Editorial

Big Data Research in Chronic Kidney Disease

Xiao-Xi Zeng[1], Jing Liu[2], Liang Ma[3], Ping Fu[1,3]

[1]West China Biomedical Big Data Center, Sichuan University, Chengdu, Sichuan 610041, China
[2]Division of Nephrology, West China School of Medicine, Sichuan University, Chengdu, Sichuan 610041, China
[3]Division of Nephrology, Kidney Research Institution, West China Hospital of Sichuan University, Chengdu, Sichuan 610041, China

Key words: Chronic Kidney Disease; Computational Biology; Databse; Machine Learning

With a worldwide estimated prevalence of 8–16%, chronic kidney disease (CKD) is a major noncommunicable disease: it substantially contributes to premature mortality and loss of disability-adjusted life years.[1,2] The variety in terms of causes, progression mechanisms, and histopathological manifestations creates challenges for early diagnosis and effective interventions with CKD.[3] In addition, CKD is a major drain on health resources: in 2015, CKD and end-stage renal disease (ESRD) spend Medicare (the United States) over $98 billion.[4] China also faces a great financial burden owing to the increasing prevalence of CKD.

The definitions and boundaries of big data in health are still debatable.[5] However, the US National Institute of Standards and Technology defines big data as consisting of extensive datasets (in terms of volume, variety, velocity, or variability) that require a scalable architecture for efficient storage, manipulation, and analysis.[6] In addition to conventional data resources (e.g., electronic medical records, observational cohorts, and medical claims), environmental, behavioral, image, wearable device, social media, and multiomics data have been used for data-driven research for CKD.

As well as actual physical data, big data refer to the techniques used for analyzing multidimensional data sets,[7] such as artificial intelligence (including machine learning for structured data and natural language processing (NLP) for unstructured data), to reveal clinically relevant information from massive amounts of data.[7,8] Progress in cross-disciplinary collaborations of medicine, mathematical modeling, machine learning, and bioinformatics has led to novel mechanisms; it has helped in targeting intervention strategies for CKD that can facilitate precise risk predictions, early diagnosis, clinical decision analysis, and cost-effective interventions.[9,10]

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.4103/0366-6999.245275

Growth of Big Data and Innovative Analytic Methods in Chronic Kidney Disease Research

One clear benchmark for big data is volume. In 2011, the data of US health-care system alone amounted to 150 exabytes (10^{18}). Before long, the data will reach zettabyte and yottabyte levels worldwide.[11] Big data for medical research can be obtained from administrative and claims data, population statistics and disease surveillance data, real-world data, research data, registries, mobile medical devices, and patient-reported information. In addition, data that are not conventionally considered direct health-care information may also be collected and incorporated into medical research and applications, such as search engine queries,[9] social media data,[12] and environmental data.[13]

Large-volume databases for CKD research in the United States include the following: the National Health and Nutrition Examination Survey; United States Renal Data System; Kaiser Permanente; and Veterans Affairs Healthcare System. Those databases are widely used and support investigations into the disease burden, risk factors, outcomes, and medical resource consumption with CKD. In China, a national cross-sectional study investigated the prevalence of CKD.[14] The study covered 47,204 participants from 13 provinces; it reported the prevalence as 10.8%, and it demonstrated that CKD is a major public health concern in China. Subsequently, according to data of China's Hospital Quality Monitoring

Address for correspondence: Dr. Ping Fu, Division of Nephrology, Kidney Research Institution, West China Hospital of Sichuan University, Chengdu, Sichuan 610041, China
E-Mail: fupinghx@163.com

Received: 23-05-2018 **Edited by:** Ning-Ning Wang
How to cite this article: Zeng XX, Liu J, Ma L, Fu P. Big Data Research in Chronic Kidney Disease. Chin Med J 2018;131:2647-50.

