Check for updates

Big Data in Nephrology

Navchetan Kaur[1,2], *Sanchita Bhattacharya*[1,2] *and Atul J. Butte*[1,2,3] ✉

Abstract | A huge array of data in nephrology is collected through patient registries, large epidemiological studies, electronic health records, administrative claims, clinical trial repositories, mobile health devices and molecular databases. Application of these big data, particularly using machine-learning algorithms, provides a unique opportunity to obtain novel insights into kidney diseases, facilitate personalized medicine and improve patient care. Efforts to make large volumes of data freely accessible to the scientific community, increased awareness of the importance of data sharing and the availability of advanced computing algorithms will facilitate the use of big data in nephrology. However, challenges exist in accessing, harmonizing and integrating datasets in different formats from disparate sources, improving data quality and ensuring that data are secure and the rights and privacy of patients and research participants are protected. In addition, the optimism for data-driven breakthroughs in medicine is tempered by scepticism about the accuracy of calibration and prediction from in silico techniques. Machine-learning algorithms designed to study kidney health and diseases must be able to handle the nuances of this specialty, must adapt as medical practice continually evolves, and must have global and prospective applicability for external and future datasets.

Big data are defined by the intertwined orders of magnitude of volume, variety, velocity, and veracity of data (FIG. 1). The term big data is generally used to describe large, complex data sets that require high-end computing solutions for their storage and analysis. In medicine, big data present an excellent opportunity to innovate and improve healthcare through the development of data-driven applications for use in research and clinical practice. Such applications have already been used to guide policies for refinement of dialysis care[1], to improve monitoring of medical products and drugs by regulatory authorities[2], to detect diabetic retinopathy in retinal fundus photographs[3] and to detect patients at a high risk of clinical deterioration in large hospital settings and therefore enable interventions to reduce mortality[4].

Data science could be critical in revolutionizing future medical practice by enabling early disease prediction and thereby providing opportunities for prevention as well as by facilitating personalized therapy and patient involvement in data-driven treatment decisions[5]. In addition, big data approaches could yield innovative applications for use in clinics and research laboratories[6]. The field of nephrology has not yet fully explored the enormous potential of big data. However, this field had the earliest experience in systematically collecting patient-level data on kidney failure, chronic kidney disease (CKD), kidney replacement therapies (KRTs) and transplantation in national registries[7]. The United States Renal Data System (USRDS), the Scientific Registry of Transplant Recipients, the UK Renal Registry (UKRR) and the European Renal Association–European Dialysis and Transplant Association (ERA–EDTA) Registry regularly report trends in disease burden, efficacy of standard of care and transplantation practices[8].

The increasing digitization of healthcare has led to a huge influx of clinical data from sources including electronic health records (EHRs), medical billing and insurance claims[9]. In addition, questionnaire-based national population surveys such as the National Health and Nutrition Examination Survey (NHANES) and National Inpatient Sample (NIS) and other large epidemiological and cohort studies, such as those conducted by the CKD-Prognosis Consortium, are eminent sources of data on the incidence and prevalence of disease[10–13]. Advances in wearable technologies and mobile health devices enable digital surveys of lifestyle and behaviour in real-time and generate huge amounts of data on compliance and healthcare delivery. In parallel, multiplexing of laboratory tests and omics technologies have enabled researchers to obtain high-throughput data from studies of molecular genetics and biology, such as genome-wide association studies and studies of single nucleotide variants, single-cell genomics, transcriptomics, metabolomics and proteomics[14].

Linkage of genomic and clinical data sources has created biomedical big data[15]. In the past few years the number of open-access platforms hosting clinical, research and clinical trials data related to various domains, including nephrology, has exponentially increased[16].

[1]*Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA.*

[2]*Department of Pediatrics, University of California, San Francisco, CA, USA.*

[3]*Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA, USA.*

✉*e-mail: atul.butte@ucsf.edu*

https://doi.org/10.1038/s41581-021-00439-x

Key points

- Big data in nephrology can provide essential information about kidney disease burden, molecular mechanisms, novel risk factors and therapeutic targets.
- Artificial intelligence and machine-learning approaches that utilize big data could be used for a variety of applications in nephrology, including early diagnosis and prognosis, as well as clinical decision-support systems for personalized selection of therapy.
- Data curation and standardization enable interoperability, facilitate consolidation and exchange of high-quality data from different sources, create independence from manufacturers and ease competition as comparable products are offered by all market players.
- Sources of big data in nephrology include patient registries, population surveys, electronic health records, open-access clinical trials, mobile health devices and molecular data repositories.
- Large-scale acquisition of annotated molecular and clinical data, together with advances in machine learning approaches, open-source computational packages, affordable computation power and cloud storage, will all facilitate more novel data-driven approaches in nephrology.
- Challenges for the utilization of big data in nephrology include issues relating to data governance and protection, siloed datasets, data heterogeneity, small sample sizes and a lack of consistent research funding.

In addition, machine learning applications provide a myriad of opportunities for researchers and clinicians to use huge datasets and identify actionable clinical insights that cannot be detected by the human brain.

In this Review, we discuss the large patient-oriented data sources that are available in the field of nephrology, and how these data sets can be utilized in nephrology research and clinical practice.

Data types, curation and standards

Five major sources of big data are available in the field of nephrology: patient registry and epidemiology data, EHRs and administrative data, clinical trial data, data from mobile health devices and molecular data (FIG. 2). The data elements from these data sources are stored in various schemas and can be structured, unstructured or semi-structured. Unstructured data include text, images, sounds and videos, structured data adhere to a human-readable, well-defined tabular format and semi-structured data are data in machine-readable formats such as JavaScript Object Notation or Extensible Markup Language.

Data curation ensures the accuracy, consistency, credibility and timeliness of these different data types so that high value data are available for analytics and research. The data curation pipeline starts with extraction, transformation and loading. After data extraction, the transformation process converts data or information from one format to another, usually from the format of a source system into the required format of the destination system. Several manual and automated frameworks are then followed for data collation, preservation and utilization[17]. This process detects inconsistencies, missing values, outliers and similarities as well as executing data standardization and data harmonization. Common data models (CDMs) that have standardized schema are implemented by institutions to store and represent their data. Popular CDMs for clinical data include Sentinel[18], the Patient-Centered Outcomes Research Network CDM, informatics for integrating biology in the bedside (I2B2), and the Observational Medical Outcomes Partnership (OMOP)[19–21].

Harnessing big data across a variety of data types requires interoperability. FAIR (Findable, Accessible, Interoperable, Reusable) data principles have been formulated to enhance the ability of machines to automatically find and use data while retaining their value[22]. These kinds of principles are implemented by using data standards that involve the definition of data elements, data interchange formats, terminologies and knowledge representation (Supplementary Table 1). These standards are usually developed through community-driven consensus with a major emphasis on uniformity in salient features, criteria, methods, processes and practices. Data standardization creates independence from manufacturers, as goods and services are developed based on consensus and have consistent quality. Standardization also

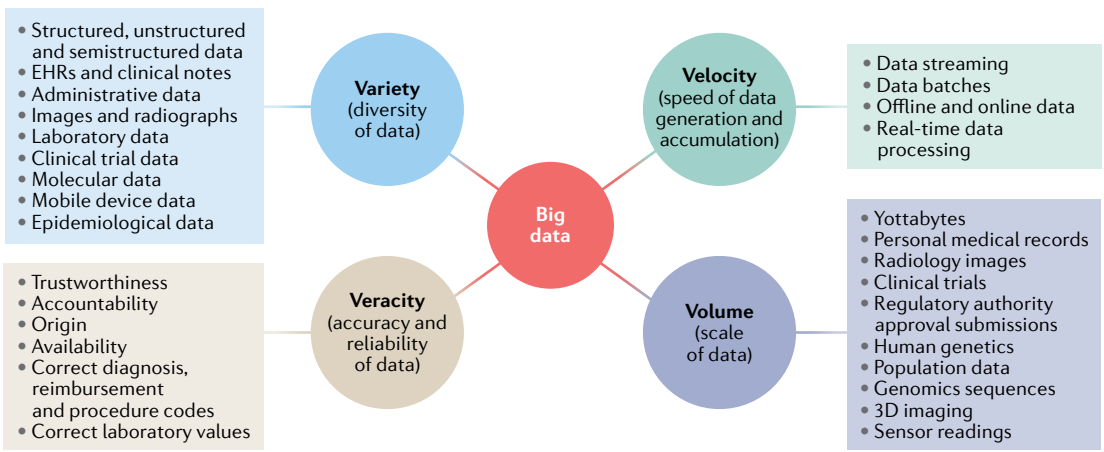

Fig. 1 | **Characteristics of big data.** Big data are characterized by variety, velocity, veracity and volume[5]. The variety results from the collection of structured, semi-structured and unstructured data from diverse sources. The velocity represents the high speed of data generation and accumulation. The veracity depends on the level of inconsistency and uncertainty in the data and the volume represents the enormous size of the datasets. Thus, distinct methods and specialized algorithms are required to exploit the potential of big data. EHRs, electronic health records.

