Lecture 5: Dealing with confounding DAGs and stratification

Brice Ozenne[1,2] - brice.mh.ozenne@gmail.com

[1] Section of Biostatistics, Department of Public Health, University of Copenhagen

[2] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

26 January 2023

Recap'

3 measures of disease frequency

- **Prevalence**: proportion of people with a disease

$$\hat{\pi} = \frac{\text{"number of people with the disease"}}{\text{"number of people"}}$$

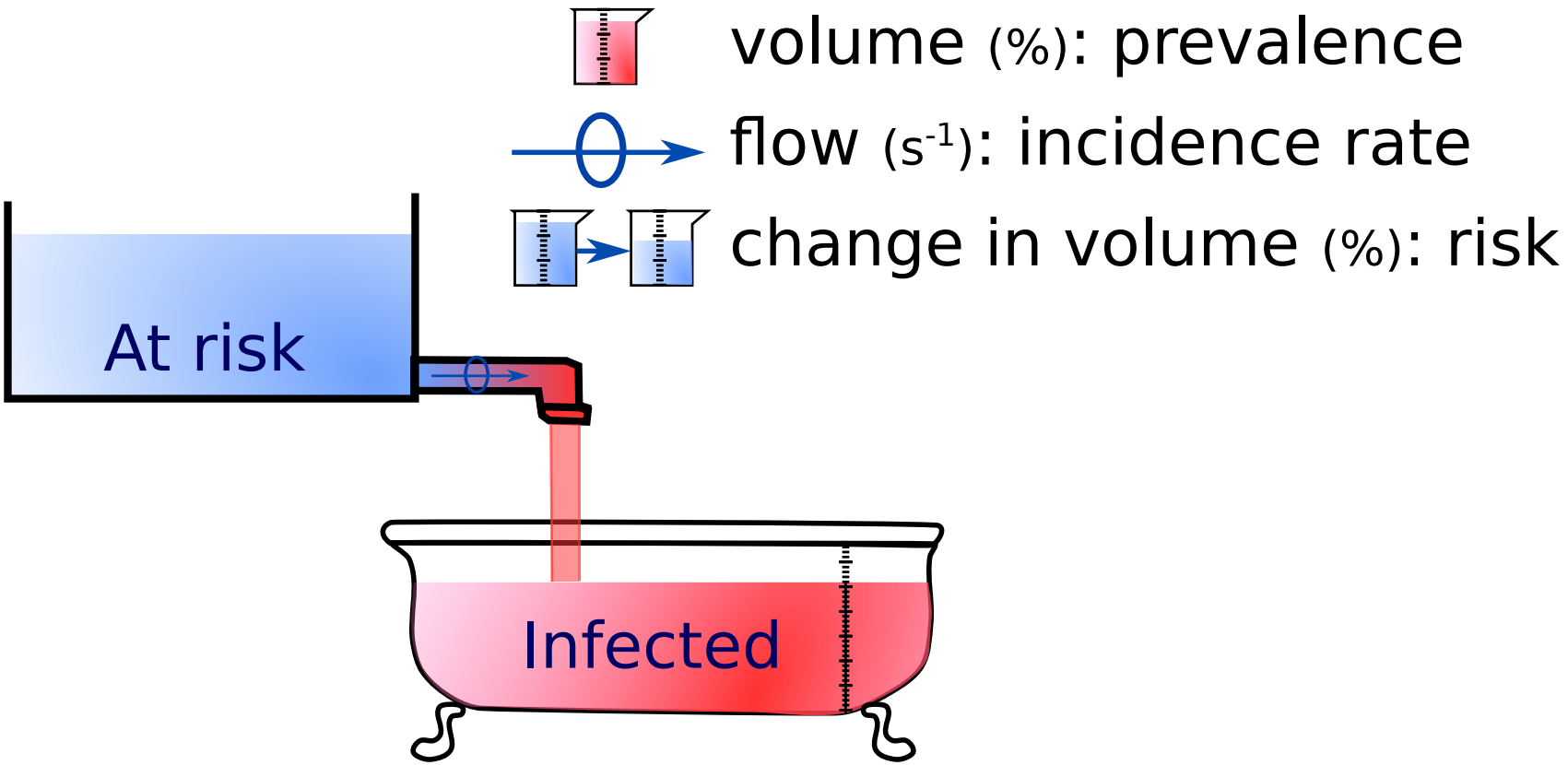
- **Incidence rate**: frequency of disease occurrence over period τ
  ⚠ unit: time^{-1}, e.g. person-year.

$$\widehat{\lambda}_\tau = \frac{\text{"number of new cases"}}{\text{"cumulative at risk time"}}$$

- **Risk**: probability of disease occurence between time 0 and τ

$$\widehat{r}(\tau) = \frac{\text{"number of new cases"}}{\text{"number of person at risk"}}$$

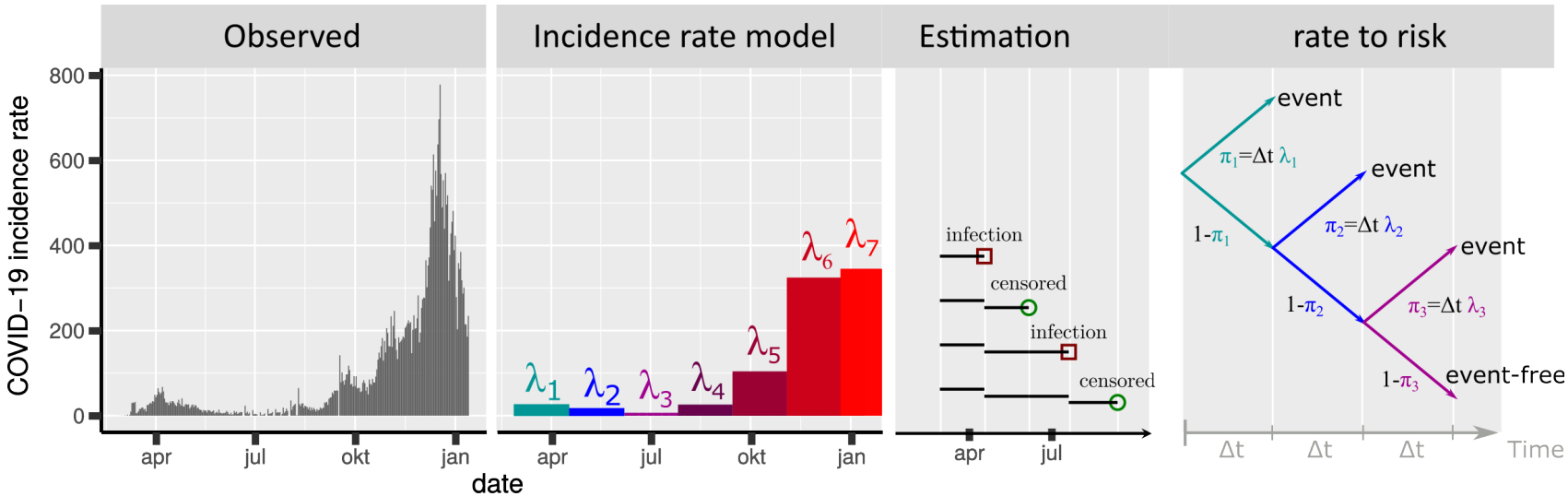
Risk rate relationship (1/2)



Assuming constant incidence rate:

- $r(\tau) = \exp(-\lambda\tau)$

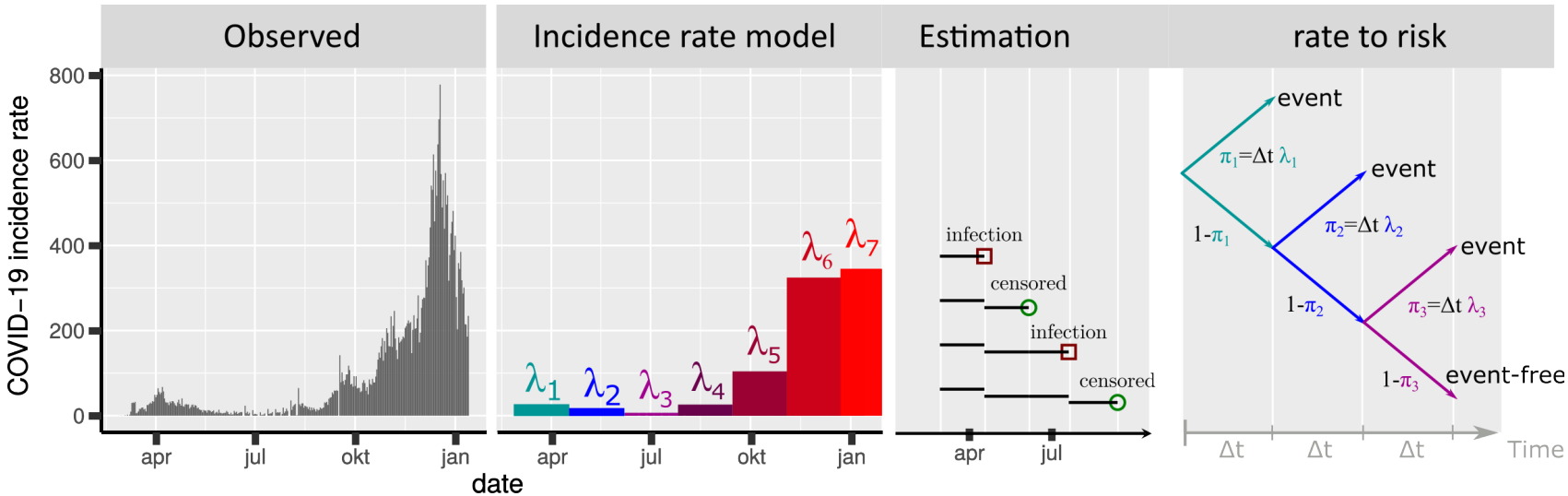
Risk rate relationship (2/2)



With varying incidence rates (3 time intervals):

$$r(\tau) = 1 - (1 - \lambda_1 \Delta t)(1 - \lambda_2 \Delta t)(1 - \lambda_3 \Delta t)$$
$$\approx 1 - \exp(-(\lambda_1 + \lambda_2 + \lambda_3)\Delta t)$$

Risk rate relationship (2/2)



With varying incidence rates (3 time intervals):

$$r(\tau) = 1 - (1 - \lambda_1 \Delta t)(1 - \lambda_2 \Delta t)(1 - \lambda_3 \Delta t)$$
$$\approx 1 - \exp(-(\lambda_1 + \lambda_2 + \lambda_3)\Delta t)$$

$\rightarrow$ useful to deal with right-censoring!

Comparing disease frequency across 2 groups

Group 2 vaccinated vs. Group 1 non-vaccinated

- **risk difference**: $RD(\tau) = r_2(\tau) - r_1(\tau)$
- **relative risk**: $RR(\tau) = \frac{r_2(\tau)}{r_1(\tau)}$
- **odds ratio**: $OR(\tau) = \left(\frac{r_2(\tau)}{1-r_2(\tau)}\right) \Big/ \left(\frac{r_1(\tau)}{1-r_1(\tau)}\right)$

Null hypothesis of **identical** risks: $RD = 0$, $RR = 1$, $OR = 1$

Estimation and confidence intervals: see `L2-summary.pdf`

Comparing disease frequency across 2 groups

Group 2 vaccinated vs. Group 1 non-vaccinated

- **risk difference**: $RD(\tau) = r_2(\tau) - r_1(\tau)$
- **relative risk**: $RR(\tau) = \frac{r_2(\tau)}{r_1(\tau)}$
- **odds ratio**: $OR(\tau) = \left(\frac{r_2(\tau)}{1-r_2(\tau)}\right) \Big/ \left(\frac{r_1(\tau)}{1-r_1(\tau)}\right)$

Null hypothesis of **identical** risks: $RD = 0$, $RR = 1$, $OR = 1$

Estimation and confidence intervals: see `L2-summary.pdf`

$\rightarrow$ how to account for covariates? Which covariates to consider?

