Yale SCHOOL OF PUBLIC HEALTH

DEMYSTIFYING SELECTION BIAS LECTURE

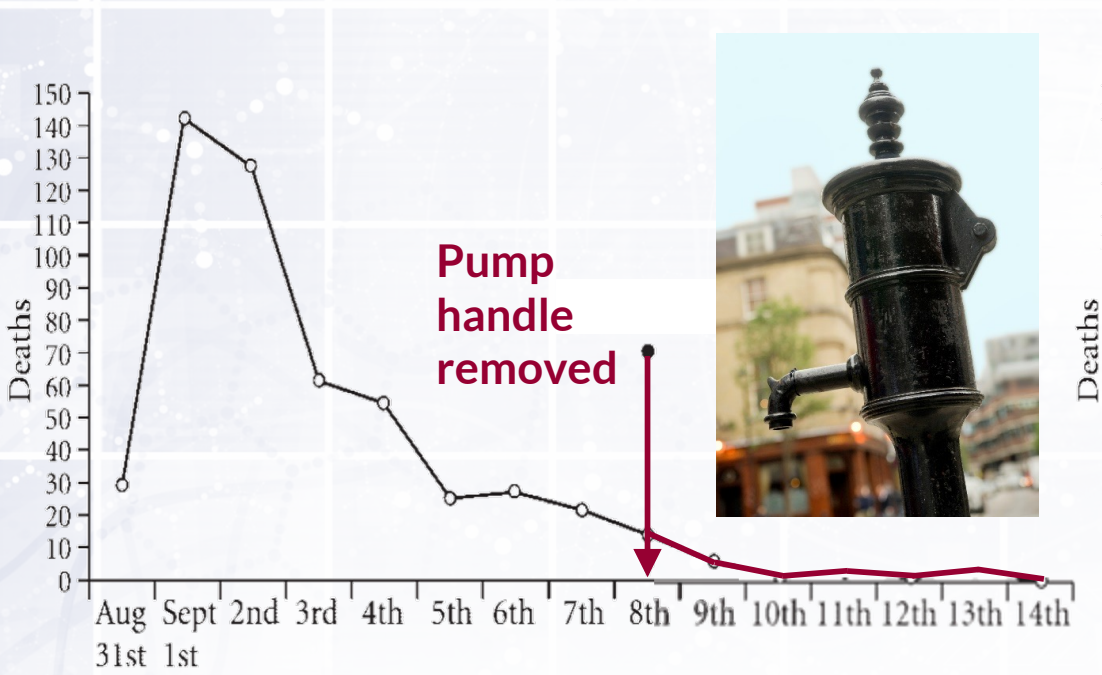
Peter WG Tennant, PhD

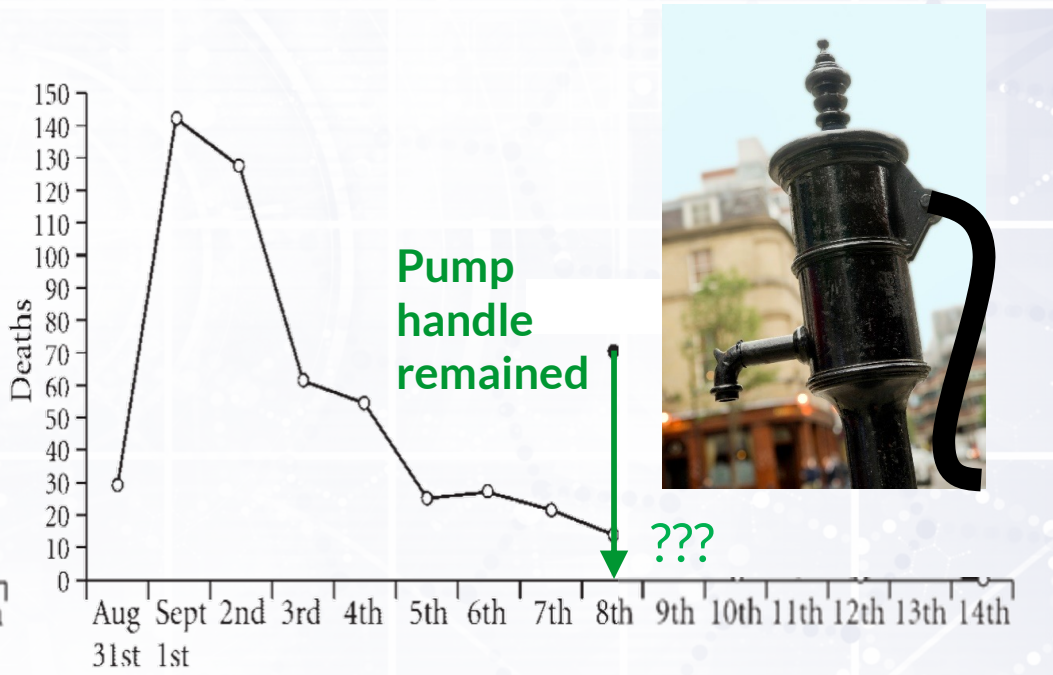
George Saden Visiting Associate Professor

PETERWGTENNANT

@PWGTennant

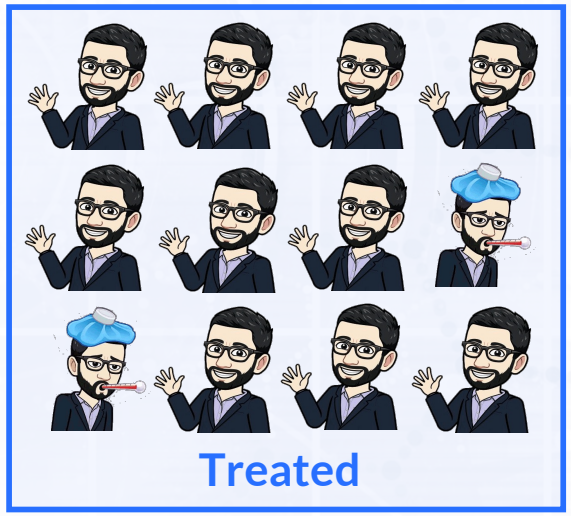
- We cannot *know* the effect on an exposure on an outcome because we cannot rewind time and see what would have happened if the exposure had been different
  - This is known as the **fundamental problem of causal inference**





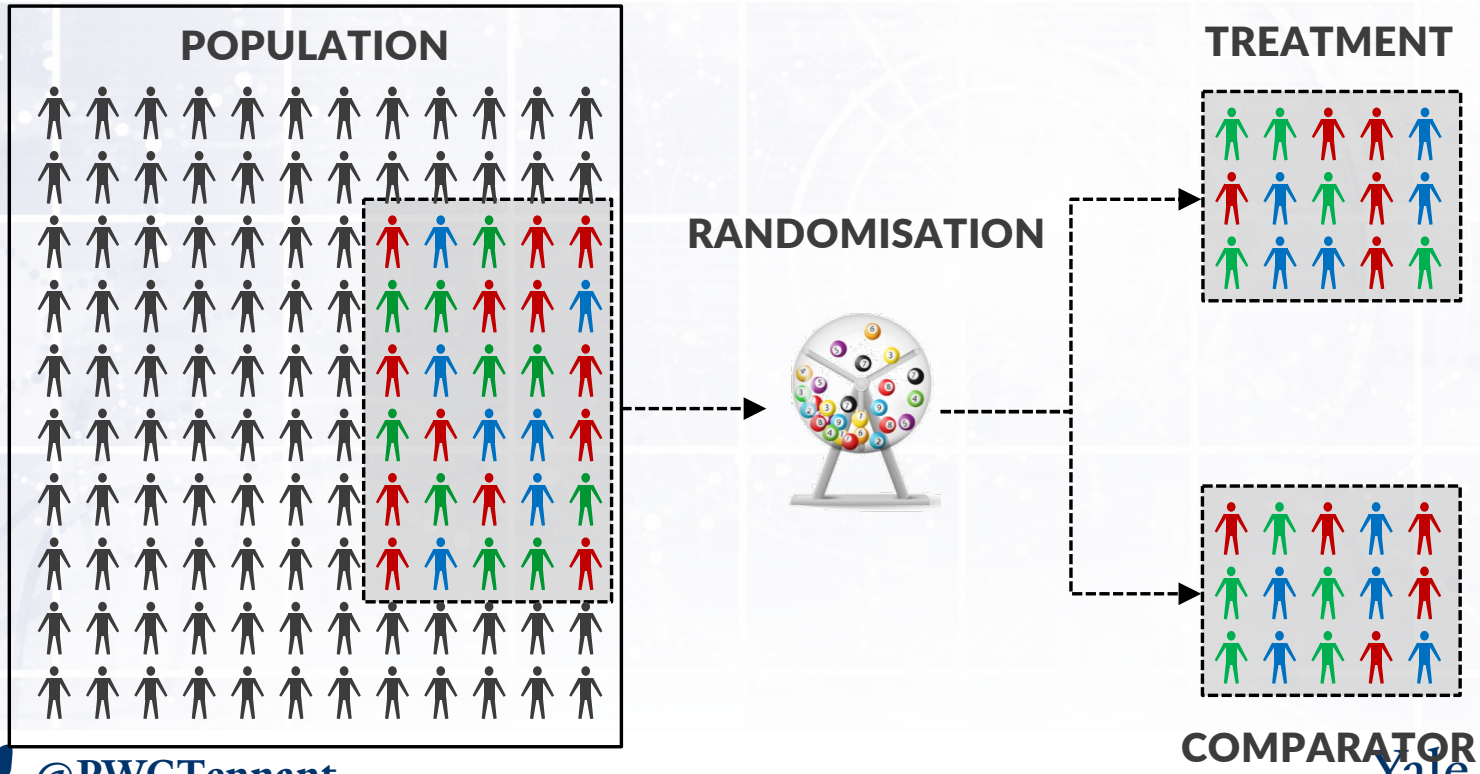
BACKGROUND RECAP

- We can *estimate* the **average causal effect** of an exposure on an outcome by comparing outcomes between groups with the same average risks of the outcome at the time of exposure assignment
- i.e. by comparing outcomes between two groups that are **exchangeable**



