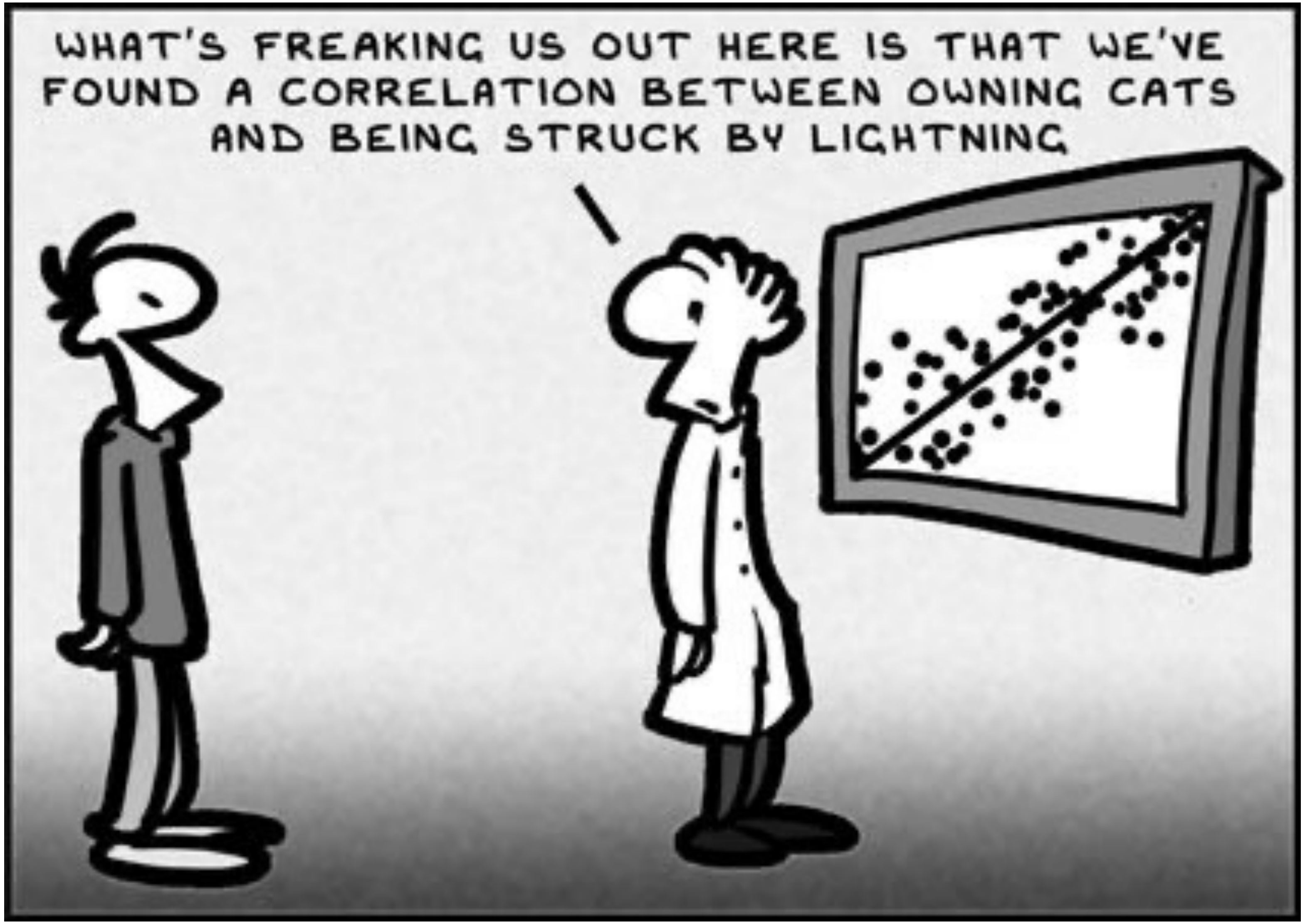

WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

# BIAS & CONFOUNDING

# Not all associations are causal or independent

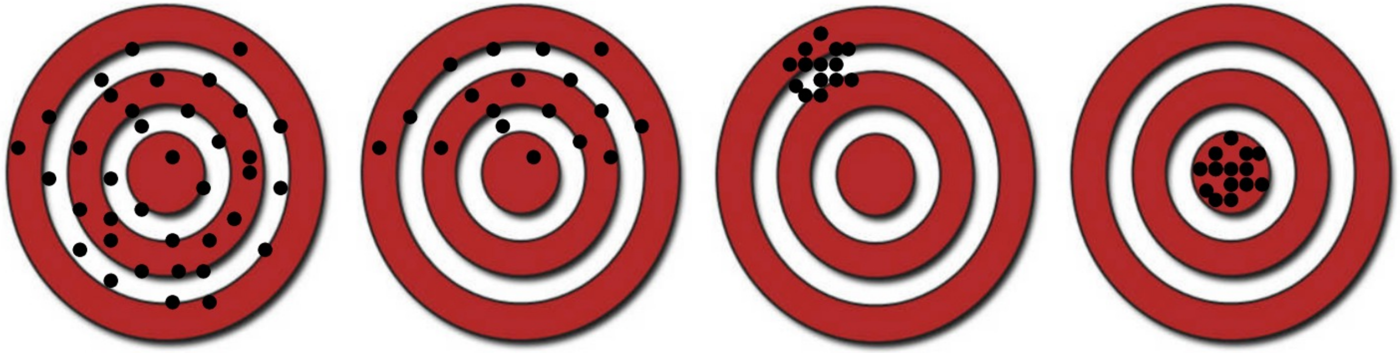
- Random error (chance)
- Confounding
- Bias
  - *Selection bias*
  - *Information bias*

**Systematic error**

- Real association

"Be as sceptical as possible as to reason for an association"

# Precision and (Internal) Validity



| | | | | |
|---|---|---|---|---|
| **Random Error** | large | large | small | small |
| **Systematic Error** | small | large | large | small |
| **Terminology** | imprecise,<br>valid | imprecise,<br>invalid | precise,<br>invalid | **precise,<br>valid<br>→ accurate** |

# Random error (chance)

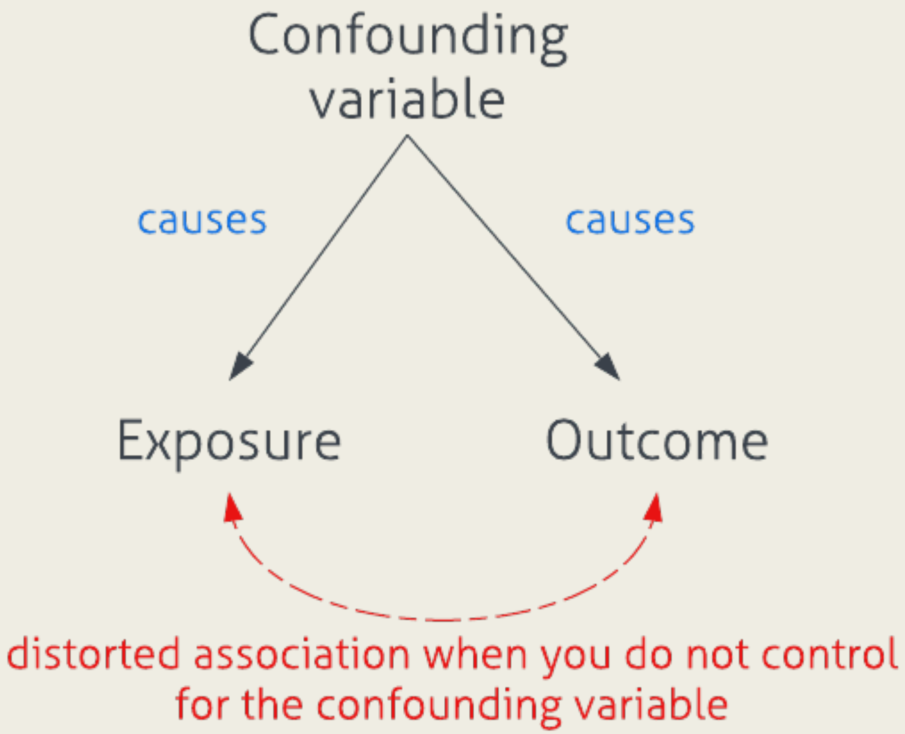
- Collect information from a sample to estimate a statistic (measure of occurrence) or association (measure of effect) for the whole population
  - *Random samples from the same population will give different results – due to chance*
  - *Results may not be same as true population result*

# Systematic error

- Confounding
- Selection bias
- Information bias

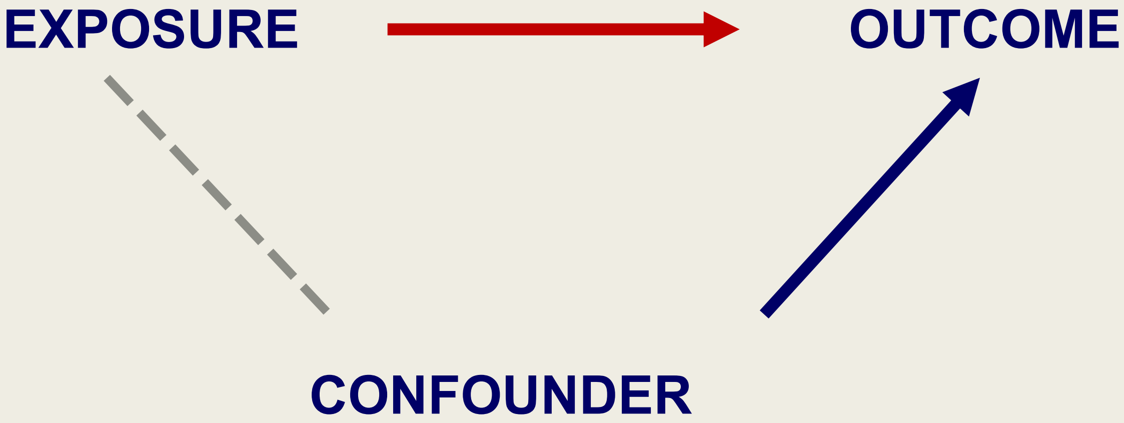
# What is confounding?

- Confounding occurs when an association (or lack of) is **distorted** by another factor
- Is there an alternative explanation for the observed exposure-disease association?
- In observational studies, people who are exposed to a particular risk factor may also have other characteristics in common that influence their risk of the disease


Confounding variable
causes
causes
Exposure
Outcome
distorted association when you do not control for the confounding variable

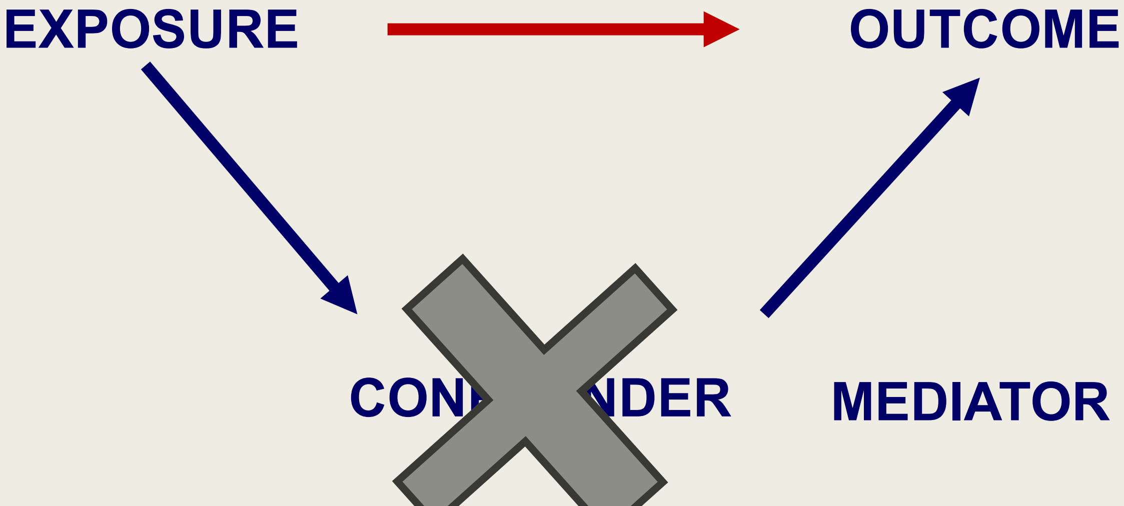
# Classical definition of a confounder

- Must be associated with the exposure
- Must be a ‘risk factor’ (cause) for the disease
- Must NOT be on the causal pathway between exposure and disease