Program for today

Why (mostly) worry about the bias

Definition of a causal effect

Identify bias using a graphical representation:

- introduction of directed acyclic graphs (**DAGs**)
- definition of **confounder**, **collider**, mediator, risk factor

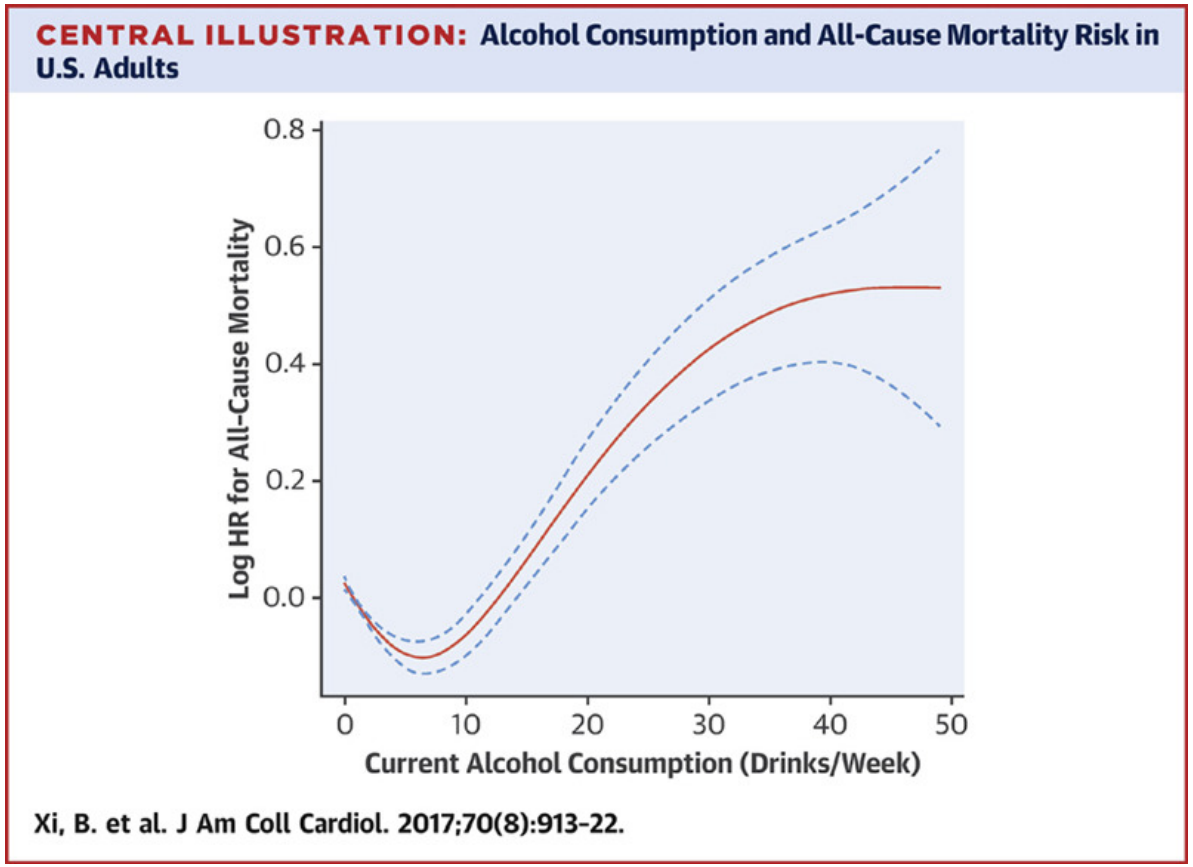
Controling for confounding:

- randomization, restriction
- **stratification** (full vs. common effect)

Alcohol J shape paradox

Prior knowledge:

- lifetime alchool consumption influences the risk of death



- is light alcohol consumption beneficial?

Error decomposition

what can go wrong?

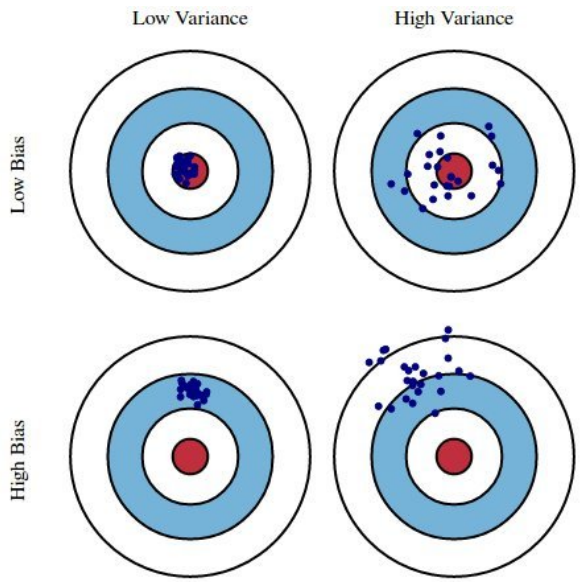


Illustration: bias vs. variance

Aim: relate healthy status Y to lifetime alchool consumption X

- $Y = \beta X + \varepsilon$

But we only observe the current alchool consumption Z:

- $Z = X + \xi$ (proxy for X)

Illustration: bias vs. variance

Aim: relate healthy status Y to lifetime alchool consumption X

- $Y = \beta X + \varepsilon$

But we only observe the current alchool consumption Z:

- $Z = X + \xi$ (proxy for X)

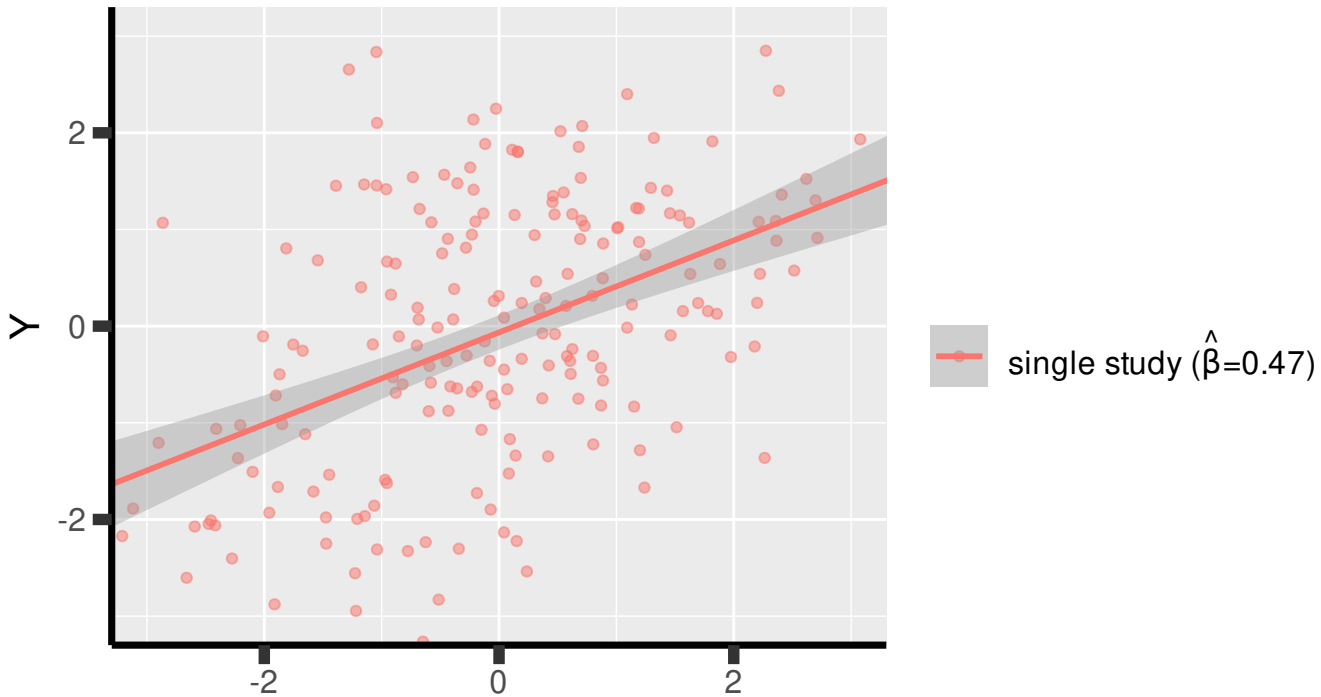


Illustration: bias vs. variance

Aim: relate healthy status Y to lifetime alchool consumption X

- $Y = \beta X + \varepsilon$

But we only observe the current alchool consumption Z:

- $Z = X + \xi$ (proxy for X)

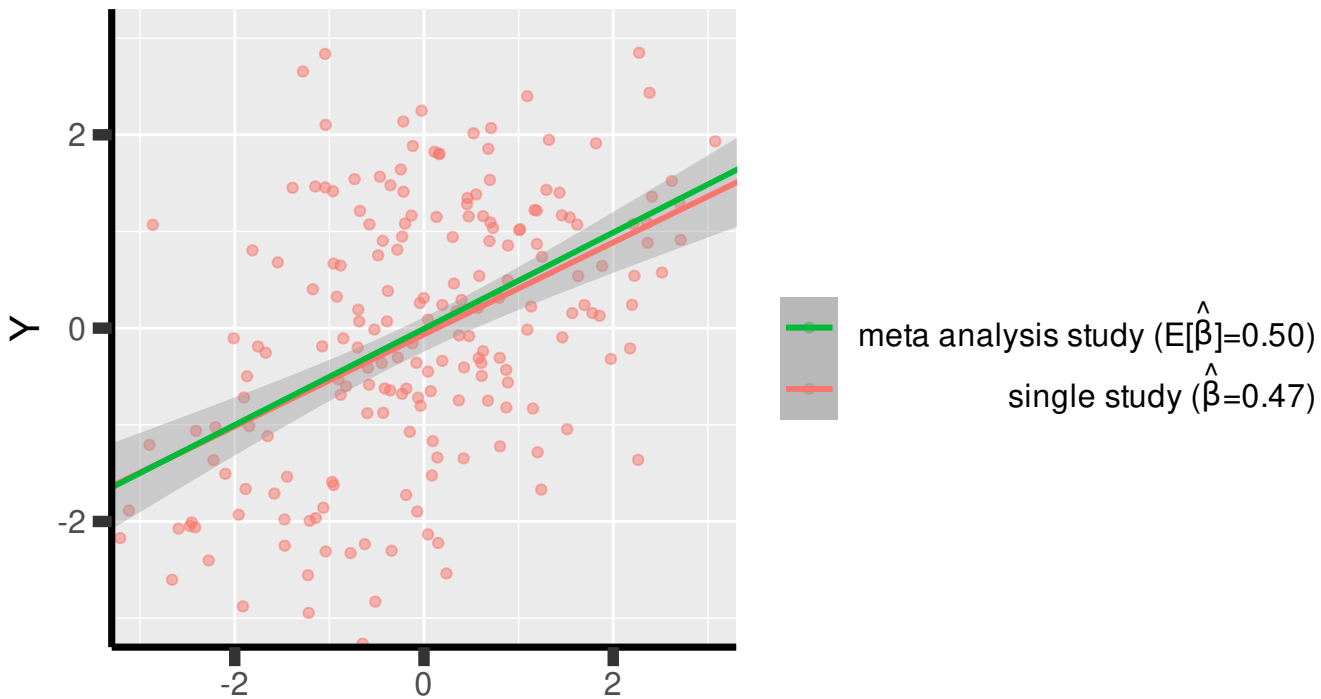


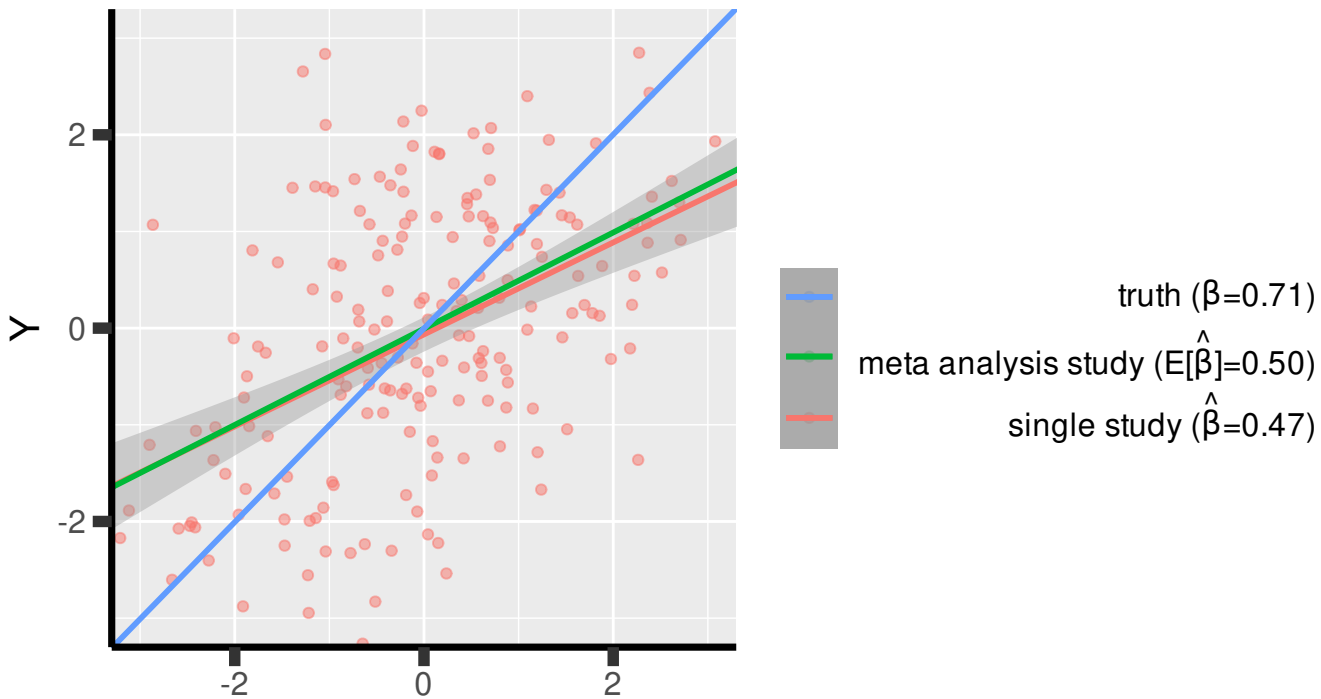
Illustration: bias vs. variance

Aim: relate healthy status Y to lifetime alchool consumption X

- $Y = \beta X + \varepsilon$

But we only observe the current alchool consumption Z:

- $Z = X + \xi$ (proxy for X)



Error decomposition

- β: parameter of interest (also called population parameter)
  deterministic quantity (i.e. fixed value)
- $\widehat{\beta}$: estimated value
  random quantity (i.e. vary from study to study)
- $\mathbb{E}\left[\widehat{\beta}\right]$: expected estimated value
  deterministic quantity (i.e. fixed value)

Error decomposition

- β: parameter of interest (also called population parameter)
  deterministic quantity (i.e. fixed value)
- $\widehat{\beta}$: estimated value
  random quantity (i.e. vary from study to study)
- $\mathbb{E}\left[\widehat{\beta}\right]$: expected estimated value
  deterministic quantity (i.e. fixed value)

The error can be decomposed in two terms:

$$\widehat{\beta} - \beta = \underbrace{\widehat{\beta} - \mathbb{E}\left[\widehat{\beta}\right]}_{\text{sampling error}} + \underbrace{\mathbb{E}\left[\widehat{\beta}\right] - \beta}_{\text{bias}}$$

Error decomposition

Bias: systematic difference between estimated and true parameter.

