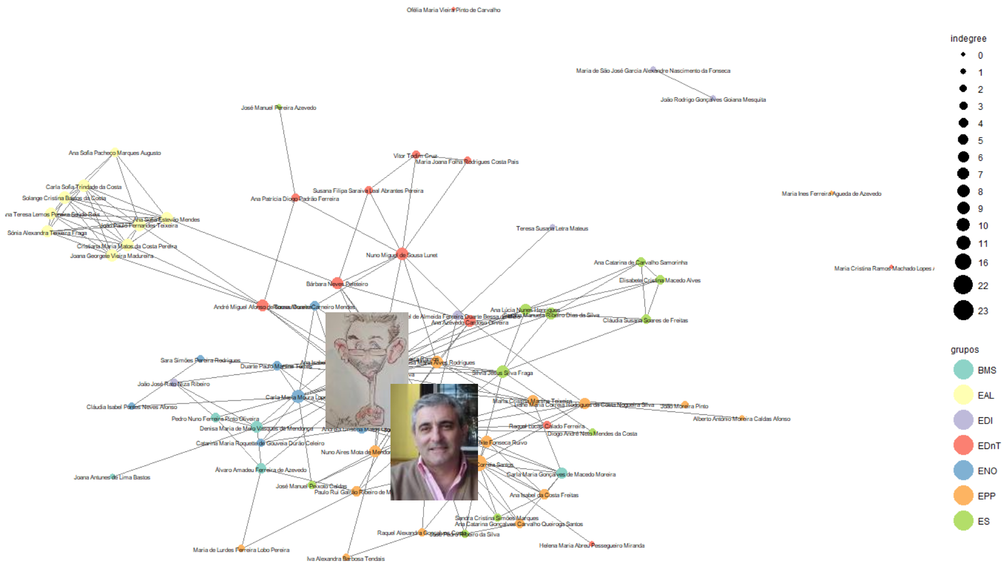
Epidemiology and the causal enquire: the role of statistics

Milton Severo

Epidemiology Keywords



Who works with whom



Statistical association vs. causation

“Correlation does not equal causation”

SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?
CORRELATION DOES NOT IMPLY CAUSATION.
SEND
Dilbert.com
DilbertCartoonist@gmail.com
11-28-11 ©2011 Scott Adams, Inc./Dist. by Universal Uclick

Figure 14-3. Another example of association or causation. (DILBERT © 2011 Scott Adams. Used by permission of UNIVERSAL UCLICK. All rights reserved.)

Statistical association vs. causation

“Correlation does not equal causation”

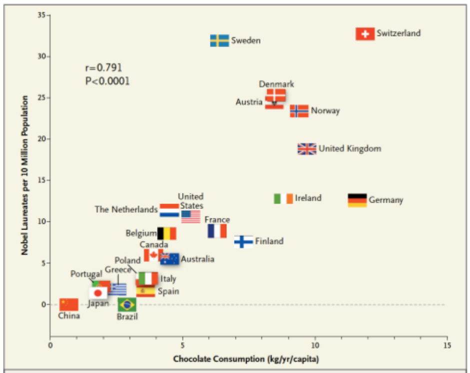


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Aspects of associations to look for when assessing causality (aka causality “criteria”)

Over the 20th century: Bradford Hill, Surgeon General, IARC

- Magnitude of effect
- Temporality
- Experimental evidence
- Dose-response relation
- Biological plausibility
- Consistency
- Specificity

Epidemiology as observation

- In empirical science in general:

Hypothesis → Observation → Structure (causes)

- In epidemiology:

Hypothesis about causal relation between exposure and outcome

→ Group-level comparison of outcome frequency between exposed and unexposed groups

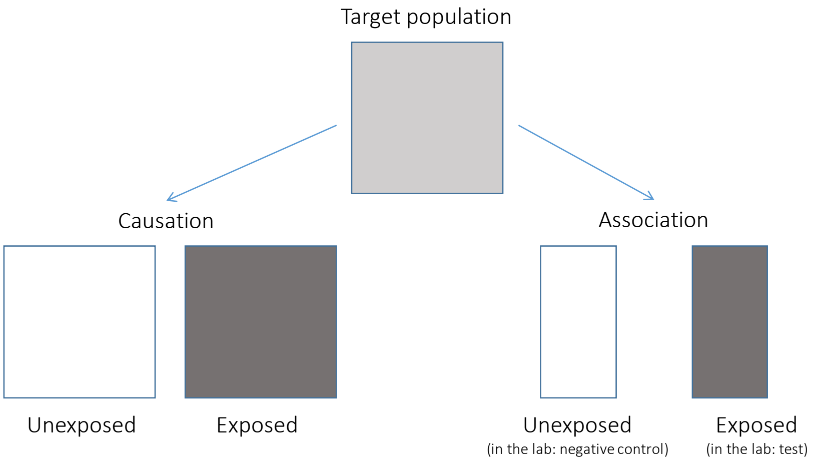
→ Inference about causation

Causal effect

The impossible contrast between the outcome of a single unit, say an individual, if assigned the experimental treatment, and the outcome of that same individual if concurrently assigned the reference treatment.

Neyman 1923

We need experiment because causation is not observable



Target population

Causation

Association

Unexposed

Exposed

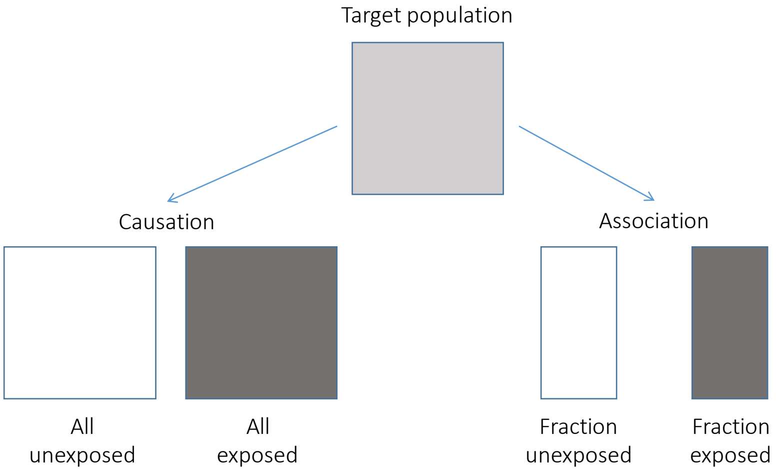
Unexposed
(in the lab: negative control)

Exposed
(in the lab: test)

Exposure allocated at random

Common to all empirical sciences

And, in epidemiology, we cannot even conduct lab experiments



Target population

Causation

Association

All
unexposed

All
exposed

Fraction
unexposed

Fraction
exposed

In observational epidemiology:
Are groups exchangeable?
Bias, confounding

The randomized controlled trial paradigm

- In epidemiology, randomization can be seen as a means of obtaining the observed contrast as close as possible to the counterfactual ideal.
- If we assume perfect randomization and no random error, both groups are as similar as possible with regard to measured and unmeasured factors.
- The probability of developing the outcome among the unexposed group equals the probability of developing the outcome in the exposed group had the latter not been exposed (counterfactual).

The randomized experiment paradigm for observational studies

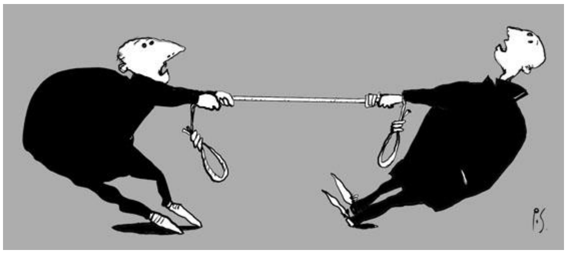
An observational study can be seen as a conditionally randomized experiment in which:

1. The interventions are not assigned by the investigators
   - but hopefully are well defined
2. The conditional probabilities of exposure are not chosen by the investigators
   - but hopefully can be estimated from the data and are not zero
3. Exchangeability is not guaranteed
   - but, based on investigators' expert knowledge, is assumed conditional on measured covariates (aka confounders)

M. Hernan

What is the effect of postmenopausal hormone therapy (HT) on coronary heart disease?

Observational vs. Experimental



Epidemiology. 2008 Nov;19(6):766-79. doi: 10.1097/EDE.0b013e3181875e61.

Postmenopausal hormone therapy (HT): observational studies vs. RCTs

- Observational studies
  - >30% **lower risk in current HT users** compared with never users
  - e.g., HR 0.68 in Nurses' Health Study (Grodstein et al. *J Women's Health 2006)*
- Randomized trials
  - >20% **higher risk in initiators of HT compared** with noninitiators
  - HR 1.24 in Women's Health Initiative (Manson et al. *NEJM 2003)*

So, studies asked different questions!