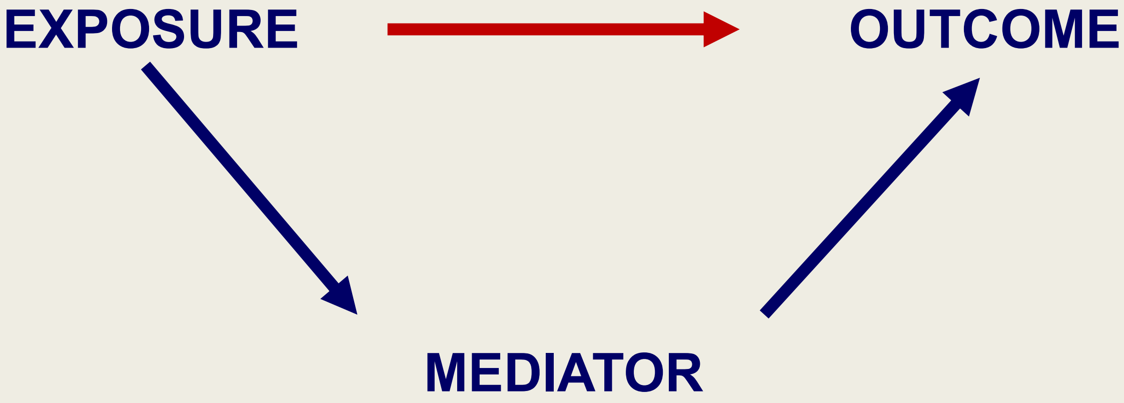
# Classical definition of a confounder

- Must be associated with the exposure
- Must be a 'risk factor' (cause) for the disease
- Must NOT be on the causal pathway between exposure and disease



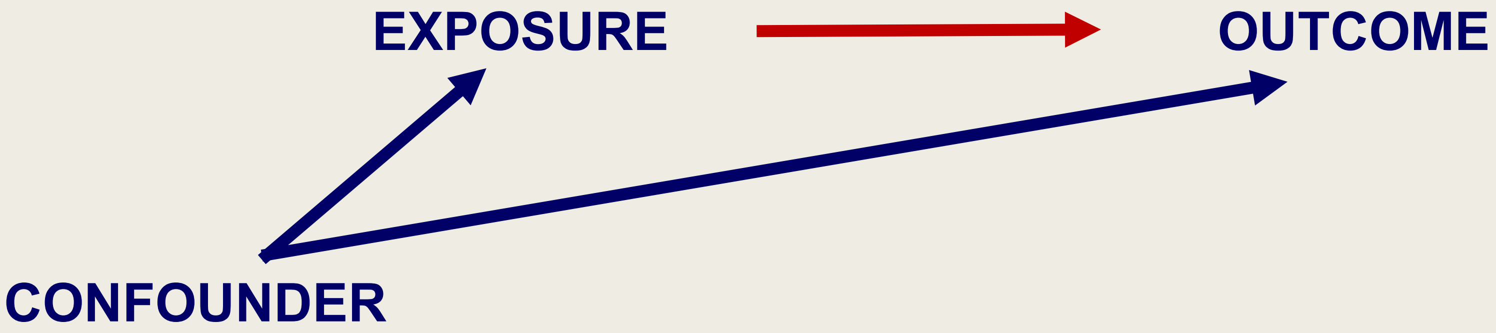
# Total and direct effects

- We are usually interested in the total effect of the exposure
- If you adjust for the effect of the mediator then you are no longer estimating total effect but the direct effect of exposure
- Mediation analysis very complex and difficult – be warned

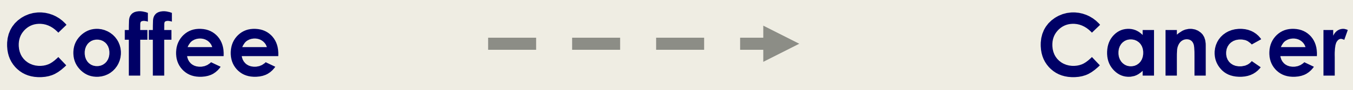


# Alternative definition of a confounder

- Common cause of the exposure and outcome



# Does drinking coffee cause cancer?



Coffee

Cancer

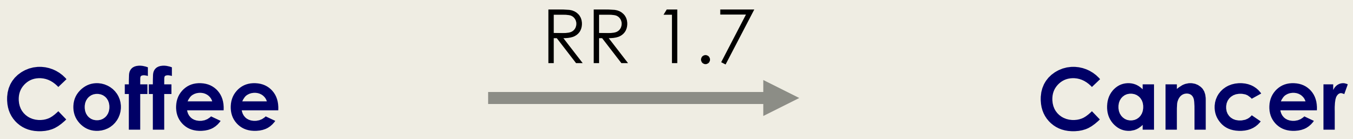
| | Cancer | No cancer | Row total |
|---|---|---|---|
| Coffee drinker | 105 | 11395 | 11500 |
| Non-coffee drinker | 45 | 8455 | 8500 |

Risk of cancer in coffee drinkers = 105/11500

Risk of cancer in non-coffee drinkers = 45/8500

Risk ratio = 1.72

# Does drinking coffee cause cancer?

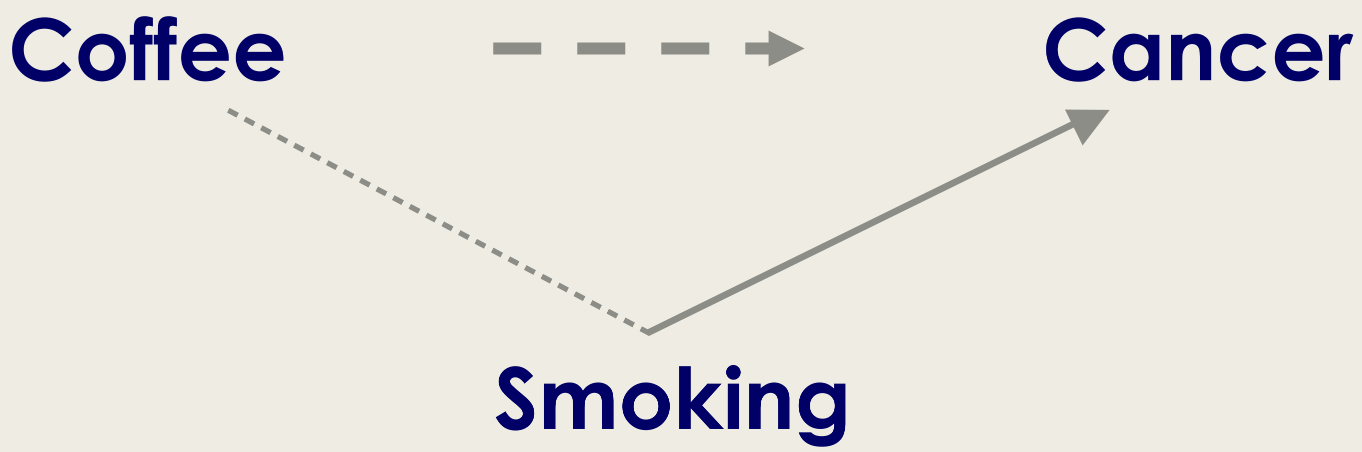


**Coffee**

RR 1.7

**Cancer**

# Does drinking coffee cause cancer?



Could this observed association be confounded by smoking?

# How to explore the confounding effect?

Stratified analysis by smoking status

Risk ratio = 1.0

Non-smokers (N=15,000)

| | Cancer | No cancer | Row total |
|---|---|---|---|
| Coffee drinker | 25 | 7475 | 7500 |
| Non-coffee drinker | 25 | 7475 | 7500 |

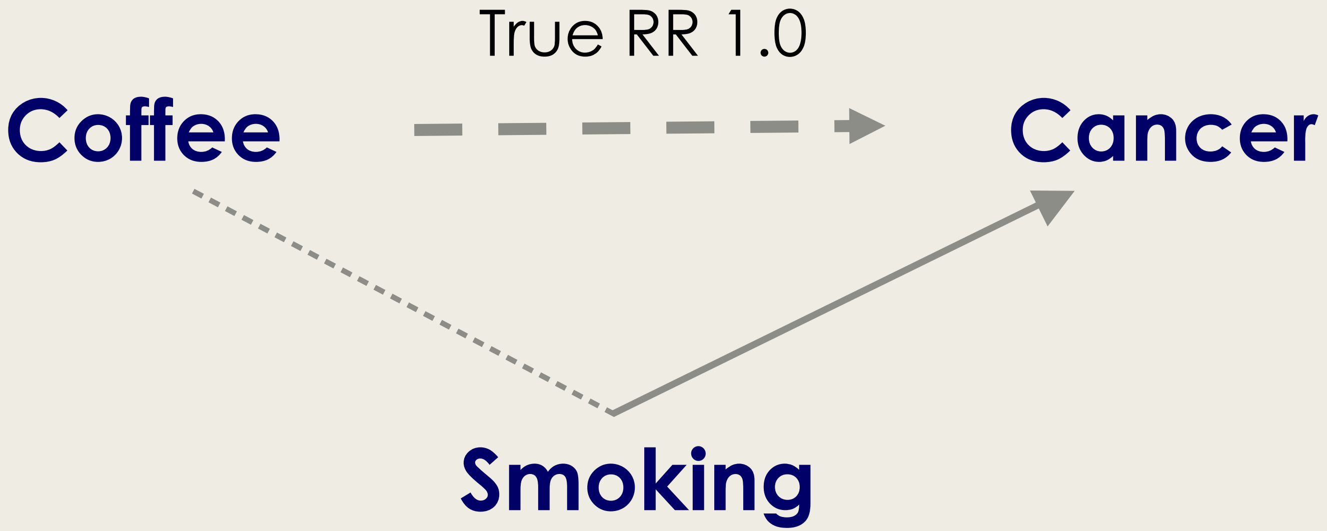
Risk ratio = 1.0

Smokers (N=5,000)

| | Cancer | No cancer | Row total |
|---|---|---|---|
| Coffee drinker | 80 | 3920 | 4000 |
| Non-coffee drinker | 20 | 980 | 1000 |

Proportion who smoke in coffee-drinkers: 35% vs. proportion who smoke in non-coffee drinkers: 12%  (~ 3x higher)

Proportion with cancer in those who smoke: 2% vs. proportion with cancer of those who do not smoke: 0.3% (~ 6x higher)


True RR 1.0

Coffee

Cancer

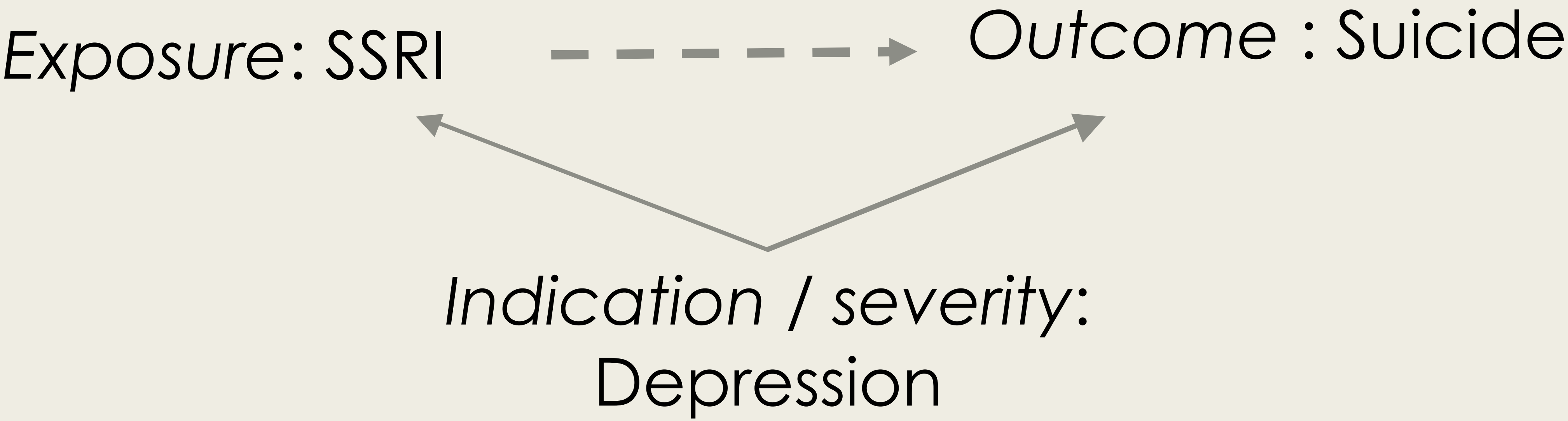
Smoking

# Interpretation

- Smoking status acting as a confounder in this study of cancer risk from coffee drinking
- Without controlling for smoking, it appears that coffee drinkers are ~70% more likely to develop cancer (RR 1.7)
- After stratification by smoking status, the analysis shows no increased risk of cancer for coffee drinkers (removed the effect of smoking)
- The **apparent** association resulted from
  - *coffee drinkers are more likely to smoke*
  - *smoking is a strong risk factor for cancer*

