folkehelseinstituttet

# DAGs intro

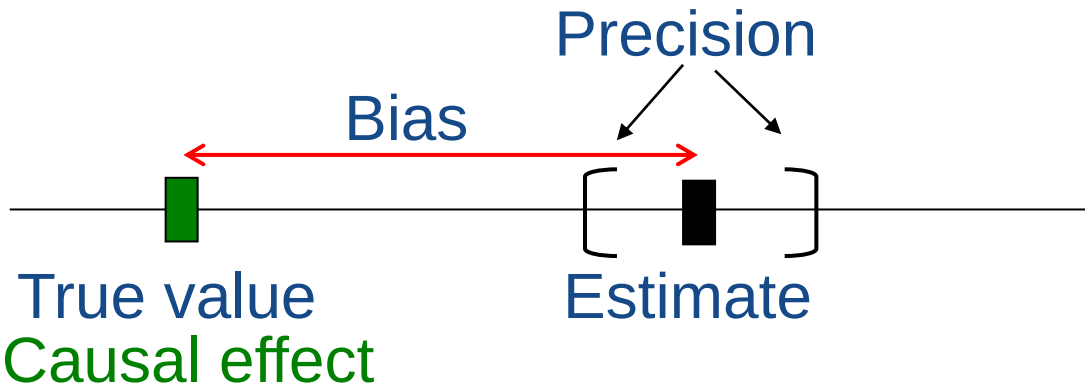
2h

DirectedAcyclicGraph

*Hein Stigum*

# Precision and bias

- Estimate effect of exposure on outcome

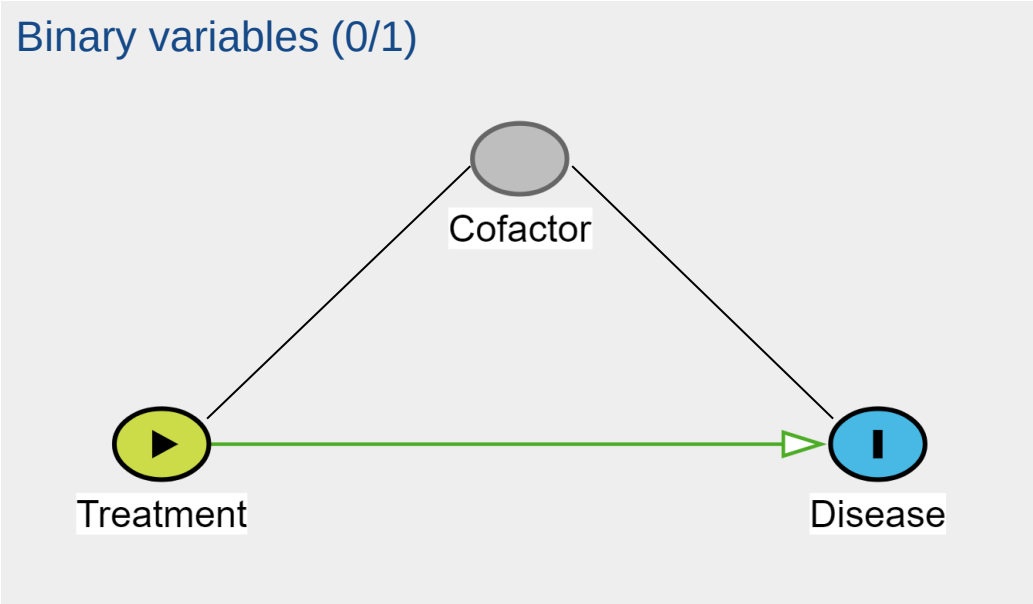



- Precision: random error
  - sample size and variance
- Bias: systematic error
  - confounding, selection bias, measurement error

Often tradeoff between lack of bias and precision

DAGs only show bias (yes/no)

# Motivating example




| | | |
|---|---|---|
| logistic D trt | OR=1.3 | 30% more disease if treated |
| logistic D trt C | OR=0.7 | 30% less disease if treated |

What is the correct analysis?

Need causal information to answer that

Causal Graph (DAG)

*Replace lines with arrows*

# Agenda

- DAGs introduction
  - Confounder, Collider, Mediator
- Causal thinking
- Estimation vs Prediction models
- Drawing and Analyzing DAGs
  - DAGitty

Causal versus casual

# CONCEPTS

(Rothman et al. 2008; Veieroed et al. 2012

# god-DAG

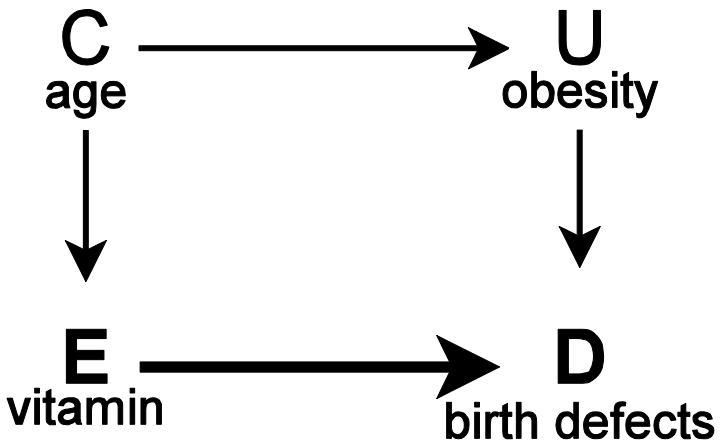
## Causal Graph:

Node = variable

Arrow = cause

E=exposure, D=disease

DAG=Directed Acyclic Graph




## Read of the DAG:

Causality = arrows

Associations = paths

Independencies = no paths

Arrow missing in the DAG!

## Estimations:

E-D association has two parts:

| | | |
|---|---|---|
| E→D | causal effect | *keep open* |
| E←C→U→D | bias | *try to close* |

E←[C]→U→D Condition (adjust) to close

→→→→Time →→→→

# Association and Cause

## Association

E
Yellow fingers

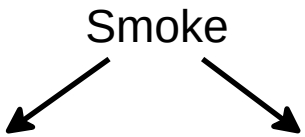
D
Lung cancer

## 3 possible causal structure

1 Yellow fingers → Lung cancer — Cause

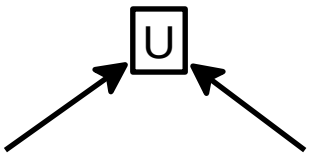
(reverse cause)

