KI REPORTS
KIReports.org

Public Data in Nephrology: A Researcher's Roadmap

Eoin Daniel O'Sullivan[1,2,3], Monica Suet Ying Ng[1,4,5], Venkat Vangaveti[6,7], Nicholas Matigian[8] and Andrew John Mallett[2,6,7]

[1]Kidney Health Service, Royal Brisbane and Women's Hospital, Herston, Queensland, Australia; [2]Institute for Molecular Bioscience, The University of Queensland, Herston, Queensland, Australia; [3]QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; [4]Faculty of Medicine, The University of Queensland, Herston, Queensland, Australia; [5]Conjoint Internal Medicine Laboratory, Chemical Pathology, Pathology Queensland, Herston, Queensland, Australia; [6]College of Medicine and Dentistry, James Cook University, Townsville, Queensland, Australia; [7]Townsville Institute of Health Research and Innovation, Townsville University Hospital, Townsville, Queensland, Australia; and [8]QCIF Bioinformatics, Brisbane, Queensland, Australia

Kidney Int Rep (2025) **10,** 1609–1612; https://doi.org/10.1016/j.ekir.2025.04.010

KEYWORDS: bioinformatics; data science; open-source databases; public datasets; research

Introduction

Access to high-quality data is essential for advancing nephrology research; however, many researchers face significant barriers to access these resources. The cost of data acquisition and the complexity of navigating data repositories can hinder equitable access, particularly for those in smaller institutions or developing regions. This may be compounded by a lack of awareness of the range of multidisciplinary data sources among researchers, in part because of increasing research subspecializations; for example, basic scientists may not be familiar with the range of clinical datasets. Paywalls may present an additional obstacle, and it is noteworthy that of the top 20 Nephrology journals, only 4 are entirely open access, with 15 taking a hybrid approach.[1]

Nephronexus.com is an online directory of publicly available research data in nephrology. Here, we review this resource, which has organized data according to biological scale starting at the largest data scale, national registries and large patient cohorts, before moving to tissues, protein expression, transcriptomics and concluding at the gene level. By identifying and collating these repositories and datasets, and highlighting their key features, we aim to provide a helpful resource to researchers, to empower them to leverage existing data, maximizing resource efficiency, and promoting collaborative efforts in the nephrology community. By enhancing access, we aim to facilitate more equitable research opportunities, improve the quality of *in silico* studies, encourage maximum utilization of data, greater collaboration between disciplines, and help drive forward the understanding and treatment of kidney diseases.

Correspondence: Eoin Daniel O'Sullivan, Kidney Health Service, Royal Brisbane and Women's Hospital, Herston, Queensland, Australia. E-mail: eoin.osullivan@health.qld.gov.au

Patient-Level Data

Patient-level data from registries and disease-specific cohorts offer a wealth of opportunities for nephrology research. Examples include epidemiological studies to estimate disease prevalence and incidence across populations, enabling the identification of high-risk groups and tracking disease progression over time, and between-group comparisons to reveal disparities or opportunities.[2-6] Such data are also invaluable for identifying risk factors, because patient demographics, comorbidities, and lifestyle information allow researchers to explore associations with disease onset and progression and outcomes.

Longitudinal analysis is another major application of patient-level data. Registries, which often collect information over extended periods, enable researchers to track important long-term outcomes such as the progression to chronic kidney disease, the initiation of dialysis, or patient mortality. This real-world data is essential for evaluating the effectiveness of treatments and interventions outside controlled clinical trials, providing insights into how therapies and policies perform across diverse patient populations.[7-9]

These data can drive clinical decision support systems by enabling the development of predictive models tailored to individual patient profiles. Such models facilitate personalized medicine approaches, where treatments are customized based on patient-specific factors. Risk stratification tools such as the kidney failure risk equation, developed from these datasets, assists clinicians in predicting disease progression and planning care accordingly, and can even help enrich trial recruitment to improve event rate and power.[S1–S3]

Finally, registries and disease-specific cohorts are indispensable in clinical trials, both for identifying eligible participants and for assessing whether trial results are generalizable to broader patient populations. By embedding clinical trials within existing health care structures, researchers can conduct more cost-effective trials, leveraging preexisting data to facilitate recruitment and outcome measurement.[S4–S6] This embedded trial design can significantly reduce costs and improve trial efficiency. In Supplementary Table S1, we present multiple international registries of patients with kidney disease (Supplementary references). Disease-specific clinical cohorts are described in Supplementary Table S2.