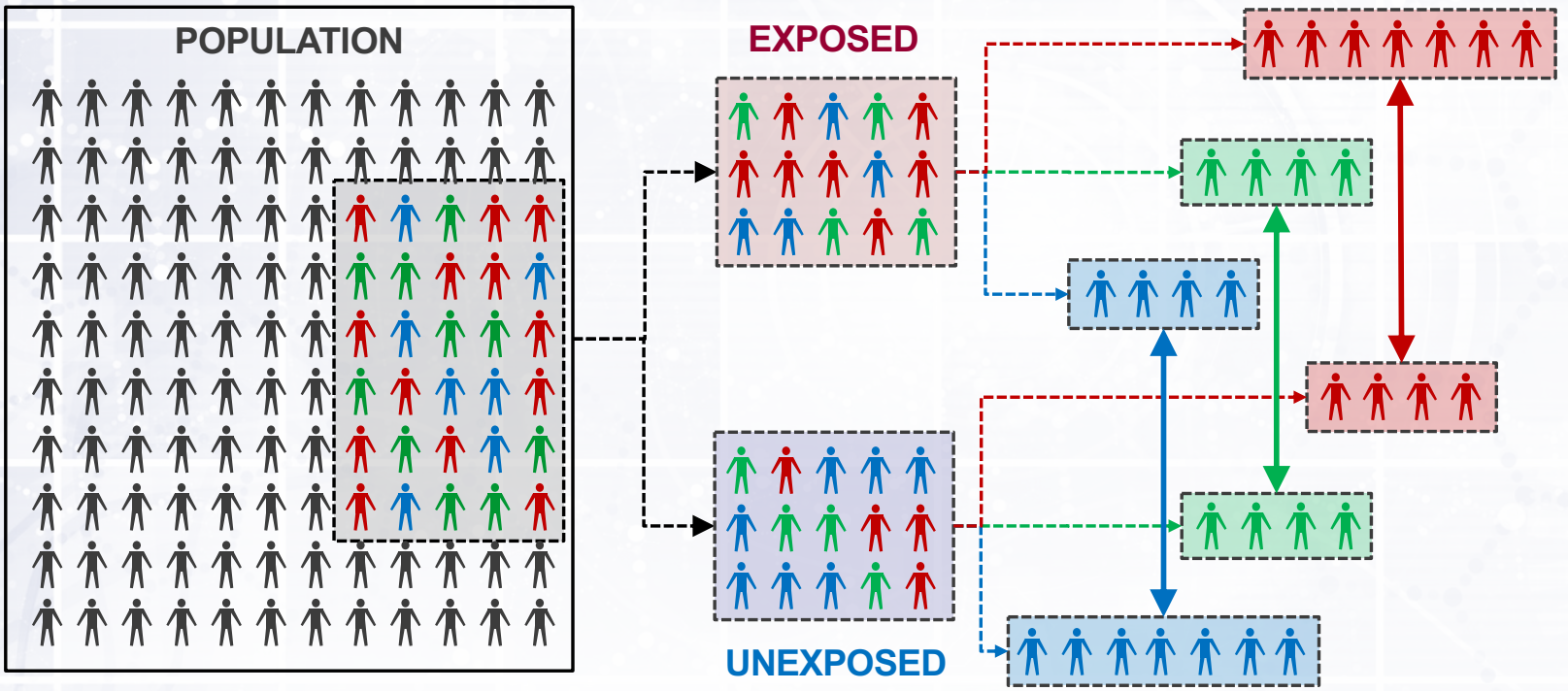
UNCONDITIONAL EXCHANGEABILITY

- The most reliable way to obtain exchangeable units of analysis is through randomisation
- If exposure is assigned **at random** it cannot be related to probability of outcome, and causal effect can be estimated from difference in observed outcomes



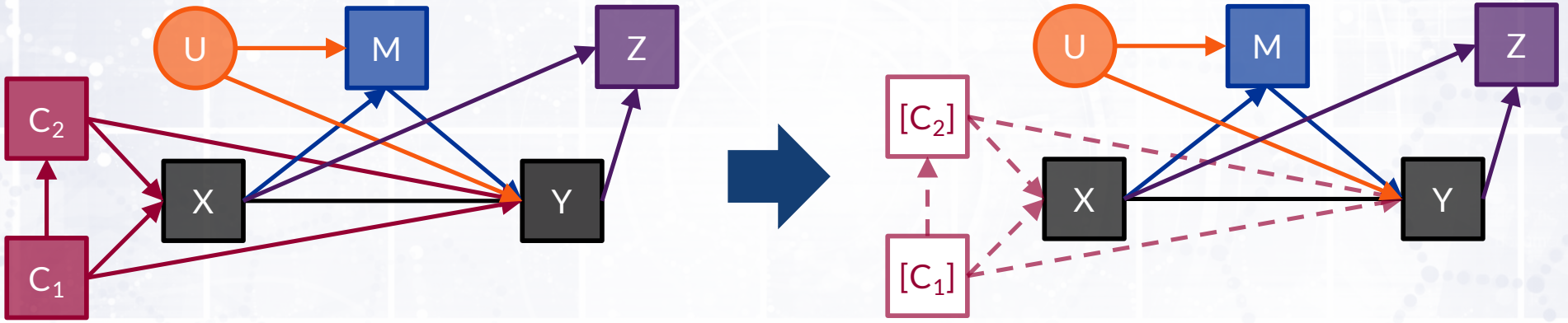
CONDITIONAL EXCHANGEABILITY

- Without randomisation we must aim for **conditional exchangeability**
- I.e. we must identify **conditional subgroups** within which the exposure was effectively assigned at random



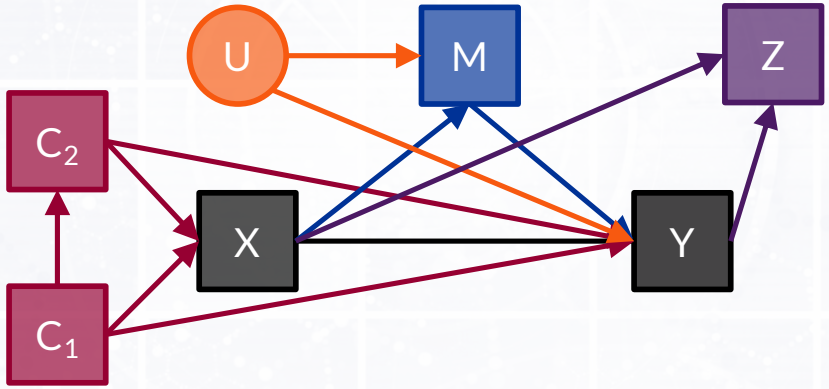
RECAP: IDENTIFYING EXCHANGEABLE GROUPS OF ANALYSIS

- Once we have clearly defined our target causal effect of interest (the **estimand**), causal diagrams such **directed acyclic graphs (DAGs)** help us to identify which variables need conditioning to obtain **exchangeable groups**



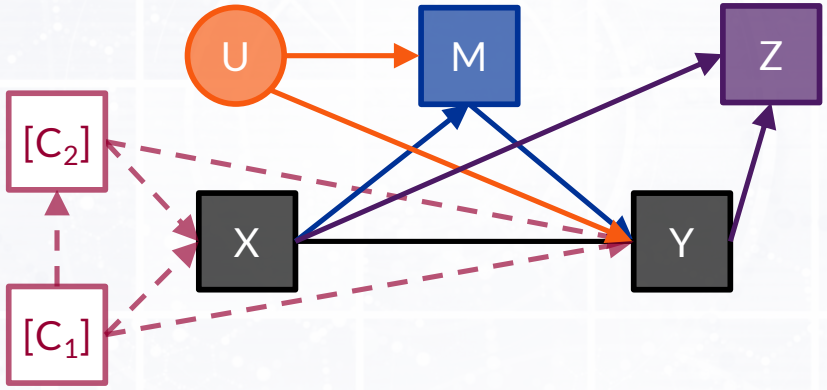
RECAP: THE AVERAGE CAUSAL EFFECT

- To estimate the **average causal effect** of **X** on **Y**
  - Want all **causal paths** to be **open**
  - Want all **confounding paths** to be **closed**
  - Want all **collider paths** to be **closed**



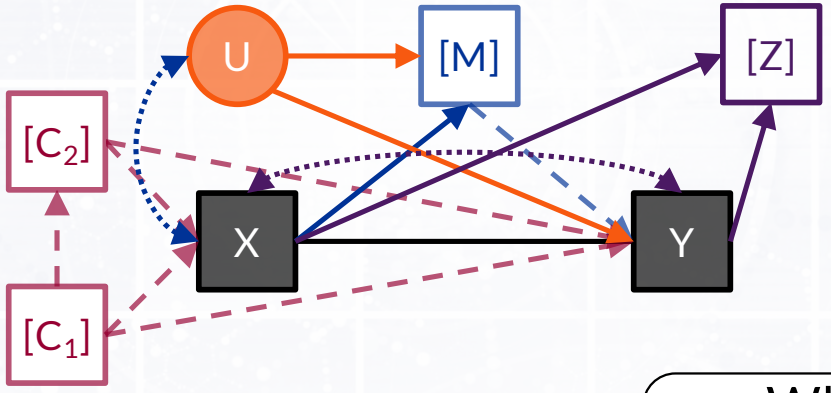
- To estimate the **average causal effect** of **X** on **Y**
  - Want all **causal paths** to be **open**
  - Want all **confounding paths** to be **closed**
  - Want all **collider paths** to be **closed**

DO:
Condition on
confounders (C_1 and C_2)

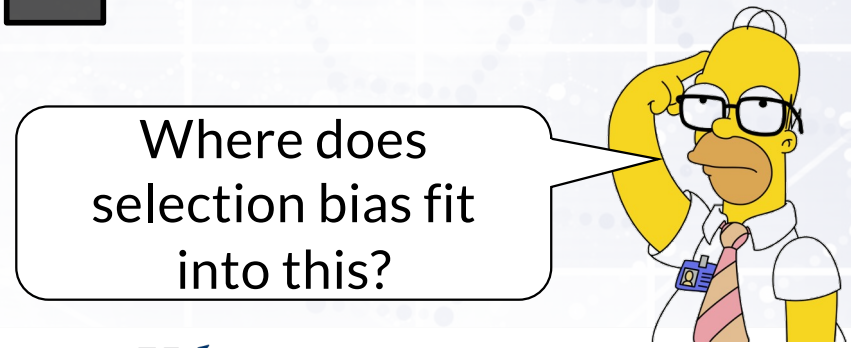


- To estimate the **average causal effect** of **X** on **Y**
  - Want all **causal paths** to be **open**
  - Want all **confounding paths** to be **closed**
  - Want all **collider paths** to be **closed**