2




Confounder

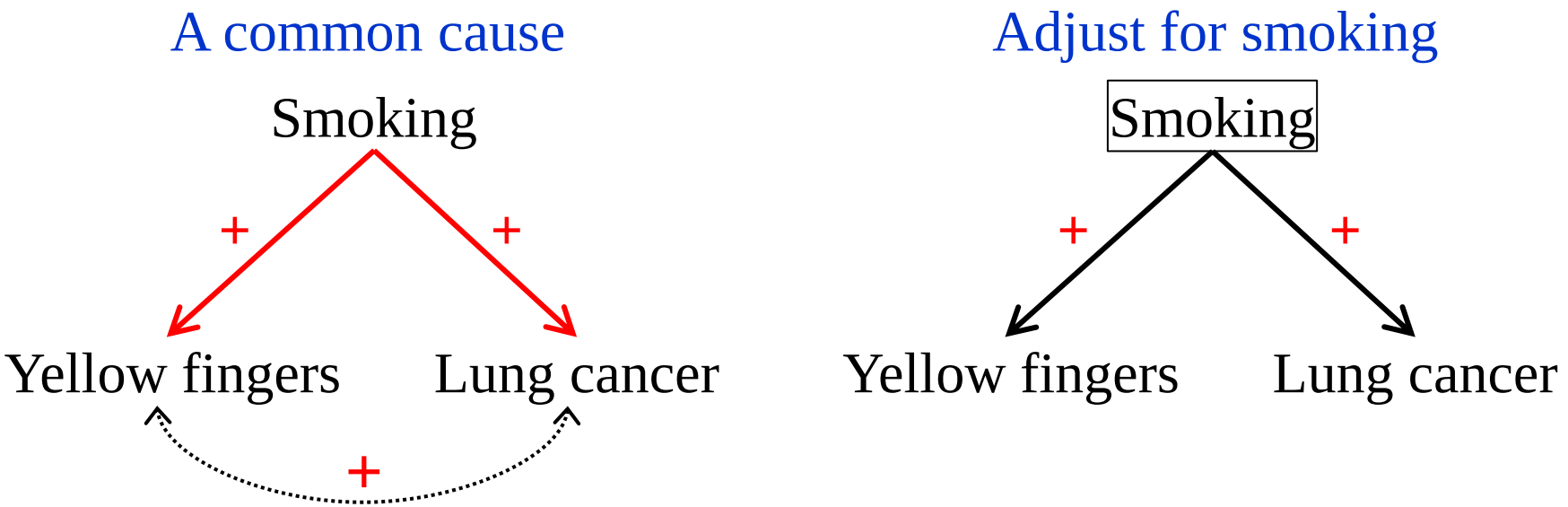
3




Collider

+ more complicated structures

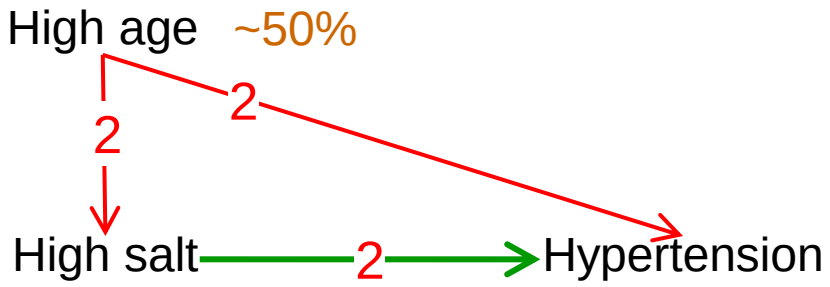
# Confounder idea




- Confounding:
- A common cause of exposure and disease
- Conditioning on a confounder removes the bias
- Condition = (restrict, stratify, regression adjust)
- Paths

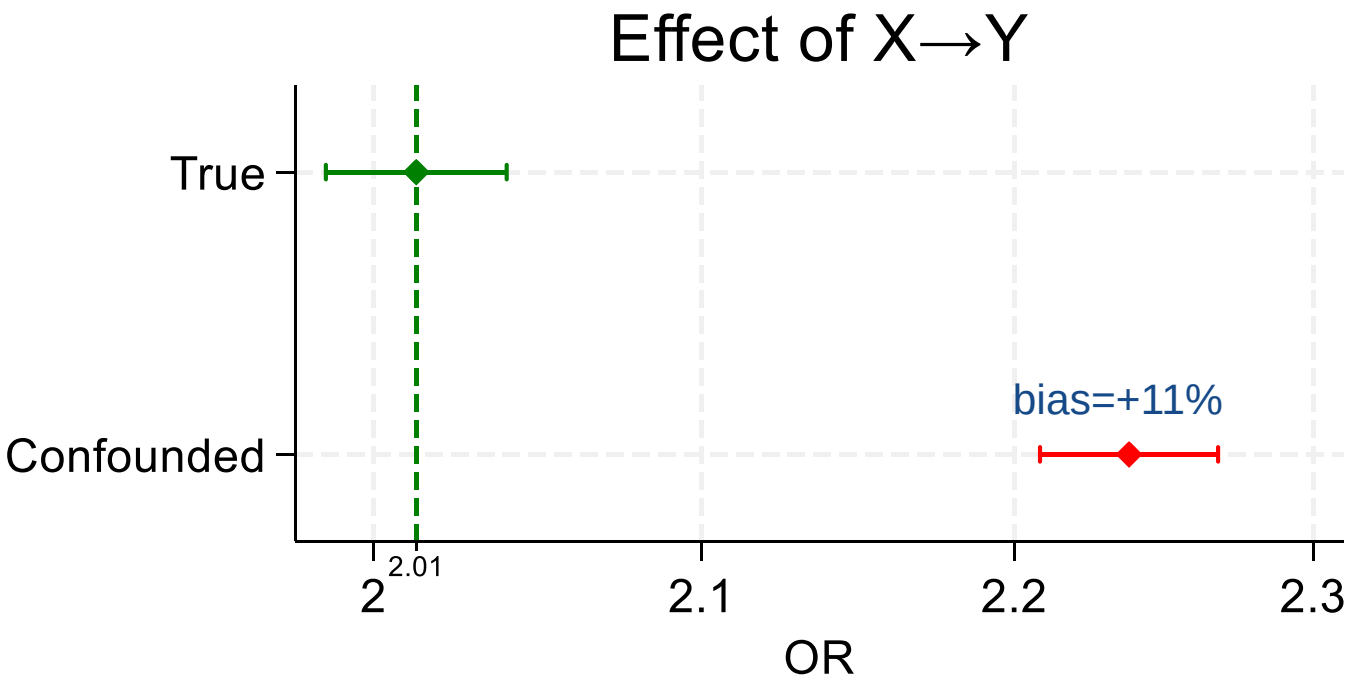
- Simplest form

# Confounder Example




Binary
OR

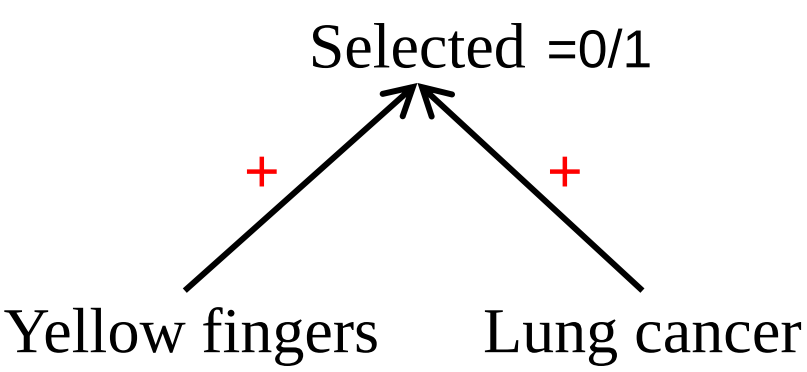
*Go to Syntax*



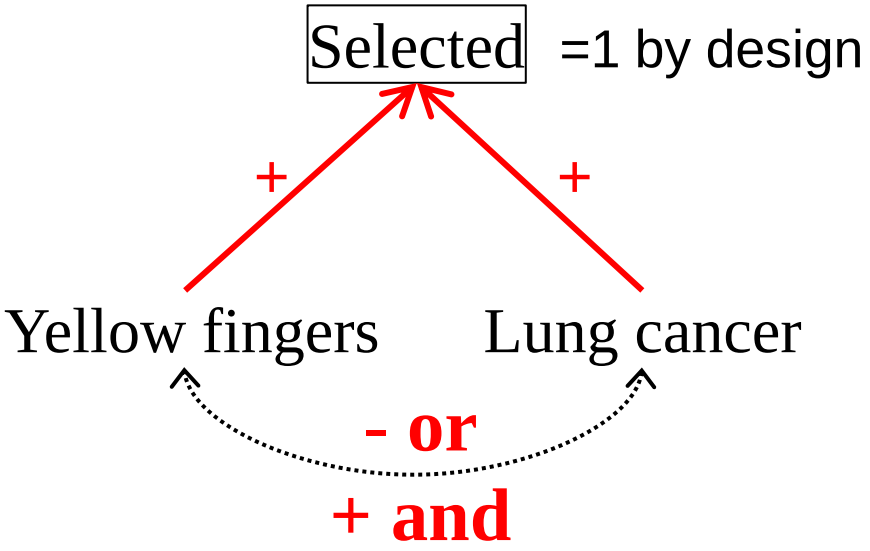

N= 1,000,000, X=logistic(.3) Y=logistic(.1) C=rbinomial(1,.5)
xy=2 cx=2 cy=2

# Collider idea

## Two causes for selection to study




## Selected subjects



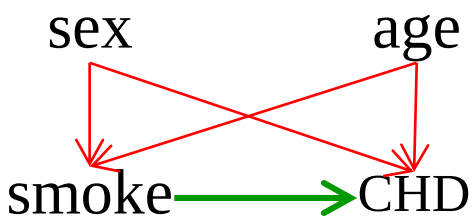

- Collider:
- A common effect of exposure and disease
- Conditioning on a collider induces bias
- "And" and "or" selection leads to different bias
- Paths

- Simplest form

(Hernan, Hernandez-Diaz et al. 2004)

