# Big data and medical research in China

**Luxia Zhang and colleagues** discuss the development of big data in Chinese healthcare and the opportunities for its use in medical research

The quantity of data that is routinely generated and collected have increased greatly in the past decade, as has our ability to analyse and interpret these data, particularly in medicine. China's large population and universal healthcare system provide rich sources of data, and interest in the application of big data to medicine has grown in the past few years. It is hoped that the combined use of large data resources and new technologies will solve many existing medical problems and provide better evidence for decision making.[1]

### What do we mean by big data?

Big data has been defined as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."[2]

Digital healthcare data are now common. Large numbers of medical data are generated through medical records, regulatory requirements, and medical research.[3] Worldwide, the number of data are projected to double every two years, which will result in 50 times more data in 2020 than in 2011.[4]

In addition to data volume,[5] variety and velocity are also important for usability—comprising the 3Vs of big data. The variety comes from the multiple sources of data (box 1), both structured and unstructured, which reflect the whole health and disease process.

Medical data are also being combined with information from social media, occupational information, geographical location, and economic and environmental data.[6] Integrating all these information sources into datasets that can be analysed is key to utilising big data. In addition, the speed at which big data are generated and processed should meet the real time demands of preventing and managing disease.

Recently, veracity has been added as a goal of big data,[7] although some argue that big data are difficult to validate and can never be completely accurate.[5 8] Nonetheless, to make the best use of big data, quality is important.

An important concept of big data is that assembly of the data is not the purpose. Instead, data must be analysed, interpreted, and acted on. Therefore, to get the best value from big data, new technologies and analytical methods (eg, machine learning) are needed and the information generated must be evaluated for clinical effectiveness and translated into tools for use in clinical practice.[9]

**KEY MESSAGES**

- The application of big data to health and medicine is a national priority for China
- Several initiatives to promote big data have been started by the government and researchers
- The use of big data and new data technologies has the potential to improve medical research and the understanding of health, and disease

**Box 1: Sources of medical big data**

- Administrative and claims data
- Routine population statistics and major disease surveillance data
- Real world data, such as electronic medical records, medical imaging, and data from health examinations
- Research data, including biomarkers, and multiomic information from clinical trials or cohort studies
- Registries (eg, of devices, procedures, and diseases)
- Data from mobile medical devices
- Data reported by patients

### What data are gathered in China and how?

Promoting the use of big data in medicine is a national priority in China. In June 2016, the State Council of China issued an official notice on the development, and use of big data in the healthcare sector.[10] The council acknowledged that big data in health and medicine were a strategic national resource and their development could improve healthcare in China, and it set out programmatic development goals, key tasks, and an organisational framework.

After regional health data centres were established in Shanghai and Ningbo, the National Health and Family Planning Commission announced in 2016 that China would establish more regional and national centres and industrial parks that focused on big data in health and medicine as part of a national pilot programme to make more meaningful use of these data.[11] Four cities in Fujian and Jiangsu provinces in eastern China were chosen as the pilot sites, and the centres are now in construction. The goal is to integrate the following datasets:

- Regional health data, including claims data from nationally funded basic health insurance that covers over 95% of the Chinese population[12]
- Administrative data from local health offices
- Data from public health services of the Chinese Center for Disease Control and Prevention, especially for women and children, and for surveillance networks of the main non-communicable diseases
- Birth and death registries
- Electronic medical records from hospitals, including primary, secondary, and tertiary hospitals.

China is already making use of big data. The country's personal identification system could be used to link data from various sources. Medical claims data from the national social insurance system have been used to generate a 5% sampling database and an overall database covering over 0.6 billion beneficiaries in the past five years, which are available to scientific researchers. Applications to use these data are managed by organisations such as the Chinese Health Insurance Research Association; there is no public access.