$\rightarrow$ stable across replication studies, here $\mathbb{E}\left[\widehat{\beta}\right] = \frac{\beta}{1+\sigma_\xi^2/\sigma_X^2}$

σ_ξ^2 variance of the mismatch between X and Z

Sampling error: random fluctuation in the estimated quantity

$\rightarrow$ due to the finite number of samples, here $\mathbb{V}ar\left[\widehat{\beta}\right] = \frac{\sigma_\varepsilon^2}{n\sigma_Z^2}$

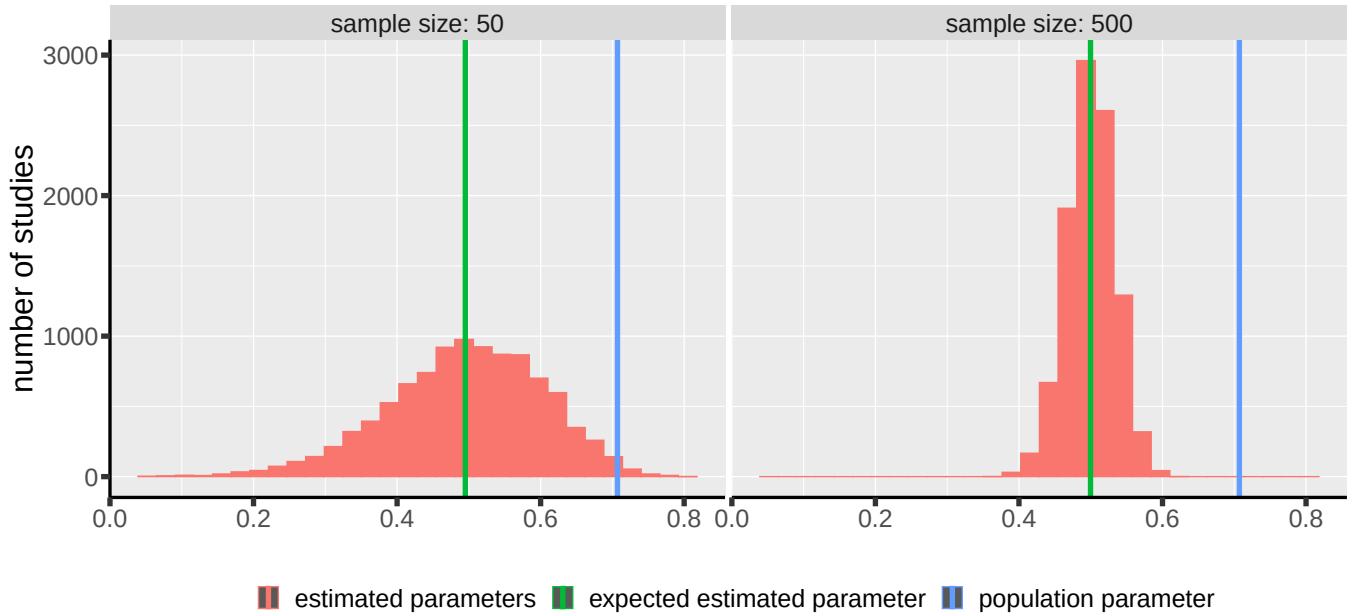
$\rightarrow$ differ from study to study

$\rightarrow$ can be estimated

The error can be decomposed in two terms:

$$\widehat{\beta} - \beta = \underbrace{\widehat{\beta} - \mathbb{E}\left[\widehat{\beta}\right]}_{\text{sampling error}} + \underbrace{\mathbb{E}\left[\widehat{\beta}\right] - \beta}_{\text{bias}}$$

Impact of the sample size



Sampling error can be reduced by:

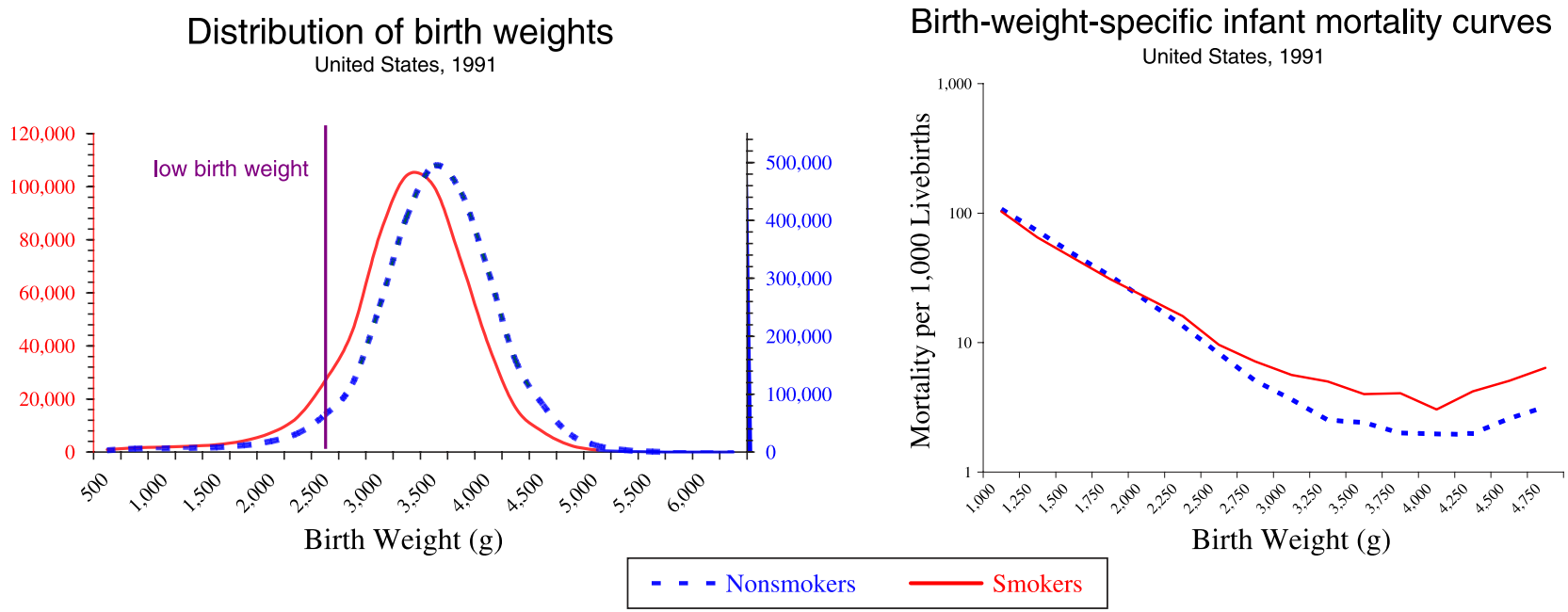
- replicating a study with a larger size
- pooling data from several studies

Primary concern is (generally) the bias

Birth weight paradox

Birth weight (BW) is a strong predictor of infant mortality

- investigators stratify on BW when evaluating risk factors



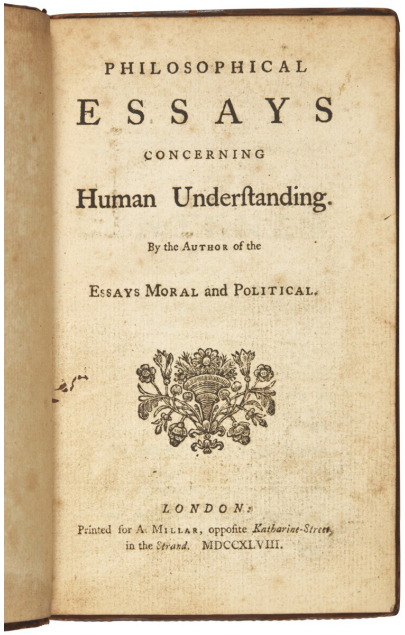
This leads to an apparent paradox (Hernández-Díaz et al., 2006)

- is maternal smoking beneficial? Sometimes beneficial?

Causality

what do we mean by 'beneficial'?

or (positive/negative) 'causal effect'?


PHILOSOPHICAL

ESSAYS

CONCERNING

Human Understanding.

By the Author of the

Essays Moral and Political.

LONDON:

Printed for A. Millar, opposite *Katharine-Street*, in the *Strand*. MDCCXLVIII.

Causation in the epidemiological literature

Various definitions for a cause of death:

- **Production**: play an essential part in death.
- **Necessary cause**: without which death cannot occur.
- **Sufficient component cause**: guarantees death will occur (alone or in conjunction with other causes).
- **Probabilistic cause**: increases the probability of death
- **Counterfactual cause**: makes a difference in death occurence when it is present compared with when it is absent, while all else is held constant.

(adapted from Parascandola and Weed (2001))

Causation in the epidemiological literature

Various definitions for a cause of death:

- **Counterfactual cause**: makes a difference in death occurence when it is present compared with when it is absent, while all else is held constant.

(adapted from Parascandola and Weed (2001))

Counterfactual outcomes

- outcome $Y \in \{0, 1\}$
- exposure $E \in \{0, 1\}$

Example: baby i died within a year ($Y_i = 1$)
and its mother was smoking ($E_i = 1$)

Counterfactual outcomes

- outcome $Y \in \{0, 1\}$
- exposure $E \in \{0, 1\}$

Example: baby i died within a year ($Y_i = 1$)
and its mother was smoking ($E_i = 1$)

- potential outcome Y^E

had his mother not smoked, he would be alive ($Y_i^{E=0} = 0$)
had his mother smoked, he would have died ($Y_i^{E=1} = 1$)

Counterfactual outcomes

- outcome $Y \in \{0, 1\}$
- exposure $E \in \{0, 1\}$

Example: baby i died within a year ($Y_i = 1$)
and its mother was smoking ($E_i = 1$)

- potential outcome Y^E

had his mother not smoked, he would be alive ($Y_i^{E=0} = 0$)
had his mother smoked, he would have died ($Y_i^{E=1} = 1$)

Consistency assumption (well defined intervention)

$Y^{E=e} = y$ when observing outcome y under exposure e

⚠ not well defined when the outcome depends on other subject exposure (e.g. risk of COVID without vaccination)

Counterfactual definition of a causal effect

- Individual causal effect:

$$\beta_i = Y_i^{E=1} - Y_i^{E=0}$$

"A cause of a disease event is an event [...] without which the disease event either would not have occurred at all or would not have occurred until some later time" (Rothman and Greenland, 2005)

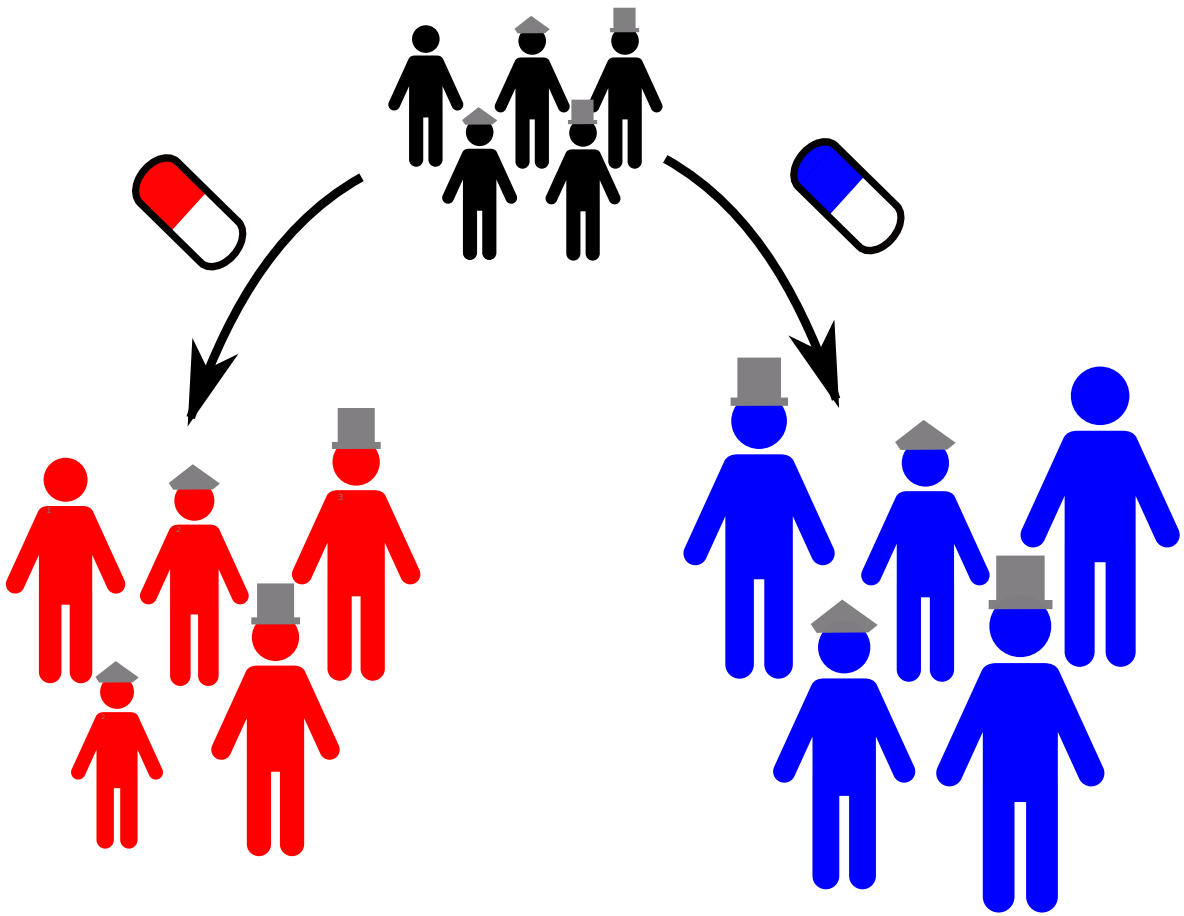
- Average causal effect:
  (average the individual causal effect over the population)

$$\beta = \mathbb{E}[\beta_i] = \mathbb{P}\left[Y^{E=1} = 1\right] - \mathbb{P}\left[Y^{E=0} = 1\right]$$

