Big Data and Stroke

6th April 2022

Dr Benjamin Bray MBChB MD MPH FFCI

Principal, Health Analytics, Lane Clark & Peacock

Honorary Senior Clinical Lecturer, King's College London

→ ben.bray@lcp.uk.com

There is no formal definition for what counts as big data in stroke research

Your smartphone has space to store ~200 GB of data

Surveys

Small observational studies

Often much less than ~ 1 GB

Clinical trials

Large electronic health records databases ~2-50 GB

Large real-world studies

25 brain CTs ~ 1 GB

Brain imaging data

Whole genome sequencing data

1 whole genome ~ 150 GB

Increasing size

Data is "big" when the amount of data starts to become a problem!

There are five main sources of big data used for stroke research

Registries and biobanks

A registry is an ongoing record about a **health condition within a specific population**

Observational cohorts

A **research study** set up to follow up study participants over time and observe their health outcomes

Administrative data

Data collected to **manage healthcare systems** or monitor the health of populations

Electronic health records

Data collected to **manage patients' healthcare**

Patient generated data

Data **generated directly by patients**, typically using apps or devices

Different types of data have different strengths and weaknesses

	Considerations	Examples
Registries and biobanks	• Might have been set up specifically for research (e.g. many biobanks) but might also have been set up for non-research reasons (e.g. healthcare quality improvement) • Biobanks may contain unique sources of data (e.g. tissue samples, 'omics) • May have limited follow up • Often smaller sample size	• Riks-Stroke (Sweden) • SSNAP (UK) • Get With The Guidelines (USA) • UK Biobank (UK)
Observational cohorts	• Set up for research • Typical focus on epidemiological questions	• Rotterdam Study (NL) • Framingham Study (USA)
Administrative data	• Main focus is on capturing healthcare activity (e.g. hospitalisations, GP attendances, surgeries) • Useful for studies with a health economic focus (healthcare resource utilisation) • Often very large sample size	• Claims Databases (US, Germany) • SNDS (France)
Electronic medical records	• Clinical records generated at the point of patient care • May include data on diagnoses, prescriptions, clinical events, lab data, imaging • Data not captured in a structured format may need to be enhanced or made usable (e.g. natural language processing, machine learning for brain CT and MRI images)	• CPRD (UK)
Patient generated data	• A growing area of data but not very well established so far • Electronic surveys, social media data, smartphone apps, devices e.g. smart watches	• Apple ResearchKit (Global)

What sort of research questions are suitable for a big data observational study?

- **Epidemiology** of disease (e.g. risk factors, predictors of stroke outcomes)
- Understanding how patients are **treated and their outcomes in real world care**
- **Comparative effectiveness** of treatments
- **Safety** and adverse events of medicines (especially rare events)
- Understanding the **effectiveness of interventions that are difficult to test in a RCT**
- Informing **trial design** and conduct (e.g. power calculations, site selection)
- **Hybrid RCTs** (e.g. single armed trials or pragmatic trials using real world data collection)
- Studies about **rare events** or uncommon side effects of treatments

Big data can be frustrating and has a lot of limitations for research

You have **no control** over what data items are collected

The data is much "**messier**" than data from a trial - lots of data cleaning required

