Opinion

Big Data in Chronic Kidney Disease: Evolution or Revolution?

Abbie Kitcher [1], UZhe Ding [1], Henry H. L. Wu [2,*] and Rajkumar Chinnadurai [1,3]

1 Department of Renal Medicine, Northern Care Alliance NHS Foundation Trust, Salford M6 8HD, UK
2 Renal Research Laboratory, Kolling Institute of Medical Research, Royal North Shore Hospital & The University of Sydney, Sydney, NSW 2065, Australia
3 Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9PL, UK
* Correspondence: honlinhenry.wu@health.nsw.gov.au; Tel.: +61-9926-4751

Abstract: Digital information storage capacity and biomedical technology advancements in recent decades have stimulated the maturity and popularization of "big data" in medicine. The value of utilizing big data as a diagnostic and prognostic tool has continued to rise given its potential to provide accurate and insightful predictions of future health events and probable outcomes for individuals and populations, which may aid early identification of disease and timely treatment interventions. Whilst the implementation of big data methods for this purpose is more well-established in specialties such as oncology, cardiology, ophthalmology, and dermatology, big data use in nephrology and specifically chronic kidney disease (CKD) remains relatively novel at present. Nevertheless, increased efforts in the application of big data in CKD have been observed over recent years, with aims to achieve a more personalized approach to treatment for individuals and improved CKD screening strategies for the general population. Considering recent developments, we provide a focused perspective on the current state of big data and its application in CKD and nephrology, with hope that its ongoing evolution and revolution will gradually identify more solutions to improve strategies for CKD prevention and optimize the care of patients with CKD.

Keywords: big data; machine learning; nephrology; chronic kidney disease; prediction models; outcomes; personalized medicine; primary prevention; treatment

In 1965, Gordon Moore described Moore's law, which rightfully predicted the exponential growth of computational capacity. Subsequently, the cost of 1 MB of storage has dropped from USD 1331 to less than USD 0.01 in the past 5 decades [1]. The drastic improvement in digital information storage capacity over the past few decades has led to a propagation in the size and number of available datasets. The result of these advancements is "big data"—colossal and complex data sets that are impossible to process with traditional methods. Big data can be defined by the three Vs—volume, velocity, and variety—initially described in 2001 by Doug Laney. Veracity and value were later added on to form the 'five Vs' in describing big data. The value of big data does not simply reside in its sheer volume, but rather from the analytical processes which can uncover and explore hidden patterns and correlations, and provide better insight and accuracy in the prediction of future events. Predicting potential trajectories in healthcare is imperative as it will aid governing bodies to decide upon longstanding investments and implement effective health policies.

Chronic Kidney Disease (CKD) is a progressive non-communicable disease that affects >10% of the general population worldwide, with 843.6 million individuals being in CKD stages 1–5 [2]. The Global Burden of Disease Studies show that CKD has surfaced as one of the leading causes of worldwide mortality since 1990 [3], and that all-age mortality rate related to CKD rose by 41.5% between 1990 and 2017. In that period, CKD also climbed in rank among the leading causes of death, from 17th in 1990 to 12th in 2017 [4]. Based on a study forecasting life expectancy, Kyle et al.'s model predicted that by 2040, deaths related

to CKD diagnosis will rise to 2.2 million per year in a best-case scenario and even further to 4 million in the worst-case scenario [5]. The cost involved in the care for CKD patients is getting higher—many patients have other comorbidities that necessitate multidisciplinary team care, risk of medical complications that require hospitalisation, and the potential need for dialysis when they reach end-stage kidney disease, which drives up the cost significantly. In the United States alone, the spending for Medicare beneficiaries with kidney disease by 2015 was close to USD 100 billion [6]. Given the immense cost of looking after CKD patients, it is therefore not surprising that there is a huge variation between disability-adjusted life years (DALYs) caused by CKD, more so in countries which are in the lower socio-demographic index quintiles [7].