Positivity assumption

Non-0 probability of receiving either treatment

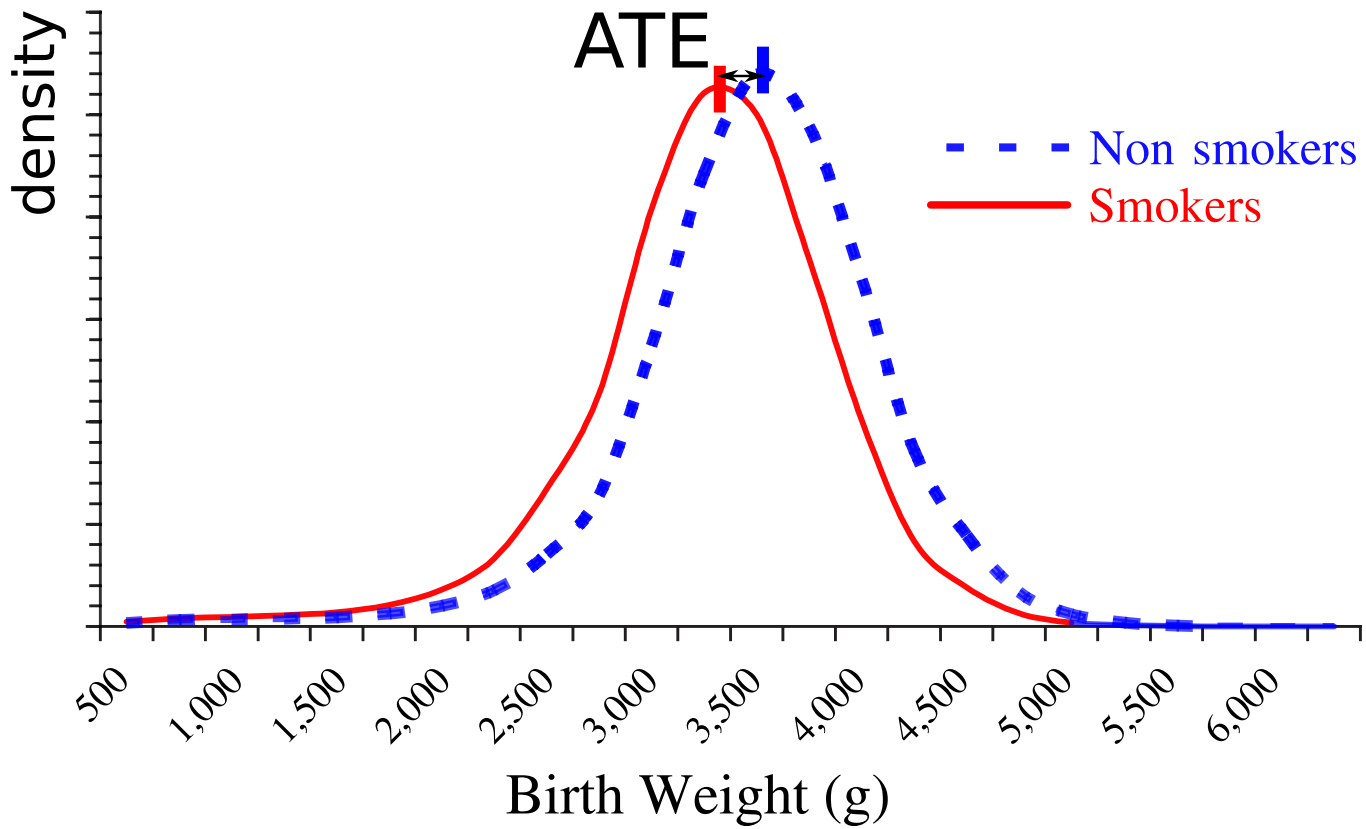
Average causal effect - illustration



⚠ we only observe either $Y_i^{E=0}$ or $Y_i^{E=1}$!

Causal effect with stochastic events

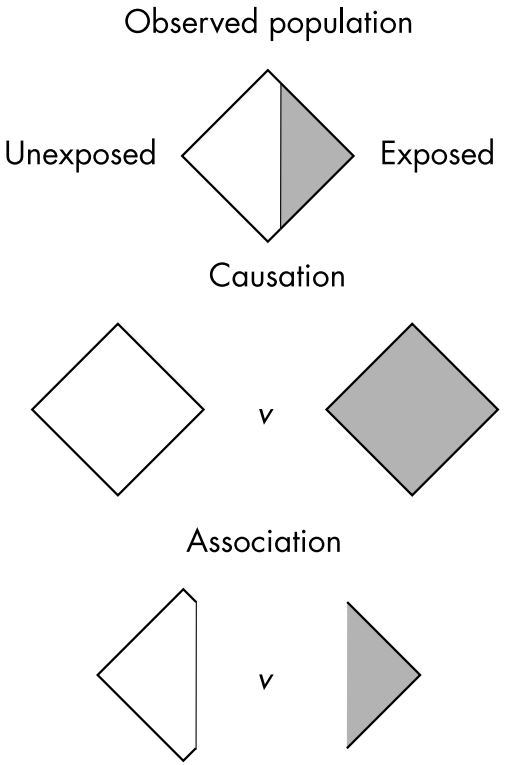
Causal effect is modeled through change in distribution (instead of value)



Estimation of the average causal effect

Use the observed probability

$$\widehat{\beta} = \mathbb{P}[Y = 1|E = 1] - \mathbb{P}[Y = 1|E = 0]$$



(Hernán, 2004)

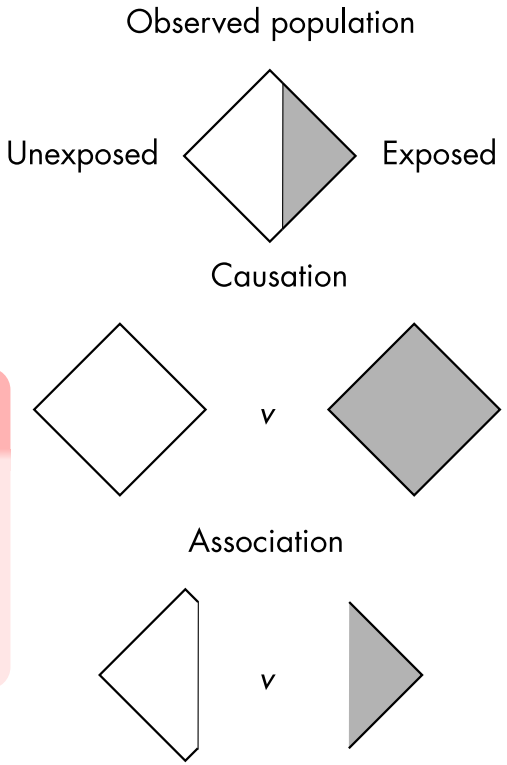
Estimation of the average causal effect

Use the observed probability

$$\widehat{\beta} = \mathbb{P}[Y = 1|E = 1] - \mathbb{P}[Y = 1|E = 0]$$

Exchangeability assumption

The actual exposure does not predict the counterfactual outcome
$Y^{E=e} \perp\!\!\!\perp E$



(Hernán, 2004)

Estimation of the average causal effect

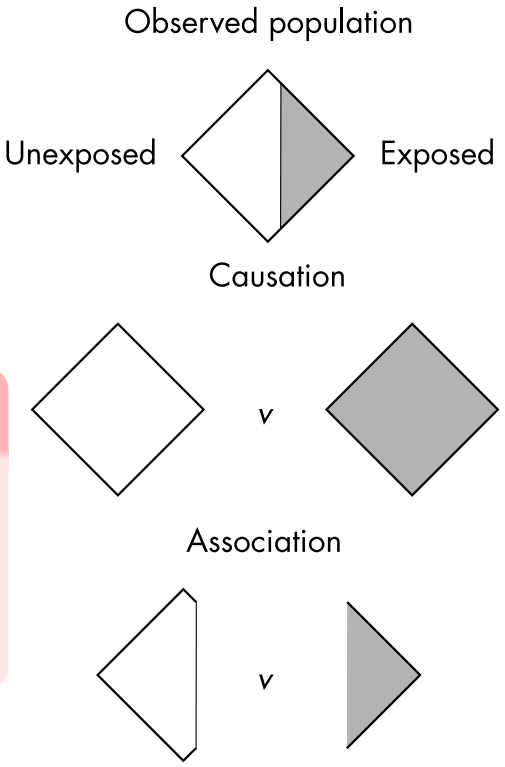
Use the observed probability

$$\widehat{\beta} = \mathbb{P}[Y = 1|E = 1] - \mathbb{P}[Y = 1|E = 0]$$

Exchangeability assumption

The actual exposure does not predict the counterfactual outcome
$Y^{E=e} \perp\!\!\!\perp E$

Are low BW babies of smokers vs non-smokers exchangeable?

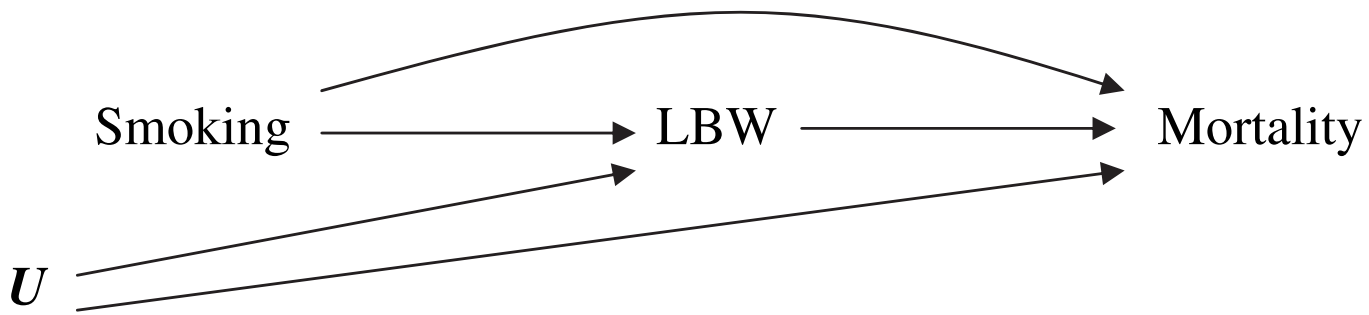


(Hernán, 2004)

DAGs

graphical representation of a system of variables

graphical criteria for exchangeability



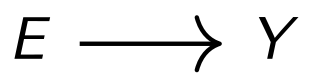
Causal associations (1/2)

$$E \longrightarrow Y$$

- changing E changes the distribution of Y

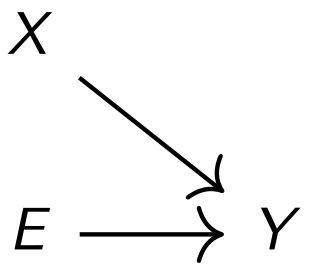
With the treatment, the risk of stroke is divided by 2 (distribution of "time to stroke" shifted toward longer times)

Causal associations (1/2)



- changing E changes the distribution of Y

```
With the treatment, the risk of stroke is divided by 2
(distribution of "time to stroke" shifted toward longer
times)
```



- for at least one x, changing E changes the distribution of Y when X is fixed at x.

```
With this preventive treatment, the risk of stroke:
  - is divided by 2 for patients with diabetes
  - unchanged otherwise
```

Causal associations (2/2)

$$E \longrightarrow M \longrightarrow Y$$

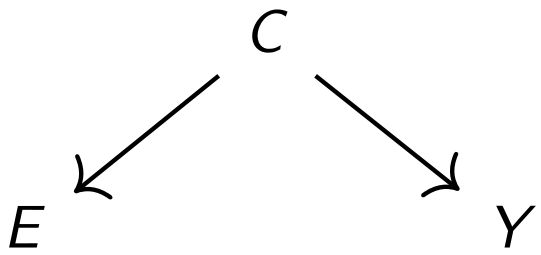
- E changes the distribution of M; **that** change in distribution of M changes the distribution of Y.

```
Preventive treatment
→ reduces your blood's ability to clot
→ decreases the risk of stroke.
```

Non-causal associations (Fork)

Causal: Getting older lead to higher risk of death and gray hair.

