> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
>
> – Sir Ronald Fisher[6]

Research Design

BRSM

Measurement in the behavioral sciences

Measurement

Define the property you want to study

Find a way to *detect* that property

Examples

Aggression (*operational definition*? *Measure*?)

Intelligence (*operational definition*? *Measure*?)

Productivity in the office (*operational definition*? *Measure*?)

Age (how do you measure this? Depends.. Developmental psych? Consumer research?)

Operational definition

- A working definition of what a researcher is measuring

In this task, "on target" is +/- 10% of the goal distance.

Golf Slides from Lassiter Speller (OSU).

Terminology

- ***A theoretical construct***. This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- ***A measure***. The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- ***An operationalisation***. The term "operationalisation" refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- ***A variable***. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

Variable types: scales of measurement

- Nominal
- Ordinal
- Interval
- Ratio

Nominal scale

Categorical

e.g. Eye color, sex

Does not make sense to say one is greater than the other

Also does not make sense to average them (e.g. average eye color?!)

Nominal scale

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

Ordinal Scale

- Slightly more structured than nominal: now you can order the variables in some sensible way

Here's an more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change, and I ask them to pick one of these four statements that most closely matches their beliefs:

> 1. Temperatures are rising, because of human activity
> 2. Temperatures are rising, but we don't know why
> 3. Temperatures are rising, but not because of humans
> 4. Temperatures are not rising

Natural ordering of the options

- Relative to some ground truth (e.g. scientific evidence), statement 1>2>3>4

So, let's suppose I asked 100 people these questions, and got the following answers:

	Number
(1) Temperatures are rising, because of human activity	51
(2) Temperatures are rising, but we don't know why	20
(3) Temperatures are rising, but not because of humans	10
(4) Temperatures are not rising	19

- How do we group these responses for analysis?
- If it is an ordinal scale measurement, there are some sensible ways to do this and others that don't make sense
- Again, the average does not make sense: the average endorsed statement here is 1.97

