Directed acyclic graphs - The view of a clinical scientist

Jay Brophy MEng MD FRCP FACC FCCS FCAHS PhD

Nov 3 2021

Conflicts of Interest

I have **no known conflicts** associated with this presentation and to the best of my knowledge, am **equally disliked** by all pharmaceutical and device companies

http://www.nofreelunch.org/

Objectives

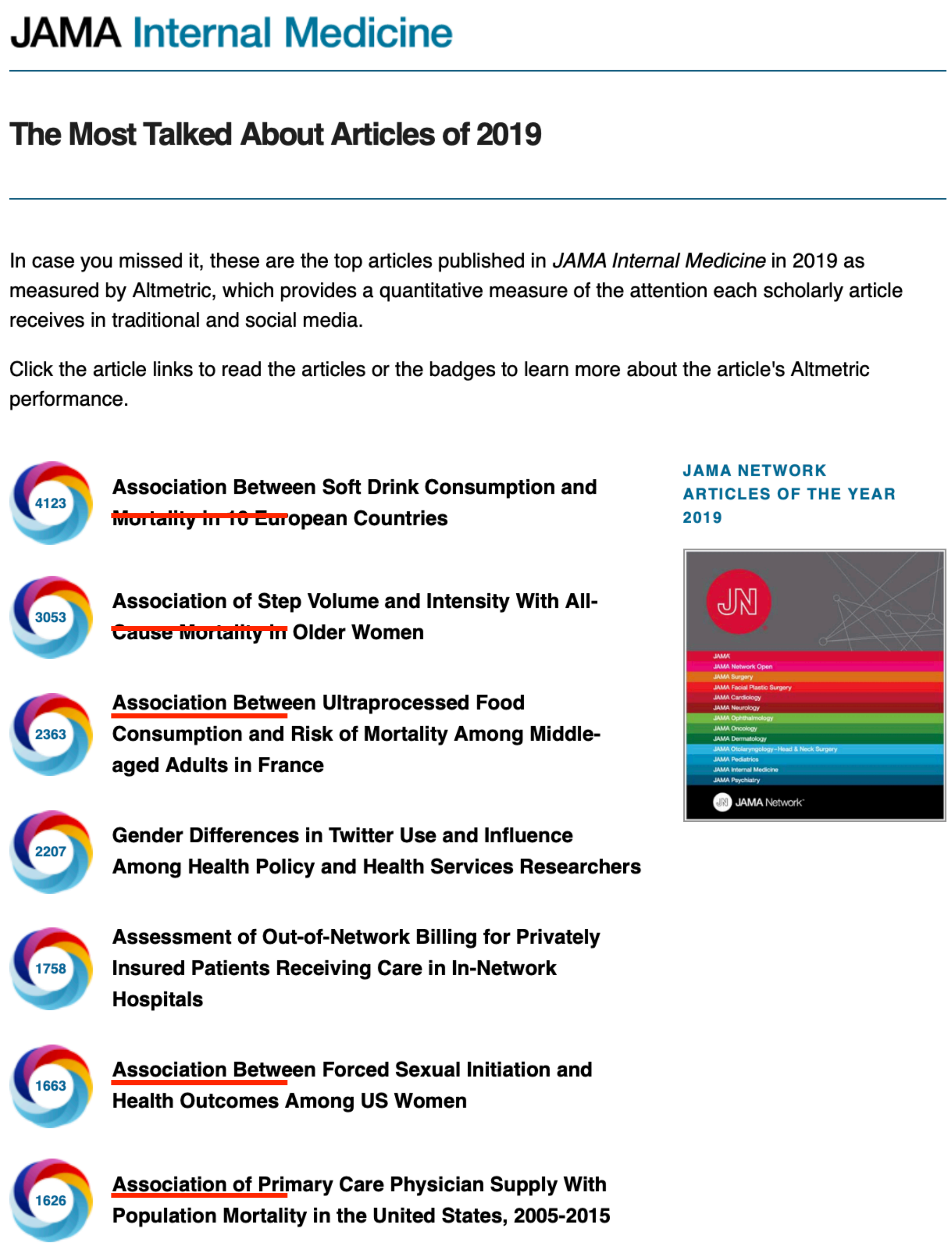
1. Operationalize Directed Acyclic Graphs (DAGs)
2. Appreciate the insights into confounding and selection bias provided by DAGs
3. Examples to appreciate the importance of DAGs (and their encoded substantive knowledge) on the road to causal inference

Felix, qui potuit rerum cognoscere causa - Vigil (29BC)

“Fortunate is he, who is able to know the causes of things”

Background

What we get versus what we want



JAMA Internal Medicine

The Most Talked About Articles of 2019

In case you missed it, these are the top articles published in *JAMA Internal Medicine* in 2019 as measured by Altmetric, which provides a quantitative measure of the attention each scholarly article receives in traditional and social media.

Click the article links to read the articles or the badges to learn more about the article's Altmetric performance.

4123 Association Between Soft Drink Consumption and Mortality in 10 European Countries

3053 Association of Step Volume and Intensity With All-Cause Mortality in Older Women

2363 Association Between Ultraprocessed Food Consumption and Risk of Mortality Among Middle-aged Adults in France

2207 Gender Differences in Twitter Use and Influence Among Health Policy and Health Services Researchers

1758 Assessment of Out-of-Network Billing for Privately Insured Patients Receiving Care in In-Network Hospitals

1663 Association Between Forced Sexual Initiation and Health Outcomes Among US Women

1626 Association of Primary Care Physician Supply With Population Mortality in the United States, 2005-2015

JAMA NETWORK ARTICLES OF THE YEAR 2019

CAUSES OF ASSOCIATIONS

- **Treatment (T) causes Outcome (Y)**
- Y causes T (reverse causality)
- T and Y share a common cause (confounding)
- Induced by conditioning on a common effect of T and Y (selection bias)
- Random fluctuations

Conventional statistical paradigm versus DAGs

- "The object of **statistical methods** is the **reduction of data**" (Fisher 1922) -> a parsimonious mathematical description of the joint distribution of observed variables
  - Good statistical processes can describe the data but say nothing about the data generating process and **can't answer causal questions**

- **DAGs** (AKA causal diagrams) **characterize causal structures** compatible with the observations & **assist in drawing logical conclusions** about the statistical relations
  - Help understand confounding, selection bias, covariate selection, over adjustment, instrumental variable analyses & avoid making errors about the statistical relations

Canonical study # 1

- Study with 350 exposed to a drug and 350 controls

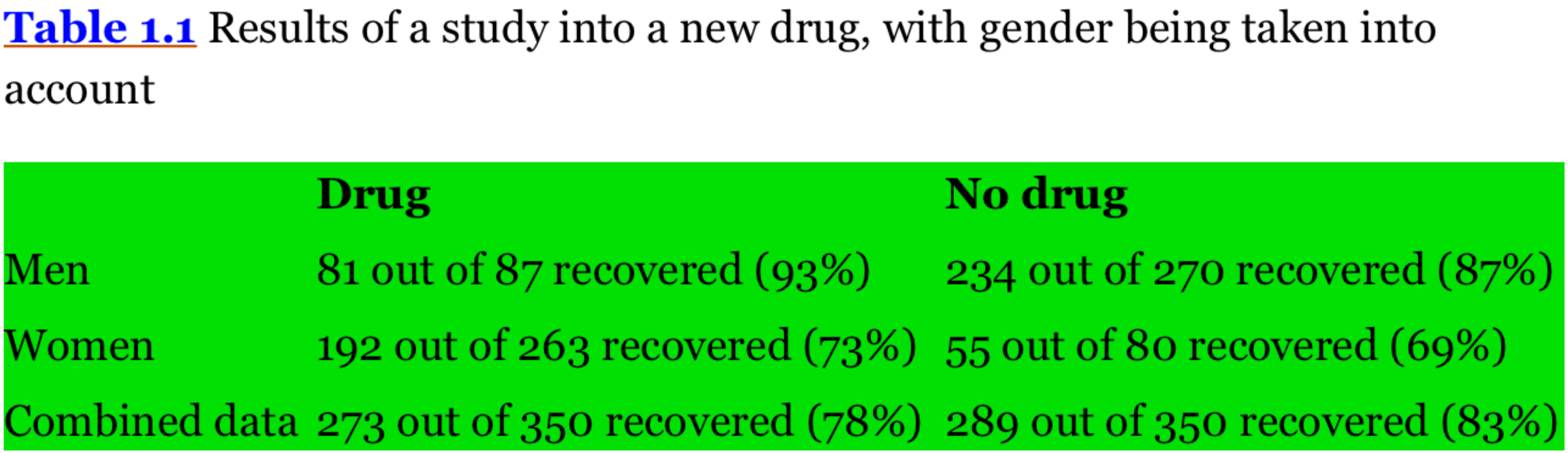


Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- Does the drug work? **Overall** population or gender **subgroups**?
- Since it works in men and women, makes no sense to say it doesn't work if gender is unknown
- Is it a general rule that more specific subgroups should always take precedence over the marginal?

Canonical study # 2

- A different experiment with a different drug that lowers BP but it also with toxic side effects, gives the **same data**

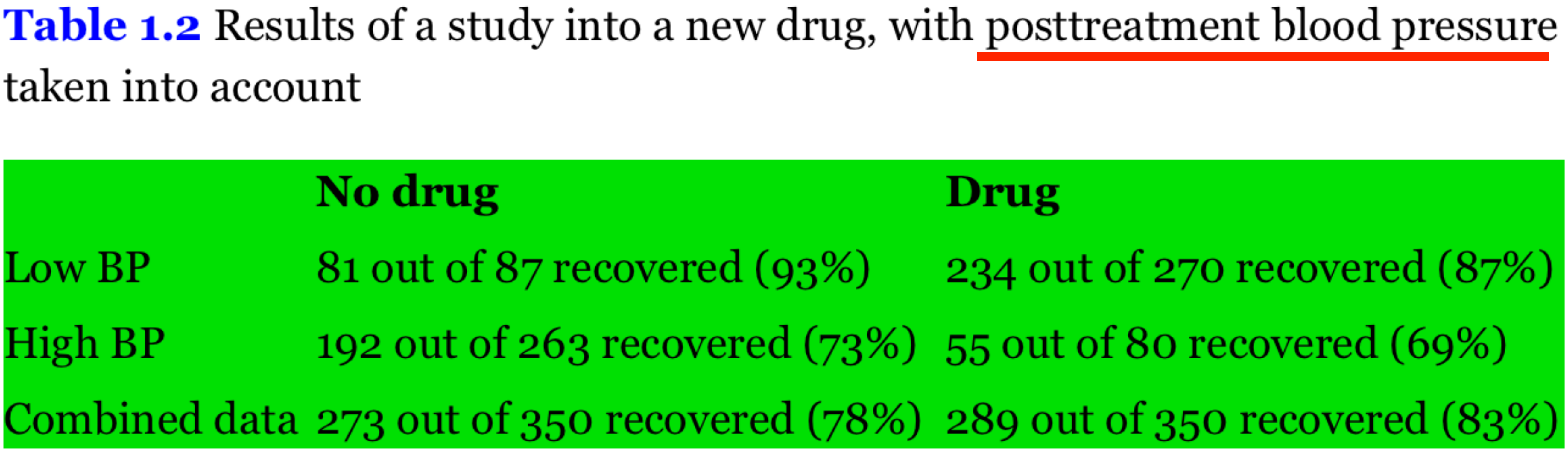


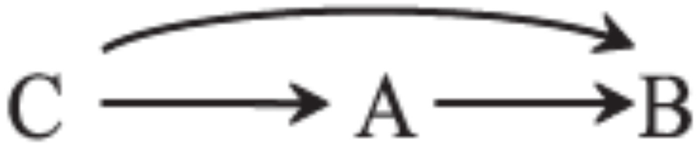
Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	**Drug**
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- Does the drug work? **Overall** population or specific **subgroups**?
- Why is **aggregate** data more informative here, same data as before?
- By stratifying, don't see the positive drug effects from BP lowering, capturing mostly negative toxic effects

Resolving the Paradox

- In the first experiment,



- where C = gender in #1

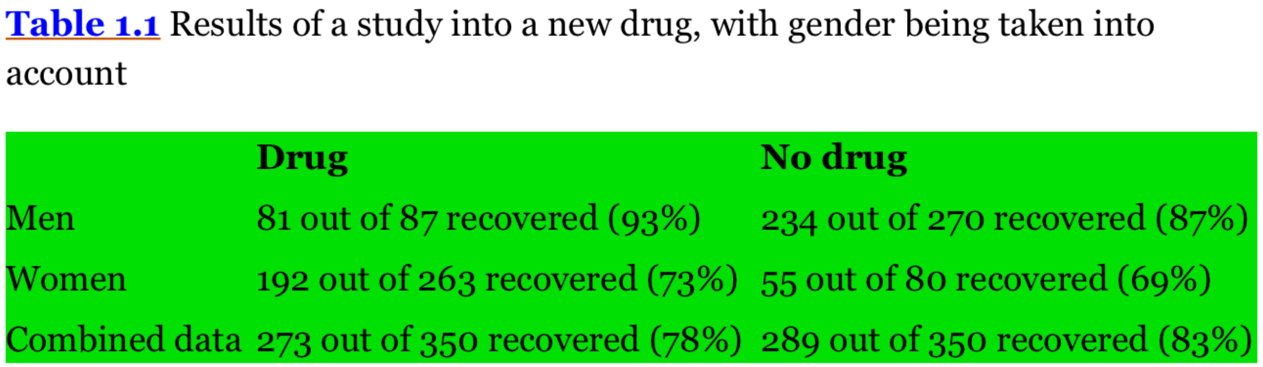
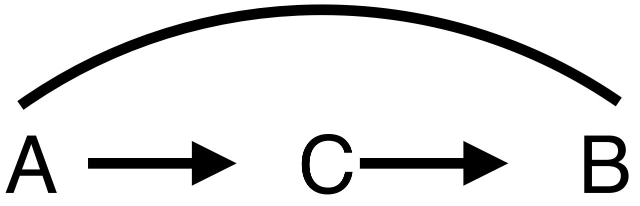


Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	**No drug**
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

the second experiment



C = low BP in #2

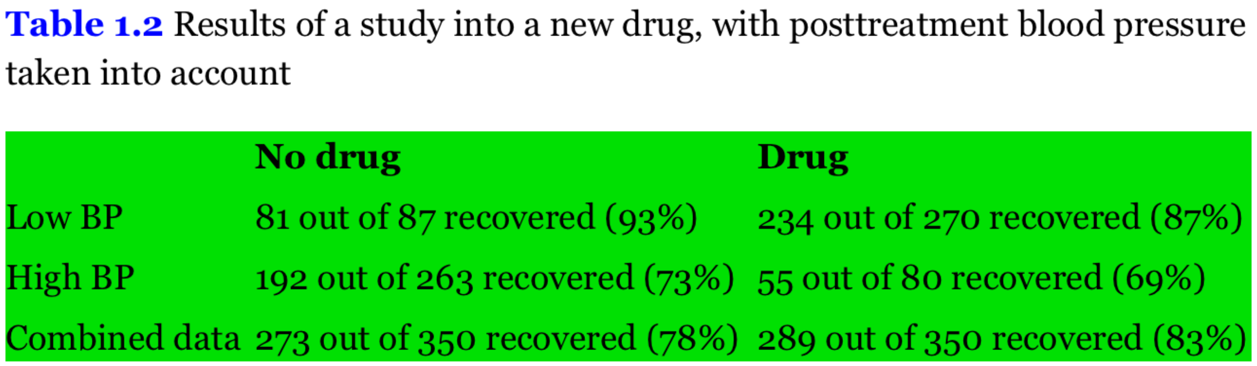


Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	**Drug**
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- Experiment #1 C is a confounder and need to adjust
- Experiment #2 C is in the causal pathway and adjusting creates bias
- Causal interpretations can only be made by the sensible inclusion of external judgement or evidence
- 2X2 tables alone express no causal information

Causal inference

Knowing a cause means being able to predict the consequences of an intervention (***What if I do this?***)

Knowing a cause means being able to construct unobserved counterfactual outcomes. (***What if I had done something else?***)

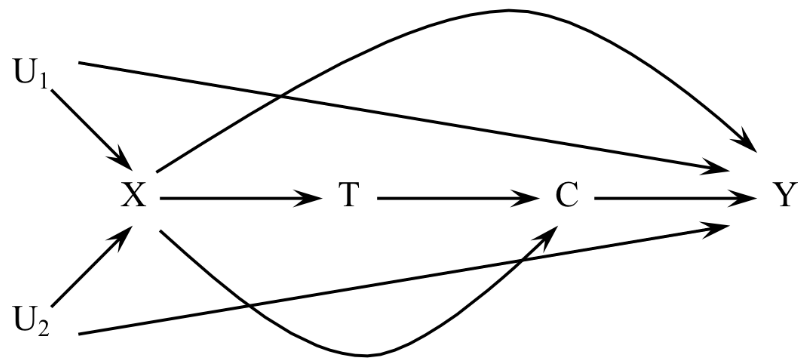
DAG principles of operation

Directed acyclic diagrams - raison d'etre

DAGs encode qualitative a priori subject matter knowledge and consideration of the causal model may provide clarity in interpreting statistical coefficients and causal inferences

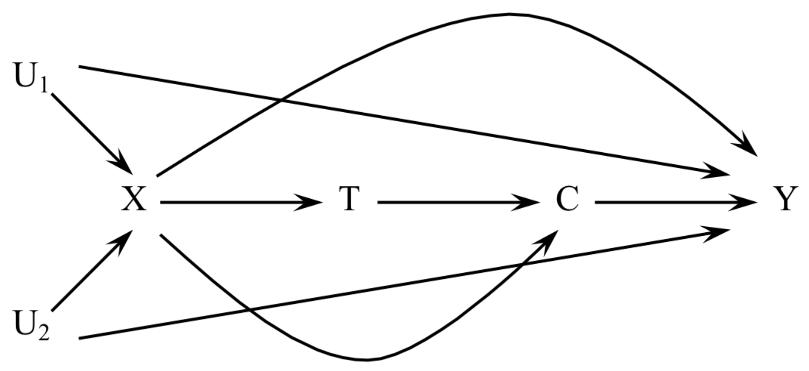
Corollary: Assumption - free causal inference doesn't exist

DAGs - Help identify causal effects



- Non-parametric visual representations of the joint distribution
- **Variables** are depicted as **nodes** and connected by **arrows**
- Acyclic (the future can’t predict the past)
- **Missing lines** strongest assumption, variable independence.
- Include all common causes of any 2 variables & all variables involved in data generation - observed or unobserved
- Contain both causal and non-causal pathways
- **Help identify causal effects by deriving testable implications of a causal model**

More DAG Terminology



Path is a sequence of non-intersecting adjacent edges X->T->C or U_2->Y<-C<-T

Causal path: a path in which all arrows point away from T to outcome Y; T->C->Y

Total causal effect of a treatment on an outcome consists of all causal paths connecting them

Non-causal path: path connecting T and Y in which at least one arrow points against flow of time T<-X->Y

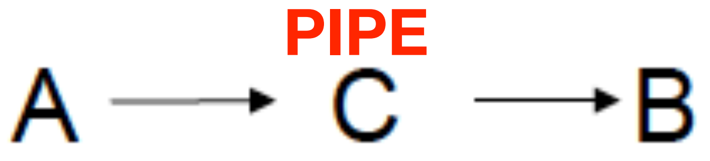
Descendants of a node: all nodes directly or indirectly caused by the node; desc(T) = {C,Y}

Children of a node: all nodes directly caused by the node; child(T) = {C}

Ancestors of a node: all nodes directly or indirectly causing the node; an(T) = {X, U_1, U_2}

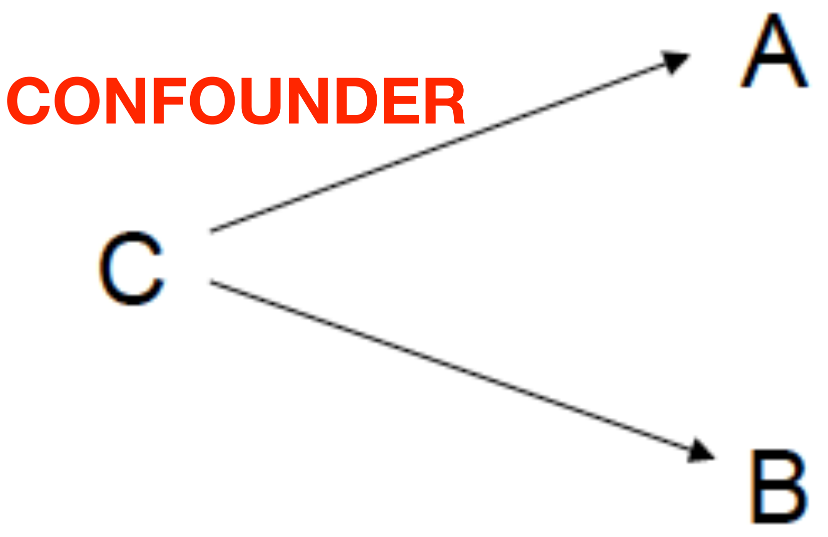
Collider variable along a path with 2 arrows pointing in U->X<-U_2

DAGs Between Two Variables



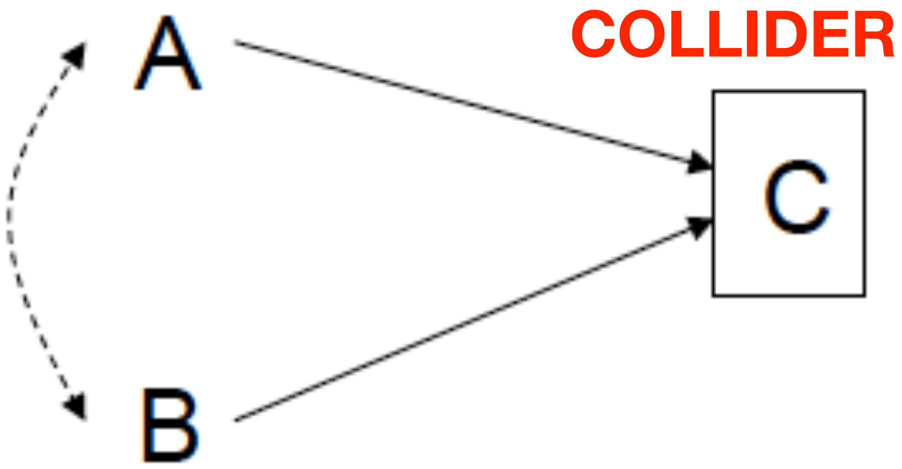
(1) Direct and indirect causation

$A \not\perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B|C$



(2) Common cause confounding

$A \not\perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B|C$



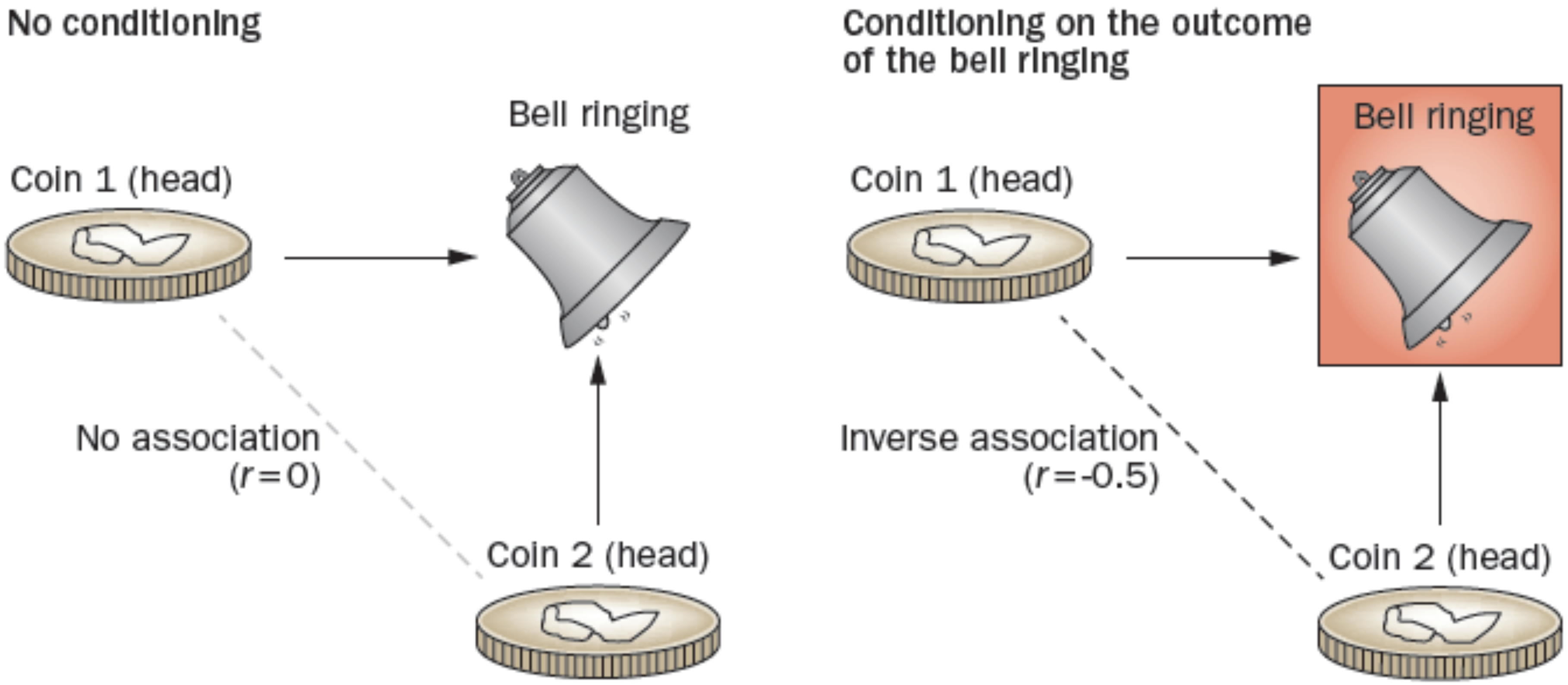
(3) Conditioning on a common effect (“collider”): Selection

$A \perp\!\!\!\perp B$ and $A \not\perp\!\!\!\perp B|C$

←------→ : non-causal (spurious) association. □ : conditioning.

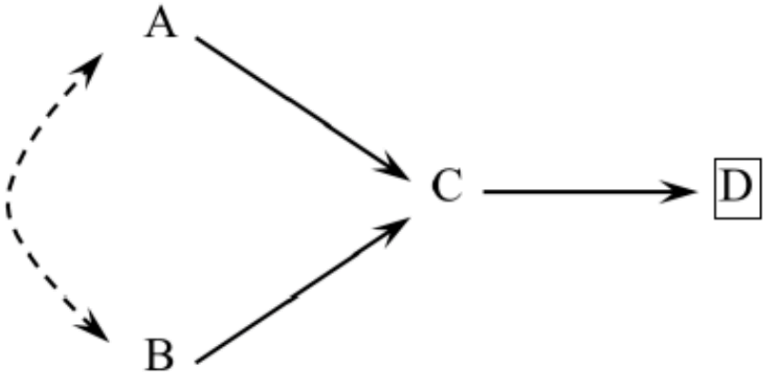
Conditioning on a common effect

- Bell rings whenever either coin comes up heads on a toss of both
- Obviously if bell rang and we know Coin 1 was tail -> Coin 2 was heads

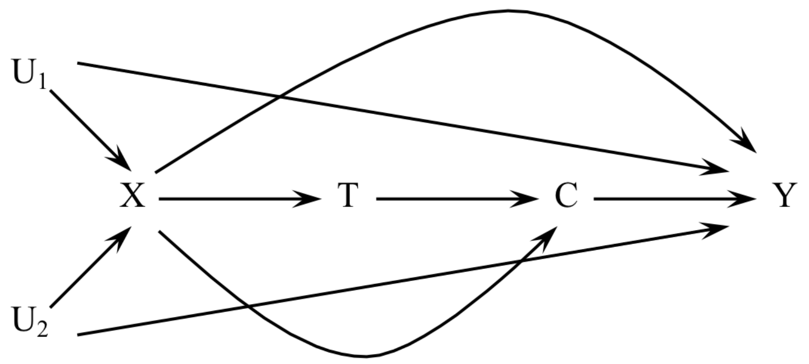


Conditioning on a common effect induces a negative correlation between two causes or ‘risk factors’