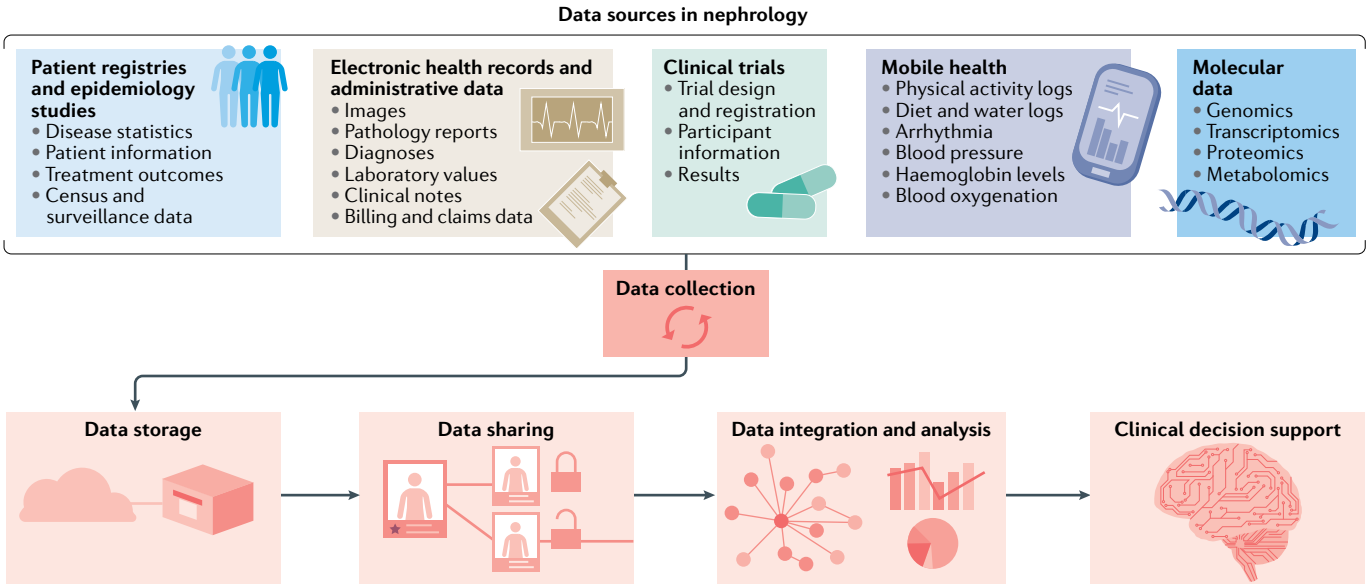

Fig. 2 | **Workflow for use of data for clinical decision support systems in nephrology.** The five major sources of data in nephrology are patient registries and epidemiological studies, electronic health records and administrative claims, clinical trials, mobile health devices and molecular studies. These data must first be curated and stored and can then be shared with stakeholders such as clinicians, researchers, biotech and pharmaceutical companies and governments. After data integration using techniques such as harmonization and normalization, the data can be analysed to build models that facilitate clinical decision support using artificial intelligence and machine-learning approaches. This process continues as new data are extracted from different sources.

eases competition between market players, as comparable products are offered at industry set standards[23,24].

Patient registries and epidemiology data

Large volumes of data collected by patient registries, national and international health surveys and large cohort studies constitute big data in healthcare. Data from these sources can help to describe the natural history, epidemiology and burden of a disease and capture treatment sites as well as regional or national variations in treatment and outcomes. Such data can be used to evaluate the safety, quality and value of patient care and could potentially help researchers to develop new hypotheses about disease mechanisms or treatment approaches that could inform healthcare policy and potentially lead to improvements in the quality of care[25].

Patient registries use organized and standardized methods to systematically collect observational data about specific groups of patients managed in routine clinical practice and typically have a predetermined objective, for example, monitoring quality of care for patients with kidney failure. They are the oldest big data source in nephrology as they were first started in the early 1960s to collect data on patients undergoing dialysis[26] (FIG. 3). National and international health surveys such as NHANES, the US National Health Interview Survey and the European Health Examination Survey are generally conducted to provide prevalence estimates for several chronic conditions[27].

Large prospective cohort studies such as the Dialysis Outcome and Practice Patterns Study (DOPPS) are another important source of big data. DOPPS began collecting data on patients undergoing haemodialysis in facilities in the USA in 1996 and was expanded to include facilities in Japan, Germany, France, Italy and the UK by 1999. In addition to haemodialysis, the study now also collects data on CKD and peritoneal dialysis. The DOPPS Practice Monitor was started in 2010 to monitor trends in dialysis and patient care and has contributed to changes in policies such as revision of the Food and Drug Adminstration (FDA) label for prescription of erythropoietin stimulating agents to patients with CKD[28,29]. In 2012, the International Network of CKD Cohort Studies was established by the International Society of Nephrology to promote collaborative research and improve understanding of the natural history of the disease[30]. This network comprises prospective, observational studies of patients with CKD that contribute study and individual-level data for meta-analyses.

Large volume datasets for CKD in the US are available from NHANES, USRDS[31], the Kaiser Permanente CKD registry[32] and the Veterans Affairs EHR system. USRDS is jointly funded by the Health Care Financing Administration and the National Institutes of Health (NIH)–National Institute of Diabetes and Digestive and Kidney Diseases and uses data collected by the Centers for Medicare and Medicaid Services. These data are used to evaluate the effects of changing dialysis regimens, the quality of patient care, epidemiological trends and cost-effectiveness of treatment regimens[33]. In addition to producing an annual data report, USRDS allows access to the data through the RenDER data extraction and referencing system.

The Partners Healthcare System has developed an EHR-based CKD registry that includes >65,000 patients in academic and community hospitals and outpatient

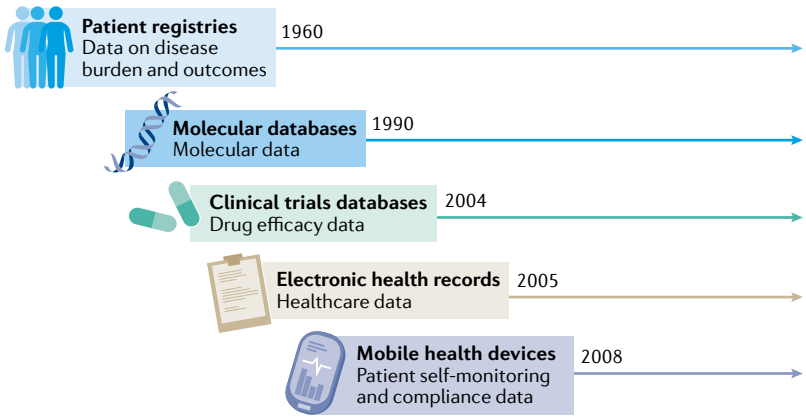

Fig. 3 | **Timeline of the availability of big data in nephrology.** Big data have been collected in the field of nephrology since the early 1960s when the European Renal Association–European Dialysis and Transplant Association established the first patient registry[26]. In the early 1990s the advent of internet and databases enabled molecular data to be made available to researchers. Clinical trials databases such as clinicaltrials.gov and the European Union Drug Regulating Authorities Clinical Trials Database were established in the early 2000s to share summaries of clinical trials. Electronic Health Records were adopted in the USA and UK in 2005 (REFS[138,139]). Data from mobile health devices, including smartphones, wearable activity trackers, blood-pressure cuffs, pulse oximeters and glucometers, began to become available in 2008 (REFS[140,141]). All of these data sources have evolved over time and their availability has led to the advent of new fields, for example, electronic health records and claims data provide the foundation for clinical informatics. Big data, coupled with extensive computational power, are now a major contributor to digital health.

care facilities in Massachusetts, USA[34]. This registry collects data on patient care with the aim of identifying opportunities for improvement and population health strategies. These data include rates of annual proteinuria testing, vascular access placement, transplantation referrals and outpatient nephrology contact. Other notable CKD registries in the USA include CURE-CKD, which is a collaboration between the University of California Los Angeles Health and Providence St Joseph Health[35], and the Cleveland Clinic CKD registry[36]. The National Kidney Foundation (NKF) is also developing a CKD registry called the National Kidney Foundation Patient Network.

The UKRR was established by the UK Renal Association with the primary aim of collecting data from all adult UK renal centres to improve the care of patients with kidney failure on KRT and kidney transplant recipients. The UKRR also collects data on all cases of acute kidney injury (AKI) in primary and secondary care and on CKD stages 4 and 5 with the aim of understanding progression to kidney failure[37,38]. A joint collaboration between the UKRR and Kidney Care UK, Think Kidneys, is leading several projects to establish local and national data collection systems for AKI and facilitate research to improve AKI care. The ERA–EDTA collects data on patients with kidney failure from the national and regional kidney disease registries of more than 36 countries in Europe and bordering the Mediterranean Sea, covering a general population of 686.9 million[39].

The Australia and New Zealand Dialysis and Transplant Registry collects data on all patients treated with KRT in Australia and New Zealand[40]. These data include demographic details, primary kidney disease, type of KRT, process measures and a variety of outcomes. The Registry Of Kidney Diseases is a clinical quality registry that collects data from patients with kidney diseases across Australia. The primary objective of the Registry Of Kidney Diseases is to identify patients during the early stages of CKD, monitor the course of their disease and assess the quality of care that they receive.