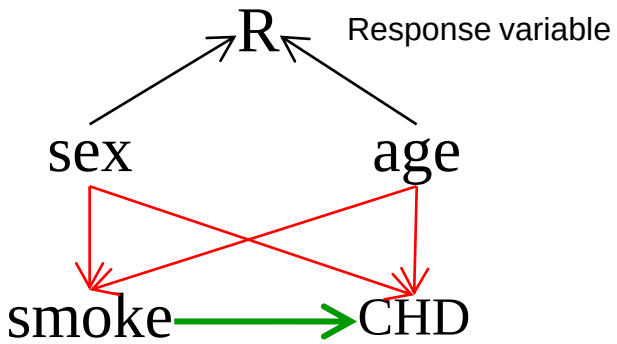
# Selection bias in a DAG

Draw DAG




Coronary Heart Disease

Add variable R= 1 if respons, 0 if not




Add causes of response

Females more willing to participate

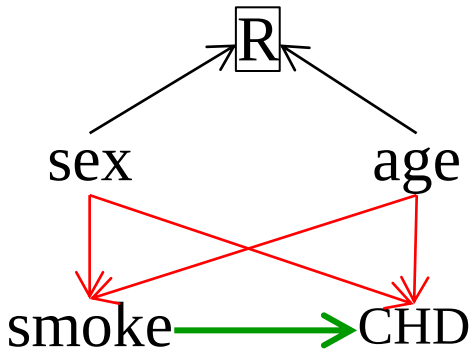
Old people less willing to participate

Smoke and CHD does not affect participation directly

Condition on R

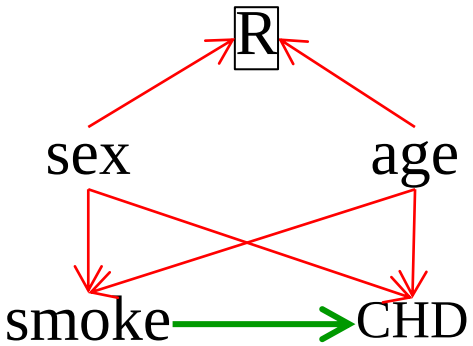
Only R=1 available

A new non-causal path opens




Hernan 2004

# Adjusting for Selection bias




| Paths | Type | Status |
|---|---|---|
| smoke→CHD | Causal | Open |
| smoke←sex→[R]←age→CHD | Non-causal | Open |
| + two confounder paths | | |

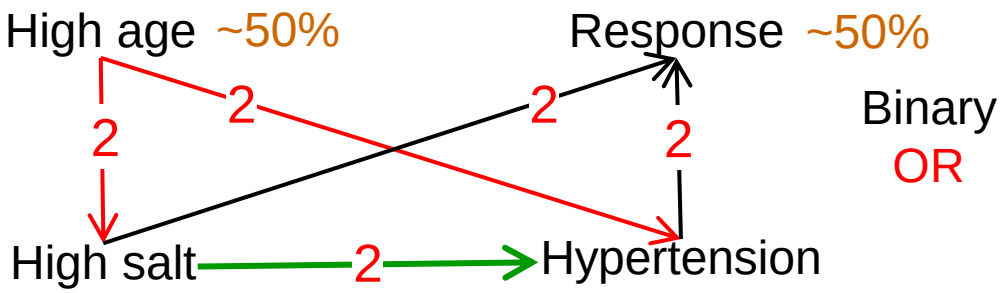
Adjusting for sex or age or both removes the collider stratification bias (selection bias)

A full understanding of selection bias requires an extended DAG theory ("separation theory")

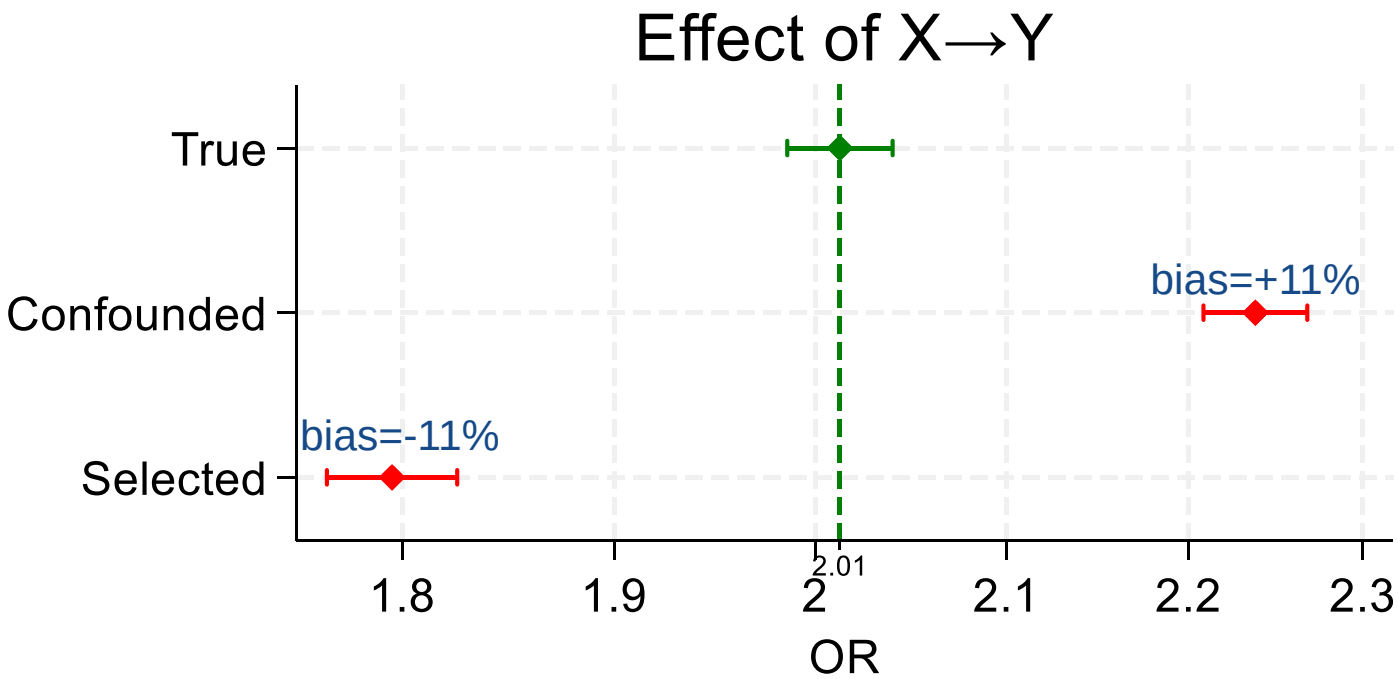
*MF 9570 Causal Inference*

Greenland and Pearl 2011

# Collider bias (response/selection) ex.




*Go to Syntax*

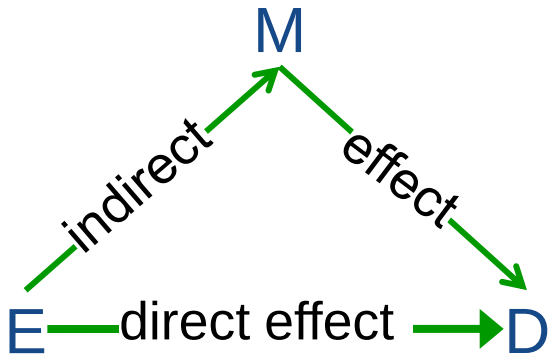



N= 1,000,000, X=logistic(.3) Y=logistic(.1) C=rbinomial(1,.5) R=log
xy=2 cx=2 cy=2

Greenland, S. 2003

# Mediator idea

- Have found a cause (E)
- How does it work?
  - Mediator (M)
  - Paths




$$Total\ effect = indirect + direct$$

$$Mediated\ proportion = \frac{indirect}{total}$$

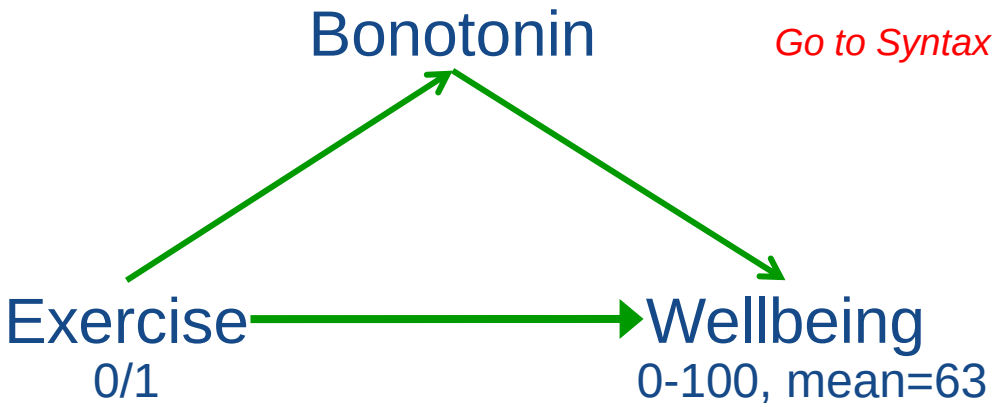
Controlled direct and indirect effects — old
Natural direct and indirect effects — new