Even conditioning on descendant of C can lead to a spurious association

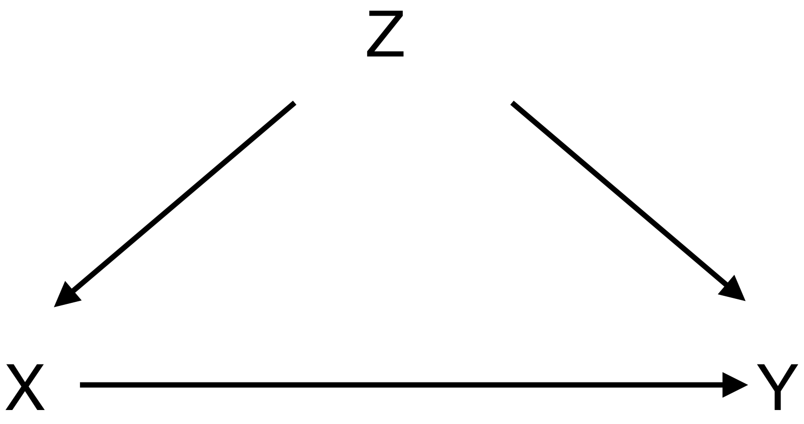


More DAG Terminology

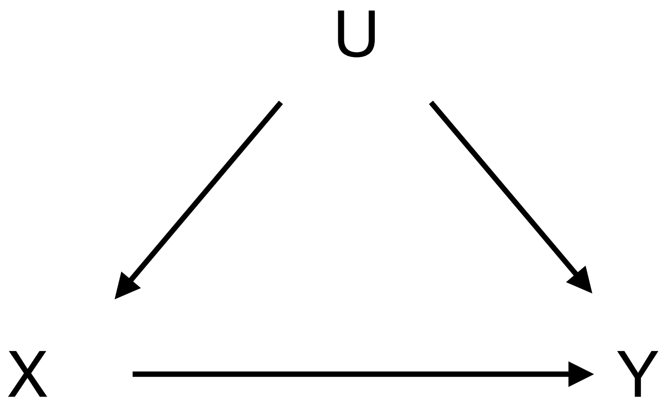


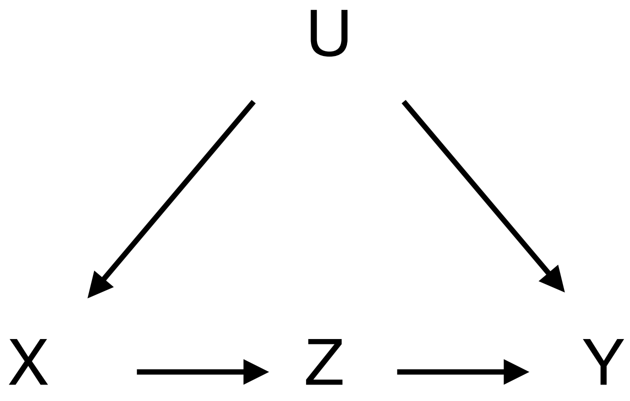
- "Blocked" (d-separated) paths don't transmit associations
- "Unblocked" (d-connected) paths may transmit association
- Three blocking criteria
  - Conditioning on a non-collider blocks a path
  - Conditioning on a collider, or a descendent of a collider, unblocks a path
  - Not conditioning on a collider leaves a path "naturally" blocked.
- Implication:
  - If X and Y are d-separated by Z along all paths in a DAG, then X is statistically independent of Y conditional on Z in every distribution compatible with the DAG
  - If X and Y are not d-separated by Z along all paths in the DAG, then X and Y are dependent conditional on Z in at least one distribution compatible with the DAG

Estimating a causal effect



- Backdoor criteria
- Z is a sufficient set
  - (1) no variable in Z is a descendant of X and
  - (2) every path between X and Y that contains an arrow into X is blocked by Z.

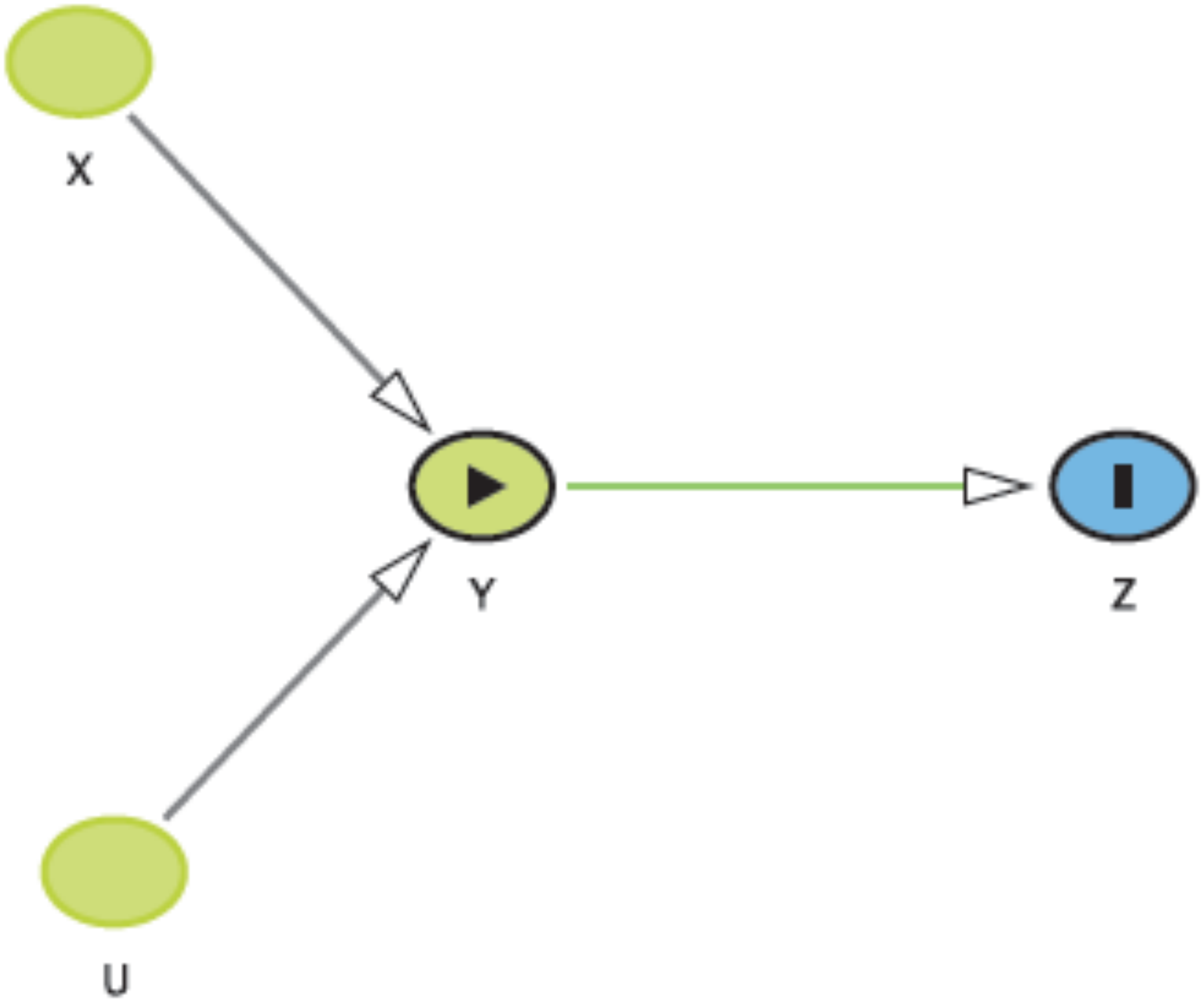




- Front door criteria
- Z is a sufficient set
  - Z intercepts all directed paths from X to y
  - No unblocked paths from X to Z
  - All backdoor paths from Z to Y are blocked by X

What is this “simple” DAG implying?

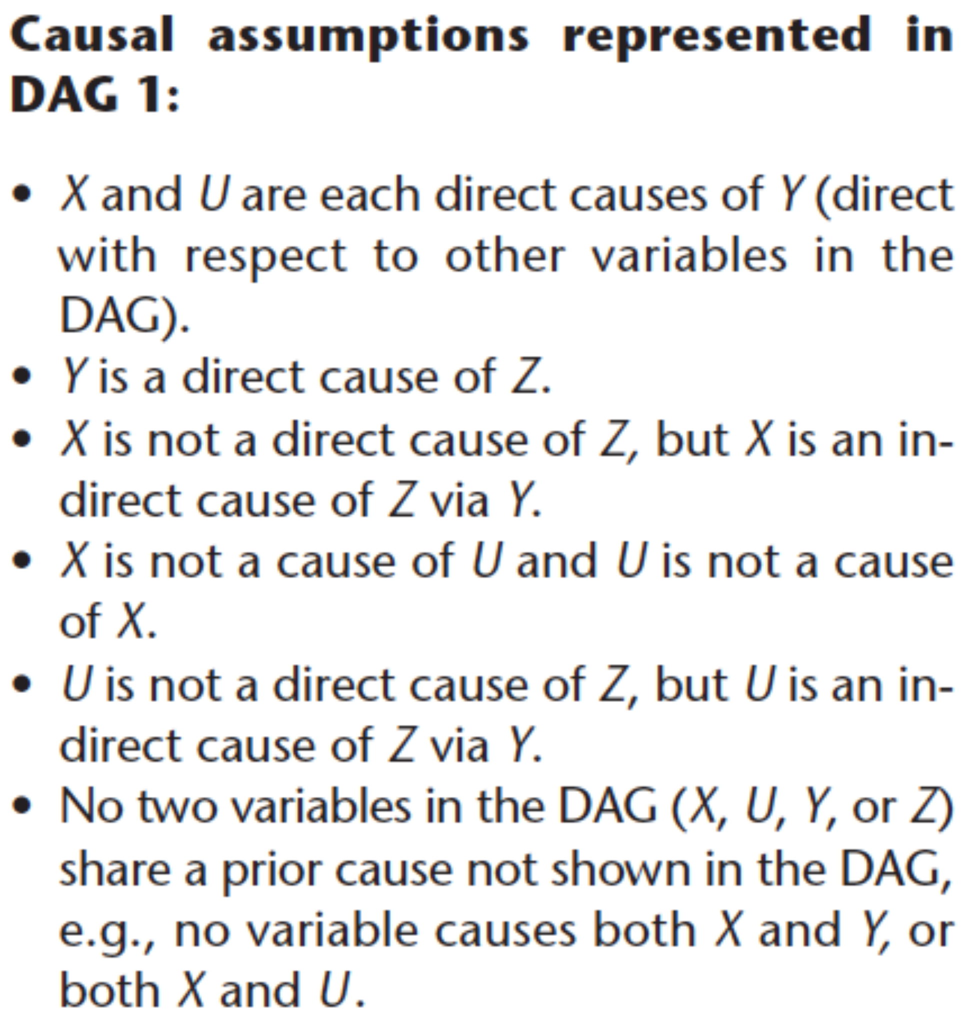
- What are the contained assumptions & statistical implications of this model?



Would you believe at least 16 assumptions and statistical implications!

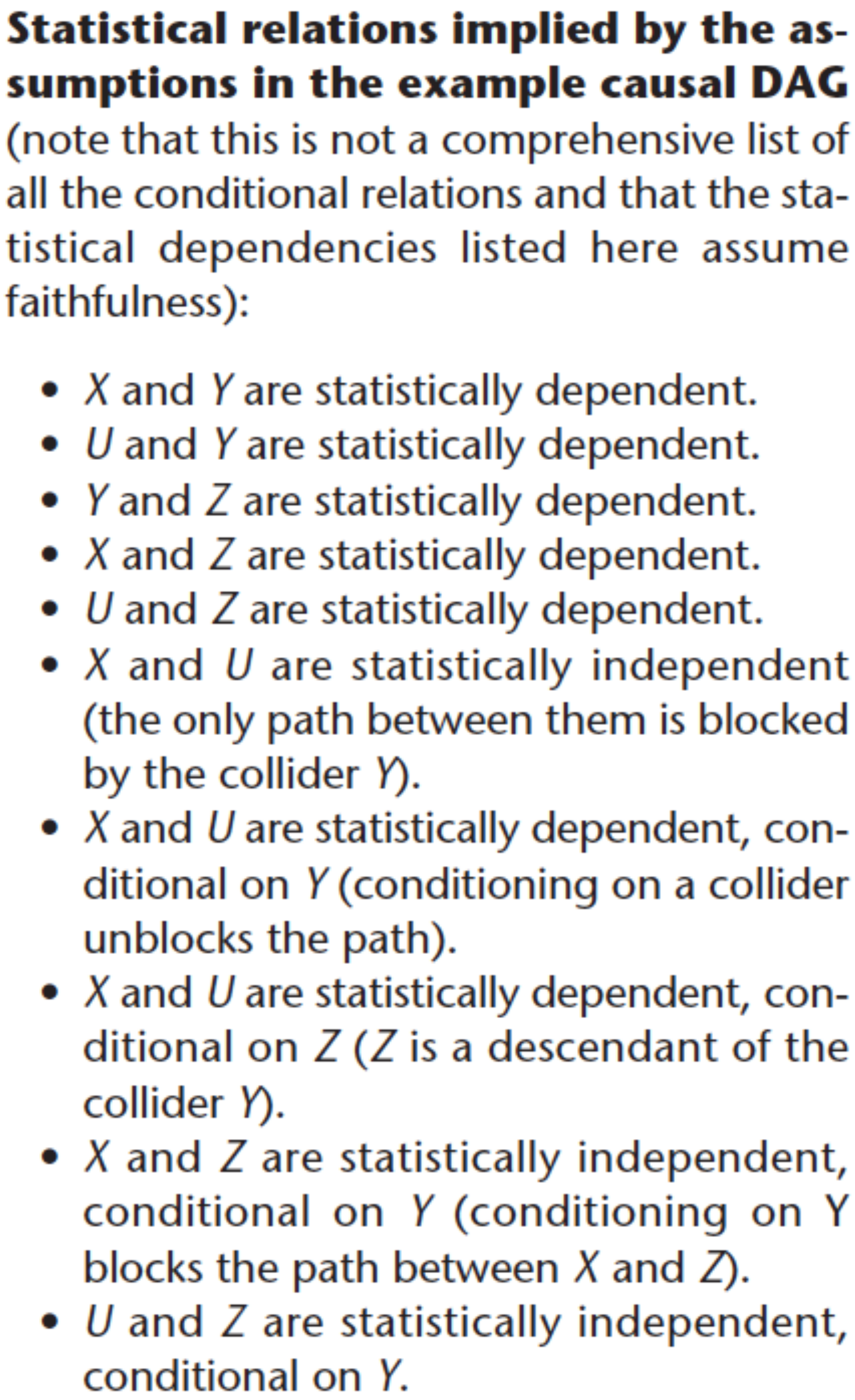
It is saying quite a lot!

What are the contained assumptions & statistical implications of this model?



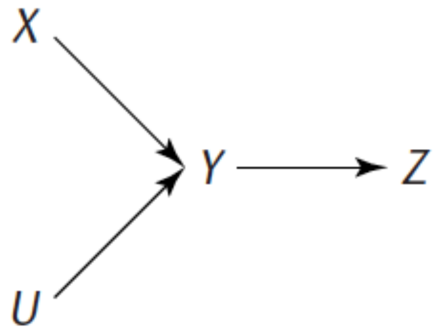
Causal assumptions represented in DAG 1:

- *X* and *U* are each direct causes of *Y* (direct with respect to other variables in the DAG).
- *Y* is a direct cause of *Z*.
- *X* is not a direct cause of *Z*, but *X* is an indirect cause of *Z* via *Y*.
- *X* is not a cause of *U* and *U* is not a cause of *X*.
- *U* is not a direct cause of *Z*, but *U* is an indirect cause of *Z* via *Y*.
- No two variables in the DAG (*X*, *U*, *Y*, or *Z*) share a prior cause not shown in the DAG, e.g., no variable causes both *X* and *Y*, or both *X* and *U*.