# Confounding by indication

- Good clinical practice on part of the medical practitioner
  - *Indication for treatment or severity of disease commonly predict the initiation of treatments*
  - *Indication for treatment and disease severity are associated with the outcome of interest*

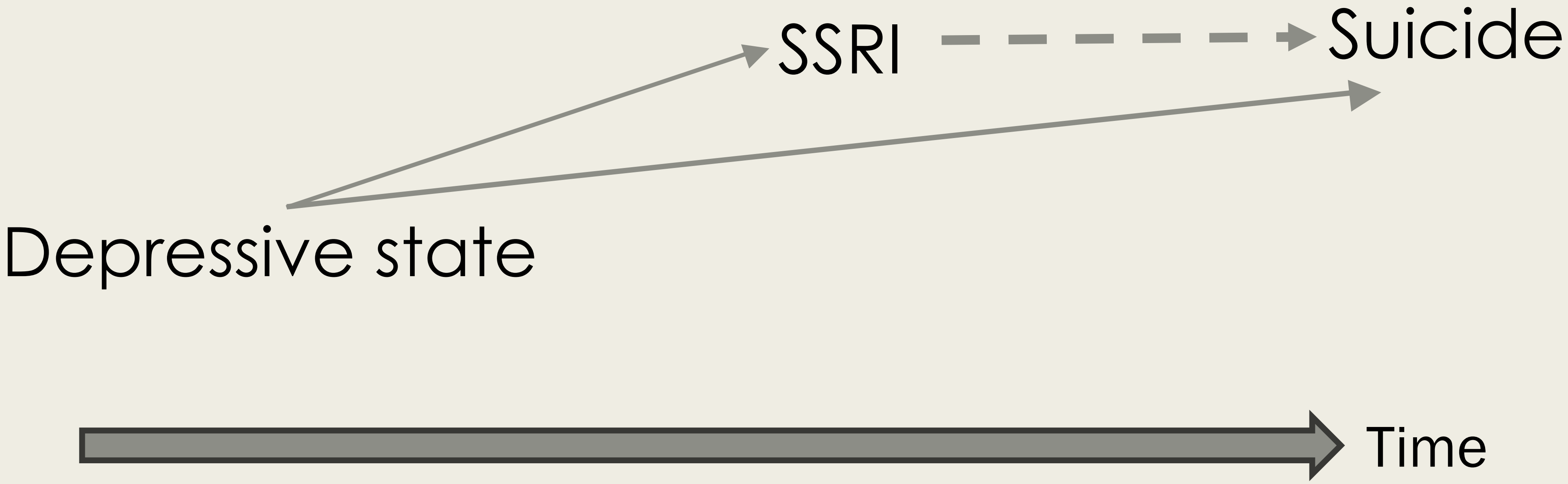


*Exposure*: SSRI

*Outcome* : Suicide

*Indication / severity:*
Depression

# Confounding by indication



# Addressing confounding

- Design Stage
- Analysis Stage

- NEED TO THINK AHEAD !

# Design stage

Two main options:

1. **Randomisation**
   - Participants are randomly allocated to exposure or control group
   - Ideal method, since it controls for known and unknown confounders
   - Ensures the two groups are similar in all respects except for exposure
   - Only be used in intervention (experimental studies)

2. **Restriction**
   - Restrict study population so all participants are similar with respect to confounding variable
   - Results can only be generalised to restricted population
     - if age and sex are known to be strong confounders then restrict study to one age group and one sex

# Analysis stage

Three main approaches:

1. **Stratification**
   - Compare exposed and unexposed within stratum of confounding variable
   - Pool all the stratum results to obtain an overall result
2. **Standardisation**
   - Particular type of stratification – compare disease rates in cohort with general population
3. **Statistical modelling- use a DAG ;-)**
   - Fit a multivariable model to data to account for multiple confounders simultaneously
   - *For each of these we must collect the data on potential confounders at the design stage!!*

# Be warned

- Extent to which confounding can be controlled depends on how accurately we have measured the confounder
  - *e.g. if age bands are too broad then we may not fully adjust for age*

- If we have error in the measurement of the confounder we will not fully adjust

- Both these lead to …. **“residual confounding”**

# To summarise

- Confounding causes a distortion in the association we are examining
- Confounder must be associated with exposure and outcome and not on the causal pathway (common cause of exposure and outcome)
- Can address confounding at design or analysis stage
- Several different approaches, stratification is a very common one
- Produces an adjusted RR
- Compare adjusted RR with crude RR to determine whether factor was a confounder

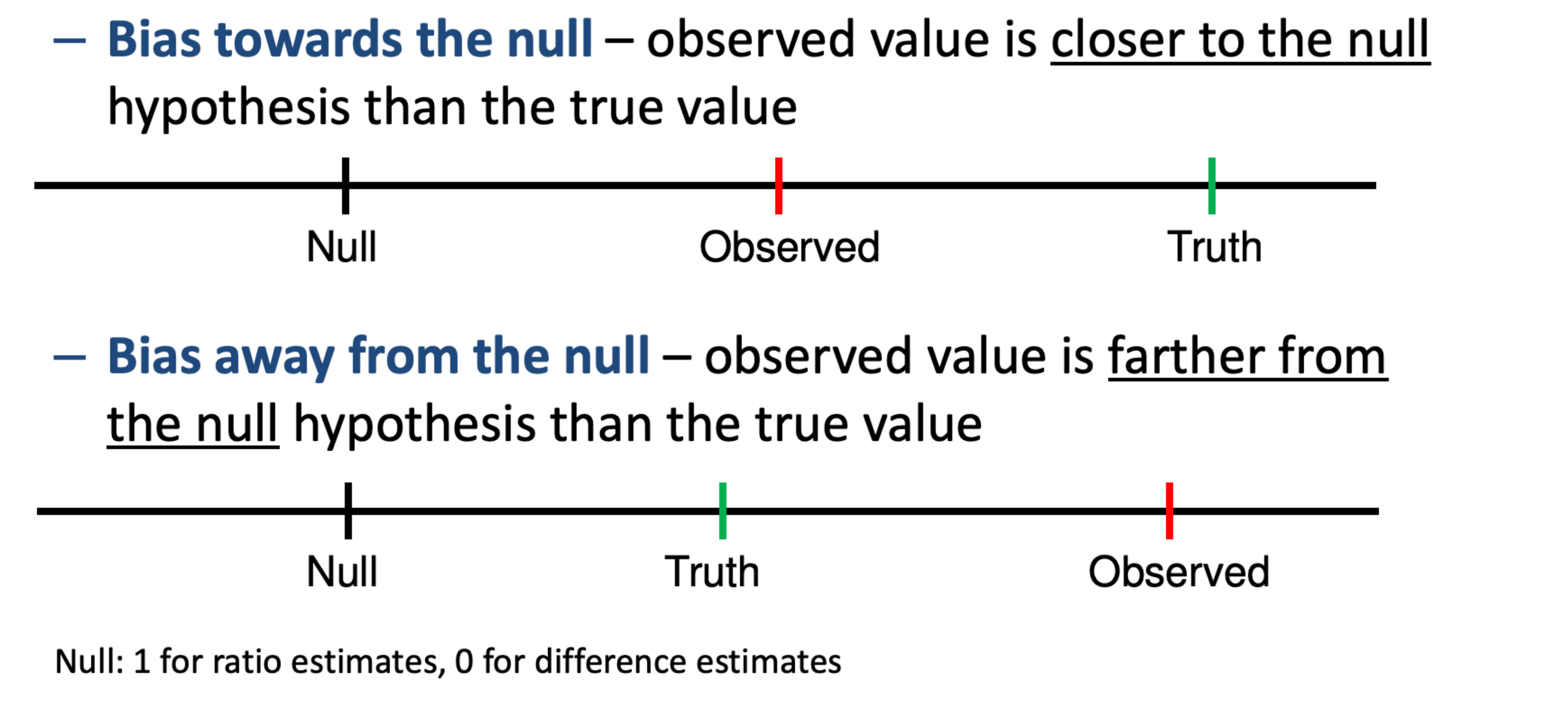
# Bias

“**Systematic deviation** of results or inferences from **truth. …** An error in the **conception** and **design** of a study **-** or in the **collection**, **analysis**, **interpretation**, **reporting**, **publication**, **or review of data -** leading to results or conclusions that are systematically **(as opposed to randomly)** different from truth.”

**Dictionary of Epidemiology**, **2008**

# Bias

- Can affect all types of studies but observational studies especially those based on routinely collected data are particularly vulnerable
- Bias has direction



- **Bias towards the null** – observed value is closer to the null hypothesis than the true value

- **Bias away from the null** – observed value is farther from the null hypothesis than the true value

Null: 1 for ratio estimates, 0 for difference estimates

# Types of bias

- Selection bias
  - *Results from the selection and retention of the study population*
- Information bias
  - *Results from poor measurement of study variables – exposure, outcome, confounding variables*
  - *Measurement error or misclassification*
  - *Differential or non-differential*

# Selection bias

- Type of systematic error which results from
  - *procedures used to select study participants*
    - E.g., TB prevalence survey: convenience sampling at hospital so that only those at highest risk of having TB participate → higher (biased) estimate of TB burden for city
  - *factors that influence participation/retention in the study*
    - E.g., TB prevalence survey: waived written consent (thumbprint) as undocumented migrant workers would not participate
    - LTFU important in longitudinal studies and can also bias RCTs!

- Bias of the estimated effect of an exposure on outcome due to conditioning on a common effect of the exposure and the outcome (**collider bias**)

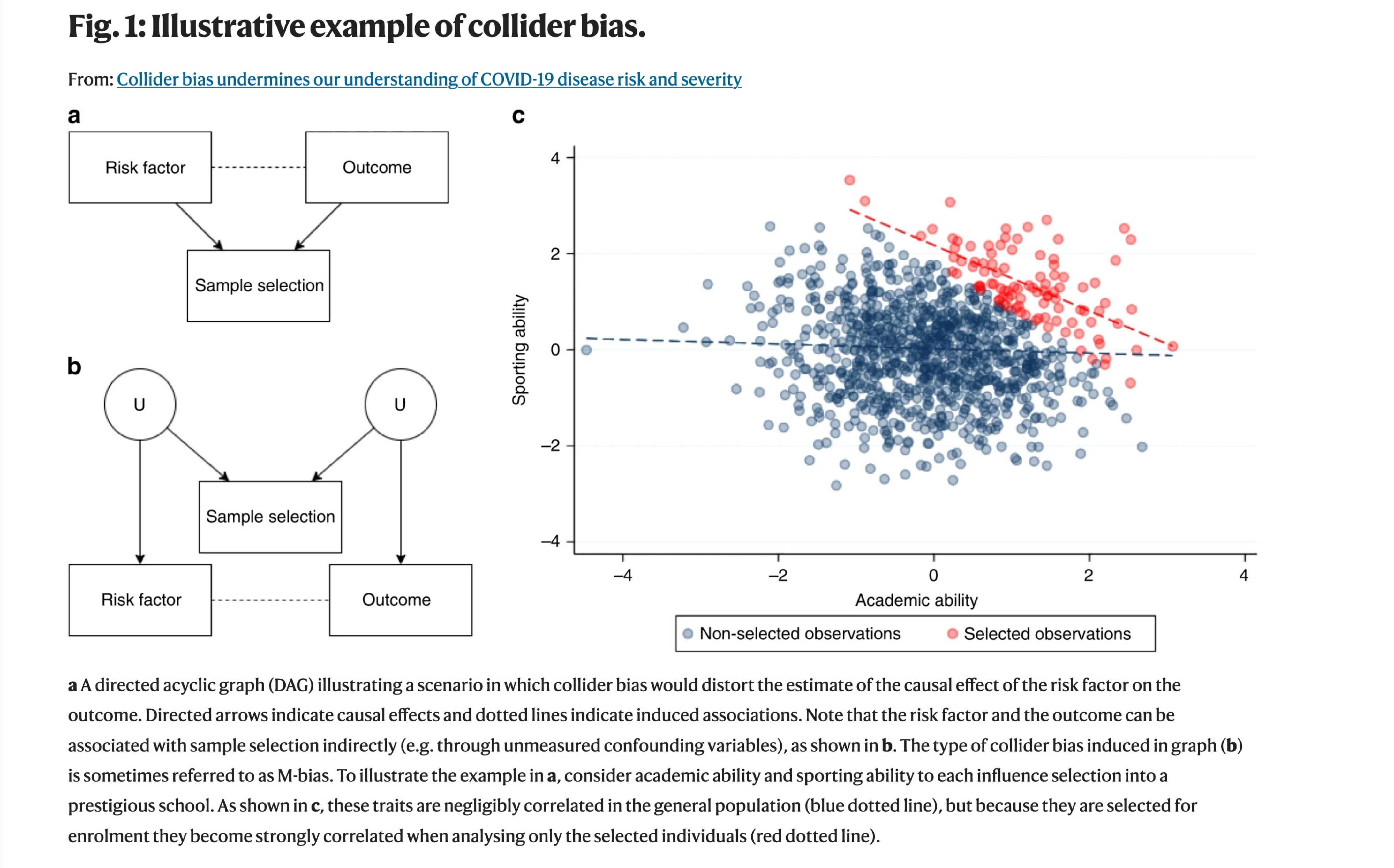
# Selection bias

- common consequence of selection bias
  - association between exposure and outcome among those selected for analysis differs from the association among those eligible