Most kidney disease registries provide broad data on treatment, including peritoneal dialysis and haemodialysis (Supplementary Table 2). However, some registries, such as the Danish Registry on Regular Dialysis and Transplantation, the Finnish Registry for Kidney Diseases, the Scottish Renal Registry, the French Language Peritoneal Dialysis Registry, the UKRR, the Argentina Register of Chronic Dialysis and the Uruguayan Registry of Dialysis also allow external access to patient-level data with KRT modality and submodality as well as outcomes data. The Global Health Data Exchange portal is a comprehensive source for descriptions of these registries[41].

Several registries focus on rare kidney diseases. For example, the Rare Kidney Stone Consortium has established international registries for primary hyperoxaluria, cystinuria, adenine phosphoribosyl transferase deficiency and Dent disease[42]. The NephCure Kidney Patient Registry collects patient-reported information from patients with nephrotic syndrome in 33 countries worldwide in a centralized location that is easily accessible to researchers.

Some patient-initiated registries also exist, including PatientsLikeMe and their research initiative DigitalMe, which offer disease-related data that are voluntarily shared by patients. This registry includes data from patients with CKD, polycystic kidney disease, medullary cystic disease, AKI and kidney transplant recipients. Descriptive data on symptoms, treatment regimens, demographics, perceived effectiveness of treatments and adverse effects is freely available in DigitalMe.

EHRs and healthcare administrative data

Healthcare providers maintain EHR systems for record keeping and to systematically deliver standard of care to patients. EHRs contain patient health information, including administrative and billing data, demographics, clinical notes, vital signs, medical histories, diagnoses, medications, immunization dates, allergies, radiology images, pathology images, laboratory values and test results. These data are either quantitative (e.g. laboratory values) or qualitative (e.g. clinical notes) and captured in structured or unstructured formats. Healthcare administrative data and claims data are collected for medical billing and administrative purposes and can be obtained from private and public health insurance providers. Data from healthcare systems and insurance claims can be an excellent source to improve knowledge as they facilitate more complete case reporting and therefore provide additional information that can enable longitudinal disease analysis[43,44].

Three main types of EHR databases exist: operational databases, clinical data repositories and clinical data warehouses. Operational databases contain EHRs that are used for patient care, whereas clinical data repositories extract, integrate and organize data from all of the

operational databases that are used by an organization[45]. Clinical data warehouses are extensive databases that are created for the long-term storage and analysis of data for the organization. The storage format for EHR and administrative data is defined by the data models used by the organizations that store the data[18]. Several commonly used data models are supported by organizations such as integrating biology in the bedside, Observational Health Data Sciences and Informatics, Sentinel and Patient-Centered Outcomes Research Network[19,20,46].

Most EHR data is accessible only to affiliates of the respective organizations where it is collected. However, the Medical Information Mart for Intensive Care database is openly accessible to researchers and widely used as a benchmarking dataset for validation of computational methods[47]. The latest release of this database, Medical Information Mart for Intensive Care-III, integrates the de-identified clinical data of patients admitted to Beth Israel Deaconess Medical Center in Boston, MA, USA. These data include vital signs, medications, laboratory measurements, observations and clinical notes, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay and survival data. Other publicly available EHR datasets include the Columbia Open Health Dataset, EHR data curated in the UK Biobank, Centers for Medicare and Medicaid Services Data Entrepreneurs' Synthetic Public Use File and Accrual to Clinical Trials Network[48–51]. Another source of EHR data that may be accessed by the community is the All of Us Research Program[52], which as of October 2020 had data available from 316,760 participants with 145 kidney-related conditions. Claims data from private stakeholders are often proprietary and available through either commercial vendors or private healthcare systems[53]. The HCUP National Inpatient Sample (NIS) is a large de-identified database of all payer inpatient care, including non-Federal hospitals in the USA.

EHR and claims data comprise patient phenotypes and treatment records and have the potential to establish new patient stratification principles and for identifying disease correlations. These data can be used to study the process of differential diagnosis and gain unique insights into patient treatment and outcomes. This type of data may greatly expand the capacity to generate new knowledge and act as a source for knowledge dissemination to inform treatment approaches for patients with multiple chronic illnesses. In addition, information on the timeliness of diagnosis and care can be used to study system-level barriers in healthcare. Electronic CKD phenotypes have been defined on the basis of EHR data for the timely identification of patients[54,55]. Going beyond passive data collection, EHRs and claims data can actively influence decision making at the point of care by presenting clinicians with information about underutilized or overutilized therapies. In addition, integration of computerized physician order entry systems with EHRs can enhance the decision support system for checking drug–disease interactions, individualized dosing for patients with AKI or CKD, and recommendations on laboratory testing during drug use[56,57]. EHRs can also be leveraged to conduct randomized clinical trials (RCTs) by identifying participants during routine clinical care and enrolling them with non-restrictive eligibility criteria or to emulate various types of trials such as comparing a single treatment with no treatment or a combinatorial intervention with no intervention or making parallel comparisons between two treatment strategies[58].

EHR-driven genomic research approaches are aimed at integrating EHR data with genetic data for improved understanding of genotype–phenotype relationships and for the diagnosis of rare diseases[59,60]. EHR data can also be added to existing surveillance data to enhance the quality and breadth of information available on equity, efficiency and clinical uncertainty[61]. Furthermore, real-time electronic health care information from medical devices may be integrated into the EHR system. For example, a 2019 study used a deep learning approach and longitudinal EHR datasets to identify factors that are associated with the risk of developing AKI and to perform continuous AKI prediction[62]. EHR data have also been used to detect the progression of diabetic kidney disease before patients develop microalbuminuria, to identify patients on dialysis who are at a high risk of death and to predict the risk of CKD in patients with diabetes[63–65].

The quality of EHR and claims data is dependent on the quality of the primary clinical data, the data linkages that are available and the derived variables that are assessed. The major limitation of these data sources is that they can only capture the information that is prompted by the system[66]. Not all health settings or systems use the same standards or policies for documenting medical care in an EHR. For example, in the USA, the majority of haemodialysis is provided by standalone centres that are not associated with hospitals or multispecialty clinics[67]. About 70% of haemodialysis care is provided by large dialysis organizations that typically have proprietary EHRs. These dialysis-focused EHRs have limited interoperability with other systems. A collaboration among federal agencies, state and local health departments, healthcare providers and EHR vendors named Digital Bridge is aimed at enabling automated flow of information through several systems[68]. The digital bridge partnership is working with eight health departments across the USA to pilot the program. Such collaboration between the subdomains of a specialty may alleviate the challenges posed by the mosaic of healthcare systems to deliver care and will lead to the availability of more comprehensive data.

Clinical trials

Clinical trials are used to inform decision-making regarding therapeutic interventions, but several ongoing data-related challenges exist, including publication bias, limited sampling, selective outcome reporting and lack of primary data release. The information that is publicly available is difficult to identify and access owing to poor indexing of regulatory documents and results portals are not easily accessible[69]. Moreover, far fewer RCTs are conducted in nephrology than in other medical subspecialities, such as oncology or cardiology[70,71]. Open-access clinical trial data provide opportunities to study the pharmacokinetics, pharmacodynamics and

Deep learning
A type of machine learning that uses multiple layers to progressively extract higher level features from the input layer of the model. Common deep learning algorithms include convolutional neural networks, recurrent neural networks, general adversarial networks and autoencoders.

safety profiles of new and older drugs. Such studies have the potential to lead to improved understanding of therapeutic options for patients with kidney diseases[72]. Data collected in clinical trials are often temporal and multidimensional and consist of medical histories, relevant vital and clinical measurements, imaging, demographics, laboratory tests, questionnaires, adverse events, concomitant medications, exposures and hospitalizations. Data from different trials can be aggregated and integrated with routinely collected data from disease-specific registries or EHRs for further analysis to advance the field using a data-driven approach. In addition, as mentioned above, routinely collected data can be used for large-scale enrolment of patients in clinical trials as well as for efficient, low-cost follow-up and recording of outcomes in the real world[73]. Thus, a vast hidden potential exists for exploring clinical trial data and discovering new insights into treatments[74].

Clinical trial data-sharing initiatives, including academic consortia and pharmaceutical companies, have pioneered transparency efforts by establishing electronic portals through which participant-level clinical trial data may be requested, shared and analysed. These data can be accessed by researchers to enable them to validate the trial findings, stimulate further inquiry and potentially lead to life-saving results[75,76]. Clinical trial data are available from several platforms[77]. For example, ClinicalTrials.gov is a registry of clinical trial protocols and a database of results. Information is presented in a tabular format with minimal narrative and limited result summaries. This resource includes information on participant flows (e.g. numbers of patients who are enrolled, those who drop out of the trial and those who are excluded from the analysis), baseline patient characteristics, outcome measures and adverse events. As of April 2021, a total of 9,257 studies relating to kidney diseases were listed on clinicaltrials.gov, of which 4,471 were completed, 1,439 were recruiting and 453 were active and not recruiting. A prediction system for kidney allograft loss, known as the iBox score, was developed using data from two clinical trials listed on clinicaltrials.gov and EHR data for kidney transplant recipients in French hospitals[78].