Hernan 2011.

Effect estimates (HR) for the same question (HT→CHD?):

	WHI (real trial)	NHS (observational analysed as trial)*
Overall	1.23 (0.99, 1.53)	1.05 (0.82, 1.34)
Years of follow-up		
0-2	1.51 (1.06, 2.14)	1.43 (0.92, 2.23)
>2	1.07 (0.81, 1.41)	0.91 (0.72, 1.16)
Years since menopause		
<10	0.89 (0.54, 1.44)	0.88 (0.63, 1.21)
10-20	1.24 (0.86, 1.80)	1.13 (0.85, 1.49)
>20	1.65 (1.14, 2.40)	--

* adjusted for potential confounders

Hernan 2011.

How to formalize and communicate causal questions and assumptions?

Causal diagrams – directed acyclic graphs (DAGs)

- Graphs (causal graphs or direct acyclic graphs) are considered useful for causal inference
    - Helpful for identifying which variables to control for
    - Make assumptions explicit

The simplest DAG

- This is a direct graph, which shows that A affects Y.

$$A \longrightarrow Y$$

- This is a undirect graph, which shows that A and Y are associated with each other

$$A \longleftrightarrow Y$$

What type of study would enable this?

Question 1

A) Cohort study
B) Case-Control study
C) Cross -sectional study
D) Experimental study

Question 1

A) Cohort study
B) Case-Control study
C) Cross -sectional study
D) Experimental study

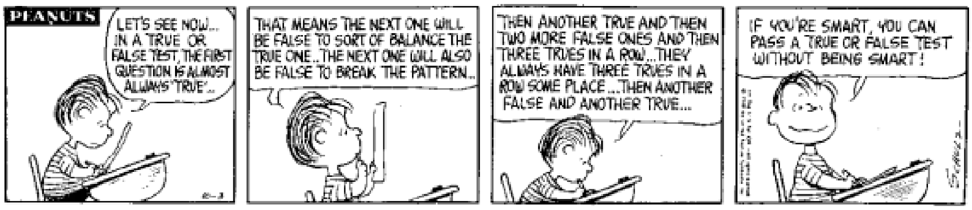
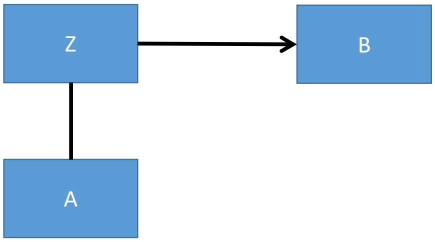

PEANUTS
LET'S SEE NOW... IN A TRUE OR FALSE TEST, THE FIRST QUESTION IS ALMOST ALWAYS 'TRUE'...
THAT MEANS THE NEXT ONE WILL BE FALSE TO SORT OF BALANCE THE TRUE ONE...THE NEXT ONE WILL ALSO BE FALSE TO BREAK THE PATTERN...
THEN ANOTHER TRUE AND THEN TWO MORE FALSE ONES AND THEN THREE TRUES IN A ROW...THEY ALWAYS HAVE THREE TRUES IN A ROW SOME PLACE...THEN ANOTHER FALSE AND ANOTHER TRUE...
IF YOU'RE SMART, YOU CAN PASS A TRUE OR FALSE TEST WITHOUT BEING SMART!

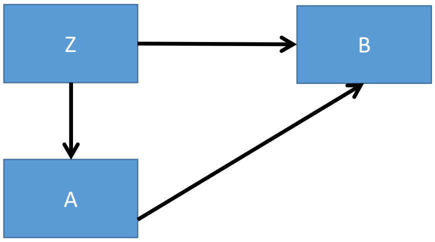
Figure 7-2. How to predict the next patient's treatment assignment in a randomized study. (PEANUTS © UFS. Reprinted by permission.)

Direct Acyclic Graphs (DAGs)

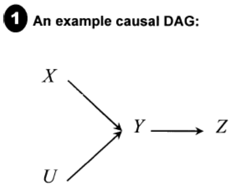
- No undirected paths



- No cycles



Causal diagrams – directed acyclic graphs (DAGs)



Paths

- A sequence of lines (edges) between two variables, regardless of direction of arrows

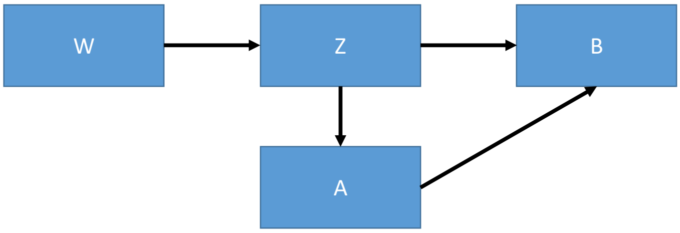
Descendants

- The direct or indirect effects of a variable

Colliders

- Common effect of two variables in a path: where the arrows ‘collide’.
- The two causes must both be “on the path”.
- Any variable on a path that is not a collider is a “non-collider”.

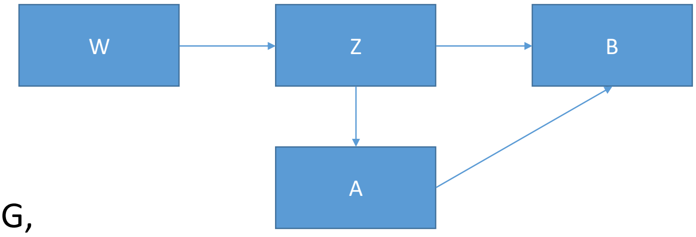
Paths



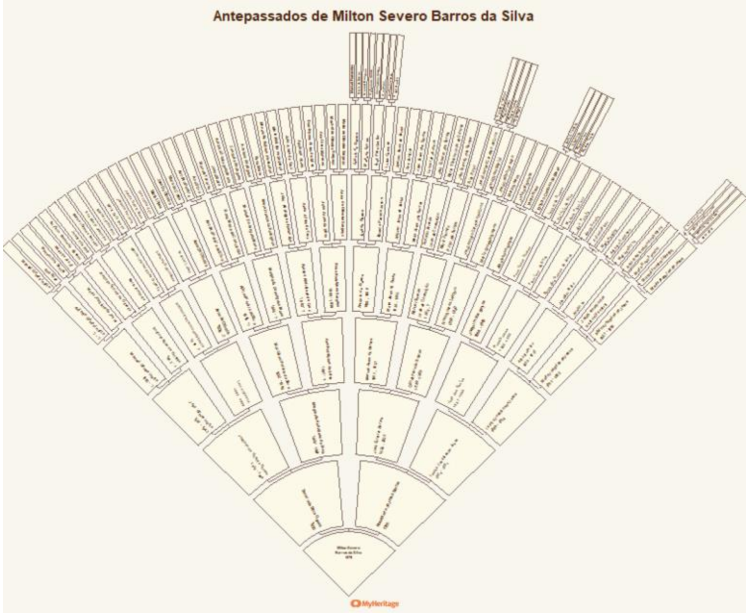
- A path is a way to get from one vertex to another, travelling along edges (regardless of direction of arrows)
  - There are two paths from W to B
  - W->Z->B and W->Z->A->B
  - There is one path from Z to W

Descendants

- Parents, children, ancestors, descendants



- In this DAG,
  - W is Z’s parent
  - Z is a child of W
  - B is a descendant of W
  - W is a ancestor of D
  - B has two parents (but we can have more than two)



So my family tree is DAG!!!!!!!