System, the pattern for CKD has changed: diabetes has become the leading cause of CKD.[15] In addition, supported by the China-WHO Biennial Collaborative Projects 2014–2015, the China Kidney Disease Network (CK-NET) was established under the leadership of Drs. Lu-Xia Zhang and Hai-Bo Wang, based on the efforts of Professor Hai-Yan Wang.[16] CK-NET covers over 19.5 million patients from China's class 3 hospitals. CK-NET summarizes patient-level data from standardized discharge summaries; it highlights information that has not previously been reported, such as that related to epidemiology, treatment, costs, and other aspects pertinent to CKD.[16] Besides, large-size biobanks also serve as basic information sources for CKD research. KADOORIE Biobank, which was launched in 2004, has recruited 500,000 people from 10 regions of China (five urban and five rural) to assess the effects of risk factors for common chronic diseases. Its resources range from questionnaires, physical measurements at baseline, and long-term follow-up survey data to laboratory assays (including genotyping, metabolomic, and blood biochemistry data).[17] The Chinese Cohort Study of CKD has enrolled and followed up on 3000 predialysis CKD patients in Mainland China; that cohort study has also been used to explore the underlying mechanisms of CKD and adverse outcomes.[18] All these big data researches characterized CKD epidemiology in China, which is essential for health policymaking and health resource allocation planning.

Another feature of integrating big data in CKD could be the variety in data types. One example is the wide use of environmental data. In several studies, long-term exposure to air pollutants was evaluated by means of land-use regression and spatiotemporal models that utilized satellite remote-sensing aerosol optical depth data.[13,19,20] The association between air pollution and incidence of CKD and declining glomerular filtration rate was investigated using a generalized additive logistic model, time-varying linear mixed-effects regression model, and Cox proportional hazard models. The results showed that air pollution could be a nonconventional risk factor in the incidence[13,20] and progression of CKD.[19,20]

With respect to the development of artificial intelligence techniques, clinical notes and images are also used in kidney research. Singh *et al*.[21] undertook a concept-wide association study of clinical notes to determine new predictions of ESRD. The concepts were extracted from existing clinical notes using NLP tools; they were evaluated as predictors using proportional subdistribution hazards regression. Novel predictors were identified, such as high-dose ascorbic acid and fast food. In another study about predicting the outcomes in kidney transplant patients,[22] Banff lesion scores from the pathology reports and vital signs were extracted from unstructured text fields using proprietary NLP solutions in IBM Watson Content Analytics. Structured data have also been obtained from electronic medical records, the United Network for Organ Sharing database, and hospitals' own transplant databases. Predictive models for graft loss and mortality have been developed from both structured and unstructured data formats. The results demonstrate that the big data approach significantly adds efficacy in predicting adverse outcomes. By means of digital pathology applied to kidney tissue slides, Pedraza A *et al*.[23] used convolutional neural network classification to identify glomerulus and nonglomerulus segments. On average, the accuracy with this approach attained 99.95%, which underlines the promising application of machine vision in kidney histopathology.

With regard to speed, practice and research have benefited from the real-time collection of patient-level data. The acute kidney injury (AKI) system is one example of such an application based on the clinical data collected in routine clinical practice: the use of real-time data can improve the early detection of AKI and permit timely therapeutic interventions.[24] For advanced CKD, some researchers have developed a smartphone-based self-management system as an adjunct to the normal care. The system collects patients' behavior elements in real time and generates personalized patient messages based on prebuilt algorithms. If predefined treatment thresholds are met or critical changes occur, alerts are sent to providers.[22] To identify CKD patients with uncontrolled blood pressure (BP), Greenberg *et al*.[25] proposed a measurement system that incorporates data from the billing system, structured fields in the electronic health records, and free-text physician notes using NLP. To take action toward improving BP control and for completion of additional data, a point-of-care paper worksheet is given to the physician when such patients are presented. Using NLP in some systems has been found to produce benefits with regard to medication errors and control of BP.[22,25]

Multiomics technology enriches the data sources and helps improve analytic techniques with respect to data variety in CKD research. High-resolution analytic omics platforms (such as genomics, proteomics, peptidomics, transcriptomics, and metabolomics) and machine learning methods have been of tremendous help in the following: elucidating the molecular map of diverse interactions, signaling and regulation, and identifying CKD-related biomarkers and targeting different molecules with high precision.[3,9] For example, genome-wide association studies (GWASs) based on big data have gradually appeared and been refined. Gene analysis and consequent single-nucleotide polymorphism (SNP) analysis, adjusted for clinical characteristics from the data of 1293 African Americans, have been used to examine the causal association between racial disparities and CKD.[26] A strong association between CKD and apolipoprotein L1 renal-risk variants became evident. With a Chinese Han population of over 10,000 participants, GWAS identified *TNFSF13* as a susceptible gene of IgA nephrology.[27] Subsequently, an advanced verification test of that association was conducted among 2000 participants using SNPs and the phenotype level of the *TNFSF13* gene.[28] Studies have also focused on the association of renal function with the gut microbiome, amino acid metabolomic profiling, and renal microRNA and RNA profiles.