DO:
Condition on
confounders (C_1 and C_2)

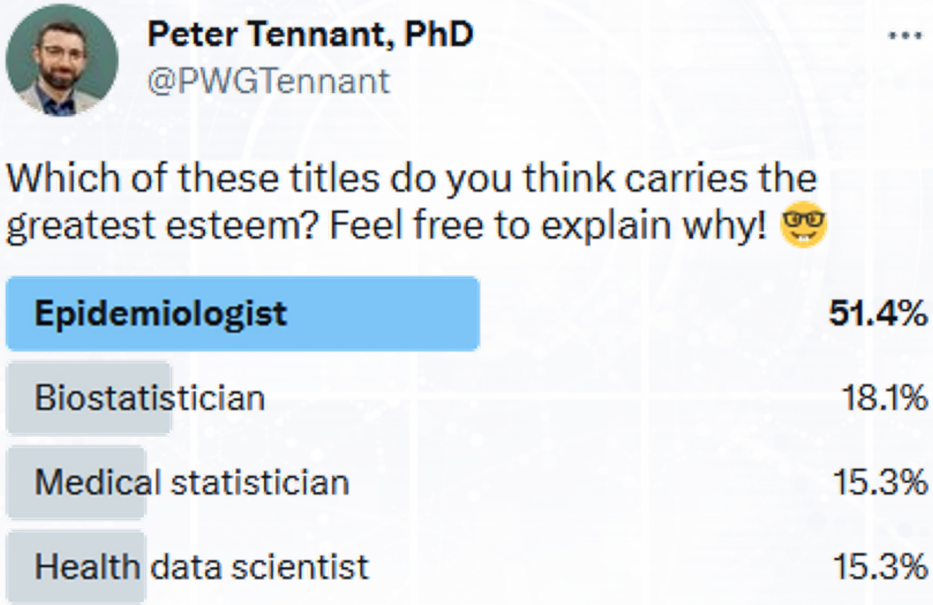


DO NOT:
Condition on **colliders**
(**M** and **Z**)



SELECTION BIAS IN DESCRIPTIVE RESEARCH

- It has long been known that an ‘unrepresentative’ survey can produce an ‘unrepresentative’ estimate of a particular measure of occurrence.
- This is exemplified by social media polls, which tell you what your echo chamber of followers think



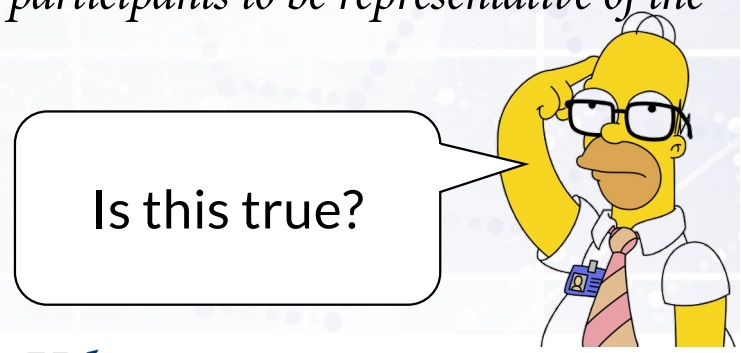
REASONS FOR A NON-REPRESENTATIVE SAMPLE

- A non-random analytical sample can occur for many reasons:
  - A non-random **sampling strategy**
    - i.e. who is approached to take part
  - Non-random **participation**
    - i.e. who agrees to take part
  - Non-random **attrition**
    - i.e. who decides to stay in the study
  - Non-random **missingness**
    - i.e. who has complete data available

SELECTION BIAS IN CAUSAL RESEARCH

- The potential for bias in causal research due to sampling and participation has historically been less well understood
- This is exemplified by a statement on the UK Biobank website until 2020
  - UK Biobank is a large, well funded cohort – launched in 2006 – that initially approached **9 million adults** to participate but only **~500,000** agreed, a **5.5% response ratio**
- They said:
  - "*UK Biobank is not representative of the general population… with evidence of a 'healthy volunteer' selection bias… However, the large sample size and heterogeneity of exposure measures allow for valid…inferences of associations between exposures and health outcomes that are generalizable to the wider population*"
  - "*Valid assessment of exposure-disease relationships… do not require participants to be representative of the population*"

biobank uk
Improving the health of future generations



A CLEAR DEFINITION OF SELECTION BIAS

- Definitions of **selection bias** remain quite varied and contradictory within the wider community
- Causal inference methods have however revolutionised understanding, and knowledge continue to advance
- In 2022, Lu et al proposed a unified definition that clarifies the two main issues that contribute to selection bias

Toward a Clearer Definition of Selection Bias When Estimating Causal Effects

Haidong Lu,[a] Stephen R. Cole,[b] Chanelle J. Howe,[c] and Daniel Westreich[b]

Abstract: Selection bias remains a subject of controversy. Existing definitions of selection bias are ambiguous. To improve communication and the conduct of epidemiologic research focused on estimating causal effects, we propose to unify the various existing definitions of selection bias in the literature by considering any bias away from the true causal effect in the referent population (the population before the selection process), due to selecting the sample from the referent population, as selection bias. Given this unified definition, selection bias can be further categorized into two broad types: type 1 selection bias owing to restricting to one or more level(s) of a collider (or a descendant of a collider) and type 2 selection bias owing to restricting to one or more level(s) of an effect measure modifier. To aid in explaining these two types—which can co-occur—we start by reviewing the concepts of the target population, the study sample, and the analytic sample. Then, we illustrate both types of selection bias using causal diagrams. In addition, we explore the differences between these two types of selection bias, and describe methods to minimize selection bias. Finally, we use an example of "M-bias" to demonstrate the advantage of classifying selection bias into these two types.

Keywords: Selection bias; Collider bias; Effect measure modification; Effect heterogeneity; Causal diagram; Internal validity; External validity; Epidemiologic research

(*Epidemiology* 2022;33: 699–706)

When estimating causal effects, selection bias remains a subject of controversy in epidemiology.[1] The definition of selection bias is not as clear as that of confounding or information bias. This controversy and ambiguity may stem from the fact that in the literature selection bias has sometimes been considered a threat to internal validity, while at other times it has been considered a threat to external validity.[2–4] To improve communication and the conduct of epidemiologic research focused on estimating causal effects, we propose a definition of selection bias with two types.

This article is organized as follows. First, we review the concepts of the target population, the study sample and the analytic sample and provide a refined definition of selection bias. Next, we describe two types of selection bias: type 1 selection bias owing to restricting to one or more level(s) of a collider (or a descendant of a collider), and type 2 selection bias owing to restricting to one or more level(s) of an effect measure modifier. Then, we use an example to demonstrate the importance of classifying selection bias into these two types. Last, we describe the utility of defining selection bias as having two types and conclude with a brief discussion.

Before proceeding, it will be useful to state the following assumptions. First, we assume that causal consistency is satisfied.[5,6] That is, here we will not consider interference[7] or multiple treatment versions.[6] For clarity and simplicity, hereafter we also assume no confounding of the exposure-outcome relationship, no measurement bias, and no random variability. For causal diagrams, we consider only four types of variables: binary exposure E, binary outcome D, selection S, and covariates L (e.g., L1, L2). Throughout, S = 1 means selection into the sample. It should also be

Lu et al 2022 Epidemiology

A CLEAR DEFINITION OF SELECTION BIAS

- Lu et al define selection bias as:

  *A systematic divergence between the true causal effect in the **target population** and the estimate obtained in the **analytical sample***

- The **target population** is the population in which you wish to estimate your **target estimand**

- Your estimate is **generalizable** to that population and has **target validity** if your study has:

  - **Internal validity**: satisfied if the estimate in your analytical sample matches the true causal effect in your study sample
  - **External validity:** satisfied if the true causal effect in your study sample matches the true causal effect in the target population

TWO TYPES OF SELECTION BIAS

- Selection bias can affect *both* **internal validity** *and* **external validity**

Type 1 Selection bias

Harms interval validity

Type 2 Selection bias

Harms external validity

TYPE 2 SELECTION BIAS

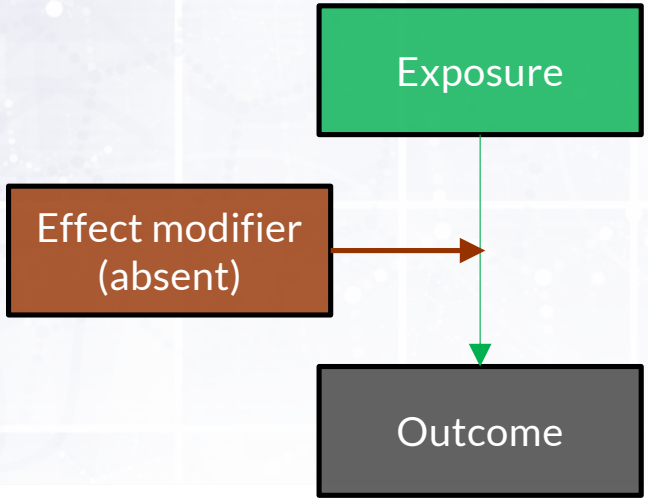
Type 2 Selection bias

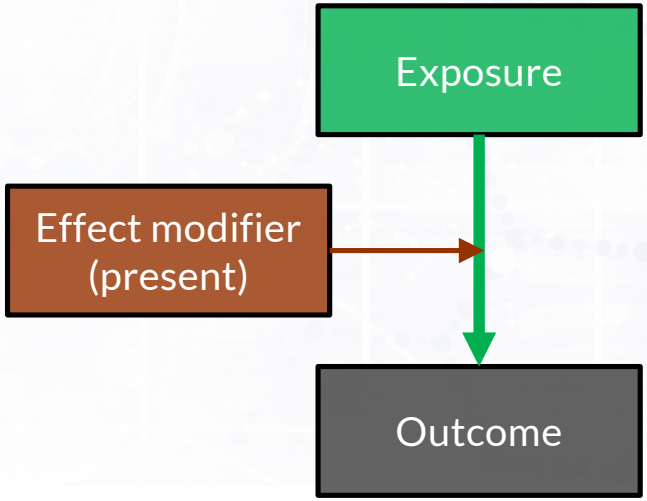
Harms external validity

- **Type 2 selection bias** (AKA **generalizability bias**) is the form of selection bias that harms the **external validity** of your causal effect estimate
  - Your estimate is valid in your sample, but does not generalize to you population
- It affects both **observational studies** and **randomised experimental studies**
- It occurs when the causal effect of your exposure on your outcome *varies* according to other characteristics,
  - I.e. when there is **effect measure modification** or **treatment effect heterogeneity**

