Data Week Online 2020

Making data work for everyone

bristol.ac.uk/golding

Data Week Online 2020

The Jean Golding Institute

- A central hub for data science and data-intensive research
- One of 5 University of Bristol research institutes
- Connect multidisciplinary experts across the University and beyond
- Events, training, funding, Ask JGI, The Alan Turing Institute

Our priorities

1. Societal challenges
2. Data visualisation
3. Reproducibility & data governance
4. Fundamental research

Making data work for everyone

bristol.ac.uk/golding

Data Week Online 2020

Date	Event	Speaker
Monday 15 June	Data science and COVID 19 & Data Week Introduction	Kate Robson Brown, JGI Director
Monday 15 June	Intermediate Python	Advanced Computing Research Centre
Tuesday 16 June	Talk: Working at and with The Turing Institute: experiences as a Fellow	Jon Crowcroft, Turing Fellow & University of Cambridge
Tuesday 16 June	Talk: increasing engagement with data	Michael Green, Luna 9
Tuesday 16 June	Introduction to data analysis in Python	Advanced Computing Research Centre
Wednesday 17 June	Do you want to be a data Rockstar?	Luke Stoughton, The Information Lab
Wednesday 17 June	Applied data analysis in Python	Advanced Computing Research Centre
Thursday 18 June	Talk: New data on COVID-19 is undermined by old statistical problems	Gibran Hemani, University of Bristol
Thursday 18 June	Managing sensitive research data: from planning to sharing	Library Research Services
Thursday 18 June	Introduction to deep learning	Advanced Computing Research Centre
Friday 19 June	Deep Learning for Health and Life Sciences	Valerio Maggio, University of Bristol
Friday 19 June	Tour of the Tidyverse	Max Kronborg, Mango Solutions
Friday 19 June	Best practices in software engineering	Advanced Computing Research Centre

Making data work for everyone

bristol.ac.uk/golding

New data on Covid-19 is undermined by old statistical problems

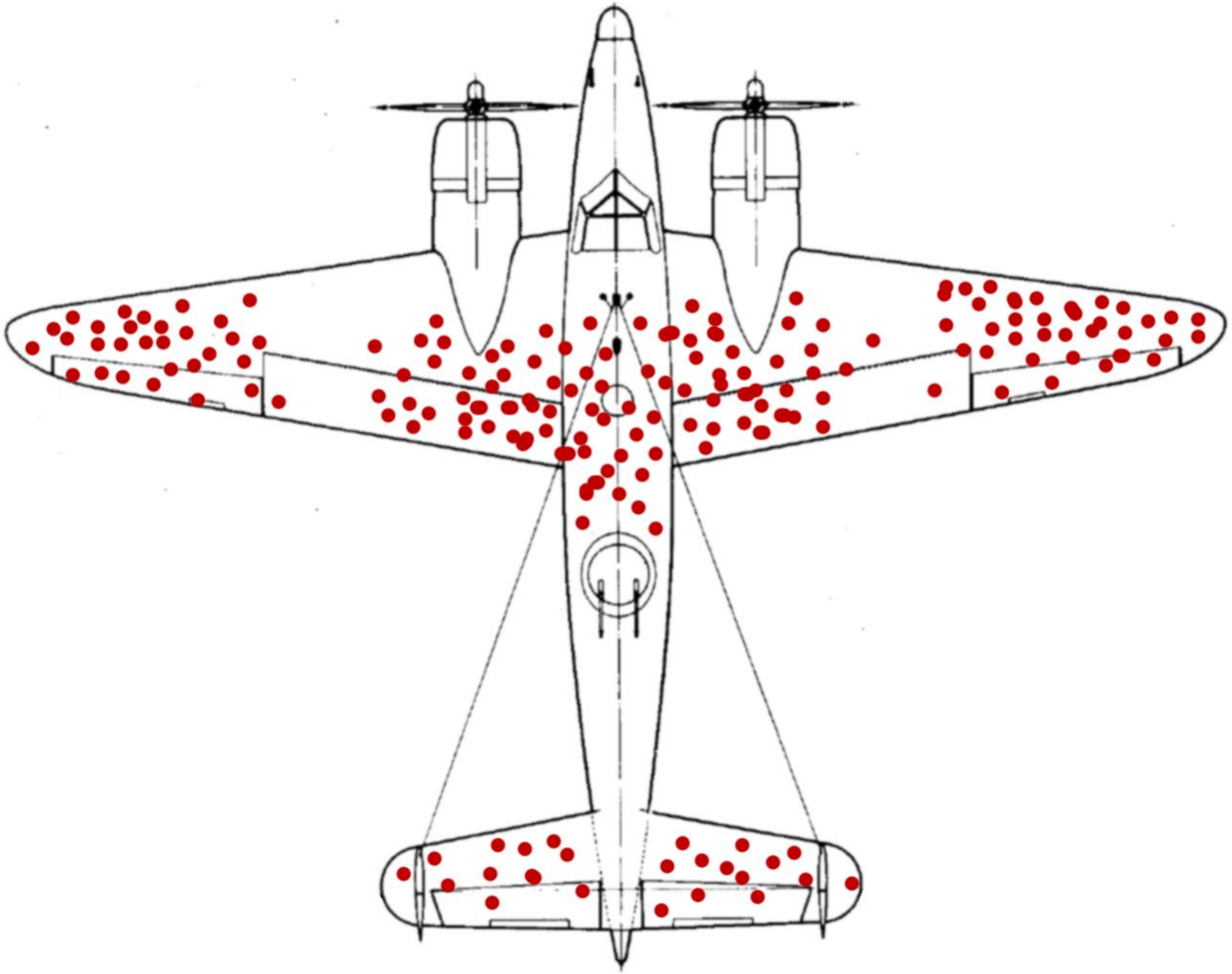
Gibran Hemani

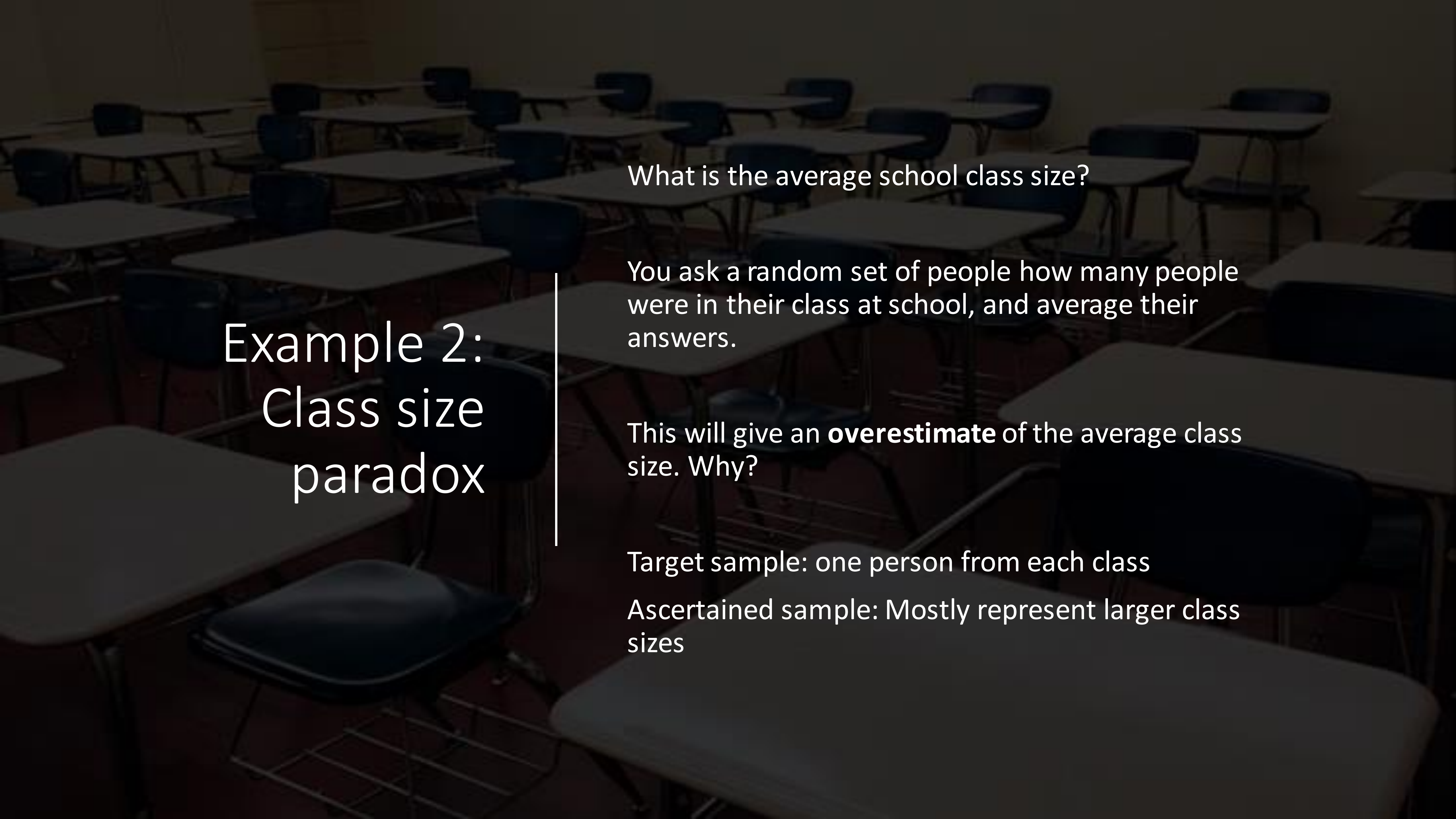
MRC Integrative Epidemiology Unit, University of Bristol

Outline

- Selection bias
- COVID-19
  - Prevalence
  - Causal factors
  - Predictors
- Mitigating the problem

Example 1: Where
to put the armor?





