INVITED FEATURE: STATISTICS PRIMER

What is 'big data' and what do biostatisticians and data managers want you to know about it?

Andrew R Gray, Dave Barson, Claire M Cameron, Ari Samaranayaka, Robin M Turner

Introduction

Imagine that you're a general practitioner and are seeing a new patient. They have some important questions for you about their health. Luckily for you, they have arrived at their appointment with a very large collection of data to help. This includes:

1. results of many laboratory tests, some recent and some from long ago (most from a reputable local laboratory; some consumer point of care, like spot glucose; and some that look to have been ordered over the internet and that you're not familiar with);
2. several years' worth of heart rate, SpO_2, sleep, activity, and location data from their wrist-worn tracker and smart device;
3. an assortment of images (including photographs of them as a young child, just in case they are useful, alongside a recent MRI and an impressive collection of X-rays—where did they get all of these from?); and
4. months of audio recordings of their sleep (again, just in case).

They also have less detailed data on half a dozen family members in case it's genetic or caused by their shared environment. Every few minutes, they remember something else. They either email it to you so you can look at it later or bring it up on their device to show you in person. This reminds them to show you their current heart rate, which is becoming increasingly elevated.

Here you have large volumes of data that are of a variety of types and coming in at great velocity. This is indicative of the modern world, where a lot of data is being captured, often in detail, even if we're not fully aware of it. We need to carefully think about its relevance and its quality, as well as its ethical implications. It would take a lot of time to try to work through which parts of this patient's data were useful and what they indicate. And this is just for this one patient, let alone the many patients you might see in a day.

It's easy to feel overwhelmed in an analogous research situation where we have what has been evocatively named 'big data'. This term has been in use since the late 1990s. It means different things to different people and its meaning has evolved over time. By the end of this (not big) primer, we hope that you feel more comfortable identifying big data, understanding its opportunities and risks, and deciding whether and how it can help you in answering your own research questions.

Defining and identifying 'big data'

What exactly does 'big data' mean to data managers and biostatisticians? We could start with the traditional view, which is based around the three Vs: volume (how much), variety (how many types), and velocity (how fast the data are coming in).[1] Not all of these are required in any given instance. They have since been joined by a selection from veracity, value, variability, vulnerability, and others. As far as we can tell, the current record number of words starting with V that have been collected into a list for big data is 56![1]

We pragmatically define big data to mean data that overwhelm current quantitative approaches. In short, we're going to have to develop some aspect(s) of our methods or technology as we go. What is today's 'big' data could be next year's 'normal' once we have well-established methods to handle them. This will become evident from some history discussed below.

There is no one definition of big data,[2] but a useful starting point is: "data sets…so large and complex that they require advanced…data storage, management, analysis, and visualization technologies."[3] A less obviously helpful quote, but still applicable in some instances, is: "a dataset that is too big to fit on a screen."[4] Finally, "a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology that provokes extensive utopian and dystopian rhetoric"[5] hints at the important social and ethical concerns about big data outside of collecting, managing, and analysing it.

Big data throughout history

While the term 'big data' is relatively recent, the idea of being overwhelmed by or uncertain about what to do with data has a long history. John Snow (not the one from Game of Thrones) looked at cases of cholera in London in 1854 by plotting them on a map showing the local water pumps. This could be seen as big data in action. This visualisation was not a well-established method at the time. Other early examples include Graunt's Bills of Mortality, Halley's lifetable, and Florence Nightingale's analyses of Crimean War hospital data. As a particularly clear example, the United States Census of 1880 took about eight years to process, leading to estimates of up to 13 years for the 1890 Census, given the higher population. The following census was planned for 1900, so there was some time pressure! The development of Hollerith's electromechanical tabulator led to the processing taking about two years. And if we want to go further back, there's the example of the Babylonian Astronomical Diaries which collect around 600 years of data (c. 652 BCE–61 BCE for surviving tablets) on weather, current events, grain prices, and many other topics. How do you analyse data like that?