EFFECT MEASURE MODIFICATION

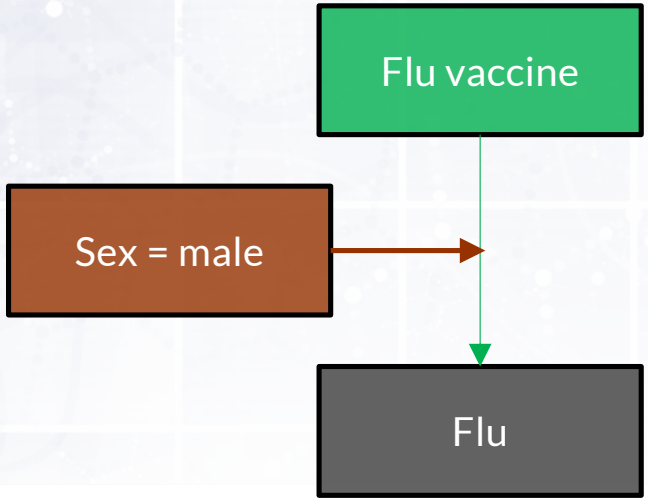
- **Effect measure modification** occurs when the effect of the exposure on the outcome depends on one or more other variables (known as **effect modifiers**)
- This means there is **no single 'causal effect'** of the exposure; the effect in a population will depend on the presence and balance of the effect modifiers
- If the balance of effect modifiers in your study sample are not the same as in the target population, then the average causal effect will differ from the **population average treatment effect (PATE)**

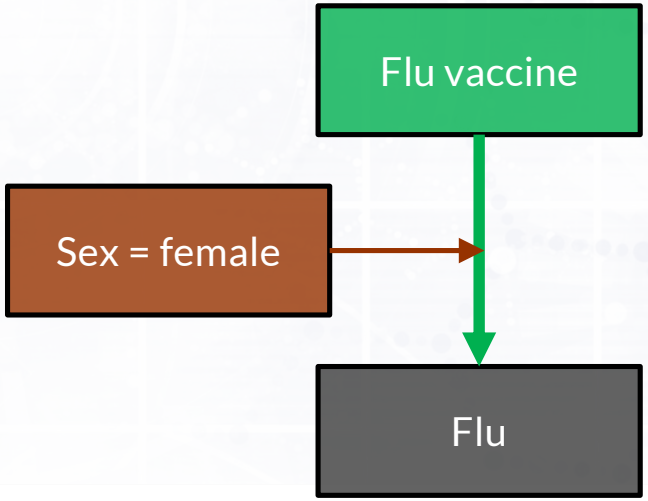




EXAMPLE: EFFECT MEASURE MODIFICATION

- **Example:** Flu vaccine and subsequent infection with flu
- The flu vaccine may be more effective in women (~50%) than men (~40%)
- Suppose your **target population** included 500 women and 500 men
  - **PATE** would be 45%
- But if your sample included 80% men and 20% women, then the (**sample average treatment effect, SATE**) would be 42% - a biased estimate of the **PATE**





- You can estimate the **PATE** from an unrepresentative sample using two equivalent methods:

  - **Standardisation** involves estimating the '***conditional***' causal effects in each subgroup, and then taking a weighted average of these to obtain the '***marginal***' effect in your population

	Vaccine effectiveness	Proportion in population	Effect x proportion
Men	0.40	0.5	0.2
Women	0.50	0.5	0.25
Target population			**0.45**

  - **Inverse probability of selection weighting** (IPSW) involves reweighting those in your sample to create a **pseudopopulation** with the same profile of effect modifiers as your target population

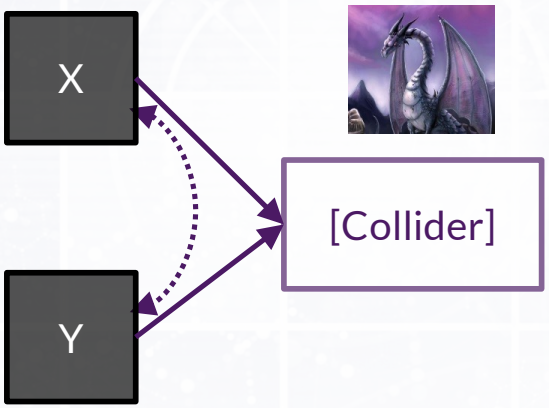
	Effectiveness	N (sample)	P(sample)	IPSW	N (Pseudo)	Effect * proportion
Men	0.40	80	0.16	1/0.16 = 6.25	80*6.25 = 500	0.4 * 500/1000 = 0.2
Women	0.50	20	0.04	1/0.04=25	20*25 = 500	0.5 * 500/1000 = 0.25
Target population		**100**			**1000**	**0.45**

- In practice, there are many challenges to correcting **for type 2 selection bias**
  - There will usually be *many* effect modifiers
  - They may not all be measured in your sample
    - You cannot estimate conditional effects if you don't have the variable!
  - They may not all be measured/knowable in your target population
    - You cannot standardize/reweight to something you don't know!
  - Some groups may be completely missing
    - You cannot estimate or reweight a conditional effect from a sample size of zero
- You need a diverse sample and a clear idea of how the key effect modifiers deviate between your sample and your target population

Type 1 Selection bias

Harms interval validity

- **Type 1 selection bias** (AKA **collider bias**) is the form of selection bias that harms the **internal validity** of your causal effect estimate
  - **RECALL: Collider bias** occurs when we **condition** on a **collider** to open a **collider path** between our **exposure** and **outcome**

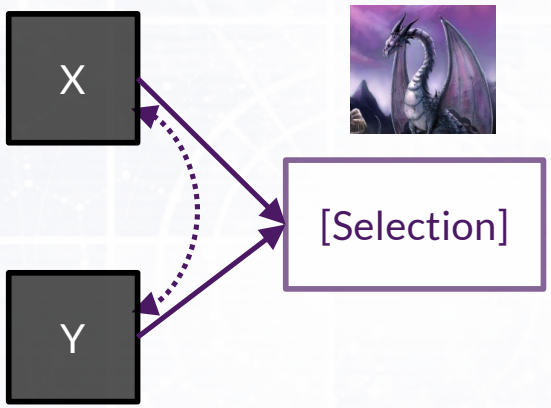


- It commonly affects multivariable regression models, where it can be avoided with appropriate variable selection

Type 1 Selection bias

Harms interval validity

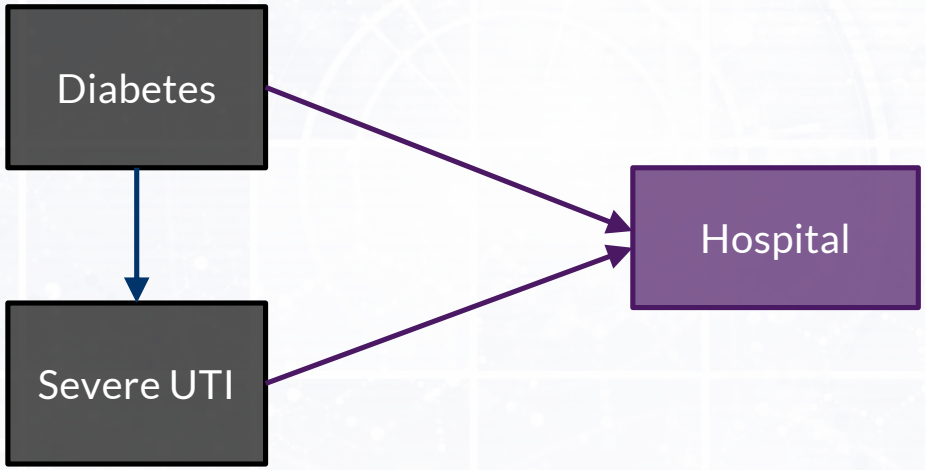
- **Type 1 selection bias** arises because **selection** into your sample is a **collider** on a path between your **exposure** and **outcome**



Type 1 selection bias can produce distorted associations even when your sample is the same as the people you are interested in!

Example: Berkson's bias

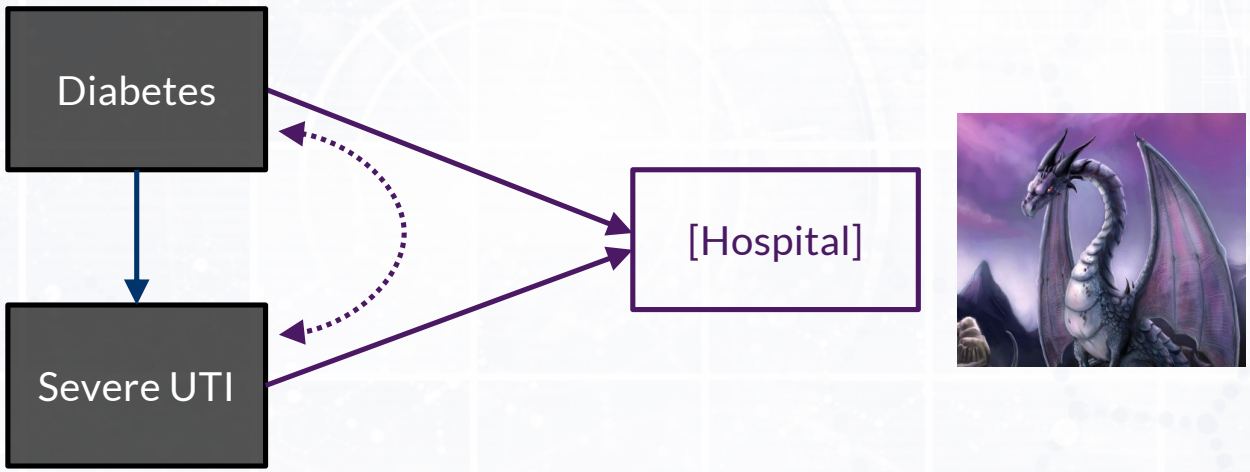
- **Diabetes** causes frequent and **severe urinary tract infections**
- We want to understand how much more common severe UTIs are in people with **diabetes** than those without