Value of Big Data Analytics in Chronic Kidney Disease Research

The aforementioned studies demonstrate the value of big data in CKD research. Big data can provide essential information about disease burden, molecular mechanisms, novel risk factors, and therapeutic targets. In this way, big data can help toward providing more effective prevention, earlier diagnosis, and more precise interventions.

According to McKinsey's report, big data – if used creatively and effectively – may lead to annual reductions of over $300 billion in the US health-care sector; most of that would be in the form of decreased health-care expenditure.[29] Another field where the value of big data has been demonstrated with regard to health policymaking is modeling and health economic evaluation using real-world data. That has been found to be time-saving, and it has potential to optimize clinical pathways and improve hospital management and the medical insurance system. Decision modeling combined with real-world data and medical knowledge can be used to predict the future prevalence of CKD in a given population.[30] That method can also be applied in health economic analysis. The American Diabetes Association and American College of Cardiology/American Heart Association Task Force recommend testing urinalysis and creatinine in patients with diabetes[31] or hypertension.[32] These recommendations are supported by modeling analysis, which has shown these tests to be cost-effective in high-risk populations, including tests for diabetes and hypertension.[33] Data science is widely used in medical insurance. One example of the application in nephrology is the ESRD prospective payment system (PPS) project. Following the report "End-stage Renal Disease Payment System: Results of Research on Case-Mix Adjustment for an Expanded Bundle" (submitted by the University of Michigan Kidney Epidemiology and Cost Center), Centers for Medicare and Medicaid Services (CMS) finalized the case-mix and facility-level adjustments for the ESRD PPS in the CY2011. Further data were collected and analyzed to support later refinement of the CMS ESRD payment system.

Opportunities and Challenges for Big Data in Chronic Kidney Disease

We have now entered the era of big data. Policies and initiatives have been announced to advance biomedical big data research and application in both developed and developing countries.[34] Quite a few instances of this kind of development can be cited, such as the Federal Big Data Research and Development Strategic Plan in the United States and Guidelines for Promoting and Standardizing the Healthcare Big Data Application and Development in China. The situation of CKD in China is characterized by a heavy disease burden in a large developing country; it is one of the most suitable places where biomedical big data should be applied.

However, fully utilizing the value of big data to support CKD research presents challenges. First, efforts have been made to encourage data sharing and accessibility to some national health databases, such as the National Scientific Data Sharing Platform for Population and Health; however, platforms where individual-level information is updated in a timely manner and can be freely accessed by scholars need to be constructed or improved. Second, health information is individual sensitive information according to China's "Information Security Technology – Personal Information Security Specification". Thus, when collecting, transferring, analyzing, sharing, and reporting health-related data, it is necessary to carefully balance the benefit of gains and risk to security and privacy. In China, there are national-level regulations that provide detailed guidance about medical data disclosure. However, data sharing could be more secure, and medical institutions need to be more willing to collaborate with outside partners in performing productive multidisciplinary research. The third challenge lies in the quality of data and techniques of data analysis. For example, Cisek *et al*.[3] concluded that there is a lack of satisfactory algorithms for multidimensional data modeling in clinically relevant predictive models for accurate elucidation of kidney disease. Fragmentary, diverse, and uncategorized data in mass information storage can result in difficulties when processing and analyzing information islands with complex and heterogeneous structures.

Despite all the above challenges, big data for CKD is in an era of opportunity, and it needs mature technology and policy supports. To provide better care and better health through cross-disciplinary efforts, building a database for CKD research is a top priority in addition to collecting and analyzing health-care information from a multidimensional perspective.