Interval scale

Both interval and ratio scales: numerical value now can be interpreted directly

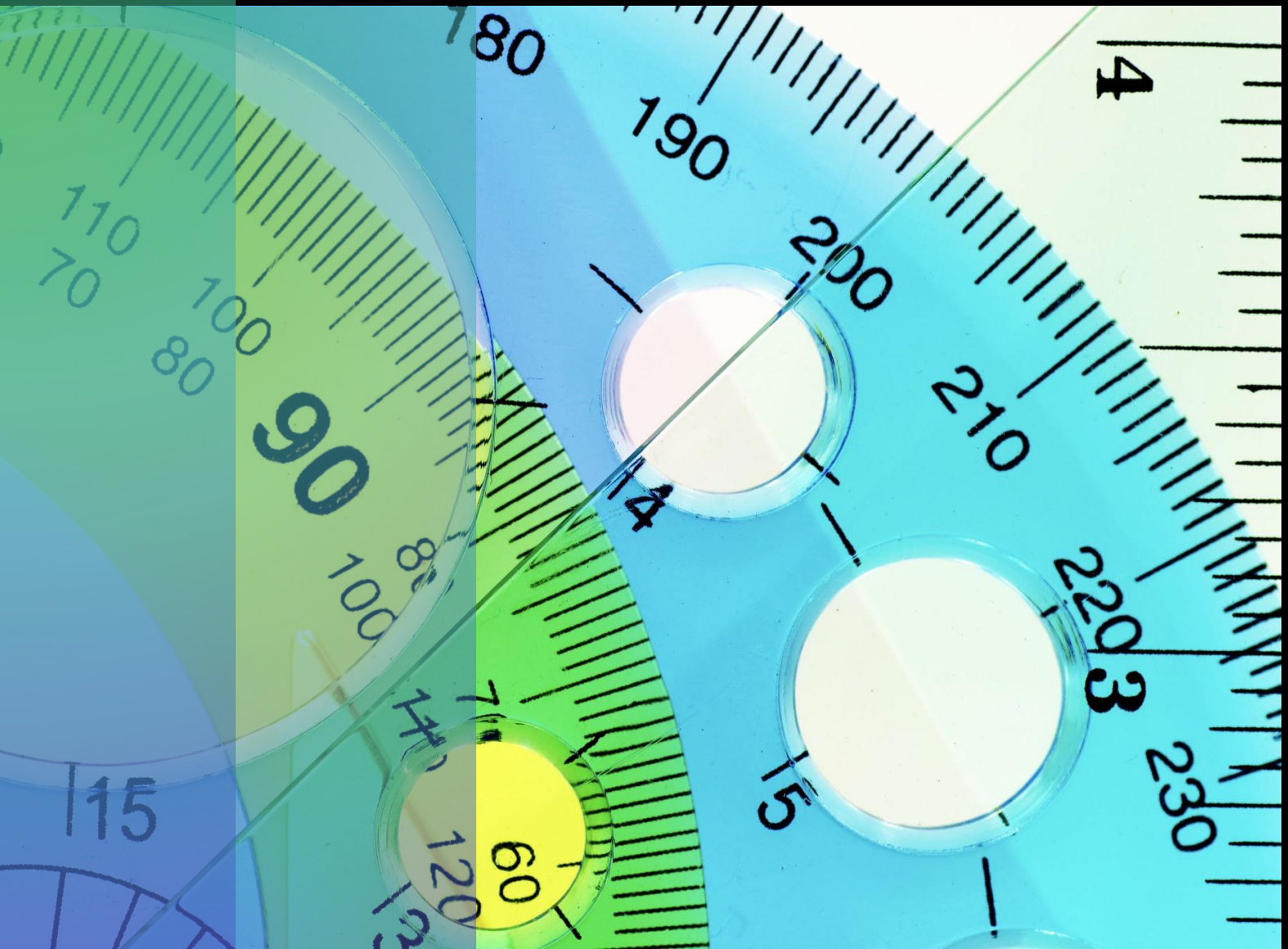
Interval: differences between numbers make sense, but there is no natural "zero" on this scale

Addition and subtraction make sense, but not multiplication or division

e.g. temperature. Difference between 20 and 17 deg celsius = 3 degrees. The same as the difference between 30 and 27 degrees. There is no natural zero, just an arbitrary point (freezing point) chosen as a reference.

Psych e.g. student attitudes as a function of time elapsed since joining date – the year of entry is an interval scale measurement

Averages, medians, etc make sense: the average temperature for the month



Ratio scale

Zero means zero
Can divide
e.g. Reaction times (e.g. I'm twice as fast as you)

Continuous vs discrete variables

- A ***continuous variable*** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A ***discrete variable*** is, in effect, a variable that isn't continuous. For a discrete variable, it's sometimes the case that there's nothing in the middle.

Examples? -- what type of scale? Discrete or continuous?

- RTs?
- Year in which participants were born?
- Temperature?
- Your mode of transport to work?
- Place attained in a race?

- RTs – ratio scale and continuous
- Year in which participants were born – interval scale and discrete
- Temperature – interval scale and continuous
- Your mode of transport to work? - nominal and discrete
- Place attained in a race? - ordinal and discrete

Continuous vs discrete variables

Table 2.1: The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

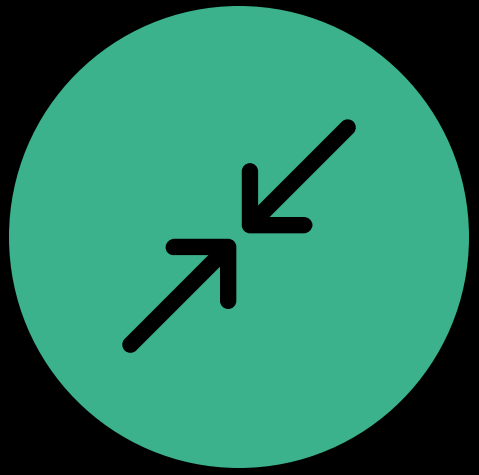
Real world variables may not always adhere to these classifications

- Likert scale
- Choose from the following options. You feel happy today:
- What scale is this?