DAGs

- All common causes of two or more variables in the diagram have to be explicit, regardless of whether or not they are observed
- The diagram should be parsimonious – causes of only one of the vertices (variables) should not be included
- Unknown or unmeasured causes can and should be represented

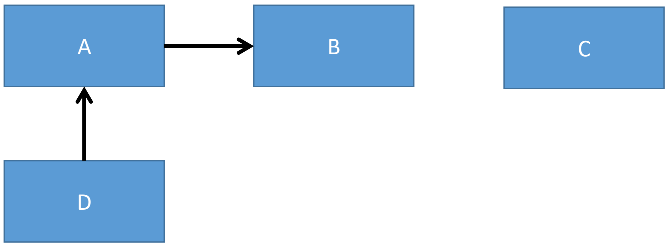
DAGs

- **Where is the statistic in this drawings?**

DAGs and Probability Distributions

- DAGs encodes assumptions about dependencies between variables
- A DAG will tell us:
  - Which variables are independent from each other
  - Which variables are conditionally independent from each other
  - i.e., ways that we can factor and simplify the joint distribution

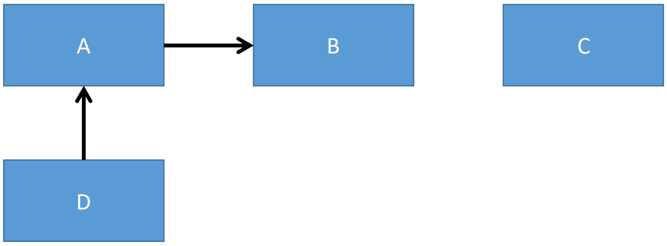
DAGs and Probability Distributions



- This DAG implies:
  - P(C|A,B,D)=P(C)
  - P(B|A,C,D)=P(B|A)
  - P(B|D)≠P(B)
  - P(D|A,B,C)=P(D|A)

Decomposition of Joint distribution

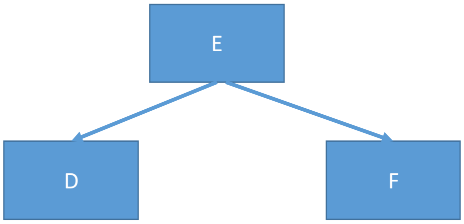
- We can decompose the joint distribution by sequential conditioning only on sets of parents
  - Start with roots (nodes with no parents)
  - Proceed down the descendant line, always conditioning on parents



  - P(A,B,C,D)=P(C)P(D)P(A|D)P(B|A)

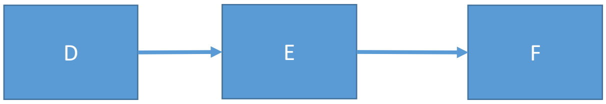
Type Paths

- Forks



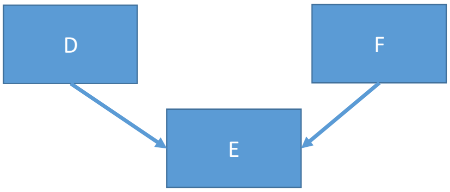
D and F will be associated

- Chains



D and F will be associated

- Inverted Fork



D and F will **NOT** be associated

In which circumstances are A and Y causally related? And statistically associated?

$$\begin{tikzcd}A \arrow[r] & Y\end{tikzcd}$$

Aspirin (A) and AMI (Y)

$$\begin{tikzcd}L \arrow[r] \arrow[rr, bend left] & A & Y\end{tikzcd}$$

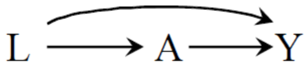
Smoking (L), Lighter (A), AMI (Y)

Does knowing A improve the prediction of Y?

Backdoor

- Backdoor paths from treatment to outcome are paths A to Y that travel through arrows going into A:
- Here, A<- L ->Y is backdoor path from A to Y.
- Backdoor path confounded relationship between A and Y

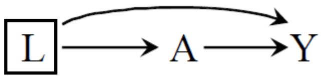
Confounding - structure



An observed statistical association between A and Y can be due to:

- A being a cause of Y
- L being a common cause of A and Y

In DAGish: A ← L → Y is an unblocked path – unless we condition on L (e.g. restrict, adjust, stratify):



Example

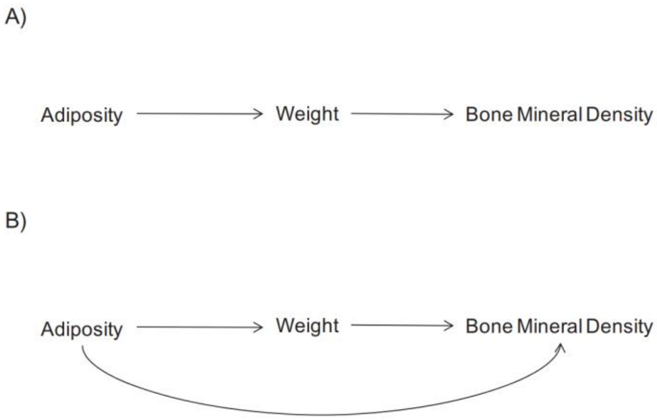
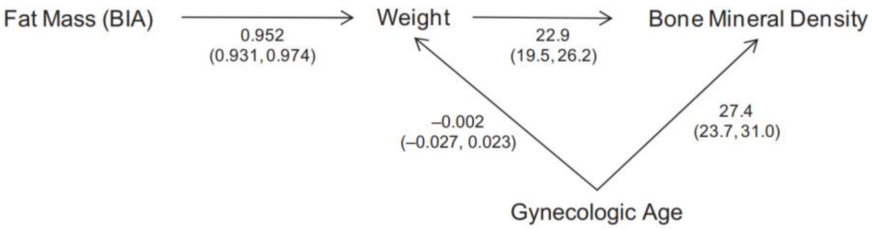
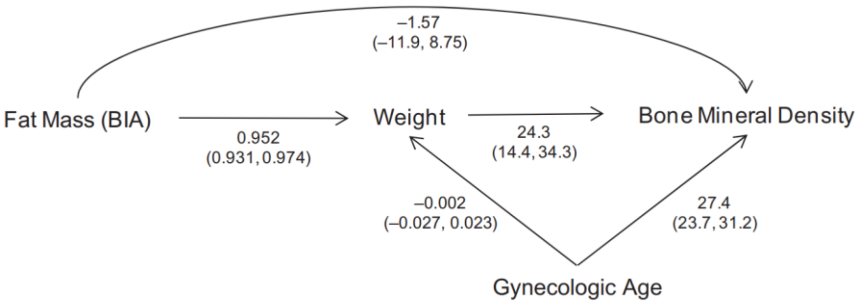


Figure 1. Two hypothesized mechanisms linking adiposity with bone mineral density in female adolescents: A) an overall effect totally mediated by weight and B) an overall effect with direct and indirect components.

- *American Journal of Epidemiology, 2011*

Example





Example

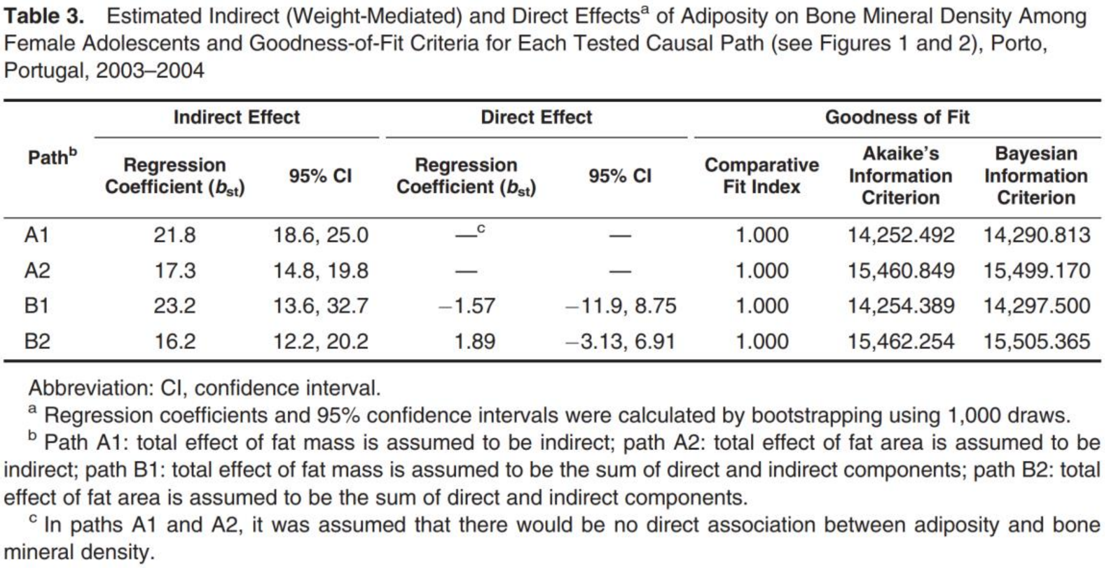


Table 3. Estimated Indirect (Weight-Mediated) and Direct Effects[a] of Adiposity on Bone Mineral Density Among Female Adolescents and Goodness-of-Fit Criteria for Each Tested Causal Path (see Figures 1 and 2), Porto, Portugal, 2003–2004

Path[b]	Indirect Effect		Direct Effect		Goodness of Fit		
	Regression Coefficient (b_{st})	95% CI	Regression Coefficient (b_{st})	95% CI	Comparative Fit Index	Akaike's Information Criterion	Bayesian Information Criterion
A1	21.8	18.6, 25.0	—[c]	—	1.000	14,252.492	14,290.813
A2	17.3	14.8, 19.8	—	—	1.000	15,460.849	15,499.170
B1	23.2	13.6, 32.7	−1.57	−11.9, 8.75	1.000	14,254.389	14,297.500
B2	16.2	12.2, 20.2	1.89	−3.13, 6.91	1.000	15,462.254	15,505.365

Abbreviation: CI, confidence interval.