Statistical relations implied by the assumptions in the example causal DAG (note that this is not a comprehensive list of all the conditional relations and that the statistical dependencies listed here assume faithfulness):

- *X* and *Y* are statistically dependent.
- *U* and *Y* are statistically dependent.
- *Y* and *Z* are statistically dependent.
- *X* and *Z* are statistically dependent.
- *U* and *Z* are statistically dependent.
- *X* and *U* are statistically independent (the only path between them is blocked by the collider *Y*).
- *X* and *U* are statistically dependent, conditional on *Y* (conditioning on a collider unblocks the path).
- *X* and *U* are statistically dependent, conditional on *Z* (*Z* is a descendant of the collider *Y*).
- *X* and *Z* are statistically independent, conditional on *Y* (conditioning on Y blocks the path between *X* and *Z*).
- *U* and *Z* are statistically independent, conditional on *Y*.



DAG additional insights

Confounding Evaluation

- Common strategies to decide whether a variable is a confounder rely mostly on statistical criteria.
  - checking if classic confounding definition is + (causally associated with the outcome, non-causally or causally associated with the exposure & not an intermediate variable on the causal pathway)
  - compare stratified to marginal effect estimates
  - compares adjusted & unadjusted effect estimates
  - **automatic variable selection** - letting multiple regression sort it out or “Let the data speak” - (IMHO, if the data are speaking to you, time to acknowledge some mental health issues)

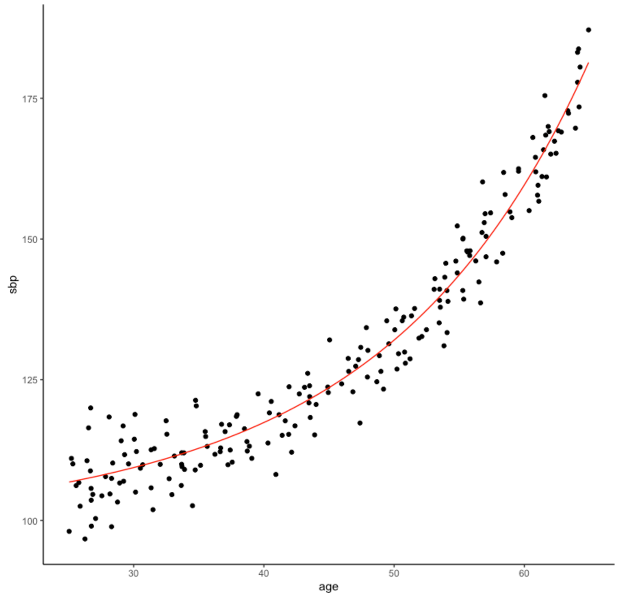
- Regression models alone insufficient
  - offer no distinction of causes from confounders
  - often ignore residual confounding, measurement error & missing data
  - may contain **causal misinformation** (Table 2 fallacy Am J Epidemiol. 2013;177(4):292-8)

- All these strategies may lead to bias

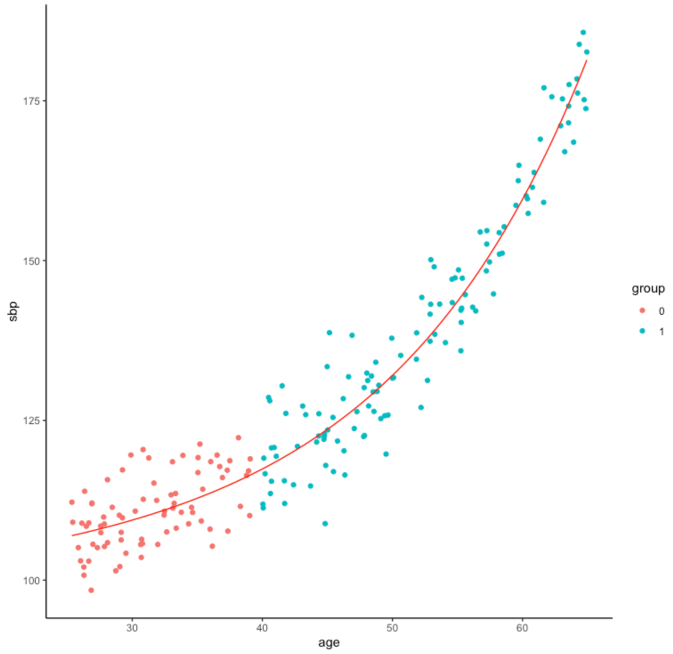
Automated statistical software

A Generate data SBP =f(age), ∐ group

$$SBP = 99 + 0.1 * age + \exp(age / 15)$$



B unexposed group younger



Now what if propose a linear regression: SBP = a + b.age + c.group

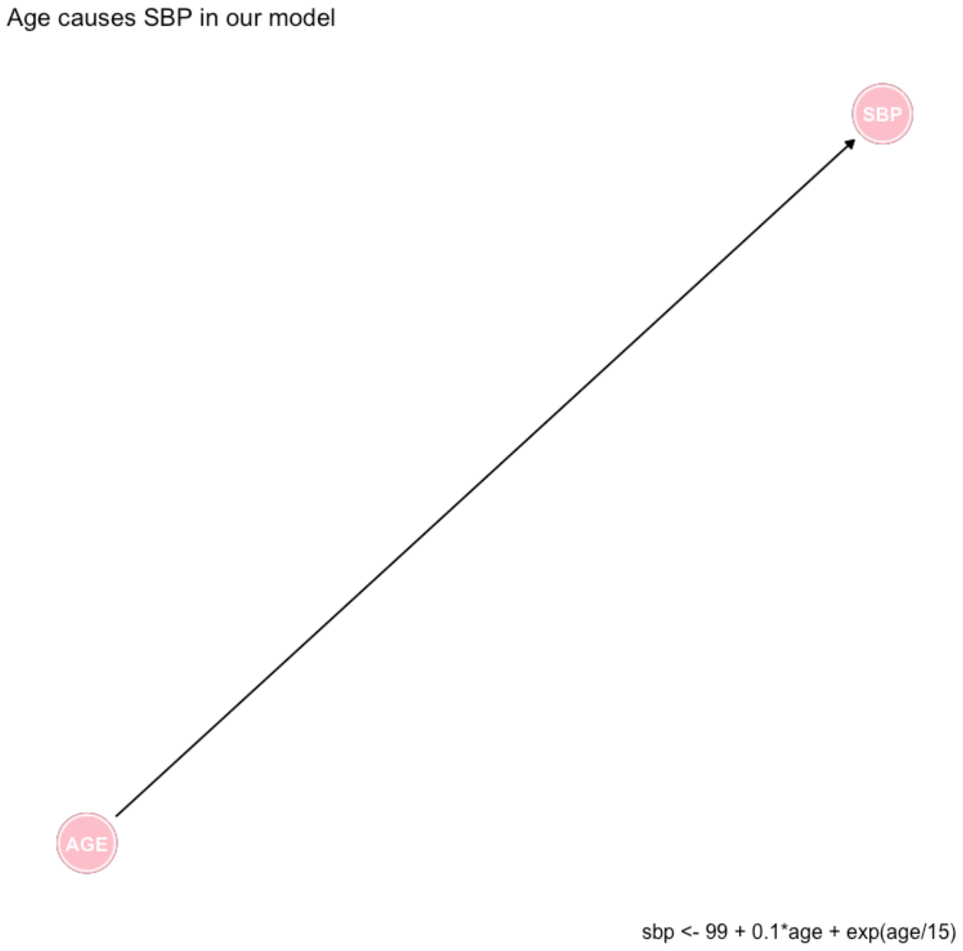
```
lm(formula = sbp ~ age + as.numeric(drug), data = dat)
Coefficients:
              Estimate    Std. Error    t value       Pr(>|t|)
(Intercept)   40.7        2.7           14.98         < 2e-16
age           2.2         0.08          26.88         < 2e-16
group         -14.7    2.0           -7.31            6.6e-12
```

“Controls for age” -> a spurious statistically difference in SBP & exposure groups, yet data generated with no group exposure effect

p-values will not pick the causally correct model

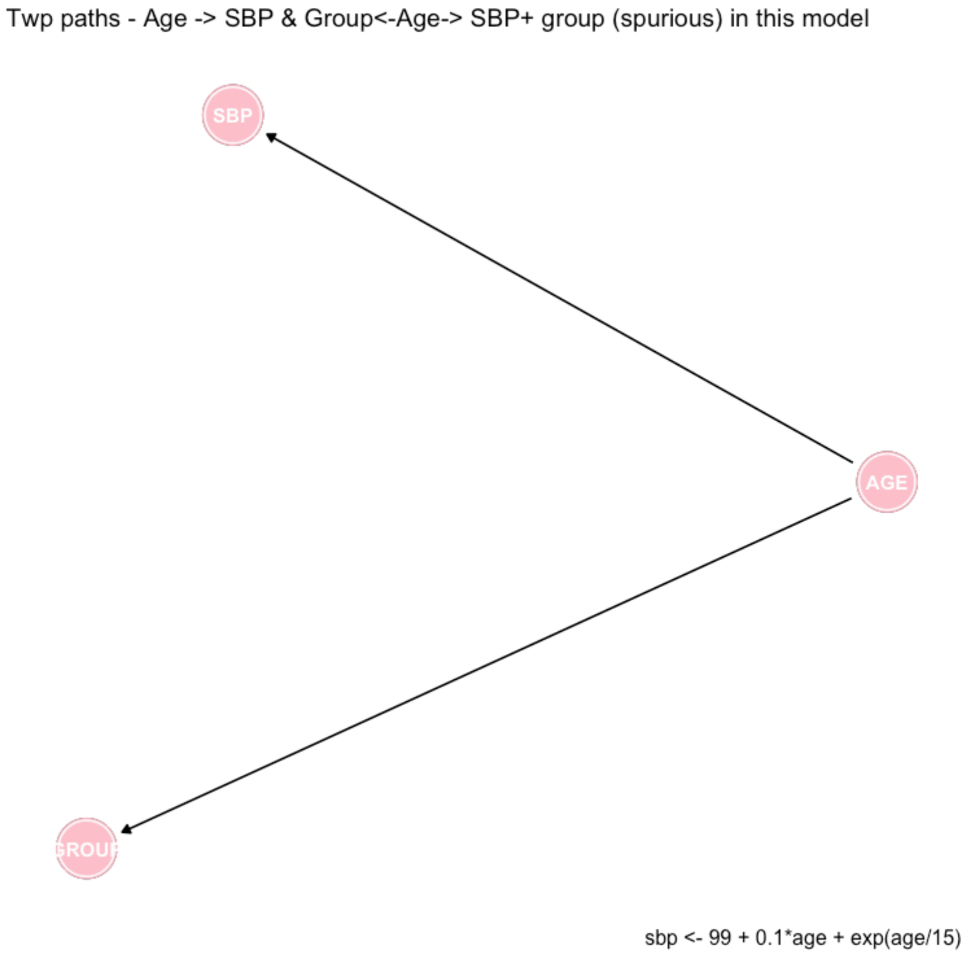
Can DAGs help explain this phenomena?

Generated causal model


Age causes SBP in our model

sbp <- 99 + 0.1*age + exp(age/15)

Automated

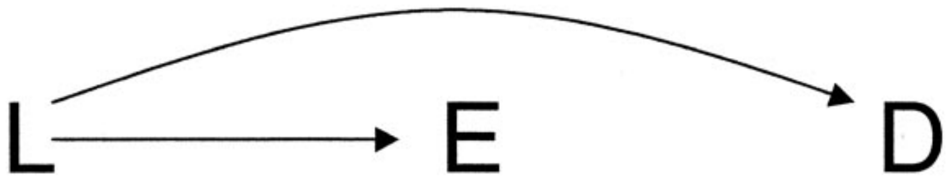

Twp paths - Age -> SBP & Group<-Age-> SBP+ group (spurious) in this model

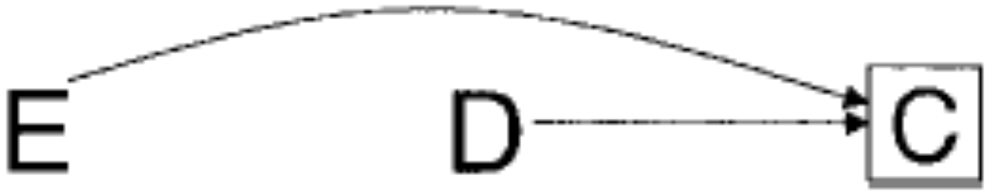
sbp <- 99 + 0.1*age + exp(age/15)

Only 1 causal path in our generated model - Age -> SBP
Adding group adds a second spurious path Group <- Age -> SBP

Selection bias & confounding

- Two important biases, not always easy to distinguish
- Terminology can be confusing - cf what is the difference between “confounding by indication” vs. “selection bias”?
- One way to distinguish is with DAGs
  - Presence of common causes -> “confounding”
  - Conditioning on common effects -> “selection bias”



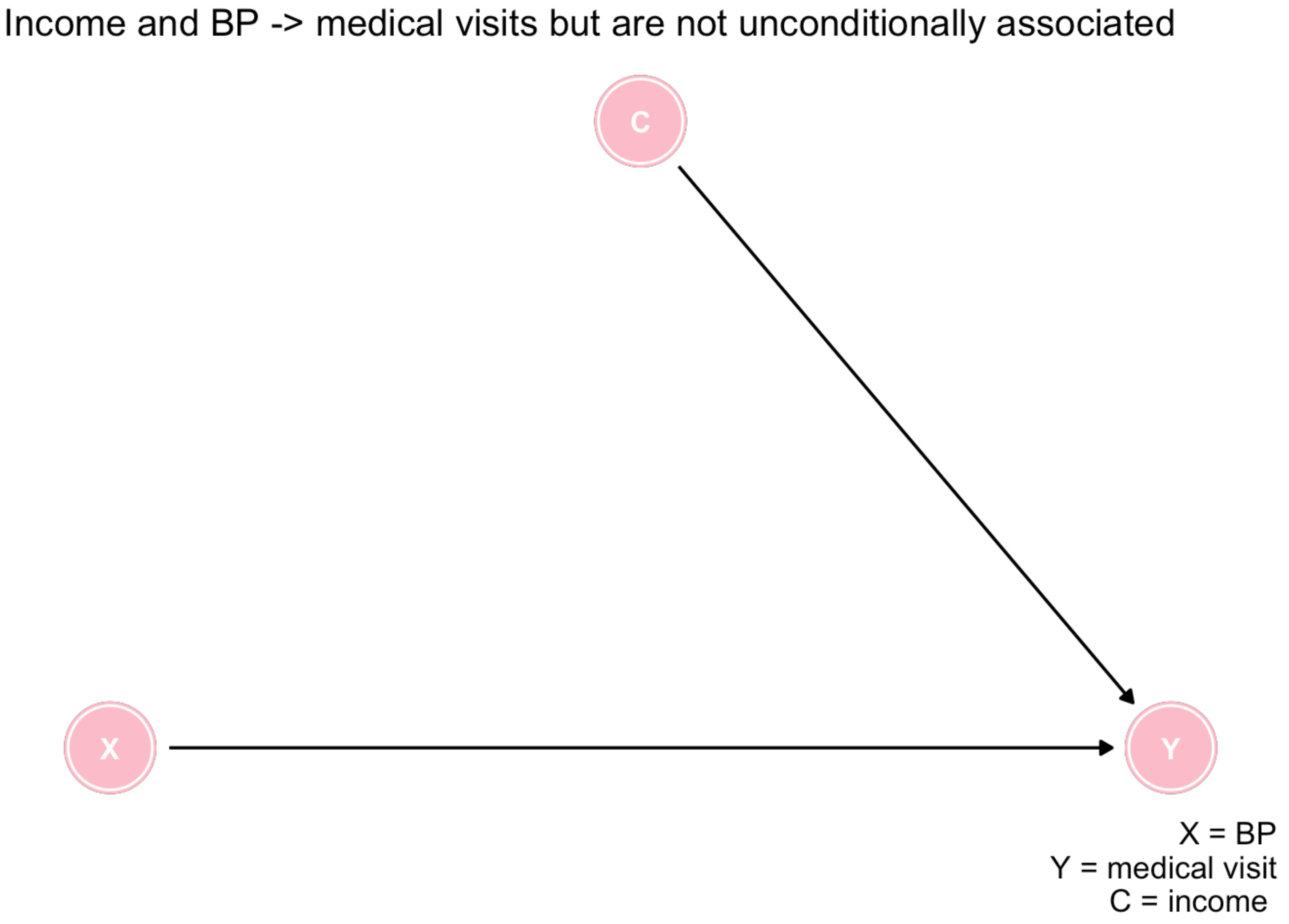


- **Confounding - state of nature; Selection bias - artifact of research process**
- Result of both is noncomparability (also referred to as lack of exchangeability) between the exposed and the unexposed