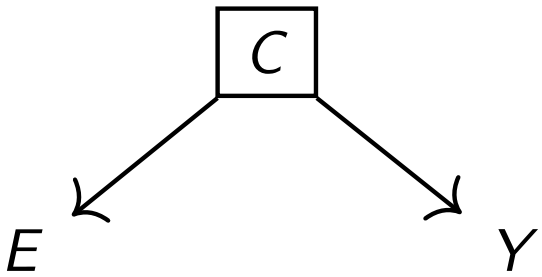
- unconditional



(open path)

Non-causal: Gray hair is associated with a higher risk of death.

- conditional



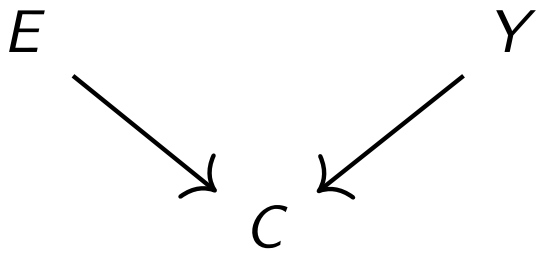
(closed path)

Causal: At a given age, there is no association between gray hair and risk of death.

Non-causal associations (Inverted Fork)

Causal: to be in this hospital (C), you must either have diabetes (E) or prostate cancer (Y).

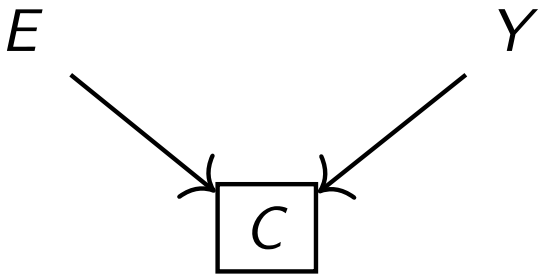
- unconditional



(closed path)

Causal: diabetes and prostate cancer are two unrealted conditions

- conditional



(open path)

Non-causal: among in-hospital patient there is a (negative) assocation between diabetes and prostate cancer

DAG

Directed:

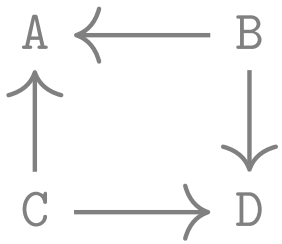
- each edge is oriented, i.e. represent a causal relationship.

Acyclic:

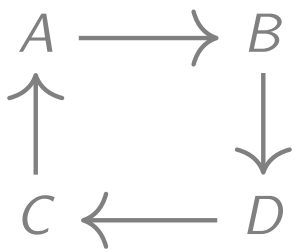
- does not contain any cycle

Graph:

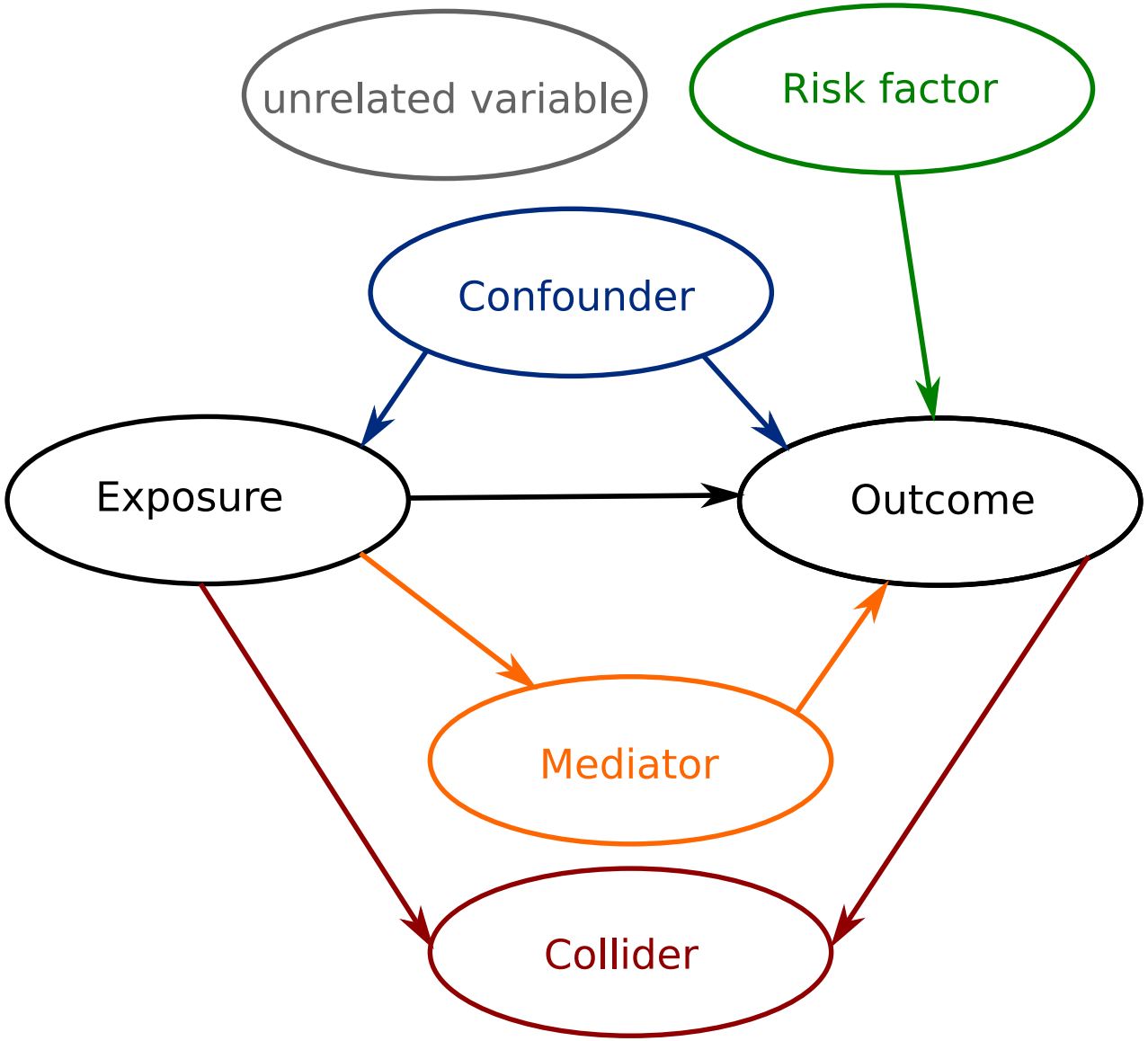
- graphical representation composed of vertices (variables) and edges (connection between variables).

Example: 

is a DAG

but not $A - B - C$ nor 

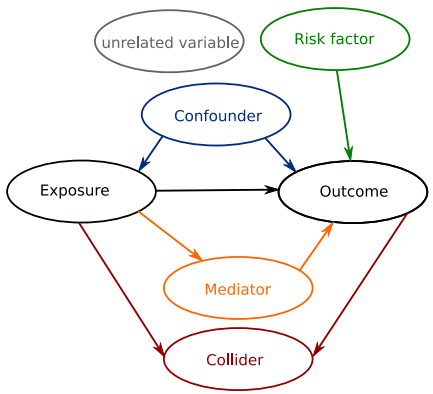
Nomenclature of the variables in a simple DAG



What to control for?

We would like use Z to:

- leave all directed paths between E and Y unperturbed
- block all spurious paths between E and Y
- create no new spurious paths between E and Y

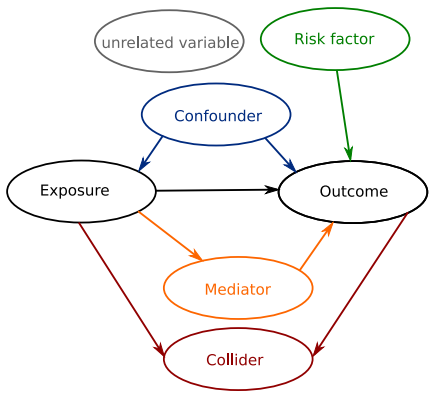


What to control for?

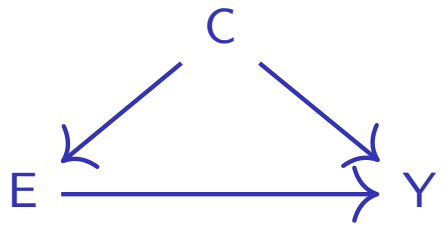
We would like use Z to:

- leave all directed paths between E and Y unperturbed
- block all spurious paths between E and Y
- create no new spurious paths between E and Y

- **Risk factor**: yes - efficiency gain
- **Confounder**: yes - otherwise bias
- **Collider**: no - otherwise bias
- **Mediator**: depends on the question:
  - adjustment: direct causal effect
  - no adjustment: total causal effect
- **Unrelated variable**: if possible not



Assessing the presence of confounding



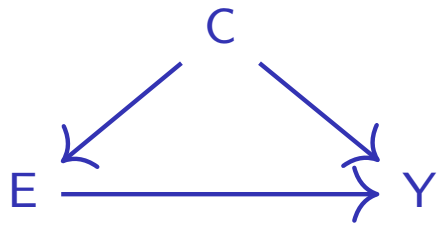
So far:

- a priori knowledge to decide on confounding (i.e. create the DAG)

What about using the data at hand?

- testing for C-Y or C-E association
- if not statistically significant ...

Assessing the presence of confounding



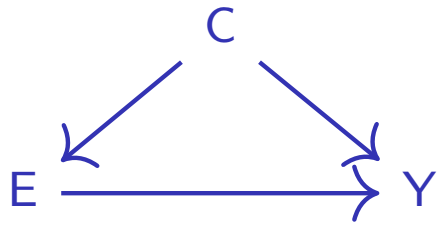
So far:

- a priori knowledge to decide on confounding (i.e. create the DAG)

What about using the data at hand?

- testing for C-Y or C-E association
- if not statistically significant . . . this does not help to decide!
  ⚠ Absence of evidence is not evidence of absence ⚠

Assessing the presence of confounding



So far:

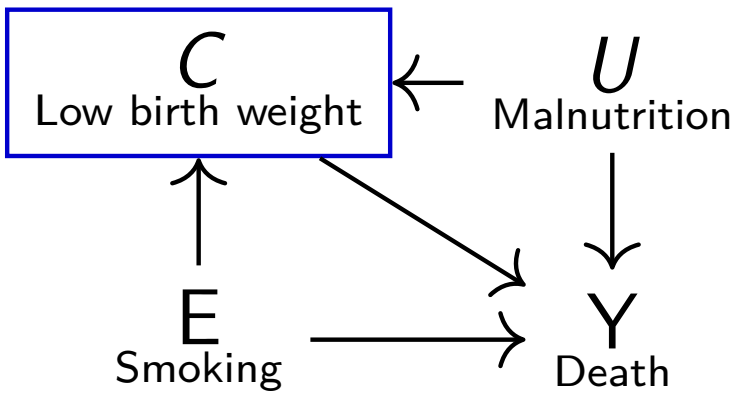
- a priori knowledge to decide on confounding
  (i.e. create the DAG)

What about using the data at hand?

- testing for C-Y or C-E association
- if not statistically significant . . . this does not help to decide!
  ⚠ Absence of evidence is not evidence of absence ⚠
- you can instead look at the confidence interval
  (narrow around 0?)