The EMA maintains the European Union Drug Regulating Authorities Clinical Trials Database and a subset of data is made available through the European Union Clinical Trials Register[79]. Users have access to descriptions of phase II to phase IV adult clinical trials and paediatric clinical trials. Similarly, the WHO International Clinical Trials Registry Platform makes data on clinical trials accessible through the International Clinical Trials Registry Platform search portal. Cochrane Kidney and Transplant maintains a register of studies that include RCTs in kidney diseases populated from a variety of sources[80]. As of April 2021, the register contained >28,000 reports from nearly 15,000 studies related to kidney diseases.

The Vivli clinical trial data-sharing platform is a global third-party platform that hosts clinical trial data from academia, pharmaceutical companies, non-profit organizations and other private organizations with search functionality and data request services[81]. OpenTrials is a collaborative effort between Open Knowledge International and the University of Oxford DataLab that uses an open database for all available structured data and documents on clinical trials, threaded together by individual trial[82]. The Yale University Open Data Access project serves as a model for data dissemination and independent analysis of clinical trial data[83]. The Yale University Open Data Access project was designed to promote wider access to clinical trial data, increase transparency, protect against industry influence and accelerate the generation of new knowledge.

Several public–private partnerships have also been established for clinical trials data. For example, the Duke Clinical Research Initiative has partnered with Bristol-Myers Squibb to expand access to clinical trial information from Bristol-Myers Squibb-sponsored studies. Duke Clinical Research Initiative nephrology leads several trials to define optimal approaches to patient management across the continuum of kidney diseases[84]. The information is made available through Supporting Open Access for Researchers initiatives and includes protocols, full clinical study reports, and de-identified patient-level data and study-level data for medicine and indications approved in the USA and/or Europe. Clinical trial data are also shared by the Wellcome Trust through ClinicalStudyDataRequest.com and by the Immune Tolerance Network. ClinicalStudyDataRequest.com has datasets on CKD, kidney failure and kidney transplantation, whereas the Immune Tolerance Network primarily shares clinical trial data related to autoimmune diseases, including some kidney diseases such as primary membranous nephropathy, lupus nephritis and organ transplantation.

Several government-funded programs facilitate data sharing from multi-centre large consortiums. ImmPort is one of the largest open-source repositories and distribution portals for encouraging re-use of immunological research data generated by the National Institute of Allergy and Infectious Diseases, other NIH programmes and privately funded studies[85]. This portal houses data from around 20 Clinical Trials in Organ Transplantation involving paediatric and adult kidney transplant recipients and several kidney-related studies. ImmPort also includes data on living kidney donors from clinical trials and observational studies and provides a resource to study the outcomes of these individuals with follow-up records for up to 40 years after donation[86]. Data from other trials that study kidney involvement in autoimmune diseases such as granulomatosis with polyangiitis, lupus and type 1 diabetes are also curated in ImmPort.

Mobile devices and wearable sensors

The ubiquity of mobile devices and wearable activity trackers provides an opportunity to integrate traditional medical models with data on social and behavioural determinants of health obtained directly from patients with the potential to influence disease management[87]. Wearable sensors can facilitate self-management, have a crucial role in the collection of longitudinal data and enable easy identification of non-compliance. Mobile health devices can also be deployed to collect data for

interventional strategies, including physical activity and diet monitoring, non-invasive blood pressure measurement through in-skin optical devices and routine haemoglobin monitoring during haemodialysis[88–90]. Electrocardiogram data collected through unobtrusive sensors in clothing, wristwatches, armbands and adhesive patches are especially relevant for patients on haemodialysis who are at an increased risk of arrhythmia. Sensors that use dual-wavelength LEDs and photodetectors to monitor blood oxygenation and/or body fluids such as urine could potentially be utilized during dialysis sessions to generate data to inform disease management and intervention strategies.

Mobile health applications are still in their infancy and rigorous evidence of their clinical utility is currently lacking. Nonetheless, these data are valuable and can provide insights into the ability of patients to perform routine activities without assistance, as well as the effectiveness of therapies such as KRT[91]. All of Us collects mobile health data on heart rate and blood pressure[52], but large repositories of mobile health data from patients with kidney disease are not yet publicly available. Such data could potentially contribute substantially to the generation of data-driven insights into complex scenarios such as the management of CKD and comorbidities. To date, mobile health interventions in patients with CKD or on dialysis have shown slightly favourable results, with some improvements in laboratory values, vital signs and components of overall quality of life[92,93]. However, mobile health technology is anticipated to lead to big data-based intervention strategies that could potentially result in the digital transformation of healthcare delivery.

Molecular data repositories

High-throughput sequencing and other multiplex technologies capture the molecular complexity of diseases and enable precision medicine through individualized diagnosis and treatment. Acute and CKDs are highly variable and multifactorial, with several molecular factors having a substantial role in their pathology, severity and rate of progression[94,95]. The genetic architecture and causal variants that underlie common and rare kidney diseases as well as aspects of kidney function such as glomerular filtration rate and tubular transport of electrolytes[96] are being elucidated. In addition to genetic variants, modifier genes, oligogenic modifier effects and carrier status contribute to disease phenotypes and are being studied extensively. Massive datasets have resulted from large-scale molecular studies that offer in-depth exploration of genomics, epigenomics, transcriptomics, proteomics and metabolomics using various techniques, including mRNA and microRNA (miRNA) arrays, next-generation sequencing, single-cell sequencing and mass spectrometry[97–99]. These datasets are stored either in general repositories or in nephrology-specific databases[100] (Supplementary Table 3).

General repositories. One of the oldest data resources available with a particular focus on gene–phenotype relationships is the Online Mendelian Inheritance in Man (OMIM) database, which catalogues >25,000 entries comprising genes, phenotypes and relationships between genotypic and phenotypic variations curated from peer-reviewed biomedical literature[101]. OMIM has a counterpart named Online Mendelian In Animals, which is an online database of genes, inherited disorders and traits in more than 135 animal species[102]. Several studies have used OMIM to correlate genes with kidney phenotypes such as hypoplasia, dysplasia, agenesis, kidney failure and proteinuria. One such study identified 731 unique disease entries related to kidney phenotypes corresponding to mutations in 258 genes[103]. Another study used a targeted exomic sequencing approach using panels of disease genes identified using the OMIM database to investigate the genetic basis of familial kidney diseases[104].

The Database of Genotypes and Phenotypes (dbGaP) is a NIH-sponsored archival repository that curates and disseminates raw sequencing data produced by studies investigating the interaction of genotypes and phenotypes[105]. The study summaries and related documents, as well as aggregate-level data, can be accessed freely on the dbGaP website[106]. In addition, individual-level data can be obtained by researchers who fulfil the controlled-access application requirements, state their research objectives and demonstrate their ability to adequately protect the data. As of April 2021, dbGaP curated 1,393 molecular datasets, including 18 for kidney carcinoma, 4 for IgA nephropathy and focal segmental glomerulosclerosis, 4 for inherited kidney diseases and 1 for CKD.

Publicly available databases containing data on differential expression of genes, miRNA, proteins and metabolites measured in specific diseases or in patients with multimorbidities are also available in the NIH National Center for Biotechnology Information Gene Expression Omnibus (GEO) and the European Bioinformatics Institute expression atlas[107,108]. In April 2021, the GEO database contained >114,505 items relating to the kidney, consisting of 218 datasets with 108,804 samples from 75 platforms. Differential gene expression results for about 1,000 genes that are involved in kidney diseases are included in the European Bioinformatics Institute expression atlas. Another large resource for protein expression data is the Human Protein Atlas, which consists of millions of high-resolution immunohistochemistry images that show the spatial distributions of proteins in normal human tissues, cancers and cell lines[109]. A transcriptional analysis of Human Protein Atlas data showed that 14,823 of 19,670 genes were expressed in the kidney; 413 of these genes showed elevated expression in kidney tissue and 53 were enriched in the kidney[110].

Nephrology-specific repositories. In 2006, the Human Kidney and Urine Proteome Project database was launched as an early initiative of the Human Proteome Organization. This database hosts experimental mass spectrometry data from human kidneys (glomerulus, medulla, proximal and distal tubules) and the urine proteome (normal, proteinuric, exosomes)[111]. In November 2019, the project announced its second generation with a primary focus on pathology and urine biomarkers from clinical samples.

The Urine Protein Biomarkers Database 2.0 is curated from the literature and contains data for 54 kidney diseases and 462 urinary proteins that have been associated with AKI, CKD, kidney allograft rejection, diabetic kidney disease and/or polycystic kidney disease[112]. Data are available with six major types of variables, including protein name, treatment, sample type, biomarker usage and species selection from humans, canines and rodents. This database is free to use for non-profit organizations.

Nephroseq compiles publicly available transcriptomics data from the blood and kidney tissues, such as glomeruli and tubules, of patients with 16 different conditions: arterial hypertension, CKD, nephrosclerosis, vasculitis, IgM nephropathy, IgA nephropathy, membranoproliferative glomerulonephritis, focal segmental glomerulosclerosis, kidney transplant rejection, kidney graft dysfunction, hypertension and lupus. The classic edition has 34 datasets with 2,105 combined samples and gene expression profiles for humans and mice coupled with a pre-analysis pipeline and data visualizations. The association of transcriptomics data with detailed clinical information makes this database unique. An updated version, Nephroseq v5, is awaited.