Strong conditions of non-confounding

Extra Material>Direct and Indirect effects

(VanderWeele 2016)

# Mediator Example




*Go to Syntax*

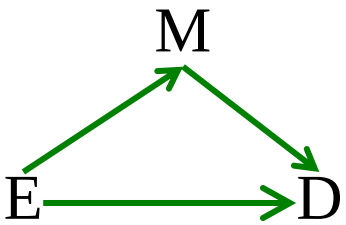
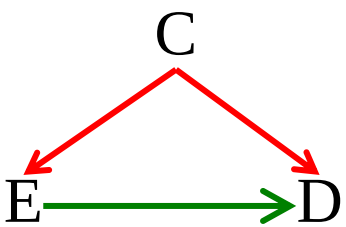
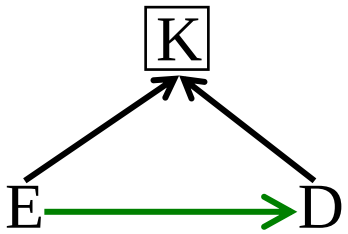
| Total | 12.7 |
|---|---|
| Indirect | 9.8 |
| Direct | 2.9 |

- Total effect
  - regress wellbeing exercise
  - Total=12.7
- Direct and Indirect effects
  - mediate (wellbeing) (bonotonin) (exercise)
  - Indirect=9.8, Direct=2.9
  - Mediated proportion=77%

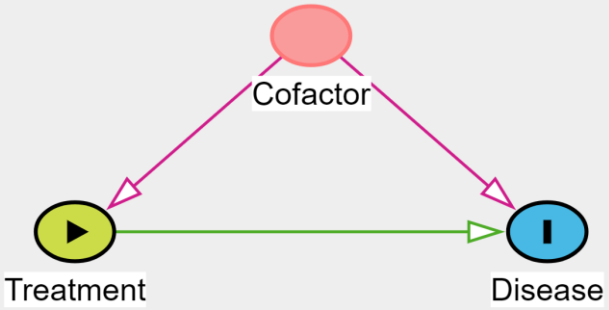
# Concepts: Summing up

Associations visible in data. Causal structure from outside the data.

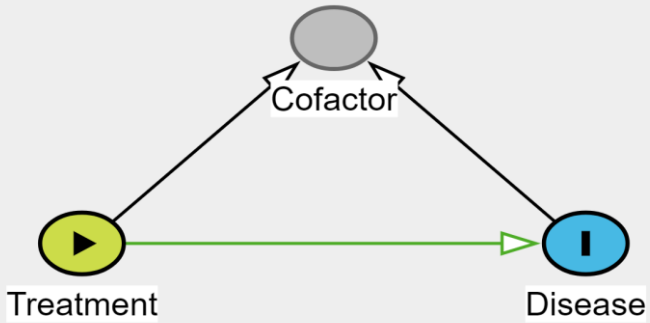
DAG: no arrow means (conditional) independence

| | Type | Adjustment |
|---|---|---|
|  | Cause | |
|  | Cause with Mediator | |
|  | Cause with Confounder | adjust |
|  | Cause with Collider | *adjust for separating variables* |

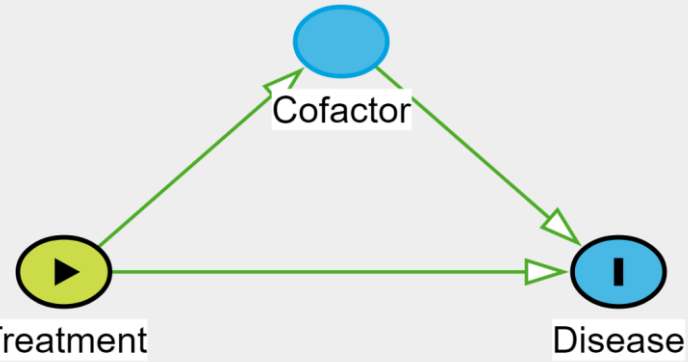
# Motivating Example




Confounder: adjust
logistic D trt C OR=0.7




Collider: not adjust
logistic D trt OR=1.3




Mediator:
regress D trt Total effect*
regress D trt C Direct effect*
Indirect effect= Total-Direct

linear regression model
regress D trt, robust

* strong assumptions of
no unmeasured confounding

# Causal thinking in analyses

# Pre DAG

- Aim (in introduction)
  - "We want to estimate the association between E and D"
- Adjust or not
  - Use statistical criteria
- Present results (Table 2)
  - Table of all estimates associations from one model

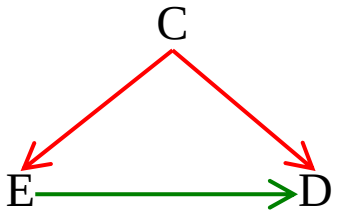
# Aims in papers

- Standard aim (in introduction)
  - "We want to estimate the association between E and D"
- Problems
  - Imprecise — many E-D association
  - Why adjust — gives no rationale for adjusting
- Solution
  - "Our target parameter is the effect of E on D"

(Lederer et al. 2018)

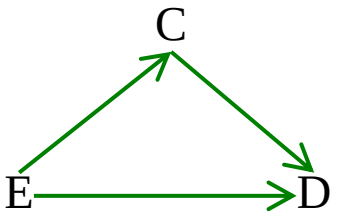
# Adjust or not for C

Cause:




Confounder

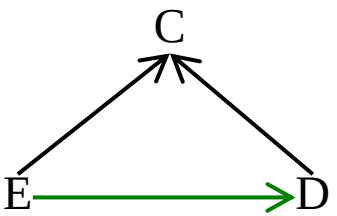
Adjust




Mediator

Direct: adjust

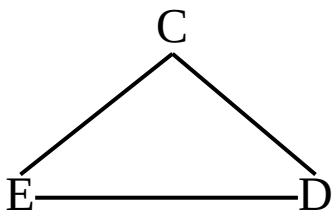
Total: not adjust




Collider

Not adjust

Association:




Statistical criteria:

likelihood ratio, AIC, 10% change in estimate

cannot differentiate between

Confounder, Mediator or Collider

Need causal model to do a proper analysis

DAGs variable selection: close all non-causal paths

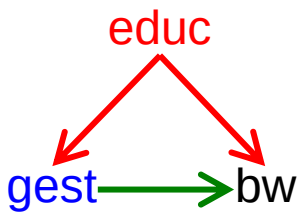
(Robins 2001)

# Table 2 fallacy, gestation age and birth weight

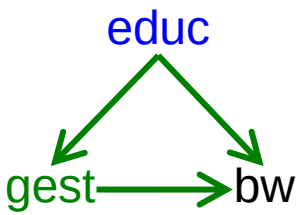
- Pre DAGs: report all covariate effects from one model
- Post DAGs:
  - report only exposure effect
  - separate models for other covariates

Exposure:

Model 1
gest




Model 2
educ




| Variable | model 1 | model 2 |
|---|---|---|
| gest | 202 | |
| educ | ~~113~~ | 278 |

educ confounder
adjust

gest mediator
not adjust

(Westreich and Greenland 2013)

folkehelseinstituttet

# Causal thinking: Summing up

- Make a clear causal aim.
- Data driven analyses do not work. Need **causal** information from outside the data. (Data driven prediction models OK though).
- Reporting table of adjusted associations from one model can be misleading. Report one exposure to one outcome.

