Title: Opportunities and challenges of using big data for global health

Peng Jia[1,2,3], Hong Xue[2,3,4], Shiyong Liu[5], Hao Wang[6,7], Lijian Yang[8,9], Therese Hesketh[10,11], Lu Ma[3,12], Hongwei Cai[3,13], Xin Liu[3,12], Yaogang Wang[14], Youfa Wang[2,3,15*]

[1]GeoHealth Initiative, Department of Earth Observation Science, Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

[2]International Initiative on Spatial Lifecourse Epidemiology (ISLE)

[3]Global Health Institute, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi, China

[4]Department of Health Behavior and Policy, School of Medicine, Virginia Commonwealth University, Richmond, Virginia, USA

[5]Research Institute of Economics and Management, Southwestern University of Finance and Economics, Chengdu, China

[6]State University of New York, Albany, New York, USA

[7]School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangdong, China

[8]Center for Statistical Science, Tsinghua University, Beijing, China

[9]Department of Industrial Engineering, Tsinghua University, Beijing, China

[10]Institute for Global Health, University College London, London, UK

[11]School of Public Health, Zhejiang University, Hangzhou, China

[12]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi, China

[13]Department of Network Information, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

[14]Department of Health Service Management, School of Public Health, Tianjin Medical University, Tianjin, China

[15]Fisher Institute of Health and Well-Being, Department of Nutrition and Health Sciences, College of Health, Ball State University, Muncie, Indiana, USA

Current word: 1,450; References: 18

Corresponding author information

Youfa Wang, MD, PhD, MS
Global Health Institute, Xi'an Jiaotong University Health Science Center
Xi'an, Shaanxi, 710061, China
Phone: +86-29-82657395
Email: youfawang@gmail.com

Funding

This work was funded in part by research grants from the US-based China Medical Board (Grant No. 16-262), the United Nations International Children's Emergency Fund (UNICEF), and the National Natural Science Foundation of China (Grant No. 91746205, 71673199), and funding from Xi'an Jiaotong University and University of Twente. We also thank Lorentz Centre, the Netherlands Organisation for Scientific Research, the Royal Netherlands Academy of Arts and Sciences, the Chinese Centre for Disease Control and Prevention, and the West China School of Public Health of Sichuan University for funding and supporting research activities of the International Initiative on Spatial Lifecourse Epidemiology (ISLE).

Opportunities and challenges of using big data for global health

The past two decades have witnessed the burgeoning of enormous digital technologies and data collected via countless channels. They are combined in numerous ways in different fields, including epidemiology, mHealth and modeling of health systems, with the intention to improve human health (e.g., clinical decision support, electronic medical record management)[1-6]. However, this is a new interdisciplinary area where no single scientific discipline knows how to take full advantage of these data and technologies to solve health problems[1].

A workshop was organized by the Global Health Institute of Xi'an Jiaotong University to discuss issues related to using big data in global health efforts. This was convened on 26 November 2018 during the 2nd Belt and Road Initiative Global Health International Congress in Xi'an, China[7]. During the workshop, the participants explored the utilization of big data in infectious and chronic disease research and health practice and policy in Belt and Road countries.

The workshop and this summary report drew on perspectives from a wide range of stakeholders including leading scientists in the fields of big data analytics, digital technologies, spatial science, biostatistics, artificial intelligence, public health, epidemiology, clinical nutrition, health policy, and systems science. The participants sought to identify critical issues and research priorities in big data in China and other Belt and Road countries, with the aim of agreeing on a mutual agenda for understanding and utilizing the potential of big data and digital health better to improve health outcomes.

Box 1 lists the top 10 priorities (among all questions discussed during the workshop) for advancing the applications of big data in future public health research and practice. They emerged out of the discussion and represented the consensus of perspectives from the experts attending the workshop.

Box 1. Top 10 priorities for advancing the applications of big data in future public health research and practice.

1. Why do we need to collect big data?
2. How can we integrate big data and digital technologies?
3. How can we reach agreement on data sharing protocols with big data holders?
4. How can big data be collected and linked to other data sources without violating individual confidentiality?
5. How can we ensure the quality of big data?
6. How can we facilitate the development of big data through industry-academic co-operation?
7. How can we make cost-effective use of big data?
8. How can we secure funding for big data research?
9. What changes need to be made in the field of statistics to meet the demand for big data analysis?
10. How can we utilize big data to improve health outcomes across Belt and Road countries?

1. *Why do we need to collect big data?* Many existing data sources in various areas have provided large volumes of information that are potentially useful in global health research and practice, such as Health Information Exchange[8] and remote sensing satellite archives[9,10]. Thinking ahead about rationales for collecting a certain type of big data would aid in proposing innovative, high-quality scientific hypotheses on the basis of those data. Therefore, before contemplating any major collection of big data, one should consider the potential use and interpretation of those data in relation to real-world health problems. This would involve improved understanding among researchers, those collecting the data, and other collaborating parties.

