Collider scope: when selection bias can substantially influence observed associations

Marcus R Munafò , Kate Tilling, Amy E Taylor, David M Evans, and George Davey Smith

Presentation by
Lucrezia Patruno, Manuel Vimercati, Francesco Craighero

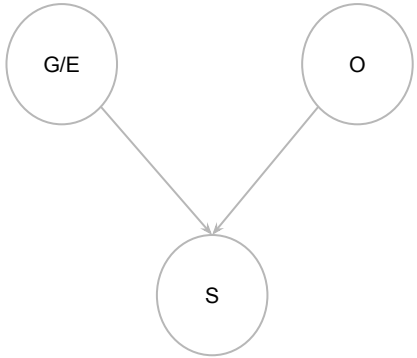
Paper Focus

Large scale cross sectional and cohort studies aim at finding **associations between variables** such as genetic/environmental factors (G/E) and a specific outcome (O).

Samples in association studies must be representative of the intended study population.

However, the same variables may also contribute in the predisposition to be part of the sample (S=1). This phenomenon is called **selection bias** and happens for different reasons:

- Unrepresentativeness of the population at inception.
- Attrition from the study (also influenced by the same variables).
- A subset of the original sample is selected for further analysis.



Paper Focus

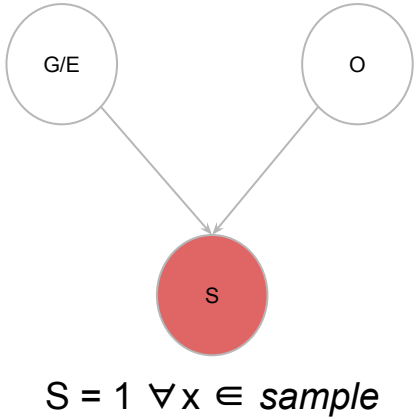
Selection bias results in **collider bias**.

General opinion: collider bias has an effect on

(1) representativeness

(2) prevalence estimates,

but is negligible in association studies.



Paper Focus

Selection bias results in **collider bias**.

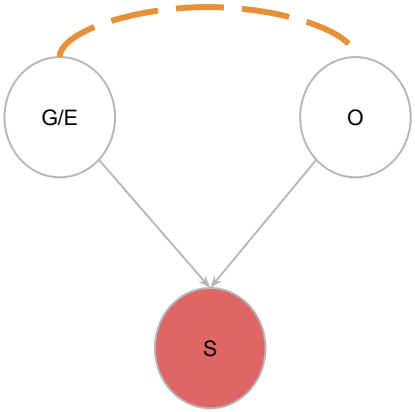
General opinion: collider bias has an effect on

(1) representativeness

(2) prevalence estimates,

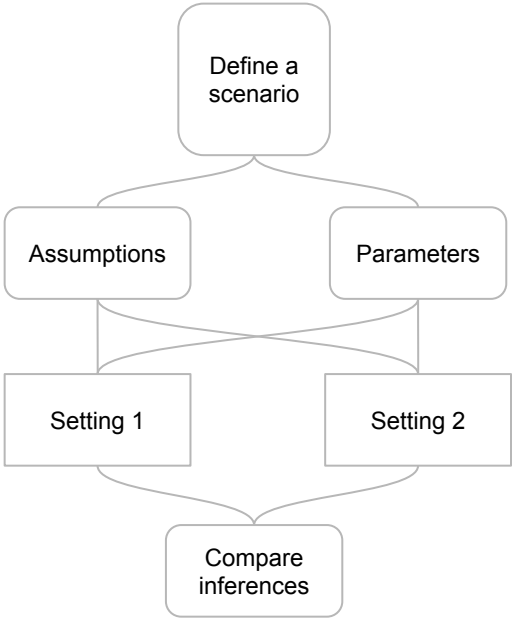
but is negligible in association studies.