- Embracing '**big data**' we decide to estimate this in a **hospital admission data**

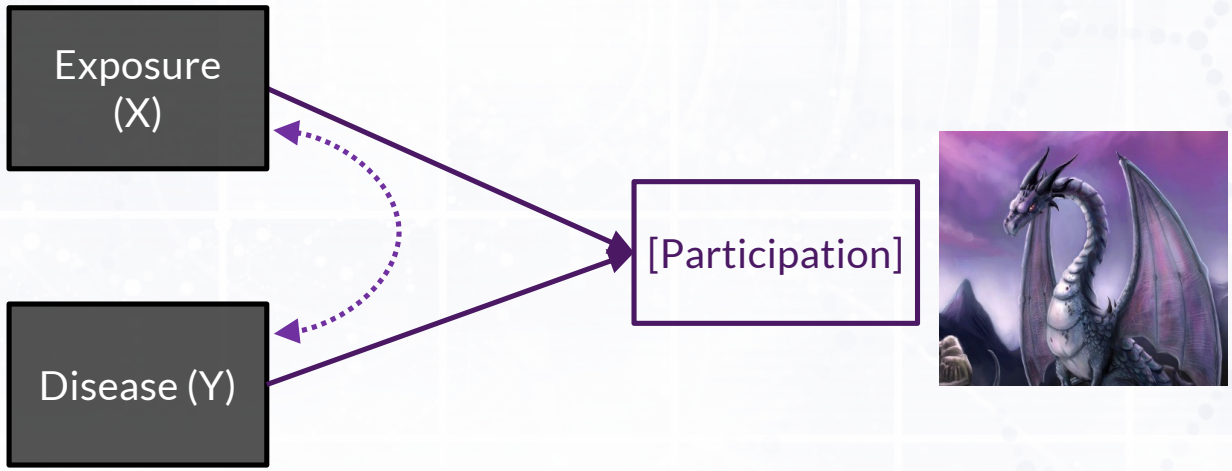
Example: Berkson's bias

- **Diabetes complications** and **severe UTIs** are two competing reasons for **[attending hospital]**
- By looking in **[hospital records]**, we open **Diabetes** ← . . . → **UTI**



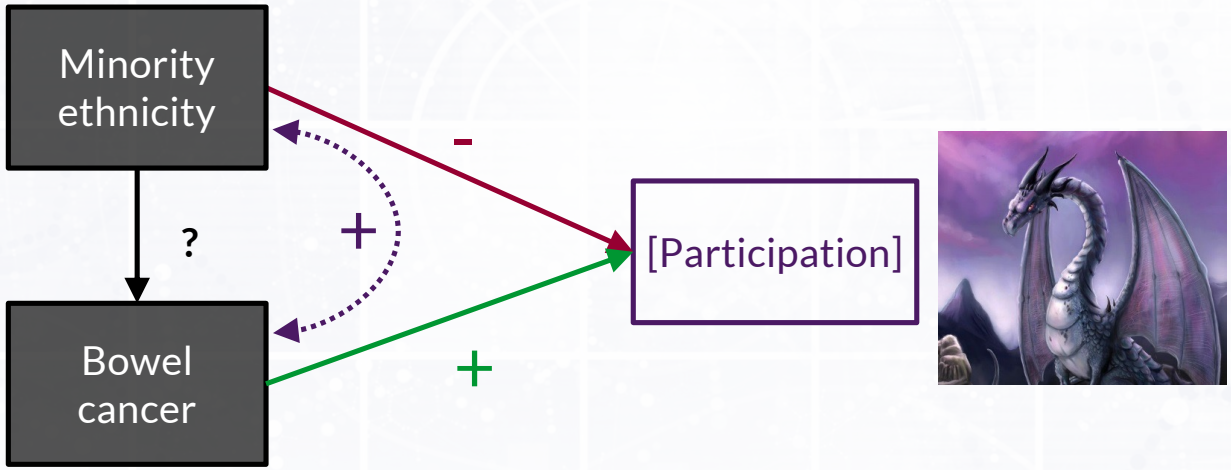
- In this scenario, diabetes appears ***protective*** of **severe UTI** – hence the **paradox**

- **Berkson's bias** is a risk for any study in health record data where the exposure and outcome directly influence 'presence' or 'measurement' within the data
- **Type 1 selection bias** is also a risk for **case-control studies**, a type of retrospective study in which participants are recruited *after the outcome occurs*
- Cases are typically keener to participate, creating a strong path between the outcome and participation
- Any path between the exposure and participation will cause selection bias



SELECTION BIAS IN RETROSPECTIVE STUDIES

- **Example:** Ethnicity and bowel cancer
  - **Minority ethnic groups** are **less likely** to participate
  - **Bowel cancer cases** are **more likely** to participate
  - Minority ethnicity appears **associated** with higher risk of bowel cancer!

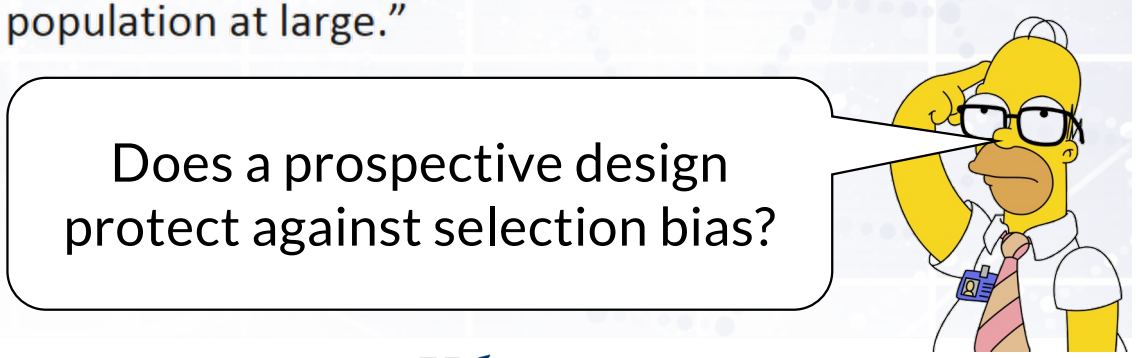


WHAT ABOUT PROSPECTIVE STUDIES?

UK Biobank is not representative of the general population on a variety of sociodemographic, physical, lifestyle and health-related characteristics, with evidence of a 'healthy volunteer' selection bias. As a result, UK Biobank is not a suitable resource for deriving generalizable disease prevalence and incidence rates. However, the large sample size and heterogeneity of exposure measures allow for valid scientific inferences of associations between exposures and health outcomes that are generalizable to the wider population.

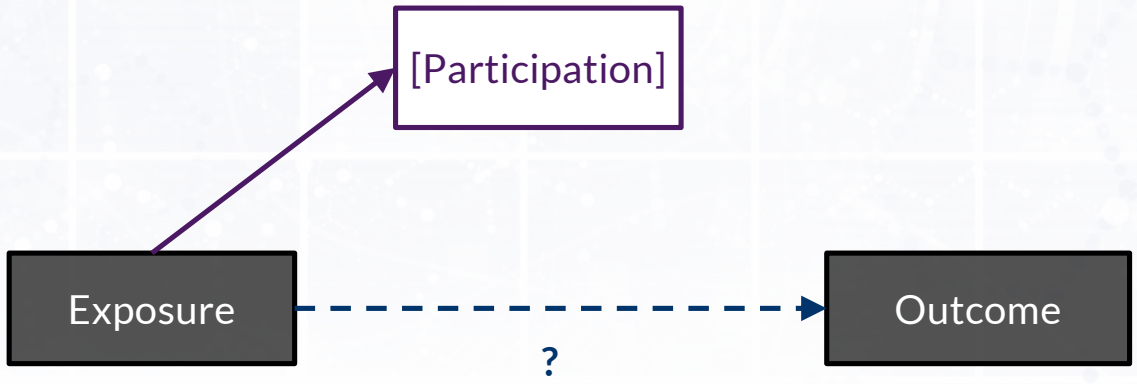
We advise that, where appropriate, publications that use UK Biobank data include a statement clarifying that "while UK Biobank participants are not representative of the general population (and hence cannot be used to provide representative disease prevalence and incidence rates), valid assessment of exposure-disease relationships are nonetheless widely generalizable and do not require participants to be representative of the population at large."