[a] Regression coefficients and 95% confidence intervals were calculated by bootstrapping using 1,000 draws.

[b] Path A1: total effect of fat mass is assumed to be indirect; path A2: total effect of fat area is assumed to be indirect; path B1: total effect of fat mass is assumed to be the sum of direct and indirect components; path B2: total effect of fat area is assumed to be the sum of direct and indirect components.

[c] In paths A1 and A2, it was assumed that there would be no direct association between adiposity and bone mineral density.

Example

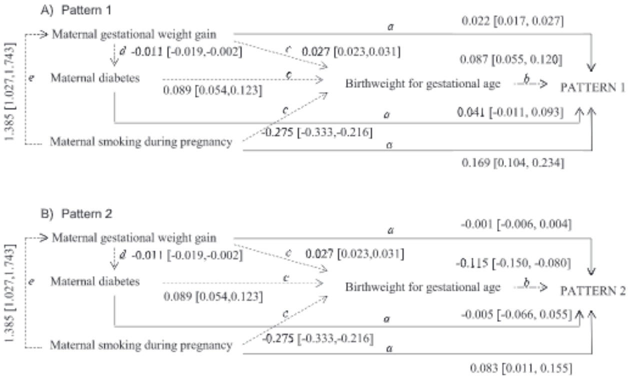


Figure 1 Causal diagram for the effects of prenatal exposures on body fat patterns identified by principal component analysis at 7-year-old children[1,2].

[1]Direct effects correspond to the intrauterine programming effects (solid arrows) and the indirect effects correspond to the effects mediated by other exposures (dashed arrows).

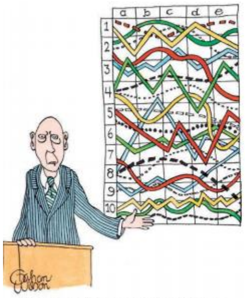
[2]Adjustment sets for each regression model:

[a]Adjusted for the other two prenatal exposures, birthweight for gestational age, maternal pre-pregnancy body mass index, age and educational level at birth.

[b]Adjusted for the prenatal exposures, maternal pre-pregnancy body mass index, age and educational level at birth.

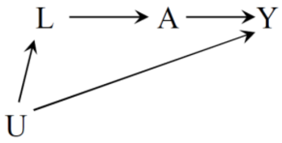
[c]Adjusted for the other two prenatal exposures, and maternal pre-pregnancy body mass index. [d]Adjusted for maternal pre-pregnancy body mass index and age at birth.

[e]Adjusted for maternal pre-pregnancy body mass index, age and educational level at birth.



"I'll pause for a moment so you can let this information sink in."

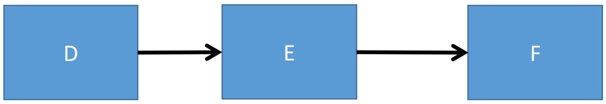
What are the confounders and how to deal with confounding?



A: aspirin; Y: stroke; L: heart disease; U: atherosclerosis
(Confounding by indication)

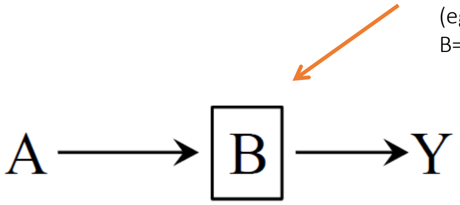
Blocking

- Paths can be blocked by conditioning on variables (vertices) in the path
- Consider the path:



- If we condition on E (a node in the middle of chain), we block the path from D to F

Is there an association between A and Y in each level of B? If we know B, does knowing A improve the prediction of Y?



Condition
(eg: restriction to
B=0)

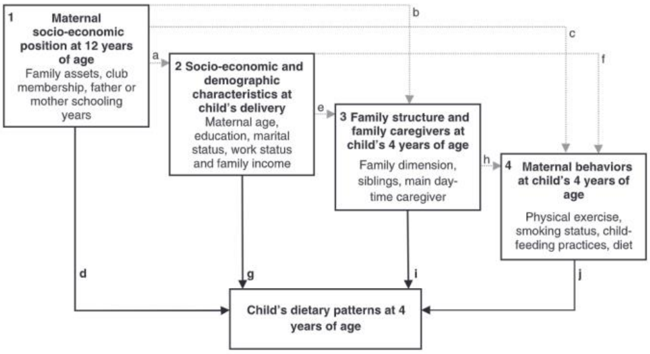
Aspirin (A), platelet aggregation (B), AMI (Y)

A and Y are marginally associated but conditionally independent, given B

Blocking

- Associations on fork (confunding) can also be block
- Consider the path:
- A<-G->B
- If we condition on G, this path from A to B is blocked.

Example 2



1 Maternal socio-economic position at 12 years of age
Family assets, club membership, father or mother schooling years

2 Socio-economic and demographic characteristics at child's delivery
Maternal age, education, marital status, work status and family income

3 Family structure and family caregivers at child's 4 years of age
Family dimension, siblings, main day-time caregiver

4 Maternal behaviors at child's 4 years of age
Physical exercise, smoking status, child-feeding practices, diet

Child's dietary patterns at 4 years of age

Fig. 1. Theoretical framework of family and maternal determinants of children's diet. This figure depicts the theoretical framework defined for analysis in the present study adapted prom previously published models (UNICEF, 1990; Victora *et al.* 1997; Davison & Birch, 2001; Solar & Irwin, 2010). Arrows represent theoretical causal relationships between determinants of children's dietary patterns. Dashed grey lines represent possible indirect effects in the pathway between levels of determinants. Solid black lines represent the direct effects of factors, after adjustment for determinants in preceding levels that is not mediated by subsequent ones, but that may be explained by other factors (unknown or unmeasured). 1. Socio-economic position at mothers' 12 years may exert an effect on children's diet through socio-economic and demographic characteristics at child's delivery (a), through its influence on subsequent family characteristics (b), through maternal behaviours (c) and/or through unknown or unmeasured determinants (d). 2. Socio-economic and demographic characteristics at child's delivery may have an effect on children's diet through subsequent family characteristics (e), through maternal behaviours (f) and/or through unknown or unmeasured factors (g). 3. Family characteristics at child's 4 years of age may have an effect on children's dietary patterns through their influence on maternal behaviours (h) and/or through unknown or unmeasured determinants (i). 4. In this conceptual framework, maternal behaviours would then influence children's dietary patterns directly and/or through other unknown or unmeasured factors (j). In the present study, we were particularly interested in the overall effects and the direct effects (highlighted in bold, d, g, i and j).

Maternal & Child Nutrition, 2017

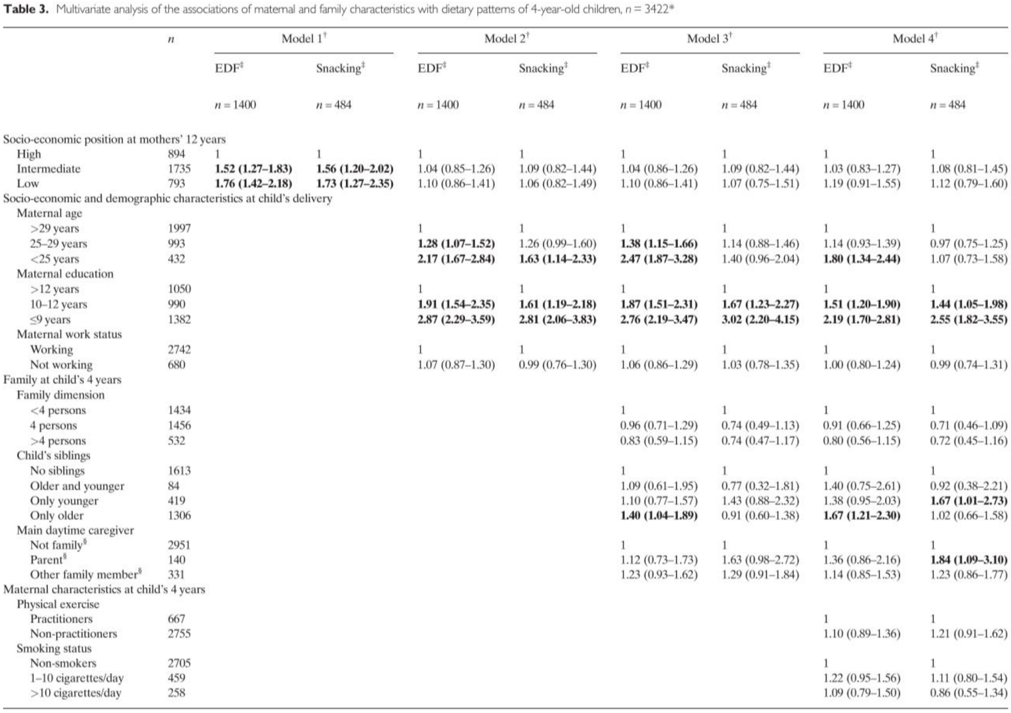


Table 3. Multivariate analysis of the associations of maternal and family characteristics with dietary patterns of 4-year-old children, *n* = 3422*