2. *How can we integrate big data and digital technologies?* Digital technologies, such as crowdsourcing[11,12] and the Internet of Things[13,14], can not only facilitate big data collection, but also reduce the cost of data collection. The functioning of digital technologies can also benefit

from big data to improve human health. For example, big data can better guide more precise applications of digital technologies in relevant areas, e.g., monitoring individuals' health status, and diagnosing, treating, predicting, and even preventing diseases, in both clinical and community contexts. In this era of team science and transdisciplinary collaboration, there is more need than ever before for the integration of big data and digital technologies to realize strategic global health goals[15].

3. *How can we reach agreement on data sharing protocols with big data holders?* The key stakeholders, including citizen representatives and research ethics boards on both sides of data collection and request, need to discuss potential problems caused by data sharing in global health research. This will also provide a good environment for research ethics staff to update their knowledge in the era of digital health. Data security is another concern in the field, which demands special attention during the process of data sharing not only between data holders and researchers, but also between different data owners. Data protection laws in some countries have already been tightened because of big data[16].

4. *How can big data be collected and linked to other data sources without violating individual confidentiality?* Without linking with identified individuals, big data can only be aggregated to map some areal patterns. Being increasingly collected everywhere, they can make a greater contribution to improving human health, if appropriately linked to all places and moments over the life course of individuals, also referred to as spatial life course epidemiology[8]. Therefore, big data-based research should be conducted with ethics prioritized and strictly regulated, which requires in-depth discussion among multiple stakeholders and is already the case in some countries.

5. *How can we ensure the quality of big data?* In the era of big data, there should be protocols for the collection and quality control of big data. By doing so, the whole process could be standardized and transparent to data users, and data quality issues cannot be propagated to

health research. This aspect should be incorporated into the training of human resources from the beginning. Also, there are some challenges encountered when deploying digital technologies at the individual level. For example, the installation, operation, and maintenance of sensors or terminals in low-education populations and in disadvantaged regions might encounter hard-to-imagine predicaments (e.g., people with low education cannot use a simple terminal even after many instructions; data may be missing due to lack of telecom infrastructure in disadvantaged regions).

6. *How can we facilitate the development of big data through industry-academic co-operation?* The industry that is mainly responsible for collecting data can benefit from academia, with regard to use of big data for scientific inquiry based on big data. The co-operation between the developers of big data and academic users should be promoted, enabling synergistic contributions to research and education activities, and thus providing real data for researchers to better investigate and understand what is happening in the real world. For example, many advanced sensor technologies have appeared in the industry first, although greater demand for tracking routine behaviors and health status always exists in health research areas. However, it is not cost-effective to utilize these sensor technologies in rarely funded research programs, to collect behavioral and exposure data only among a limited number of people. Such practices cannot solve health problems at a scale that makes more sense.

7. *How can we make cost-effective use of big data?* We should take advantage of national- and local-level opportunities to build platforms for big data (i.e., allowing for more political control), for example, collaborating with province- and municipal-level data collection mechanisms, such as the State-owned Assets Supervision and Administration Commission of the State Council (SASAC) in China. Often such organizations have the resources and operation to collect related data. In addition to reducing the cost of data collection and maintenance, this will also aid in developing data quality control and assessment, as well as data sharing systems.

8. *How can we secure funding for big data research?* Adequate and sustainable funding is critical for big data research. We should convey the health-related significance of big data research to relevant governmental agencies and private partners, so that big data research can be reasonably funded by solving problems they are facing. For example, Quarantine and Inspection Bureaus, Centers for Disease Control and Prevention, and hospitals can all benefit from the linkage of big data for control of spread of infectious diseases.

9. *What changes need to be made in the field of statistics to meet the demand for big data analysis?* Both theories and methods of handling big data need to be incorporated into the field of statistics, so traditionally trained statisticians could be better geared toward new data types and structures. Some big data analytics may not stem from classical statistical theory, with less than optimal performance. Systems modelling and machine learning are two promising options that may play an important role in this field[1,17].

10. *How can we utilize big data to improve health outcomes across Belt and Road countries?* Belt and Road countries have diverse cultures, religions, lifestyles (e.g., dietary behaviors), economic development, natural environments, physical environments (e.g., built and food environments), and disease burdens. This provides some excellent and unique opportunities for research besides many related challenges. Efforts to improve health outcomes in these countries need to be based on data collected from citizens in different countries, instead of simple, homogeneous hypotheses. Global, multidisciplinary, academic platforms, such as the Belt and Road Initiative Global Health International Congress, have been established to promote dialogues and collaborations among Belt and Road countries[7]. The community-based participatory research approach[18], for example, could be adapted to an international setting to promote active involvement of country representatives while shaping research and intervention strategies.