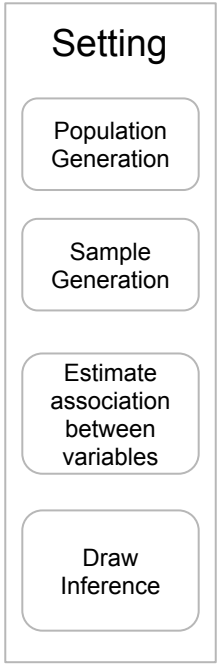
Claim of the authors: collider bias may result in the identification of **spurious/biased associations**.



$S = 1 \ \forall x \in sample$

Study the collider bias with simulations





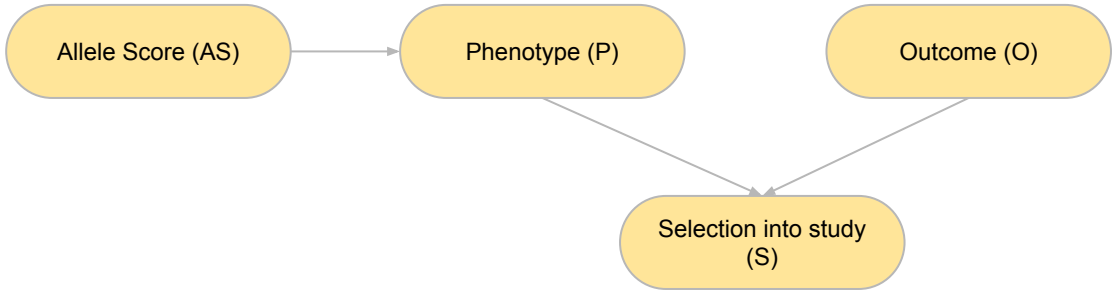
Design the Simulation

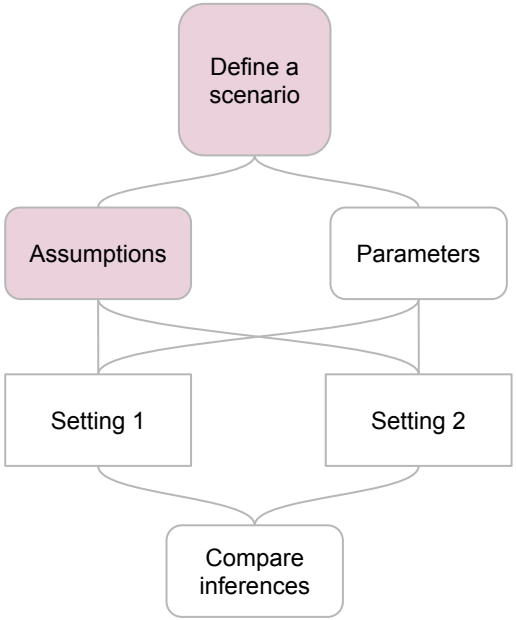
Scenario: simulated allele score (***AS***), phenotype (***P***) and outcome (***O***):

- ***P*** and ***O*** influence selection (***S***) into the study.
- NO association between ***AS*** and ***O*** in the population.
- Population of 9,000,000 individuals.
- Selection ***S*** of of 500,000 individuals.

Assumptions:

- All variables are normally distributed with σ = 1.
- ***P*** and ***O*** have independent (and equal) effects on ***S***.





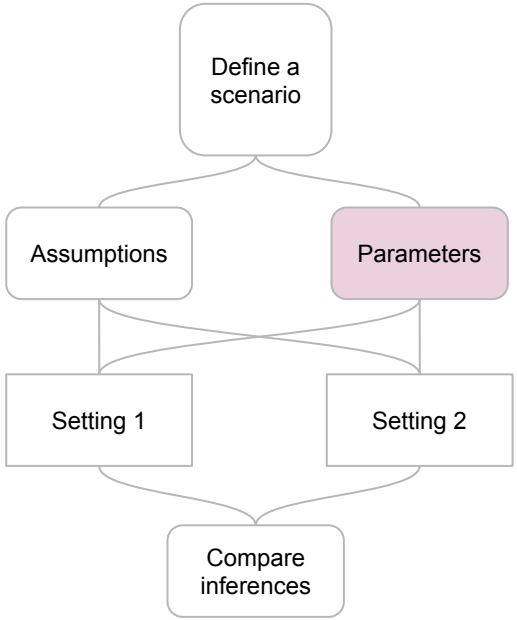
Settings Parameters

Each setting is identified by three parameters, i.e. ***OR***, ***r***, ***C***:

- Association between P/O and S constrained by the odds ratio (***OR***) for missingness:

$$OR = \frac{P(S = 0 \mid P = p + \sigma)}{P(S = 0 \mid P = p)} = 1.2, 1.5, 1.8$$

  where σ is the standard deviation.



Settings Parameters

Each setting is identified by three parameters, i.e. ***OR***, ***r***, ***C***:

- Association between P/O and S constrained by the odds ratio (***OR***) for missingness:

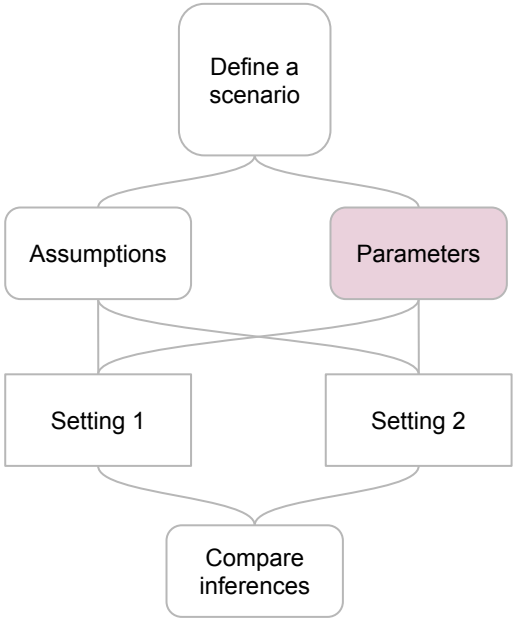
$$OR = \frac{P(S = 0 \mid P = p + \sigma)}{P(S = 0 \mid P = p)} = 1.2, 1.5, 1.8$$

  where σ is the standard deviation.

- The correlation ***r*** between AS and P:

$$\boldsymbol{r}(AS, P) = 0.05, 0.1, 0.15, 0.20, 0.30$$

  Correlation was set to mimic the real values found in a real dataset of the UK biobank.



Settings Parameters

Each setting is identified by three parameters, i.e. ***OR***, ***r***, ***C***:

- Association between P/O and S constrained by the odds ratio (***OR***) for missingness:

$$OR = \frac{P(S = 0 \mid P = p + \sigma)}{P(S = 0 \mid P = p)} = 1.2, 1.5, 1.8$$

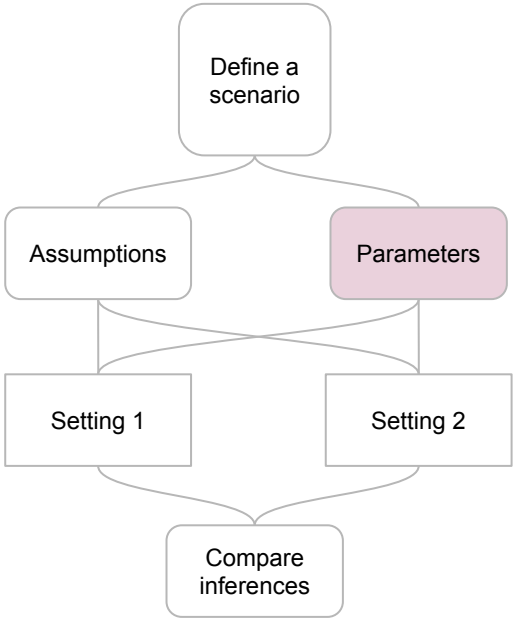
  where σ is the standard deviation.