Need causal model to do a proper analysis

# Estimation- versus Prediction Models

# Purpose of regression

- Estimation — **DAGs**, bias, precision
  - Estimate effect of exposure on outcome adjusted for other covariates
  - *Estimate the effect of smoking on lung cancer*

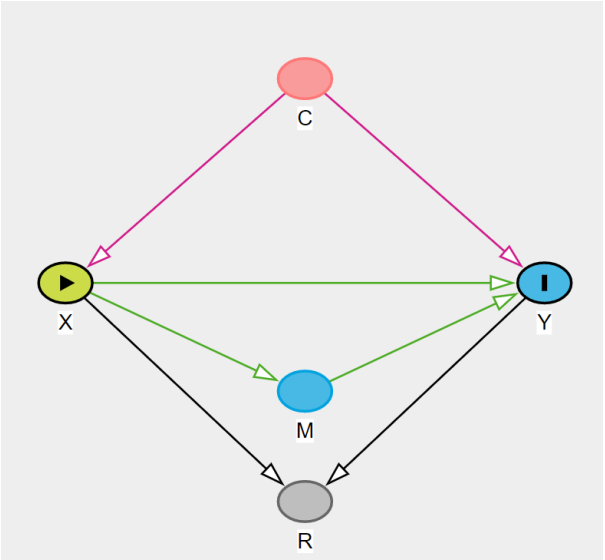
- Prediction — Predictive power, model fit, $R^2$
  - *Predict outcome by exposures*
  1. Estimate model (CHD and age, sex, cholesterol, blood pressure, …)
  2. Predict CHD risk using age, sex, cholesterol, blood pressure, …

# Estimation vs Prediction

## Table of ORs

| Variable | m1 | m2 | m3 | m4 |
|---|---|---|---|---|
| X | 2.22 | 2.00 | 1.80 | 1.68 |
| C | | 1.99 | 2.00 | 2.00 |
| M | | | 2.01 | 2.01 |
| R | | | | 1.94 |

## DAG




## Table of model fit (low values best)

| Model | df | AIC |
|---|---|---|
| m1 | 2 | 393176.4 |
| m2 | 3 | 387965.6 |
| m3 | 4 | 382628 |
| m4 | 5 | 379202.7 |

- Analyze as Estimation Model
  - Use the DAG: what is correct model for the **total effect of X on Y**?
- Analyze as Prediction Model
  - Use the AIC: What is the **best fitting model**?

# Summing up so far

- Estimation and prediction modeling differ
  - Estimation: DAG
  - Prediction: Best fitting model

  } Variable selection

- Remarks
  - We use *model fit* in estimation models:
    - Compare linear and non-linear dose response
    - Include interaction terms
  - Often *predict* results from estimation models:
    - To show dose-response
    - Marginal (standardized) results

# Drawing DAGs

## with DAGitty

(Textor, Hardt et al. 2011)

# DAGitty commands

Search: DAGitty  Kinder egg: Draw, Analyze, Test

- Draw new model
  - Model>New model
- New variables, arrow
  - click  new variable (fill in name)
  - click 1, click 2  arrow
- Set status: Hold pointer over variable and hit on the keyboard:
  - e  exposure
  - o  outcome
  - u  unobserved
  - a  adjusted
  - r  rename
  - d  delete

folkehelseinstituttet

# Country of origin and HPV vaccination

- School based vaccination program for girls
- Started in 2009. 2018 boys included.
- Vaccination uptake: 80%

| Variable | Contrast | OR |
|---|---|---|
| Country | *Asia/Norway* | 1.8 |
| Mothers age | *>35/<25* | 0.7 |
| Income | *high/low* | 1.4 |
| Year | *2014/2009* | 2.7 |

Adjusted ORs

80%*1.25=100%

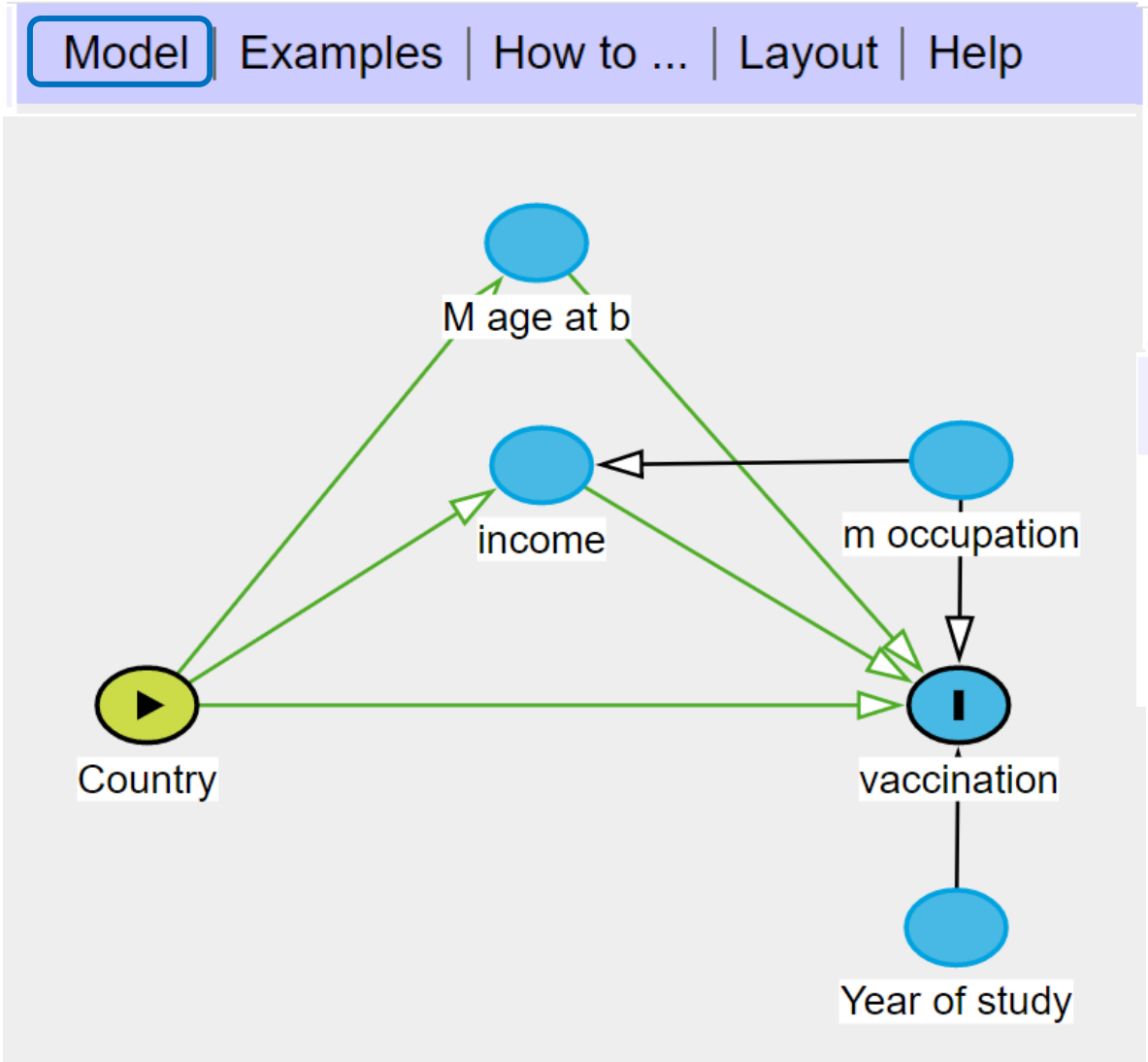
Simplified from a real study

More variables adjusted for

Discussion: Occupation unmeasured confounder

No DAG!