Hernán et al. A Structural Approach to Selection Bias. Epidemiology 15(5):p 615-625, September 2004. | DOI: 10.1097/01.ede.0000135174.63482.43

# Collider bias

- Goes by many names…
- Selection bias, sampling bias, ascertainment bias, Berkson’s paradox
- Colliders become an issue when they are conditioned upon in analysis, as this can distort the association between the two variables influencing the collider
- It is possible to distort the association between two variables that do not directly influence the collider
  - If the factors that influence sample selection themselves influence the variables of interest, the relationship between these variables of interest can become distorted.
  - This is sometimes referred to as M-bias due to the shape of the DAG



# Fig. 1: Illustrative example of collider bias.

From: Collider bias undermines our understanding of COVID-19 disease risk and severity

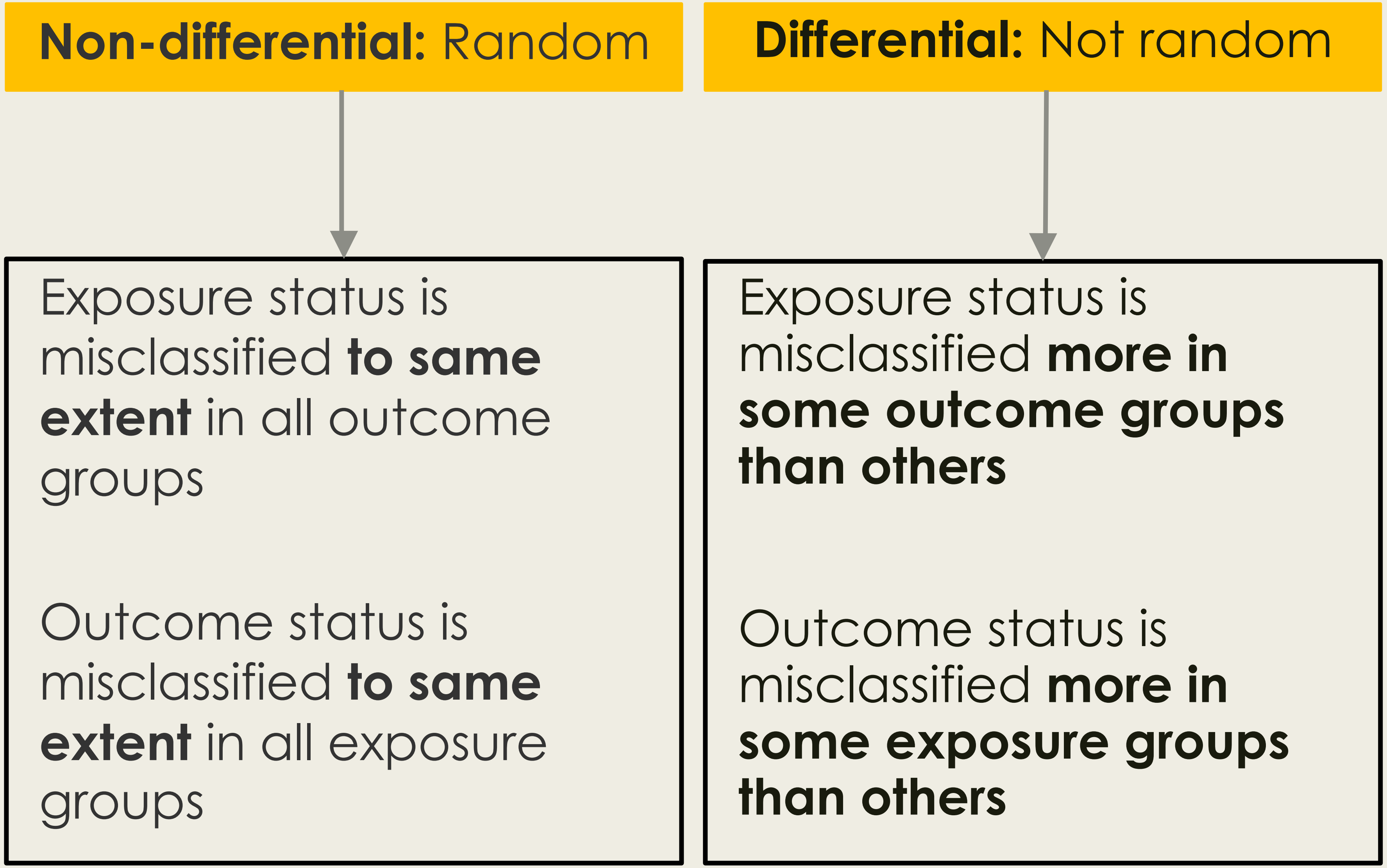
**a** A directed acyclic graph (DAG) illustrating a scenario in which collider bias would distort the estimate of the causal effect of the risk factor on the outcome. Directed arrows indicate causal effects and dotted lines indicate induced associations. Note that the risk factor and the outcome can be associated with sample selection indirectly (e.g. through unmeasured confounding variables), as shown in **b**. The type of collider bias induced in graph (**b**) is sometimes referred to as M-bias. To illustrate the example in **a**, consider academic ability and sporting ability to each influence selection into a prestigious school. As shown in **c**, these traits are negligibly correlated in the general population (blue dotted line), but because they are selected for enrolment they become strongly correlated when analysing only the selected individuals (red dotted line).

Griffith, G.J., Morris, T.T., Tudball, M.J. *et al.* Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* **11**, 5749 (2020)

# Information bias

- Occurs due to poor measurement (classification) of study variables (exposure, outcome, confounders )
- Particularly problematic when using secondary data
  - *Primary data: collected for research purposes*
  - *Secondary data: collected for clinical, administrative, or payment purposes*

- Distinguish two basic types of information bias
  - *Non-differential - Misclassification between groups is approximately equal*
  - *Differential - Amount of misclassification differs between groups*



**Non-differential:** Random

Exposure status is misclassified **to same extent** in all outcome groups

Outcome status is misclassified **to same extent** in all exposure groups

**Differential:** Not random

Exposure status is misclassified **more in some outcome groups than others**

Outcome status is misclassified **more in some exposure groups than others**

# SAMPLING

# Sampling

- **Objective**: To make inferences about a population from data contained in a sample

- To investigate the properties of a population, we collect data on a selection of members of the population
  - *Logistics and cost*

# What does it mean to survey?

- An investigation about the characteristics of a given population
- Collect data from a sample of the given population
- Systematic use **of statistical methodology** to estimate their characteristics

- Surveys are used in many fields such as:
  - *Health*
  - *Politics and social concerns*
  - *Business*
  - *Human interest*

Survey a random sample of the population in Karachi to estimate burden of (infectious) pulmonary TB

# Effective sampling

1. Define target population
2. Define source population
3. Identify a sampling frame (list)
4. Select individuals from sampling frame
   - *Each individual in the sampling frame = sampling unit*
   - Sampling units form the study population

- **Internal validity:** how well do the study findings relate to the source population?
- **External validity:** how well can the study findings can be extrapolated to the target population?

# Types of sampling

- Probability and non-probability of sampling

- Random selection – employs use of random process
  - *Simple, systematic or stratified*
  - *Equal or unequal probability sampling*

- Non-probabilistic methods
  - *Convenience sampling / purposive sampling*

# Sources of Error in Surveys

- There are two main types of error in surveys:
  - ***sampling error***
  - ***non-sampling error***

- ***Sampling error*** is the variation/error in the result associated with examining a sample (and not the whole population)
  - *This can be quantified but only for* ***random samples***
    - With true probability samples sampling error is reduced by having larger samples
    - In non-probability sampling, the degree to which the sample differs from the population is unknown.

# Non-Sampling Error (systematic error)

- Error due to the result of factors other than random error associated with taking a sample, i.e. due to a flaw at some stage in the survey process from design to implementation

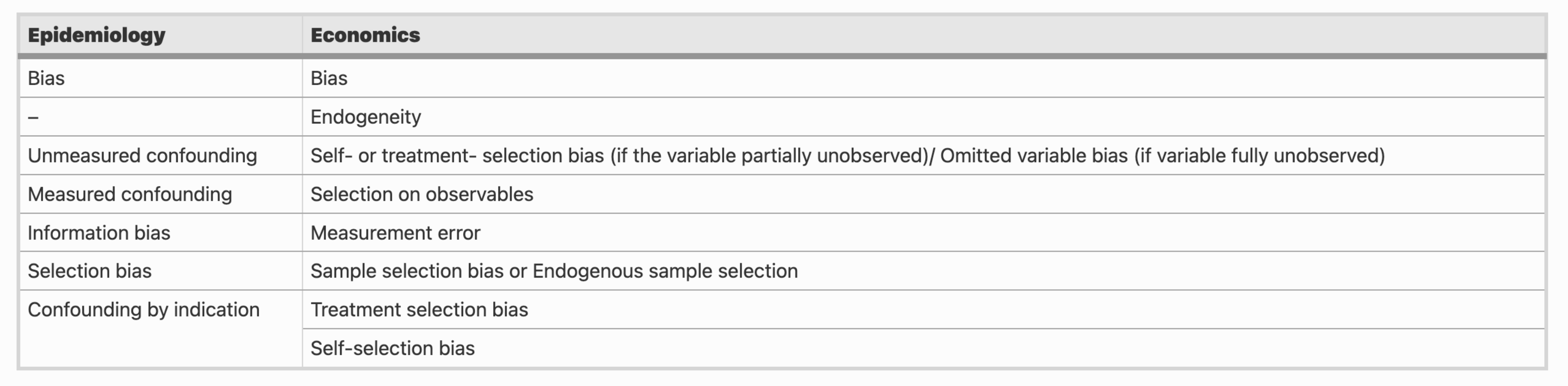
- Try to avoid at design stage
  - *Karachi TB prevalence survey*
    - Asked for waiver of written consent → illiteracy, undocumented migrants and security concerns
    - Avoid potentially missing a section of the population in Karachi who had the highest risk of having undiagnosed pulmonary TB

# Sources of non-sampling error

- **Non-Response**, i.e. failure to obtain a measurement on one or more study variables.
  - **non-participation** in the survey (i.e. no data recorded for that person/empty rows in the data) and
  - **missing answers**, so there is incomplete data (incomplete columns in the data)
- **Selection bias, i.e. an "unrepresentative" sample being used**
  - Use of random sampling makes this much less likely to occur
- **Interviewer bias**
  - This arises when different responses would be given to different interviewers asking the same question. Control this by interviewer training.
- **Inadequate sampling frame ("coverage error"):**
  - If the sampling frame has left out a key group from the target population

