

# Leveraging Data Science for a Personalized Haemodialysis

Miguel Hueso<sup>a</sup> Lluís de Haro<sup>b</sup> Jordi Calabia<sup>c</sup> Rafael Dal-Ré<sup>d</sup> Cristian Tebé<sup>e</sup>  
Karina Gibert<sup>f</sup> Josep M. Cruzado<sup>a</sup> Alfredo Vellido<sup>f</sup>

<sup>a</sup>Department of Nephrology, Hospital Universitari Bellvitge, and Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Spain; <sup>b</sup>Functional Competence Center, Information Systems, Institut Català de la Salut, Barcelona, Spain; <sup>c</sup>Department of Nephrology, Hospital Universitari Dr. Josep Trueta, Girona, Spain; <sup>d</sup>Health Research Institute, Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid, Madrid, Spain; <sup>e</sup>Biostatistics Unit, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Spain; <sup>f</sup>Intelligent Data Science and Artificial Intelligence (IDEAI) Research Center, Universitat Politècnica de Catalunya (UPC BarcelonaTech), Barcelona, Spain

## Keywords

Data science · Haemodialysis · Personalized medicine · Artificial intelligence · Machine learning · Pragmatic clinical trials

## Abstract

**Background:** The 2019 Science for Dialysis Meeting at Bellvitge University Hospital was devoted to the challenges and opportunities posed by the use of data science to facilitate precision and personalized medicine in nephrology, and to describe new approaches and technologies. The meeting included separate sections for issues in data collection and data analysis. As part of data collection, we presented the institutional ARGOS e-health project, which provides a common model for the standardization of clinical practice. We also pay specific attention to the way in which randomized controlled trials offer data that may be critical to decision-making in the real world. The opportunities of open source software (OSS) for data science in clinical practice were also discussed. **Summary:** Precision medicine aims to provide the right treatment for the right patients at the right time and is deeply connected to data science. Dialysis patients are highly dependent on technology to live, and their treatment generates a huge volume of data that has to be analysed. Data science has emerged

as a tool to provide an integrated approach to data collection, storage, cleaning, processing, analysis, and interpretation from potentially large volumes of information. This is meant to be a perspective article about data science based on the experience of the experts invited to the Science for Dialysis Meeting and provides an up-to-date perspective of the potential of data science in kidney disease and dialysis.

**Key messages:** Healthcare is quickly becoming data-dependent, and data science is a discipline that holds the promise of contributing to the development of personalized medicine, although nephrology still lags behind in this process. The key idea is to ensure that data will guide medical decisions based on individual patient characteristics rather than on averages over a whole population usually based on randomized controlled trials that excluded kidney disease patients. Furthermore, there is increasing interest in obtaining data about the effectiveness of available treatments in current patient care based on pragmatic clinical trials. The use of data science in this context is becoming increasingly feasible in part thanks to the swift developments in OSS.

© 2020 The Author(s).  
Published by S. Karger AG, Basel

Miguel Hueso  
Department of Nephrology, Hospital Universitari  
Bellvitge and Bellvitge Research Institute (IDIBELL), C/Feixa llarga, s/n  
ES-08907 L'Hospitalet de Llobregat (Spain)  
mhueso@idibell.cat

Alfredo Vellido  
Intelligent Data Science and Artificial Intelligence (IDEAI) Research Center  
Universitat Politècnica de Catalunya (UPC BarcelonaTech), C. Jordi Girona, 1-3  
ES-08034 Barcelona (Spain)  
avellido@cs.upc.edu

## Introduction

Computerization and digitalization are a wave that has swept, in a very short period of time, not only human labour at large but also all areas of science. In the intersection of both, healthcare is a paramount example of this phenomenon, which is radically altering the way medical care is dealt with and provided. At least part of the challenges faced by the healthcare sector have, therefore, become computer science-related problems, including medical data analysis beyond statistics using, for instance, machine learning (ML) and artificial intelligence (AI) methods [1]. The latter involves all sorts of novel social implications, including interrelated issues such as compliance with legislation, ethics and fairness, privacy and anonymity, as well as interpretability and explainability, to name a few [2].

An aspect of this phenomenon is that healthcare is quickly becoming a data-dependent endeavour, in a process that involves digital networks, fast advances in medical data acquisition methods, and the widespread adoption of electronic medical records (EMRs) [3]. These changes do not come about without their share of pain, as reported, for instance, in Ash et al. [4]. In the case of physicians who must routinely use EMRs, it could take the form of report content impoverishment, reduction of personal engagement with patients, or even de-skilling [5]. This goes beyond the classical difficulties and barriers that are to be found in the introduction of any new technology. Networked computerization, together with the *datafication* of medical science and practice, signifies a systematic shift in procedures and protocols at all levels of medical practice. The fact that medicine and healthcare are adopting computer science innovations also means they necessarily lag behind current practice in the latter.