# DAGitty interface




Independent: ⊥

Given: |

A independent of B given C:

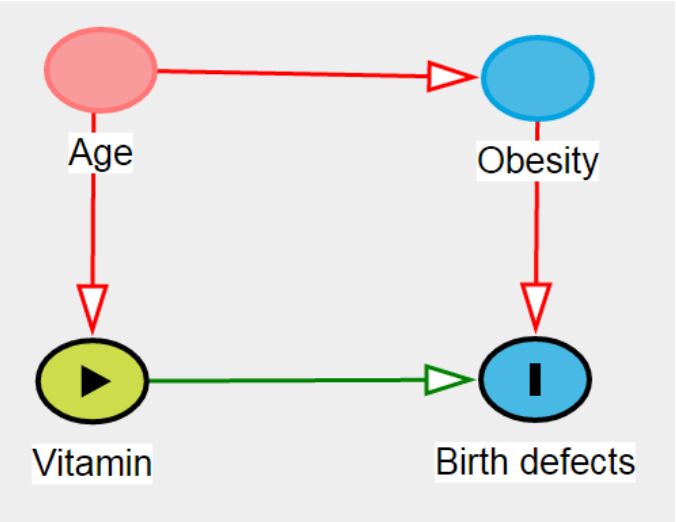
A ⊥ B | C

# Example: Vitamin and Birth Defects

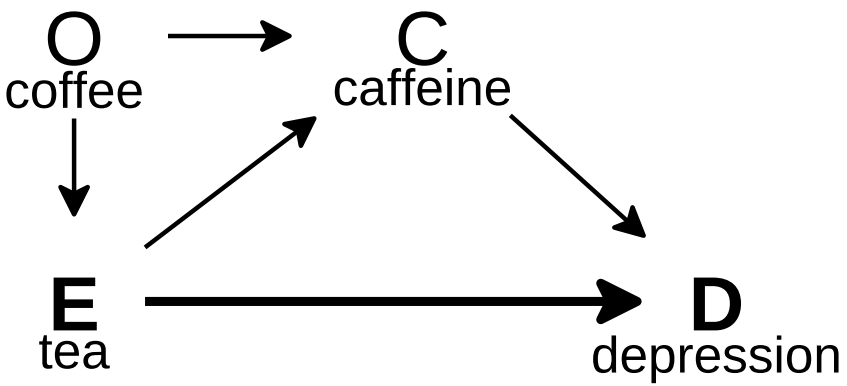
## Draw the Vitamin-Birth defects DAG (as shown)

- Use Obesity as an observed variable (the default).
  - Interpret the “Causal effect identification”
  - Interpret the “Testable implications”
- Add an arrow from Age to Birth defects
  - Interpret the “Causal effect identification”
  - Interpret the “Testable implications”

Question:
Is obestity a confounder?



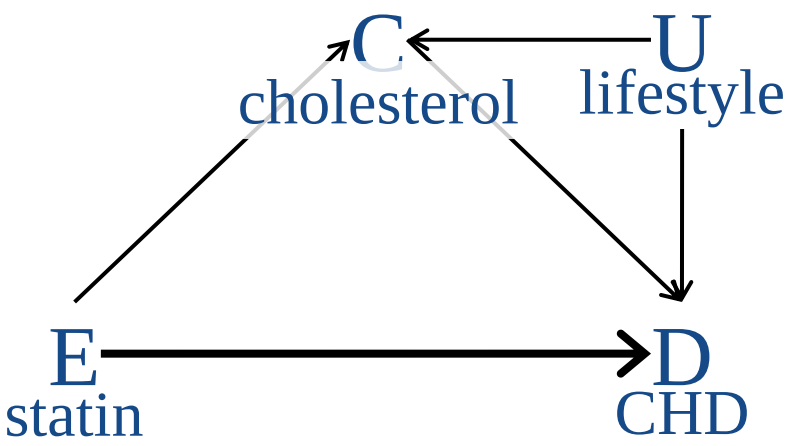
# Example: Tea and depression




1. Draw the DAG in DAGitty.
2. You want the **total effect** of tea on depression. What would you adjust for?
3. You want the **direct effect** of tea on depression. What would you adjust for?
4. Is caffeine an intermediate variable or a variable on a confounder path?

Hintikka et al. 2005

# Example: Statin and CHD




1. Draw the DAG in DAGitty.
2. You want the **total effect** of statin on CHD. What would you adjust for?
3. If lifestyle is unmeasured, can we estimate the **direct effect** of statin on CHD (not mediated through cholesterol)?
4. Is cholesterol an intermediate variable or a collider?

folkehelseinstituttet

# Summing up

- Data driven analyses do not work. Need (causal) information from outside the data.
- DAGs are intuitive and accurate tools to display that information.
- Paths show the flow of causality and of bias and guide the analysis.
- DAGs clarify concepts like confounding and selection bias, and show that we can adjust for both.

**Better discussion based on DAGs**

**Draw your assumptions before your conclusions**

# Recommended DAG reading

- Resources cited in DAGitty (Books, Papers, YouTube)
- Books
  - Hernan, M. A. and J. Robins. *Causal Inference, What If*. 2011
  - Rothman, K. J., S. Greenland, and T. L. Lash. *Modern Epidemiology,* 2008.
  - Morgan and Winship, *Counterfactuals and Causal Inference*, 2009
  - Pearl J, *Causality – Models, Reasoning and Inference*, 2009
  - Veierød, M.B., Lydersen, S. Laake,P. Medical Statistics. 2012
- Papers
  - Greenland, S., J. Pearl, and J. M. Robins. *Causal diagrams for epidemiologic research,* Epidemiology 1999
  - Hernandez-Diaz, S., E. F. Schisterman, and M. A. Hernan. *The birth weight "paradox" uncovered?* Am J Epidemiol 2006
  - Hernan, M. A., S. Hernandez-Diaz, and J. M. Robins. *A structural approach to selection bias,* Epidemiology 2004
  - Berk, R.A. *An introduction to selection bias in sociological data*, Am Soc R 1983
  - Greenland, S. and B. Brumback. *An overview of relations among causal modeling methods,* Int J Epidemiol 2002
  - Weinberg, C. R. *Can DAGs clarify effect modification?* Epidemiology 2007

# References


- Burgess S. 2014. Sample size and power calculations in mendelian randomization with a single instrumental variable and a binary outcome. International Journal of Epidemiology 43:922-929.
- Chen L, Davey SG, Harbord RM, Lewis SJ. 2008. Alcohol intake and blood pressure: A systematic review implementing a mendelian randomization approach. PLoS Med 5:e52.
- Didelez V, Sheehan N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res 16:309-330.
- Greenland S, Robins JM, Pearl J. 1999. Confounding and collapsibility in causal inference. Statistical Science 14:29-46.
- Greenland S, Pearl J. 2011. Adjustments and their consequences -collapsibility analysis using graphical models. Int Stat Rev 79:401-426.
- Hafeman DM, Schwartz S. 2009. Opening the black box: A motivation for the assessment of mediation. International Journal of Epidemiology 38:838-845.
- Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. 2002. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. AJE 155:176-184.
- Hernan MA, Hernandez-Diaz S, Robins JM. 2004. A structural approach to selection bias. Epidemiology 15:615-625.
- Hintikka J, Tolmunen T, Honkalampi K, Haatainen K, Koivumaa-Honkanen H, Tanskanen A, et al. 2005. Daily tea drinking is associated with a low level of depressive symptoms in the finnish general population. European Journal of Epidemiology 20:359-363.
- Lange T, Hansen JV. 2011. Direct and indirect effects in a survival context. Epidemiology 22:575-581.
- Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey SG. 2008. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. Stat Med 27:1133-1163.
- Pearl J. 2012. The causal mediation formula-a guide to the assessment of pathways and mechanisms. Prev Sci 13:426-436.
- Robins JM, Greenland S. 1992. Identifiability and exchangeability for direct and indirect effects. Epidemiology 3:143-155.
- Robinson LD, Jewell NP. 1991. Some surprising results about covariate adjustment in logistic-regression models. Int Stat Rev 59:227-240.
- Rothman KJ, Greenland S, Lash TL. 2008. Modern epidemiology. Philadelphia:Lippincott Williams & Wilkins.
- Sheehan NA, Didelez V, Burton PR, Tobin MD. 2008. Mendelian randomisation and causal inference in observational epidemiology. PLoS Med 5:e177.
- Sjolander A, Dahlqwist E, Zetterqvist J. 2016. A note on the noncollapsibility of rate differences and rate ratios. Epidemiology 27:356-359.
- Textor J, Hardt J, Knuppel S. 2011. Dagitty a graphical tool for analyzing causal diagrams. Epidemiology 22:745-745.
- Tritchler D. 1999. Reasoning about data with directed graphs. Stat Med 18:2067-2076.
- VanderWeele TJ. 2009. Mediation and mechanism. Eur J Epidemiol 24:217-224.
- VanderWeele TJ. 2014. A unification of mediation and interaction: A 4-way decomposition. Epidemiology 25:749-761.
- VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. 2014. Methodological challenges in mendelian randomization. Epidemiology 25:427-435.
- VanderWeele TJ. 2016. Mediation analysis: A practitioner's guide. Annual Review of Public Health, Vol 37 37:17-32.
- Veieroed M, Lydersen S, Laake P. 2012. Medical statistics in clinical and epidemiological research. Oslo:Gyldendal Akademisk.
- Westreich D, Greenland S. 2013. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. AJE 177:292-298.
- Xing C, Xing GA. 2010. Adjusting for covariates in logistic regression models. Genet Epidemiol 34:937-937.