# Warning

- Be aware of terminology
- Clarify! Don’t worry about being annoying

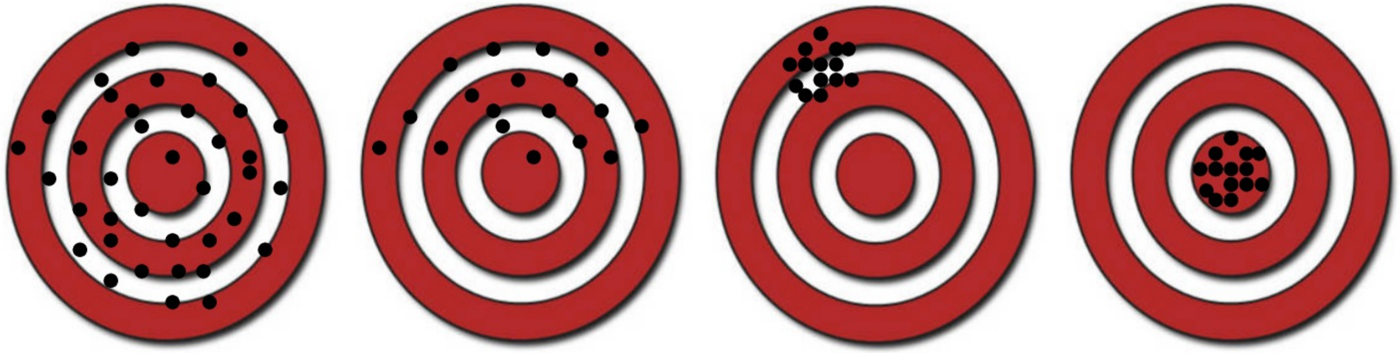


| Epidemiology | Economics |
|---|---|
| Bias | Bias |
| – | Endogeneity |
| Unmeasured confounding | Self- or treatment- selection bias (if the variable partially unobserved)/ Omitted variable bias (if variable fully unobserved) |
| Measured confounding | Selection on observables |
| Information bias | Measurement error |
| Selection bias | Sample selection bias or Endogenous sample selection |
| Confounding by indication | Treatment selection bias |
|  | Self-selection bias |

Varga *et al.* Bias? Clarifying the language barrier between epidemiologists and economists. *Health Serv Outcomes Res Method* (2022). https://doi.org/10.1007/s10742-022-00291

# INTERNAL VALIDITY

# Precision and (Internal) Validity

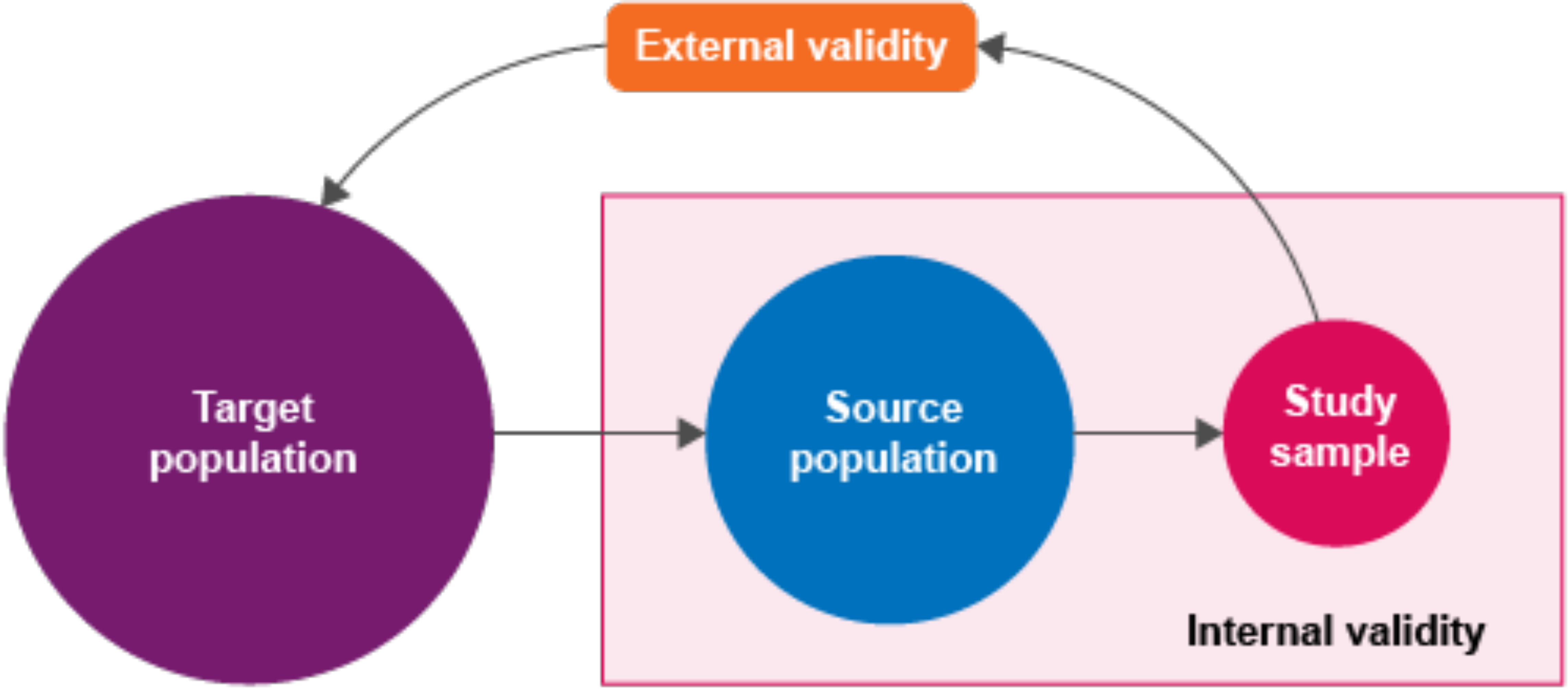


| | | | | |
|---|---|---|---|---|
| **Random Error** | large | large | small | small |
| **Systematic Error** | small | large | large | small |
| **Terminology** | imprecise, valid | imprecise, invalid | precise, invalid | **precise, valid → accurate** |

# EXTERNAL VALIDITY

A result is **externally valid** if the true effect in the study sample is unbiased for the true effect in the target population

A result is **internally valid** when the effect *estimated* in the study sample is unbiased for the *true* effect in that sample


External validity
Target population
Source population
Study sample
Internal validity

# CONCEPT OF TARGET VALIDITY

# “Hierarchy” of study designs

- Different study designs are best for different questions

- **Target validity** = Internal validity + external validity
  - the overall validity of a causal effect estimate in the specific target population of interest.
  - RCT conducted in a white MSM under-30 (low external validity; high internal validity) vs
  - Observational study conducted in a sample of population where all groups are represented (high external validity; low internal validity)

- What kind of evidence are you aiming to produce?
  - If you are producing evidence more relevant to public health policy change then need to consider internal validity AND external validity

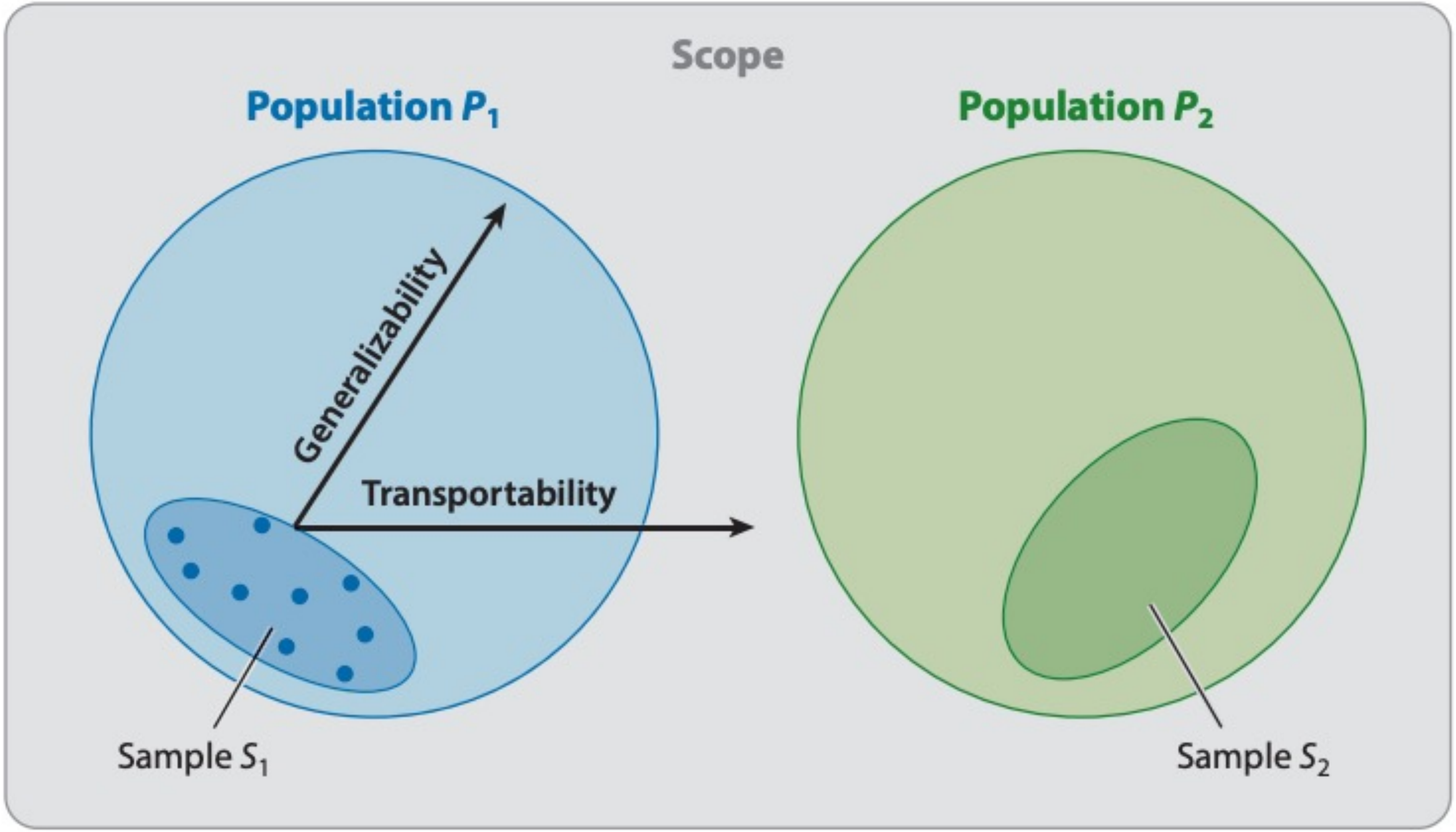
Daniel Westreich and others, Target Validity and the Hierarchy of Study Designs, *American Journal of Epidemiology*, Volume 188, Issue 2, February 2019, Pages 438–443, https://doi.org/10.1093/aje/kwy228

# Why is external validity a problem?

- If study sample is not the same as the target population → cannot assume that the true causal effect in a study sample will be the same as the true effect in the population

- It is near-universally overlooked that estimates of causal effects obtained from a study sample are only well-defined if they include **specific reference to a target population** in which they are said to apply

# GENERALISIBILITY VS TRANSPORTABILITY


Scope
Population $P_1$
Population $P_2$
Generalizability
Transportability
Sample $S_1$
Sample $S_2$