Since 2016, many academic research projects using these national datasets have been approved to evaluate the current and future clinical and economic burden of chronic diseases such as cardiovascular disease, diabetes, kidney disease, and chronic obstructive pulmonary disease. Furthermore, other national administrative databases, including the national standardised discharge summary of inpatients and the national death registry, with hundreds of millions of patient records, have been used by medical and public health researchers.[13 14]

China is also focusing on personalising medicine. Since 2016, the Ministry of Science and Technology has initiated and funded many "precision medicine" projects under the national key research and development programme. A centralised and integrated data platform for precision medicine is being developed, which will store all patient/population data as well as biosamples collected from a series of large cohort studies and from biobanks. The platform is expected to include at least 0.7 million participants, 0.4 million from the general population and 0.3 million from patients with major non-communicable diseases. China's large population base and centralised governance mean that very large sample sizes can be reached, which is of great value to personalised medicine initiatives.

As well as the government-led projects, Chinese academic medical societies are leading data-sharing initiatives (box 2). In October 2017, the School of Public Health at Peking University announced the launch of the China Cohort Consortium (chinacohort.bjmu.edu.cn/home). Currently 20 cohorts with more than 2 million participants are included. The activities of the consortium include using common data models for data harmonisation, performing individual participant data meta-analyses, and generating new cohorts. Furthermore, disease based data sharing platforms, including for cardiovascular disease, stroke, cancer, and kidney disease, have been established by medical specialists with the support of the government. For example, the China Kidney Disease Network (kidney.net.cn), which launched in 2015, integrates various sources of data on kidney disease and uses new analytic techniques to provide evidence for healthcare policy, strengthen academic research, and promote effective disease management.[15]

**Box 2: Current projects applying big data to medicine in China**

**Government led**

- Development of regional health data centres with pilot programmes in four cities in Fujian and Jiangsu provinces (http://en.nhfpc.gov.cn/2016-10/24/c_70420.htm)
- Opening of existing national administrative, claims, death registry, and other databases for academic use
- Promotion of precision medicine by supporting cohort studies and integrated data platforms (http://www.most.gov.cn/tztg/201603/t20160308_124542.htm)

**Researcher initiated**

- China Cohort Consortium (http://chinacohort.bjmu.edu.cn/home)
- China Kidney Disease Network (kidney.net.cn)
- Others funded by the government include cardiovascular disease (eg, China Cardiovascular Surgery Registry), stroke (eg, Chinese National Stroke Registry), and cancer (eg, National Central Cancer Registry of China)

## What are the challenges and what needs to be done?

### Electronic record systems

Electronic medical records, whether collected by one organisation or for individual patients across organisations, are not commonly used for research in China. They are primarily used for clinical practice and largely contain unstructured data. Although over 90% of hospitals in China use electronic records, accessibility to and quality of the data are not optimal.

Adoption of individual electronic health records has been impeded by incompatibility between different hospital systems. China has over 300 commercial providers of hospital information systems with various technical structures and data standards. Furthermore, healthcare systems are not required to exchange data with each other. Some regions are planning to establish regional electronic health records but most are in preliminary stages. To overcome these problems, the interoperability of electronic records needs to be improved, especially for data structures, data standards, and data transfer agreements. Health authorities, hospitals, and electronic record companies must agree on how to improve hospital information systems. Technologies that can integrate data from different sources are also needed. In addition, the government should introduce policies to strengthen data exchange and integration across organisations.

### Lack of medical terminology system

The lack of a widely adopted and consistently implemented medical terminology system is another problem for using big data in medical research. For example, since 2002, the use of the International Classification of Diseases (ICD-9, and more recently ICD-10) was mandated by the National Health and Family Planning Commission for all hospital patients. However, the growth of hospital information systems has resulted in many variations in the coding of other clinical terms beyond diagnosis, making data exchange difficult. Widely accepted terminology systems, such as the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT), the Unified Medical Language System (UMLS), or the General Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine (GALEN), are not available in China. By integrating and distributing key terminology, classification, and coding standards in medicine, these systems promote more effective and interoperable biomedical information systems and services, including electronic health records. More effort is needed to resolve linguistic differences between Chinese and English beyond the existing translation of terms.