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

- Nominal? (hint: is there a natural ordering? If so, it can't be nominal)
- Ratio? (hint: is there a natural "zero"?)
- Ordinal or interval. Which one is it?
- Can we prove that everybody treats the difference between 1. and 2. the same as the difference between 4. and 5.?
- In practice, most people treat the likert scale as an interval scale since many participants treat the entire scale seriously (but this is very much dependent on the task and context).

Is the measurement any good?



RELIABILITY: HOW REPEATABLE?

VALIDITY: HOW ACCURATE IS IT IN RELATION TO WHAT YOU WANT TO MEASURE?

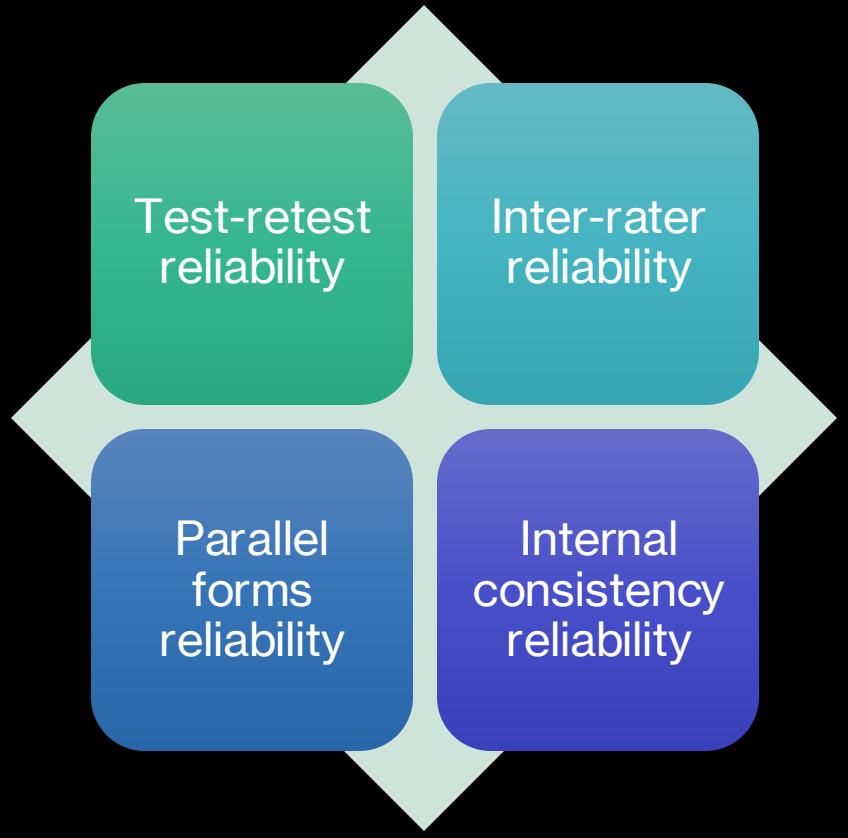
Reliability

Will we produce the same thing repeatedly?

E.g.

- Weighing machine: day 1 = 90 kgs, day 2 = 110 kgs unreliable!
- Psychology example:
  - Want to measure depression
  - Operational definition: Number of times you hang out with family and friends (lower = depression)
  - Measurement in July vs Nov
  - Reliable?