Analyzing complex DAGs

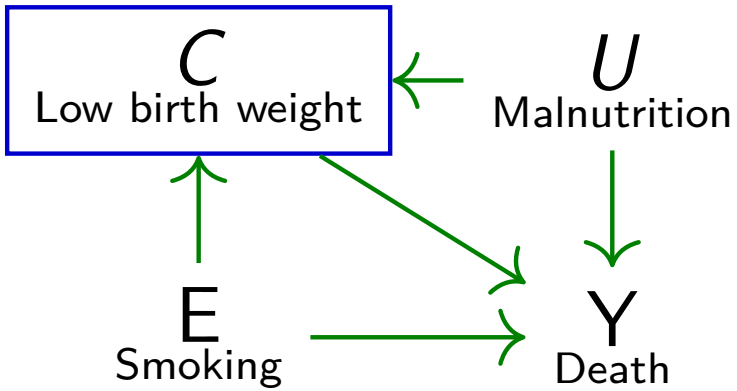
Here is a possible DAG for the birth weight paradox:



Procedure to assess causality:

- list all undirected paths from E to Y
- decide whether it is or not a causal path
- check that: - all causal paths are open/unblocked
  - all non-causal paths are closed/blocked

DAGs path by path - conditional on smoking

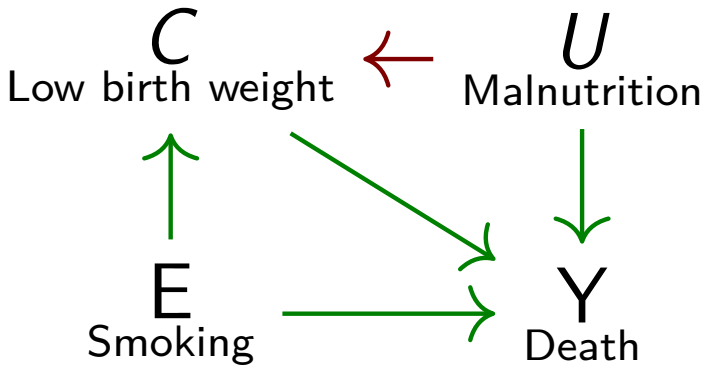


Path	Type of path	Status of the path
$E \rightarrow Y$	Causal	Open
$E \rightarrow \boxed{C} \rightarrow Y$	Causal	Closed
$E \rightarrow \boxed{C} \leftarrow U \rightarrow Y$	Non-causal	Open

because

$E \rightarrow \boxed{C} \leftarrow U$	Non-causal	Open
$\boxed{C} \leftarrow U \rightarrow Y$	Non-causal	Open

DAGs path by path - unconditional

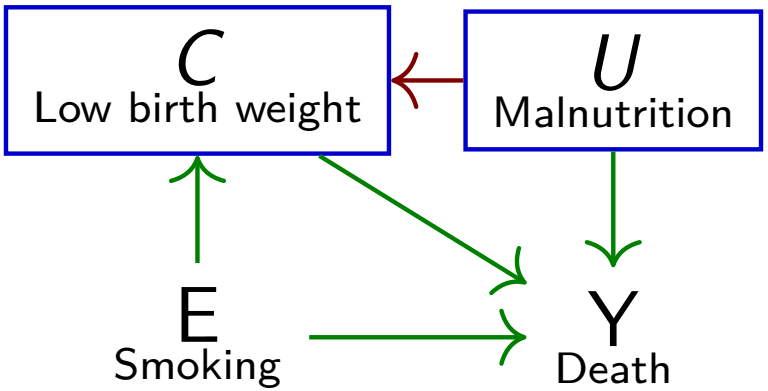


Path	Type of path	Status of the path
$E \rightarrow Y$	Causal	Open
$E \rightarrow C \rightarrow Y$	Causal	Open
$E \rightarrow C \leftarrow U \rightarrow Y$	Non-causal	Closed

because

$E \rightarrow C \leftarrow U$	Non-causal	Closed
$C \leftarrow U \rightarrow Y$	Non-causal	Open

DAGs path by path - adjustment

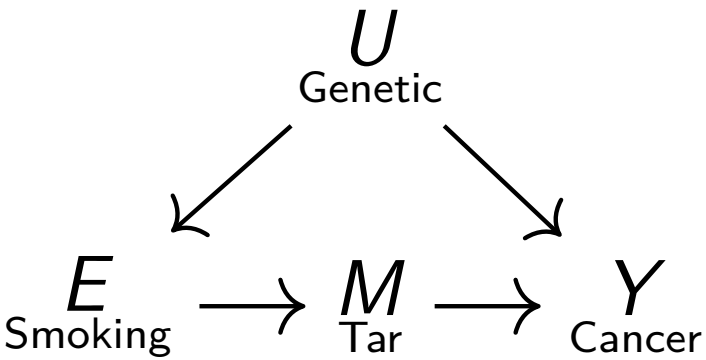


Path	Type of path	Status of the path
$E \rightarrow Y$	Causal	Open
$E \rightarrow \boxed{C} \rightarrow Y$	Causal	Closed
$E \rightarrow \boxed{C} \leftarrow \boxed{U} \rightarrow Y$	Non-causal	Closed

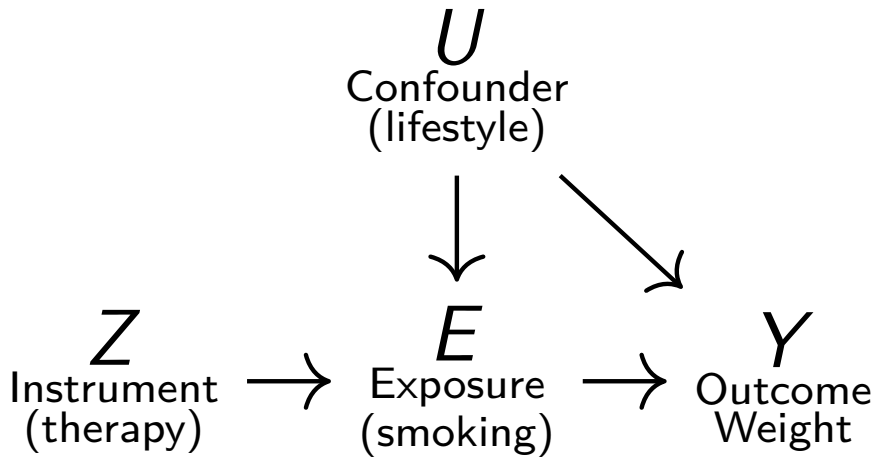
because

$E \rightarrow \boxed{C} \leftarrow \boxed{U}$	Non-causal	Open
$\boxed{C} \leftarrow \boxed{U} \rightarrow Y$	Non-causal	Closed

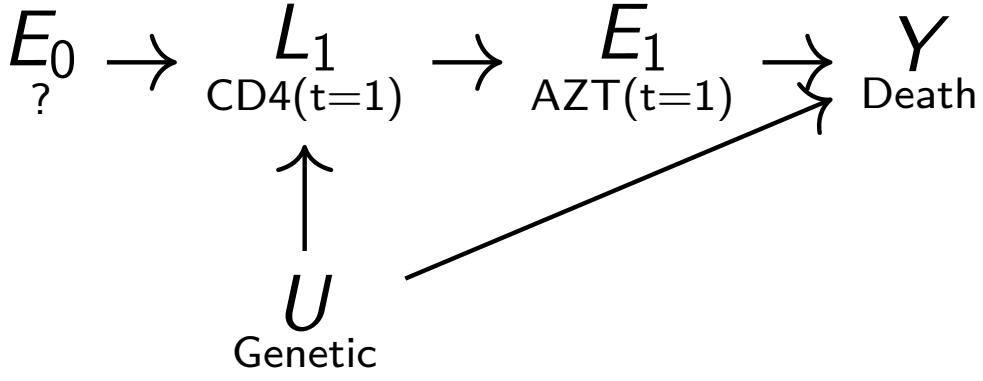
Handling confounding using causal inference



Front door criteria



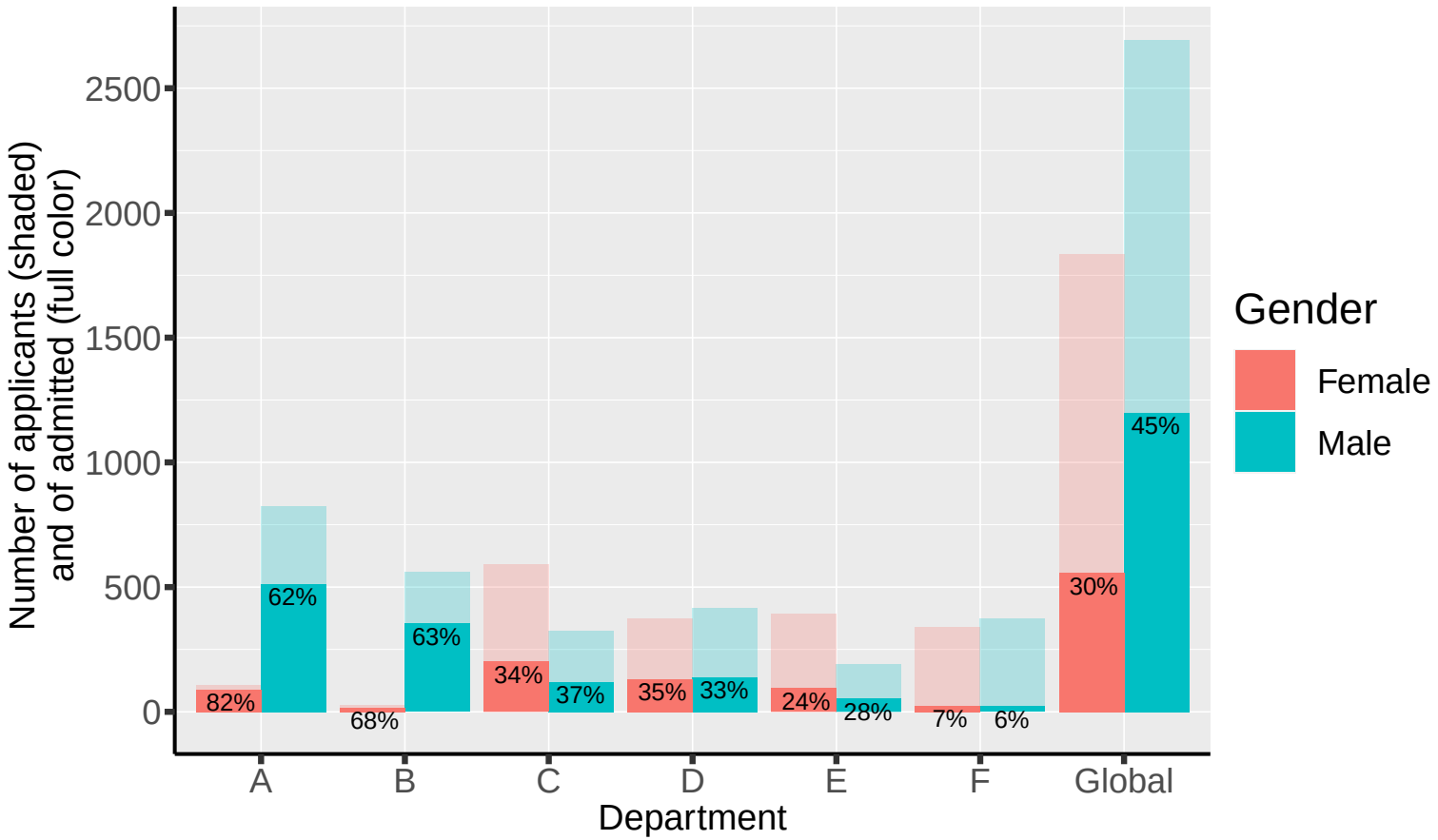
Instrumental variables



Time varying confounding

Simpson paradox

Graduate school admissions to UC Berkeley, fall of 1973 (Bickel et al., 1975)

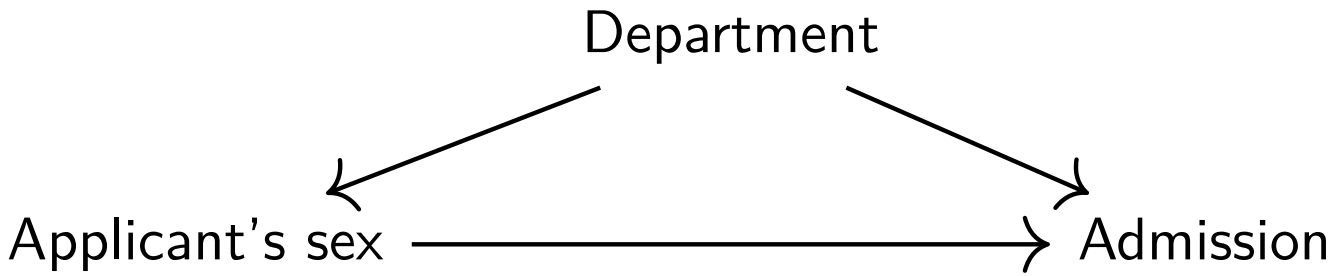


- Are females less likely to be admitted than males?

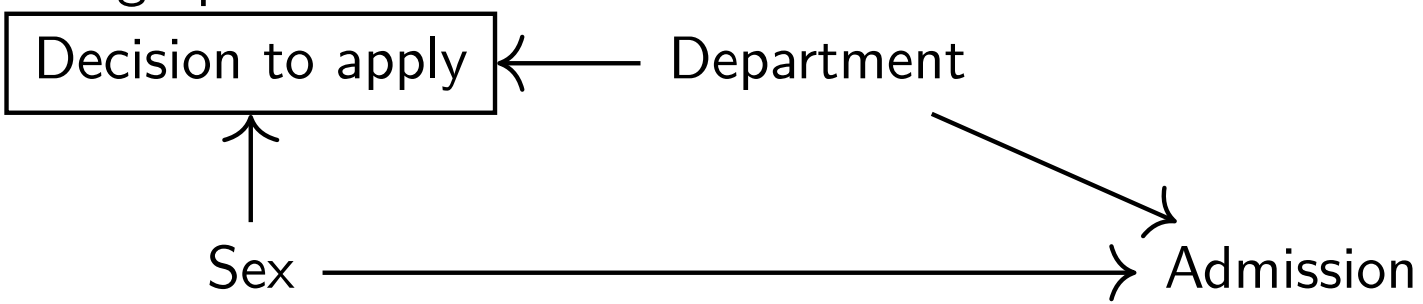
DAG of the Simpson paradox

Simpson paradox: confounding (+ collider)

Simplified graph:



Full graph:



How to adjust for "Department" in the analysis?

Controlling for confounding:

- by design
- using stratification

Restriction

Only include participants with a specific value of a variable.

