Big Data Reality Check (BDRC) for public health: to what extent the environmental health and health services research did meet the ‘V’ criteria for big data? A study protocol

Pui Pui Tang 1, I Lam Tam 1, Yongliang Jia 2,3, Siu-wai Leung 4,5

ABSTRACT

Introduction Big data technologies have been talked up in the fields of science and medicine. The V-criteria (volume, variety, velocity and veracity, etc) for defining big data have been well-known and even quoted in most research articles; however, big data research into public health is often misrepresented due to certain common misconceptions. Such misrepresentations and misconceptions would mislead study designs, research findings and healthcare decision-making. This study aims to identify the V-eligibility of big data studies and their technologies applied to environmental health and health services research that explicitly claim to be big data studies.

Methods and analysis Our protocol follows Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P). Scoping review and/or systematic review will be conducted. The results will be reported using PRISMA for Scoping Reviews (PRISMA-ScR), or PRISMA 2020 and Synthesis Without Meta-analysis guideline. Web of Science, PubMed, Medline and ProQuest Central will be searched for the articles from the database inception to 2021. Two reviewers will independently select eligible studies and extract specified data. The numeric data will be analysed with R statistical software. The text data will be analysed with NVivo wherever applicable.

Ethics and dissemination This study will review the literature of big data research related to both environmental health and health services. Ethics approval is not required as all data are publicly available and involves confidential personal data. We will disseminate our findings in a peer-reviewed journal.

PROSPERO registration number CRD42021202306.

INTRODUCTION

Environment affects human health

The quality of the environment is a powerful determinant of human health. Environmental health refers to the influence of environmental factors on human health.1 2 The environmental factors that are attributable to diseases of high global burden have been classified into physical, chemical, biological, social and other factors.3 Recent studies reported that 25% of deaths between 2006 and 2016 were attributable to environmental causes.4 5 Environmental health once focused on the prevention of infectious diseases; however, the focus has shifted toward chronic diseases such as cardiovascular diseases.6 The WHO estimated that ischaemic heart disease is the biggest killer worldwide (16% of total deaths) in 2019.7 The reduction of environmental risks such as PM2.5 levels in the air was associated with decreases in cardiovascular mortality.8 Although the US Environmental Protection Agency was established in the 1970s and aimed to manage air, water and soil pollution control and remediation, air pollution continues to be the largest environmental health risk to public health in the US.9 The simple statistical learning methods to study the linkage of environmental health risk factors and potential diseases have been replaced by the big data analysis methods, which improves the accuracy of evaluation, evolves our understanding of the relationship between environmental and human health, and enhances public health policies that
reduce populations at risk. For example, since previous studies demonstrated that reduction in air pollution exposure can improve respiratory health, a recent study used a deep neural network method to identify the environmental health risk factors of acute respiratory diseases.\(^{10}\) Additionally, a recent study launched a national long-term birth cohort study that aimed to use big data to investigate environmental influences on children’s health and address the need for public health strategies to reduce the burden of environment-related diseases.\(^{11}\)

### Health services in public health

The implementation of health services focuses on individuals, and the target of public health is the health condition of the overall population.\(^{12-15}\) Offering health services inside or outside the clinical stage to the population is a critical strategy for developing and maintaining public health.\(^{15}\) Health services need to be sufficiently available, high-quality, cost-affordable, and accessible.\(^{16-18}\) Both health services and public health strive to prevent and intervene against acute diseases, chronic diseases, injuries, and risks.\(^{12}\) Diverse health services (eg, data monitoring, primary healthcare, illness screening, injury protection, drug development, and health promotion) and professional personnel collaboratively benefit the health of the public, and the public health services system plays an irreplaceable role in the management of individual health.\(^{12,16,19}\)