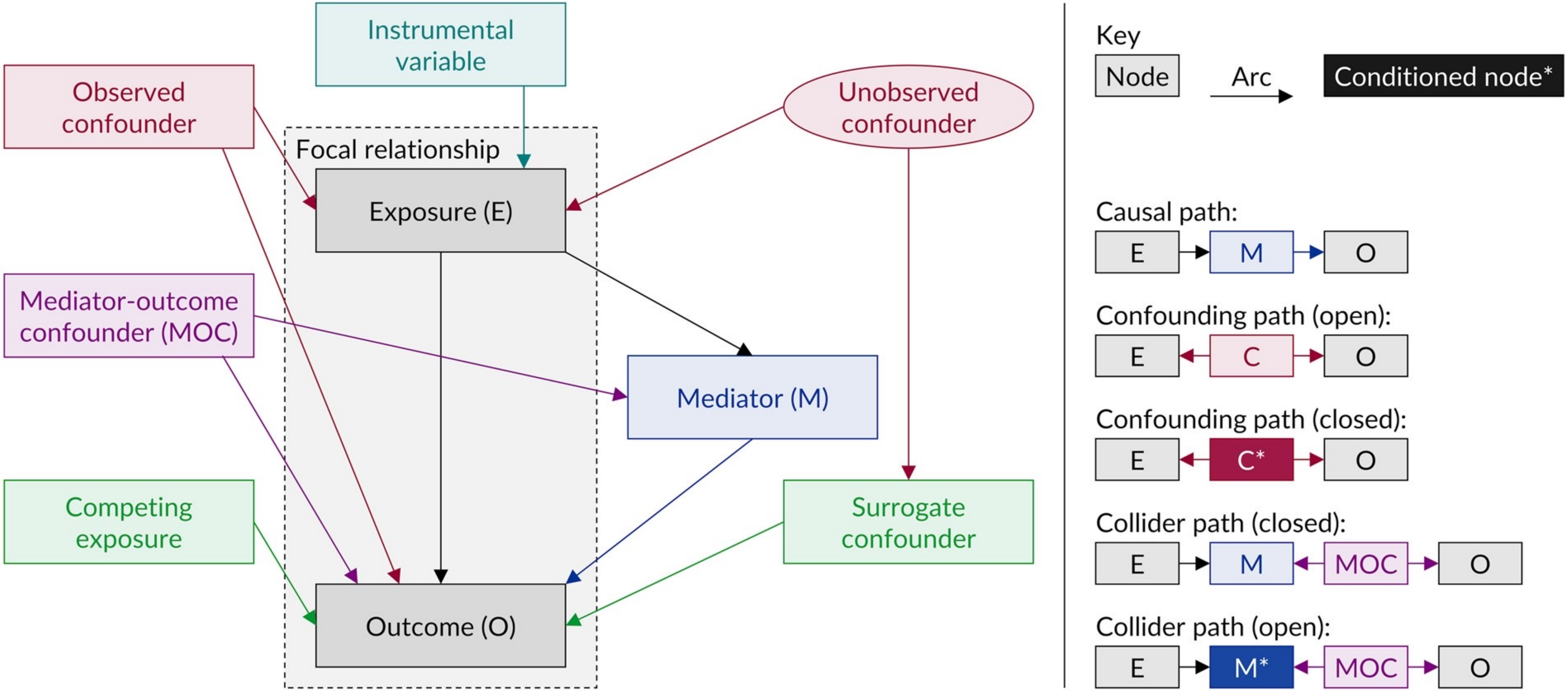
# DAGS IN MORE DETAIL

# DAGs

- DAGs are **non-parametric diagrammatic representations of the assumed data-generating process** for a set of variables in a specified context
- Variables and their measurements are depicted as nodes connected by unidirectional arcs (or arrows; hence ‘directed’) depicting the hypothesized relationships between them
- An arc between two nodes denotes the assumed existence and direction of a causal relationship
- It does not specify the sign (positive or negative), magnitude (large or small), shape (linear or non-linear) or form of that relationship (hence ‘non-parametric’)

Peter W G Tennant et al., Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations, *International Journal of Epidemiology*, Volume 50, Issue 2, April 2021, Pages 620–632, https://doi.org/10.1093/ije/dyaa213

**Figure 1** Illustration of the main components of a DAG, the most common types of contextual variables and the most common types of paths



# Structural approach to systematic error

- You will get unbiased estimate for what you are trying to estimate (‘estimand’) by closing all paths that are **not causal** and leaving all causal paths open

- Confounding
  - *Failure to condition on a common cause*
- Collider-stratification (selection) bias
  - *Conditioning on a common effect*
- Over adjustment bias
  - *Conditioning on a variable (or marker of a variable) that lies along the causal pathway*

# FORMAL CAUSAL INFERENCE METHODS

# Causal inference frameworks

1. Potential Outcome (PO) framework, associated with the work by Donald Rubin, building on the work on RCTs from the1920s by Ronald Fisher and Jerzey Neyman
2. Structural causal modelling framework including Directed Acyclic Graphs (DAGs) and do-calculus, much of it associated with work by Judea Pearl and his collaborators

These frameworks are complementary, with different strengths that make them particularly appropriate for different questions

# Causal identification conditions

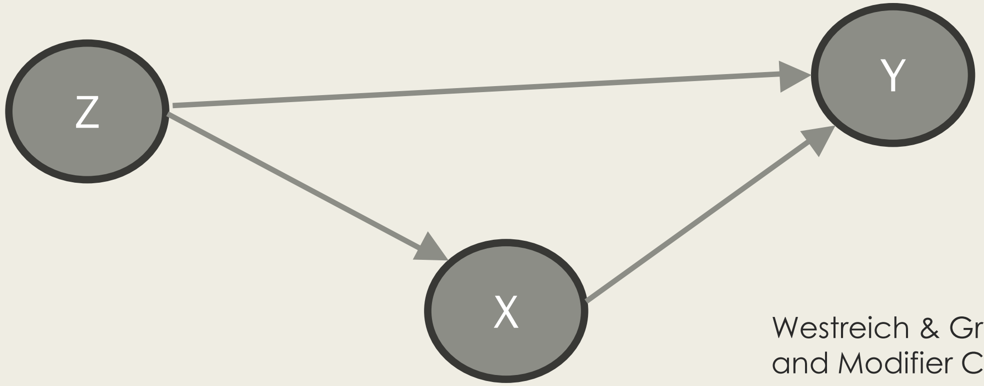
- Temporality
- Causal consistency
- Exchangeability
  - *Conditional exchangeability with positivity*
- No measurement error

- Machine learning can help to ensure correct model specification with fewer assumptions
- A WHOLE FIELD OF RESEARCH OUT THERE!

# TABLE 2 FALLACY

# A conceptual mismatch: Table 2 Fallacy

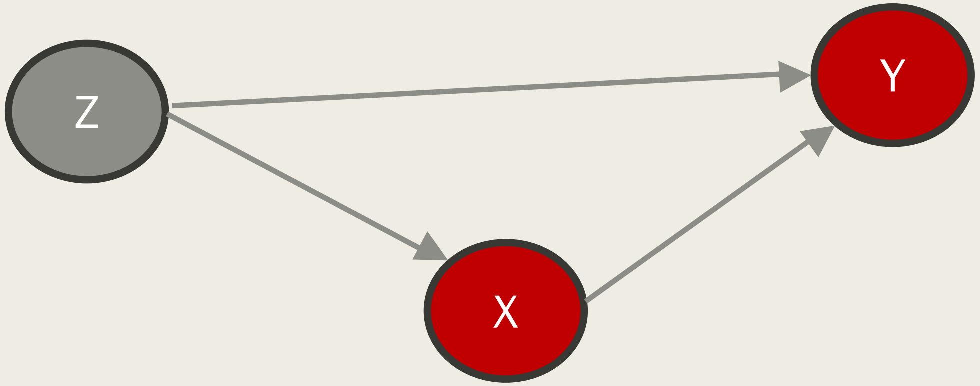
- Be wary of the ‘risk factor’ analysis
- Table 1 in most epidemiological articles =  baseline characteristics
- Table 2 reports the results of a multivariable regression model i.e., multiple adjusted effect estimates from a single model



Westreich & Greenland, The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients, *American Journal of Epidemiology*, Volume 177, Issue 4, 15 February 2013, Pages 292–298, https://doi.org/10.1093/aje/kws412

# A conceptual mismatch: Table 2 Fallacy

- Table 2 reports the results of a multivariable regression model i.e., multiple adjusted effect estimates from a single model
- Run our regression model: Y ~ X + Z
- Coefficient of X estimates the causal effect of X on Y
- But in Table 2 we also present the coefficient of Z



- Interpret the coefficient as the effect of Z on Y
- Can you see the problem?

# An example

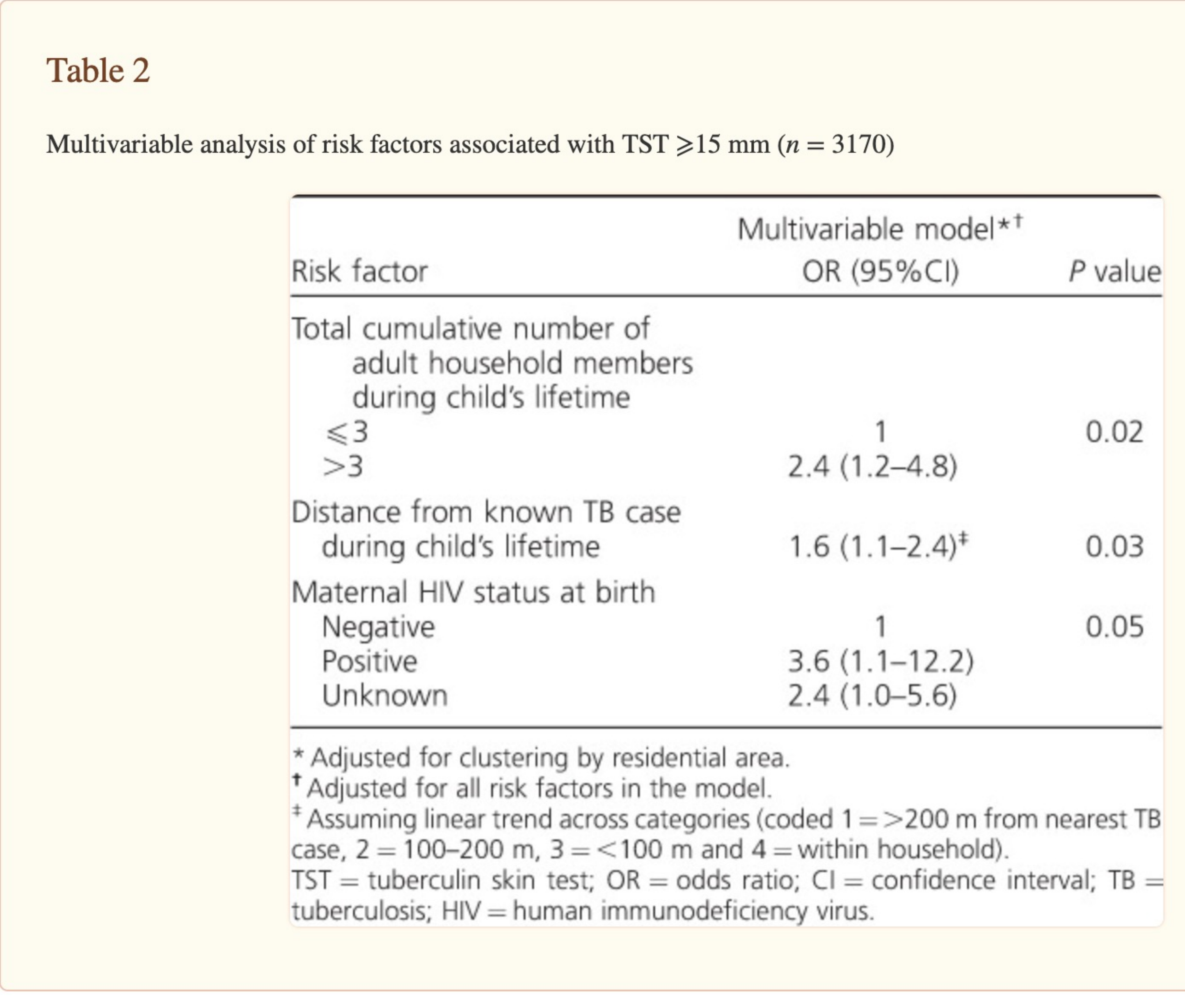


Table 2

Multivariable analysis of risk factors associated with TST ⩾15 mm (*n* = 3170)

| Risk factor | Multivariable model*† <br> OR (95%CI) | *P* value |
|---|---|---|
| Total cumulative number of adult household members during child's lifetime | | |
| ⩽3 | 1 | 0.02 |
| >3 | 2.4 (1.2–4.8) | |
| Distance from known TB case during child's lifetime | 1.6 (1.1–2.4)‡ | 0.03 |
| Maternal HIV status at birth | | |
| Negative | 1 | 0.05 |
| Positive | 3.6 (1.1–12.2) | |
| Unknown | 2.4 (1.0–5.6) | |

* Adjusted for clustering by residential area.

† Adjusted for all risk factors in the model.

‡ Assuming linear trend across categories (coded 1 = >200 m from nearest TB case, 2 = 100–200 m, 3 = <100 m and 4 = within household).

TST = tuberculin skin test; OR = odds ratio; CI = confidence interval; TB = tuberculosis; HIV = human immunodeficiency virus.