In these cases, methodological advances were needed for data that was too much, too complex, came in too fast, or was otherwise unmanageable. The important work done by Snow, Graunt, Halley, or Nightingale can now be reproduced using off-the-shelf software alongside standard epidemiological and statistical methods. These examples would no longer qualify as 'big data'—we have the methodology!

Along with methodological developments, modern hardware and software have contributed to solving some previous big data problems. For example, the development of faster computers, better algorithms, and more sophisticated statistical software allows us to use advanced statistical methods with a copy of Stata.[6] When some of us started our training in statistics, some analyses that are now straightforward to perform were simply infeasible.

Contemporary examples of 'big data'

The UK Biobank and Aotearoa New Zealand's (A-NZ) Integrated Data Infrastructure (IDI) are two excellent examples of contemporary big data. These can require us to develop new approaches to answer our research question. The IDI is a collection of linked administrative datasets. The UK Biobank was created to facilitate research, but without specific research questions in mind. What both datasets have in common is that they are large, broad, evolving, and challenging to work with. They require careful planning and thoughtful weighing up of strengths and limitations. See Table 1 for a brief overview of some of the features of each.

If you're interested in publications that have used these datasets, you could start with some from us and our colleagues using the UK Biobank[7] and the IDI[8].

Another contemporary dataset used in research is InterRAI. This is included in the IDI and is also available separately. This dataset records assessments for people before they receive home or residential care, and continues while they remain in care. InterRAI contains hundreds of thousands of assessments. This is definitely a very large dataset. However, it is one where we already have the required methods, so not what we would usually describe as big data.

What big data do well

Big data often contains data on large numbers of people. This sometimes allows researchers to get very precise answers to research questions, such as investigating associations between diseases and exposure/risk factors. This can be true even for rare exposures or outcomes because of the sheer size of the dataset. It can also allow us to integrate seemingly incompatible or normally unrelated sources of data. Creating a study to collect these data could be too expensive and time-consuming. In other cases, we could extract or process data in ways we couldn't realistically do by hand.

There are many important questions that we cannot investigate without grappling with big data. As a first example, understanding a rare genotype (say with 0.05% prevalence, equating to around 250 people in the UK Biobank) and its links to relatively common health conditions cannot be done with a standard cross-sectional or case-control study. As a second, quantitatively exploring intergenerational economic privilege in A-NZ and its links with health, both overall and by ethnicity, would be very challenging without the IDI. A third example, inspired by a study the first author is involved with, involves video data. In this case, wearable cameras were used to record activity, diet, and social interactions. The goal is to understand how these relate to sleep quality. This would present substantial challenges even for modest numbers of participants. Manually coding the videos by watching them quickly becomes laborious. Being able to automatically code these events from video footage would be a game-changer. We still don't know exactly how to work with video data in some situations, which makes this big data, according to our earlier definition.

The results from statistical analyses involving big data might look like those from more conventional analyses (see our earlier articles on p-values[9] and confidence intervals[10]), but they could also require coming up with new ways of visualising data or results.

When small/medium-sized data is better

Big isn't always better. Having data on hundreds of thousands or millions of people sounds impressive. However, the recruitment of wealthier and healthier volunteers to the UK Biobank has led to some spurious associations between variables. The much smaller Dunedin Study,[11] which started with 1037 children, maintains outstanding relationships with the Study members, who come back for regular high-quality and detailed assessments. This minimises the impact of many biases.

Smaller and simpler data might also allow you to use simpler statistical approaches. You'll often hear advice from biostatisticians regarding the design of your study, including the sample size.[12] This includes to make the data management and analysis that answers your research question(s) as simple as possible.