biobank uk
Improving the health of future generations

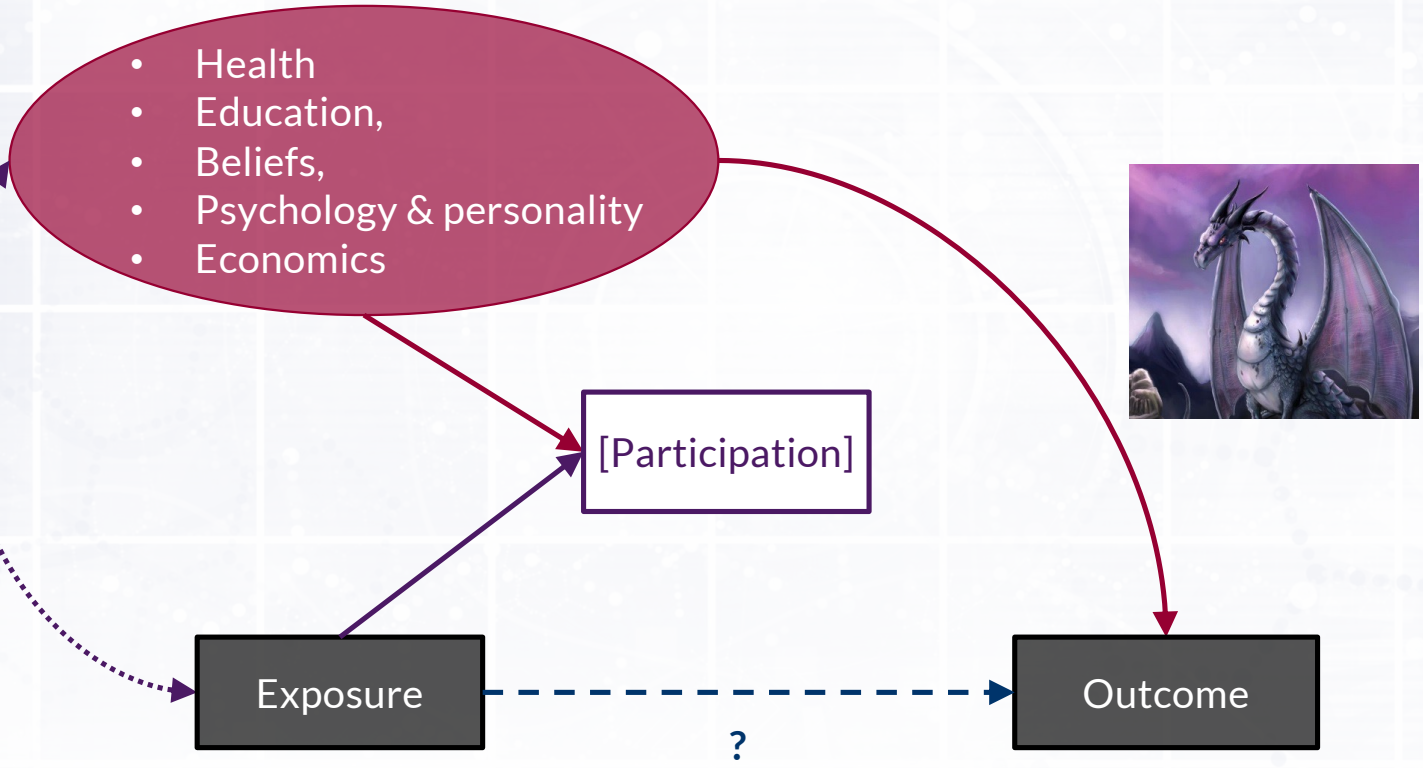


SELECTION BIAS IN PROSPECTIVE STUDIES

- In **prospective observational studies**, participation precedes the outcome and cannot therefore *cause* participation – this is certainly a strength
- However, if the exposure precedes participation, it is possible that it influences participation



- There are **many determinants of participation**
- If any of these also cause the outcome…

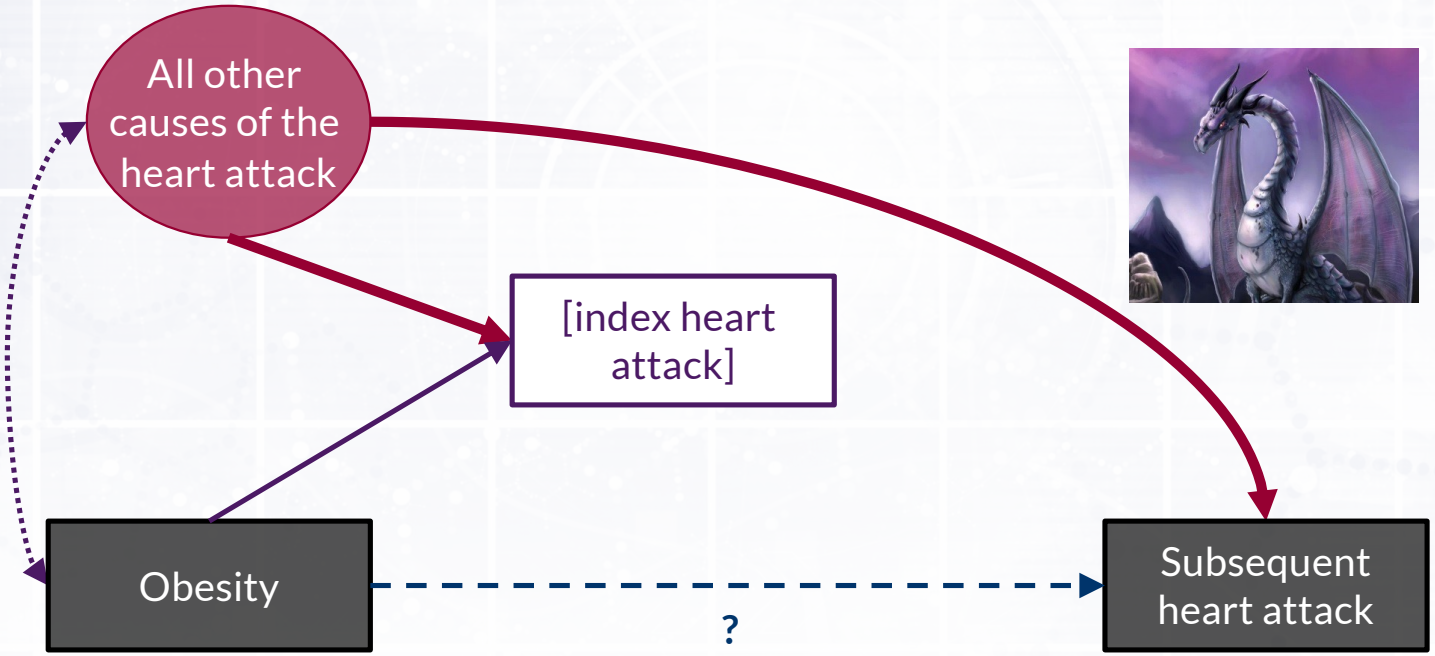


INDEX EVENT BIAS

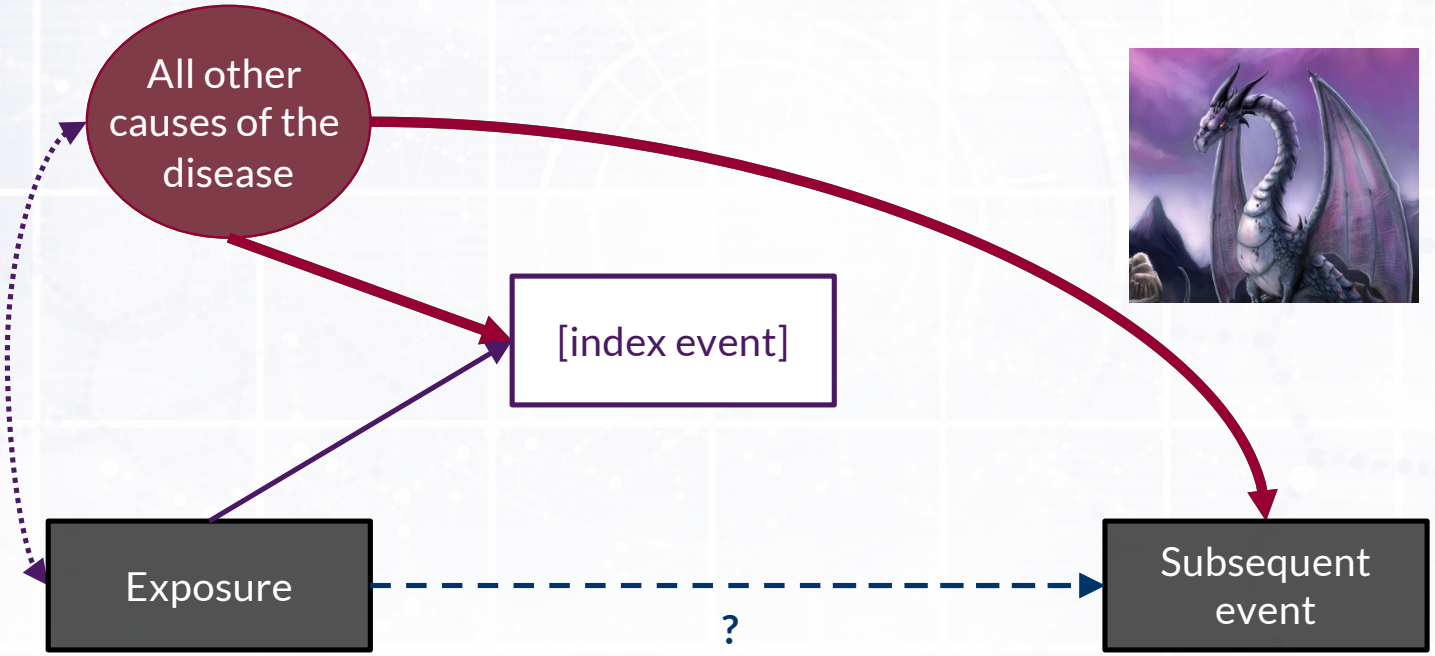
- This is particularly problematic for many clinical prognosis studies, where developing the disease is a *requirement* for participation
- If the exposure causes the 'index event' (e.g. developing the disease), there will likely be strong backdoor paths via all others causes of the disease - this is known as **index event bias**

INDEX EVENT BIAS

- **Example**: **Obesity** and **heart disease**
- **Obesity** causes **heart disease**, leading to a higher risk of a **first heart attack**
- Among **those who have had a heart attack**, effect of obesity is biased by backdoor path, perhaps even appearing protective (**obesity paradox**)

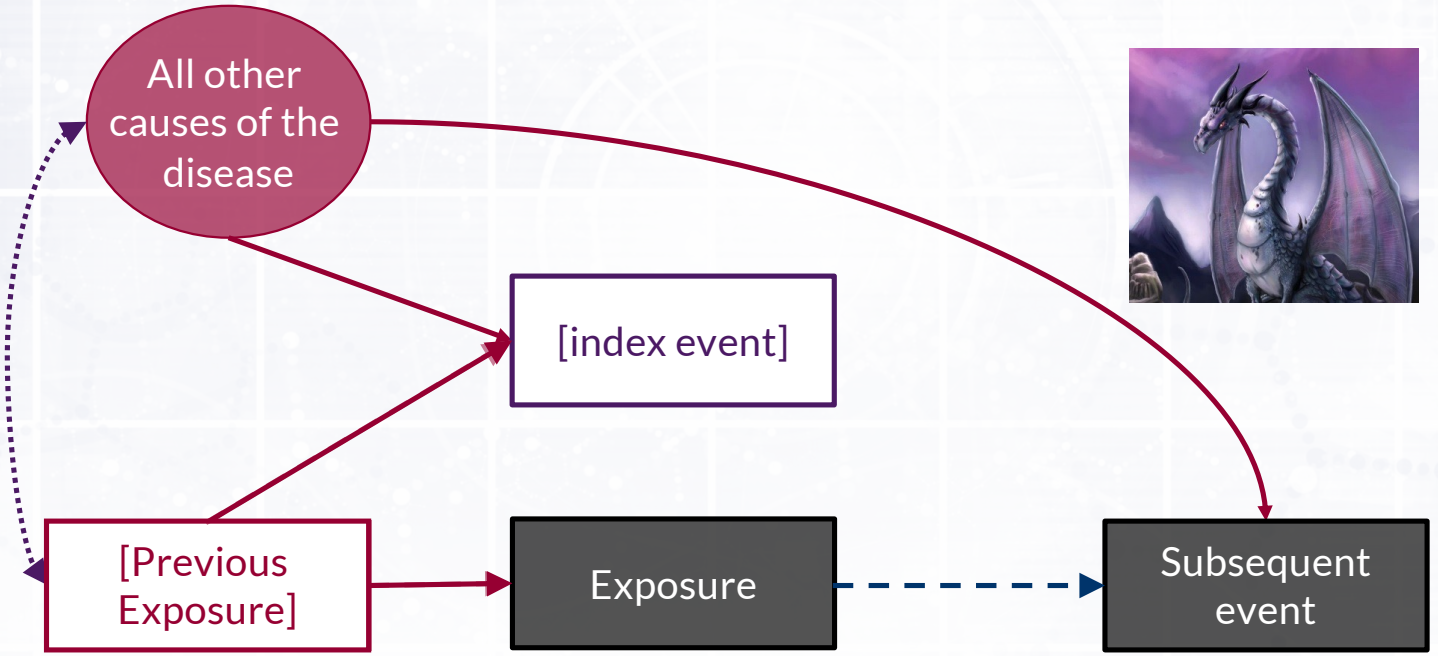


- **Solution**: The exposure must not be related to participation!
- The time zero for our exposure should occur *after* sample eligibility (i.e. after index event)