Selection bias

- Occurs when exposure and a disease outcome both affect participation in the study.
  - **Enrolment** if the variables affect initial participation (typically case-control studies)
  - **Withdrawal** if there are differential losses to follow-up (cohort studies & RCTs)
- Classic examples -
  - Berkson, healthy-worker bias, volunteer bias, selection of controls into case-control studies, differential loss-to-followup, depletion of susceptibles, incidence - prevalence, and nonresponse (complete case - informative censoring)
- Selection bias is often is **difficult to identify & frequently overshadowed by other bias but remains ubiquitous**

Understanding selection bias (colliders)



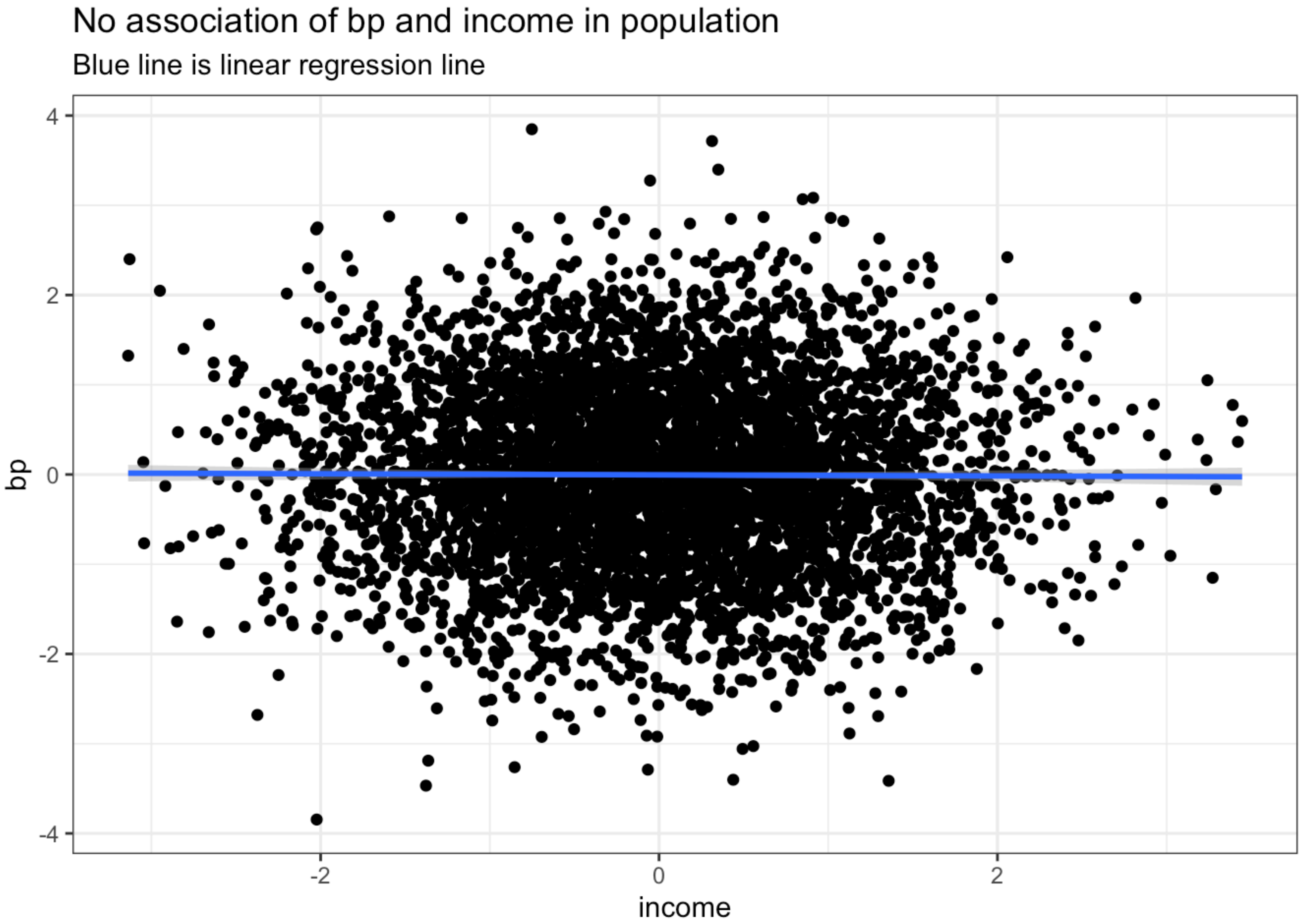
Income and BP -> medical visits but are not unconditionally associated

X = BP
Y = medical visit
C = income

```
dag <- dagitty::dagitty("dag {
  X -> Y
  C ->  Y
              }")

coordinates( dag ) <-  list(
  x=c(X=1, C=3, Y=5),
  y=c(X=1, C=3, Y=1) )

dag <- ggdag::tidy_dagitty(dag)
ggdag::ggdag(dag, layout = "circle") +
  ggdag::theme_dag_blank(plot.caption = element_text(hjust = 1)) +
  ggdag::geom_dag_node(color="pink") + ggdag::geom_dag_text(color="white") +
  ggtitle("Income and BP -> medical visits but are not unconditionally associated") +
  labs(caption = "X = BP\nY = medical visit\nC = income ")
```

R code

Understanding selection bias (colliders)

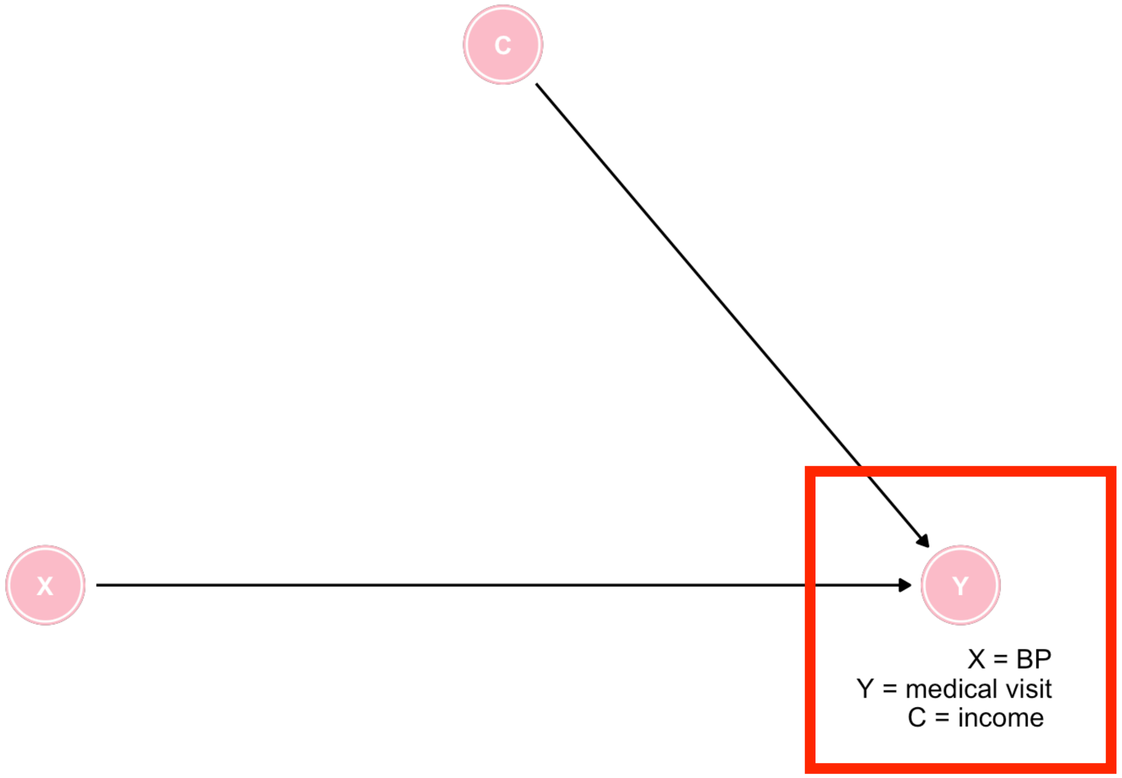


R code

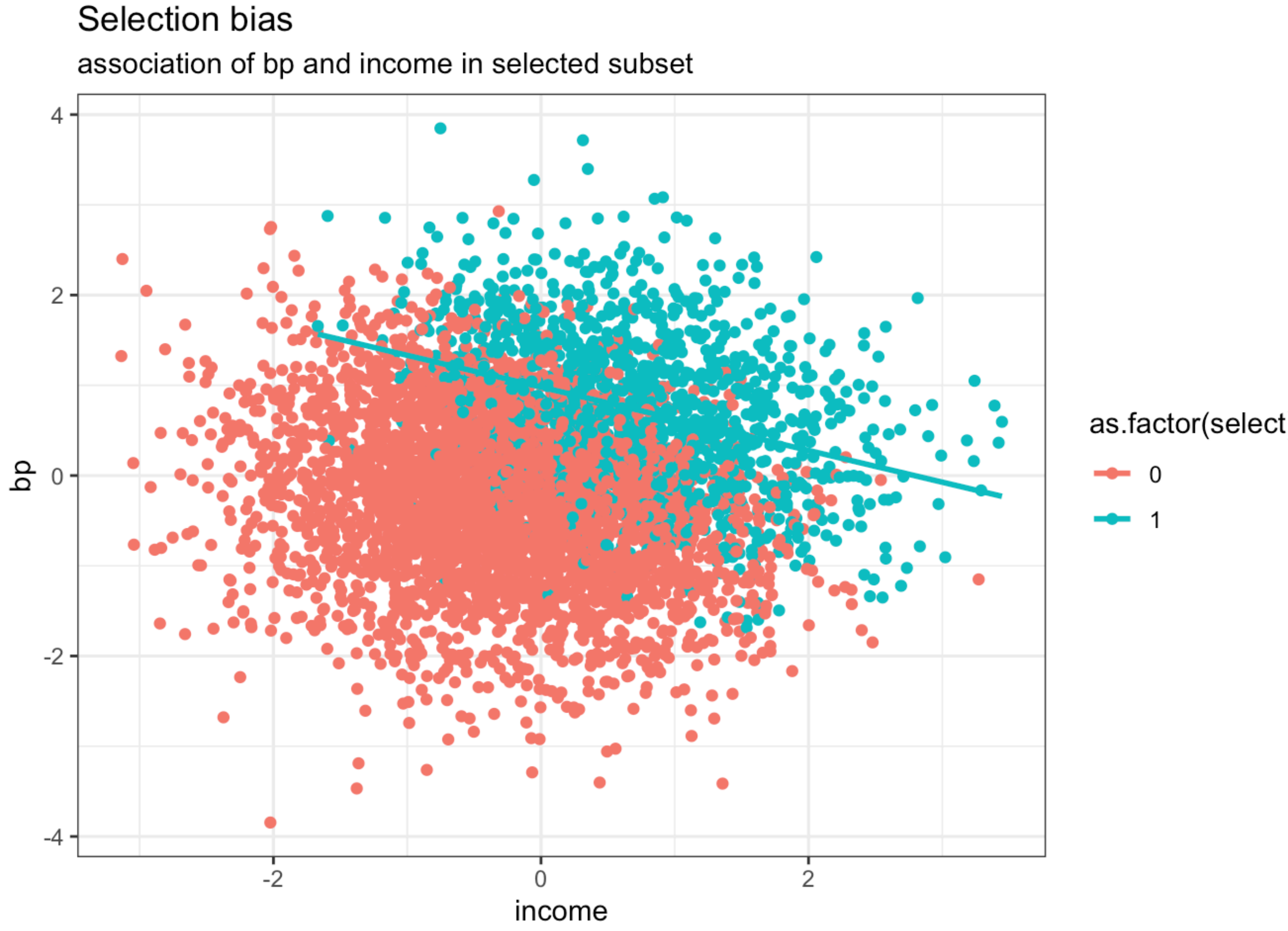
```
n = 5000
set.seed(123)

income <- rnorm(n)      #simulate independent income and bp data
bp <- rnorm(n)

ggplot(data.frame(income,bp), aes(income, bp)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x) +
  labs(title = "No association of bp and income in population", subtitle = "Blue line is linear regression line") +
  theme_bw()
```

Understanding selection bias (colliders)

Income and BP -> medical visits but are not unconditionally associated



X = BP
Y = medical visit
C = income


Selection bias

association of bp and income in selected subset

R code

```
logitVisit <- -2 + 2*income + 2*bp
pVisit <- 1/(1+exp(-logitVisit))
# easier to use inverse function expit locfit::expit(logitVisit)
visit <- rbinom(n, 1, pVisit)

dPop <- data.table::data.table(income, bp, visit)
dSample <- dPop[visit == 1]

ggplot(dPop, aes(income, bp, color=as.factor(visit))) +
   geom_point() +
   geom_smooth(data= dSample, method = "lm", se = FALSE) +
   labs(title = "Selection bias", subtitle = "association of bp and income in selected subset") +
   theme_bw()
```