Need to handle **missing data** correctly

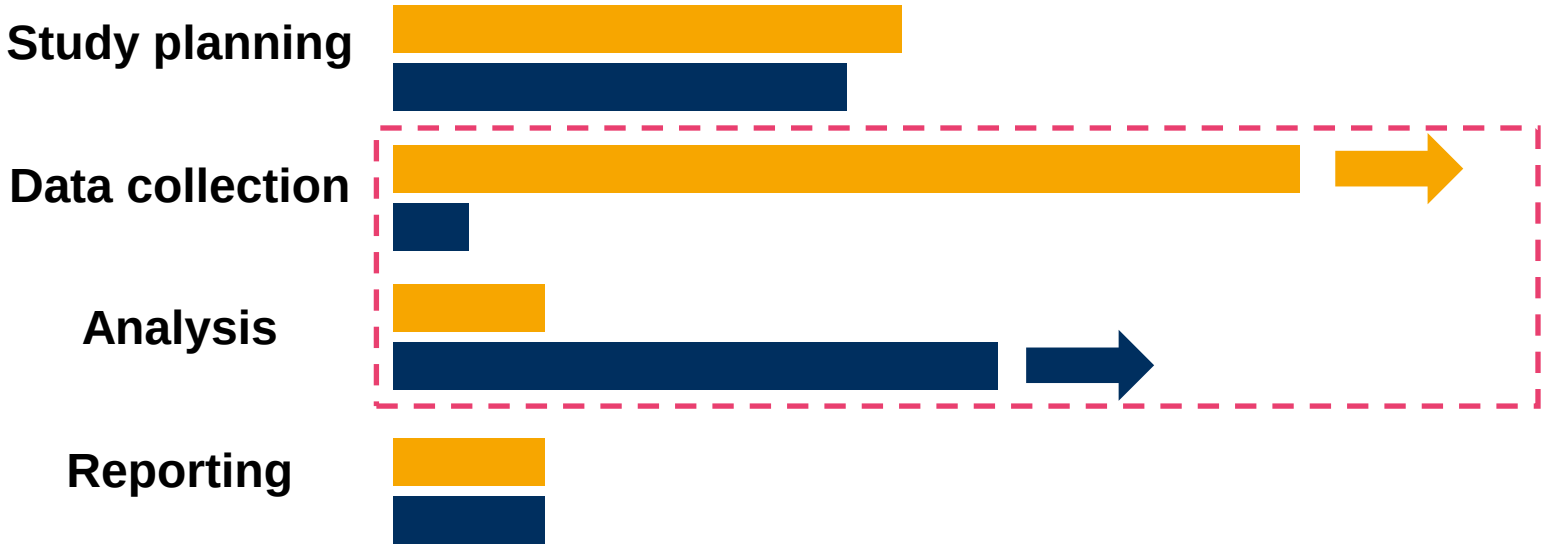
Accessing the data and getting permission to use it for your study can be very time consuming

Need to put a lot of thought into the **study design and interpretation**…*doing bad research is easier when you have huge sample sizes!*

Your ideal endpoints are almost certainly **not going to be available** in the data

Large size can lead to **practical challenges** with storing and/or analysing the data