- DAG: condition on the variable, remove arrow to descendants

`Example: only include females in the study`

- done in nearly all studies. Balance between:
  - controling confounding
  - feasibility, generalizability

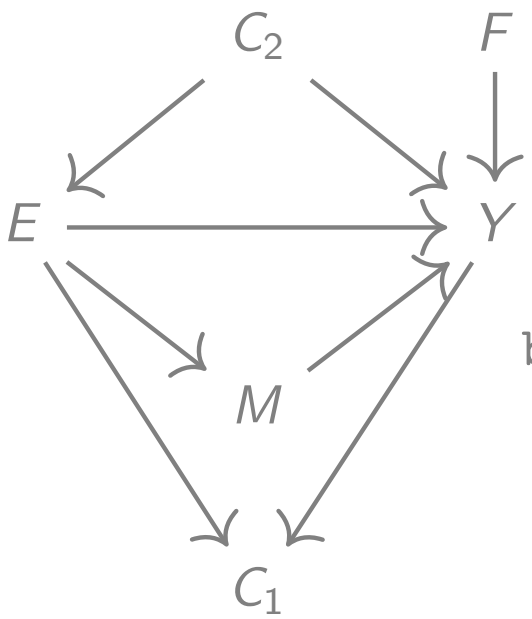
- control for **known** confounding
  ⚠ residual confounding is possible
  ⚠ make sure the variable is not a collider! `Berkson paradox`

Randomization

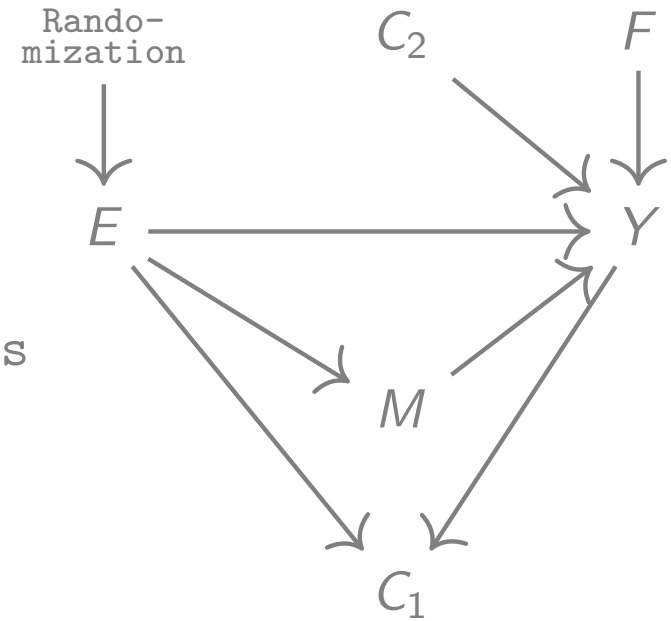
The exposure is randomly allocated among participants

- DAG: "removes" all arrows directed to the exposure variable

Example:



becomes



✔ control for **known** and **unknown** confounders

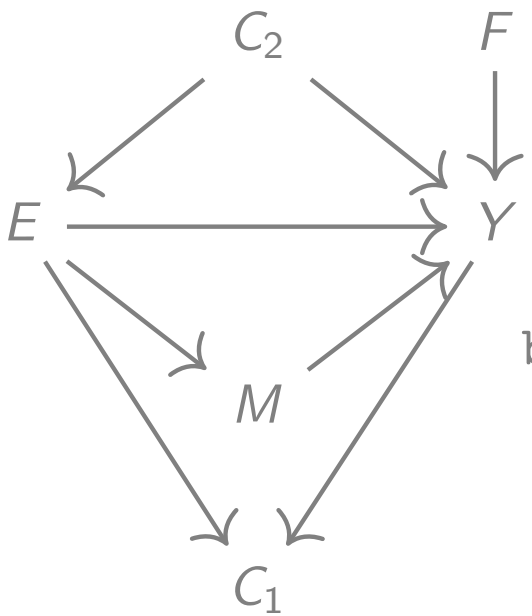
✘ can be complex/expensive/unethical to carry-out

Randomization

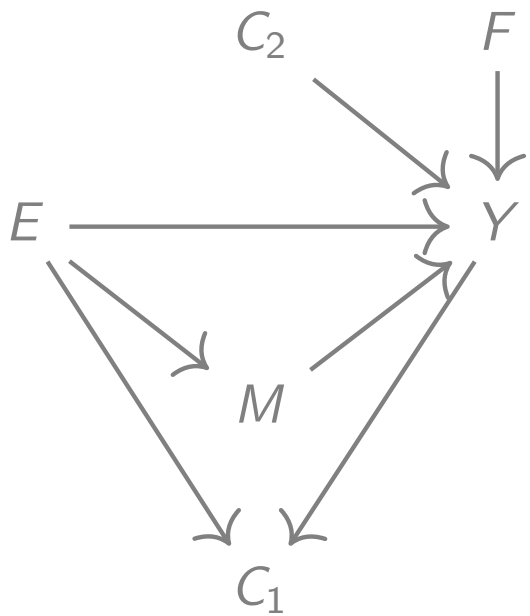
The exposure is randomly allocated among participants

- DAG: "removes" all arrows directed to the exposure variable

Example:



becomes



✔ control for **known** and **unknown** confounders

✘ can be complex/expensive/unethical to carry-out

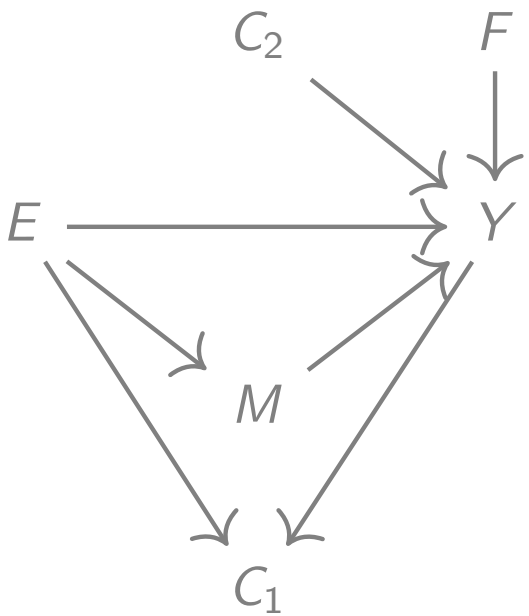
Over adjustment for confounding

After randomization, adjusting:

✔ on risk factors (e.g. C_2, F) may reduce the sampling error (more efficient estimator)

✘ on colliders or mediators (e.g. M, C_1) can lead to bias "over adjustment"

$\rightarrow$ be careful about post randomization variables



"Full" stratification - example

Estimate the prevalence/incidence rate/risk for each exposure and confounder value.

"Full" stratification - example

Estimate the prevalence/incidence rate/risk for each exposure and confounder value.

	Female			Male		
Department	N_F	D_F	$\widehat{\pi}_F$	N_M	D_M	$\widehat{\pi}_M$
All	1835	557	30.35%	2691	1198	44.52%

becomes

	Female			Male		
Department	N_F	D_F	$\widehat{\pi}_F = \frac{D_F}{N_F}$	N_M	D_M	$\widehat{\pi}_M = \frac{D_M}{N_M}$
A	108	89	82.41%	825	512	62.06%
B	25	17	68%	560	353	63.04%
C	593	202	34.06%	325	120	36.92%
D	375	131	34.93%	417	138	33.09%
E	393	94	23.92%	191	53	27.75%
F	341	24	7.04%	373	22	5.9%

"Full" stratification - strata-specific tests

Null hypothesis ($\mathcal{H}_0$)

- same probability for males and females in all strata

Alternative hypothesis ($\mathcal{H}_1$)

- probability for males and females differs in at least one strata

[1] could also be the ratio between probabilities is far away from 1

"Full" stratification - strata-specific tests

Null hypothesis $(\mathcal{H}_0)$

- same probability for males and females in all strata

Alternative hypothesis $(\mathcal{H}_1)$

- probability for males and females differs in at least one strata

Intuitive test:

- reject the null if the difference in probability between men and female is large in any strata [1]

Dept.	$\widehat{\pi}_F$	$\widehat{\pi}_M$	$\widehat{\pi}_F - \widehat{\pi}_M$	p-value
A	82.41%	62.06%	20.35%	$1.4\,10^{-5}$
B	68%	63.04%	4.96%	0.88
C	34.06%	36.92%	-2.86%	0.40
D	34.93%	33.09%	1.84%	0.60
E	23.92%	27.75%	-3.83%	0.32
F	7.04%	5.9%	1.14%	0.59

[1] could also be the ratio between probabilities is far away from 1

"Full" stratification - strata-specific tests

Null hypothesis $(\mathcal{H}_0)$

- same probability for males and females in all strata

Alternative hypothesis $(\mathcal{H}_1)$

- probability for males and females differs in at least one strata

Intuitive test:

- reject the null if the difference in probability between men and female is large in any strata [1]

Dept.	$\widehat{\pi}_F$	$\widehat{\pi}_M$	$\widehat{\pi}_F - \widehat{\pi}_M$	p-value	adjusted p-value
A	82.41%	62.06%	20.35%	$1.4\,10^{-5}$	$8.6\,10^{-5}$
B	68%	63.04%	4.96%	0.88	1.0
C	34.06%	36.92%	-2.86%	0.40	1.0
D	34.93%	33.09%	1.84%	0.60	1.0
E	23.92%	27.75%	-3.83%	0.32	1.0
F	7.04%	5.9%	1.14%	0.59	1.0

[1] could also be the ratio between probabilities is far away from 1

Likelihood

Likelihood of observing the data given the model parameters, e.g.:

	Female			Male		
Department	N_F	D_F	$\widehat{\pi}_F$	N_M	D_M	$\widehat{\pi}_M$
All	1835	557	30.35%	2691	1198	44.52%

$$\mathcal{L}(\pi_F, \pi_M) = (\pi_F)^{D_F} (1 - \pi_F)^{N_F - D_F} (\pi_M)^{D_M} (1 - \pi_M)^{N_M - D_M} \in [0, 1]$$

Likelihood

Likelihood of observing the data given the model parameters, e.g.:

	Female			Male		
Department	N_F	D_F	$\widehat{\pi}_F$	N_M	D_M	$\widehat{\pi}_M$
All	1835	557	30.35%	2691	1198	44.52%

$$\mathcal{L}(\pi_F, \pi_M) = (\pi_F)^{D_F} (1 - \pi_F)^{N_F - D_F} (\pi_M)^{D_M} (1 - \pi_M)^{N_M - D_M} \in [0, 1]$$

- here a likelihood of 1 would indicate that our model perfectly explain the data
- we usually look for the parameter value maximizing the likelihood

"Full" stratification - Likelihood ratio test

Null hypothesis ($\mathcal{H}_0$)

- same probability for males and females in all strata
- $\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_0})$: likelihood under non-stratified model

Alternative hypothesis ($\mathcal{H}_1$)

- probability for males and females differs in at least one strata
- $\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1})$: likelihood under stratified model

Likelihood ratio test (LRT)

- is the "fit" significantly better for the stratified model:

$$2\left(\log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1})\right) - \log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_0})\right)\right) \text{ large ?}$$

Under $\mathcal{H}_0$, it follows a χ^2_6: large means > 12.59

"Manual" LRT (1/2)

- Likelihood under the alternative:

$$\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1}) = \prod_{k=1}^{6} \left(\widehat{\pi}_{F,k}^{D_{F,k}} (1 - \widehat{\pi}_{F,k})^{N_{F,k} - D_{F,k}} \widehat{\pi}_{M,k}^{D_{M,k}} (1 - \widehat{\pi}_{M,k})^{N_{M,k} - D_{M,k}} \right)$$

$$= (82.41\%)^{89} (1 - 82.41\%)^{108-89} (62.06\%)^{512} (1 - 62.06\%)^{825-512}$$

$$\times \ldots$$

$$\times (7.04\%)^{24} (1 - 7.04\%)^{341-24} (5.9\%)^{373} (1 - 5.9\%)^{373-22}$$

"Manual" LRT (1/2)

- Likelihood under the alternative:

$$\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1}) = \prod_{k=1}^{6} \left(\widehat{\pi}_{F,k}^{D_{F,k}} (1 - \widehat{\pi}_{F,k})^{N_{F,k} - D_{F,k}} \widehat{\pi}_{M,k}^{D_{M,k}} (1 - \widehat{\pi}_{M,k})^{N_{M,k} - D_{M,k}} \right)$$

$$= (82.41\%)^{89}(1 - 82.41\%)^{108-89}(62.06\%)^{512}(1 - 62.06\%)^{825-512}$$

$$\times \ldots$$

$$\times (7.04\%)^{24}(1 - 7.04\%)^{341-24}(5.9\%)^{373}(1 - 5.9\%)^{373-22}$$

- Likelihood under the null ($\frac{89+512}{108+825} \approx 0.6441$):

$$\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_0}) = \prod_{k=1}^{6} \left(\widehat{\pi}_{k}^{D_k} (1 - \widehat{\pi}_{k})^{N_k - D_k} \right) = (64.41\%)^{601}(1 - 64.411\%)^{933-601}$$

$$\times \ldots$$

$$\times (6.44\%)^{46}(1 - 6.44\%)^{714-46}$$

"Manual" LRT (2/2)

- log-likelihood under $\mathcal{H}_0$: $\log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_0})\right) = -2594.5$
- log-likelihood under $\mathcal{H}_1$: $\log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1})\right) = -2583.6$
- Difference ("improvement"):

$$\log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1})\right) - \log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_0})\right) = 10.9$$

- Test statistic:

$$2 * \log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_1})\right) - 2 * \log\left(\mathcal{L}(\widehat{\Theta}_{\mathcal{H}_0})\right) = 21.7$$

- Significance threshold: 12.59

Likelihood ratio test - In R

Dataset (first three lines):

```
df[1:3,]
```

```
  Gender Dept   N   D
1   Male    A 825 512
2 Female    A 108  89
3   Male    B 560 353
```

Fit model under $\mathcal{H}_0$ and $\mathcal{H}_1$:

```
e.H1 <- glm(cbind(D,N-D) ~ Gender*Dept,
            data = df, family = binomial(link="logit"))
e.H0 <- glm(cbind(D,N-D) ~ Dept,
            data = df, family = binomial(link="logit"))
```

Likelihood ratio test:

```
anova(e.H0, e.H1, test = "LRT")
```

"Full" stratification - summary

Estimating a separate effect in each strata is:

- a flexible approach - minimal assumptions
- not completely straightforward to interpret and report:

  ⚠ (possibly) different effect in each strata

Statistical inference

- strata-specific tests:

  ✔ intuitive, show in which strata rejection occured

  ✘ not optimal (in term of statistical power)

- likelihood ratio test:

  ✔ implemented in standard software

  ✘ can be hard to interpret - reason for rejection?