- Don't do this
- If you are interested in each of these 'risk factors' then draw a DAG for each
- 3 different models rather than putting them all together in one model

# EFFECT MEASURE MODIFICATION & INTERACTION

# Again terminology – *arrghh!!*

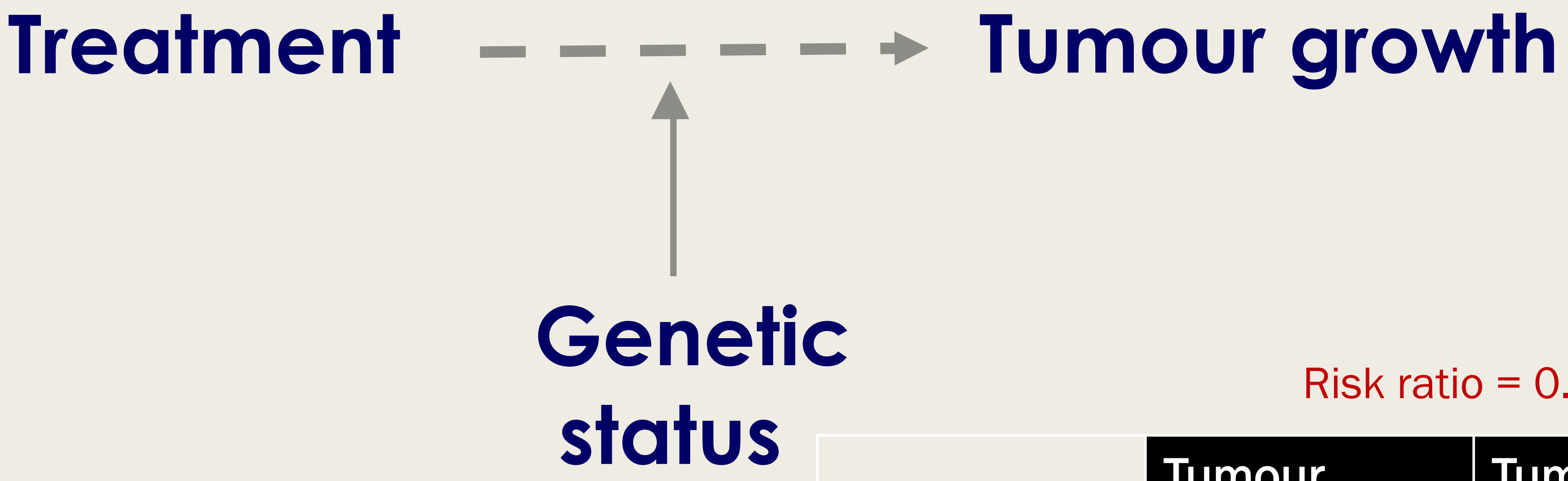
- Effect modification
  - *Effect measure modification*
- Interaction
  - *Statistical interaction*
  - *Causal (biological) interaction*
- Heterogeneity of effects
- Treatment effect heterogeneity
- Moderation

# My preferred definitions

- Effect measure modification is different to causal interaction
- Effect measure modification is concerned with a single causal effect of an exposure on an outcome and whether it differs in the different observed levels of an additional variable
- Causal interaction (to be distinguished from statistical interaction) involves assessing two independent and exposures and examining their joint causal effects
  - *More complex*
  - *Need to confront causal identification conditions for both exposures*
- Statistical interaction is a term relevant to regression modelling
  - *Can be interpreted as either casual interaction or effect measure modification* ***depending on the assumptions*** *made*

# Effect measure modification

Is the effect of a treatment on tumour growth modified by genetic status?



Risk ratio = 0.75

| | Tumour growth + | Tumour did not grow | Row total |
|---|---|---|---|
| Treatment | 30 | 370 | 400 |
| Control | 160 | 1440 | 1600 |

# Effect measure modification

Stratified analysis by genetic status

Risk ratio = 1.0

Genetic status- (N=1,600)

| | Tumour growth + | Tumour did not grow | Row total |
|---|---|---|---|
| Treatment | 20 | 180 | 200 |
| Control | 140 | 1260 | 1400 |

Risk ratio = 0.5

Genetic status+ (N=400)

| | Tumour growth + | Tumour did not grow | Row total |
|---|---|---|---|
| Treatment | 10 | 190 | 200 |
| Control | 20 | 180 | 200 |

# Interpretation

- No effect in those that were negative for the specific genetic status
- Highly protective in those that were positive for the genetic status – reduced risk of tumour growth by 50%
- Average treatment effect is misleading (RR = 0.75)
  - *Results are best presented stratifies*
- Facilitates personalized medicine

# Don’t get confused between confounding and effect measure modification

- Uses stratification by the third variable
- Examines the stratum-specific effect measure
- Very different epidemiological concepts

# TIME ZERO

# What do I mean about time zero?

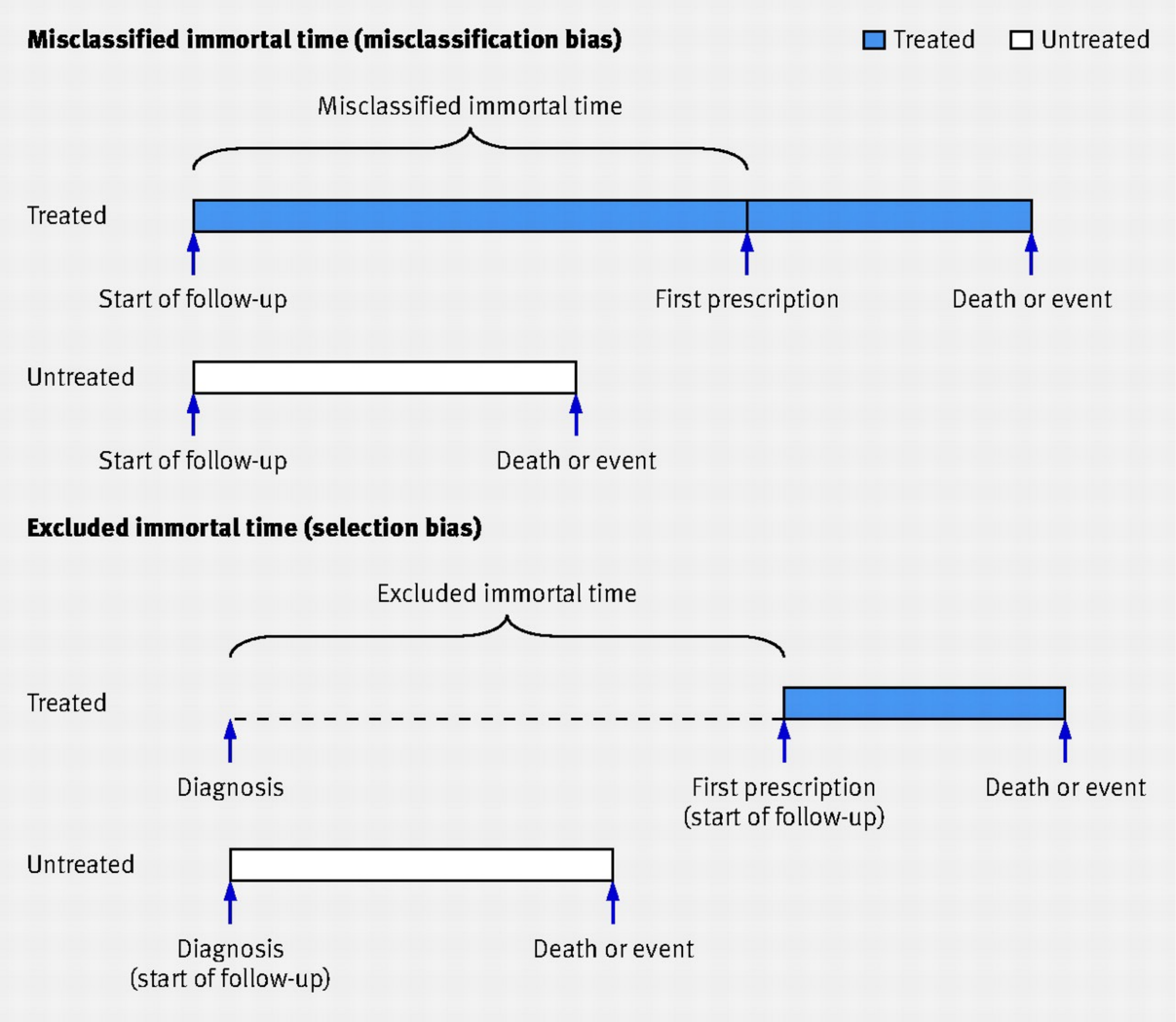
- The time at which the follow-up starts in a longitudinal study = $t_0$
- Make sure that it is the same for both groups you are comparing
- Be especially wary if using routine health records of the risk of **immortal time bias**

- Immortal time refers to a period of follow-up during which by design → death or the study outcome **cannot** occur

Hernán et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016 Nov;79:70-75. PMCID: PMC5124536.

# Immortal time bias

Time-dependent analyses where person-years of follow-up are classified as untreated until first prescription



Target trial emulation using observational cohort data

Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol.*2016;183:758–64

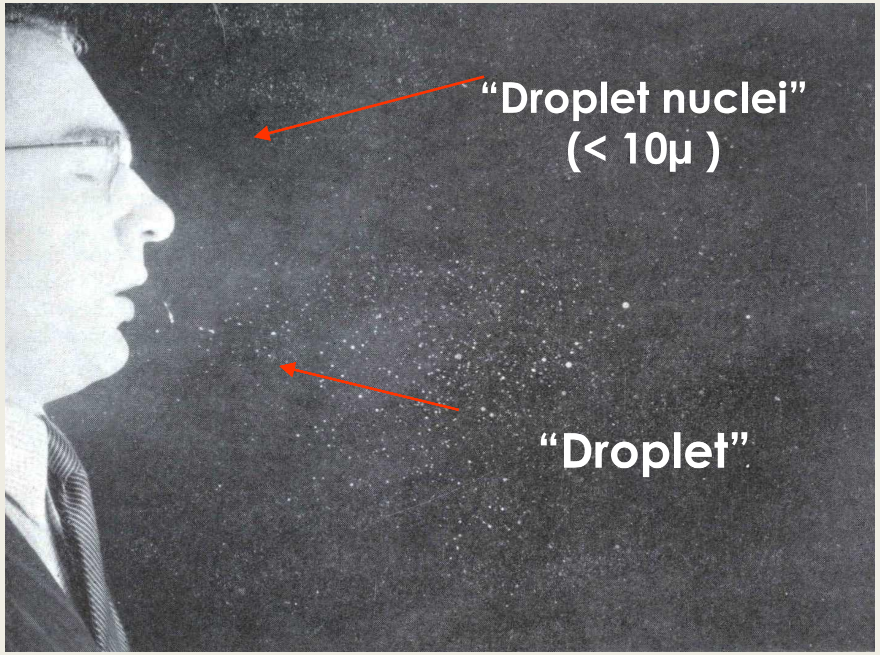
Lévesque et al. Problem of immortal time bias in cohort studies *BMJ* 2010; 340 :b5087 doi:10.1136/bmj.b5087

# METHOD FOR THINKING

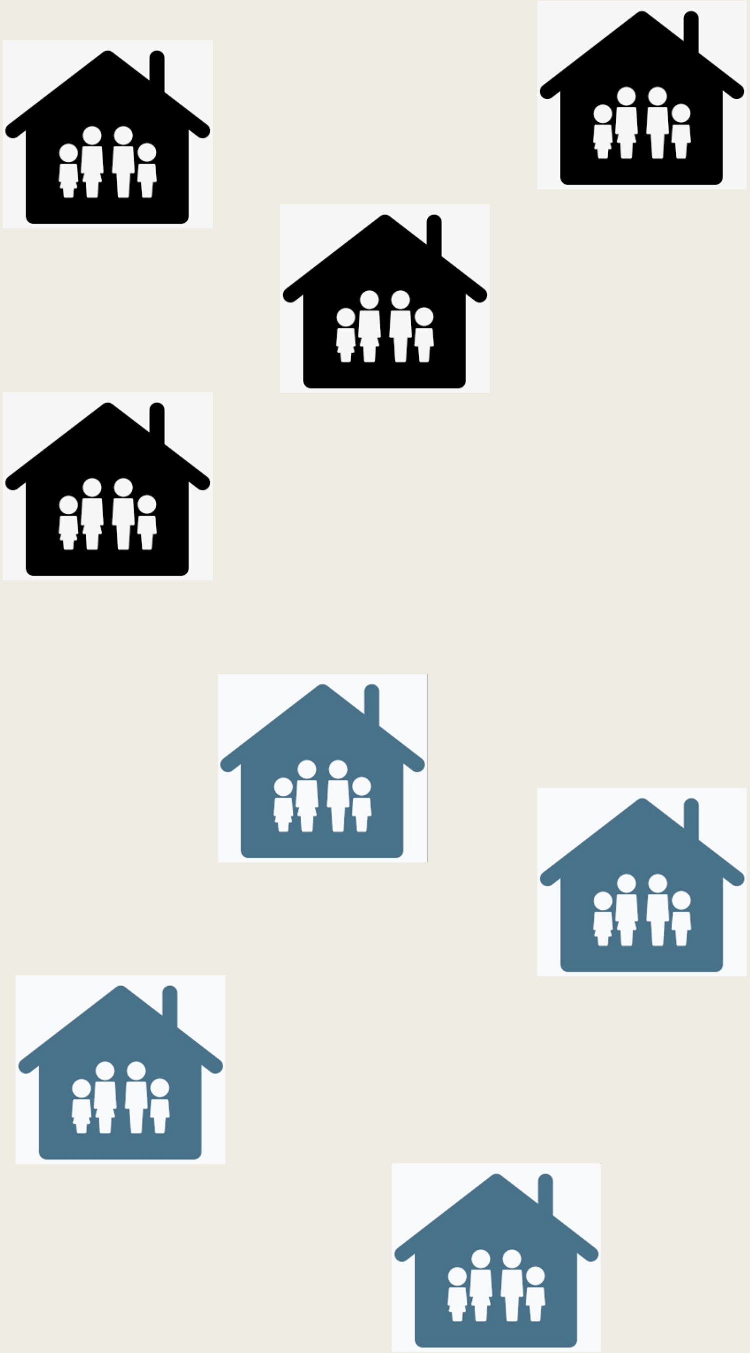
# Real-life example of using epidemiological thinking

- **Problem**:
  - People with TB have symptoms so they seek care, get diagnosed and treated
  - Control strategy dependent on symptomatic TB
  - TB prevalence surveys have found many individuals who are ‘well’ but have grown *M.tuberculosis* from their sputum sample = ‘subclinical’ TB

- **Research Qs**:
  - Are these individuals ‘infectiousness’?



