Editorial

Release of the national healthcare big data in China: a historic leap in clinical research

Zhongheng Zhang

Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China
Correspondence to: Zhongheng Zhang. Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, No. 3, East Qingchun Road, Hangzhou 310016, China. Email: zh_zhang1984@hotmail.com.

Received: 13 January 2017; Accepted: 24 January 2017; Published: 09 February 2017.
doi: 10.21037/amj.2017.02.03
View this article at: http://dx.doi.org/10.21037/amj.2017.02.03

At the beginning of 2017, the National scientific data sharing platform for population and health (NSDSPPH) released 49.1 TB data on Chinese population and health. The data comprise 280 million observations involving the fields of biomedicine, basic medicine, clinical medicine, public health, traditional Chinese medicine, pharmacology, population and reproductive health (http://www.ncmi.cn/1). NSDSPPH is a major project of the national science and technology infrastructure. It was launched in 2003. Clinical investigators may be interested in clinical data contained in NSDSPPH. The clinical data center is jointly established by Peking Union medical college hospital and Chinese PLA general hospital. After more than 10 years of endeavor, the clinical data center has become more and more sophisticated. Clinical data were collected from hospitals nationwide with strict criteria for data entry. Data are presented in several forms such as raw data, summary data and statistical report. The clinical data are accessible to the public and researchers. Additionally, the center also provides services of statistical analysis and data mining. Numerous scientific papers have been published using the database. Some examples include the correlation between waist circumference and respiratory function in teenagers (1), epidemiological study on coexisting prehypertension and prediabetes in northern China (2), and the Association between γ-glutamyltransferase and prehypertension (3).

Medical big data is a big treasure that can provide valuable information for scientific researches (4,5). Investigators can test their ideas and hypothesis by using the big data. Although China has the largest population, as well as the patient population in the word, there is little voice in the research field of clinical medicine. In major medical journals, especially the clinical journals, most high-impact articles are reported based on data from western countries. As a result, Chinese patients receive medical treatment based on guidelines derived from western countries. It is largely unknown whether results or conclusions based on western populations are generalizable to Chinese patients. The cause of this contradictory situation is partly due to the lack of awareness of the importance of heath care data exploration. Furthermore, clinical studies based on China are seldom accessible to other researchers. In other words, collected data are seldom openly accessible to the public and other investigators. The situation is that while the government invests large amount of funds on the research projects and the funded research group collects a large body of healthcare-related data, the researchers may only exploit a minority of these data and selected results are reported. This is a waste of data and funds. If the data are openly accessible, other researchers can perform data mining, individual-patient level meta-analysis from a distinct perspective. In other words, more useful information can be extracted from the dataset for medical decision making. In an analogy to gold mine, the value of big data can never be fully exploited unless it is openly accessible to researchers. Data opening is also in concordance with the international open data campaign, many journals such *PloS One* and *British Medical Journal* have declared that original articles published in their journals must include a statement on data availability (6,7).

However, datasets released at this initial stage are not without limitations. For example, there is a dataset

established based on critically ill patients treated in the intensive care unit (http://124.207.187.26:8080/lab/index.jsp). The number of observation is very small that there are less than 200 patients and/or observations. The recorded variables are also limited with less than 100 pieces of items. Recorded variables include demographics, physical variables, laboratory findings, severity scores, antibiotics and outcomes. However, the recording is over simplified. For example, only the name of antibiotics is available and there is no information regarding the time of initiation and discontinuation, dosage and administration route. Laboratory variable is simply presented as a numeric value and there is no information on the time of measurement. Usually, investigators are interested in the serial changes of laboratory findings, which is indicative of therapeutic efficacy, disease progression and recovery. With lack of these multi-dimensional data, it is difficult to establish a learning algorithm for medical decision making. Missing values are also prevalent in this dataset, which may impose challenges in statistical analysis (8). It appears that the ICU dataset was from a single hospital, which significantly compromised the generalizability of the results. However, we have to appreciate that the opening of the dataset is a historic leap in clinical research in China. We cannot expect a perfect dataset at its initial phase.

Acknowledgements

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *AME Medical Journal*. The article did not undergo external peer review.

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037amj.2017.02.03). Dr. Zhang serves as an unpaid Associate Editor-in-Chief of AME Medical Journal from Jan 2017 to Jan 2019.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Feng K, Chen L, Han SM, et al. Ratio of waist circumference to chest circumference is inversely associated with lung function in Chinese children and adolescents. Respirology 2012;17:1114-8.
2. Wu J, Yan WH, Qiu L, et al. High prevalence of coexisting prehypertension and prediabetes among healthy adults in northern and northeastern China. BMC Public Health 2011;11:794.
3. Qin X, Tang G, Qiu L, et al. Association between γ-glutamyltransferase and prehypertension. Mol Med Rep 2012;5:1092-8.
4. Zhang Z. Big data and clinical research: focusing on the area of critical care medicine in mainland China. Quant Imaging Med Surg 2014;4:426-9.
5. Zhang Z. Big data and clinical research: perspective from a clinician. J Thorac Dis 2014;6:1659-64.
6. Henry D, Fitzpatrick T. Liberating the data from clinical trials. BMJ 2015;351:h4601.
7. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, et al. Public availability of published research data in high-impact journals. PLoS One 2011;6:e24357.
8. Zhang Z. Missing values in big data research: some basic skills. Ann Transl Med 2015;3:323.

doi: 10.21037/amj.2017.02.03
Cite this article as: Zhang Z. Release of the national healthcare big data in China: a historic leap in clinical research. AME Med J 2017;2:19.