	n	Model 1†		Model 2†		Model 3†		Model 4†	
		EDF‡	Snacking‡	EDF‡	Snacking‡	EDF‡	Snacking‡	EDF‡	Snacking‡
		n = 1400	*n* = 484	*n* = 1400	*n* = 484	*n* = 1400	*n* = 484	*n* = 1400	*n* = 484
Socio-economic position at mothers' 12 years									
High	894	1	1	1	1	1	1	1	1
Intermediate	1735	**1.52 (1.27–1.83)**	**1.56 (1.20–2.02)**	1.04 (0.85–1.26)	1.09 (0.82–1.44)	1.04 (0.86–1.26)	1.09 (0.82–1.44)	1.03 (0.83–1.27)	1.08 (0.81–1.45)
Low	793	**1.76 (1.42–2.18)**	**1.73 (1.27–2.35)**	1.10 (0.86–1.41)	1.06 (0.82–1.49)	1.10 (0.86–1.41)	1.07 (0.75–1.51)	1.19 (0.91–1.55)	1.12 (0.79–1.60)
Socio-economic and demographic characteristics at child's delivery									
Maternal age									
>29 years	1997			1	1	1	1	1	1
25–29 years	993			**1.28 (1.07–1.52)**	1.26 (0.99–1.60)	**1.38 (1.15–1.66)**	1.14 (0.88–1.46)	1.14 (0.93–1.39)	0.97 (0.75–1.25)
<25 years	432			**2.17 (1.67–2.84)**	**1.63 (1.14–2.33)**	**2.47 (1.87–3.28)**	1.40 (0.96–2.04)	**1.80 (1.34–2.44)**	1.07 (0.73–1.58)
Maternal education									
>12 years	1050			1	1	1	1	1	1
10–12 years	990			**1.91 (1.54–2.35)**	**1.61 (1.19–2.18)**	**1.87 (1.51–2.31)**	**1.67 (1.23–2.27)**	**1.51 (1.20–1.90)**	**1.44 (1.05–1.98)**
≤9 years	1382			**2.87 (2.29–3.59)**	**2.81 (2.06–3.83)**	**2.76 (2.19–3.47)**	**3.02 (2.20–4.15)**	**2.19 (1.70–2.81)**	**2.55 (1.82–3.55)**
Maternal work status									
Working	2742			1	1	1	1	1	1
Not working	680			1.07 (0.87–1.30)	0.99 (0.76–1.30)	1.06 (0.86–1.29)	1.03 (0.78–1.35)	1.00 (0.80–1.24)	0.99 (0.74–1.31)
Family at child's 4 years									
Family dimension									
<4 persons	1434					1	1	1	1
4 persons	1456					0.96 (0.71–1.29)	0.74 (0.49–1.13)	0.91 (0.66–1.25)	0.71 (0.46–1.09)
>4 persons	532					0.83 (0.59–1.15)	0.74 (0.47–1.17)	0.80 (0.56–1.15)	0.72 (0.45–1.16)
Child's siblings									
No siblings	1613					1	1	1	1
Older and younger	84					1.09 (0.61–1.95)	0.77 (0.32–1.81)	1.40 (0.75–2.61)	0.92 (0.38–2.21)
Only younger	419					1.10 (0.77–1.57)	1.43 (0.88–2.32)	1.38 (0.95–2.03)	**1.67 (1.01–2.73)**
Only older	1306					**1.40 (1.04–1.89)**	0.91 (0.60–1.38)	**1.67 (1.21–2.30)**	1.02 (0.66–1.58)
Main daytime caregiver									
Not family§	2951					1	1	1	1
Parent§	140					1.12 (0.73–1.73)	1.63 (0.98–2.72)	1.36 (0.86–2.16)	**1.84 (1.09–3.10)**
Other family member§	331					1.23 (0.93–1.62)	1.29 (0.91–1.84)	1.14 (0.85–1.53)	1.23 (0.86–1.77)
Maternal characteristics at child's 4 years									
Physical exercise									
Practitioners	667							1	1
Non-practitioners	2755							1.10 (0.89–1.36)	1.21 (0.91–1.62)
Smoking status									
Non-smokers	2705							1	1
1–10 cigarettes/day	459							1.22 (0.95–1.56)	1.11 (0.80–1.54)
>10 cigarettes/day	258							1.09 (0.79–1.50)	0.86 (0.55–1.34)

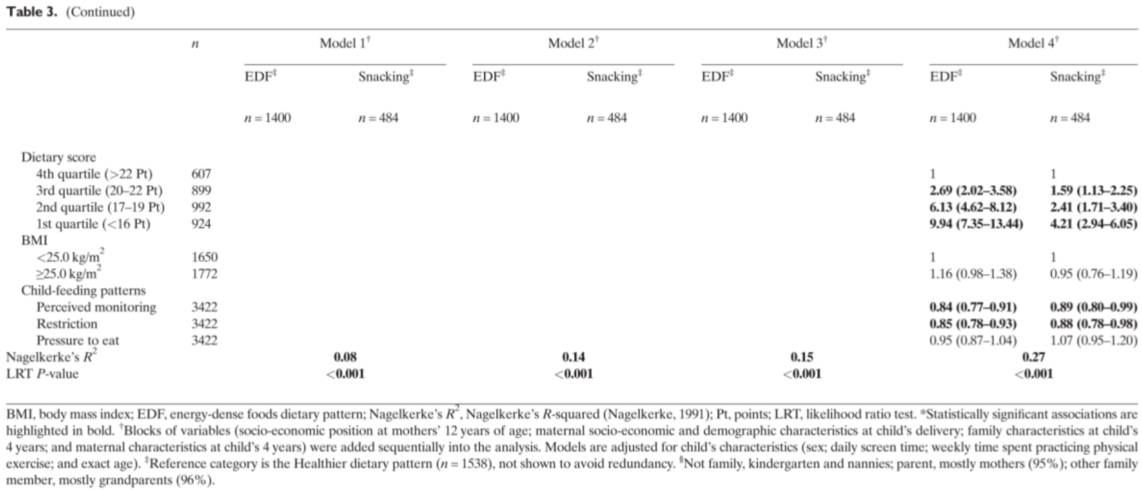


Table 3. (Continued)

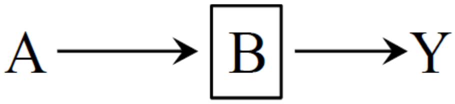
	n	Model 1[†]		Model 2[†]		Model 3[†]		Model 4[†]	
		EDF[‡]	Snacking[‡]	EDF[‡]	Snacking[‡]	EDF[‡]	Snacking[‡]	EDF[‡]	Snacking[‡]
		$n = 1400$	$n = 484$	$n = 1400$	$n = 484$	$n = 1400$	$n = 484$	$n = 1400$	$n = 484$
Dietary score									
4th quartile (>22 Pt)	607							1	1
3rd quartile (20–22 Pt)	899							**2.69 (2.02–3.58)**	**1.59 (1.13–2.25)**
2nd quartile (17–19 Pt)	992							**6.13 (4.62–8.12)**	**2.41 (1.71–3.40)**
1st quartile (<16 Pt)	924							**9.94 (7.35–13.44)**	**4.21 (2.94–6.05)**
BMI									
<25.0 kg/m^2	1650							1	1
≥25.0 kg/m^2	1772							1.16 (0.98–1.38)	0.95 (0.76–1.19)
Child-feeding patterns									
Perceived monitoring	3422							**0.84 (0.77–0.91)**	**0.89 (0.80–0.99)**
Restriction	3422							**0.85 (0.78–0.93)**	**0.88 (0.78–0.98)**
Pressure to eat	3422							0.95 (0.87–1.04)	1.07 (0.95–1.20)
Nagelkerke's R^2		**0.08**		**0.14**		**0.15**		**0.27**	
LRT P-value		**<0.001**		**<0.001**		**<0.001**		**<0.001**	

BMI, body mass index; EDF, energy-dense foods dietary pattern; Nagelkerke's R^2, Nagelkerke's R-squared (Nagelkerke, 1991); Pt, points; LRT, likelihood ratio test. *Statistically significant associations are highlighted in bold. [†]Blocks of variables (socio-economic position at mothers' 12 years of age; maternal socio-economic and demographic characteristics at child's delivery; family characteristics at child's 4 years; and maternal characteristics at child's 4 years) were added sequentially into the analysis. Models are adjusted for child's characteristics (sex; daily screen time; weekly time spent practicing physical exercise; and exact age). [‡]Reference category is the Healthier dietary pattern ($n = 1538$), not shown to avoid redundancy. [§]Not family, kindergarten and nannies; parent, mostly mothers (95%); other family member, mostly grandparents (96%).