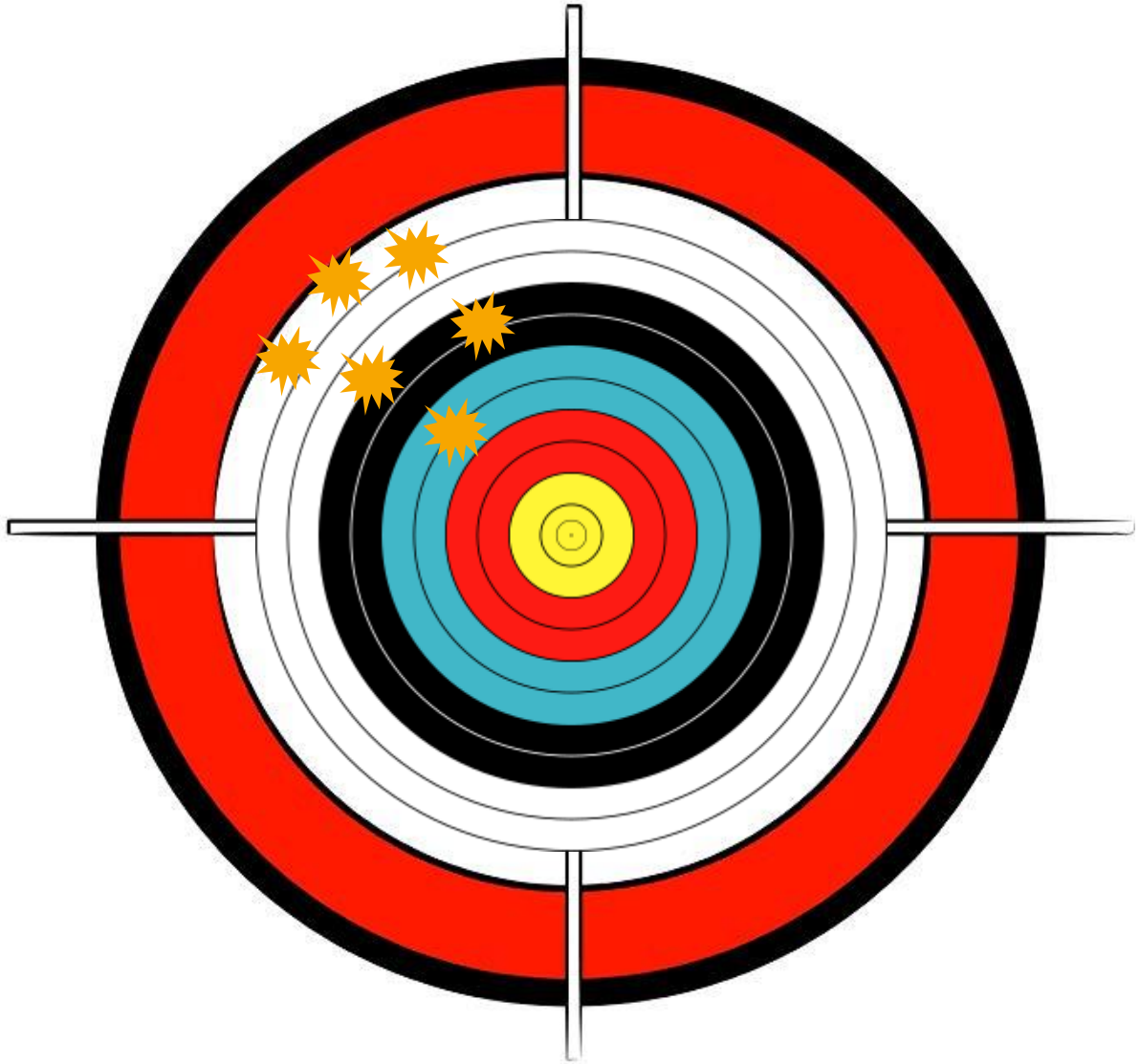
Big data studies are typically quicker and much less expensive than RCTs, but often need more complex data analysis



Studies are not necessarily more reliable or correct when they use big data

LCP INSIGHT CLARITY ADVICE

! *Large sample sizes **increase precision** but **do not reduce bias***



Recommend using the "target trial" methodology when designing observational studies addressing a causal question

"If this was an RCT, how would I design this? Can I replicate the features of the RCT using observational data?"

- Define appropriate inclusion and exclusion criteria
- Appropriate controls (e.g.don't include patients who would never have been eligible for the active treatment anyway)
- Correct identification of key points in the timeline e.g. the time from which you start measuring the outcomes (*the index date*)
- Defining the endpoints accurately
- Reference for more detail:

  Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Hernán MA, Robins JM. Am J Epidemiol. 2016 Apr 15;183(8):758-64*

Some dos and don’ts of big data stroke research

DO	DON’T
✓ Take care to design the study appropriately	X Mine the data for “interesting” results
✓ Build in suitable timelines for data access and analysis	X Use hypothesis tests without a good reason
✓ Pre-register the study protocol	X Forget to involve patients
✓ Use tools like Git Lab to manage and share your programming code	
✓ Find a good team to work with – data science is a team sport requiring a range of skills (domain/clinical expertise, epidemiology, statistics, programming)	
✓ Follow the relevant EQUATOR Network reporting guidelines: https://www.equator-network.org/	
✓ Publish your disappointing/negative findings	

Questions?