Data regarding incidence, morbidity, prevalence, mortality, and rate of recovery that are obtained from the health services system not only inform the data monitoring and surveillance in epidemiology, but also serve as references for the development of health promotion policies and strategies.\(^{14}\) The construction of sufficient and efficient health services is necessary for response and recovery during public health emergencies.\(^{18}\)

### What is big data?

The term big data was first used at Silicon Graphics in the mid-1990s according to Diebold,\(^{20}\) Cox et al\(^{21}\) first used big data in publications on data-intensive computing in 1997. The term big data was defined in various ways in recent publications. Lane\(^{22}\) highlighted three dimensions (volume, velocity, and variety) of big data in a research note in 2001. Moreover, De Mauro et al\(^{23}\) characterised big data as high volume, high velocity, and high variety (the 3Vs criteria). Specially designed technologies and analytics are required to transform such data into valuable information.

### Use of big data in environmental health research

Data-intensive research in environmental health has grown,\(^{24}\) especially over the last two decades.\(^{25}\) These studies require large datasets. For instance, studies on air pollution\(^{26}\) have processed terabytes of data to identify air pollution as the largest killer worldwide.\(^{27}\) Meanwhile, deaths due to chemical pollution and soil pollution have also been increasing.\(^{28}\) Studies on mobile health would require several gigabytes of GPS data.\(^{29}\) As big data became a buzzword in public health research, some researchers may not aware of or did not follow its definitions. It is not uncommon for non-trivial datasets to be called big data. As a result, non-Veligible (zero V) big data research on environmental health exists that could mislead healthcare decision-makers. This is the first study that seeks to identify the V-eligibility of big data studies and their technologies applied to environmental health research. It is anticipated that the findings will promote V-eligible big data research on environmental health.

### Use of big data in health services research

Big data and its technologies change rapidly and improve the efficiency of various data workflows, such as data collection, processing, utilisation, and management in health services; examples include electronic health records (EHR), digital health applications, research studies, and so on.\(^{30-34}\) Based on a preliminary search that was conducted in the Web of Science (WOS), 34% of all published big data research in public-health-related categories (ie, primary healthcare, public, environmental and occupational health, healthcare sciences and services, health policy and services) covered health services. Developing big data technologies and architectures such as Internet of Things (IoT) based patient monitoring system,\(^{35}\) which facilitate the continued exploration of health services such as the performance of operations, development of personalised medicine, evaluation of policies, reduction of medical expenses,\(^{36}\) disease prevention, enhancement of overall service capacity and communication among service providers, mediators and receivers.\(^{36,37}\)

Recent research\(^{38-41}\) identified the major challenges of big data including data structure, data security, data standardisation, data storage and transfer, and training of data analysts. For example, data management in health services was hindered by difficulties in sharing EHRs due to slow standardisation, non-compliance with standards, and a lack of expertise. In addition, big data is not easy to manage using traditional database technology designed for well-structured data. Moreover, enabling artificial intelligence investigations into big data analytics would require even better technologies.\(^{30,42}\)

### Big data have outgrown traditional data technologies

Big data analytics, like that for traditional data, aims to generate information from data but with different technologies.\(^{43}\) Even if parts of big data were stored in traditional databases, yet multiple sources and fast growth of such data (eg, from the web, social networks, sensor networks, scientific experiments and others) in terms of types, sizes, timeliness, and complexity would require big data analytics.\(^{44,45}\) Integrating various forms of data, that is, structured, semi-structured, and unstructured data, would also require big data analytics,\(^{46}\) rather than the ordinary Structured Query Language designed for managing relational databases.\(^{43}\) Therefore, data-intensive computing
tools, for example, Hadoop and Spark, are commonly used for big data processing. There have been unstructured data management systems for healthcare data; for example, Luo et al57 developed a double-reading/entry system for extracting key-value data items (ie, structured data) from the unstructured medical records (ie, texts, drawings, laboratory test results and physicians’ notes) and for curating semi-structured EHR databases.