Question 2

- Socio-economic position at mother 12 years is considering that the DAG is correctly defined:
  - A) not associated to dietary pattern at 4 years
  - B) associated to dietary pattern at 4 years but is not causal effect
  - C) Causal effect of dietary pattern at 4 years

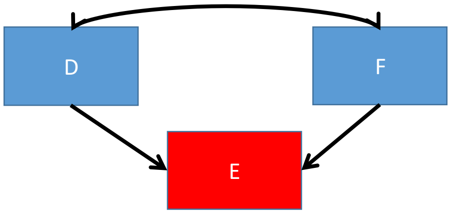
Confounders cannot be intermediate steps



Aspirin (A), platelet aggregation (B), AMI (Y)

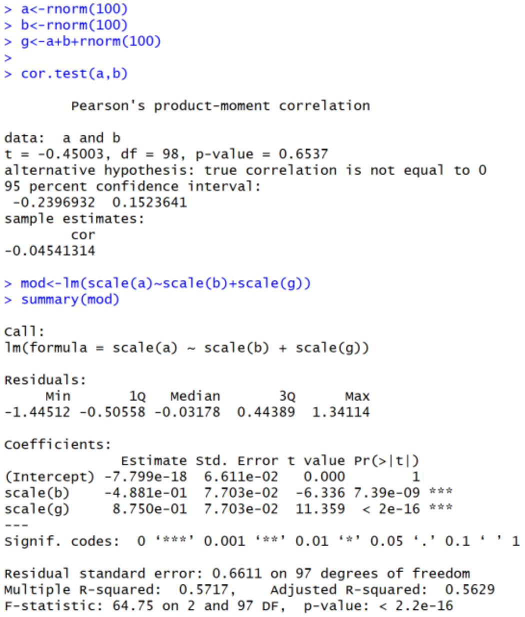
Blocking

- The opposite situation occurs if collider is conditioned on



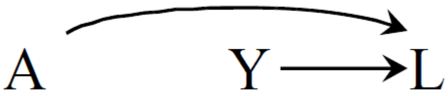
- Consider the path:
- A->G<-B
- Here, A and B are not associated via this path
- However, conditioning on G induces an association between A and B
  - Opens door between A->B

Blocking



```
> a<-rnorm(100)
> b<-rnorm(100)
> g<-a+b+rnorm(100)
>
> cor.test(a,b)

        Pearson's product-moment correlation

data:  a and b
t = -0.45003, df = 98, p-value = 0.6537
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2396932  0.1523641
sample estimates:
        cor
-0.04541314

> mod<-lm(scale(a)~scale(b)+scale(g))
> summary(mod)

Call:
lm(formula = scale(a) ~ scale(b) + scale(g))

Residuals:
     Min       1Q   Median       3Q      Max
-1.44512 -0.50558 -0.03178  0.44389  1.34114

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.799e-18  6.611e-02   0.000        1
scale(b)    -4.881e-01  7.703e-02  -6.336 7.39e-09 ***
scale(g)     8.750e-01  7.703e-02  11.359  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6611 on 97 degrees of freedom
Multiple R-squared:  0.5717,	Adjusted R-squared:  0.5629
F-statistic: 64.75 on 2 and 97 DF,  p-value: < 2.2e-16
```

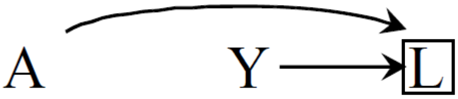
Collider: In which circumstances are A and Y causally related? And statistically associated?



Genotype (A), smoking (Y), AMI (L)

Does knowing A improve the prediction of Y?

Adjusting for a collider causes bias



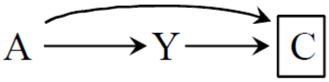
If L=1: in individuals with AMI, knowing that they do not smoke modifies the probability of them having a risk genotype

Genotype (A), smoking (Y), AMI (L)

Selecting L=1 (AMI present) originates an open path: A→L←Y, i.e., a source of statistical association – but we are only interested in the causal A→Y, which is null

RR causal ≠ RR association

Selection bias in study design – conditioning on a common effect

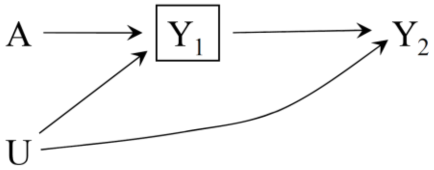


A: Folic acid; Y: congenital heart defect; C: fetal death

Selecting C=0 (live births) originates 2 open paths: A→Y and A→C←Y, i.e., two sources of statistical association – but we are only interested in the causal A→Y

RR causal ≠ RR association
How to solve?

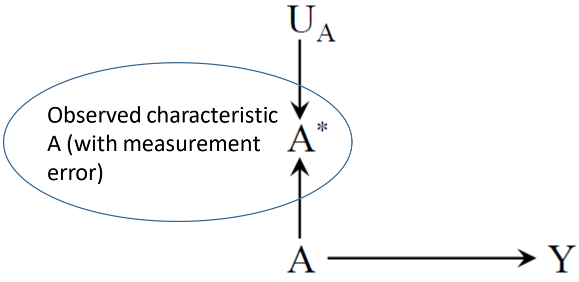
Selection bias in study analysis – conditioning on an intermediate step



A: parental education, Y1: young adult's education, Y2: young adult's back pain,
U: unmeasured organic condition

Conditioning on Y1 unblocks the path A→Y1←U→Y2, i.e. induces a statistical association when there is no causal relation from A to Y2

Information/misclassification bias



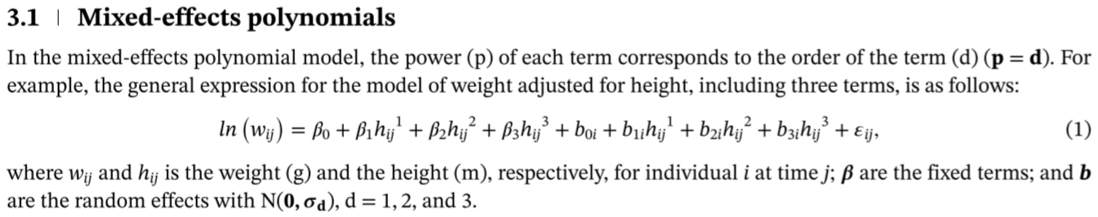
Measurement error of characteristic A

Observed characteristic A (with measurement error)

Construct

THE SCIENTIFIC COMMUNITY IS DIVIDED. SOME SAY THIS STUFF IS DANGEROUS, SOME SAY IT ISN'T.
Richter

Figure 20-4. Dealing with scientific uncertainty. (© The New Yorker Collection 1988. Mischa Richter from cartoonbank.com. All rights reserved.)

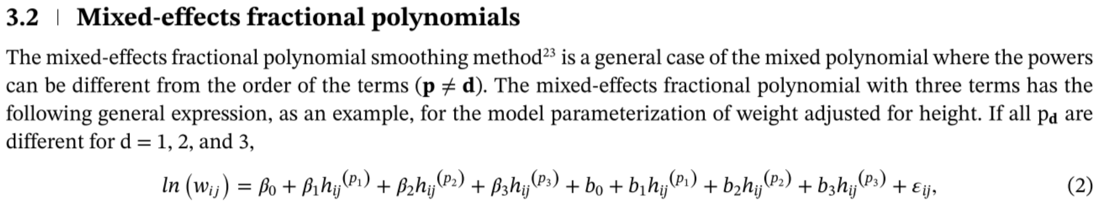


3.1 | Mixed-effects polynomials

In the mixed-effects polynomial model, the power (p) of each term corresponds to the order of the term (d) (**p** = **d**). For example, the general expression for the model of weight adjusted for height, including three terms, is as follows:

$$ln\left(w_{ij}\right) = \beta_0 + \beta_1 h_{ij}^{\ 1} + \beta_2 h_{ij}^{\ 2} + \beta_3 h_{ij}^{\ 3} + b_{0i} + b_{1i} h_{ij}^{\ 1} + b_{2i} h_{ij}^{\ 2} + b_{3i} h_{ij}^{\ 3} + \varepsilon_{ij}, \tag{1}$$

where w_{ij} and h_{ij} is the weight (g) and the height (m), respectively, for individual i at time j; $\boldsymbol{\beta}$ are the fixed terms; and $\boldsymbol{b}$ are the random effects with N($\mathbf{0}, \boldsymbol{\sigma}_{\mathbf{d}}$), d = 1, 2, and 3.