- The correlation ***r*** between AS and P:

$$\boldsymbol{r}(AS, P) = 0.05, 0.1, 0.15, 0.20, 0.30$$

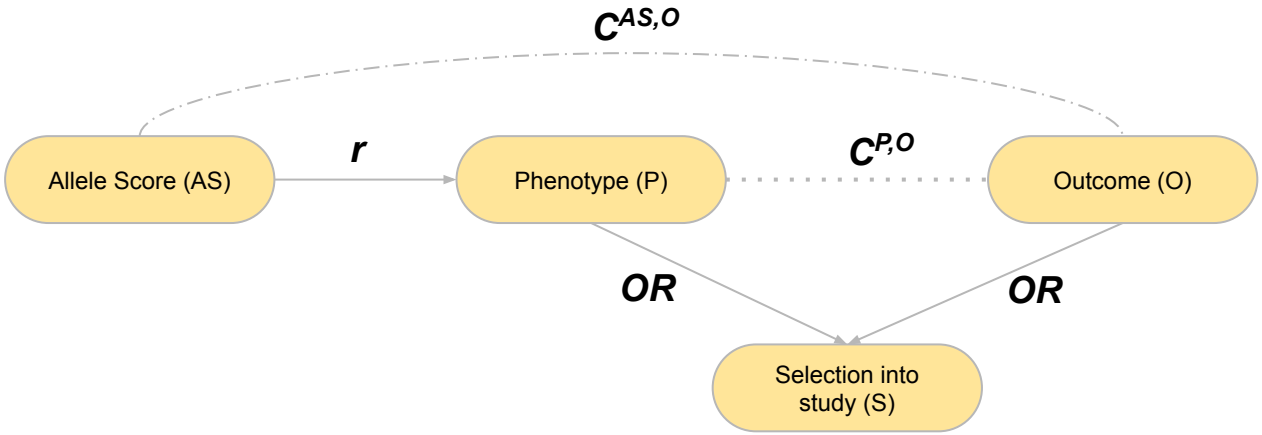
  Correlation was set to mimic the real values found in a real dataset of the UK biobank.

- Causal effect of P on O in the original population:
  - Null association, regression coefficient ***C*** = 0. (Setting 1)
  - Positive association, ***C*** = 0.1. (Setting 2)



Final inference and Experiment Recap

Check if the association between ***AS*** and ***O*** exists in both settings,
i.e. if the correlation coefficient $C^{AS,O}$ is not null in the selected sample with 500,000 elements.



Experimental Settings

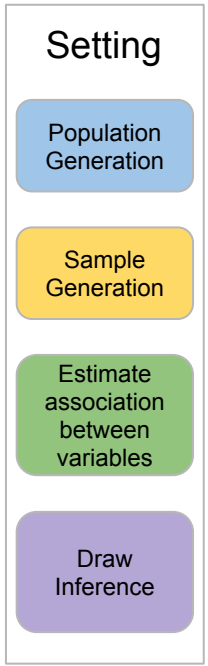
C = 0, forall combinations of ***OR*** and ***r***

C = 0.1, forall combinations of ***OR*** and ***r***

Sample the population 100 times

Compute the regression coefficients between ***AS*** and ***O***
and the respective 95% confidence intervals

Check the existence of an association between ***AS*** and ***O***,
i.e. the confidence interval of the the regression coefficient does not contain 0