### Current medical practice patterns

Medical practice patterns and the infrastructure of health systems in China also impede the meaningful use of big data. The lack of an established referral system and the heterogeneity in the quality of healthcare contribute to "medical migration," when patients travel to different provinces and cities to seek medical care. In the current Chinese medical system, it is almost impossible to track a patient through electronic record systems for clinical purposes as there is no unified national platform that can consolidate all the data from all healthcare institutions in China. The main barrier to conducting a "deep patient" study,[16] where machine learning is used to predict future adverse events using medical data, is obtaining the longitudinal data and outcomes of each patient from electronic records. Furthermore, the wide differences in medical practice raise concerns about the veracity of data.

### Data quality

The problems described above affect the quality of big data. It has been shown that, when the quality of clinical data is higher, big data analytics produce more valid, stable, and clinically useful results.[17] However, it is difficult to validate high volume datasets. One way of dealing with the data quality problem is to examine the characteristics of the database and judge which variables are likely to be relatively accurate—for example, expenditure from claims data—and to answer questions based on those variables. Improving the veracity of data requires an ongoing and joint effort by multiple sectors to rigorously examine the validity, representativeness, and completeness of data.

### Privacy concerns

Although privacy is an extremely important topic for big data in health and medicine, there is no specific law or guidance on this in China. Regulation from authorities and research standards about privacy protection are needed that do not jeopardise the completeness of data that can be used.

## Opportunities to improve health

The use of big data in medicine includes public health promotion (disease monitoring and population management), healthcare management (quality control and performance measurement), drug and medical device surveillance, routine clinical practice (risk prediction, diagnosis accuracy, and decision support), and research.[19]

The existing mandatory national administrative databases in China produce big data that can easily be used to monitor trends in major diseases and provide evidence for policy making in healthcare. New data analytics, such as machine learning, to replace much of the work of radiologists and anatomical pathologists, can also be used and is an active area of research in China.[18] However, for applications that need detailed and high quality clinical information and long term follow-up, such as predicting long term outcomes and providing support for clinical decisions, the data systems in China need to be developed further.

In China, discussion on big data in medicine has focused on how to collect, store, integrate, and manage data and has been led by computer scientists, and the health information industry. However, the future of big data in medicine is in using new analytic techniques such as machine learning to answer clinical questions, educating doctors and policy makers to understand big data, and promoting the use of tools generated by big data and big data technologies that support clinical decision making.

## Conclusion

China's national campaign to promote the application of big data in health and medicine is likely to change medical research, medical practice, and the development of the healthcare industry in the near future. Despite the great interest in big data, we advocate following Confucian doctrine to ensure that we obtain true value for medicine—that is, to learn extensively, inquire carefully, think deeply, discriminate clearly, and practise faithfully.

We thank Alan Leichtman (Arbor Research Collaborative for Health, and University of Michigan) and Roseanne Yeung (University of Alberta) for their constructive suggestions and editing. We also thank Fan Liu (former chief information officer of Peking University People's Hospital and Peking University International Hospital) for comments on electronic record systems.

**Contributors and sources:** LZ is a renal epidemiologist and the executive deputy director of Peking University, Center for Data Science in Health and Medicine. HW is the founding director and key architect of several national medical databases in China. QL is a principal investigator focusing on artificial intelligence in health and medicine. MHZ is a nephrologist with substantial experience in experimental research and population based studies in China. QMZ is the Academician of the Chinese Academy of Engineering and the Chief Scientist of the 973 National Fundamental Program in China. His main interest is the translational study of cancer. LZ and HW contributed equally to this work and are the guarantors. This article arose from discussions about the status and future directions of big data in health and medicine in China, and the relationship with traditional medical studies.

**Competing interests:** We have read and understood BMJ policy on declaration of interests and declare that the article was funded by the World Health Organization (WHO Reference 2014/435380-0), the National Key Technology R&D Program of the Ministry of Science and Technology (2016YFC1305400), and the University of Michigan Health System-Peking University Health Science Center Joint Institute for Translational and Clinical Research (BMU20140479).