Example 2: Class size paradox

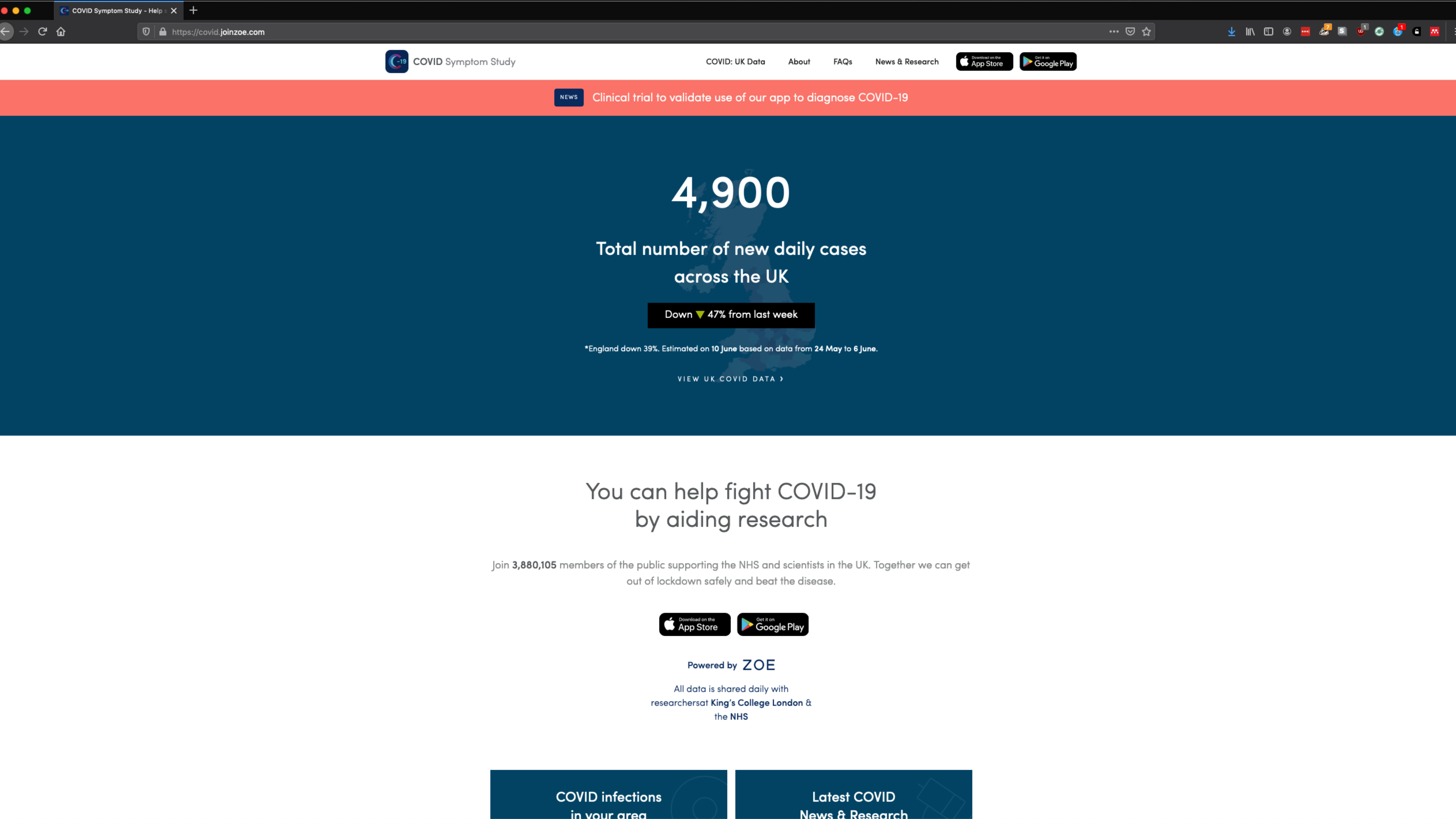
What is the average school class size?

You ask a random set of people how many people were in their class at school, and average their answers.

This will give an **overestimate** of the average class size. Why?

Target sample: one person from each class

Ascertained sample: Mostly represent larger class sizes



Clinical trial to validate use of our app to diagnose COVID-19

4,900

Total number of new daily cases across the UK

Down ▼ 47% from last week

*England down 39%. Estimated on **10 June** based on data from **24 May** to **6 June**.

VIEW UK COVID DATA ›

You can help fight COVID-19 by aiding research

Join **3,880,105** members of the public supporting the NHS and scientists in the UK. Together we can get out of lockdown safely and beat the disease.

Powered by ZOE

All data is shared daily with
researchersat **King's College London** &
the **NHS**

COVID infections in your area

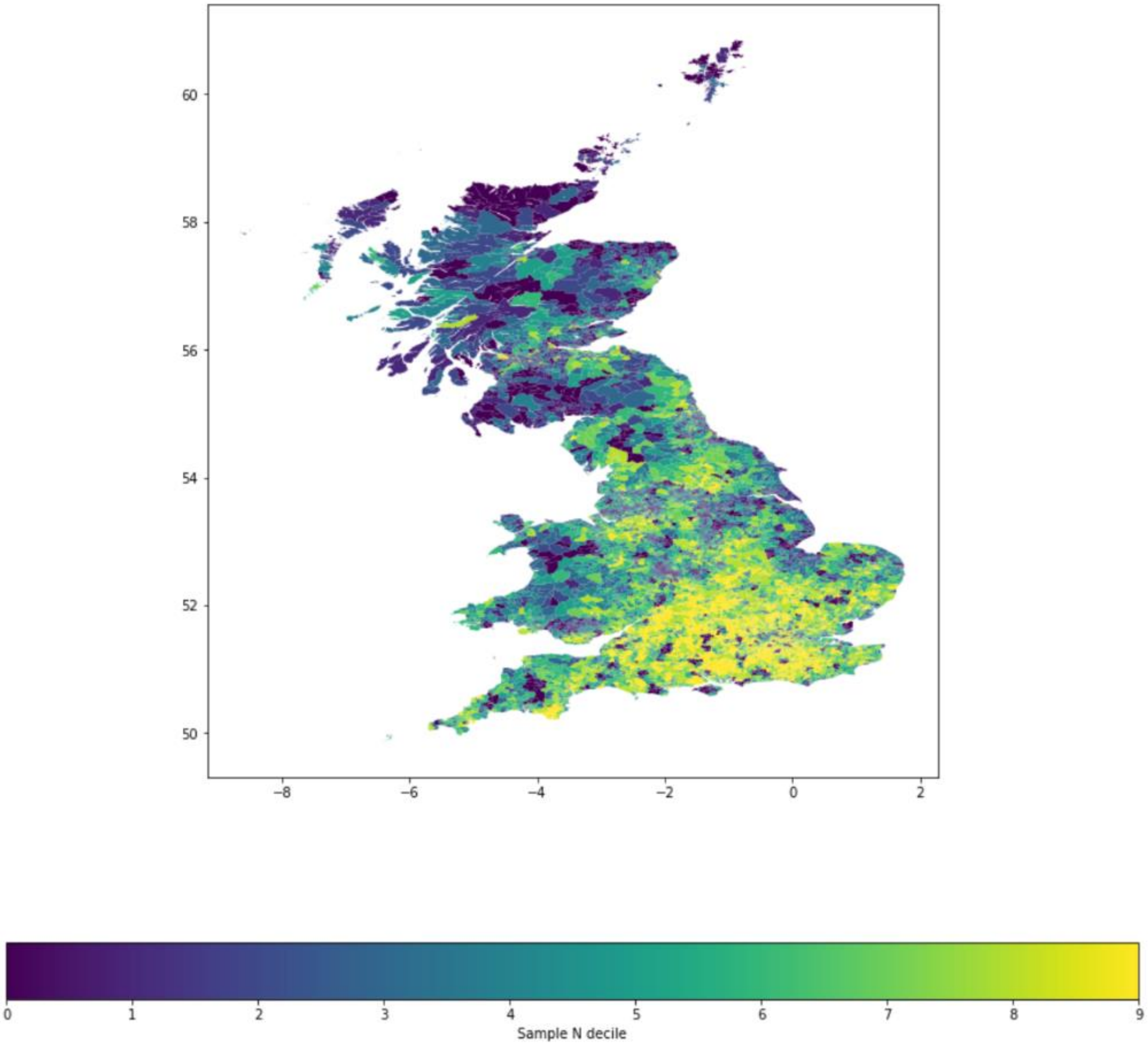
Latest COVID News & Research

Epidemiology team, assemble!

- Gibran Hemani
- Neil Davies
- Gemma Sharp
- Gareth Griffith
- Annie Herbert
- Tim Morris
- Amanda Hughes
- Ruth Mitchell
- Luisa Zuccolo
- Giulia Mancano
- Zoe Reed

MRC | Integrative Epidemiology Unit

Geographic patterning of app usage

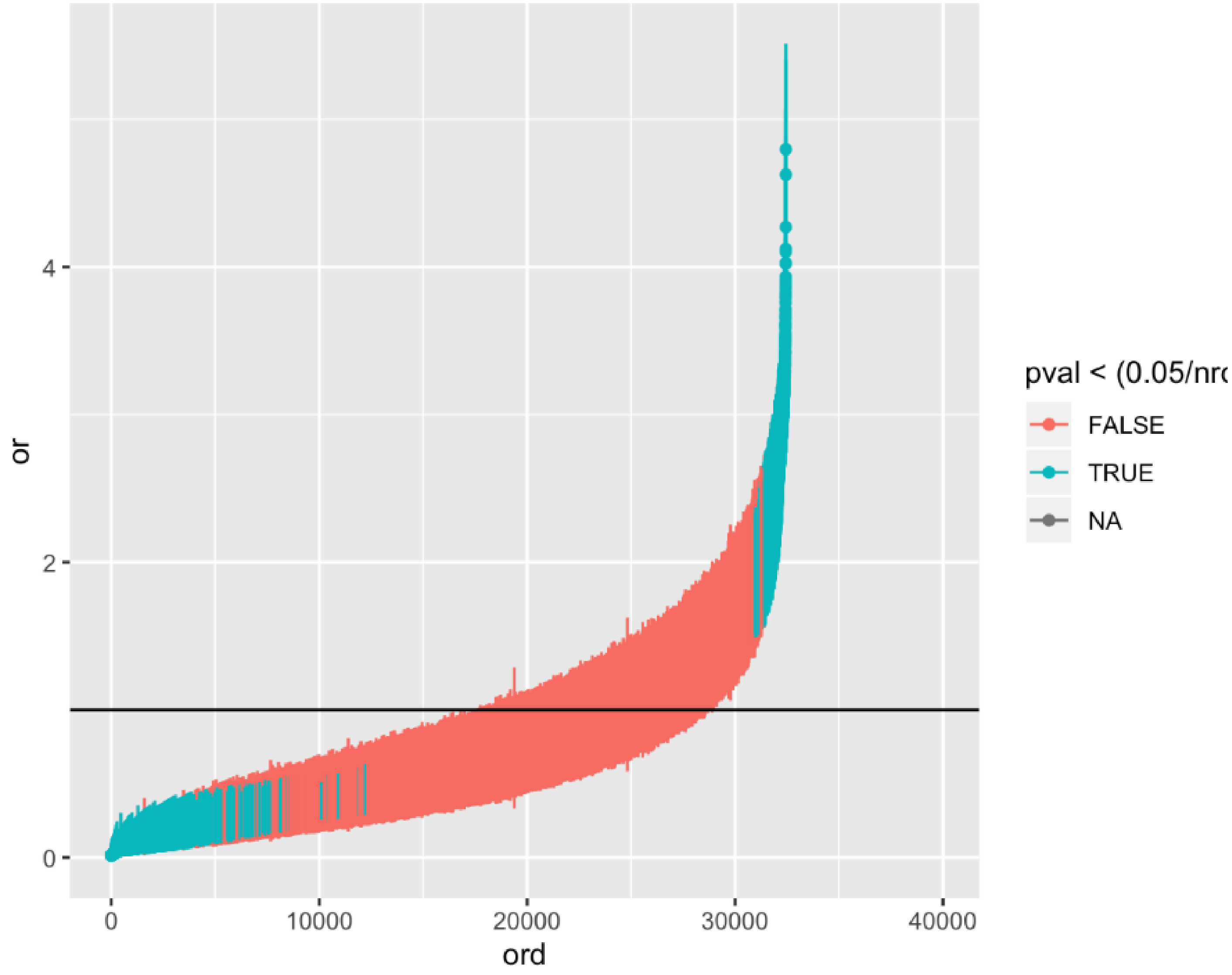


Analysis performed by Gareth Griffith

Socioeconomic status of region explains 20% of the geographic ascertainment

Take home message:

- the samples in the data are not representative of the general population
- All the ways in which they are not representative is not known.
- The rest of this talk is to explore why this should be taken seriously



Estimates of COVID-19 disease prevalence

3-11th May, UK

ONS survey:	**133k (95%CI: 62k to 250k)**
COVID symptom tracker:	**225k (95%CI: 210 to 240)**
Difference:	**92k (95%CI: 19k to 165k)**

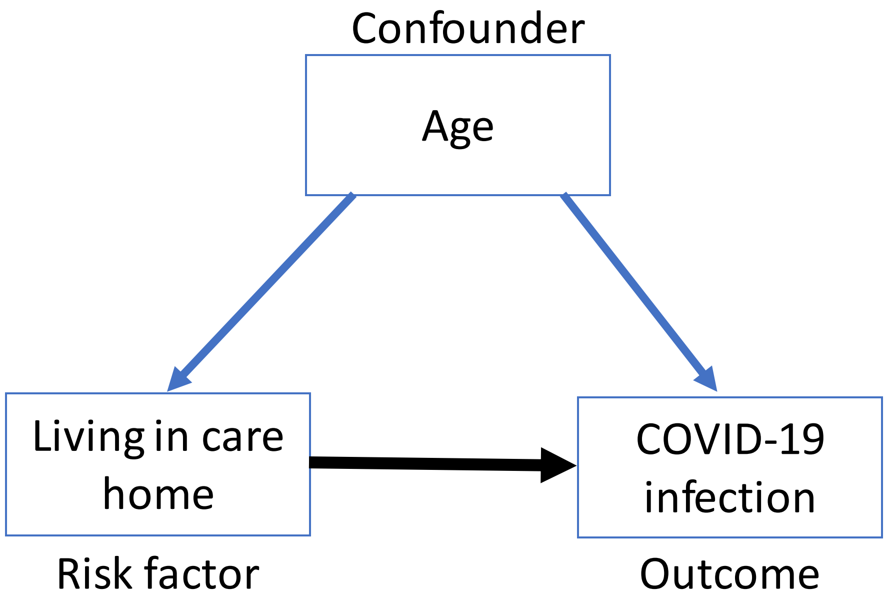
App only estimates **symptomatic infections**,
whereas ONS estimates **all infections**

Assuming 30% of infections are symptomatic: App data over-represents COVID infections by **5-6x** compared to the general population

Analysis performed by Neil Davies

Risk factors for COVID-19

Directed acyclic graphs (DAGs)



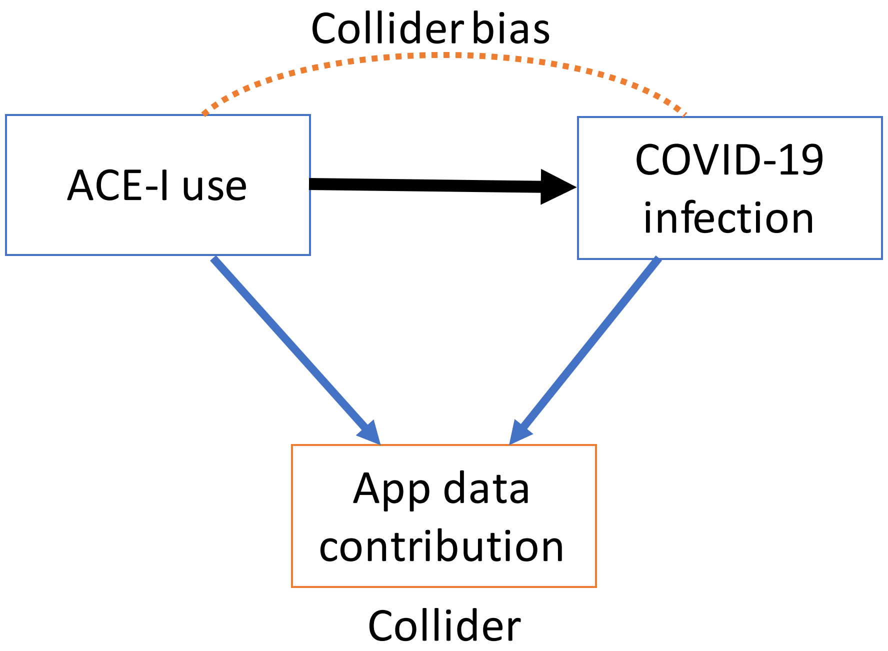
An arrow represents a (causal) association

Either because you believe it exists, or because you are hypothesizing it exists

DAGs are used to lay out what your assumptions are when conducting an analysis

Often the absence of an arrow is very informative in terms of modelling assumptions

Do ACE inhibitors increase risk of COVID-19 infection?

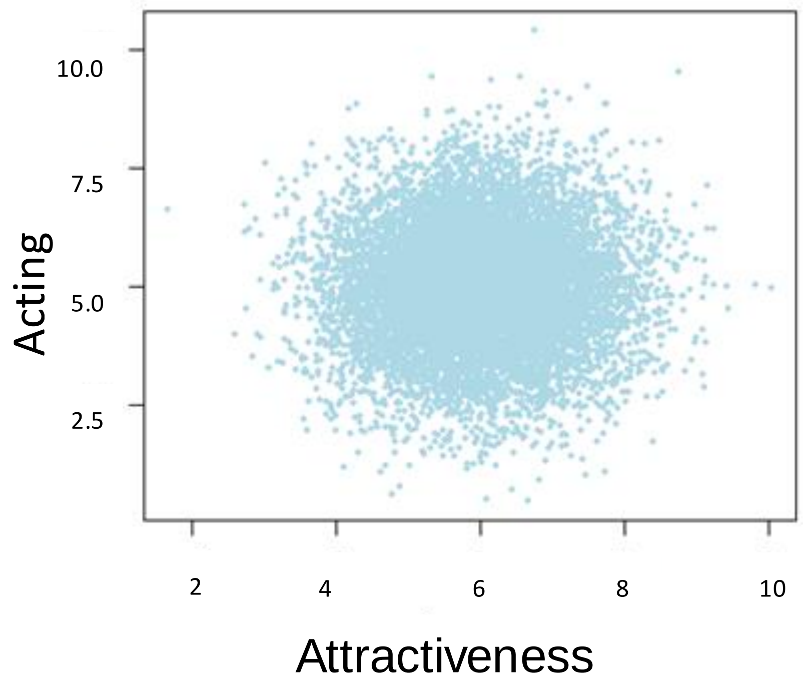


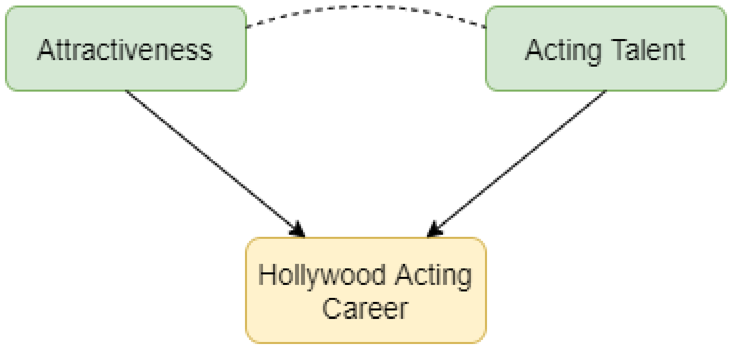
ANNOYINGLY, if two factors influence selection of participants into a sample, they become correlated

Say those two factors are a hypothesized risk factor and an outcome

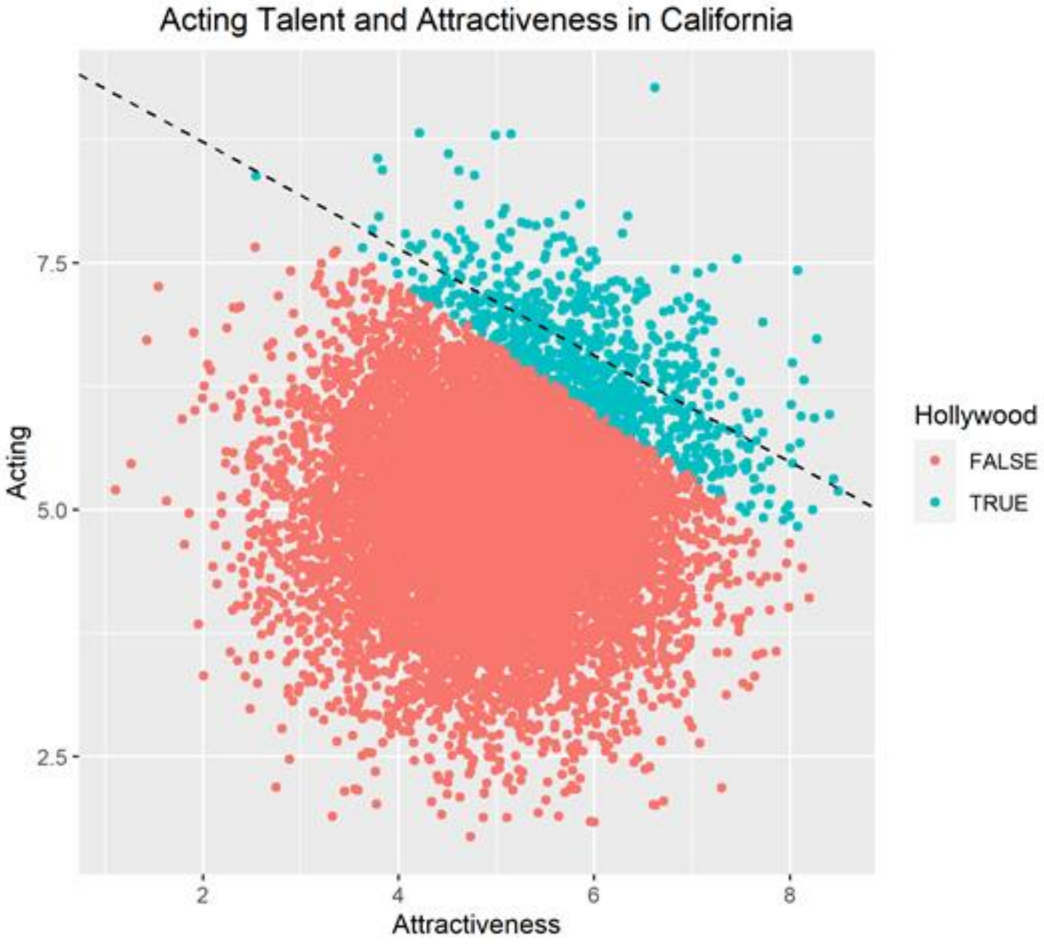
Therefore, estimating the effect of the risk factor on the outcome is biased

