BIG DATA BASICS

National Collaborating Centre for Infectious Diseases
Centre de collaboration nationale des maladies infectieuses

In public health, the use of big data for planning is still emerging and questions are common. Here, we begin with basic questions: what is 'big data', what sets these data apart from other data, and how are big data being used?

WHAT IS 'BIG DATA'?

Big data are not new but continue to change with advancements in technology. This has led to many different definitions and a lack of conceptual clarity. One common definition for big data is this:

Very large and diverse datasets that are analysed computationally at high velocity to reveal patterns, trends and associations.[1,2]

WHAT SETS BIG DATA APART? – THE '3 Vs'

Most big data definitions refer to three characteristics.

1 VOLUME

A continuously increasing amount of data can be collected and stored. [1,2,3]

New technology has enabled researchers to collect and analyse massive datasets.

In 2017, more than 2.5 quintillion bytes of data were created each day, which was 300 times more than in 2005.[4]

2 VARIETY

There is an abundance of big data sources, which fall into 3 categories: [1,2,3]

Health data can now be extracted from a variety of sources: The Internet of Things, mobile devices, social media, search engine queries, or e-commerce data.[5]

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

E.g. Excel files of pre-defined fields with demographic information or location data from cell phones.[6]

E.g. Photos, videos, audio files, social media content, websites, satellite imagery, open-ended survey responses.[6]

Incorporating several datasets within an analysis allows researchers to enrich their findings and identify new trends and associations.[2]

Less than 20% of all data are structured and unstructured data make up an increasing proportion.[6]

3 VELOCITY

The speed at which data are generated is high.[1,2,3]

Cell phones and wearable devices provide high velocity, continuously produced data collected in real time.[7]

Data are generated at vastly different speeds, depending on the source.

HOW ARE BIG DATA USED?

HIGH VOLUME DATA

A recent study that analysed the content of 2.8 million tweets related to the COVID-19 pandemic helps public health to understand the public's information needs and respond faster and more appropriately.[8]

MULTIPLE DATA SETS

Studies that bring together data on commercial air travel, location and mobility of people, and travel restrictions help predict the spread of COVID-19 and assess the effects of travel restrictions compared to other public health measures.[9,10]

HIGH VELOCITY DATA

A study that analysed cell phone data helps public health unravel the relationship between peoples' movements and socioeconomic factors, and to understand how public health restrictions may affect different populations differently and at different times.[11]

REFERENCES

1. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu. Rev. Public Health. 2018 Apr 1;39:95-112.
2. Fuller D, Buote R, Stanley K. A glossary for big data in population and public health: discussion and commentary on terminology and research methods. J Epidemiol Community Health. 2017 Nov 1;71(11):1113-7.
3. Ylijoki O, Porras J. Perspectives to definition of big data: a mapping study and discussion. J. Innov. Manag. 2016 May 4;4(1):69-91.
4. Herschel R, Miori VM. Ethics & big data. Technology in Society. 2017;49:31-36.
5. Jia Q, Guo Y, Wang G & Barnes SJ. Big data analytics in the fight against major public health incidents (including COVID-19): a conceptual framework. Int. J. Environ. Res. Public Health. 2020;17(17): 6161.
6. Marr B. What's the difference between structured, semi-structured and unstructured data? Forbes. 2019 Oct 18.
7. Badr HS, Du H, Marshall M, Dong E, Squire MM, & Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. The Lancet Infect. Dis.. 2020; 20(11):1247-1254.
8. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J.Med. Internet Res. 2020;22(4):e19016.
9. Watts A, Au NH, Thomas-Bachli A, Forsyth J, Mayah O, Popescu S. & Bogoch II. Potential for inter-state spread of Covid-19 from Arizona, USA: analysis of mobile device location and commercial flight data. J. Travel Med. 2020, 1–3.
10. Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395–400.
11. Long JA, Ren C. Associations between mobility and socio-economic indicators vary across the timeline of the Covid-19 pandemic. Computers, environment and urban systems. 2022 (91)1.