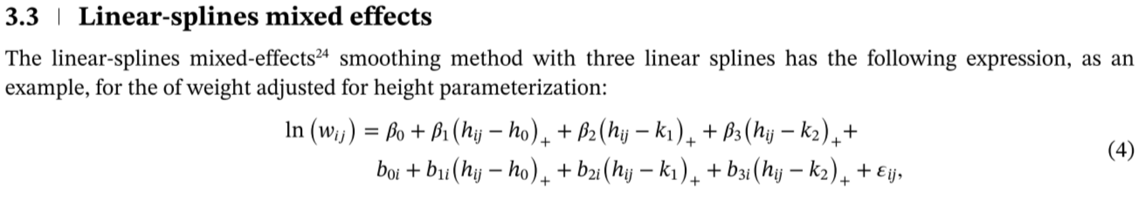


3.2 | Mixed-effects fractional polynomials

The mixed-effects fractional polynomial smoothing method[23] is a general case of the mixed polynomial where the powers can be different from the order of the terms ($\mathbf{p} \neq \mathbf{d}$). The mixed-effects fractional polynomial with three terms has the following general expression, as an example, for the model parameterization of weight adjusted for height. If all $p_{\mathbf{d}}$ are different for d = 1, 2, and 3,

$$ln\left(w_{ij}\right) = \beta_0 + \beta_1 h_{ij}^{\ (p_1)} + \beta_2 h_{ij}^{\ (p_2)} + \beta_3 h_{ij}^{\ (p_3)} + b_0 + b_1 h_{ij}^{\ (p_1)} + b_2 h_{ij}^{\ (p_2)} + b_3 h_{ij}^{\ (p_3)} + \varepsilon_{ij}, \tag{2}$$

Statistics in medicine, 2019



3.3 | Linear-splines mixed effects

The linear-splines mixed-effects[24] smoothing method with three linear splines has the following expression, as an example, for the of weight adjusted for height parameterization:

$$\begin{aligned}\ln\left(w_{ij}\right) = \beta_0 + \beta_1\left(h_{ij} - h_0\right)_+ + \beta_2\left(h_{ij} - k_1\right)_+ + \beta_3\left(h_{ij} - k_2\right)_+ + \\ b_{0i} + b_{1i}\left(h_{ij} - h_0\right)_+ + b_{2i}\left(h_{ij} - k_1\right)_+ + b_{3i}\left(h_{ij} - k_2\right)_+ + \varepsilon_{ij},\end{aligned} \tag{4}$$

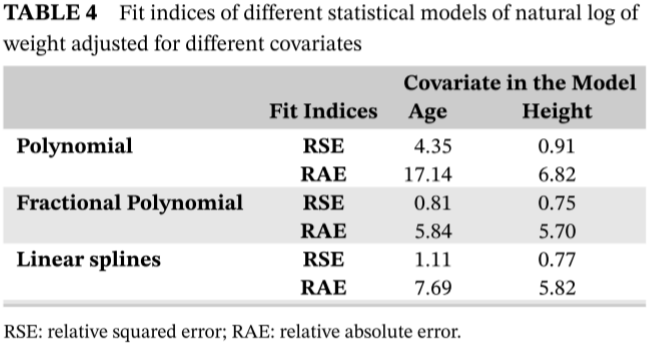


TABLE 4 Fit indices of different statistical models of natural log of weight adjusted for different covariates

		Covariate in the Model	
	Fit Indices	**Age**	**Height**
Polynomial	**RSE**	4.35	0.91
	RAE	17.14	6.82
Fractional Polynomial	**RSE**	0.81	0.75
	RAE	5.84	5.70
Linear splines	**RSE**	1.11	0.77
	RAE	7.69	5.82

RSE: relative squared error; RAE: relative absolute error.

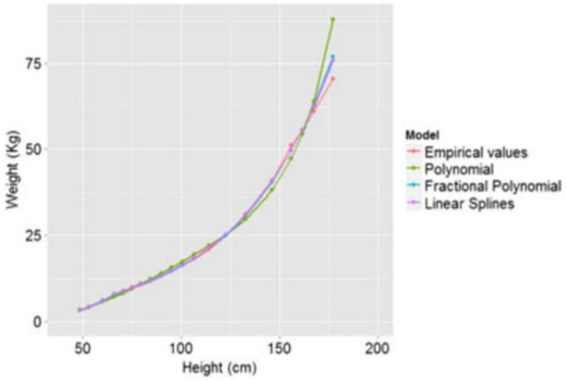
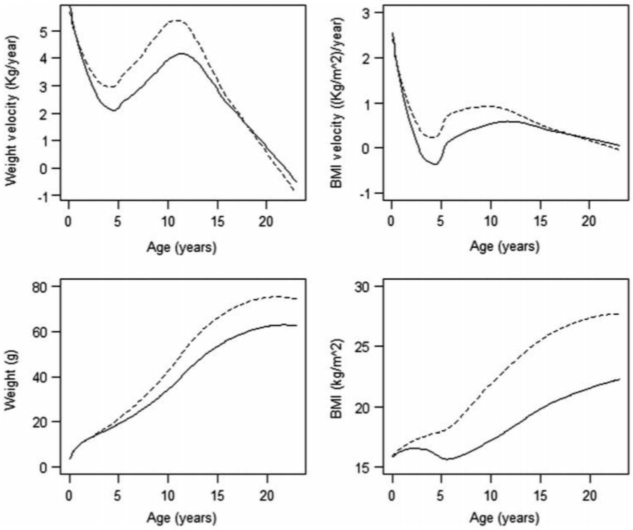


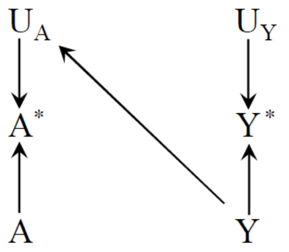
FIGURE 2 Growth curves for weight-for-height estimated from the three different statistical models (polynomial, fractional polynomial and linear splines), compared to the empirical mean values of the sample [Colour figure can be viewed at wileyonlinelibrary.com]



International Journal of Obesity **2015**

Figure 1. Weight velocity (top left), BMI velocity (top right), mean values of predicted weight (bottom left) and predicted BMI (bottom right) throughout age, according to the two trajectories identified. Solid line: 'Average BMI growth'; dashed line: 'Higher BMI growth'.

Information/misclassification bias



A: drug; Y: dementia (self-reported medication asked by interview)
or
A: alcohol intake during pregnancy; Y: birth defect

What are the traditional names of these biases?