The V-criteria for big data

What makes big data different from ordinary data? The main differences lie in the so-called ‘V’ characteristics of big data. Laney originally suggested that volume, variety, and velocity are the three basic ‘Vs’ that characterise big data, as originally suggested by META Group. Then big data’s characteristics surpassed the 3Vs. Yin et al48 stated that volume, variety, velocity, veracity, and value and are the 5Vs of big data. Andreu-Perez et al49 stated that volume, variety, velocity, veracity, value, and variability are the 6Vs that are applicable to health data research. Among these characteristics, volume, variety, velocity, and veracity constitute the most popular criteria for big data. Therefore, these 4Vs criteria (table 1) were adopted in this review.

Volume, or the magnitude of data generated every second, describes how big the datasets are. The size of big data should span terabytes and petabytes. There have already been data that exceed zettabytes, and the 10 or near real time. In environmental health research, an unbounded sequence of event processing in real time is needed. A data stream is a flow of data generated and analysed to create, capture, process and store data. Given the rapid growth of data integration, a real-time processing solution is needed. A data stream is an unbounded sequence of event processing in real time or near real time.

Variety or the structural heterogeneity of a dataset, refers to the coexistence of various types of data (ie, structured, semi-structured, and unstructured data). In contrast to tabular data and other well-structured data, images, audio, and videos are unstructured data that require customised analysis. Extensible Markup Language and JavaScript Object Notation are useful techniques for managing semi-structured data.

Velocity refers to the speed and rate at which data are generated and analysed to create, capture, process and store data. Given the rapid growth of data integration, a real-time processing solution is needed. A data stream is an unbounded sequence of event processing in real time or near real time.

Veracity, or the trustworthiness of the data, refers to the reliability and provenance of the data. For obvious reasons such as the sheer volume of data, big data must be cleaned and harmonised in an automated manner.

Kitchin et al67 found that few environmental health datasets fulfilled the 3Vs (ie, very large volume, fast and continuous velocity, and wide variety). Some datasets, such as sensors for pollution and sound, only produce gigabytes of data per year. In addition, insufficient data quality, precision, and timeliness of the IoT-generated big IoT data presented challenges to the processing of big data. Although social sensing has widespread usage in mobile devices, its reliability (ie, veracity) still requires slow human verification.

Big data research in medicine is often misrepresented

Traditional IT approaches to data management are no longer suitable for managing large unstructured data and processing data-intensive tasks. For example, Excel and SPSS were not designed for handling big data but are often used in non-V-eligible big data research as the main (if not only) data management and processing tools. Common visualisation software (eg, Pajek and Cytoscape for visualisation of the social/biological networks of above-average sizes) was not designed for big data, but were often mislabelled in non-V-eligible research as big data tools. Genuine big data research requires special data engineering to facilitate the acquisition, access, processing, analysis, mining, modelling, and so on. In particular, ensemble analysis, association analysis, high-dimensional analysis, deep analysis, precision analysis, and divide-and-conquer analysis have been proposed as the six major strategies and technologies that enable big data research.

Objectives

This study aims to identify the V-eligibility of big data studies on and their technologies applied to environmental health and health services research that explicitly claimed to be big data studies.

METHODS

This protocol follows Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) and was registered with the PROSPERO International Prospective Register of Systematic Reviews, dated 4 January 2021. Based on the actual extractable data types from the included studies, scoring review and/or systematic review will be compared in their V-eligibility. Wherever applicable, the reporting of results will follow PRISMA for Scoping Reviews (PRISMA-ScR), PRISMA 2020 and Synthesis Without Meta-analysis (SWIM) guideline. The PRISMA-P checklist for this protocol is available as a separate supplementary document (online supplemental appendix 1). The actual start date of this
review was 1 December 2020. The anticipated completion date is 31 March 2022.