Other challenges with 'big data'

When analysing datasets, data managers, biostatisticians, and subject-matter experts all need to think carefully about how the results of analyses could be misused. You might have heard of Māori data sovereignty. This is the recognition that Māori data should be governed by Māori. Indigenous statistics is an important topic. While the IDI can be used to address important questions about Māori, it could also be misused. Big data doesn't create problems here in itself, but it does magnify the potential adverse consequences if study results are used inappropriately. For data that involves Māori, we strongly recommend building a relationship with Māori to minimise the risk of ethnicity

Table 1: Selected 'V' characteristics for the UK Biobank and the Integrated Data Infrastructure.

	UK Biobank	Integrated Data Infrastructure
Volume (variables × participants × repeated measures)	Data on about 500,000 participants, with questionnaire, healthcare, genomic, and other sources of data, including 15 million biological samples. The full dataset takes over 30 million gigabytes to store.	Longitudinal data on over nine million people who were at some stage resident in Aotearoa New Zealand. This involves core administrative collections and surveys on topics such as health, education, and justice. These mostly start in the 1990s or early-2000s, but births, deaths, and marriages go all the way back to 1840. Other sources have only been added recently.
Variety	A range of data types are collected, including standard text and numbers, images, and genomic data.	More like the types of data we're used to working with in spreadsheets.
Velocity	New releases are roughly every three months, usually based on a theme. There are some updates in between these as well. Participants can withdraw their data.	Refreshed around three times per year with some additional updating in between.
Veracity and validity	Well-established data cleaning rules are routinely applied. Data can be easily linked through the design of the study.	Linkage of data is both deterministic (identifiers exactly match) and probabilistic (the identifiers probably refer to the same individual). Linkage is not perfect, and some data cannot be reliably connected.
Value	Over ten thousand publications have been generated using these data, suggesting that it has helped many researchers to address their research questions.	Allows many questions to be addressed that would otherwise be impractical/impossible.

being used in ways that could cause problems. Similar points can be made for other groups based on ethnicity, gender, or otherwise.

Understanding how data were collected is crucial. This is especially so if you weren't the one who collected them. You need to think carefully about whether your sample represents the population you want. You also need to consider what biases might have been introduced. The quantity of data cannot be converted into quality.

Google Flu Trends (GFT) was an algorithm launched in 2008, where Google searches that involved flu-related symptoms and other terms were used to estimate actual cases of flu. But people changed their searches over time. This included feedback loops following media reports on GFT predictions. GFT was eventually 'retired' after a particularly disastrous failure with the 2012–2013 flu season. In hindsight, it deserves the accusation of "big data hubris."[13]

Tips when faced with 'big data'

Here are our top six tips:

1. Be very clear about what you want to do rather than what you can do. The temptation when faced with a mountain of data is to get as much as you can out of it. The research consequences of this approach include finding misleading or spurious correlations.
2. Make sure you have the right people in your team. Big data is an excellent opportunity for collaboration between researchers. This includes subject-matter experts, data managers, and biostatisticians. Some big data projects will need further researchers, such as Māori health experts, health economists, or machine learning experts.
3. It is important to plan to analyse your data according to what you want to achieve. If your focus is on understanding a causal link between an exposure and a disease, traditional statistical approaches are generally much easier to examine than machine learning approaches.
4. We recommend that you look for opportunities to simplify as much as possible while still addressing your research question. An accretion (layer-by-layer consolidation) approach to building up datasets can be helpful here. This means starting with two data sources, solving the problems of using them together, and afterwards, adding another dataset. You can keep repeating this process until you have everything working together or have exhausted your methods.
5. No matter whether it's big, medium-sized, or small data, think carefully about potential biases and how these could affect the analysis. High-quality boutique datasets can provide better answers to some questions than studies that are one, two, three, or more orders of magnitude larger.
6. Be prepared for the quantitative parts of your project to take longer than initially planned. Challenges often arise. The point of big data is that we're working out how to best deal with it as we go. The methods we need follow the data rather than already being in place.

What next?