In comparison to kidney disease, the use of big data in medicine has been more well-established for conditions such as skin cancer and diabetic retinopathy, where over hundreds of thousands of clinical images are fed into data-driven models which are then used for the classification and detection of the aforementioned conditions based on deep convolutional neural networks [8,9]. Another example of big data analysis being successfully utilized is in cardiology, with Loghmanpour et al. [10] demonstrating the superiority of the Bayesian network—a graphical model that is ideal for predicting probable relationships between two events—against the pre-existing traditional risk prediction model in predicting right ventricular failure following left ventricular assist device therapy. In oncology, Jang et al. [11] have also built an extensive clinical and genomic information system from several public databases that aim to aid clinicians in improving diagnostic decision-making, risk assessment, and providing targeted and precise treatment. However, a review of PubMed citations over the previous 2 decades still demonstrates that nephrology is lagging behind other specialties in terms of big data research [12]. In an analysis by Joshi et al. [13], radiology and cardiology were shown to be two of the specialties which showed a drastic increase in the numbers of United States Food and Drug Administration (FDA)-approved machine learning medical devices in the past decade, with the former taking up to 75% of the total amount. Interestingly, there were no nephrology-related machine learning medical devices listed on the FDA website at the time this review was written [14].

There have been increased efforts in the application of big data in CKD (Table 1). Having the ability to predict patient outcomes is essential to achieve targeted preventive medicine. Using traditional regression models based on large cohort studies, Tangri et al. [15] were able to formulate an equation to predict the progression of CKD patients towards end-stage kidney disease. A machine learning algorithm was developed by Ravizza et al. [16] to predict and quantify the risk of CKD progression using real-world data, demonstrating similar or even better predictive accuracy compared to using clinical trial data. Sandokii et al. [17] and Inaguma et al. [18] also replicated successful studies in using machine learning algorithms to identify risk factors and variables in AKI and CKD progression, respectively. A prediction model for end-stage renal disease in primary IgA nephropathy with a 91% success rate was developed by Schena et al. [19]. By applying deep learning techniques to a large data set of 703,872 patients, Tomasev et al. [20] were able to generate a model which had 90.2% accuracy in predicting AKIs requiring dialysis within 90 days.

Inaguma et al. [18] also replicated a similar machine learning algorithm to predict the risk factors for CKD progression. The examples above would not have been possible without pre-existing epidemiological big data. Epidemiological big datasets can come from national registries, surveillance programmes, and electronic health records. For example, the United States Renal Data System (USRDS) is a national surveillance system that compiles and evaluates demographic and clinical information for patients diagnosed with CKD [21]. Similar surveillance projects have also been replicated in Ireland and Canada, which are useful for identifying and describing the prevalence of CKD and improving the care for CKD patients [22,23]. The China Kidney Disease Network (CK-NET)

is set up to integrate and analyse data from China's national database, covering 39 million inpatient electronic records [24].

Developments in biomedical technology over recent years have led to a decrease in the costs of performing high throughput sequencing—also known as next-generation sequencing (NGS)—as well as other biomedical technologies in parallel. This has stimulated an abundance of research efforts focusing on genome-wide association studies (GWAS) and other omics data, such as proteomics (quantification of protein), metabolomics (quantification of metabolites), and transcriptomics (measurement of RNA transcripts), just to name a few. In nephrology, these multi-omics studies paved the way to building "biobanks", such as that of NEPTUNE (Nephrotic Syndrome STudy Network), ERCB (European Renal cDNA Bank), EURenOmics, C-PROBE (Clinical Phenotyping and Resource Biobank), PKU-IgAN, TRIDENT (for diabetic nephropathy), CureGN (for glomerulopathies), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and the Kidney Precision Medicine Project (KPMP) [12,25,26]. When combined with machine learning methods, they can provide clinicians with a deeper understanding of the complexity of molecular events and the pathogenesis of kidney diseases and thus lead to the development of a more precise treatment strategy [12,25,26].

The use of electronic notes and images coupled with artificial intelligence technology has been considered in nephrology research. This has resulted in the design of algorithms that could detect risk factors and identify different stages of CKD from electronic health records [27]. By feeding a convolutional neural network (CNN) with virtual slides of biopsy samples obtained from the Academia and Industry Collaboration for Digital Pathology (AIDPATH) kidney database, Pedraza et al. [28] were also able to demonstrate the encouraging application of artificial intelligence technology at a histopathological level, in which the algorithm they developed was able to achieve a level of accuracy up to 99.5% in differentiating between glomerular and non-glomerular samples. A deep learning framework that could analyse and grade digitized kidney biopsies for fibrosis was generated by using deidentified whole slide images obtained from the Kidney Precision Medicine Project (KPMP) [29].

Table 1. Completed and ongoing research studies relating to the application of big data in chronic kidney disease.