- TB is predominantly transmitted by **airborne transmission** by very small droplets
  - *Individuals with infectious pulmonary disease*
  - *Cough, sneeze, talk, sing, breath*



# How to measure ‘infectiousness’?

- It is complex and far from ideal
- Near impossible to measure infectiousness – huge measurement error
- Tests for *M.tb* infection are based on identifying a host immune response – not directly finding the ‘bug’
  - *Blood test called interferon-gamma release assay (IGRA)*
  - *V different to nasal or throat swabs for SARS-CoV-2*
- Measure the proportion of household contacts who are IGRA+ in the household of people identified with ‘subclinical’ TB
  - *Compare to proportion of household contacts who are IGRA+ in households of people identified with ‘clinical’ TB*
  - *Compare to proportion of household contacts who are IGRA+ in ‘control’ households*

# Issues to consider

- Compare subclinical TB vs clinical TB then we need to make sure that we minimize outcome misclassification
  - ***KEY assumption*** *– we are saying that any household contact who has a positive IGRA has been infected by the 'index' participant*
    - younger children more likely to have been infected by HH index case than through community/non-HH contact
    - Testing children aged 2-14yr

*subclinical*

TB-affected households

*clinical*

# Issues to consider

- Compare subclinical TB vs control household
  - *Control household provides a measure of ‘background community transmission’*
    - How do we choose our control household?
      - *Households that have never had diagnosed TB person in the household*
      - *Are these households likely to have the same background transmission risk as the household that is affected by TB?*

*subclinical*

*subclinical*

TB-affected households

*clinical*

*control*

# Issues to consider

- You find a lower proportion IGRA+ in the control HH compared to the subclinical HH
  - *living in a HH with subclinical case increases risk of IGRA-positivity (M.tb infection) by 2-fold compared to control HH*
  - *Maybe all you have picked up is that the HH contacts in control HH were at a lower risk of community transmission than the subclinical household – different social contact patterning etc*
  - *Can we collect all the variables to be able to adjust for this?*

*subclinical*

TB-affected households

*clinical*

*subclinical*

*control*

# Potentially need different study designs for the different Qs

## 1. Observational

*subclinical*

*clinical*

TB-affected households

- Collect information on potential confounders, predictors
- Degree of ‘exposure’ – closeness of contact to index person
- Biological markers of infectiousness

## 2. RCT

*subclinical*

- Ethical equipoise – should these individuals be given 6 months of treatment?
- Randomised the index person to TB Rx or placebo and follow up HH contacts
- Measure conversion from IGRA-negative to IGRA+ in HH contacts

# DAGITTY EXERCISE

# DAGitty – draw and analyse causal diagrams

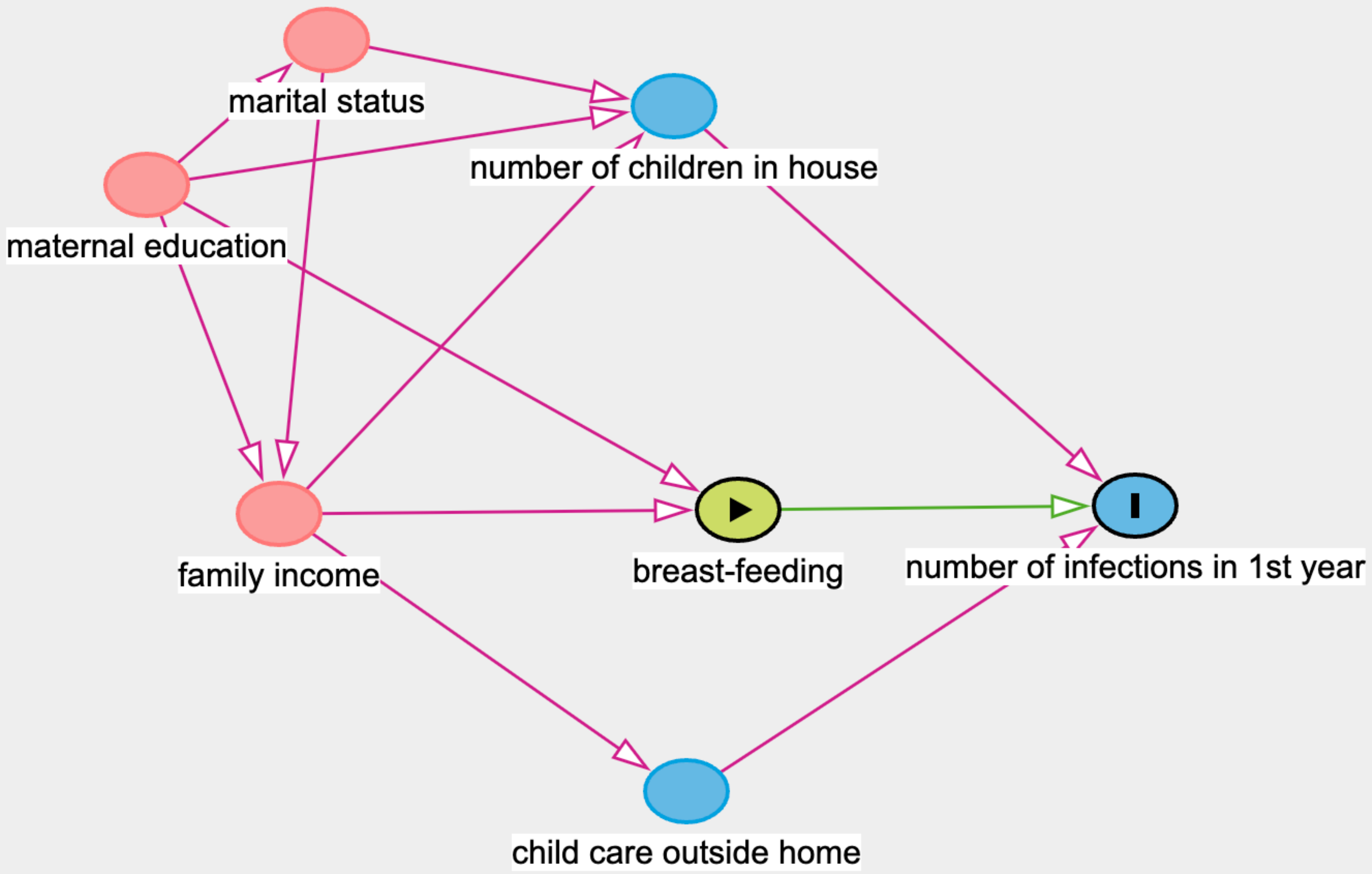
- DAGITTY

# Create a DAG for focal relationship of effect of **breastfeeding** on **number of upper respiratory tract infections in 1st year**

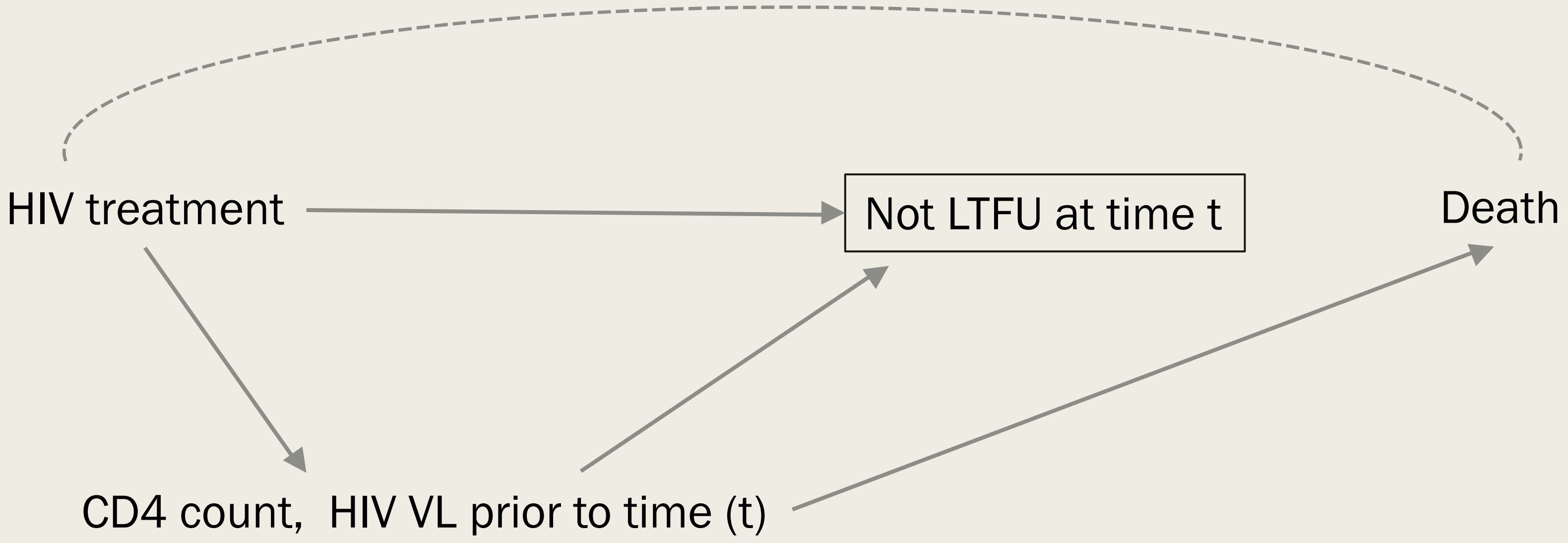
Seven variables

- Breastfed
- Num_URTI
- Childcare_outside_home
- Family_income
- Education_status_mother
- Marital_status
- Num_children_in_house

- Maternal education determines:
  - Marital status
  - Number of children in house
  - Breast-feeding status
  - Family income
- Marital status determines:
  - Family income
  - Number of children in house
- Family income determines:
  - Childcare outside of home
  - Breast-feeding status
- Number of children in house determines:
  - Number of infections in 1st year
- Childcare outside of home determines
  - Number of infections in 1st year


marital status
number of children in house
maternal education
family income
breast-feeding
number of infections in 1st year
child care outside home

# RCTs and collider bias



LTFU may have had more severe disease and died or had more side effects from medication and dropped out – conditioning on those still in the study introducing collider bias

# 10 steps to drawing a DAG like a pro!

1. Develop and state a clear research question
2. Consider and state your context
3. Draw your DAG(s) as early as possible
4. Get help - don’t draw it alone
5. Include all relevant variables
6. Draw your DAG(s) in temporal order
7. Draw forward arcs, unless confident otherwise
8. Check & update your DAG(s) against your data
9. Use your DAG(s) to inform and interpret your model
10. Share & publish your DAG(s)

Peter Tennant Leeds Causal School 2022

# DAGs are not a panacea

- Many limitations
- Rapidly expanding area of research

- Help with being explicit about causal assumptions underlying your analysis
- Help to identify potential variables that you should NOT adjust for
- Help with maintaining a principled workflow to an analysis
- Simulate data from DAGs to understand epidemiological concepts
  - *I will share the pdf of this paper (not open access)*