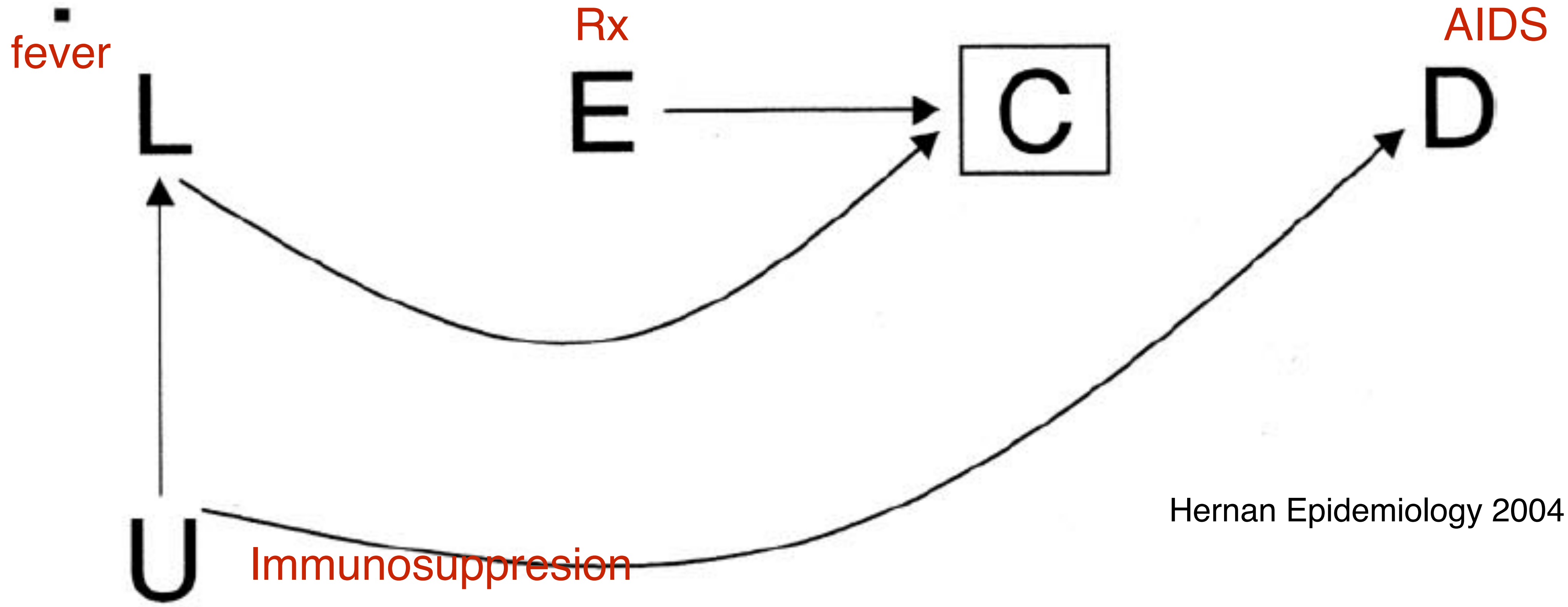
```
summary (lm(bp~income, data=dSample))

Call:
lm(formula = bp ~ income, data = dSample)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0115     0.0275    36.8   <2e-16
income      -0.3623     0.0246   -14.7   <2e-16

Residual standard error: 0.784 on 1353 degrees of freedom
Multiple R-squared:  0.138,  Adjusted R-squared:  0.138
F-statistic:  217 on 1 and 1353 DF,  p-value: <2e-16
```

Selection bias 2^o lost to follow-up

- **Selection bias also possible due to differential loss to follow-up: AKA bias due to informative censoring**

- Cohort: anti retroviral Rx (E), D (AIDS), C (censoring), U (unmeasured immunosupression level of pt which is mediated by L (fever, Sx) also not measured)

- $RR_{ED} = 1.0$ but $RR_{ED|c} \neq 1.0$ due to collider bias conditioning on C, which is a common effect of exposure E and a cause U of the outcome



Hernan Epidemiology 2004;15: 615–625

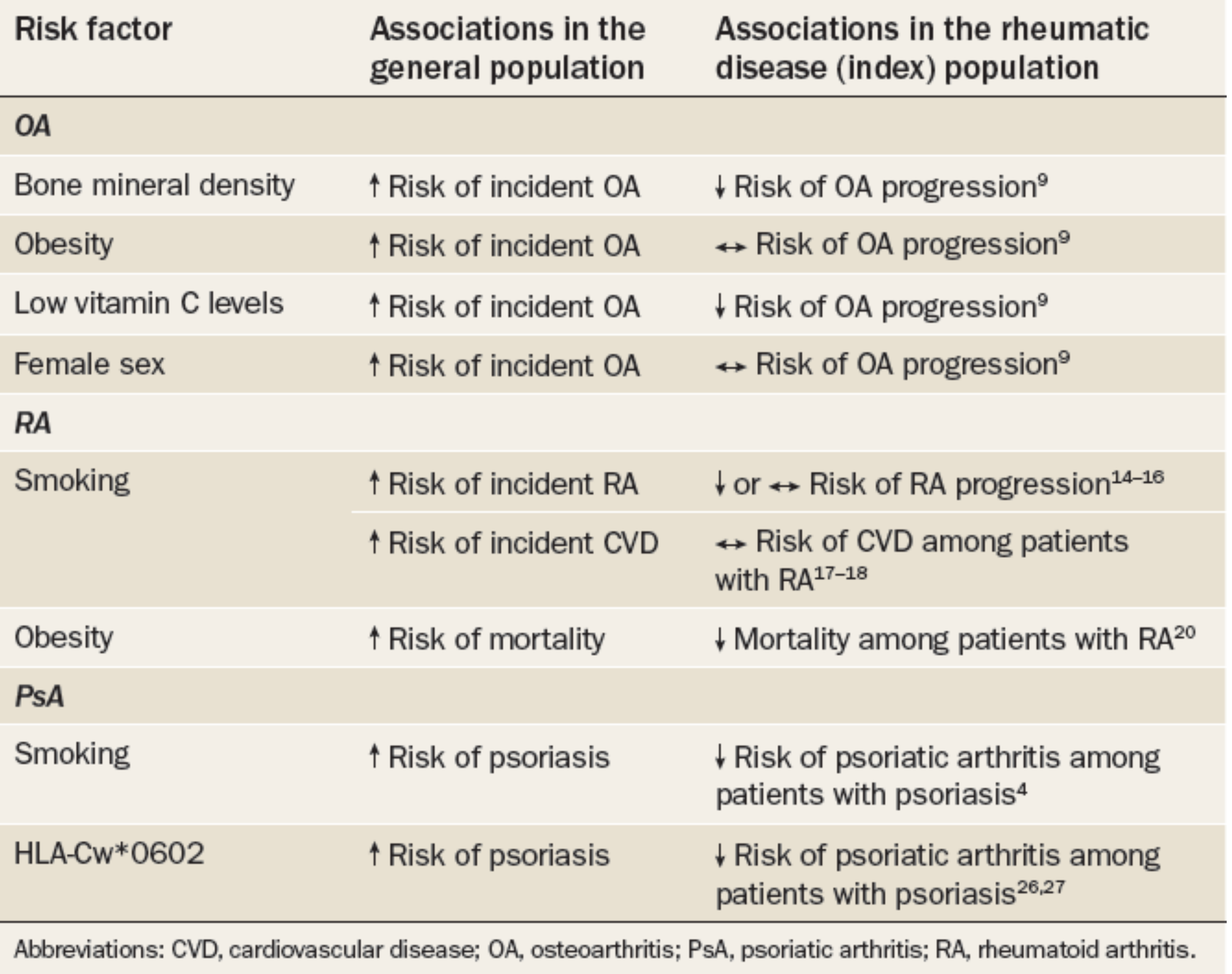
Confounder vs collider

	Confounder	Collider
Main attribute	common cause	common effect
Association	contributes to the association between its effects	does not contribute to the association between its causes
Type of path	open path	blocked path
Effect of conditioning	blocked path	open path
Bias before conditioning?	Yes, confounding bias	No
Bias after conditioning?	No	Yes, colliding bias

Examples

Index event (collider stratification) bias

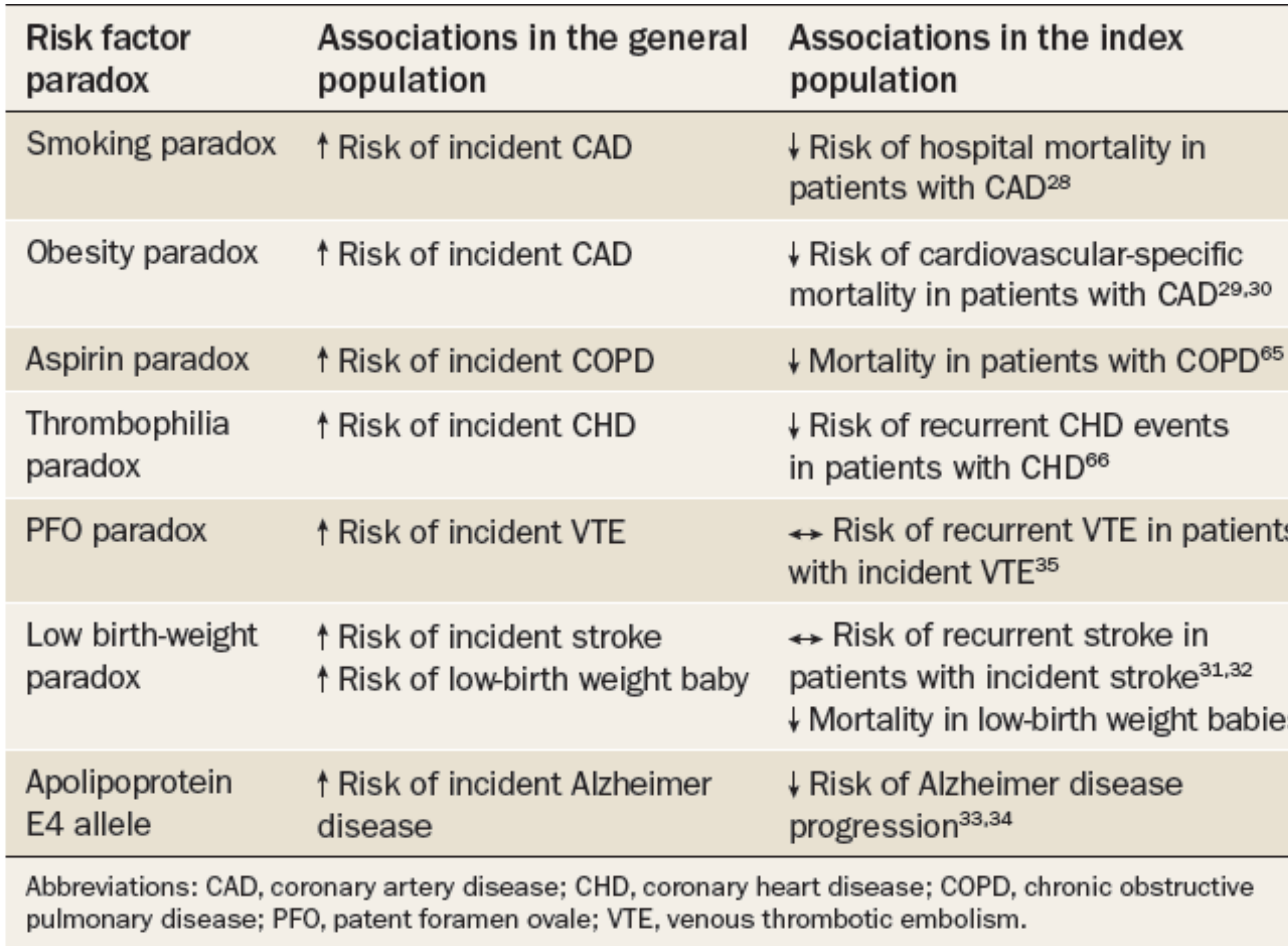
Rheumatic diseases



Risk factor	Associations in the general population	Associations in the rheumatic disease (index) population
OA		
Bone mineral density	↑ Risk of incident OA	↓ Risk of OA progression[9]
Obesity	↑ Risk of incident OA	↔ Risk of OA progression[9]
Low vitamin C levels	↑ Risk of incident OA	↓ Risk of OA progression[9]
Female sex	↑ Risk of incident OA	↔ Risk of OA progression[9]
RA		
Smoking	↑ Risk of incident RA	↓ or ↔ Risk of RA progression[14–16]
	↑ Risk of incident CVD	↔ Risk of CVD among patients with RA[17–18]
Obesity	↑ Risk of mortality	↓ Mortality among patients with RA[20]
PsA		
Smoking	↑ Risk of psoriasis	↓ Risk of psoriatic arthritis among patients with psoriasis[4]
HLA-Cw*0602	↑ Risk of psoriasis	↓ Risk of psoriatic arthritis among patients with psoriasis[26,27]

Abbreviations: CVD, cardiovascular disease; OA, osteoarthritis; PsA, psoriatic arthritis; RA, rheumatoid arthritis.

Cardiac diseases



Risk factor paradox	Associations in the general population	Associations in the index population
Smoking paradox	↑ Risk of incident CAD	↓ Risk of hospital mortality in patients with CAD[28]
Obesity paradox	↑ Risk of incident CAD	↓ Risk of cardiovascular-specific mortality in patients with CAD[29,30]
Aspirin paradox	↑ Risk of incident COPD	↓ Mortality in patients with COPD[65]
Thrombophilia paradox	↑ Risk of incident CHD	↓ Risk of recurrent CHD events in patients with CHD[66]
PFO paradox	↑ Risk of incident VTE	↔ Risk of recurrent VTE in patients with incident VTE[35]
Low birth-weight paradox	↑ Risk of incident stroke ↑ Risk of low-birth weight baby	↔ Risk of recurrent stroke in patients with incident stroke[31,32] ↓ Mortality in low-birth weight babie
Apolipoprotein E4 allele	↑ Risk of incident Alzheimer disease	↓ Risk of Alzheimer disease progression[33,34]

Abbreviations: CAD, coronary artery disease; CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; PFO, patent foramen ovale; VTE, venous thrombotic embolism.

