International Journal for Quality in Health Care, 2021, **33(3)**, 1–2
DOI: https://doi.org/10.1093/intqhc/mzab134
Advance Access Publication Date: 12 September 2021
Perspective

The intersection of big data and epidemiology for epidemiologic research: The impact of the COVID-19 pandemic

CHUNLEI TANG[1,2,3], JOSEPH M. PLASEK[1,2], SUHUA ZHANG[4], YUN XIONG[5], YANGYONG ZHU[5], JING MA[2], LI ZHOU[1,2], and DAVID W. BATES[1,2,3]

[1]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA
[2]Harvard Medical School, Boston, MA, USA
[3]Clinical and Quality Analysis, Partners HealthCare System to Mass General Brigham, Boston, MA, USA
[4]Department of Kidney Disease, Suzhou Kowloon Hospital, Jiangsu, China
[5]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China

Address reprint requests to: Suhua Zhang, Department of Kidney Disease, Suzhou Kowloon Hospital, 118 Wansheng Street, Suzhou Industrial Park, Suzhou, Jiangsu 215021, China. Fax: +86 13912792640; E-mail: suhua.zhang@suhsmu.edu.cn

Abstract

Big data epidemiology facilitates pandemic response by providing data-driven insights by utilizing big data tools that differ from traditional methods. Aspects regarding 'garbage in, garbage out', such as insufficient data, inaccessibility of data, missing data, uncertainty in handling data and bias in analysis or common findings are addressable by combining techniques across disciplines.

Key words: public health, health-care system, patient safety, big data, epidemiologic studies, COVID-19

The coronavirus disease 2019 (COVID-19) pandemic has pushed infectious disease epidemiology and disease modeling to the forefront of the general public's consciousness due to its profound impact, necessitating cultural changes steeped in focus on location and safety, such as universal masking, vaccination and social distancing. The COVID-19 pandemic has resulted in a rapidly increasing volume of available, digitized, global data fueling an increased pervasiveness of data-driven decision-making. These perspectives include (i) an epidemiologic perspective that reflects the basic yet essential descriptive information of the 4Ws: 'What (virus), Who (person), Where (transmission location), and When (time)', (ii) an informatics perspective that often focuses on the logistics of data capture, and (iii) a methodological perspective focused on study designs and analytical techniques. Historically, asking the right question(s) in the early stages of descriptive epidemiological studies was necessary when following a hypothetical purpose-driven paradigm. With the spread of COVID-19 around the globe, adopting a data-driven paradigm may help internalize and coalesce knowledge at the speed at which new COVID-19 data are becoming globally available and at which relevant research studies are being published. For example, population-level thinking starts for a change. Population-level thinking combined with spatiotemporal data analysis methods has great potential to transform the fields of big data and epidemiology, respectively. By combining population-level thinking with spatial and temporal scales, researchers rethink a spatiotemporal range following data as it crosses borders [1] and reclassify population to find appropriate classes for certain known (or unknown) rules among many real-world objects. However, the sudden rise of big data in the public health sphere might lead to an increase in misconceptions and misunderstandings due to its novel use. Among them, 'garbage in, garbage out' representing big data epidemiological studies is particularly prominent.

In the Dictionary of Epidemiology [2], epidemiological methods' role is in reducing negative '… health-related events, states, and processes' in a population. Data used for observational epidemiological research and public health purposes generally fit one of six descriptions: (i) specifically designed surveys and questionnaires, (ii) administratively collected clinical records (clinical trials and clinical practice), (iii) measures of participants' genomic or metabolomic biology, (iv) participant measurements captured automatically by a medical device, (v) measures of participant geospatial context as environmental representations, and (vi) measures compiled from the Internet such as from social media postings. These six data categories are theoretically sufficient for epidemiological studies, such as surveillance of infectious disease outbreaks, contact tracing, and the development of diagnostic tests and therapeutics.

Big data are first and foremost technical terminologies built around the notion of the 5Vs: volume, velocity, variety, veracity and value. It is superficial to argue merely about the size of the data as being big; but what makes COVID-19 unique among emerging infectious diseases is the scope and velocity of newly available data, for which many traditional epidemiological methods lack timeliness and scalability to

Received 6 March 2021; Editorial Decision 2 September 2021; Revised 30 August 2021; Accepted 9 September 2021

handle. Our prior research suggests that the field of epidemiology has been approximately 5–8 years behind in adopting emerging techniques proposed in the data science literature [3]. The adoption of these data techniques for public health research may be appropriate. A more complete picture of big data is that it is composed of three critical elements: 'data, technology, and application [4]', which all work toward augmenting the decision-making of public health professionals by following data as it crosses borders [5].

The fields of big data and epidemiology share a common thread of rigorous analysis of quantitative data but bring different perspectives and tools to solve complex public health problems [6]. Implementing big data effectively is an art. A big data epidemiological study may be poorly conceived if it lacks necessary creativity; for example, pursuing too precise of a model but ignoring necessary abstraction. Advanced knowledge in study design, measurement and casual inference are part of the identity of an epidemiologist; biostatisticians have unique tools and theories appropriate for smaller datasets; data scientists have unique tools for analyzing high-dimensional data; and informaticians have the skills needed to communicate across these fields and understand what techniques from each discipline are necessary for a particular study [7]. Data source selection checked by collaboration and cross-training is one way to avoid 'garbage in'; for example, two retracted COVID-19 studies from the *Lancet* [8] and *N. Engl. J. Med.* [9], as the data could not be audited. Other aspects regarding 'garbage in, garbage out', such as insufficient data, inaccessibility of data, missing data, uncertainty in handling data and bias in analysis or common findings are addressable by combining techniques across disciplines.

Big data represent valuable and increasingly necessary tools to facilitate effective local responses to pandemics, rather than the curse of 'garbage in, garbage out'. Only with full availability and use of such approaches will the future of big data epidemiological research be genuinely sustainable.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

Contributorship

All authors provided substantial contribution to the conception and design of this work, and helped draft and revise the manuscript. All the authors are accountable for the integrity of this work.

References

1. Leondes CT ed. *Expert Systems, Six-Volume Set: The Technology of Knowledge Management and Decision Making for the twenty-first Century*. 3rd edn. Boston, MA: McGraw-Hill Education, 2016, 736.
2. Porta MS ed. *A Dictionary of Epidemiology*. 6th edn. New York, NY: Oxford University Press, 2014,376.
3. Hidden for blinding purposes.
4. Hidden for blinding purposes.
5. Hidden for blinding purposes.
6. Goldstein ND, LeVasseur MT, McClure LA. On the convergence of epidemiology, biostatistics, and data science. *Harvard Data Sci Rev* 2020.
7. Gange SJ, Golub ET. From smallpox to big data: the next 100 years of epidemiologic methods. *Am J Epidemiol* 2015;**183**: 423–6.
8. Mehra MR, Ruschitzka F, Patel AN. Retraction - hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31324-6/fulltext (5 June 2021, date last accessed).
9. Mehra MR, Ruschitzka F, Henry TD *et al.* Retraction: cardiovascular disease, drug therapy, and mortality in covid-19. *N Engl J Med* 2020. https://www.nejm.org/doi/full/10.1056/NEJMc2021225 (5 June 2021, date last accessed).