Toy example:
Let’s assume no relationship between **attractiveness** and **acting ability** in the general population





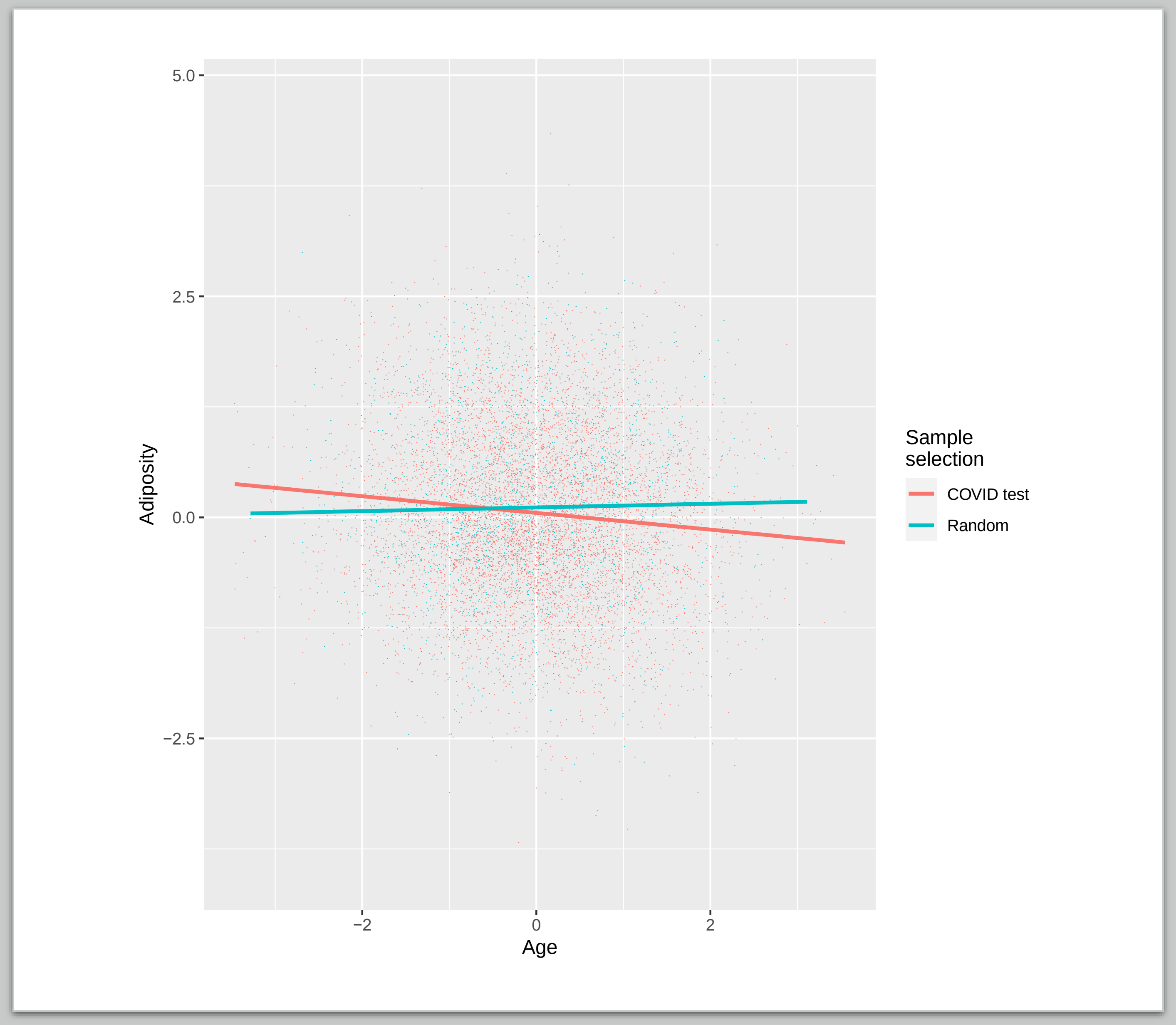
What would the relationship look like if you restricted the analysis only amongst Hollywood actors?



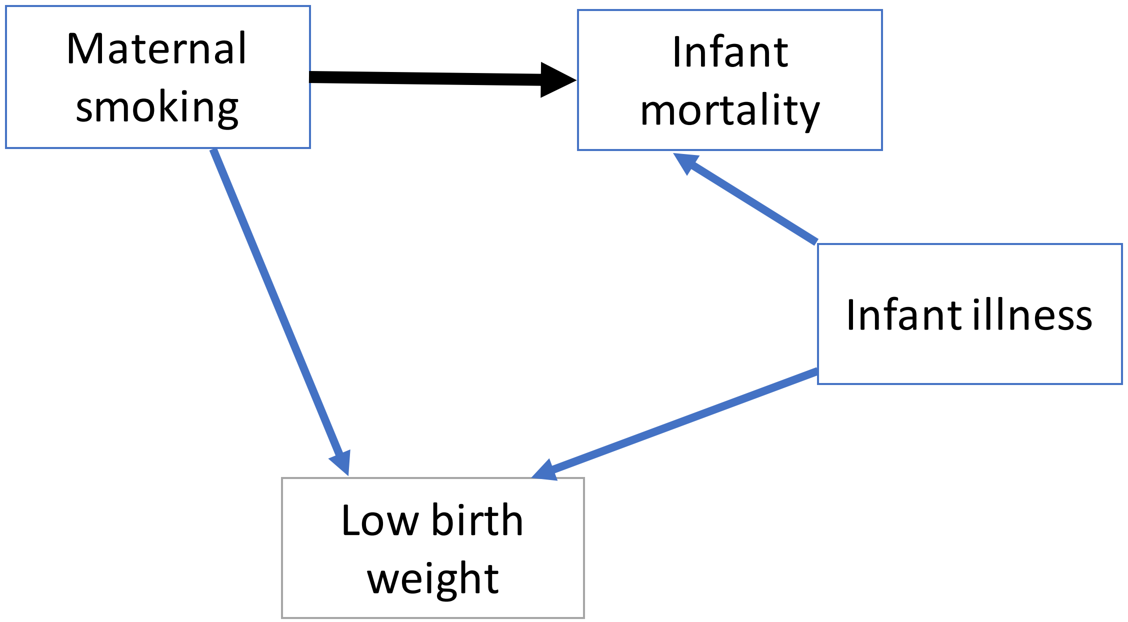
In Hollywood, being physically attractive is inversely related to being a talented actor.

Reality tends to be a bit more subtle

- Relationship between age and obesity (‘adiposity’) is distorted in people who have been tested for COVID-19



Among low birthweight infants, those whose mothers smoked during pregnancy are less likely to die than those whose mothers did not smoke



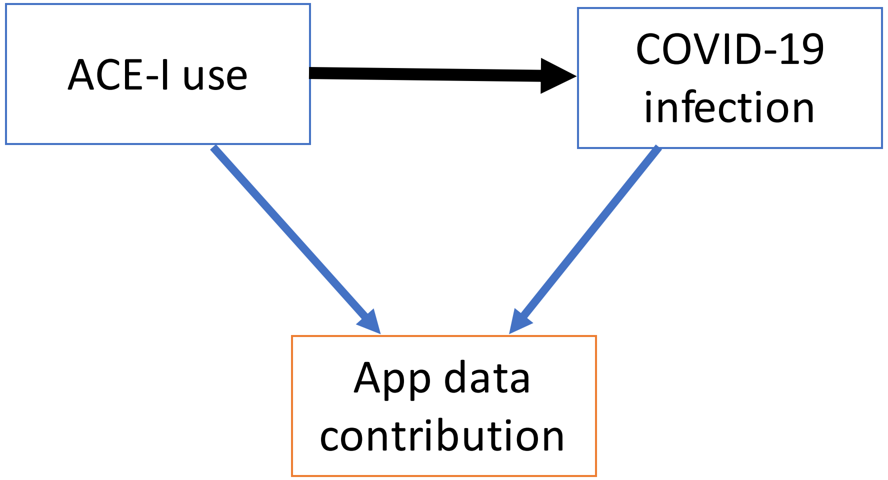
Intuition:

Maternal smoking probably does increase infant mortality a bit, but severe illnesses are likely much worse for infant mortality

A by-product of both smoking and illnesses is low birthweight

By analyzing only amongst the low birth weight babies, those that were *selected* due to smoking exposure probably have much better prospects than those selected due to more severe reasons for low birthweight.

Back to COVID-19 and ACE-inhibitors…



First data freeze (end of March), ~1 million samples and 38k self reported having COVID-19
Odds ratio = **4.1 (95% CI 3.8-4.5)**

Second data freeze (early April), ~2.2 million samples
Odds ratio = **1.95 (95% CI 1.89-2.03)**

When analysed amongst individuals tested for COVID-19
Odds ratio = **0.82 (95% CI 0.61-1.10)**

Take home message: each of these samples have different sample selection pressures