**Provenance and peer review:** Commissioned; externally peer reviewed.

**Luxia Zhang**, professor[1,2]

**Haibo Wang**, researcher[3,4]

**Quanzheng Li**, associate professor[5]

**Ming-Hui Zhao**, professor[1,6]

**Qi-Min Zhan**, professor[7]

[1]Renal Division, Department of Medicine, Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China

[2]Peking University, Center for Data Science in Health and Medicine, Beijing, China

[3]Clinical Trial Unit, First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China

[4]China Standard Medical Information Research Center, Shenzhen, China

[5]MGH & BWH Center for Clinical Data Science, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

[6]Peking-Tsinghua Center for Life Sciences, Beijing, China

[7]Peking University, Health Science Center, Beijing, China

**Correspondence to:** L Zhang
zhanglx@bjmu.edu.cn

1 Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375:1216-9. doi:10.1056/NEJMp1606181

2 Gartner. Big data. https://www.gartner.com/it-glossary/big-data/

3 Auffray C, Balling R, Barroso I, et al. Making sense of big data in health research: towards an EU action plan. *Genome Med* 2016;8:71. doi:10.1186/s13073-016-0323-y

4 Austin C, Kusumoto F. The application of big data in medicine: current implications and future directions. *J Interv Card Electrophysiol* 2016;47:51-9. doi:10.1007/s10840-016-0104-y

5 Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. *Biomed Res Int* 2015;2015:639021. PubMed doi:10.1155/2015/639021http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=26137488&dopt=Abstract

6 Fernández-Luque L, Bau T. Health and social media: perfect storm of information. *Healthc Inform Res* 2015;21:67-73. doi:10.4258/hir.2015.21.2.67

7 Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. *JMIR Med Inform* 2016;4:e38. doi:10.2196/medinform.5359

8 Ward JC. Oncology reimbursement in the era of personalized medicine and big data. *J Oncol Pract* 2014;10:83-6. doi:10.1200/JOP.2014.001308

9 Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016;13:350-9. doi:10.1038/nrcardio.2016.42

10 The State Council. The People's Republic of China. China to boost big data application in health and medical sectors. 2016 http://english.gov.cn/policies/latest_releases/2016/06/24/content_281475379018156.htm

11 National Health and Family Planning Commission of the PRC. China to build health care big data centers, industrial parks. 2016 http://en.nhfpc.gov.cn/2016-10/24/c_70420.htm

12 Shan L, Wu Q, Liu C, et al. Perceived challenges to achieving universal health coverage: a cross-sectional survey of social health insurance managers/administrators in China. *BMJ Open* 2017;7:e014425. doi:10.1136/bmjopen-2016-014425

13 Zhang L, Long J, Jiang W, et al. Trends in chronic kidney disease in China. *N Engl J Med* 2016;375:905-6. doi:10.1056/NEJMc1602469

14 Zhou M, Wang H, Zhu J, et al. Cause-specific mortality for 240 causes in China during 1990-2013: a systematic subnational analysis for the Global Burden of Disease Study 2013. *Lancet* 2016;387:251-72. doi:10.1016/S0140-6736(15)00551-6

15 Zhang L, Wang H, Long J, et al, China Kidney Disease Network (CK-NET) 2014 annual data report. *Am J Kidney Dis* 2017;69(suppl 2):S1-S149. doi:10.1053/j.ajkd.2016.06.011

16 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016;6:26094. doi:10.1038/srep26094

17 Altman RB, Ashley EA. Using "big data" to dissect clinical heterogeneity. *Circulation* 2015;131:232-3. doi:10.1161/CIRCULATIONAHA.114.014106

18 The State Council. The People's Republic of China. China issues guideline on artificial intelligence development. 2017. http://english.gov.cn/policies/latest_releases/2017/07/20/content_281475742458322.htm

Cite this as: *BMJ* 2018;360:j5910
http://dx.doi.org/10.1136/bmj.j5910