Different ways to measure reliability



Test-retest reliability

CONSISTENCY OVER TIME

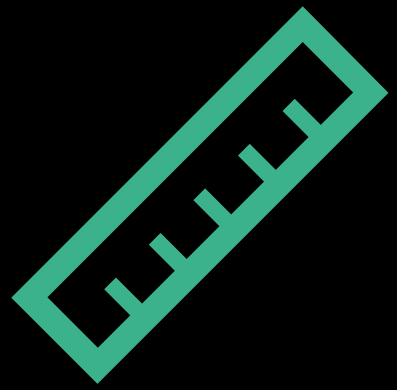
DO WE GET THE SAME RESULTS WHEN WE TEST AT ANOTHER TIME?

Inter-rater reliability

CONSISTENCY ACROSS PEOPLE

IF SOMEONE ELSE DOES THE MEASUREMENT, WILL WE GET THE SAME RESULT?

Parallel forms reliability



Consistency across theoretically-equivalent measurements

If I use a different weighing scale, do I get the same weight measurement?

Internal consistency reliability

CONSISTENCY ACROSS DIFFERENT PARTS WITH THE SAME FUNCTION

IF QUESTIONS ON FLUID INTELLIGENCE SPREAD ACROSS THE IQ TEST ALL GIVE SIMILAR ESTIMATES OF MY INTELLIGENCE, THE TEST HAS INTERNAL CONSISTENCY

Think about the evaluation components of this course

How good is the internal consistency of the evaluations? (problem sets + quizzes + projects)

How about within quizzes or any given component?

Experimental variables

Independent variable: (IV) the variable that
is manipulated Examples: amount of light, exposure to a loud noise, drug

Dependent variable: (DV) the variable that is measured to see if the independent variable had an effect. Examples: Plant growth, change in heart rate, anxiety scores

Table 2.2: The terminology used to distinguish between different roles that a variable can play when analysing a data set. Note that this book will tend to avoid the classical terminology in favour of the newer names.

role of the variable	classical name	modern name
to be explained	dependent variable (DV)	outcome
to do the explaining	independent variable (IV)	predictor

Modern terminology

- We're using the predictors to make guesses about the outcome

Experimental Research

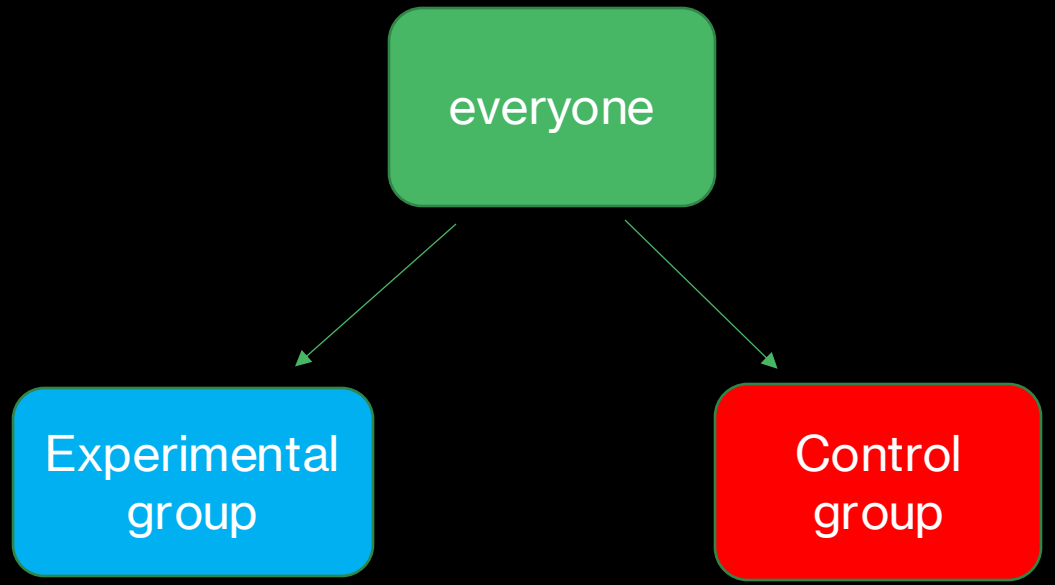
- The experimenter controls everything
- Manipulates the predictors and sees how the outcome changes

Practical issues

We cannot possibly think of ALL the predictors that can influence the outcome

How do we solve this issue?