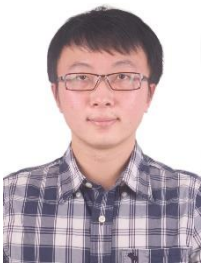
These ten research questions/priorities represent the consensus of perspectives from a cohort of leading scientists in multiple fields related to big data and digital health, and are intended to serve as a start of a discussion on related research priorities in big data and digital technologies for improving global health.

References

1. Wang Y, Xue H, Liu S. Applications of systems science in biomedical research regarding obesity and noncommunicable chronic diseases: opportunities, promise, and challenges. *Advances in nutrition (Bethesda, Md)* 2015; **6**(1): 88-95.
2. Wang Y, Xue H, Esposito L, Joyner MJ, Bar-Yam Y, Huang TT-K. Applications of complex systems science in obesity and noncommunicable chronic disease research. Oxford University Press; 2014.
3. Wang Y, Xue H, Huang Y, Huang L, Zhang D. A systematic review of application and effectiveness of mHealth interventions for obesity and diabetes treatment and self-management. *Advances in Nutrition* 2017; **8**(3): 449-62.
4. Xue H, Slivka L, Igusa T, Huang T, Wang Y. Applications of systems modelling in obesity research. *Obesity reviews* 2018; **19**(9): 1293-308.
5. Wang H. Big Data the current status, opportunities, and challenges. *Chinese Journal of Clinical Laboratory Management (Electronic Editon)* 2017; **5**(01): 30-5.
6. Jia P, Xue H, Yin L, Stein A, Wang M, Wang Y. Spatial Technologies in Obesity Research: Current Applications and Future Promise. *Trends Endocrinol Metab* 2019.
7. Jia P, Wang Y. Global health efforts and opportunities related to the Belt and Road Initiative. *Lancet Glob Health* 2019; **7**(6): e703-e5.
8. US Government. Health Information Exchange. https://www.healthit.gov/topic/health-it-basics/health-information-exchange2019).
9. Jia P, Stein A. Using remote sensing technology to measure environmental determinants of non-communicable diseases. *International journal of epidemiology* 2017; **46**(4): 1343-4.
10. Jia P, Stein A, James P, et al. Earth Observation: Investigating Noncommunicable Diseases from Space. *Annu Rev Public Health* 2019.
11. Goodchild MF. Citizens as sensors: the world of volunteered geography. *GeoJournal* 2007; **69**(4): 211-21.
12. Jia P. Integrating Kindergartener-Specific Questionnaires With Citizen Science to Improve Child Health. *Front Public Health* 2018; **6**: 236.

13. Hossain MS, Muhammad G. Cloud-assisted industrial internet of things (iiot)–enabled framework for health monitoring. *Computer Networks* 2016; **101**: 192-202.
14. Suciu G, Suciu V, Martian A, et al. Big Data, Internet of Things and Cloud Convergence--An Architecture for Secure E-Health Applications. *J Med Syst* 2015; **39**(11): 141.
15. Jia P. Spatial lifecourse epidemiology. *Lancet Planet Health* 2019; **3**(2): e57-e9.
16. Jia P, Lakerveld J, Wu J, et al. Top 10 Research Priorities in Spatial Lifecourse Epidemiology. *Environ Health Perspect* 2019; **127**(7): 74501.
17. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *The New England journal of medicine* 2019; **380**(14): 1347-58.
18. O'Fallon LR, Dearry A. Community-based participatory research as a tool to advance environmental health sciences. *Environ Health Perspect* 2002; **110 Suppl 2**: 155-9.

First author



Peng Jia is a faculty member at the University of Twente. He coined the term “Spatial Lifecourse Epidemiology” and founded the International Initiative on Spatial Lifecourse Epidemiology (ISLE) to facilitate this area. He received B.Eng in environmental engineering, two M.S. in spatial science and spatial epidemiology, and Ph.D. in health geography. He uses statistical, spatial, location-based, and artificial intelligence technologies to conduct spatial lifecourse epidemiologic research. He is also expert in planning health-care resource allocation, and optimizing hierarchical health-care systems.

Corresponding author

Youfa Wang is John and Janice Fisher Endowed Chair of Wellness, Associate Director of Fisher Institute at Ball State University; founding dean of Xi’an Jiaotong University Global Health Institute; previously, Professor and Department Chair in State University of New York at Buffalo, Center Director and Associate Professor at Johns Hopkins University. As an international known expert, he has served many leadership roles including as WHO and UN consultant, Chair of American Society for Nutrition Nutritional Epidemiology Section; fostered transdisciplinary/international collaborations.