Research Study (Author(s), Journal, Country of Publication, Year of Publication if Specific Details Available)	Summary of Findings and Conclusions
Tangri et al. [15], JAMA, Canada, 2011	• Development and validation of prediction models included 3449 patients and 4942 patients, respectively, from 2 independent Canadian cohorts • A model using routine lab tests can accurately predict the risk of kidney failure in chronic kidney disease patients
Ravizza et al. [16], Nature Medicine, Switzerland, 2019	• Data from 417,912 individual electronic health records were used for the study • Predictive analytic algorithms taught using real world data were shown to be equivalent, if not more accurate, than those taught using clinical trial data
Inaguma et al. [18], PLoS One, Japan, 2020	• Machine-learning-based model included 118,584 patients obtained from an electronic medical records system • Increased urine tendency was found to be a risk factor for rapid decline in kidney function

Table 1. *Cont.*

Research Study (Author(s), Journal, Country of Publication, Year of Publication if Specific Details Available)	Summary of Findings and Conclusions
Pedraza et al. [28], Medical Image Understanding and Analysis, 2017	• Digital/virtual slides were obtained from the AIDPATH (Academia and Industry Collaboration for Digital Pathology) kidney database—a compilation of kidney tissue cohorts from institutions and labs around Europe • Accuracy of convolutional neural networks was observed at 99.95% in differentiating glomerular and non-glomerular samples
Shang et al. [27], NPJ Digital Medicine, United States, 2021	• An algorithm with a 95% positive predictive value in identifying CKD cases in Electronic Health Records (EHR) • The algorithm was validated in EHR from more than 5 institutions and over 1.3 million patients from the Columbia Clinical Data Warehouse
NEPTUNE (Nephrotic Syndrome STudy Network) United States, study due for completion in 2024	• Established to collect long-term observational data with corresponding biological specimens from 1200 patients with nephrotic syndrome across 44 separate institutions in North America
ERCB (European Renal cDNA Bank) database study Germany, ongoing	• A consortium of more than 2600 anonymized kidney biopsies with matching genomic analysis developed from the collaboration of multiple kidney research centres across Europe
EURenOmics database Germany, 2012–2017 Multiple Publications Refer to https://eurenomics.eu/publications/index.html, accessed on 1 February 2023.	• A consortium built with data from more than 15,000 patients to study the pathogenesis of rare nephropathies and to explore new treatment therapies
C-PROBE (Clinical Phenotyping and Resource Biobank) United States, study due for completion in 2025	• Prospective observational study aiming to collect clinical phenotyping of up to 1600 kidney disease patients, laboratory, and histopathology samples
TRIDENT (Transformative Research in diabetic nephropathy) United States, study due for completion in 2023	• Prospective observational study aiming to collect laboratory and histopathology samples combined with high-throughput genomic analysis for patients with diabetic nephropathy
CureGN database study United States and Europe, ongoing	• A multi-centre international consortium of both children and adults with glomerular disease aiming to identify and understand epidemiology, genetics, biomarkers, and patient-related outcomes

Ultimately, potential applications of big data and big data analysis in nephrology are promising, but various limitations and challenges remain. It would make sense that with more information, we would be able to identify previously unrecognized patterns, though this may also provide misleading concepts between causality and correlation. A lot of primary kidney diseases are rare diseases, and the lack of data can sometimes limit the development of accurate prediction models. A relatively smaller funding budget for nephrology research in general compared to other medical specialties has been observed historically, with less clinical trials being conducted in nephrology compared to specialties such as cardiology [30,31]. This may be a hindering factor for the application of big data and big data analysis in nephrology, given a considerable number of clinical trials exclude

patients with CKD as well [32]. It is encouraging that greater efforts have been made by international nephrology societies (e.g., the International Society of Nephrology Advancing Clinical Trials Group) to address these issues over recent years, with initiatives to garner increased industry funding, government support, and patient participation. Another key issue with big data, not only limited to nephrology, is that of 'veracity'—which is the reliability of the collected data—as large retrospective cohort data can suffer from biases, and the data from clinical trials is sometimes not representative of what occurs within the real world [33]. In the current climate where patient privacy is considered invaluable for patients, families, and the clinical team, restrictions and regulations surrounding the collection of health data from wearables, implantable devices, and smartphones remains an issue that needs to be overcome. Protecting patient confidentiality is of the utmost importance and not to be disregarded.

In summary, it appears that the utility of big data in CKD and nephrology research, and integration in clinical practice, is undergoing an evolutionary phase, albeit at a slower pace when compared to other conditions and specialties. The revolutionary aspect of this should take place at an operator level where the users of big data—data scientists, statisticians, health informatics experts, and clinicians—need to gain the skills and direction to effectively translate the findings from big data analysis into clinical practice. At a global health level, we will also need to continuously brainstorm strategies on how best to combine information from big data acquired across various demographics, and search for optimal pathways in utilizing information from big data analysis to prevent CKD and improve CKD outcomes for individuals and populations.