Randomization



Predictor (IV)

Absence of the predictor variable (or a different level of the variable)

Then compare the outcomes in the two groups

Discussion: Effect of playing violent video games on aggression

Examine the database of players provided by a gaming company

Get criminal records

Test for a difference in the records between game players and non-players

Any problems with this?

The role of confounds

- Perhaps the people playing violent video games as young children are also ones without proper parental support
- In the previous study, there was no consideration of this potential confound

The ideal experiment?

- Take a random sample from the population
- Randomly assign them into violent game-play vs peaceful game-play groups
- Monitor their lives for a few decades
- Get criminal records
- This is not exactly feasible though

So what do we do then?

Use statistics!

Incorporate confounds as covariates in your statistical models!

I.e., we still want to understand how the outcome (aggression) varies as the predictor value is changed (violent game play) but now we will first take into account what amount of the outcome is affected by the confounding variables

Validity

Confounds affect the validity of your study

Many more factors that affect the validity of a study

Important to examine those before we delve into statistical methods

Validity

- Internal validity
- External validity
- Construct validity
- Face validity
- Ecological validity

Internal validity

- The ability to draw cause and effect inferences from the data
- The effect of covid (Delta) on IQ.
- Recruit govt hospital patients. Compare with healthy controls who responded to online ads for your study.
- Internal validity?

External validity

- Generalizability of your findings
- Govt hospital COVID patients and their cognitive issues: generalizable to the rest of the population?
- A basic perception study with college undergrads?
- A study on attitudes towards psychotherapy based on CogSci students at IIITH?

Construct validity

- Are you really measuring what you want to be measuring?
- I want to understand the prevalence of depression in the student population
- I post a tweet and ask people with depression to like the tweet and others to retweet. The proportion fo students who liked the tweet = my answer. How good is my construct validity?

Face validity

DOES YOUR TEST "APPEAR" TO BE DOING THE JOB IT SAYS IT WILL DO?

DOESN'T REALLY MATTER FOR SCIENTISTS.

CAN MATTER IF YOU'RE TRYING TO CONVINCE POLICY MAKERS FOR EXAMPLE. THEN THEIR PERCEPTION ABOUT THE TEST WOULD MATTER.

Ecological validity

- Does the experiment closely mimic real-world scenarios?
- Related to external validity in that ecological validity is supposed to help us generalize the findings to real-world scenarios
- Though that is not guaranteed
- e.g. eye-witness studies in the lab lack ecological validity
- e.g. Word memory experiments
- However, insights from word memory experiments may (and do) generalize to more ecologically valid settings

Threats to validity

- Confounds – related to both predictors and outcomes in some systematic way. A threat to internal validity. Why?
- Artifacts – something about the way you did the experiment that gave you the result. A threat to external validity (but probably also internal). Why?

History effects

Something that happens during the study (or preceding) that can influence the results

Hospital stay, patient testing, 3rd day compared to 7th day. Electrode rearrangement surgery on day 5.

Maturational effects

Something that changes naturally over time that can influence your results

One big effect in psych lab experiments: waning attention, fatigue, which increases over the course of the experiment. How do you know that primacy effects are not driven by such maturational effects?

(Repeated) testing effects

Practice effects

Familiarity with the test

Better scores in session 2 compared to session 1

Selection bias

- Refers to anything that makes the groups being compared different in some potentially critical aspect
- Different proportions of males/females in the two groups in a study on aggression
- No more internal validity

Differential attrition

- If you do a long study, or a longitudinal study or any study that requires quite a bit of effort from the participants, this may be relevant.
- People drop out.
- The people dropping out are not random people.