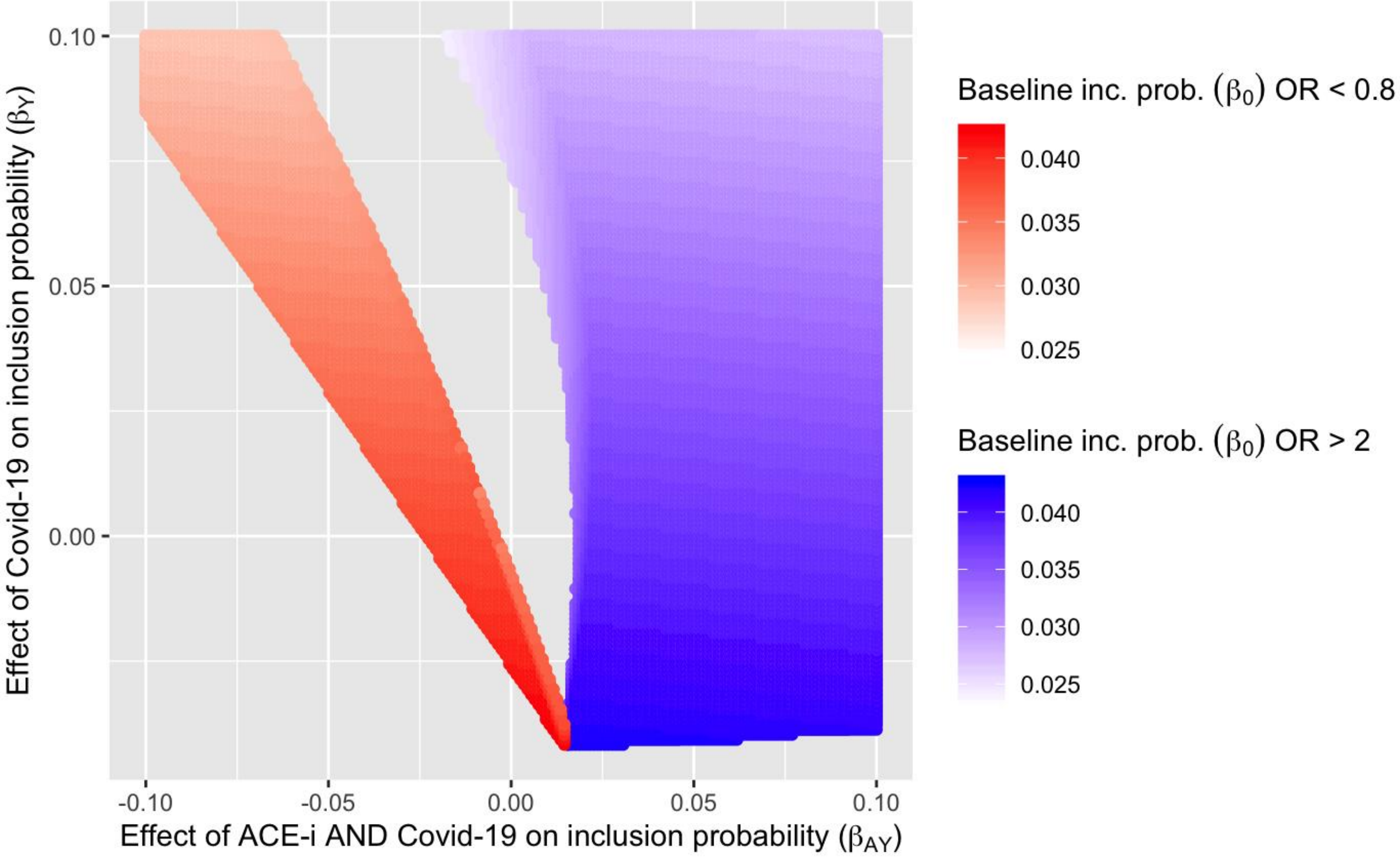
Can collider bias induce such large associations?

Probability of being in the sample as a function of ACE-I and COVID-19 status

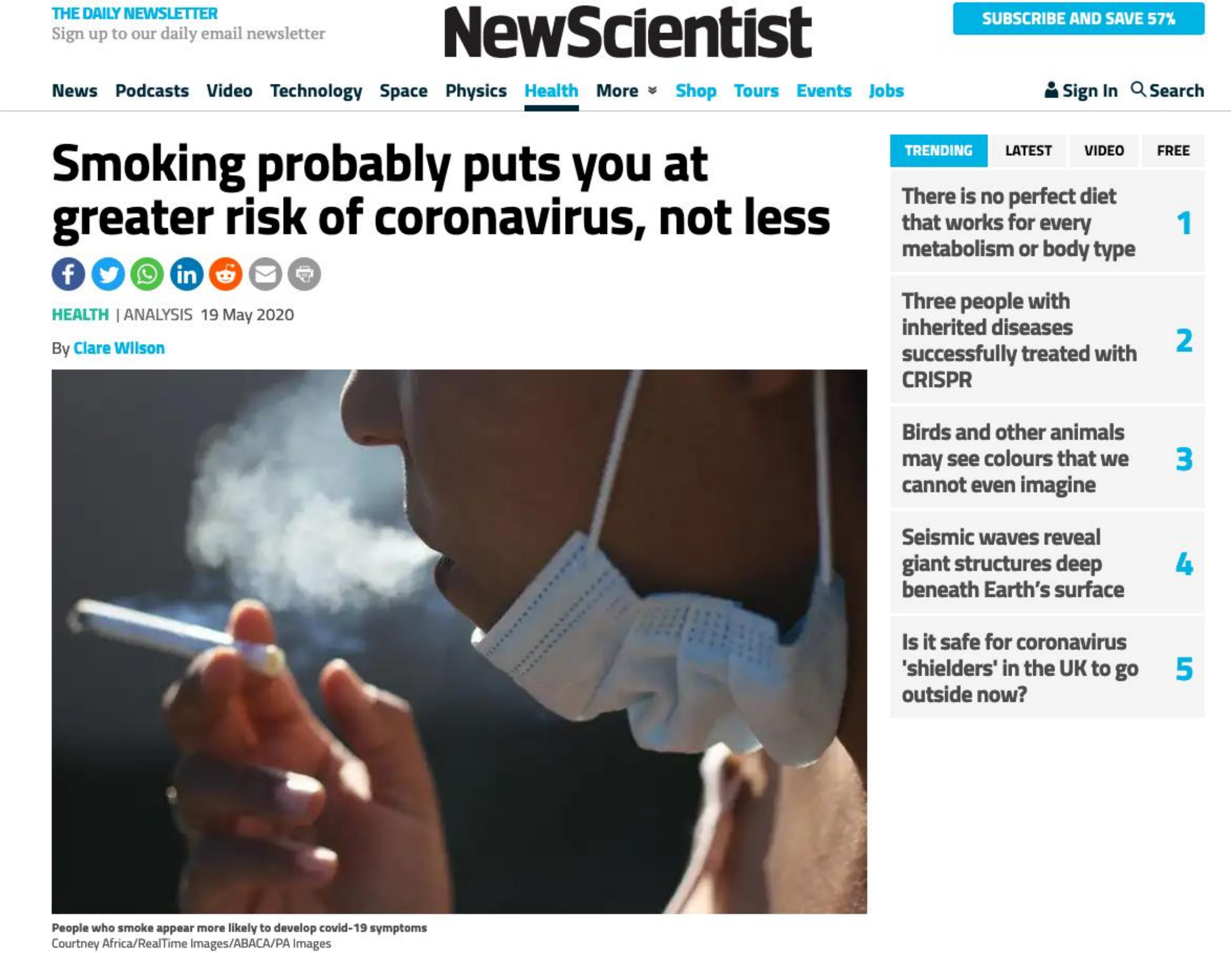
$$\mathbb{P}(S = 1|A, Y) = \beta_0 + \beta_A A + \beta_Y Y + \beta_{AY} AY$$

$$\begin{aligned}\mathbb{E}[\hat{OR}_{S=1}] &= \frac{P(Y=1|A=1,S=1)}{1-P(Y=1|A=1,S=1)} \Big/ \frac{P(Y=1|A=0,S=1)}{1-P(Y=1|A=0,S=1} \\ &= \frac{\beta_0(\beta_0+\beta_A+\beta_Y+\beta_{AY})}{(\beta_0+\beta_A)(\beta_0+\beta_Y)}.\end{aligned}$$

Groenwold, Palmer and Tilling (2019)



http://apps.mrcieu.ac.uk/ascrtain/



NewScientist

Smoking probably puts you at greater risk of coronavirus, not less

HEALTH | ANALYSIS 19 May 2020

By Clare Wilson

People who smoke appear more likely to develop covid-19 symptoms
Courtney Africa/RealTime Images/ABACA/PA Images

TRENDING LATEST VIDEO FREE

There is no perfect diet that works for every metabolism or body type 1

Three people with inherited diseases successfully treated with CRISPR 2

Birds and other animals may see colours that we cannot even imagine 3

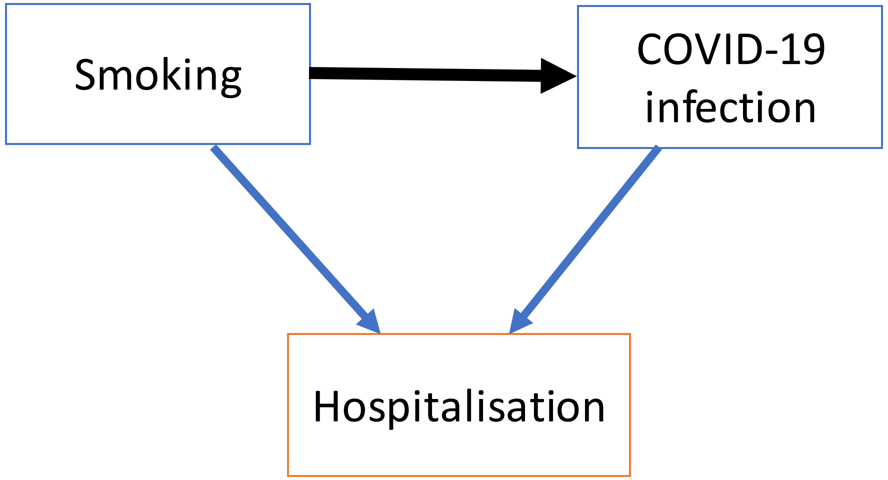
Seismic waves reveal giant structures deep beneath Earth's surface 4

Is it safe for coronavirus 'shielders' in the UK to go outside now? 5

Low rate of daily active tobacco smoking in patients with symptomatic COVID-19 Preprint v4

Makoto Miyara[1], Florence Tubach[1], Valérie POURCHER[1], Capucine Morelot-Panzini[1], Julie Pernet[1], Julien Haroche[1], Said Lebbah[1], Elise Morawiec, Guy Gorochov[2], Eric Caumes[1], Pierre Hausfater[1], Alain COMBES[1], Thomas Similowski, Zahir Amoura[1]