Study selection criteria
Human research articles that explicitly claimed to be big data research on environmental health or health services will be included. The selected studies will cover all types of research, including exploratory research, descriptive research, and experimental research. Non-human studies (e.g., animal or plant studies) will be excluded. Non-original research articles including reviews articles (e.g., literature reviews, systematic reviews, meta-analyses), book reviews, editorials, commentaries, expert opinions, monographs, case reports, case series, protocols, debates, meeting reports, guidelines, subject indexes, round table reports, and forum reports will be excluded. No restrictions will be imposed on the participants, interventions, or comparators. Besides, the eligible article must be written in English and accessible online with full text.

Data sources and search strategy
The search for environmental health research
Literature databases, including the WOS, PubMed, Medline (EBSCOhost interface), and ProQuest Central, will be searched, through the application of a search strategy specified in previous reviews that covered the same field of environmental health research.68 69 Search terms will include “environmental health”, “environmental exposure”, “environmental illness” and “environmental epidemiology”, and the term “big data”. No restrictions will be imposed to study type, language, and timespan. The full search strategies for different databases are available (online supplemental appendix 2).

The search for health services research
There are two major keywords: (1) “big data” and (2) “health services”. All conceptual categories will be formulated from keywords with the Boolean operators. The studies which meet at least one V (volume, variety, or veracity) will be included for information extraction and analysis. The eligible category of studies will be assessed for their V-eligibility. Zero V studies and the studies which do not claim to use big data will be excluded. A flow diagram for study search and selection is provided in figure 1.

For the studies that are: (1) not clearly indicated in terms of volume or less than a terabyte in volume; (2) used very simple formats (e.g., 2-D tables) of data, or only one data source; (3) not real-time or near real-time in data generation; (4) not trackable or trustworthy in data provenance, questionnaires will be sent to the authors for confirmation. The authors will have one month to reply email and/or return questionnaires.

Information extraction
The studies which meet at least one V (volume, variety, velocity, or veracity) will be included for information extraction and analysis. The text of the following six categories will be extracted from the included studies:
1. Publication information: authors(s) names, article titles, publication years, countries of the first author's research centre or organisation, and journal titles;
2. Types of articles: full research articles, short research reports, and methodology papers;
3. Vs criteria for big data: data characteristics, data collection and processing details (sample sizes, data pro-

Figure 1 Flowchart diagram for study selection.
cessing speeds, accuracy measures etc.), and analysis methodologies;

4. Categories of big data applications (examples): environmental health (major topics as covered by environmental health literature, eg, environmental pollution, environmental health hazards, environmental exposures, environmental diseases, work environment and health, vulnerable groups, and environmental health risk assessment); health services (genomics, elderly care, mental health, personalised healthcare, drug discovery, clinical research, financial benefits, etc);

5. Claimed big data techniques (examples): cluster analysis, data mining, graph analytics, machine learning, natural language processing, neural network, pattern recognition, and spatial analysis, etc;

6. Data sources (examples): electronic health records (EHR), biomarkers, health insurance claims, clinical trials, social media, wearables, sensors, and so on.

Subgroup analyses will be performed if sufficient data are available. Possible subgroups include, for example, big data applications, big data techniques, and big data sources.

Any disagreements in the information extraction process between two reviewers will be resolved through discussion with the third reviewer. Examples of the
information to be analysed by cross-tabulation are as shown in table 2.

Moreover, Nvivo, a qualitative data analysis software, will be used to manage and analyse the text data. Relevant and distinctive text excerpts will be coded as themes (nodes) for comparison among studies. Synonyms with varied morphology will be automatically handled in identifying common themes. Then, the software will perform proper text analysis according to the identified themes.

Outcomes

The primary outcomes will be categorised into types of study design, subgroups, and the V-eligibility of the included studies. The secondary outcomes will be the types of big data analytics used in the included studies. A specification of our data analysis plan is given below.

Data synthesis and analysis

Categories regarding the outcomes above will be collected and analysed by basic descriptive statistics with R software. The results will be summarised in table 3.