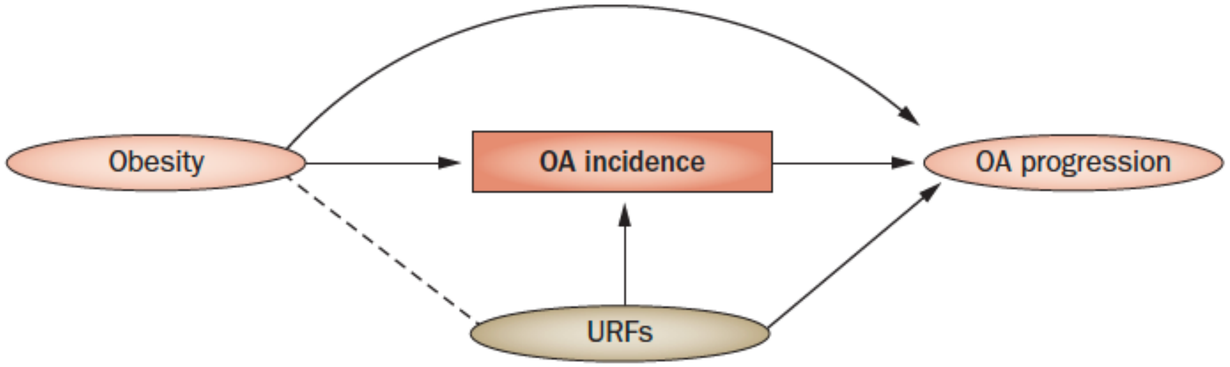
Choi, H. K. et al. Nat. Rev. Rheumatol. 10, 403–412 (2014); published online 1 April 2014; doi:10.1038/nrrheum.2014.36

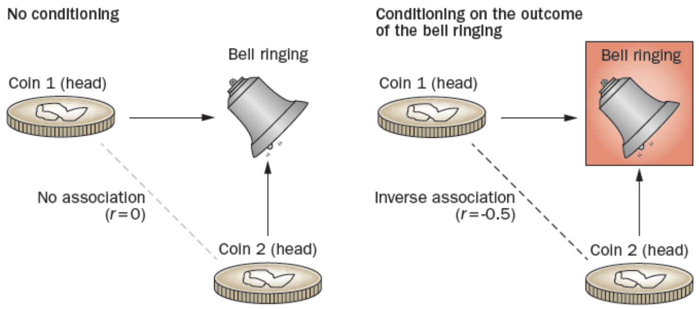
- Risk factor paradox in chronic diseases
- Well established risk factors in general population reverse their impact in these selected (index event) populations ???

The risk factor paradox - what’s going on?

“Systematic review finds little to no evidence that obesity influences the progression of osteoarthritis” Arthritis Rheum 2007 Feb 15;57(1):13-26

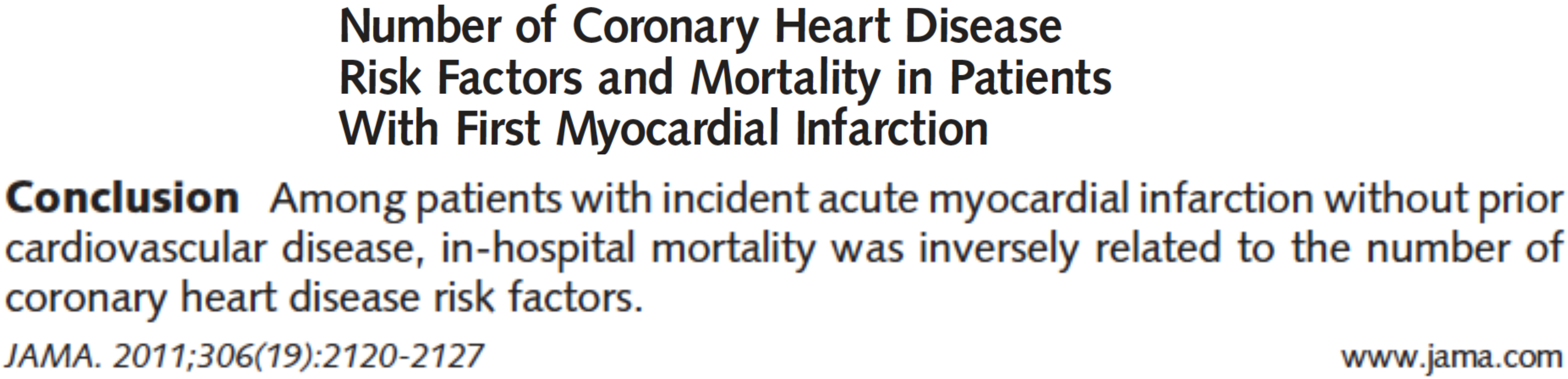
- Editors like the word “paradox” and its mention increases likelihood of publication - novel, controversial findings, easy to invent hypothetical explanations
- Causal versus a non-biological explanation?





Collider stratification bias -> spurious negative association among those risk factors with an index event (explains most “paradoxes”)

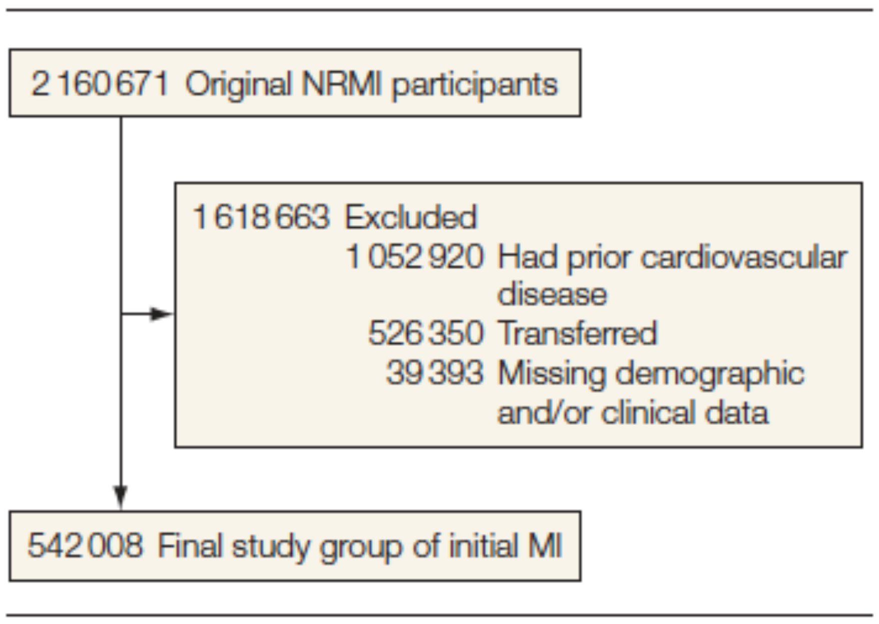
An egregious published example



Number of Coronary Heart Disease Risk Factors and Mortality in Patients With First Myocardial Infarction

Conclusion Among patients with incident acute myocardial infarction without prior cardiovascular disease, in-hospital mortality was inversely related to the number of coronary heart disease risk factors.

JAMA. 2011;306(19):2120-2127 www.jama.com



2 160 671 Original NRMI participants

1 618 663 Excluded
1 052 920 Had prior cardiovascular disease
526 350 Transferred
39 393 Missing demographic and/or clinical data

542 008 Final study group of initial MI

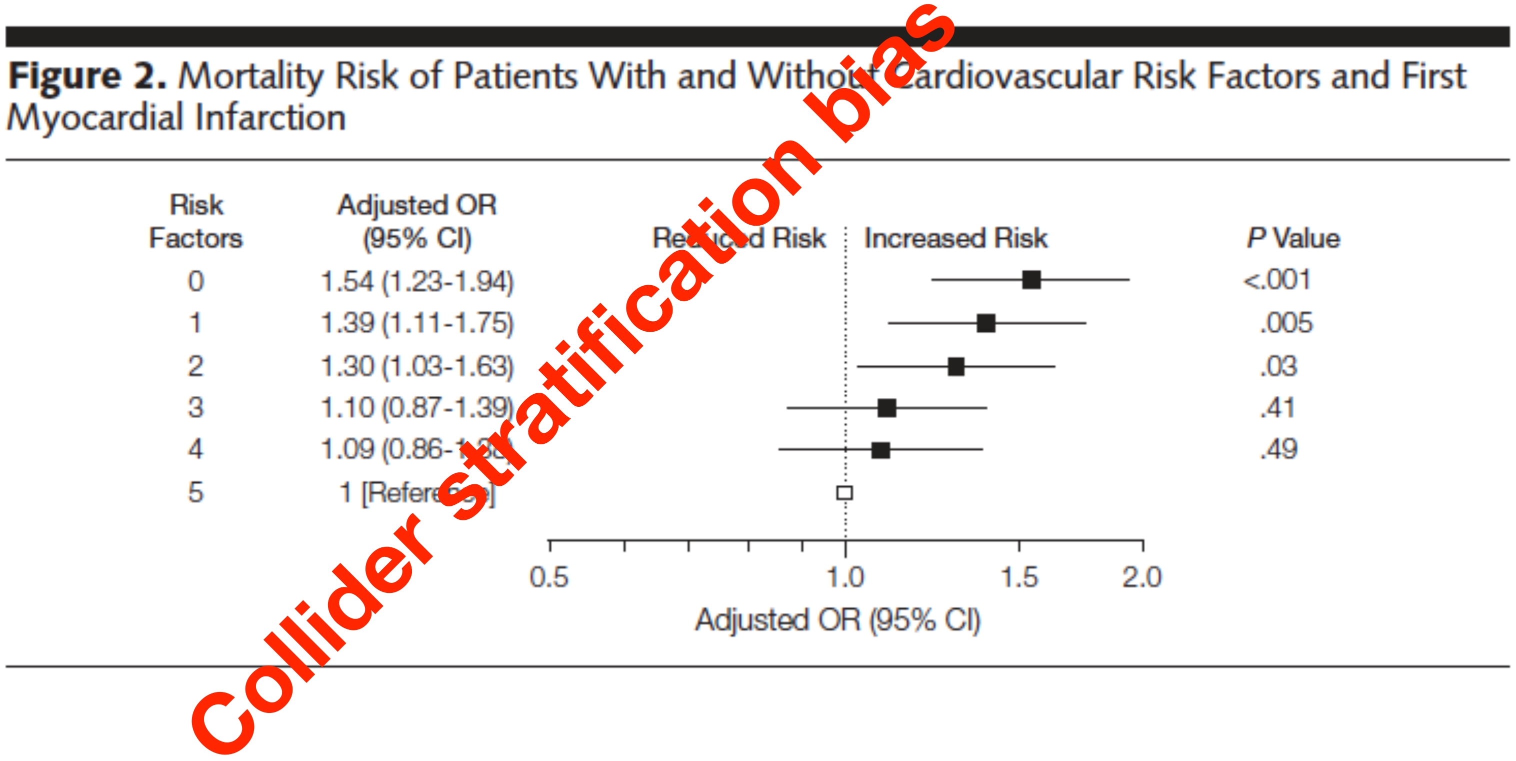


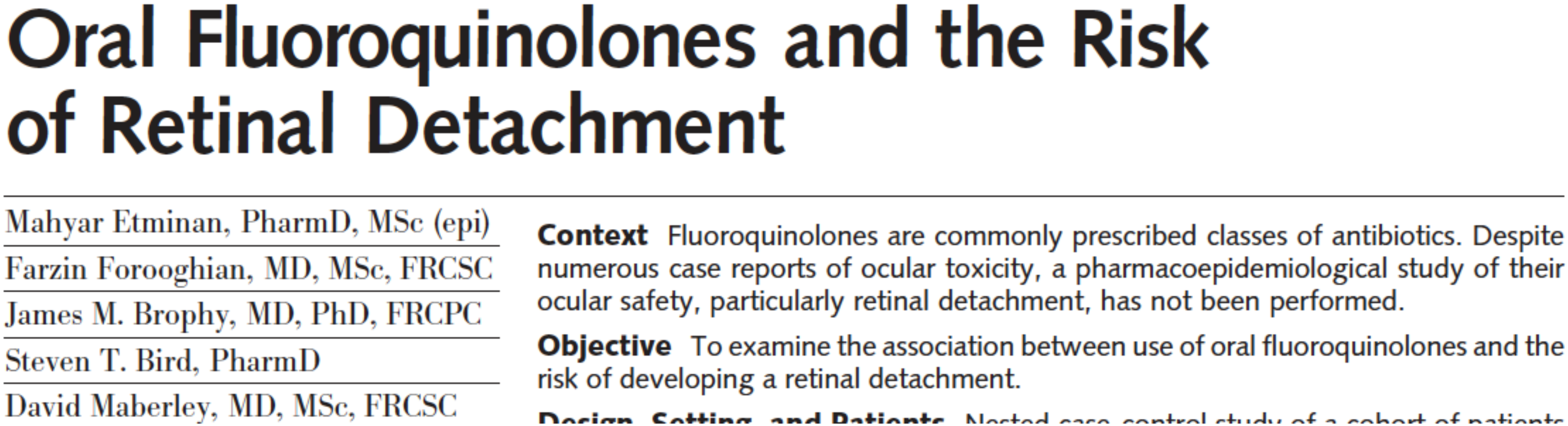
Figure 2. Mortality Risk of Patients With and Without Cardiovascular Risk Factors and First Myocardial Infarction

Risk Factors	Adjusted OR (95% CI)	*P* Value
0	1.54 (1.23-1.94)	<.001
1	1.39 (1.11-1.75)	.005
2	1.30 (1.03-1.63)	.03
3	1.10 (0.87-1.39)	.41
4	1.09 (0.86-1.38)	.49
5	1 [Reference]	

Collider stratification bias

Should we tell patients following a MI that they will do better if they increase their smoking, weight, cholesterol, BP and diabetes?