# EXTRA MATERIAL

# Direct and indirect effects

So far: Controlled (in)direct effect

limitations: no E-M interaction and only linear models

New concept: Natural (in)direct effect

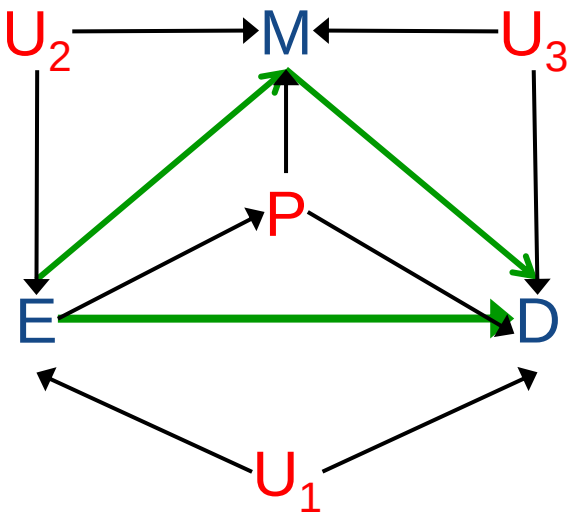
limitations: no exposure dependent confounders

Assumptions:

No unmeasured confounders ($U_1$, $U_2$, $U_3$,)

No exposure dependent confounders (P)

measured or unmeasured




**Mediation analysis requires strong assumptions**

Mediation analysis:

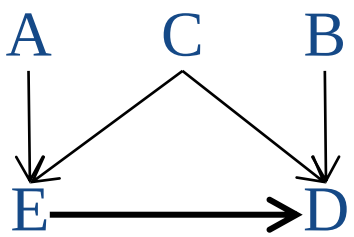
*MF 9580 Epidemiological methods,* *March*

*MF 9570 Causal Inference,* *November*

Hafeman and Schwartz 2009;
Lange and Hansen 2011;
Pearl 2012;
Robins and Greenland 1992;
**VanderWeele** 2009, **2016**

# Effects of adjustment

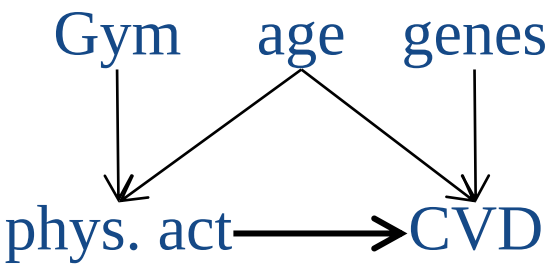
# Effects of adjustment




What variables should we adjust for?

What are the effects of adjustment?

| Variable | Adjust | Bias | Precision |
|---|---|---|---|
| A | no | bias if misspecified | reduce precision (collinearity) |
| B | maybe | no | model dependent |
| C | yes | remove confounding | model dependent |



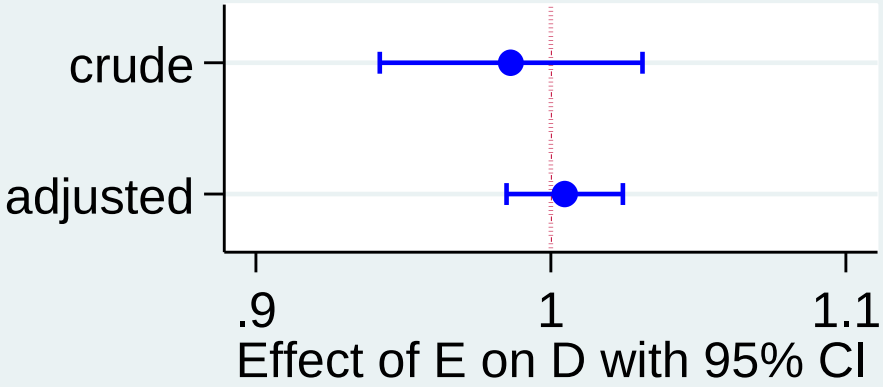
# Effects of adjustment: Precision


Should we adjust for B?

DAG: no bias from B, need not adjust

May include B to improve precision, depends on model!

## Linear regression




Including B: better precision

## Logistic regression


Including B: worse precision

OR not collapsible

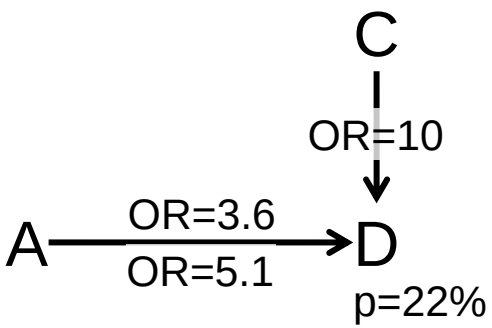
Robinson and Jewell 1991; Xing and Xing 2010

# Non-collapsibility of the odds ratio

(Greenland, Robins et al. 1999, Greenland and Pearl 2011, Sjolander, Dahlqwist et al. 2016)

# Non-collapsibility of the OR

## No confounding




## Population

**D**

| A | | 1 | 0 | sum | odds |
|---|---|---|---|---|---|
| | 1 | 470 | 530 | 1 000 | 0.89 |
| | 0 | 1 775 | 7 225 | 9 000 | 0.25 |
| | | | | 10 000 | |

OR= 3.6

## C=0

**D**

| A | | 1 | 0 | sum | odds |
|---|---|---|---|---|---|
| | 1 | 105 | 395 | 500 | 0.27 |
| | 0 | 225 | 4 275 | 4 500 | 0.05 |
| | | | | 5 000 | |

OR= 5.1

## C=1

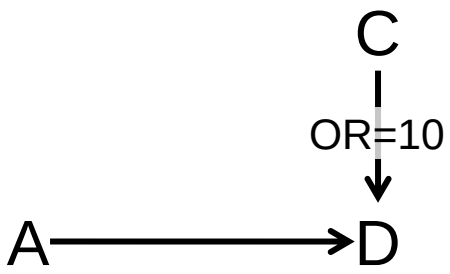
**D**

| A | | 1 | 0 | sum | odds |
|---|---|---|---|---|---|
| | 1 | 365 | 135 | 500 | 2.70 |
| | 0 | 1 550 | 2 950 | 4 500 | 0.53 |
| | | | | 5 000 | |