The text data (ie, characteristics and analysis technologies of big data) will be qualitatively analysed with Nvivo. The eligibility for big data studies will be evaluated by their identities and number of Vs (0–4) in terms of the V criteria.

All the above-mentioned synthesis and analysis processes will be undertaken by two reviewers and any disagreement will be discussed with the third reviewer for consensus. The results will be visualised with tables and graphic charts. The body of evidence will not be assessed
in this study. The examples of the result table will be performed in tables 4–10.

Patient and public involvement
No patient involved.

DISCUSSION
The present study provides the first comprehensive review of big data articles with respect to both environmental health and health services. It used the 4Vs—volume, variety, velocity, and veracity—as criteria to identify whether these studies actually worked with big data and big data analytics techniques.

This study seeks to improve the quality of future big data studies in the field of environmental health and health services. The results from the present review will enable researchers to understand how big data studies should be conducted and improve the study quality. This review is the first study to determine which big data technologies have been properly applied to environmental health research.

In reviewing the field of health services, we do not impose restrictions on diseases, conditions, or healthcare domains. This review is the first study to determine how V-eligible big data studies can make a difference to the healthcare service domains.
This study has several limitations. Firstly, research papers could be too obscure in reporting their eligibility. Email queries with online questionnaires will be sent to the authors for confirmation of information. Secondly, the present study mainly involves descriptive statistics of the extracted information to outline how big data are used in environmental health and health services research. Advanced statistics and further hypothesis-based studies will be designed after the present study. Lastly, this study does not reflect a representational picture of how big data are used in the broad field of medicine, but only that in environmental health and health services research.

ETHICS AND DISSEMINATION

Ethics approval

All data reviewed by the present study have been published and are publicly available; thus, ethics approval is not required for this study.

Publication plan

This protocol has been registered with the PROSPERO. The conduct and reporting of the review will follow PRISMA-ScR, PRISMA 2020, and SWiM. The results of this study will be disseminated through a peer-reviewed journal.

Author affiliations

1State Key Laboratory of Quality Research in Chinese Medicine, University of Macau, Macau, China
2BGI College & Henan Institute of Medical and Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, China
3Department of Obstetrics and Gynecology, The Second Affiliated Hospital of Zhengzhou University, Zhengzhou, China
4Edinburgh Bayes Centre for AI Research in Shenzhen, College of Science and Engineering, University of Edinburgh, Scotland, UK
5Center for Machine Learning and Intelligent Applications, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, People’s Republic of China

Contributors

S-wL conceived the study. All authors designed the protocol. The initial protocol of the BDRC study was drafted by PPT and ILT, and revised by YJ and S-wL. All authors read and approved the final manuscript.

Funding

Part of the work of YJ was financially supported by the Henan Institute of Medical and Pharmacological Sciences (Grant No. 2021B0113), the Zhengzhou University (Grant No. 32212456). Part of the work of S-wL was financially supported by the University of Edinburgh and the Shenzhen Institute of Artificial Intelligence and Robotics for Society (Grant No. 2021-ICP001).

Competing interests

None declared.

Patient consent for publication

Not applicable.

Provenance and peer review

Not commissioned; externally peer reviewed.

Supplemental material

This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been vetted by any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access

This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs

Pui Pui Tang http://orcid.org/0000-0003-1890-7025
Lam Tam http://orcid.org/0000-0001-8574-8963
Yongliang Jia http://orcid.org/0000-0002-2397-8539
Siu-wai Leung http://orcid.org/0000-0003-3692-9578

REFERENCES

18 Ghosevaspay TA. How could health care be anything other than high quality? Lancet Glob Health 2018;6:e1140–1.
20 Diebold FX. On the origin(s) and development of the term “big data”. PIER Work Pap 2012:12–37.


Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. *Int J Inf Manage* 2015;35:137–44.


Almeida F. Big data: concept, potentialities and vulnerabilities. *ESJ* 2018;2.