Financial support and sponsorship

This study was supported by grants from Science and Technology Department of Sichuan Province (No. 2016HH0069) and Chengdu Science and Technology Bureau (No. 2015-RK00-00252-ZF).

References

1. Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B, *et al.* Chronic kidney disease: Global dimension and perspectives. Lancet 2013;382:260-72. doi: 10.1016/s0140-6736(13)60687-x.
2. Human Development Unit EAaPR. Towards a Healthy and Harmonious Life in China: Stemming the Rising Tide of Non-Communicable Disease: WorldBank.org. Available from: http://www.worldbank.org/content/dam/Worldbank/document/NCD_report_en.pdf2011. [Last accessed on 2018 Jun 27].
3. Cisek K, Krochmal M, Klein J, Mischak H. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. Nephrol Dial Transplant 2016;31:2003-11. doi: 10.1093/ndt/gfv364.
4. Saran R, Robinson B, Abbott KC, Agodoa LY, Bhave N, Bragg-Gresham J, *et al.* US renal data system 2017 annual data report: Epidemiology of kidney disease in the United States. Am J Kidney Dis 2018;71:A7. doi: 10.1053/j.ajkd.2018.01.002.
5. Hansen MM, Miron-Shatz T, Lau AY, Paton C. Big data in science and healthcare: A Review of recent literature and perspectives. Contribution of the IMIA social media working group. Yearb Med