The Kidney and Urinary Pathway Knowledge Base is available as a web tool that integrates data from high-throughput omics datasets on kidney diseases and urine. This tool was developed as part of an e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science project to facilitate a rapid data mining platform in the context of kidney pathophysiology[113,114]. Initially, four main omics datasets were included in this knowledge base: the Higgins, Chabardes-Garonne, EuReGene and Vlahou–Mann datasets[115–118]. The Kidney and Urinary Pathway Knowledge Base also offers an interactive and flexible network visualization and exploration tool known as KUPNetViz[119]. The data are mapped to multiple kidney locations and diseases and result outputs for user queries show associations of entities with biological functions, processes, cellular components, biochemical pathways and miRNA targets.

The Chronic Kidney Disease Database is a component of the iMODE-CKD consortium, which aims to use clinical and system omics information to build a global molecular resource that will aid identification of molecular determinants of CKD by studying disease pathways. The Chronic Kidney Disease Database includes manually curated data from published multi-omics studies involving microRNA, genomics, peptidomics, proteomics and metabolomics of CKD in rat, mice and humans[120]. Data are available for >49,000 redundant molecules and >16,800 unique molecules in blood, urine and kidney tissue. Demographics and clinical data are coupled with omics data and include information on age, sex, body mass index, clinical history, blood pressure, creatinine, urea, albumin and estimated glomerular filtration rate.

The Kidney Gene Database (KGDB) catalogues information about genes implicated in human kidney diseases from the biomedical literature and gene expression databases[121]. The KGDB is part of a larger, web-based human urological gene database and provides information on gene amplification, mutations, deletions, polymorphisms, loss of heterozygosity, DNA methylation and DNA hypomethylation. For each gene, the KGDB provides information about the gene product, a tissue type gene expression profile, links to protein, mRNA and genomic DNA sequence information, relevant literature citations and cross-references to other databases.

The Renal Gene Expression database contains gene expression profiles generated by high-throughput DNA expression profiling studies of kidney disease that were identified using the GEO database[122]. The Renal Gene Expression Database has an integrated web-based platform to make data easily accessible to the kidney disease research community. Data from 88 studies focusing on 10 kidney diseases (AKI, autosomal dominant polycystic kidney disease, CKD, IgA nephropathy, diabetic kidney disease, kidney carcinoma, kidney transplantation, lupus nephritis and nephrosclerosis) are included in the database. As well as providing gene expression profiles, this database enables users to look for genes that have similar expression patterns to the gene of interest, which might suggest a common regulatory pathway.

The Autosomal Dominant Polycystic Kidney Disease Mutation Database collects data on variants of the autosomal dominant polycystic kidney disease genes, *PKD1* and *PKD2*. Data are available on mutation designation, complementary DNA change, amino acid change, mutation type and clinical significance. Other nephrology-specific molecular databases include neph QTL[123] and human kidney eQTL[124], which provide information on expression quantitative trait loci of kidney tissues

The promise and challenges of big data

Integrating the big data modalities that are available in nephrology provides an opportunity to take a comprehensive approach to personalized medicine. Big data have the potential to enable data-driven clinical decision making for patients with kidney disease. In addition, predictive models based on big data could be utilized in the nephrology clinic. The current implementation of machine-learning algorithms has barely scratched the surface of the potential applications of big data in nephrology[125] (BOX 1). In research and development, such applications could help to identify novel therapies and molecular biomarkers, decipher the causes of monogenic diseases, elucidate drivers of disease onset and progression and enable precision disease stratification and characterization. New computational approaches could also aid understanding of complex diseases by dissecting the underlying genetic, epigenetic and post-translational modifications. In the clinic, big data could contribute to artificial intelligence-aided early diagnosis and prognosis, as well as personalized selection of therapy based on the underlying causes of disease. Big data could also be used in clinical decision-support systems, to investigate drug efficacy in the real world and for risk stratification of patients with CKD or kidney failure, including those on KRT. Applications of big data in population health include studies of the comparative effectiveness of short-term and long-term interventions in patients

Box 1 | **Artificial intelligence and machine learning in healthcare**

Machine learning is a subset of artificial intelligence that encompasses algorithmic methods that can identify features and learn from data without being explicitly programmed. Machine-learning algorithms are essentially a set of mathematical procedures that explore linear and non-linear relationships between different variables and offer flexibility and scalability compared with traditional analytical approaches. Predictions made by these algorithms can be discreet or continuous depending on the application and require data pre-processing, model training and validation[142].

In healthcare, machine-learning applications can yield insights to assist in diagnosis, risk assessment, triage and treatment decisions. Major applications for artificial intelligence in healthcare either utilize structured data (such as drug prescriptions, diagnoses, comorbidities, genetic findings, laboratory values and biomarkers) or unstructured data (such as images, videos, notes, medical journals and surveys).

In general, machine learning can be divided into two categories: supervised learning and unsupervised learning. In supervised learning, prediction rules are acquired from labelled training data and then used to predict unseen data labels. Supervised learning can be used for classification and regression. In unsupervised learning, no labels are present and the goal is to find the structure of unknown input data by searching for similarities between records. Principal component and cluster analyses are examples of unsupervised learning.

Machine learning has been used in a number of studies in patients with kidney diseases[143]. For example, machine-learning approaches have been used to predict kidney function outcomes in patients with progressive IgA nephropathy[144] and in-hospital mortality of patients with acute kidney injury in intensive care units[145]. They have also been used to predict protein catabolic rate[146], optimal dialysis session time[147] and optimal erythropoietin dose for anaemia management[148] in patients on haemodialysis.

with kidney disease or kidney failure, which could inform health policy and improve quality of care as well as help to optimize resource allocation. Mobile health applications could enable continuous remote monitoring of patients and provide data to inform their management and/or aid diagnosis.

Combined efforts from researchers, clinicians and data scientists, as well as engagement from multiple stakeholders, including healthcare organizations, government and pharmaceutical and biotech industries, are required to realize the potential of big data in nephrology. However, several challenges limit implementation of data science in nephrology clinical care and research[126]. These challenges relate to data volume, ownership, privacy, heterogeneity, incompleteness and analysis approaches. Current data governance policies and frameworks are often designed by parent organizations to address these challenges and legal aspects such as data protection laws[127]. However, such policies should be transparent, should be updated in response to changing user requirements and should evolve as new technologies and research methodologies become available, while providing standardized high-quality data.

Another major challenge for data governance is to institute equitable proprietary rights among various stakeholders, including governments, care providers, patients, insurance companies, pharmaceutical companies, biotechnology companies and software vendors. The jurisdictions of these stakeholders pose further hindrances for easy data accessibility to users as they navigate through complex national and international approval protocols.

The current fragmented system with segregated data housed in solitary unrelated databases is the major barrier to combining different clinical measurements on a unified platform. This fragmentation diminishes the power of large clinical datasets and the optimization of downstream analytics. The lack of semantic integration of data elements across disparate data sources also restricts collaborations between interested groups. As a result, siloed data sets often become obsolete and unusable. This problem strongly suggests that breaking down silos could help to secure data communications and improve the efficiency of the systems[128]. Although much of the existing data in nephrology is siloed, great potential exists for transformation upon integration and expansion of these data[129]. For example, linking of molecular data to EHRs would provide information on the molecular phenotypes of kidney diseases, potentially enable targeted disease surveillance and personalized therapy and inform family counselling.

Importantly, a lack of consistent funding for nephrology research limits the generation of new data. This problem is illustrated by the fact that many nephrology-specific databases contain data that are obsolete or not frequently updated (Supplementary Table 3). Increased output of high-quality research in nephrology will be required to move the field towards data-driven approaches.

Additional challenges for the incorporation of big data into healthcare include data inconsistencies, missing elements, ambiguities and erroneous data collection practices[130]. Data heterogeneity results from measuring qualitative and quantitative variables using different scales and from disparities between different supply chain vendors that provide testing reagents and equipment, technologies and acquisition parameters. Substantial intravariability and intervariability also exists between patient phenotypes. In particular, patients with kidney disease often have comorbidities that make nephrology cohorts very heterogeneous. Another issue that is distinct from data quality is the comprehensiveness and completeness of the available data. Many clinical data sources are limited owing to a lack of information about patient prognosis, complete elements of care, outcomes including performance status, disease stages, the intent of treatment or disease burden, which is particularly relevant for comparative effectiveness studies. In nephrology, data quality, collection and curation need to be improved to incorporate more granularity, such as recording KRT submodality and reporting comprehensive outcome data and causal factors for kidney failure in patient registries and EHRs. Factors that influence the quality of primary data include how the data were collated, the skills, training and oversight of the individuals who collected the data, and whether external influences such as budget and infrastructure affected accurate data capture.

For big data projects, a key component that can potentially introduce bias into training machine learning tools is sample size[131]. In nephrology, data from ethnically diverse cohorts and from paediatric patients with genetic disorders are very limited. Predictions based on small sample sizes are unrepresentative and data on rare kidney diseases need to be enhanced in size to provide sufficient information for improved elucidation of genotype and phenotype relationships.

k-anonymity
A data anonymization technique that protects the identities of individuals using methods such as suppression and generalization. A dataset is said to have k-anonymity if the information for each individual cannot be distinguished from that of at least k-1 individuals.

l-diversity
A data anonymization approach that relies on introducing further entropy or diversity to the dataset. This model uses generalization and promotes diversity for sensitive values within a group. l-diversity is an extension of the k-anonymity model.

t-closeness
This model is a further refinement of the k-anonymity and l-diversity models. t-closeness is the maximum of the distances between the distribution of values of a sensitive attribute and that of the entire database table. An equivalence class will have t-closeness if the distance between the attribute in the class and whole table is no more than threshold t.