References

1. McCallum, J.; Blok, H. *Historical Cost of Computer Memory and Storage. Our World in Data*; Oxford Martin School, The University of Oxford: Oxford, UK, 2022; Available online: https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage?country=~OWID_WRL (accessed on 1 February 2023).
2. Kovesdy, C.P. Epidemiology of chronic kidney disease: An update 2022. *Kidney Int. Suppl.* **2022**, *12*, 7–11. [CrossRef]
3. Abubakar, I.I.; Tillmann, T.; Banerjee, A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **2015**, *385*, 117–171.
4. Bikbov, B.; Purcell, C.A.; Levey, A.S.; Smith, M.; Abdoli, A.; Abebe, M.; Adebayo, O.M.; Afarideh, M.; Agarwal, S.K.; Agudelo-Botero, M.; et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2020**, *395*, 709–733. [CrossRef]
5. Foreman, K.J.; Marquez, N.; Dolgert, A.; Fukutaki, K.; Fullman, N.; McGaughey, M.; Pletcher, M.A.; Smith, A.E.; Tang, K.; Yuan, C.W.; et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and territories. *Lancet* **2018**, *392*, 2052–2090. [CrossRef] [PubMed]
6. Saran, R.; Robinson, B.; Abbott, K.C.; Agodoa, L.Y.; Bhave, N.; Bragg-Gresham, J.; Balkrishnan, R.; Dietrich, X.; Eckard, A.; Eggers, P.W.; 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. [CrossRef]
7. Cockwell, P.; Fisher, L.-A. The global burden of chronic kidney disease. *Lancet* **2020**, *395*, 662–664. [CrossRef]
8. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]

9. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photo-graphs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]

10. Loghmanpour, N.A.; Kormos, R.L.; Kanwar, M.K.; Teuteberg, J.J.; Murali, S.; Antaki, J.F. A Bayesian Model to Predict Right Ventricular Failure Following Left Ventricular Assist Device Therapy. *JACC: Hear. Fail.* **2016**, *4*, 711–721. [CrossRef] [PubMed]

11. Jang, Y.; Choi, T.; Kim, J.; Park, J.; Seo, J.; Kim, S.; Kwon, Y.; Lee, S.; Lee, S. An integrated clinical and genomic information system for cancer precision medicine. *BMC Med. Genom.* **2018**, *11*, 34. [CrossRef]

12. Saez-Rodriguez, J.; Rinschen, M.M.; Floege, J.; Kramann, R. Big science and big data in nephrology. *Kidney Int.* **2019**, *95*, 1326–1337. [CrossRef]

13. Joshi, G.; Jain, A.; Adhikari, S.; Garg, H.; Bhandari, M. FDA approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An updated 2022 landscape. *medRxiv* **2022**. [CrossRef]

14. Food and Drug Administration, US. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. 2022. Available online: https://www.fda.gov/media/145022 (accessed on 1 February 2023).

15. Tangri, N.; Stevens, L.A.; Griffith, J.; Tighiouart, H.; Djurdjev, O.; Naimark, D.; Levin, A.; Levey, A.S. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA* **2011**, *305*, 1553–1559. [CrossRef]

16. Ravizza, S.; Huschto, T.; Adamov, A.; Böhm, L.; Büsser, A.; Flöther, F.F.; Hinzmann, R.; König, H.; McAhren, S.M.; Robertson, D.H.; et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat. Med.* **2019**, *25*, 57–59. [CrossRef] [PubMed]

17. Sandokji, I.; Yamamoto, Y.; Biswas, A.; Arora, T.; Ugwuowo, U.; Simonov, M.; Saran, I.; Martin, M.; Testani, J.M.; Mansour, S.; et al. A Time-Updated, Parsimonious Model to Predict AKI in Hospitalized Children. *J. Am. Soc. Nephrol.* **2020**, *31*, 1348–1357. [CrossRef]

18. Inaguma, D.; Kitagawa, A.; Yanagiya, R.; Koseki, A.; Iwamori, T.; Kudo, M.; Yuzawa, Y. Increasing tendency of urine protein is a risk factor for rapid eGFR decline in patients with CKD: A machine learning-based prediction model by using a big database. *PLoS ONE* **2020**, *15*, e0239262. [CrossRef]