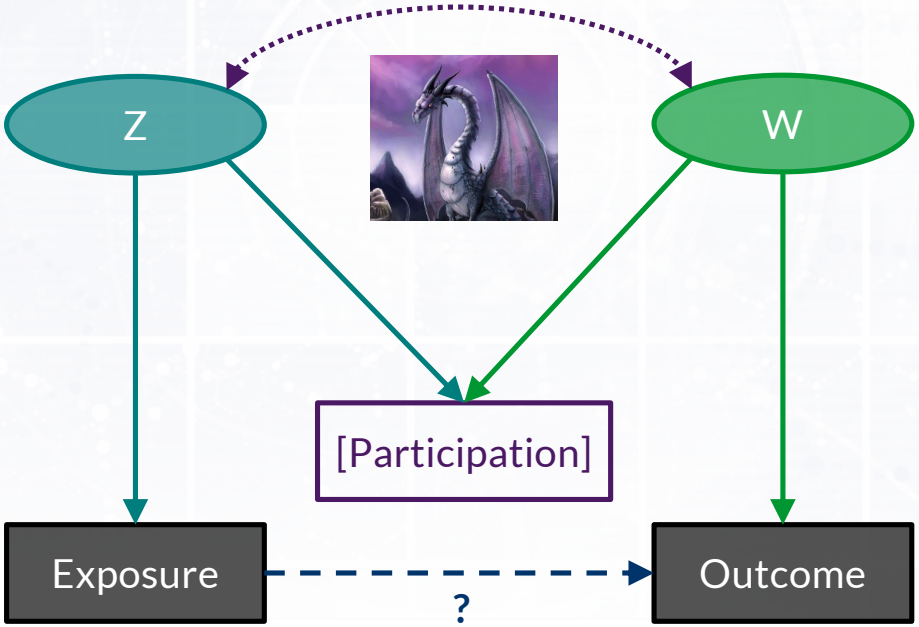


INDEX EVENT BIAS

- **Solution**: The exposure must not be related to participation!
- The time zero for our exposure should occur *after* sample eligibility (i.e. after index event)
- If your exposure is ongoing / longitudinal…
  - Could try and condition on an earlier measure and study 'change' after the index event



- Unfortunately, the exposure does not need to *cause* participation, we only need open paths between our **exposure**, **outcome**, & **participation** for potential selection bias
- E.g. If **Z** or **W** include any of **health, education, beliefs, psychology & personality, or economics** then there's the potential for collider bias via an **M-bias** pattern

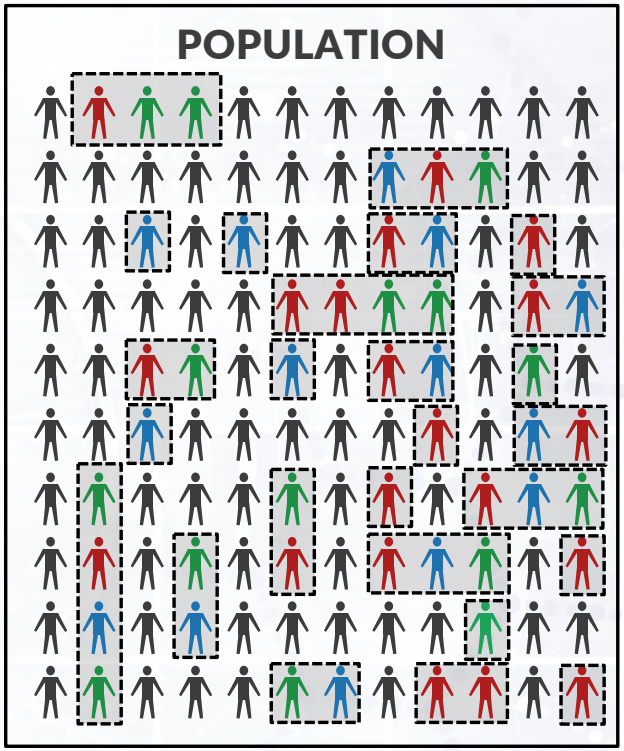


- There are several ways to reduce selection bias

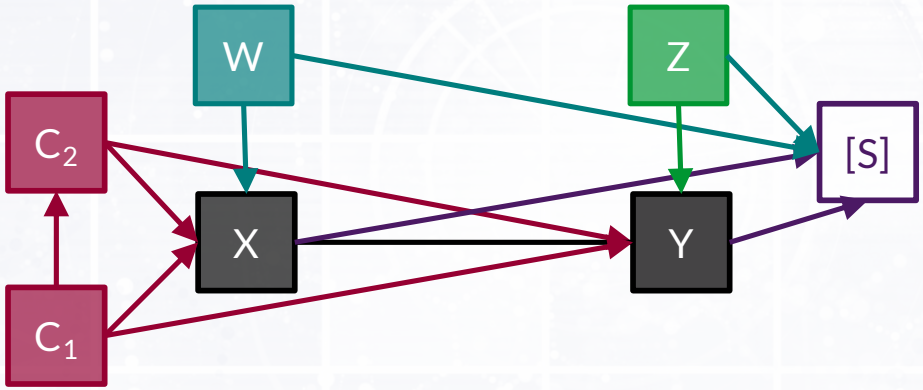
1. **Reduce selection bias by design with a thoughtful sampling strategy**
2. **Consider the determinants of selection**
3. **Condition on appropriate variables**
4. **Reweight your sample**

REDUCE SELECTION BIAS BY DESIGN

- Where possible, design the sampling strategy to take a representative (or known) sample of the target population
  - E.g. A random sample (such as 1958 National Child Development Study)
  - E.g stratified cluster sampling (2000 Millenium Cohort Study)
- Focus on maximizing participation across all key strata, particularly 'hard to reach' groups
- Consider **two-stage recruitment**, collecting basic information for all people first, before recruiting to the more burdensome study



- If your target population is not the general population, then you need to carefully consider who is and is not in your sample
  - This includes almost all routine data sources, which are typically highly selected
- The best way to do this is to include the selection process in your DAG

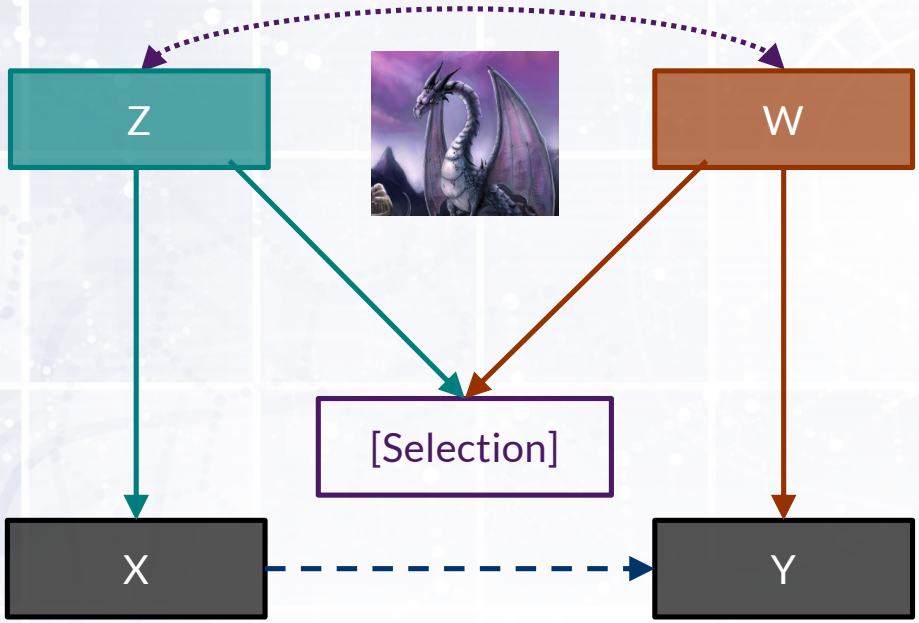


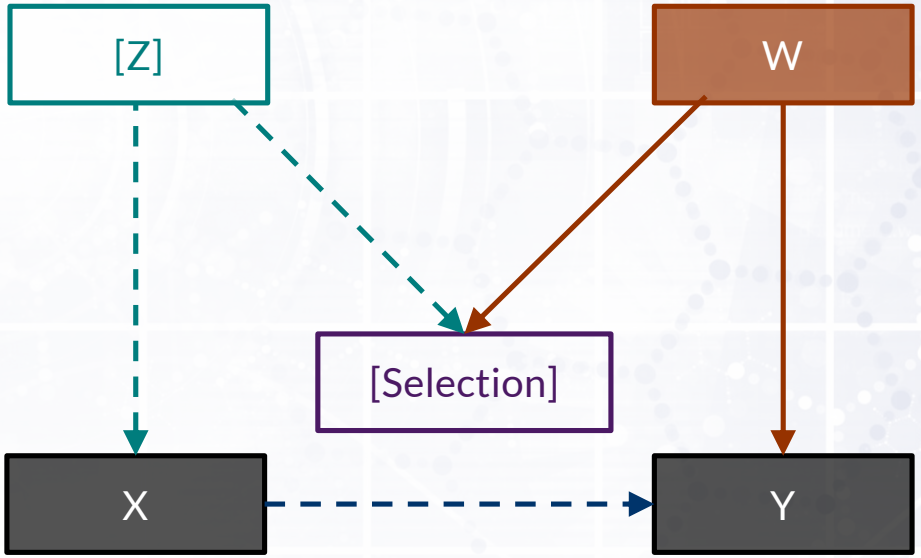
- Your estimates will experience type **1 selection bias** if the outcome causes selection or your exposure and outcome are otherwise connected by a path through selection