Tissue-Level Data

Patient-level and cohort data are limited in their ability to reveal pathophysiology or generate new biological insight. Here, tissue-level data plays a fundamental role in understanding kidney disease at the organ, cellular, and molecular scale. Through biobanks and kidney-specific tissue atlases, researchers can gain insights into the spatial organization and cellular composition of the kidney. By leveraging these detailed tissue resources, nephrology researchers can achieve deeper understanding than patient data alone.

Biobanks collect and store various biological materials, including blood, urine, kidney tissue, and genetic samples, from individuals with different kidney conditions. By linking these samples with detailed clinical data, biobanks provide researchers with a powerful tool for investigating the underlying mechanisms, biomarkers, and genetic factors associated with kidney diseases. Moreover, these resources can facilitate the development of novel diagnostic tools, prognostic indicators, and targeted therapies tailored to individual patients. These data are contained in Supplementary Table S3.

Tissue atlases in kidney research represent comprehensive catalogs of tissue of healthy and diseased kidneys. These atlases offer researchers a wealth of information about cell architecture, spatial distribution gene and protein expression, and cellular interactions within the kidney across different developmental stages, physiological conditions, and disease states. In Supplementary Table S4, we present these atlases as resources for exploring disease mechanisms, validating a researcher's own transcriptomic or genetic data, or testing preliminary hypotheses without expensive experiments.

Transcriptomic Data

RNA sequencing (RNA-Seq) has revolutionized the field of nephrology research by providing unprecedented insights into gene expression patterns and mRNA dynamics within the kidney. With over 1000 tools now publicly available for use, and increasing consensus about best practice and optimal workflow, these datasets provide rich insights for researchers who can access and analyze them. The key repositories for sequencing data include Gene Expression Omnibus (GEO), Expression Atlas, BioProject, and Array express.

GEO is specifically designed for storing and sharing gene expression data, particularly from microarray and RNA-Seq experiments. Managed by the National Center for Biotechnology Information, GEO archives high-throughput gene expression data and allows users to search, analyze, and visualize datasets. Researchers can access a vast collection of RNA-Seq data, including raw sequence reads and processed gene expression levels, from various studies related to kidney health and disease. GEO not only facilitates the sharing of data but also provides tools for meta-analysis, enabling scientists to glean insights from multiple studies. The Expression Atlas contains RNA-Seq and microarray data and is maintained by the European Bioinformatics Institute, and focuses on gene expression data across different conditions, tissues, and developmental stages.

In contrast to GEO and the European Bioinformatics Institute's Expression Atlas, BioProject provides a broader, project-oriented view of biological data. It is a National Center for Biotechnology Information resource that organizes information about scientific projects and datasets associated with the study of biological systems. It serves as a repository for various genomic and transcriptomic data, linking to multiple data types, including RNA-Seq.

Finally, ArrayExpress is another database managed by the European Bioinformatics Institute, designed specifically for storing and sharing functional genomics data, including RNA-Seq, microarray, and other high-throughput studies. ArrayExpress includes functional genomics data, encompassing not only gene expression data but also epigenomics and proteomics. Each data set in ArrayExpress comes with detailed annotations, including experimental design, sample characteristics, and platform details, ensuring that researchers can thoroughly understand the data.

These repositories are searchable with standard MESH terms,[S7] the data are often submitted to repositories to support

publication, where the focus is data storage, technical specifications of the sequencing workflow and generic clinical description of the datasets. Supplementary Table S5 is a curated, kidney-specific RNA-Seq and multiomics data relevant to nephrology research.

Genomic Data

Through high-throughput sequencing technologies and genome-wide association studies, researchers search for genetic variations, rare mutations, and polymorphisms implicated in kidney disease. These databases and projects involve sequencing of germline DNA, often in large populations, and may be linked to phenotypic or health data. Understanding normal variation is a key part of genomic analysis, and we include resources that focus on common genetic variants in the population, used to understand normal genetic diversity. We also provide a selection of key diagnostic and research cohorts. These cohorts are typically smaller, disease-focused, and may involve highly specialized research groups as described in Supplementary Table S6.

Artificial Intelligence Training Data

Computational tools are at the forefront of modern research. The 2024 Nobel Prize in Chemistry was awarded to John Jumper and Demis Hassabis from Google DeepMind for the development of Alphafold2, which is an AI model used for the prediction of protein structure.[S8,S9] With the growing application of artificial intelligence (AI) and machine learning in biomedical research, access to well-curated and annotated datasets is critical for training accurate and robust models. There are many roles for AI and machine learning approaches in nephrology and transplantation, including data analysis, discovery science, clinical prediction, and decision-making support and patient facing communication roles.