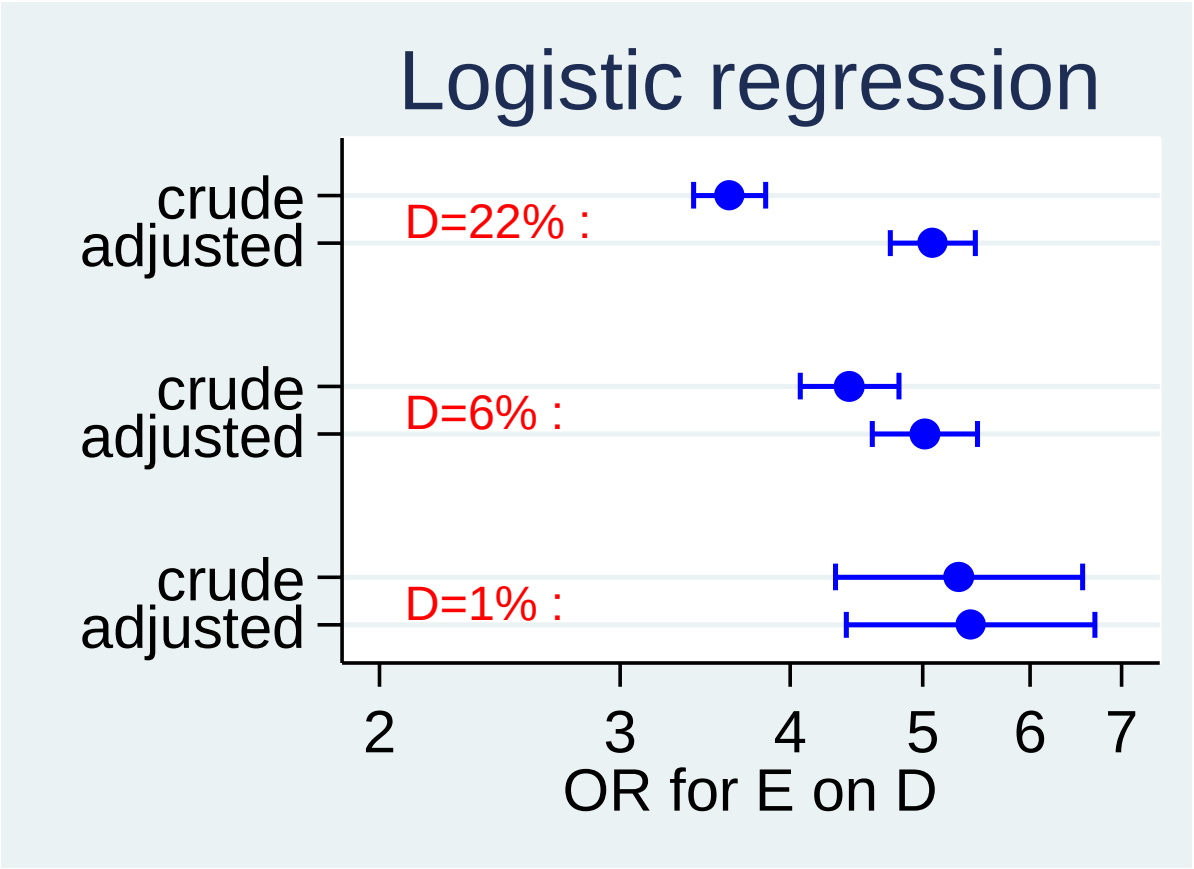
OR= 5.1

(Greenland 1996; Greenland and Pearl 2011; Martinussen and Vansteelandt 2013)

# Prevalence of D and non-collapsibility




Non-collapsibility depends on prevalence of D

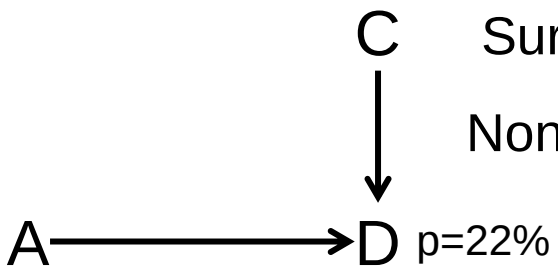



Not collapsible

Appr. collapsible

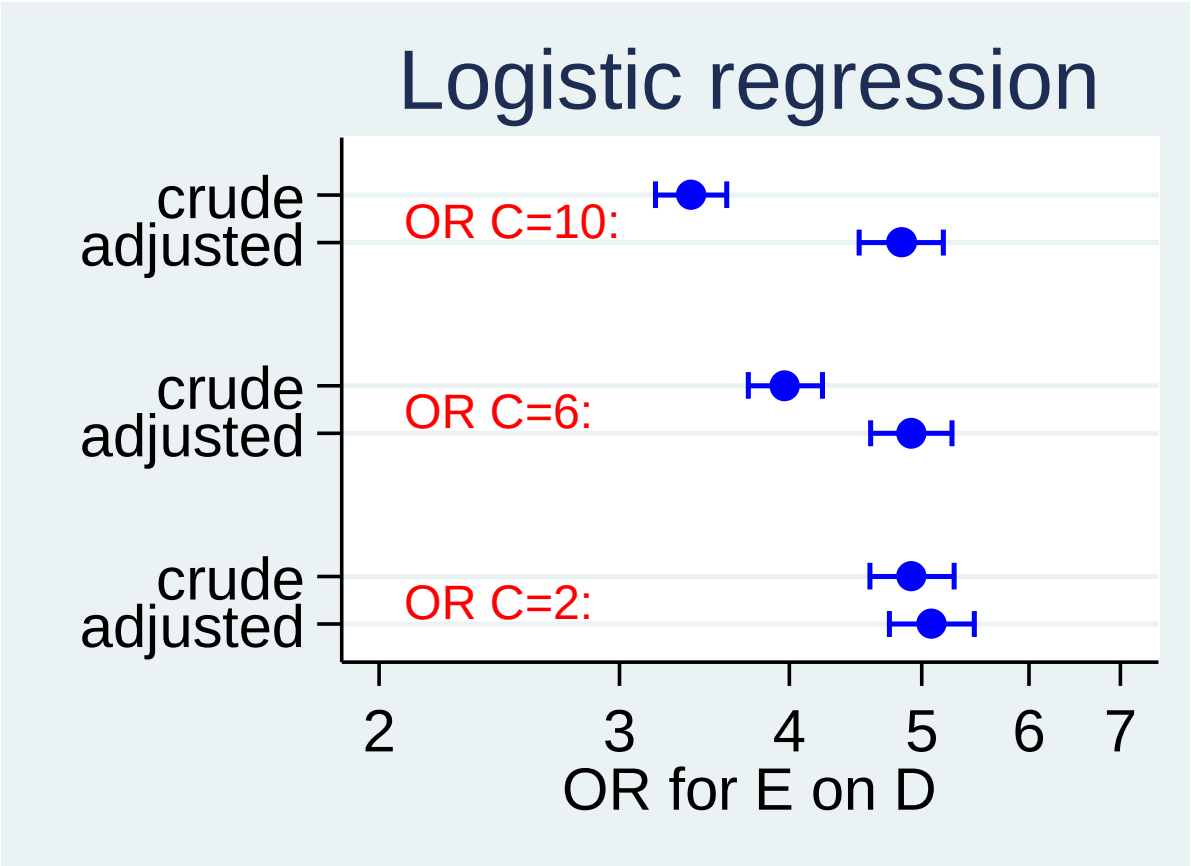
Collapsible

# Effect of C and non-collapsibility




Survival: C is Frailty

Non-collapsibility depends on effect of C of D

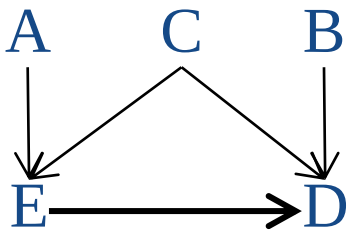



Not collapsible

Not collapsible

Collapsible

# Summing up so far




- Adjustment
  - Never adjust for A (reduce precision)
  - May adjust for B (improve precision in linear models)
  - Adjust for C (remove confounding)

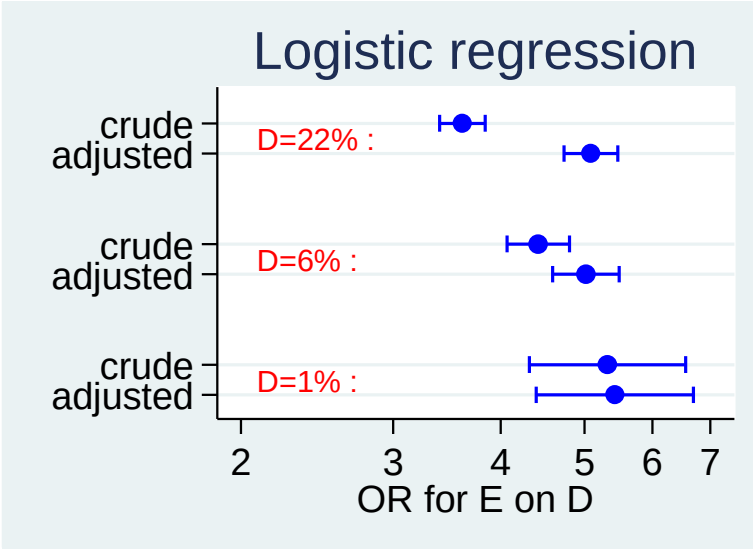
- Collapsibility
  - Collapsible measures:
    - Risk Difference (RD), Rate Difference, Risk Ratio (RR)
  - Non-collapsible measures:
    - Odds Ratio (OR), Rate Ratio (IRR, HRR)

# Do not use OR for common outcomes

## 1. Interpretation

| | RR=1.2 | RR=2 |
|---|---|---|
| Disease risk | OR | OR |
| 1 % | 1.2 | 2.0 |
| 5 % | 1.2 | 2.1 |
| 20 % | 1.3 | 2.6 |
| 40 % | 1.4 | 4.7 |

## 2. Collapsibility




## Use models for RR or RD for common outcomes

binreg D E C, or — estimates OR

binreg D E C, rr — estimates RR

binreg D E C, rd — estimates RD