Homogeneous vs heterogeneous attrition

- The rates of attrition can be the same across groups you're comparing – homogeneous attrition
- But they can also be different! - heterogeneous attrition
- Older people for instance may not carry on with a demanding task, and if you have a critical comparison between age groups, this can be a major issue

Non-response bias

- You work for a company
- You send out a survey to 1000 randomly selected email ids from your database
- Only 200 respond
- You say you chose the initial emails at random, so what's the problem?
- Again, the people who choose to respond are NOT random!

Regression to the mean

- When you select data based on an extreme value of some measure, a subsequent measurement will tend to "regress to the mean"
- Good examples in the textbook
- The children of tall people will tend to be taller than average but shorter than the parents but the children of short parents tend to be taller than the parents.
- Early studies suggested that people learn better from negative feedback than positive feedback
- But not really, it was also an artifact of regression to the mean (Kahneman & Tverseky, 1973)

Experimenter Bias

Oskar Pgungst: student at the Psychological Institute at the University of Berlin, through careful experiments, showed that Clever Hans was responding to subtle, involuntary cues from von Osten. Classic early example of experimental design in behavioral Psychology

Demand and reactivity effects

"Hawthorne" effect

The influence of lighting on factory worker productivity

But results were driven by the fact that workers did better when they thought they were being observed

Solution to both experimenter bias and reactivity effects

Double blind studies

Placebo effects

- The expectation of a positive effect even from an inert drug will sometimes make people feel better

Fraud and deception

- This part is important as they are very much related to statistical methods, inappropriate use of methods (sometimes intentionally, in order to deceive)

Data fabrication

See https://retractionwatch.com/

People make up data! Including some very high profile researchers

There are data science sleuths who detect fraud using statistical methods

Study misdesigns

Issues with study design that don't get reported

Results may be artifacts of such misdesign

e.g. surveys that are self-evident, sit back and let reactivity decide your results for you. If reviewers don't see the full surveys, this may not get detected

Data mining and post-hoc hypothesizing

- Data mining: I run 50 different variations of a model. Report only the one that worked.
- If you are honest, your statistical methods would "correct" for the 50 times you touched the data because we want to know that the result obtained is a true one that is not likely to have come about due to mere chance.
- Post-hoc hypothesizing: my initial hypothesis didn't work but as part of the data mining effort above, I found something else and reported that I had actually hypothesized it.
- Huge statistical issue when you do this because many frequentist statistical methods depend on assumptions made about the null hypothesis

Publication Bias

- Journals as well as authors do not publish negative findings
- Distorts the literature which comes to be dominated by small N but "significant" studies
- Partly led to the "replication crisis" in Psychology
- Also limits what you can learn from meta-analyses/reviews.

Summary

1. Be aware of all the different ways in which the data from a study may have issues with reliability/validity
2. Be aware of potential confounds
3. Address the confounds using statistical methods
4. Be aware of dubious practices such as data mining and post-hoc hypothesizing

Advanced topics

Article | Open Access | Published: 12 November 2020

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

Gareth J. Griffith, Tim T. Morris, Matthew J. Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, Jonathan Sterne, Tom M. Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil M. Davies & Gibran Hemani

Nature Communications **11**, Article number: 5749 (2020) | Cite this article

39k Accesses | **159** Citations | **334** Altmetric | Metrics

Abstract

Numerous observational studies have attempted to identify risk factors for infection with SARS-CoV-2 and COVID-19 disease outcomes. Studies have used datasets sampled from patients admitted to hospital, people tested for active infection, or people who volunteered to participate. Here, we highlight the challenge of interpreting observational evidence from such non-representative samples. Collider bias can induce associations between two or more variables which affect the likelihood of an individual being sampled, distorting associations between these variables in the sample. Analysing UK Biobank data, compared

Install R and RStudio

- http://cran.r-project.org/
- RStudio: http://www.RStudio.org/