Life before DAGs



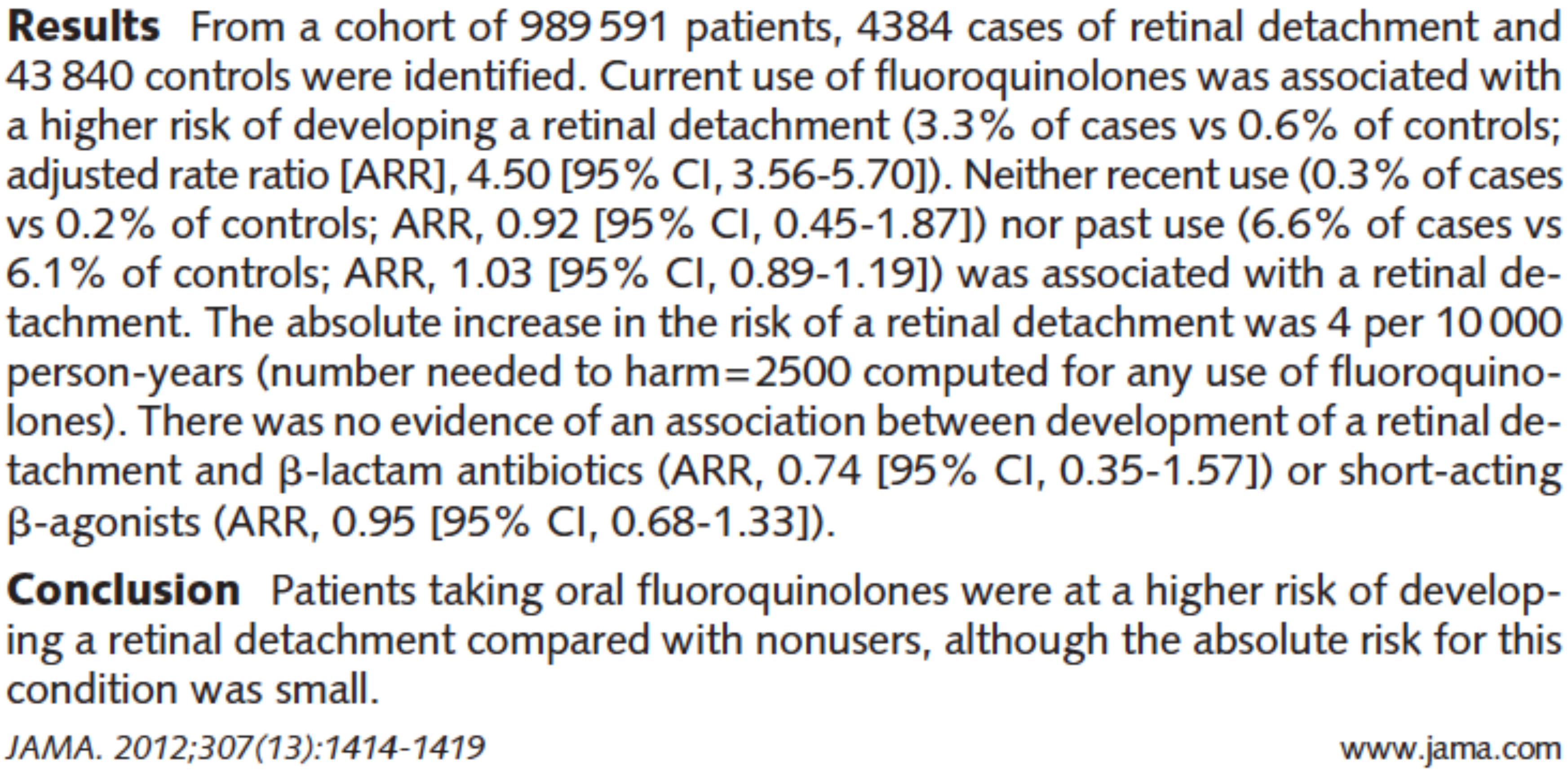
Oral Fluoroquinolones and the Risk of Retinal Detachment

Mahyar Etminan, PharmD, MSc (epi)
Farzin Forooghian, MD, MSc, FRCSC
James M. Brophy, MD, PhD, FRCPC
Steven T. Bird, PharmD
David Maberley, MD, MSc, FRCSC

Context Fluoroquinolones are commonly prescribed classes of antibiotics. Despite numerous case reports of ocular toxicity, a pharmacoepidemiological study of their ocular safety, particularly retinal detachment, has not been performed.

Objective To examine the association between use of oral fluoroquinolones and the risk of developing a retinal detachment.

Design, Setting, and Patients Nested case-control study of a cohort of patients



Results From a cohort of 989 591 patients, 4384 cases of retinal detachment and 43 840 controls were identified. Current use of fluoroquinolones was associated with a higher risk of developing a retinal detachment (3.3% of cases vs 0.6% of controls; adjusted rate ratio [ARR], 4.50 [95% CI, 3.56-5.70]). Neither recent use (0.3% of cases vs 0.2% of controls; ARR, 0.92 [95% CI, 0.45-1.87]) nor past use (6.6% of cases vs 6.1% of controls; ARR, 1.03 [95% CI, 0.89-1.19]) was associated with a retinal detachment. The absolute increase in the risk of a retinal detachment was 4 per 10 000 person-years (number needed to harm=2500 computed for any use of fluoroquinolones). There was no evidence of an association between development of a retinal detachment and β-lactam antibiotics (ARR, 0.74 [95% CI, 0.35-1.57]) or short-acting β-agonists (ARR, 0.95 [95% CI, 0.68-1.33]).

Conclusion Patients taking oral fluoroquinolones were at a higher risk of developing a retinal detachment compared with nonusers, although the absolute risk for this condition was small.

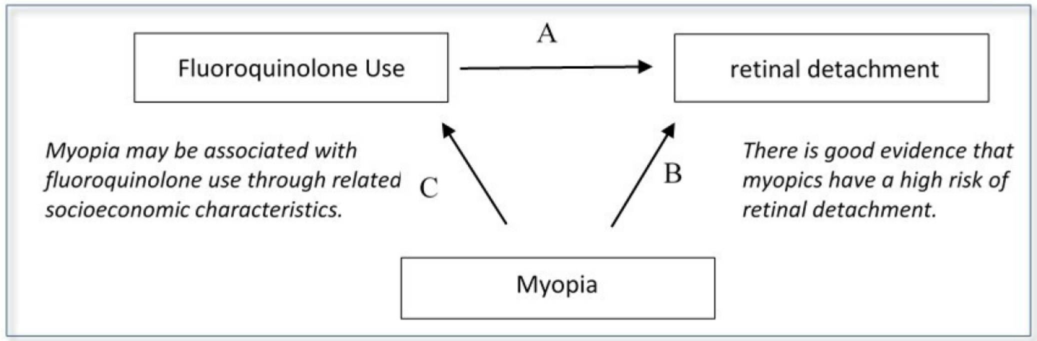
JAMA. 2012;307(13):1414-1419 www.jama.com

adjusted for age, sex, cataracts,
myopia, diabetes, # Rx,
# ophthalmic visits

Why you need a causal model

- Years later, asked to peer review a paper for Ophthalmology
- Authors present a DAG (Figure A) and praised our paper
- But their text actually described a different DAG (Figure B)

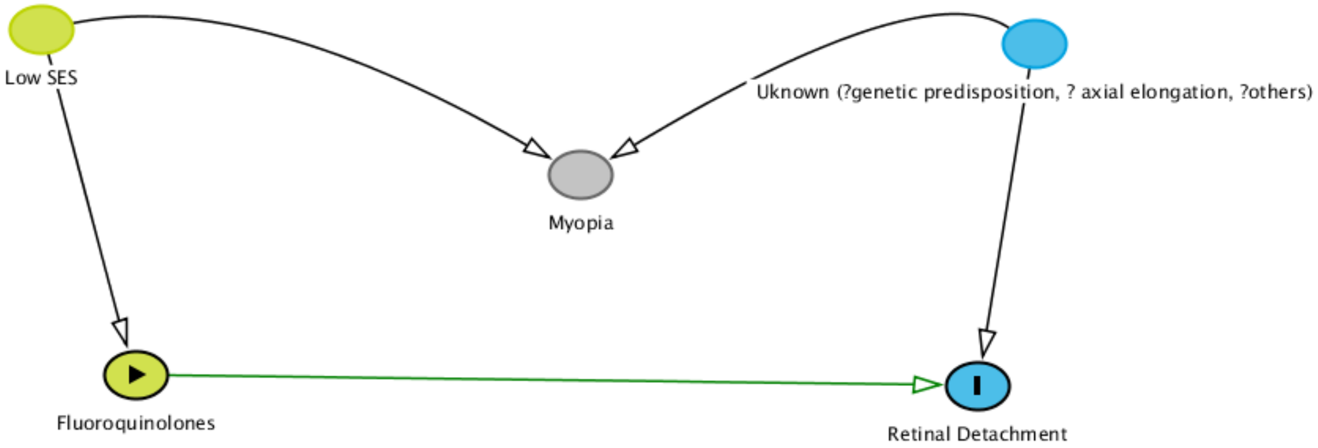
Figure A



Myopia may be associated with fluoroquinolone use through related socioeconomic characteristics.

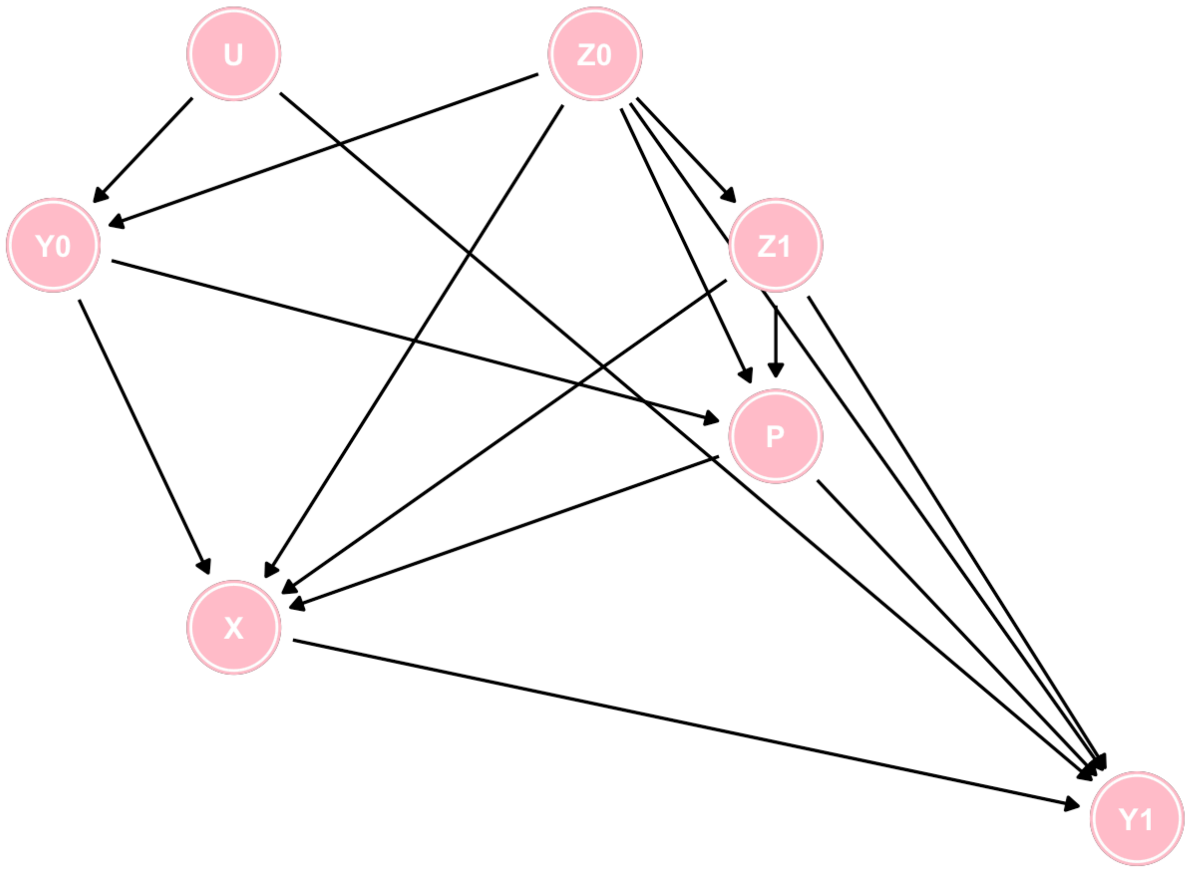
There is good evidence that myopics have a high risk of retinal detachment.

Figure B


Low SES
Uknown (?genetic predisposition, ? axial elongation, ?others)
Myopia
Fluoroquinolones
Retinal Detachment

- Should we have controlled for myopia?
- If their causal model **B** is right, myopia is not a confounder but a collider, stratifying on it, as the authors recommend (and we did) will increase, not decrease bias.
- **So maybe we got it wrong**

A Final More Complex Example - R can help



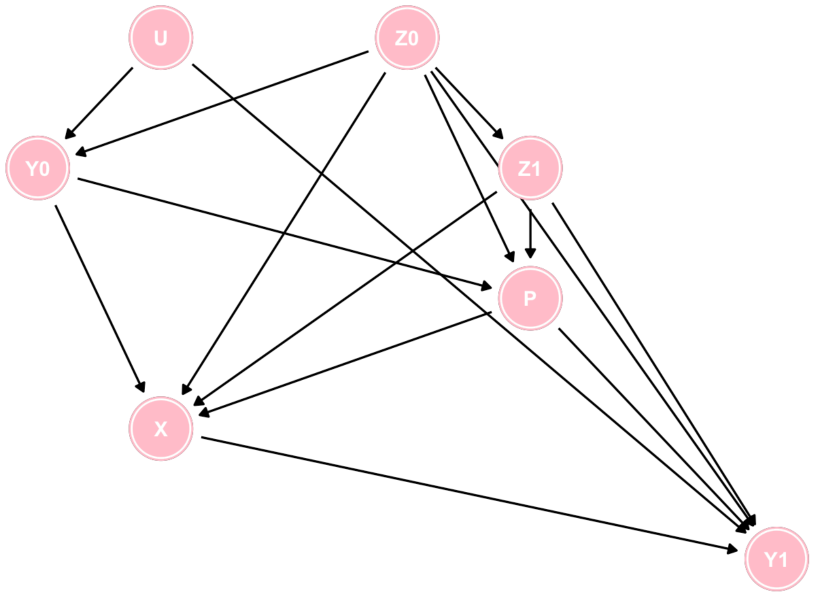
R CODE

```
  dag <- ggdag::dagify(Y1 ~ X + Z1 + Z0 + U + P,
              Y0 ~ Z0 + U,
              X ~ Y0 + Z1 + Z0 + P,
              Z1 ~ Z0,
              P ~ Y0 + Z1 + Z0,
              exposure = "X",
              outcome = "Y1")

dag %>%
  ggdag::tidy_dagitty(layout = "auto", seed = 12345) %>%
  arrange(name) %>%
  ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point() +
  geom_dag_edges() +
  geom_dag_text(parse = TRUE, label = c("P", "U", "X",
expression(Y[0]), expression(Y[1]), expression(Z[0]),
expression(Z[1]))) +
  theme_dag() +
  geom_dag_node(color="pink") +
geom_dag_text(color="white")
```

A Final More Complex Example - R can help

Questions arising from this DAG

1. How many paths are there from X to Y1?
2. How many of those paths are spurious (backdoor) paths?
3. How many of those backdoor paths are open?
4. What is the minimal set of variables to block these spurious pathways?



Questions theoretically answerable by careful attention to DAG but easier with the R dagitty package’s built-in functions

```
g <- dagitty::paths(dag, "X", "Y1")
paste0("There are ", length(g$paths), " pathways from X to Y1 and all are backdoor except for 1")
paste0("Of these backdoor pathways ", sum(g$open=="TRUE"), " are open")
paste0("The minimum adjustment sets are “, adjustmentSets(dag, "X", "Y1", type = "minimal"))
```

```
## [1] "There are 43 pathways from X to Y1 and all are backdoor except for 1”

## [1] "Of these backdoor pathways 25 are open”

## [1] "The minimum adjustment sets are "
## { P, U, Z0, Z1 }
## { P, Y0, Z0, Z1 }
```

My Bottom Line

DAGs can be super useful
on the road to causal inference

References

- Lots of excellent references - basically anything by Judea Pearl or Miguel Hernan
    - Pearl, J, M Glymour, and NP Jewell. 2016. Causal Inference in Statistics. John Wiley. Book.
    - Miguel A. Hernán, James M. Robins Causal Inference What if https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- Some of this material can be found in (Mostly Clinical) Epidemiology with R (https://bookdown.org/jbrophy115/bookdown-clinepi/)