"Common effect"

Simplified stratification:

- assume the same effect in all strata

Example with a multiplicative effect.

Statistical model:

	Female			Male		
Department	N_F	D_F	π_F	N_M	D_M	π_M
A	108	89	$\beta \times \pi_{M,A}$	825	512	$\pi_{M,A}$
B	25	17	$\beta \times \pi_{M,B}$	560	353	$\pi_{M,B}$
C	593	202	$\beta \times \pi_{M,C}$	325	120	$\pi_{M,C}$
D	375	131	$\beta \times \pi_{M,D}$	417	138	$\pi_{M,D}$
E	393	94	$\beta \times \pi_{M,E}$	191	53	$\pi_{M,E}$
F	341	24	$\beta \times \pi_{M,F}$	373	22	$\pi_{M,F}$

"Common effect"

Simplified stratification:

- assume the same effect in all strata

Example with a multiplicative effect.

Estimates: $\widehat{\beta} = 1.12$, $\widehat{\pi}_A$, ..., $\widehat{\pi}_F$.

	Female			Male		
Dep	N_F	D_F	$\widehat{\pi}_F \neq \frac{D_F}{N_F}$	N_M	D_M	$\widehat{\pi}_M \neq \frac{D_M}{N_M}$
A	108	89	$1.12 \times 63.9\% = 71.7\%$	825	512	63.9%
B	25	17	$1.12 \times 62.9\% = 70.6\%$	560	353	62.9%
C	593	202	$1.12 \times 32.4\% = 36.4\%$	325	120	32.4%
D	375	131	$1.12 \times 32.1\% = 36.0\%$	417	138	32.1%
E	393	94	$1.12 \times 23.2\% = 26.0\%$	191	53	23.2%
F	341	24	$1.12 \times 6.1\% = 6.8\%$	373	22	6.1%

Limitation of the multiplicative model

For the $k - th$ strata, the probability is:

- π_k in the "reference" group
- $\beta\pi_k$ in the other group

where β can be any positive number.

Limitation of the multiplicative model

For the $k-th$ strata, the probability is:

- π_k in the "reference" group
- $\beta\pi_k$ in the other group

where β can be any positive number.

⚠ If β is large, then $\beta\pi_k$ can be above 1!

- use a multiplicative effect on the "odd scale" instead

$$\Omega_k = \frac{\pi_{M,k}}{1-\pi_{M,k}} \iff \pi_{M,k} = \frac{\Omega_k}{1+\Omega_k}$$

$$\beta\Omega_k = \frac{\pi_{F,k}}{1-\pi_{F,k}} \iff \pi_{F,k} = \frac{\beta\Omega_k}{1+\beta\Omega_k} \in [0,1]$$

Cochran–Mantel–Haenszel test (CMH)

Simplified stratification:

- assume the same effect in all strata

CMH: multiplicative odd effect

	Female			Male		
Department	N_F	D_F	π_F	N_M	D_M	π_M
A	108	89	$\frac{\beta\Omega_A}{1+\beta\Omega_A}$	825	512	$\frac{\Omega_A}{1+\Omega_A}$
B	25	17	$\frac{\beta\Omega_B}{1+\beta\Omega_B}$	560	353	$\frac{\Omega_B}{1+\Omega_B}$
C	593	202	$\frac{\beta\Omega_C}{1+\beta\Omega_C}$	325	120	$\frac{\Omega_C}{1+\Omega_C}$
D	375	131	$\frac{\beta\Omega_D}{1+\beta\Omega_D}$	417	138	$\frac{\Omega_D}{1+\Omega_D}$
E	393	94	$\frac{\beta\Omega_E}{1+\beta\Omega_E}$	191	53	$\frac{\Omega_E}{1+\Omega_E}$
F	341	24	$\frac{\beta\Omega_F}{1+\beta\Omega_F}$	373	22	$\frac{\Omega_F}{1+\Omega_F}$

CMH models a common odd-ratio over the strata: β

Estimation

Consider the sequence of 2 by 2 tables, one for each strata k:

Group \ Outcome	Rejected	Admitted
Male	$a_k = n_{M,k} - D_{M,k}$	$b_k = D_{M,k}$
Female	$c_k = n_{F,k} - D_{F,k}$	$d_k = D_{F,k}$

The common odd-ratio is estimated by:

$$\widehat{OR}^{MH} = \frac{\sum_{k=1}^{K} \frac{a_k d_k}{n_k}}{\sum_{k=1}^{K} \frac{b_k c_k}{n_k}}$$

with $n_k = a_k + b_k + c_k + d_k$ the number of applicants per department.

Estimation

Consider the sequence of 2 by 2 tables, one for each strata k:

Group \ Outcome	Rejected	Admitted
Male	$a_k = n_{M,k} - D_{M,k}$	$b_k = D_{M,k}$
Female	$c_k = n_{F,k} - D_{F,k}$	$d_k = D_{F,k}$

The common odd-ratio is estimated by:

$$\widehat{OR}^{MH} = \frac{\sum_{k=1}^{K} \frac{a_k d_k}{n_k}}{\sum_{k=1}^{K} \frac{b_k c_k}{n_k}} = \frac{\sum_{k=1}^{K} \frac{b_k c_k}{n_k} \frac{a_k d_k}{b_k c_k}}{\sum_{k=1}^{K} \frac{b_k c_k}{n_k}} = \frac{\sum_{k=1}^{K} w_k OR_k}{\sum_{k=1}^{K} w_k}$$

with $n_k = a_k + b_k + c_k + d_k$ the number of applicants per department.

Cochran–Mantel–Haenszel test - In R

Dataset (vector of 2 by 2 tables)

```
str(UCBAdmissions)
```

```
'table' num [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 .
- attr(*, "dimnames")=List of 3
 ..$ Admit : chr [1:2] "Admitted" "Rejected"
 ..$ Gender: chr [1:2] "Male" "Female"
 ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

Test:

```
mantelhaen.test(UCBAdmissions[,2:1,])
## [,2:1,] to use males as reference
```

```
Mantel-Haenszel X-squared = 1.4269, df = 1, p-value = 0.2323
95 percent confidence interval:
 0.9431028 1.2954922
common odds ratio
        1.105343
```

"Common effect" - summary

Reporting a single effect is convenient . . .

⚠ . . . but the "common effect" is a strong assumption. ⚠

Can be checked:

- looking at the effects in the "full" stratification
- using a statistical test (e.g. Breslow-Day Test or Woolf test)

Odds ratios:

- not very intuitive
- but have nice numerical properties when working with probabilities

Summing up

What we have seen today

Definition of "a causal effect"

- consistency, positivity, exchangeability assumptions

Graphical representation of a study:

- reading and constructing **DAGs**
- definition of **confounder**, **collider**, mediator, risk factor
- using **DAGs** to decide what to adjust on
  on the validity of a study

Controling for confounding:

- by design: **randomization**, **restriction**
- using a statistical method: **stratification**
  - "full stratification" (flexible, strata-specific effects)
  - "common effect" (assumption to be checked)

Reference I

Bickel, P. J., Hammel, E. A., and O Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404.

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271.

Hernández-Díaz, S., Schisterman, E. F., and Hernán, M. A. (2006). The birth weight "paradox" uncovered? *American journal of epidemiology*, 164(11):1115–1120.

Parascandola, M. and Weed, D. L. (2001). Causation in epidemiology. *Journal of Epidemiology & Community Health*, 55(12):905–912.

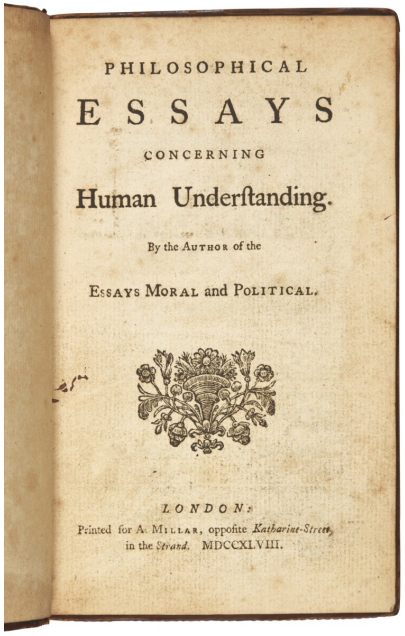
Reference II

Pearce, N. and Vandenbroucke, J. P. (2020). Educational note: types of causes. *International Journal of Epidemiology*, 49(2):676–685.

Rothman, K. J. and Greenland, S. (2005). Causation and causal inference in epidemiology. *American journal of public health*, 95(S1):S144–S150.

Definition of causality by Hume

"We may define a cause to be an object,
followed by another,
and where all the objects similar to the first
are followed by objects similar to the second.
Or in other words where,
if the first object had not been,
the second never had existed" (Hume 1748).


PHILOSOPHICAL
ESSAYS
CONCERNING
Human Understanding.
By the Author of the
Essays Moral and Political.
LONDON:
Printed for A. Millar, opposite Katharine-Street,
in the Strand. MDCCXLVIII.

One categorisation of causes (Pearce and Vandenbroucke, 2020)

Table 1. Characteristics of different types of causes

	'Fixed' states	Dynamic states	Events
Examples	Sex 'Ancestry' Genetics	Gender Ethnicity Racism[a] DNA methylation Obesity High cholesterol High blood pressure	Smoking a pack a day Racism[a] Gene therapy Exercise Diet Antihypertensives
Can we explore the mechanisms?	Yes (e.g. hormonal influences on breast cancer risk)	Yes (e.g. obesity causes chronic inflammation which increases CVD risk)	Yes (e.g. effects of exercise on development of collateral vasculature and hence on CVD)
Can we make a counterfactual contrast?	Yes (e.g. genetic comparisons)	Yes (e.g. BMI = 35 vs BMI = 25)	Yes (e.g. high exercise vs low exercise)
Can we randomize?	No (e.g. sex cannot be randomized[b])	No (e.g. obesity cannot be randomized)[c]	Yes (e.g. exercise can be randomized)
Can we intervene?	No (although we can intervene on possible mediators or take actions on intermediate states)[d]	Yes (we can carry out interventions which reduce or increase obesity)	Yes (e.g. interventions to encourage exercise)

One categorisation of causes (Pearce and Vandenbroucke, 2020)

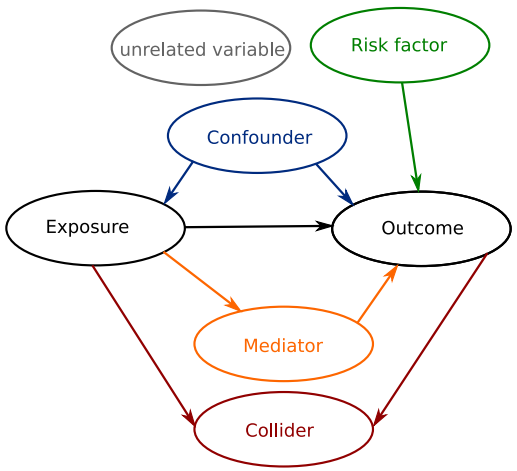
Table 1. Characteristics of different types of causes

	'Fixed' states	Dynamic states	Events
Examples	Sex 'Ancestry' Genetics	Gender Ethnicity Racism[a] DNA methylation Obesity High cholesterol High blood pressure	Smoking a pack a day Racism[a] Gene therapy Exercise Diet Antihypertensives
Can we explore the mechanisms?	Yes (e.g. hormonal influences on breast cancer risk)	Yes (e.g. obesity causes chronic inflammation which increases CVD risk)	Yes (e.g. effects of exercise on development of collateral vasculature and hence on CVD)
Can we make a counterfactual contrast?	Yes (e.g. genetic comparisons)	Yes (e.g. BMI = 35 vs BMI = 25)	Yes (e.g. high exercise vs low exercise)
Can we randomize?	No (e.g. sex cannot be randomized[b])	No (e.g. obesity cannot be randomized)[c]	Yes (e.g. exercise can be randomized)
Can we intervene?	No (although we can intervene on possible mediators or take actions on intermediate states)[d]	Yes (we can carry out interventions which reduce or increase obesity)	Yes (e.g. interventions to encourage exercise)

⚠ Using the counterfactual framework to label as 'causal' effects of fixed states is controversial.

Nomenclature of the variables

- **Risk factor**: ancestor of only the outcome
- **Confounder**: ancestor of both the exposure and the outcome
- **Collider**: descendant of both the exposure[2] and the outcome.
- **Mediator**: variable on a directed path relating the exposure to the outcome.
- **Unrelated variable**: none of the previous



[2] there must be a least one directed path relating the collider to the exposure that does not contain the outcome