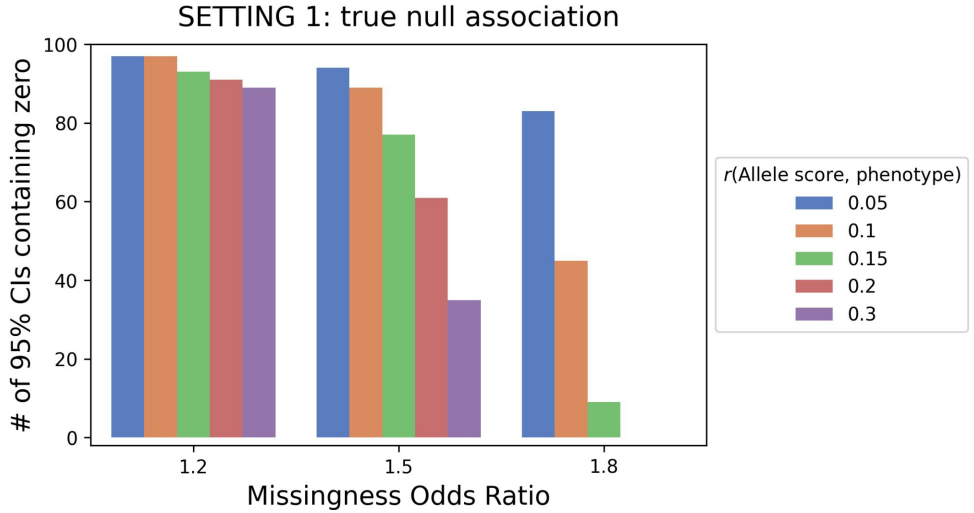


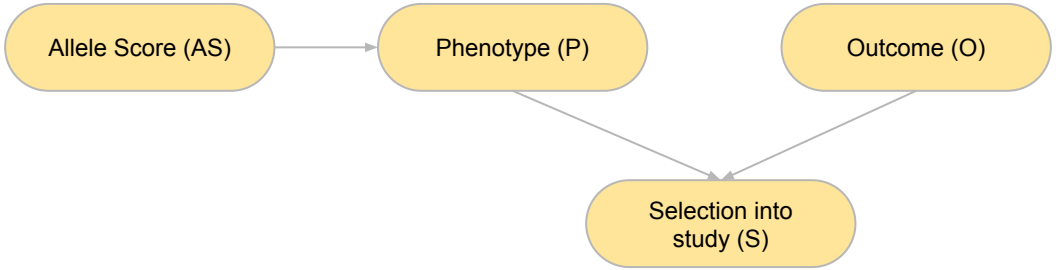
SETTING 1: true null association ($C^{P,O} = 0$)

Two trends:

- As **r increases**, the number of CIs containing 0 decreases.
- As **OR increases**, the number of CIs containing 0 decreases

For high values of both **OR** and **r** (1.8 and [.2, .3] respectively), a spurious association is always found.