The kidney community must mobilize to perform more multicentre collaborative studies and to collect more data on metrics for monitoring diseases such as AKI and CKD[132]. For example, more data are needed on patient-reported outcome measures, disease prevention, progression and management, KRT, medications, vaccinations and lifestyle factors. Greater granularity of data on these metrics could potentially lead to more reliable machine-learning models and the development of personalized evidence-based interventions. Predictive algorithms should not only recapitulate historical trends but also accurately predict future events. As the practice of medicine is constantly evolving in response to new technology, epidemiology and social phenomena, machine learning algorithms that are developed to predict patient outcomes or classify patients based on disease phenotype are essentially chasing a moving target[133,134].

Another major challenge is developing the infrastructure that is required to handle extremely high data volumes, which require incrementally scalable storage platforms. A need exists for solutions such as cloud computing that will ensure storage capacity, server management, network power and bandwidth utilities. The need for secure and centralized management of data cannot be overemphasized. It is equally imperative to keep pace with the velocity of the data to process the data stream accurately and efficiently. Most importantly, good research practice guidelines should be updated to consider how to protect patient data while supporting use of big data in research.

The rights and privacy of research participants and patients must be protected and respected by implementing data protection regulations when performing data-driven research and using big data applications in healthcare[135]. Appropriate de-identification of data and security standards should guard against data theft, manipulation and unauthorized access. Advanced privacy metrics such as k-anonymity, l-diversity and t-closeness are being introduced to enhance data protection[136]. These advances lead to additional challenges of striking the right balance between implementing new privacy measures and preventing information loss from data.

Finally, a major challenge in implementing data-driven clinical decision-support systems is a reluctance among members of the medical community to accept this paradigm shift in healthcare[137]. This reluctance might be due to a lack of confidence in the accuracy of artificial intelligence or machine-learning tools to aid decision making, as well as concerns that such tools cannot outperform medical professionals in real-world scenarios in which healthcare inequities exist. Obtaining approval for such systems from regulatory bodies is also challenging. A need exists to optimize audit analysis so that automated frameworks are more transparent in order to gain trust from the community.

Conclusions

An enormous amount of data is available that could potentially be utilized to improve kidney care and research. Big datasets could enable a shift from experience-based medicine to evidence-based medicine in nephrology, but proper analysis pipelines are required. To have a true impact, big data require integrative, multidisciplinary approaches that combine experimental, clinical and computational skills across multiple institutions, while respecting the needs and values of patients and research participants.

Published online: 30 June 2021

1. Erickson, K. F., Qureshi, S. & Winkelmayer, W. C. The role of big data in the development and evaluation of US dialysis care. *Am. J. Kidney Dis.* **72**, 560–568 (2018).
2. Adimadhyam, S. et al. Leveraging the capabilities of the FDA's sentinel system to improve kidney care. *J. Am. Soc. Nephrol.* **31**, 2506–2516 (2020).
3. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
4. Escobar, G. J. et al. Automated identification of adults at risk for in-hospital clinical deterioration. *N. Engl. J. Med.* **383**, 1951–1960 (2020).
5. Hulsen, T. et al. From big data to precision medicine. *Front. Med.* **6**, 34 (2019).
6. Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S. & Rubin, D. L. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digit. Med.* **2**, 78 (2019).
7. Liu, F. X., Rutherford, P., Smoyer-Tomic, K., Prichard, S. & Laplante, S. A global overview of renal registries: A systematic review Epidemiology and Health Outcomes. *BMC Nephrol.* **16**, 1–10 (2015).
8. Bikbov, B. 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* **395**, 709–733 (2020).
9. Friedman, D. J., Parrish, R. G. & Ross, D. A. Electronic health records and US public health: Current realities and future promise. *Am. J. Public Health* **103**, 1560–1567 (2013).
10. Murphy, D. et al. Trends in prevalence of chronic kidney disease in the United States. *Ann. Intern. Med.* **165**, 473–481 (2016).
11. Chan, L. et al. The effect of depression in chronic hemodialysis patients on inpatient hospitalization outcomes. *Blood Purif.* **43**, 226–234 (2017).
12. Cheung, A. et al. Impact of atrial fibrillation in patients with chronic kidney disease undergoing transcatheter aortic valve replacement: Insights of the Healthcare Cost and Utilization Project's National Inpatient Sample. *Cardiovasc. Revasc. Med.* **19**, 21–25 (2018).
13. Matsushita, K. et al. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. *Lancet* **375**, 2073–2081 (2010).
14. Zoccali, C., Brancaccio, D. & Nathan, M. J. Causality at the dawn of the 'omics' era in medicine and in nephrology. *Nephrol. Dial. Transplant.* **31**, 1381–1385 (2016).
15. Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the missing link for big biomedical data. *J. Am. Med. Assoc.* **311**, 2479–2480 (2014).
16. Nadkarni, G. N., Coca, S. G. & Wyatt, C. M. Big data in nephrology: promises and pitfalls. *Kidney Int.* **90**, 240–241 (2016).
17. Pezoulas, V. C. et al. Medical data quality assessment: on the development of an automated framework for medical data curation. *Comput. Biol. Med.* **107**, 270–283 (2019).
18. Danese, M. D., Halperin, M., Duryea, J. & Duryea, R. The generalized data model for clinical research. *BMC Med. Inform. Decis. Mak.* **19**, 1–13 (2019).
19. Fleurence, R. L. et al. Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* **21**, 578–582 (2014).
20. Murphy, S. N. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).
21. Klann, J. G., Joss, M. A. H., Embree, K. & Murphy, S. N. Data model harmonization for the all of us research program: transforming i2b2 data into the OMOP common data model. *PLoS One* **14**, 1–13 (2019).
22. Kush, R. D. et al. FAIR data sharing: the roles of common data elements and harmonization. *J. Biomed. Inform.* **107**, 103421 (2020).
23. Wilkinson, M. D. et al. Comment: the FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
24. Kubben, P., Dumontier, M. & Dekker, A. *Fundamentals of clinical data science*. (Springer, 2019).
25. Dreyer, N. A. & Garner, S. Registries for robust evidence. *JAMA* **302**, 790–791 (2009).
26. Jager, K. J. & Wanner, C. Fifty years of ERA-EDTA registry — a registry in transition. *Kidney Int. Suppl.* **5**, 12–14 (2015).
27. Choi, N. G., Sullivan, J. E., DiNitto, D. M. & Kunik, M. E. Health care utilization among adults with CKD and psychological distress. *Kidney Med.* **1**, 162–170 (2019).
28. Robinson, B. M., Bieber, B., Pisoni, R. L. & Port, F. K. Dialysis outcomes and practice patterns study (DOPPS): Its strengths, limitations, and role in informing practices and policies. *Clin. J. Am. Soc. Nephrol.* **7**, 1897–1905 (2012).
29. DOPPS. DPM sampling, study design, and calculation methods. *DOPPS* https://www.dopps.org/DPM/Data_Sources_Methods.pdf (2020).
30. Dienemann, T. et al. International Network of Chronic Kidney Disease cohort studies (iNET-CKD): a global network of chronic kidney disease cohorts. *BMC Nephrol.* **17**, 1–9 (2016).
31. Saran, R. et al. US Renal Data System 2019 Annual Data Report: epidemiology of kidney disease in the United States. *Am. J. Kidney Dis.* **75**, A6–A7 (2020).
32. Go, A. S., Chertow, G. M., Fan, D., McCulloch, C. E. & Hsu, C. Y. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *N. Engl. J. Med.* **351**, 1296–1305 (2004).
33. Saran, R. et al. US Renal Data System 2014 Annual Data Report: epidemiology of kidney disease in the United States. *Am. J. Kidney Dis.* **66**, A7 (2015).