Inform 2014;9:21-6. doi: 10.15265/IY-2014-0004.
6. NIST Big Data Interoperability Framework: Volume 1, Definitions. NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, September 2015. Available from: https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf. [Last accessed on 2018 Jun 27].
7. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. Int J Med Inform 2018;114:57-65. doi: 10.1016/j.ijmedinf.2018.03.013.
8. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, *et al.* Artificial intelligence in healthcare: Past, present and future. Stroke Vasc Neurol 2017;2:230-43. doi: 10.1136/svn-2017-000101.
9. Lin E, Lane HY. Machine learning and systems genomics approaches for multi-omics data. Biomark Res 2017;5:2. doi: 10.1186/s40364-017-0082-y.
10. Kern HP, Reagin MJ, Reese BS. Priming the pump for big data at sentara healthcare. Front Health Serv Manage 2016;32:15-26. doi: 10.1097/01974520-201604000-00003.
11. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. IEEE J Biomed Health Inform 2015;19:1193-208. doi: 10.1109/JBHI.2015.2450362.
12. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med 2014;63:112-5. doi: 10.1016/j.ypmed.2014.01.024.
13. Xu X, Wang G, Chen N, Lu T, Nie S, Xu G, *et al.* Long-term exposure to air pollution and increased risk of membranous nephropathy in China. J Am Soc Nephrol 2016;27:3739-46. doi: 10.1681/asn.2016010093.
14. Zhang L, Wang F, Wang L, Wang W, Liu B, Liu J, *et al.* Prevalence of chronic kidney disease in China: A cross-sectional survey. Lancet 2012;379:815-22. doi: 10.1016/s0140-6736(12)60033-6.
15. Zhang L, Long J, Jiang W, Shi Y, He X, Zhou Z, *et al.* Trends in chronic kidney disease in China. N Engl J Med 2016;375:905-6. doi: 10.1056/NEJMc1602469.
16. Zhang L, Wang H, Long J, Shi Y, Bai K, Jiang W, *et al.* China kidney disease network (CK-NET) 2014 annual data report. Am J Kidney Dis 2017;69:A4. doi: 10.1053/j.ajkd.2016.06.011.
17. Study Resource Overview of China Kadoorie Biobank. China Kadoorie Biobank (CKB), University of Oxford; 2015. Available from: http://www.ckbiobank.org/site/Study+Resources. [Last accessed on 2018 Jun 27].
18. Gao B, Zhang L, Wang H, Zhao M. Chinese cohort study of chronic kidney disease: Design and methods. Chin Med J 2014;127:2180-5.
19. Mehta AJ, Zanobetti A, Bind MA, Kloog I, Koutrakis P, Sparrow D, *et al.* Long-term exposure to ambient fine particulate matter and renal function in older men: The veterans administration normative aging study. Environ Health Perspect 2016;124:1353-60. doi: 10.1289/ehp.1510269.
20. Bowe B, Xie Y, Li T, Yan Y, Xian H, Al-Aly Z, *et al.* Particulate matter air pollution and the risk of incident CKD and progression to ESRD. J Am Soc Nephrol 2018;29:218-30. doi: 10.1681/asn.2017030253.
21. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS, *et al.* A concept-wide association study of clinical notes to discover new predictors of kidney failure. Clin J Am Soc Nephrol 2016;11:2150-8. doi: 10.2215/cjn.02420316.
22. Srinivas TR, Taber DJ, Su Z, Zhang J, Mour G, Northrup D, *et al.* Big data, predictive analytics, and quality improvement in kidney transplantation: A Proof of concept. Am J Transplant 2017;17:671-81. doi: 10.1111/ajt.14099.
23. Pedraza A, Gallego J, Lopez S, Gonzalez L, Laurinavicius A, Bueno G. Glomerulus Classification with Convolutional Neural Networks; Cham: Springer International Publishing; 2017.
24. Colpaert K, Hoste EA, Steurbaut K, Benoit D, Van Hoecke S, De Turck F, *et al.* Impact of real-time electronic alerting of acute kidney injury on therapeutic intervention and progression of RIFLE class. Crit Care Med 2012;40:1164-70. doi: 10.1097/CCM.0b013e3182387a6b.
25. Greenberg JO, Vakharia N, Szent-Gyorgyi LE, Desai SP, Turchin A, Forman J, *et al.* Meaningful measurement: Developing a measurement system to improve blood pressure control in patients with chronic kidney disease. J Am Med Inform Assoc 2013;20:e97-101. doi: 10.1136/amiajnl-2012-001308.
26. Lipkowitz MS, Freedman BI, Langefeld CD, Comeau ME, Bowden DW, Kao WH, *et al.* Apolipoprotein L1 gene variants associate with hypertension-attributed nephropathy and the rate of kidney function decline in African Americans. Kidney Int 2013;83:114-20. doi: 10.1038/ki.2012.263.
27. Yu XQ, Li M, Zhang H, Low HQ, Wei X, Wang JQ, *et al.* A genome-wide association study in han Chinese identifies multiple susceptibility loci for IgA nephropathy. Nat Genet 2011;44:178-82. doi: 10.1038/ng.1047.
28. Zhong Z, Feng SZ, Xu RC, Li ZJ, Huang FX, Yin PR, *et al.* Association of TNFSF13 polymorphisms with IgA nephropathy in a Chinese han population. J Gene Med 2017;19(6-7). doi: 10.1002/jgm.2966.
29. Big Data: The Next Frontier for Innovation, Competition and Productivity. McKinsey and Company; 2011. Available from: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation. [Last accessed on 2018 Jun 27].
30. Hoerger TJ, Simpson SA, Yarnoff BO, Pavkov ME, Ríos Burrows N, Saydah SH, *et al.* The future burden of CKD in the United States: A simulation model for the CDC CKD initiative. Am J Kidney Dis 2015;65:403-11. doi: 10.1053/j.ajkd.2014.09.023.
31. American Diabetes Association. 3. Comprehensive medical evaluation and assessment of comorbidities: Standards of medical care in diabetes-2018. Diabetes Care 2018;41:S28-37. doi: 10.2337/dc18-S003.
32. Whelton PK, Carey RM, Aronow WS, Casey DE Jr., Collins KJ, Dennison Himmelfarb C, *et al.*2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2018;71:e127-248. doi: 10.1016/j.jacc.2017.11.006.
33. Hoerger TJ, Wittenborn JS, Segel JE, Burrows NR, Imai K, Eggers P, *et al.* A health policy model of CKD: 2. The cost-effectiveness of microalbuminuria screening. Am J Kidney Dis 2010;55:463-73. doi: 10.1053/j.ajkd.2009.11.017.
34. Tian R, Yang P, Wang KZ. Joint Registration System under the background of big data. Chin Med J 2017;130:2524-6. doi: 10.4103/0366-6999.217079.