CONDITION ON APPROPRIATE VARIABLES

Careful conditioning can help reduce selection bias

- Conditioning on nodes on **collider paths** can reduce **Type 1 selection bias**



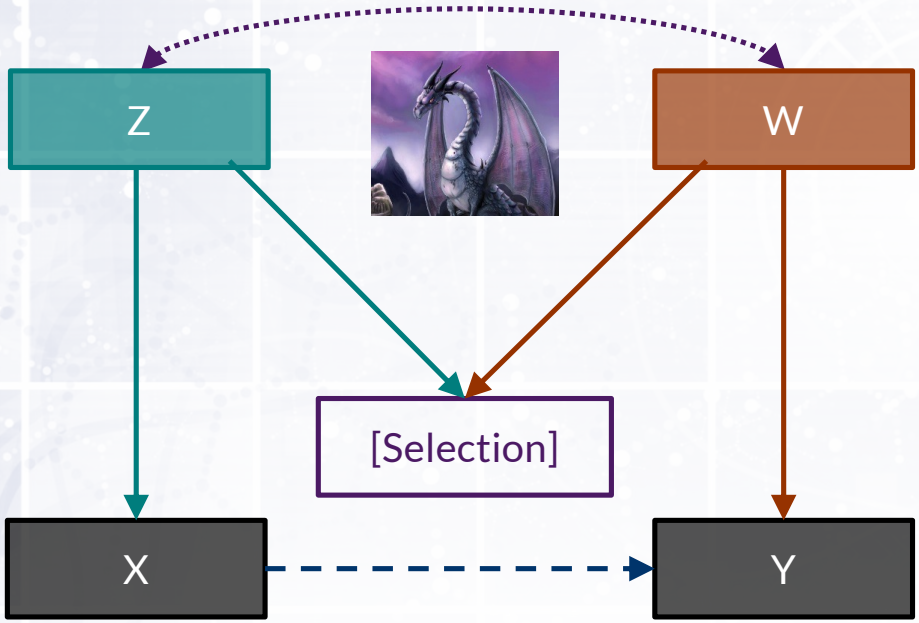


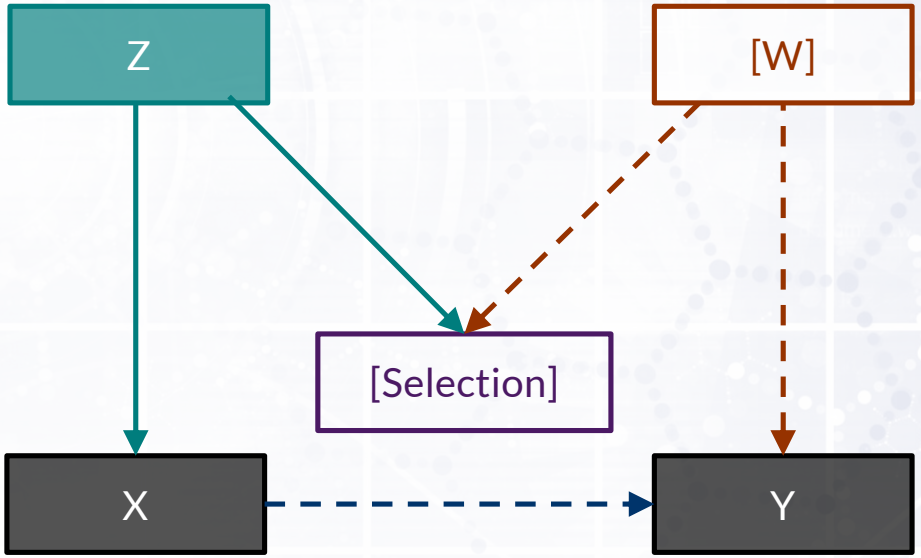
Type-1 selection bias eliminated by closing **X←[Z]→[Selection]←W→Y**

CONDITION ON APPROPRIATE VARIABLES

Careful conditioning can help reduce selection bias

- Conditioning on nodes on **collider paths** can reduce **Type 1 selection bias**
- If these nodes are also **effect modifiers** they could also help reduce **Type 2 selection bias**
- However, you must take care not to open other collider paths!

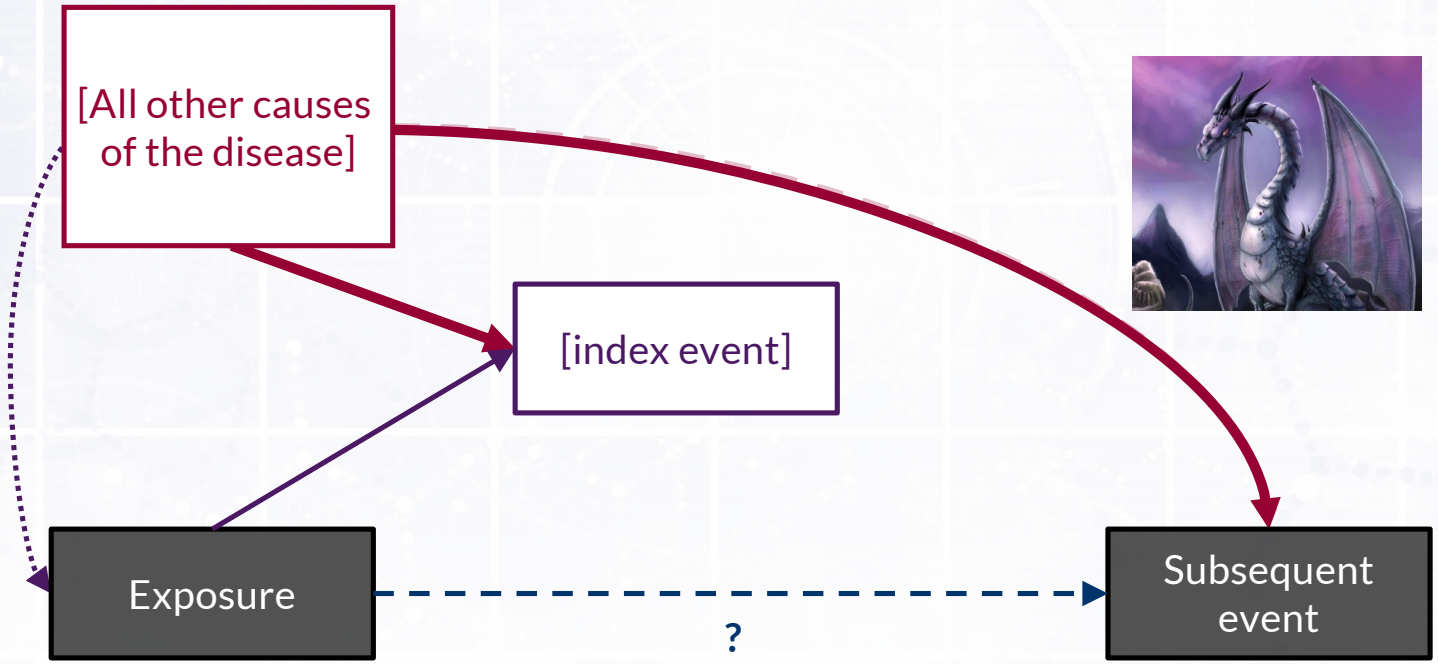




Both **Type 1** and **Type-2 selection bias** eliminated by closing **[W]→Y**

CONDITION ON APPROPRIATE VARIABLES

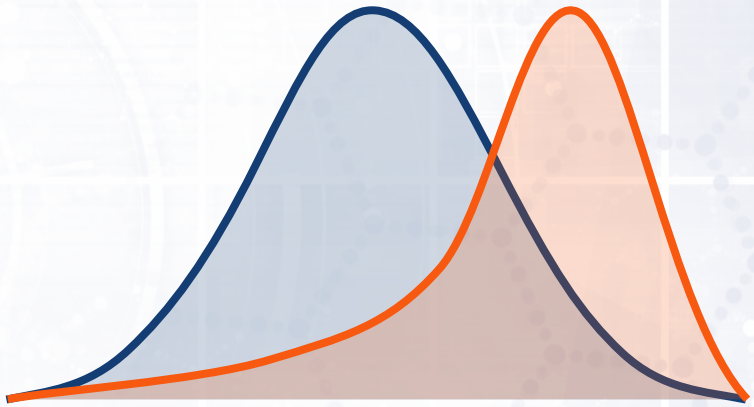
- Some of these selection path nodes will be picked up and conditioned by your confounding adjustment approach
- But, depending on the structure, you may still need to measure and condition on a lot of variables that are not strictly confounders!



- Conditioning therefore offers less help when the exposure causes selection directly

- An alternative approach is to reweight the sample (using IPSW) to produce a **pseudo-population** in which the probability of selection is unrelated to the exposure

- Reweighting is generally considered easier than conditioning when you need to account for a large number of variables



- However, it requires external information on the profile of non-participants, and sufficient diversity in your own sample to avoid extreme weights

A WARNING

- It is not always possible to ‘fix’ selection bias after the data have been collected
- If selection is determined by the outcome, an unbiased estimate cannot be obtained without external data.
  - Retrospective studies are therefore more vulnerable to severe and unresolvable selection bias
- Where selection is *impossible* in certain groups, then again this cannot be recovered
- We need to learn to recognise unsuitable data sources and/or avoid such problems during study design

SUMMARY

- Selection bias occurs when there is a divergence between the true causal effect in the target population and the estimate obtained in the analytical sample
  - **Type 1 selection bias** – also known as **collider bias** damages **internal validity**
  - **Type 2 selection bias** – also known as **generalizability bias** damages **external validity**
- The size of selection bias will depend on many factors, but studies where the exposure or outcome cause participation or with extreme non-response are likely most affected
- You should always think about the determinants of selection into your data and try to avoid selection bias by design
  - DAGs with selection nodes can be helpful
- You may be able to reduce selection biases by carefully conditioning on appropriate variables or using inverse probability weights,
  - This requires that selection is still possible within all relevant strata