Models require training and several public resources offer kidney-specific datasets. These datasets are sourced from diverse platforms and include both clinical and research data, allowing for advanced AI applications such as kidney tissue segmentation, histopathological analysis, and disease classification. In Supplementary Table S7, we provide key publicly available AI training datasets relevant to kidney research.

Conclusion

The landscape of publicly available data in nephrology offers many opportunities for researchers to advance our understanding of kidney diseases through *in silico* approaches. By leveraging public resources, such as registries, biobanks, transcriptomic datasets, and AI training data, researchers can conduct innovative studies with less cost and effort. Such resources are maintained at nephronexus.com

Using publicly available datasets can present challenges. There may be ethical and governance considerations that will need to be navigated, particularly when accessing patient-level data. Researchers may encounter data-related issues, such as inconsistent formats, incomplete datasets or limited metadata. This can hinder reproducibility and interpretation. Datasets generated by others might have specific technical biases or natural limitations based on their workflow, experimental design, and research goals. This is not always obvious to a new researcher, thereby potentially limiting their utility for secondary analyses. Finally, a degree of technical skills and computational resources will be required, which will vary based upon the nature of the data and task at hand. Despite these challenges, the benefits of using publicly available datasets outweigh the challenges in many cases. By facilitating cost-effective and wide-reaching studies, these resources help advance nephrology research, foster a more inclusive community and improve patient outcomes.

DISCLOSURE

All the authors declared no competing interests.

SUPPLEMENTARY MATERIAL

Supplementary File (PDF)

Contains 7 supplementary tables documenting the tabulated public data provided on the Nephronexus website as of April 2025 and extended list of references.

Supplementary References.

Table S1. Patient registries.

Table S2. Disease-specific cohorts.

Table S3. Biobanks.

Table S4. Tissue atlases.

Table S5. Single-cell resources.

Table S6. Genetic resources.

Table S7. AI training resources.

REFERENCES

1. Stauss M, Floyd L, Woywodt A. Weighing up open access publishing in nephrology—bronze, platinum, or fools' gold? *Kidney360*. 2023;4(11):1637–1640.
2. Yamagata K. Trends in the incidence of kidney replacement therapy: comparisons of ERA, USRDS and Japan registries. *Nephrol Dial Transplant*. 2023;38:797–799. https://doi.org/10.1093/ndt/gfac312
3. UK Renal Registry. 25th annual report-data to 31/12/2021. Accessed March 1, 2025. https://ukkidney.org/audit-research/annual-report/25th-annual-report-data-31122021
4. Robinson BM, Akizawa T, Jager KJ, Kerr PG, Saran R, Pisoni RL. Factors affecting outcomes in patients reaching end-stage kidney disease worldwide: differences in access to renal

replacement therapy, modality use, and haemodialysis practices. *Lancet Lond Engl*. 2016;388:294–306. https://doi.org/10.1016/S0140-6736(16)30448-2

5. Stel VS, Boenink R, Astley ME, et al. A comparison of the epidemiology of kidney replacement therapy between Europe and the United States: 2021 data of the ERA Registry and the USRDS. *Nephrol Dial Transplant*. 2024;39:1593–1603. https://doi.org/10.1093/ndt/gfae040
6. United States renal data system. 2024 USRDS annual data report: Epidemiology of kidney disease in the United States.National Institute of Diabetes and Digestive and Kidney Diseases. National Institutes of Health. https://usrds-adr.niddk.nih.gov/2024
7. Nagasu H, Yano Y, Kanegae H, et al. Kidney outcomes associated with SGLT2 inhibitors versus other glucose-lowering drugs in real-world clinical practice: the Japan chronic kidney disease database. *Diabetes Care*. 2021;44:2542–2551. https://doi.org/10.2337/dc21-1081
8. Stengel B, Metzger M, Combe C, et al. Risk profile, quality of life and care of patients with moderate and advanced CKD: the French CKD-REIN Cohort Study. *Nephrol Dial Transplant*. 2019;34:277–286. https://doi.org/10.1093/ndt/gfy058
9. Lash JP, Go AS, Appel LJ, et al. Chronic renal insufficiency cohort (CRIC) study: baseline characteristics and associations with kidney function. *Clin J Am Soc Nephrol*. 2009;4:1302–1311. https://doi.org/10.2215/CJN.00070109