Analysis conducted amongst hospitalized patients in a district of Paris



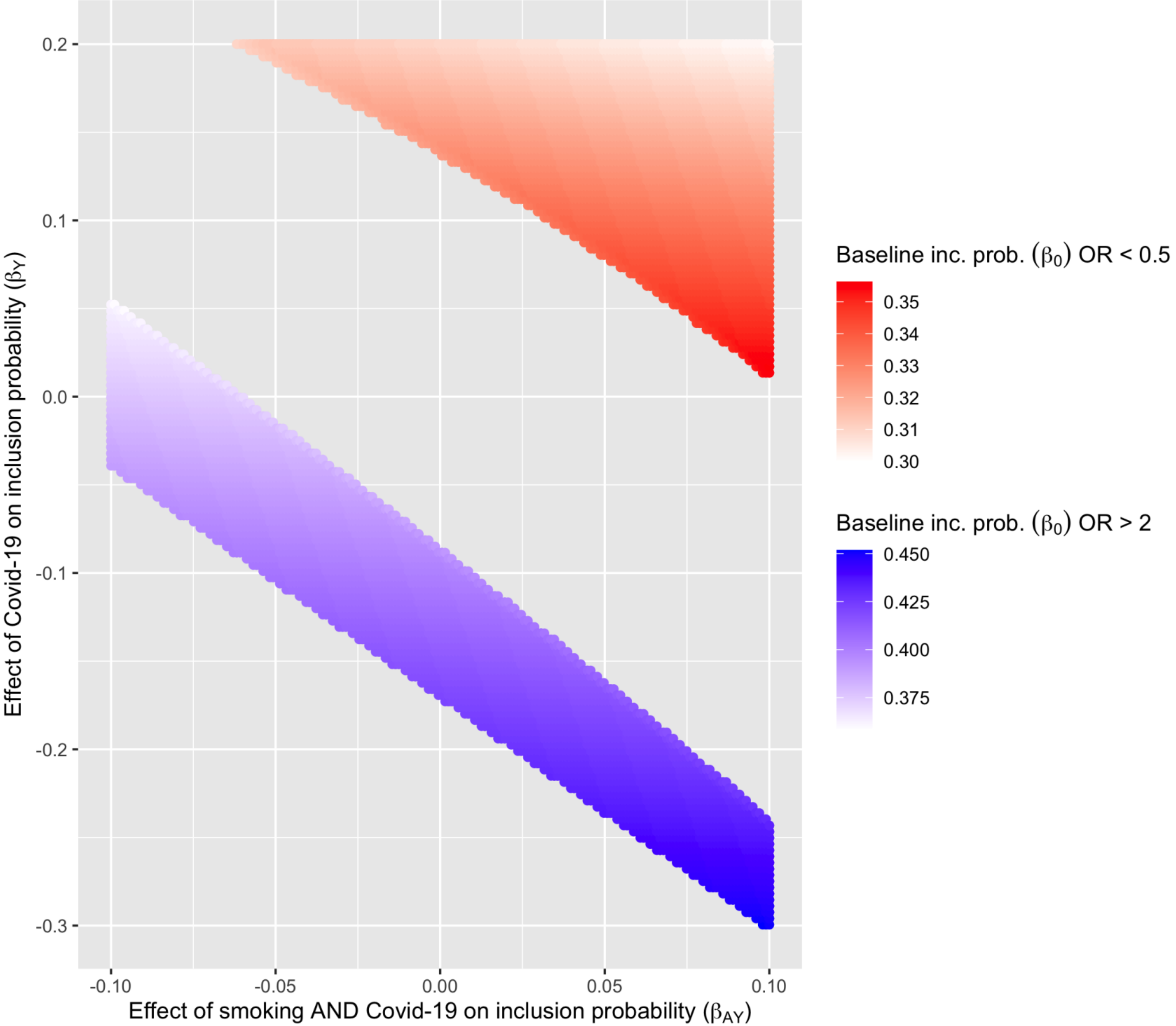
Smoking prevalence in Paris: 27%

Smoking prevalence in sample: 5%

Covid-19 prevalence in the population: unknown (allow to vary on the y axis)

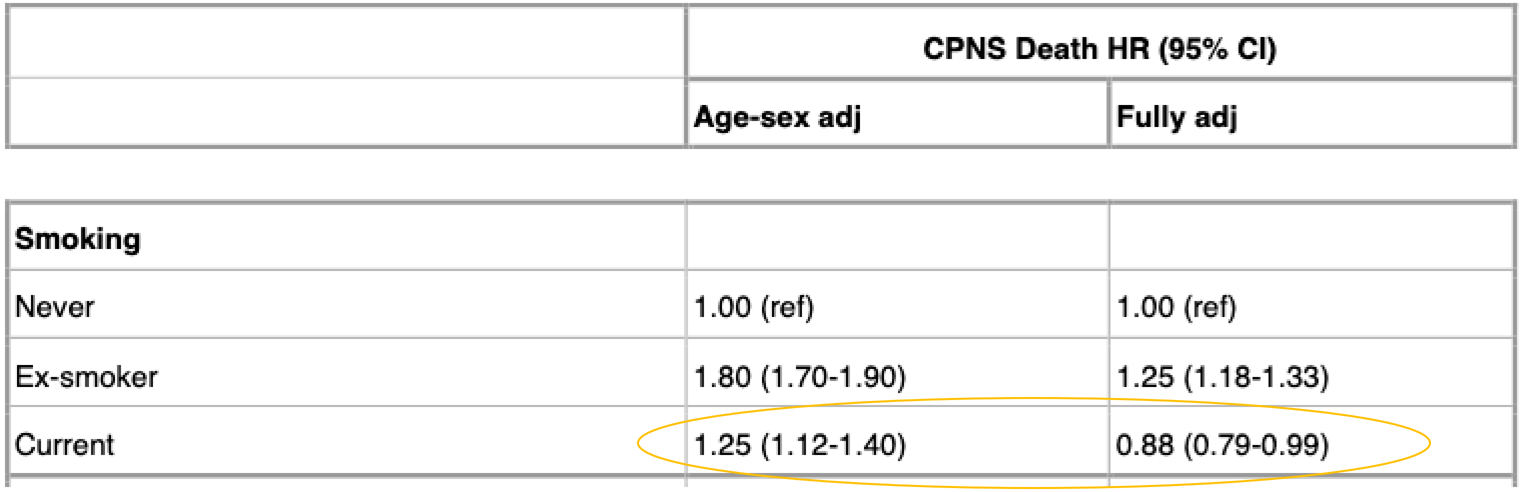
Interaction between covid-19 and smoking on sample inclusion: unknown (allow to vary on the x-axis)

40% of the parameter space gives rise to 2x protective or risk association

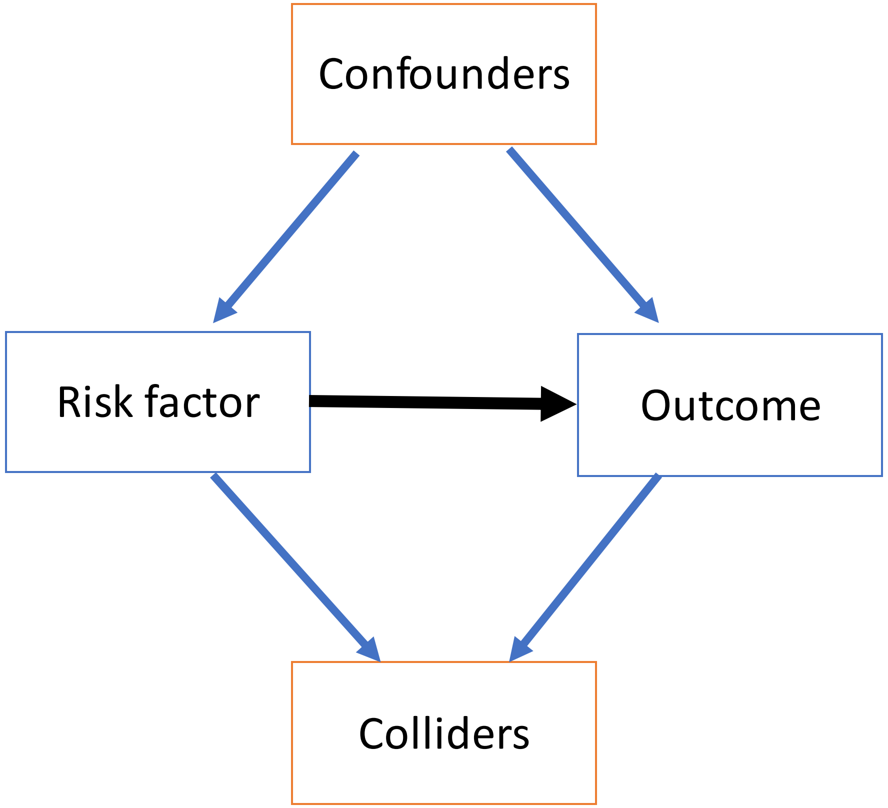


Collider bias in representative samples

OpenSAFELY analysis of influence of smoking on death from COVID-19 in 17 million primary care users



	CPNS Death HR (95% CI)	
	Age-sex adj	**Fully adj**
Smoking		
Never	1.00 (ref)	1.00 (ref)
Ex-smoker	1.80 (1.70-1.90)	1.25 (1.18-1.33)
Current	1.25 (1.12-1.40)	0.88 (0.79-0.99)



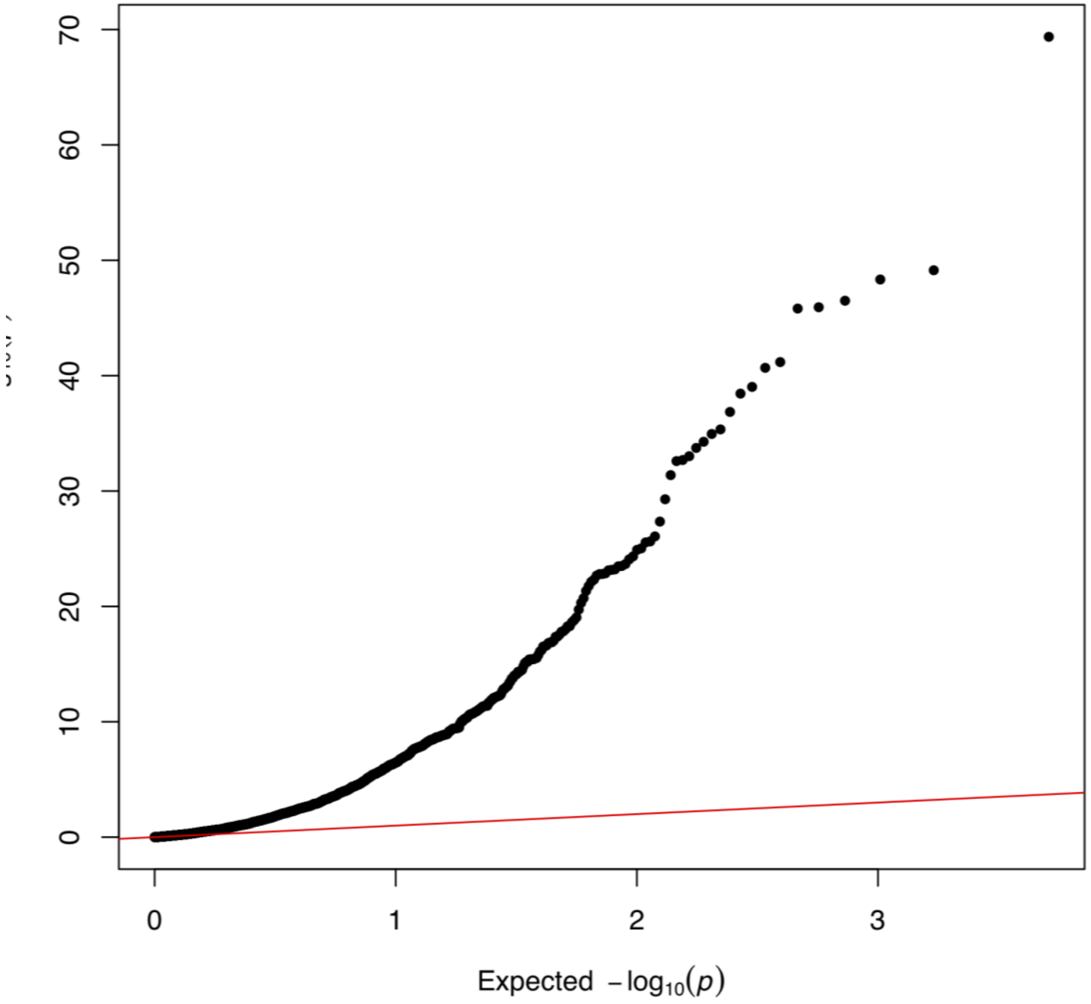
Testing for COVID-19 is non-random

In the UK-Biobank (500k individuals), about 2000 were linked to their COVID-19 test results from primary care records

Tested for an association between each of ~2400 variables and whether or not an individual received a test

850 of the variables had a strong association

Socioeconomic status, health status, behaviours, age, sex, genetic factors, etc.