34. Mendu, M. L. et al. Development of an electronic health record-based chronic kidney disease registry to promote population health management. *BMC Nephrol.* **20**, 1–11 (2019).
35. Norris, K. C. et al. Rationale and design of a multicenter Chronic Kidney Disease (CKD) and at-risk for CKD electronic health records-based registry: CURE-CKD. *BMC Nephrol.* **20**, 1–9 (2019).
36. Navaneethan, S. D. et al. Development and validation of an electronic health record-based chronic kidney disease registry. *Clin. J. Am. Soc. Nephrol.* **6**, 40–49 (2011).
37. Evans, K. et al. UK renal registry 20th annual report: introduction. *Nephron* **139**, 1–11 (2018).
38. Pyart, R. et al. The 21st UK renal registry annual report: a summary of analyses of adult data in 2017. *Nephron* **144**, 59–66 (2020).
39. Kramer, A. et al. The European Renal Association — European Dialysis and Transplant Association (ERA-EDTA) Registry Annual Report 2016: a summary. *Clin. Kidney J.* **12**, 702–720 (2019).
40. McDonald, S. P. Australia and New Zealand dialysis and transplant registry. *Kidney Int. Suppl.* **5**, 39–44 (2015).
41. Global Health Data Exchange. http://www.healthdata.org/about/ghdx (2020).
42. Rare Kidney Stone Consortium. http://www.rarekidneystones.org/ (2015).
43. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA* **309**, 1351–1352 (2013).
44. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
45. McCartney, P. R. Clinical databases: electronic health records and repositories. *MCN Am. J. Matern. Nurs.* **38**, 186 (2013).
46. Hripcsak, G. et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
47. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
48. Ta, C. N., Dumontier, M., Hripcsak, G., Tatonetti, N. P. & Weng, C. Columbia open health data, clinical concept prevalence and co-occurrence from electronic health records. *Sci. Data* **5**, 1–17 (2018).
49. Centers for Medicare & Medicaid Services. *CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)* (2019).
50. UK Biobank. *Integrating Electronic Health Records into the UK Biobank Resource*. http://biobank.ctsu.ox.ac.uk/showcase/showcase/docs/DataLinkageProcess.pdf (2014).
51. Visweswaran, S. et al. Accrual to clinical trials (ACT): a clinical and translational science award consortium network. *JAMIA Open* **1**, 147–152 (2018).
52. The All of Us Research Program Investigators. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
53. Cadarette, S. M. & Wong, L. An introduction to health care administrative data. *Can. J. Hosp. Pharm.* **68**, 232–237 (2015).
54. Nadkarni, G. N. et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu. Symp. Proc.* **2014**, 907–916 (2014).
55. Norton, J. M. et al. Development and validation of a pragmatic electronic phenotype for CKD. *Clin. J. Am. Soc. Nephrol.* **14**, 1306–1314 (2019).
56. Wilkerson, M. L., Henricks, W. H., Castellani, W. J., Whitsitt, M. S. & Sinard, J. H. Management of laboratory data and information exchange in the electronic health record. *Arch. Pathol. Lab. Med.* **139**, 319–327 (2015).
57. Mills, S. Electronic health records and use of clinical decision support. *Crit. Care Nurs. Clin. North. Am.* **31**, 125–131 (2019).
58. Abdel-Kader, K. & Jhamb, M. EHR-based clinical trials: the next generation of evidence. *Clin. J. Am. Soc. Nephrol.* **15**, 1050–1052 (2020).
59. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011).
60. Garcelon, N., Burgun, A., Salomon, R. & Neuraz, A. Electronic health records for the diagnosis of rare diseases. *Kidney Int.* **97**, 676–686 (2020).
61. Matsushita, K. et al. Cohort profile: the chronic kidney disease prognosis consortium. *Int. J. Epidemiol.* **42**, 1660–1668 (2013).
62. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
63. Makino, M. et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci. Rep.* **9**, 11862 (2019).
64. Akbilgic, O. et al. Machine learning to identify dialysis patients at high death risk. *Kidney Int. Rep.* **4**, 1219–1229 (2019).
65. Ravizza, S. et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat. Med.* **25**, 57–59 (2019).
66. Pivovarov, R., Albers, D. J., Sepulveda, J. L. & Elhadad, N. Identifying and mitigating biases in EHR laboratory tests. *J. Biomed. Inform.* **51**, 24–34 (2014).
67. Sutton, P. R. & Payne, T. H. Interoperability of electronic health information and care of dialysis patients in the United States. *Clin. J. Am. Soc. Nephrol.* **14**, 1536–1538 (2019).
68. Centers for Disease Control and Prevention. *Surveillance Strategy Report — How Sharing Data Digitally Benefits Health*. https://www.cdc.gov/surveillance/innovation/sharing-data-digitally.html (2018).
69. Krumholz, H. M. & Peterson, E. D. Open access to clinical trials data. *JAMA* **312**, 1002–1003 (2014).
70. Baigent, C. et al. Challenges in conducting clinical trials in nephrology: conclusions from a Kidney Disease — Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney Int.* **92**, 297–305 (2017).
71. Kitchlu, A. et al. Representation of patients with chronic kidney disease in trials of cancer therapy. *JAMA* **319**, 2437–2439 (2018).
72. Panchapakesan, U. & Pollock, C. Drug repurposing in kidney disease. *Kidney Int.* **94**, 40–48 (2018).
73. Herrington, W. G., Staplin, N. & Haynes, R. Kidney disease trials for the 21st century: innovations in design and conduct. *Nat. Rev. Nephrol.* **16**, 173–185 (2020).
74. Sim, I. et al. Time for NIH to lead on data sharing. *Science* **367**, 1308–1309 (2020).
75. Kiley, R., Peatfield, T., Hansen, J. & Reddington, F. Data sharing from clinical trials — a research funder's perspective. *N. Engl. J. Med.* **377**, 1990–1992 (2017).
76. Mc Cord, K. A. et al. Routinely collected data for randomized trials: promises, barriers, and implications. *Trials* **19**, 29 (2018).
77. Shlipak, M. & Stehman-Breen, C. Observational research databases in renal disease. *J Am. Soc. Nephrol.* **16**, 3477–3484 (2005).
78. Loupy, A. et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: International derivation and validation study. *BMJ* **366**, l4923 (2019).
79. Egger, G. F. et al. European Union Clinical Trials Register: on the way to more transparency of clinical trial data. *Expert Rev. Clin. Pharmacol.* **6**, 457–459 (2013).
80. Cochrane Kidney and Transplant. https://kidneyandtransplant.cochrane.org/ (2021).
81. Bierer, B. E., Li, R., Barnes, M. & Sim, I. A global, neutral platform for sharing trial data. *N. Engl. J. Med.* **374**, 2411–2413 (2016).
82. Goldacre, B. & Gray, J. Opentrials: towards a collaborative open database of all available information on all clinical trials. *Trials* **17**, 164 (2018).
83. Ross, J. S. et al. Overview and experience of the YODA project with clinical trial data sharing after 5 years. *Sci. Data* **5**, 1–14 (2018).
84. Pencina, M. J. et al. Supporting open access to clinical trial data for researchers: the Duke Clinical Research Institute-Bristol-Myers Squibb supporting open access to researchers initiative. *Am. Heart J.* **172**, 64–69 (2016).
85. Bhattacharya, S. et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5**, 1–9 (2018).
86. Chen, J. et al. Assessment of postdonation outcomes in US living kidney donors using publicly available data sets. *JAMA Netw. Open* **2**, e191851 (2019).
87. Sim, I. Mobile devices and health. *N. Engl. J. Med.* **381**, 956–968 (2019).
88. Sieverdes, J. C. Mobile health considerations for kidney disease and transplantation. *mHealth* **4**, 13–13 (2018).
89. Lambert, K., Mullan, J., Mansfield, K. & Owen, P. Should we recommend renal diet–related apps to our patients? An evaluation of the quality and health literacy demand of renal diet–related mobile applications. *J. Ren. Nutr.* **27**, 430–438 (2017).
90. Streeper, N. M., Lehman, K. & Conroy, D. E. Acceptability of mobile health technology for promoting fluid consumption in patients with nephrolithiasis. *Urology* **122**, 64–69 (2018).
91. Lunde, P., Nilsson, B. B., Bergland, A., Kværner, K. J. & Bye, A. The effectiveness of smartphone apps for lifestyle improvement in noncommunicable diseases: systematic review and meta-analyses. *J. Med. Internet Res.* **20**, 1–12 (2018).
92. Singh, K. et al. Patients' and nephrologists' evaluation of patient-facing smartphone apps for CKD. *Clin. J. Am. Soc. Nephrol.* **14**, 523–529 (2019).
93. Yang, Y., Chen, H., Qazi, H. & Morita, P. P. Intervention and evaluation of mobile health technologies in management of patients undergoing chronic dialysis: scoping review. *JMIR mHealth Uhealth* **8**, e15549 (2020).
94. Pejchinovski, M. & Mischak, H. Clinical proteomics in kidney disease: from discovery to clinical application. *Prilozi* **38**, 39–54 (2018).
95. Bullich, G. et al. A kidney-disease gene panel allows a comprehensive genetic diagnosis of cystic and glomerular inherited kidney diseases. *Kidney Int.* **94**, 363–371 (2018).
96. Groopman, E. E., Rasouly, H. M. & Gharavi, A. G. Genomic medicine for kidney disease. *Nat. Rev. Nephrol.* **14**, 83–104 (2018).
97. Groopman, E. E. et al. Diagnostic utility of exome sequencing for kidney disease. *N. Engl. J. Med.* **380**, 142–151 (2019).
98. Weiss, R. H. Metabolomics and metabolic reprogramming in kidney cancer. *Semin. Nephrol.* **38**, 175–182 (2018).
99. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 1–15 (2017).
100. Papadopoulos, T. et al. Omics databases on kidney disease: where they can be found and how to benefit from them. *Clin. Kidney J.* **9**, 343–352 (2016).
101. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, 514–517 (2005).
102. Lenffer, J. OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.* **34**, D599–D601 (2006).
103. Parsa, A. et al. Common variants in mendelian kidney disease genes and their association with renal function. *J. Am. Soc. Nephrol.* **24**, 2105–2117 (2013).
104. Mallett, A. J. et al. Massively parallel sequencing and targeted exomes in familial kidney disease can diagnose underlying genetic disorders. *Kidney Int.* **92**, 1493–1506 (2017).
105. Tryka, K. A. et al. NCBI's database of genotypes and phenotypes: DbGaP. *Nucleic Acids Res.* **42**, 975–979 (2014).
106. Wong, K. M. et al. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res.* **45**, D819–D826 (2017).
107. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets — update. *Nucleic Acids Res.* **41**, 991–995 (2013).
108. Papatheodorou, I. et al. Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* **46**, D246–D251 (2018).
109. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
110. Thul, P. J. & Lindskog, C. The human protein atlas: a spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).
111. Yamamoto, T., Langham, R. G., Ronco, P., Knepper, M. A. & Thongboonkerd, V. Towards standard protocols and guidelines for urine proteomics: a report on the Human Kidney and Urine Proteome Project (HKUPP) symposium and workshop — 6 October 2007, Seoul, Korea and 1 November 2007, San Francisco, CA, USA. *Proteomics* **8**, 2156–2159 (2008).
112. Shao, C. et al. A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database. *Mol. Cell. Proteom.* **10**, 1–8 (2011).
113. e-LICO *An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science*. http://www.e-lico.eu/ (2019).
114. Jupp, S., Klein, J., Schanstra, J. & Stevens, R. Developing a kidney and urinary pathway knowledge base. *J. Biomed. Semant.* **2**, S7 (2011).