19. Schena, F.P.; Anelli, V.W.; Trotta, J.; Di Noia, T.; Manno, C.; Tripepi, G.; D'Arrigo, G.; Chesnaye, N.C.; Russo, M.L.; Stangou, M.; et al. Development and testing of an artificial intelligence tool for predicting end-stage kidney disease in patients with im-munoglobulin A nephropathy. *Kidney Int.* **2021**, *99*, 1179–1188. [CrossRef] [PubMed]

20. Tomašev, N.; Harris, N.; Baur, S.; Mottram, A.; Glorot, X.; Rae, J.W.; Zielinski, M.; Askham, H.; Saraiva, A.; Magliulo, V.; et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat. Protoc.* **2021**, *16*, 2765–2787. [CrossRef]

21. Port, F.K.; Held, P.J. The US Renal Data System at 30 Years: A Historical Perspective. *Am. J. Kidney Dis.* **2019**, *73*, 459–461. [CrossRef] [PubMed]

22. Stack, A.G.; Casserly, L.F.; Cronin, C.J.; Chernenko, T.; Cullen, W.; Hannigan, A.; Saran, R.; Johnson, H.; Browne, G.; Ferguson, J.P. Prevalence and variation of Chronic Kidney Disease in the Irish health system: Initial findings from the National Kidney Disease Surveillance Programme. *BMC Nephrol.* **2014**, *15*, 185. [CrossRef] [PubMed]

23. Bello, A.K.; E Ronksley, P.; Tangri, N.; Singer, A.; Grill, A.; Nitsch, D.; A Queenan, J.; Lindeman, C.; Soos, B.; Freiheit, E.; et al. A national surveillance project on chronic kidney disease management in Canadian primary care: A study protocol. *BMJ Open* **2017**, *7*, e016267. [CrossRef]

24. Saran, R.; Steffick, D.; Bragg-Gresham, J. The China kidney disease network (CK-NET): "big data-big dreams". *Am. J. Kidney Dis.* **2017**, *69*, 713–716. [CrossRef] [PubMed]

25. 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.* **2015**, *31*, 2003–2011. [CrossRef]

26. Wuttke, M.; Köttgen, A. Insights into kidney diseases from genome-wide association studies. *Nat. Rev. Nephrol.* **2016**, *12*, 549–562. [CrossRef] [PubMed]

27. Shang, N.; Khan, A.; Polubriaginof, F.; Zanoni, F.; Mehl, K.; Fasel, D.; Drawz, P.E.; Carrol, R.J.; Denny, J.C.; Hathcock, M.A.; et al. Medical records-based chronic kidney disease phenotype for clinical care and "big data" observational and genetic studies. *NPJ Digit. Med.* **2021**, *4*, 1–13. [CrossRef]

28. Pedraza, A.; Gallego, J.; Lopez, S.; Gonzalez, L.; Laurinavicius, A.; Bueno, G. Glomerulus classification with convolutional neural networks. In Proceedings of the Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, 11–13 July 2017; Springer International Publishing: Berlin/Heidelberg, Germany; pp. 839–849.

29. Zheng, Y.; Cassol, C.A.; Jung, S.; Veerapaneni, D.; Chitalia, V.C.; Ren, K.Y.; Bellur, S.S.; Boor, P.; Barisoni, L.M.; Waikar, S.S.; et al. Deep-Learning–Driven Quantification of Interstitial Fibrosis in Digitized Kidney Biopsies. *Am. J. Pathol.* **2021**, *191*, 1442–1453. [CrossRef]

30. Erickson, K.F.; Hostetter, T.H.; Winkelmayer, W.C.; Olan, G.; Meyer, R.N.; Hakim, R.; Sedor, J.R. Federal Funding for Kidney Disease Research: A Missed Opportunity. *Am. J. Public Heal.* **2016**, *106*, 406–407. [CrossRef]

31. Strippoli, G.F.M.; Craig, J.; Schena, F.P. The Number, Quality, and Coverage of Randomized Controlled Trials in Nephrology. *J. Am. Soc. Nephrol.* **2004**, *15*, 411–419. [CrossRef]

32. Banerjee, D.; Lowe-Jones, R.; Damster, S.; Thomas, N.; Scholes-Robertson, N.; Tong, A.; Levin, A. International perspectives on patient involvement in clinical trials in nephrology. *Kidney Int.* **2020**, *98*, 566–571. [CrossRef]

33. Glicksberg, B.; Johnson, K.; Dudley, J.T. The next generation of precision medicine: Observational studies, electronic health records, biobanks and continuous monitoring. *Hum. Mol. Genet.* **2018**, *27*, R56–R62. [CrossRef]