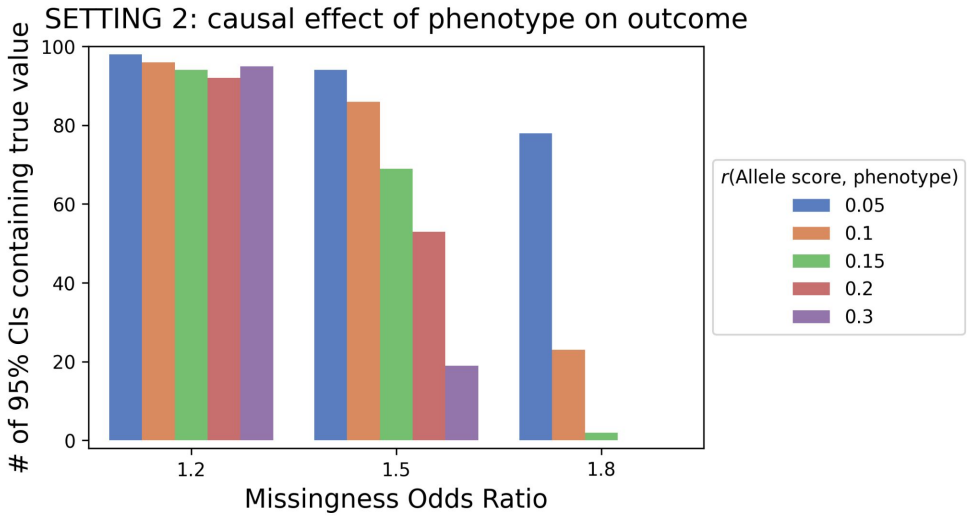


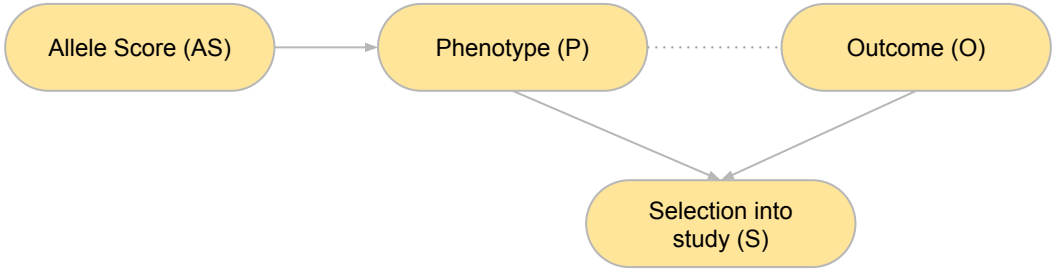


SETTING 2: $C^{P,O} = 0.1$

The **same behaviour** is observed also when an association actually exists in the original whole population.

- As **r increases**, the number of CIs containing 0 decreases.
- As **OR increases**, the number of CIs containing 0 decreases



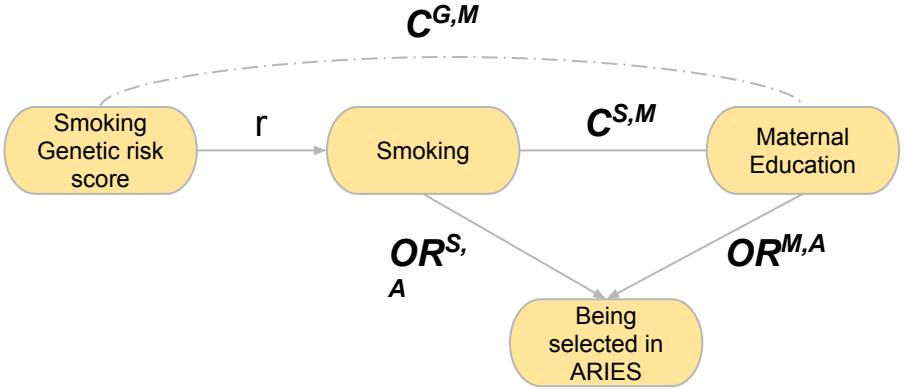


Real-word example: ALSPAC and ARIES

ALPSAC [1]: cohort of mothers-children pairs.

ARIES [2]: subset of ALPSAC. Mothers-children were selected based on the availability of DNA samples (2 for the mother).

Three variables in the study: genetic risk score, smoking (ever, never) and maternal education.

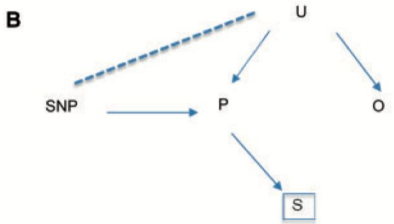


Association	OR	pval
r	1.07	0.003
$OR^{S,A}$	0.59	<0.001
$OR^{M,A}$	1.86	<0.001
$C^{S,M}$ (ALSPAC)	0.45	<0.001
$C^{S,M}$(ARIES)	0.61	0.003
$C^{G,M}$ (ALSPAC)	1.01	**0.74**
$C^{G,M}$ (ARIES)	1.20	**0.03**

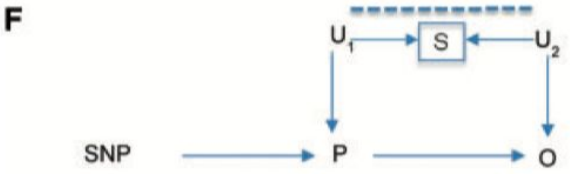
[1] Fraser, Abigail, et al. "Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort." *International journal of epidemiology* 42.1 (2013): 97-110.
[2] Relton, Caroline L., et al. "Data resource profile: accessible resource for integrated epigenomic studies (ARIES)." *International journal of epidemiology* 44.4 (2015): 1181-1190.