115. Helfand, B. T., Mendez, M. G., Pugh, J., Delsert, C. & Goldman, R. D. Maintaining the shape of nerve cells. *Mol. Biol. Cell* **14**, 5069–5081 (2003).
116. Chabardès-Garonne, D. et al. A panoramic view of gene expression in the human kidney. *Proc. Natl Acad. Sci. USA* **100**, 13710–13715 (2003).
117. Willnow, T. E. et al. The European renal genome project. *Organogenesis* **2**, 42–47 (2005).
118. Mischak, H. et al. Comprehensive human urine standards for comparability and standardization in clinical proteome analysis. *Proteom. Clin. Appl.* **4**, 464–478 (2010).
119. Moulos, P. et al. The KUPNetViz: a biological network viewer for multiple -omics datasets in kidney diseases. *BMC Bioinformatics* **14**, 235 (2013).
120. Fernandes, M. & Husi, H. Establishment of a integrative multi-omics expression database CKDdb in the context of chronic kidney disease (CKD). *Sci. Rep.* **7**, 1–11 (2017).
121. Zhao, H. et al. Kidney gene database: a curated and integrated database of genes involved in kidney disease. *J. Urol.* **172**, 2344–2346 (2004).
122. Zhang, Q. et al. Renal Gene Expression Database (RGED): a relational database of gene expression profiles in kidney disease. *Database* **2014**, 1–6 (2014).
123. Gillies, C. E. et al. An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* **103**, 232–244 (2018).
124. Qiu, C. et al. Renal compartment–specific genetic variation analyses identify new pathways in chronic kidney disease. *Nat. Med.* **24**, 1721–1731 (2018).
125. Ketchersid, T. Big data in nephrology: friend or foe? *Blood Purif.* **36**, 160–164 (2014).
126. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).
127. Kaye, J. et al. Including all voices in international datasharing governance. *Hum. Genomics* **12**, 18–23 (2018).
128. Reinholz, D. L. & Andrews, T. C. Breaking down silos working meeting: an approach to fostering cross-disciplinary STEM–DBER collaborations through working meetings. *CBE Life Sci. Educ.* **18**, 1–8 (2019).
129. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Futur. Healthc. J.* **6**, 94–102 (2019).
130. Kruse, C. S., Goswamy, R., Raval, Y. & Marawi, S. Challenges and opportunities of big data in health care: a systematic review. *JMIR Med. Inform.* **4**, e38 (2016).
131. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **20**, e262–e273 (2019).
132. Floege, J., Mak, R. H., Molitoris, B. A., Remuzzi, G. & Ronco, P. Nephrology research — the past, present and future. *Nat. Rev. Nephrol.* **11**, 677–687 (2015).
133. Pépin, J. L., Bailly, S. & Tamisier, R. Big data in sleep apnoea: opportunities and challenges. *Respirology* **25**, 486–494 (2019).
134. Adibuzzaman, M., DeLaurentis, P., Hill, J. & Benneyworth, B. D. Big data in healthcare — the promises, challenges and opportunities from a research perspective: a case study with a model database. *AMIA Annu. Symp. Proc.* **2017**, 384–392 (2017).
135. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
136. Jeon, S. et al. Proposal and assessment of a de-identification strategy to enhance anonymity of the observational medical outcomes partnership common data model (OMOP-CDM) in a public cloud-computing environment: anonymization of medical data using privacy models. *J. Med. Internet Res.* **22**, e19597 (2020).
137. Meskó, B. & Görög, M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit. Med.* **3**, 126 (2020).
138. Jha, A. K. et al. How common are electronic health records in the United States? A summary of the evidence. *Health Aff.* **25**, 496–507 (2006).
139. Brennan, S. The biggest computer programme in the world ever! How's it going? *J. Inf. Technol.* **22**, 202–211 (2007).
140. Lee Ventola, C. Mobile devices and apps for health care professionals: uses and benefits. *P T* **39**, 356–364 (2014).
141. Liu, C., Zhu, Q., Holroyd, K. A. & Seng, E. K. Status and trends of mobile-health applications for iOS devices: a developer's perspective. *J. Syst. Softw.* **84**, 2022–2033 (2011).
142. Sidey-Gibbons, J. A. M. & Sidey-Gibbons, C. J. Machine learning in medicine: a practical introduction. *BMC Med. Res. Methodol.* **19**, 1–18 (2019).
143. Niel, O. & Bastard, P. Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. *Am. J. Kidney Dis.* **74**, 803–810 (2019).
144. Geddes, C. C., Fox, J. G., Allison, M. E. M., Boulton-Jones, J. M. & Simpson, K. An artificial neural network can select patients at high risk of developing progressive IgA nephropathy more accurately than experienced nephrologists. *Nephrol. Dial. Transplant.* **13**, 67–71 (1998).
145. Lin, K., Hu, Y. & Kong, G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *Int. J. Med. Inform.* **125**, 55–61 (2019).
146. Gabutti, L. et al. Usefulness of artificial neural networks to predict follow-up dietary protein intake in hemodialysis patients. *Kidney Int.* **66**, 399–407 (2004).
147. Akl, A. I., Sobh, M. A., Enab, Y. M. & James, T. Artificial intelligence: a new approach for prescription and monitoring of hemodialysis therapy. *Am. J. Kidney Dis.* **38**, 1277–1283 (2001).
148. Barbieri, C. et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int.* **90**, 422–442 (2016).

Acknowledgements
The authors would like to acknowledge Flavio Vincenti, Sri Lekha Tummalapalli, Vivek Rudrapatna, Douglas Arneson and Zicheng Hu (all University of California, San Francisco) for their valuable suggestions for this manuscript. The authors' work was supported by the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN316201200036W), the UCSF Bakar Computational Health Sciences Institute and the UCSF Clinical and Translational Sciences Institute, supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1 TR001872. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions
N.K. researched data for the article and wrote the manuscript, S.B. contributed to data research and edited the manuscript. A.J.B. conceived the manuscript framework and reviewed and edited it before submission.

Competing interests
A.T.B. is a co-founder and consultant to Personalis and NuMedii; consultant to Samsung, Mango Tree Corporation, and in the recent past, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, CVS, Nuna Health, Assay Depot, Vet24seven, Regeneron, Sanofi, Royalty Pharma, AstraZeneca, Moderna, Biogen, Paraxel and Sutro, and several other non-health-related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease-specific foundations and associations, and health systems. A.T.B. receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. His research has been funded by NIH, Northrup Grumman (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity.

Peer review information
Nature Reviews Nephrology thanks Luxia Zhang, who co-reviewed with Chao Yang, William Herrington, Min Jun and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s41581-021-00439-x.

RELATED LINKS
Autosomal dominant polycystic kidney disease mutation database: https://pkdb.mayo.edu/
ClinicalStudyDataRequest.com: https://clinicalstudydatarequest.com/Default.aspx
Clinicaltrials.gov: https://clinicaltrials.gov/
Immune Tolerance Network: https://www.immunetolerance.org/
National Kidney Foundation Patient Network: https://www.kidney.org/nkfpatientnetwork
Nephroseq: https://www.nephroseq.org/resource/login.html
PatientsLikeMe: https://www.patientslikeme.com/
Registry Of Kidney Diseases: https://gardn.org.au/registries/registry-of-kidney-diseases/
RenDER data extraction and referencing system: https://render.usrds.org/render/xrender.phtml
Sentinel and Patient-Centered Outcomes Research Network: https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model
The HCUP National Inpatient Sample (NIS): https://www.hcup-us.ahrq.gov/nisoverview.jsp
The NephCure Kidney Network Patient Registry: https://nephcure.org/2015/12/the-nephcure-kidney-network-patient-registry-nkn/
Think Kidneys: https://www.thinkkidneys.nhs.uk/
WHO International Clinical Trials Registry Platform: https://www.who.int/ictrp/en/

© Springer Nature Limited 2021