Information/misclassification bias

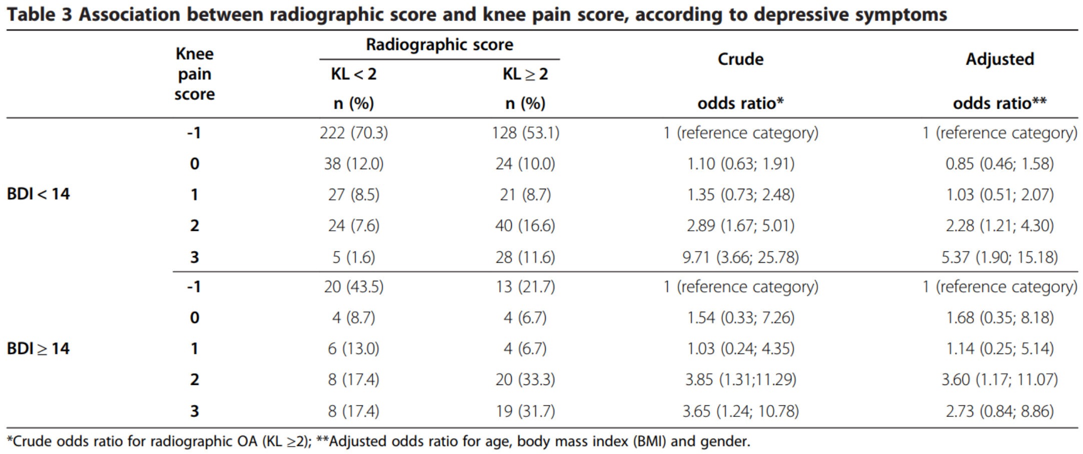
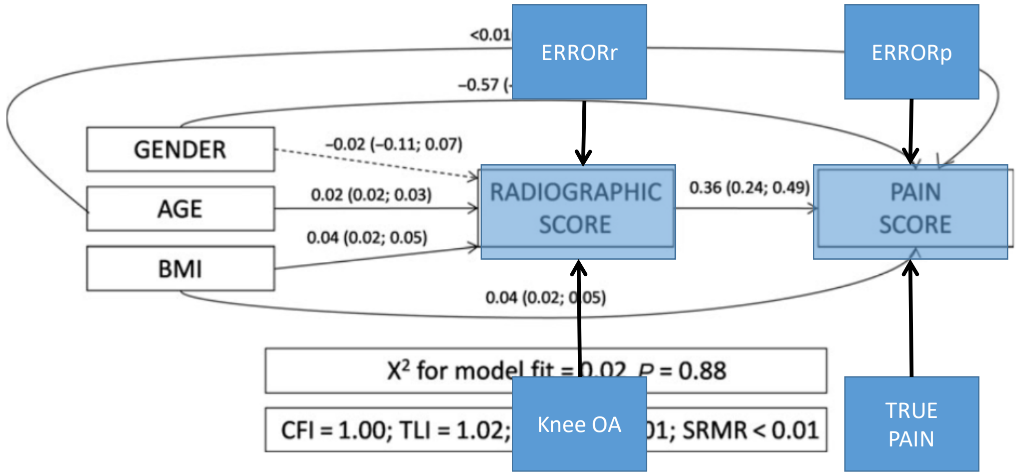


Table 3 Association between radiographic score and knee pain score, according to depressive symptoms

	Knee pain score	Radiographic score		Crude	Adjusted
		KL < 2	KL ≥ 2		
		n (%)	n (%)	odds ratio*	odds ratio**
BDI < 14	-1	222 (70.3)	128 (53.1)	1 (reference category)	1 (reference category)
	0	38 (12.0)	24 (10.0)	1.10 (0.63; 1.91)	0.85 (0.46; 1.58)
	1	27 (8.5)	21 (8.7)	1.35 (0.73; 2.48)	1.03 (0.51; 2.07)
	2	24 (7.6)	40 (16.6)	2.89 (1.67; 5.01)	2.28 (1.21; 4.30)
	3	5 (1.6)	28 (11.6)	9.71 (3.66; 25.78)	5.37 (1.90; 15.18)
BDI ≥ 14	-1	20 (43.5)	13 (21.7)	1 (reference category)	1 (reference category)
	0	4 (8.7)	4 (6.7)	1.54 (0.33; 7.26)	1.68 (0.35; 8.18)
	1	6 (13.0)	4 (6.7)	1.03 (0.24; 4.35)	1.14 (0.25; 5.14)
	2	8 (17.4)	20 (33.3)	3.85 (1.31;11.29)	3.60 (1.17; 11.07)
	3	8 (17.4)	19 (31.7)	3.65 (1.24; 10.78)	2.73 (0.84; 8.86)

*Crude odds ratio for radiographic OA (KL ≥2); **Adjusted odds ratio for age, body mass index (BMI) and gender.

Information/misclassification bias



International Journal of Rheumatic Diseases, 2017

Structural classification of bias

- Two variables are statistically associated when
  - One is a cause of the other ← our aim!
  - They share common causes – confounding
  - They share effects that have been conditioned on – selection bias
  - There is differential measurement error – information/misclassification bias

Bias – any structural association between exposure and outcome that does not arise from the causal effect of the exposure on the outcome

Rules for d-separation

- A path is d-separated by set of variables C if:
  - It contains a chain (D->E->F) and the middle part is in C
  - It contains a fork (D<-E->F) and the midle part is in C
  - It contains na inverted fork (D->E<-F) and the midle part is not in C, nor any descendants of it

D-separation, Pearl 1995 ("Moralization", Lauritzen 1990)

A path is a sequence of edges that connect two nodes in a graph. The path is said to be open or closed according to the following rules:

1. If no variable has been conditioned on, a path is blocked if and only if two edges collide along the path: L→A→Y is an open path but A→Y← L is a closed path: Y is called a collider
2. Any path that contains a non-collider that has been conditioned on is blocked: L→$\boxed{\text{A}}$→Y
3. Conditioning on a collider unblocks the path: A→$\boxed{\text{Y}}$← L
4. Conditioning on the descendant of a collider unblocks the path

Rules for d-separation

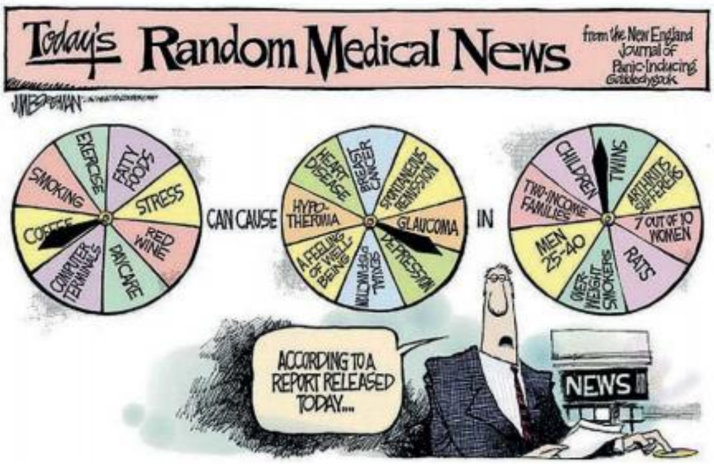
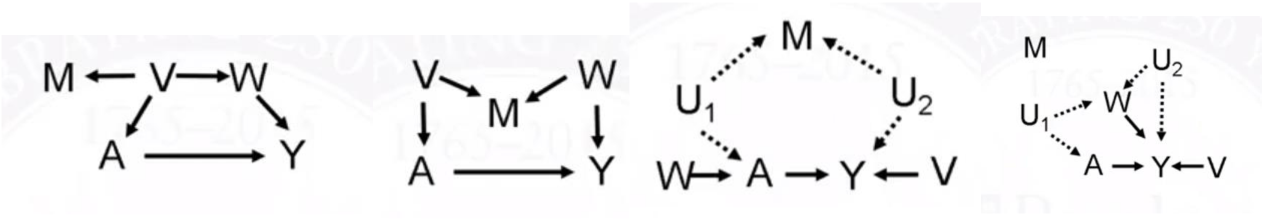


Figure 20-3. One view of the seemingly endless stream of reported risks confronting the public. (Jim Borgman. The Cincinnati Enquirer. 1997. Reprinted with special permission of King Features Syndicate.)

Variable selection

- The usual criterion would be adjusting for all variables (**Criterion 1**)
- The disjunctive cause criterion (VanderWeele 2011) (**Criterion 2**)
  - Control for all (observed) causes of exposure, outcome or both
- Researchers do not know the whole graph, but rather, the list of variables that affect the exposure or outcome



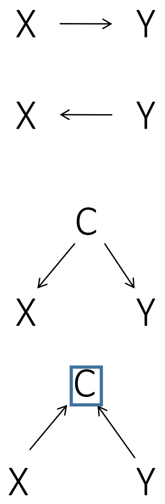
	DAG 1	DAG 2	DAG 3	DAG 4
Criterion 1	YES	YES	NO	NO
Criterion 2	YES	YES	YES	NO

Confounding – marginal and conditional independence

It is possible to identify causal relations when, between the exposure and the outcome,

- There are no common causes (RCTs)
- There are common causes but enough variables were measured that allow for blocking all backdoor paths – in that case it is said that there is confounding but no unmeasured confounding → Adjust, Stratify, Mactching, Standardize, etc.

Four ways of explaining a robust statistical association (Joffe 2010)



causation

reverse causation

common ancestor
(confounding)

conditioned common
descendant
(selection bias in design or analysis)

Causal diagrams

By representing

- Previous knowledge
- Assumptions

And applying a set of logical rules

It is possible to

- Understand the extent to which observed data are consistent with the causal model
- Predict expected statistical associations
- Detect logical problems and contradictions in data analysis

Acknowledge

- To Raquel Lucas for lending some of the slides

24 May 2019

Epidemiology and the causal enquire: the role of statistics

Milton Severo (milton@med.up.pt)