Other cases with collider bias

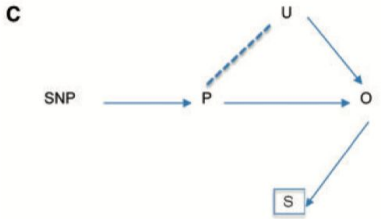
- P, O associated by environmental factors U
- P=“smoking”, O=”Alzheimer”, P→ S=”mortality”

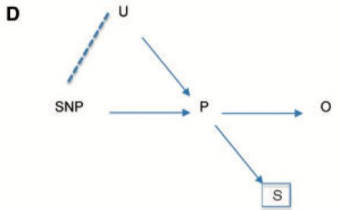


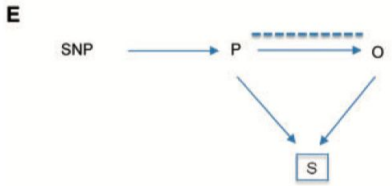
- P,O association: **biased**
- SNP, O association: **unbiased**



- SNPs directly cause O
- SNP, O association: **biased**
- P, O association in C, E: **biased**







Conclusions

Collider Bias

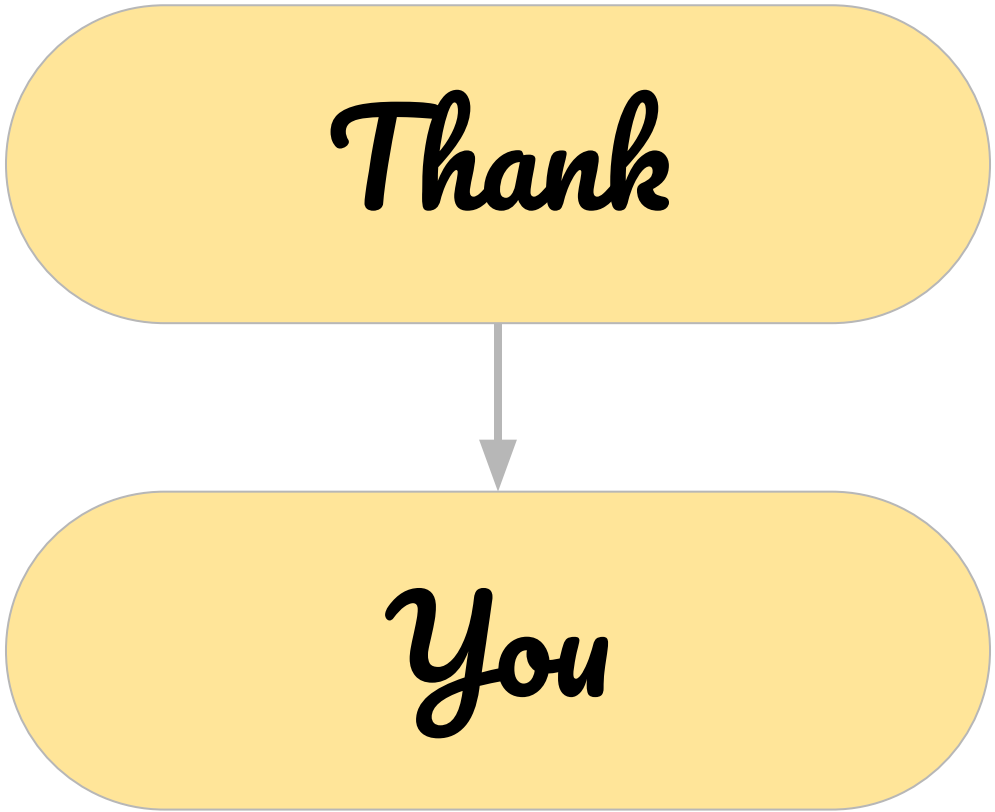
Selection/attrition can induce misleading associations

Real-world scenarios

Studies with polygenic scores are most at risk of producing misleading results

Representativeness

- It is important having representative cohorts
- Baseline data availability to investigate selection


Thank
You