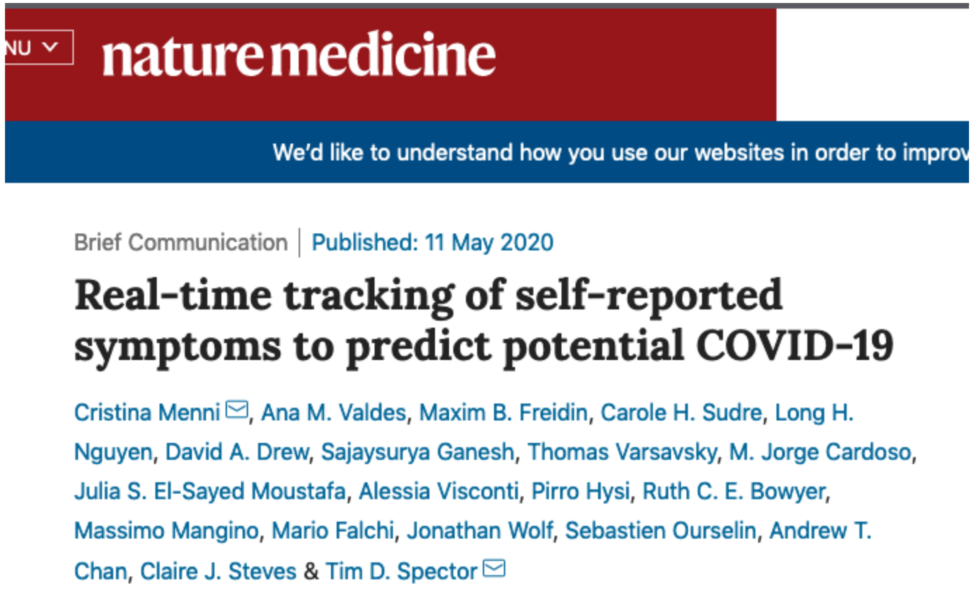
Prediction of COVID-19 outcomes

We don’t necessarily care about finding causal factors when making prediction

e.g. yellow fingers might make a good predictor of lung cancer (smoking confounds yellow fingers and lung cancer)

Colliders could help you with prediction also (e.g. if baby has low birthweight then maternal smoking predicts lower likelihood of infant mortality)

This only works if the testing sample and the training sample have the same sample selection



NU

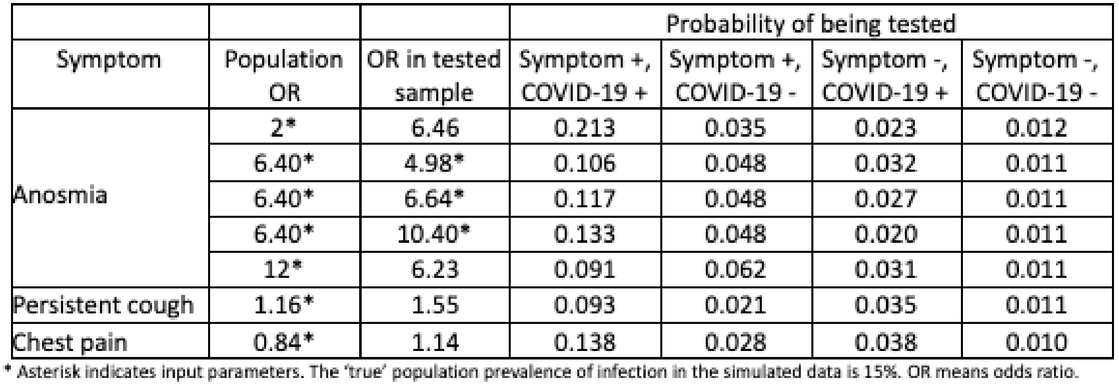
nature medicine

We'd like to understand how you use our websites in order to improv

Brief Communication | Published: 11 May 2020

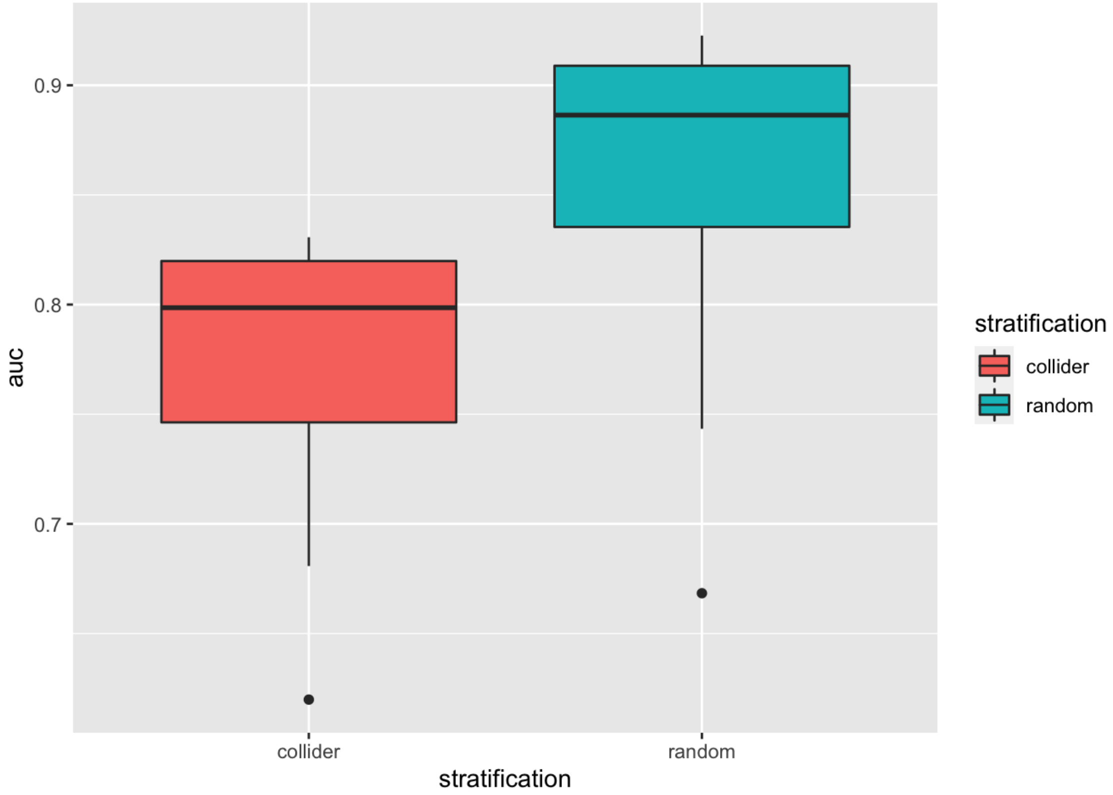
Real-time tracking of self-reported symptoms to predict potential COVID-19

Cristina Menni, Ana M. Valdes, Maxim B. Freidin, Carole H. Sudre, Long H. Nguyen, David A. Drew, Sajaysurya Ganesh, Thomas Varsavsky, M. Jorge Cardoso, Julia S. El-Sayed Moustafa, Alessia Visconti, Pirro Hysi, Ruth C. E. Bowyer, Massimo Mangino, Mario Falchi, Jonathan Wolf, Sebastien Ourselin, Andrew T. Chan, Claire J. Steves & Tim D. Spector



			Probability of being tested			
Symptom	Population OR	OR in tested sample	Symptom +, COVID-19 +	Symptom +, COVID-19 -	Symptom -, COVID-19 +	Symptom -, COVID-19 -
Anosmia	2*	6.46	0.213	0.035	0.023	0.012
	6.40*	4.98*	0.106	0.048	0.032	0.011
	6.40*	6.64*	0.117	0.048	0.027	0.011
	6.40*	10.40*	0.133	0.048	0.020	0.011
	12*	6.23	0.091	0.062	0.031	0.011
Persistent cough	1.16*	1.55	0.093	0.021	0.035	0.011
Chest pain	0.84*	1.14	0.138	0.028	0.038	0.010

* Asterisk indicates input parameters. The 'true' population prevalence of infection in the simulated data is 15%. OR means odds ratio.



Analysis performed by Matt Tudball

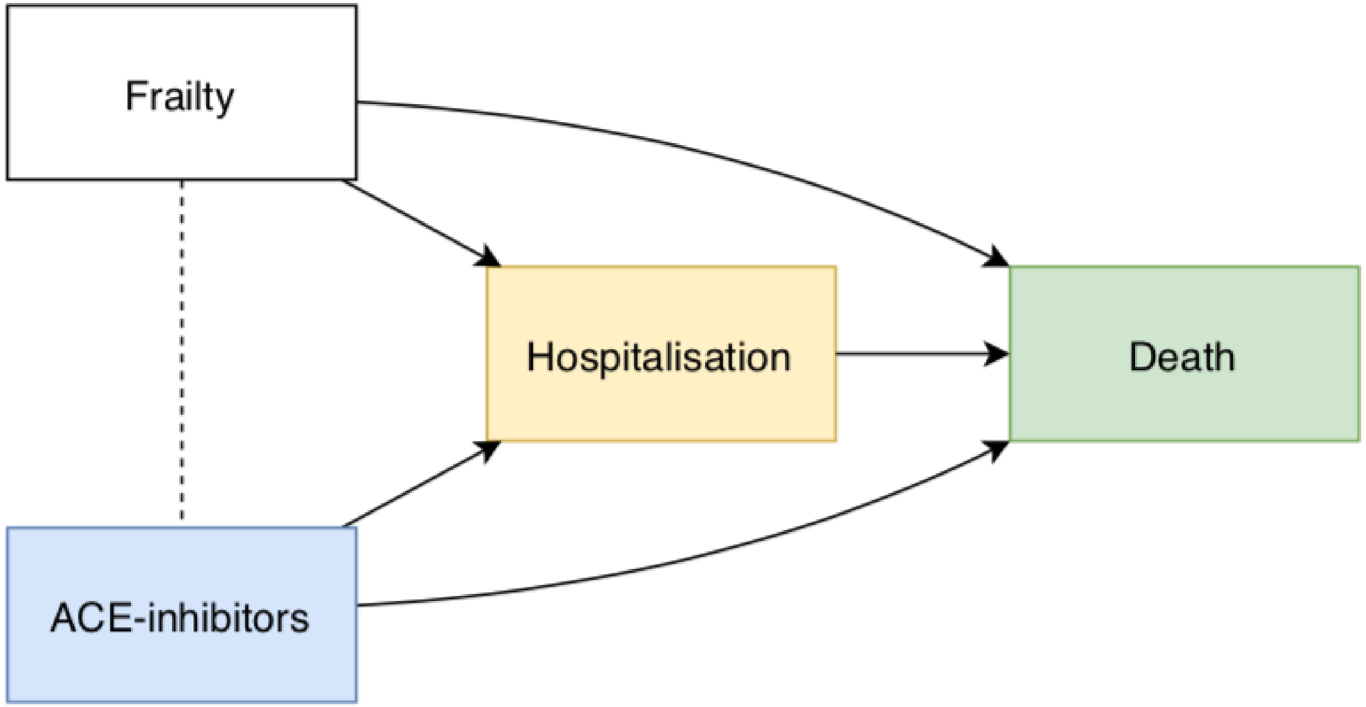
What about analyzing survival amongst those infected by COVID-19?