We've stopped talking as much as we did back in the 1990s and 2000s about 'big data'. Having more challenging data than we initially know how to make best use of has possibly become the new normal.

When planning and designing your research, we strongly recommend that you focus on quality over quantity. You still need sufficient data, but more isn't always better. It is important that you acknowledge the limitations inherent in all large and complex datasets. And, of course, we recommend that you collaborate with a data manager and a biostatistician (and other experts as appropriate).

Remember to have the right multidisciplinary team in place, that quality wins over quantity, and plan before you analyse. If you follow this advice, you'll give your team the biggest chance of research success.

References

1. Hussein AA. Fifty-six big data V's characteristics and proposed strategies to overcome security and privacy challenges (BD2). J Inf Secur. 2020;11(4):304-28. https://doi.org/10.4236/jis.2020.114019

2. Favaretto M, De Clercq E, Schneble CO, Elger BS. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. PLoS One. 2020;15(2):e0228987. https://doi.org/10.1371/journal.pone.0228987

3. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart. 2012;36(4):1165. https://doi.org/10.2307/41703503

4. Shneiderman B. Extreme visualization: squeezing a billion records into a million pixels. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York (NY): Association for Computing Machinery; 2008. p. 3-12. https://doi.org/10.1145/1376616.1376618

5. Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc. 2012;15(5):662-679. https://doi.org/10.1080/1369118x.2012.678878

6. Gray A, Cameron C, Samaranayaka A, Turner R. What statistical software should I use? And does it actually matter? NZ Med Stud J. 2022;(34):38-40. https://doi.org/10.57129/gadb4864

7. Gulick CN, Peddie MC, Cameron C, Bradbury K, Rehrer NJ. Physical activity, dietary protein and insulin-like growth factor 1: cross-sectional analysis utilising UK Biobank. Growth Horm IGF Res. 2020;55:101353. https://doi.org/10.1016/j.ghir.2020.101353

8. Satherley N, De Graaf B, Davie G, Gibb S, Teng A, Sporle A. Applying indigenous identity definitions in official health statistics: a case study using linked cancer registry data on stomach cancer. NZ Med J. 2025;138(1614):81-90. https://doi.org/10.26635/6965.6844

9. Cameron C, Samaranayaka A, Turner RM. P values: what is their significance? NZ Med Stud J. 2020;(31):48-49. https://doi.org/10.57129/gfju5536

10. Cameron C, Turner R, Samaranayaka A. Understanding confidence intervals and why they are so important. NZ Med Stud J. 2021;(33):42-3. https://doi.org/10.57129/agag5939

11. Poulton R, Guiney H, Ramrakha S, Moffitt TE. The Dunedin study after half a century: reflections on the past, and course for the future. J R Soc NZ. 2023;53(4):446-65. https://doi.org/10.1080/03036758.2022.2114508

12. Samaranayaka A, Cameron C, Turner RM. Sample size in health research. NZ Med Stud J. 2021;(32):52-4. https://doi.org/10.57129/xpsc9819

13. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science. 2014;343(6176):1203-5. https://doi.org/10.1126/science.1248506

About the authors

› Associate Professor Andrew Gray, BCom(Hons), BA, is a Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Dave Barson, BA, PGDip, DipGrad, is a Senior Research Fellow and Data Manager and Programmer in the Department of Preventive and Social Medicine, Division of Health Sciences, University of Otago.
› Associate Professor Claire Cameron, BSc(Hons), DipGrad, MSc, PhD, is the Director of the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Associate Professor Ari Samaranayaka, BSc, MPhil, PhD, is a Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Professor Robin Turner, BSc(Hons), MBiostat, PhD, is the Head of Department of Preventive and Social Medicine, Division of Health Sciences, University of Otago.

Acknowledgements

Not applicable

Conflicts of interest

Not applicable

Funding

Not applicable

Correspondence

Associate Professor Andrew Gray: andrew.gray@otago.ac.nz