Association of Inpatient Use of Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers With Mortality Among Patients With Hypertension Hospitalized With COVID-19

Peng Zhang, Lihua Zhu, Jingjing Cai, Fang Lei, Juan-Juan Qin, Jing Xie, Ye-Mao Liu, Yan-Ci Zhao, Xuewei Huang, Lijin Lin, Meng Xia, Ming-Ming Chen, Xu Cheng, Xiao Zhang, Deliang Guo, Yuanyuan Peng, Yan-Xiao Ji, Jing Chen, Zhi-Gang She, Yibin Wang, Qingbo Xu, Renfu Tan, Haitao Wang, Jun Lin, Pengcheng Luo, Shouzhi Fu, Hongbin Cai, Ping Ye, Bing Xiao, Weiming Mao, ... **See all authors**

Originally published 17 Apr 2020 | https://doi.org/10.1161/CIRCRESAHA.120.317134 | Circulation Research. 2020;126:1671–1681

ACE-inhibitors protect you once infected?



Those taking ACE-inhibitors and hospitalized for COVID-10 are a much healthier sub-group than most others hospitalized for COVID-19

Techniques to overcome collider bias

Recruit individuals who are representative of your target individuals

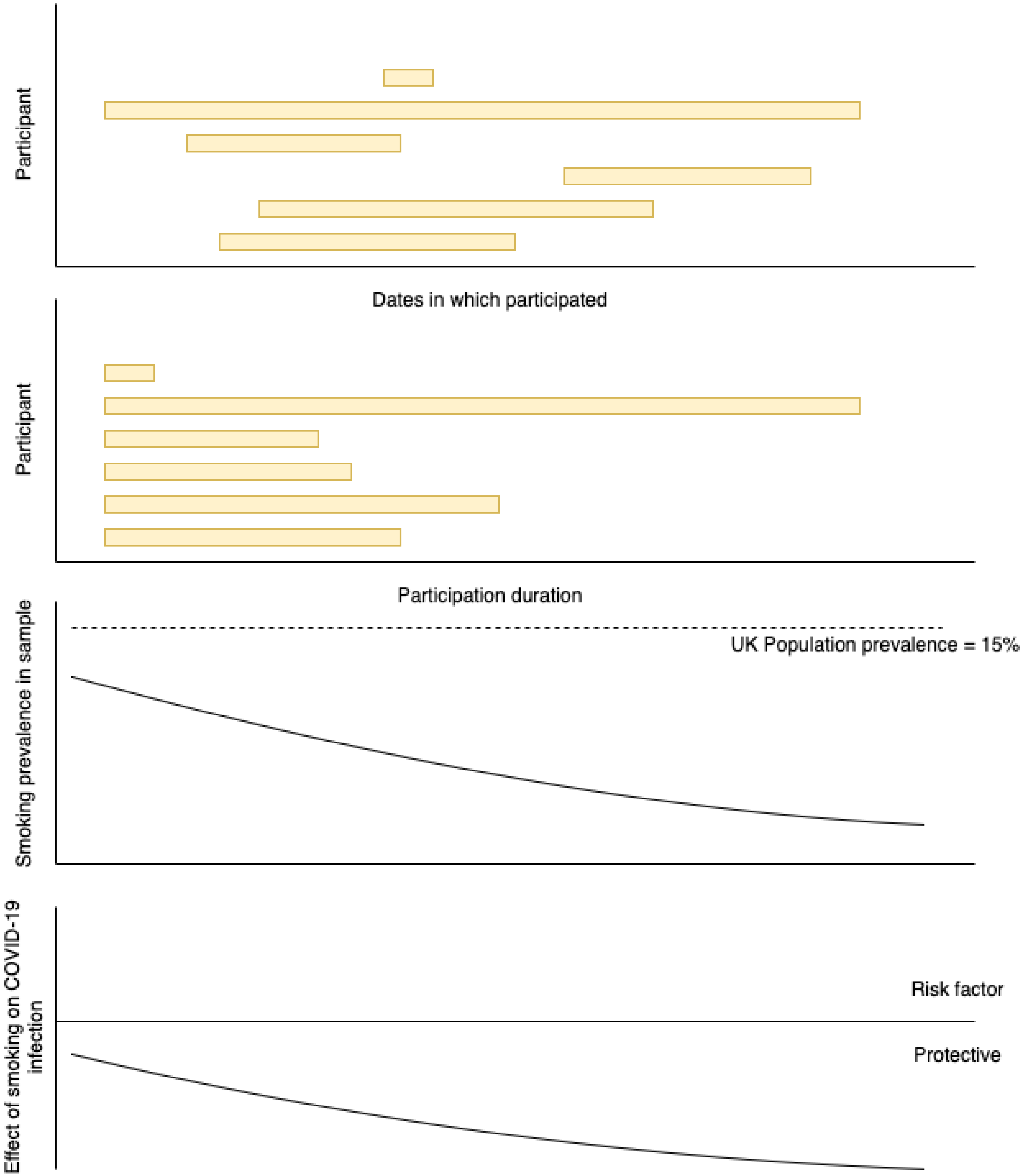
This is easier said than done

- You might invite people at random but only a particular subset responds (e.g. UK Biobank)
- You might recruit a representative sample but individuals drop out non-randomly (e.g. ALSPAC)

App participation drop out and smoking

Illustrative example (data not available)

Sensitivity analysis: extrapolate backwards to the point that participation prevalence matches population prevalence



Sensitivity analysis

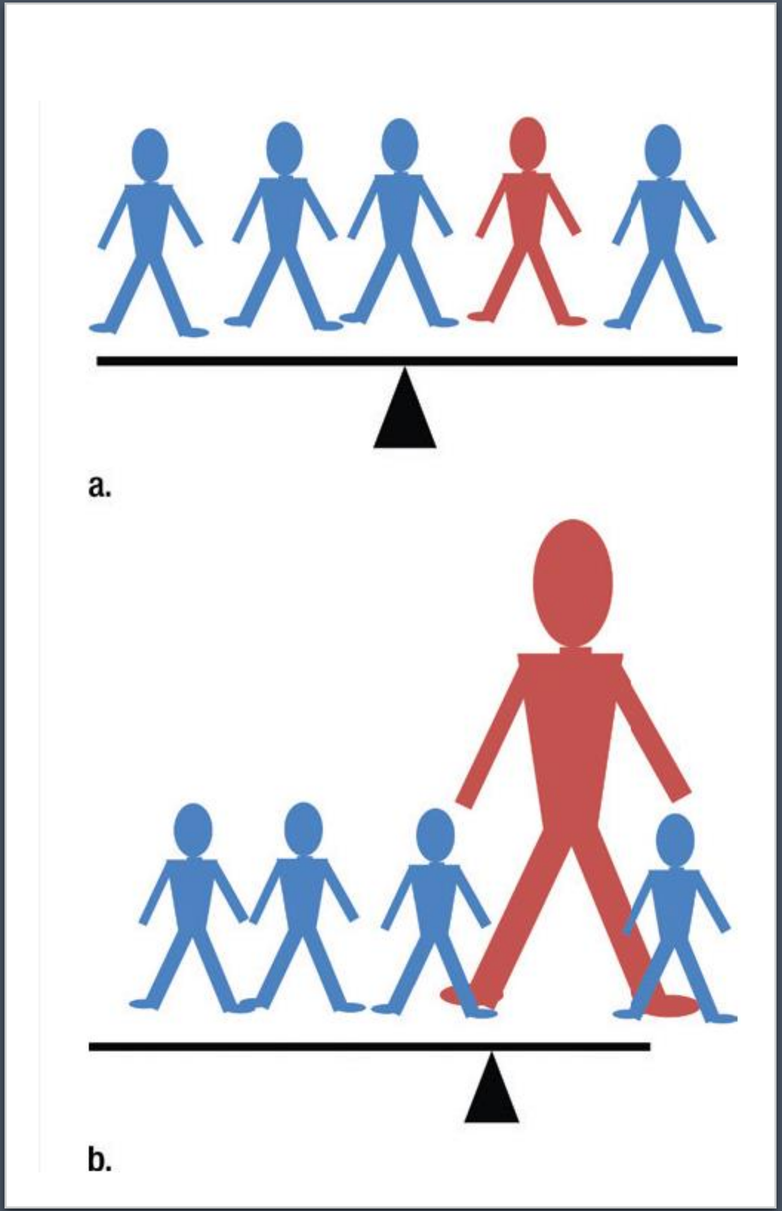
Compare population descriptive statistics against sample – is the sample representative?

How credibly could collider bias explain your association?

Perform more sophisticated analyses to redress the lack of representativeness of the sample

Weighting individuals in the analysis

- Create a model that predicts the probability of an individual being selected into the data
- Now weight individuals so that they contribute inversely to that probability (Inverse Probability Weighting IPW)
- An important sensitivity analysis, but hard to execute – do we know all the factors that influenced selection?
- Could misspecification of the probability model make things worse!?



Summary

- Epidemiological data on Covid-19 data is crucial, but across almost all study designs it’s highly non-representative (whether looking at infection or disease severity)
- This can distort associations to a large degree
- When reading any epidemiological study for COVID-19, hit “ctrl+F” and search for “collider”, “selection”, “random”, “representative” etc
- Have the authors done anything to convince you that they tried to account for non-representative sampling?

Acknowledgements

- Gareth Griffith
- Tim T Morris
- Matt Tudball
- Annie Herbert
- Giulia Mancano
- Lindsey Pike
- Gemma C Sharp
- Tom M Palmer
- Jonathan Sterne
- George Davey Smith
- Kate Tilling
- Luisa Zuccolo
- Neil M Davies

Further info

- Detailed paper on the issues discussed here: https://www.medrxiv.org/content/10.1101/2020.05.04.20090506v3

- Blog posts
  - HDRUK
  - IEUREKA

Data Week Online 2020

Share your participation

#Dataweekonline2020

Keep in touch

@JGIBristol
jgi-admin@bristol.ac.uk
bristol.ac.uk/golding

bristol.ac.uk/golding