Health data have dramatically increased in quantity and complexity over the last decade. Data science, as a compound area of expertise, has begun to be used in current clinical practice to extract knowledge and insights to guide individualized patient care from the available quantitative and qualitative information. This process of health *datafication*, part of the big data phenomenon, is fuelling a “Precision and Personalized Medicine” revolution that promises to improve diagnosis, risk assessment, and treatment of multiple diseases, so far mainly in oncology and cardiology. However, nephrology still lags behind in this process [3], despite the fact that some advances have been achieved in the application of data science and AI in the field [1]. A recent and very publicized example of this

is a deep-learning artificial neural network model for continuous risk prediction of acute kidney injury (AKI), based on a large, longitudinal dataset of electronic health records (EHRs), presented in *Nature* [6]. The proposed model was a recurrent neural network that operates sequentially over EHRs, processing the data one step at a time and building an internal memory that keeps track of relevant information available up to that point. At each time point, the model outputs a probability of AKI occurring at any stage of severity within the next 48 h. The model predicts 55.8% of all inpatients’ episodes of AKI and 90.2% of all AKIs that requires dialysis, with a lead time of up to 48 h and a ratio of 2 false alerts for every true alert. AI-based decision support systems have also been used, for instance, to recommend suitable erythropoiesis-stimulating agent doses for optimizing anaemia management in haemodialysis patients [7] and to guide the management of blood pressure, fluid volume, and dialysis dose [8].

This brief study aimed to outline and provide some perspective on some relevant topics regarding the use of data science, ML, and AI as tools for personalized haemodialysis. With this aim, we follow a thread that starts considering the problem of data collection, including an example of practical use of data science for the design of pragmatic clinical trials, and continues with topics on data storage and analysis. It also discusses the issue of open source software (OSS) for data collection and analysis in the area.

As mentioned in the abstract, the article reports the contents of the 2019 Science for Dialysis Meeting at Bellvitge University Hospital in Barcelona, Spain, which was divided into different parts, loosely related to areas in data science (often characterized as the triad of (1) data collection, (2) data modelling and analysis, and (3) interpretation-oriented tools to extract actionable knowledge suitable to support complex decision-making). Therefore, it does not intend to be a review of the field, but to provide an up-to-date comprehensive perspective of the potential of data science in kidney disease and dialysis.

## Advances in Data Collection and Storage

### *E-Health and Remote Health Models for Dialysis Units: The Nephrologist’s Point of View*

Advances in the field of information management are quickly opening multiple opportunities for development in the health sciences. The application of new technologies in the specific field of treatment of advanced CKD

offers many advantages at multiple levels. First, the nephrologist has access to intuitive software and interfaces that are fully connected to the different treatments and monitoring devices that instantly transmit physiological data (blood pressure, weight, and heart rate) and treatment information, allowing for a centralization of the information and an updated integration with the hospital medical information systems [9, 10]. This setting should contribute to a much more precise and personalized service [11], but only if the possibilities presented by data availability are properly leveraged. Second, clinical managers have access to tools that allow them monitoring resources (both human and material), clinical and productivity indicators, in order to increase efficiency and improve assistance quality and distribution of resources. Finally, tools of e-health, m-health, and tele-health provide patients and their environment with plenty of resources for the understanding of their illness, self-care, and, in some cases, self-treatment [12]. Nephrologists should be aware of these advances and play a leading role in their development, without forgetting, in many cases, to continue applying evidence-based medicine with medical responsibility and ethics.

Nephrologists must be aware that data science, as discussed in the following sections, goes beyond quantitative data to embrace the true diversity and complexity of available medical information. Much of this information is currently structured in the form of EHRs, which provide opportunities to enhance patient care and improve the identification, and are increasingly being used for research, beyond the primary purpose for which they were collected [13]. They may assist in the assessment of whether new treatments or innovation in healthcare delivery result in improved outcomes or healthcare savings. In addition, clinical managers have access to tools that allow them to monitor resources (both human and material), clinical and productivity indicators, in order to increase efficiency and improve assistance quality and distribution of resources.

#### *The ARGOS Project in E-Health: A Path towards Standardization of Clinical Practice*

Health professionals have always collected information in order to improve clinical care and to be able to learn about diseases. The current shift from traditional medical records to EMRs is profoundly transforming the way we can analyse this information. Fifteen years ago, the Catalan Health Institute started the ARGOS institutional project in Catalonia, Spain: a new system of clinical information management in hospitals. One of the great

difficulties that was faced at the onset was how to replicate in digital format what the medical doctors did on paper. Nowadays, with professionals already accustomed to healthcare electronics and with the accumulated experience, a vast universe of possibilities opens to show, collect, and analyse the data resulting from medical care. From the traditional clinical workstation (CWS), we are moving towards CWSs oriented to specific processes and pathologies, where the information is displayed intelligently, accompanying the needs of professionals always. To help achieve this goal, the ARGOS project has built a clinical dictionary, where variables are always defined in a structured way so that they can be used in any form that requires the assistance and treatment of patients and different pathologies [14, 15].

Each variable is unique, correctly defined, and coded using international standards (SNOMED CT [Systematized Nomenclature of Medicine-Clinical Terms] and LOINC [Logical Observation Identifiers Names and Codes]) and national standards (Spanish Society of Medical Radiology (SERAM) and Spanish Society of Nuclear Medicine and Molecular Imaging (SEMNiM) that will then allow technical and semantic interoperability between systems. The forms are sets of variables that can be used uniquely or incorporated into previously defined care processes, which include workflows of professionals and the predetermined rules linked to algorithms that facilitate clinical management.

The use of rules within the processes has allowed experts to use automatisms able to construct clinical notes and reports derived from each assistance act and facilitate the daily work of doctors and other clinical professionals. All these variables may come from any source: laboratories, pharmacy, medical monitoring equipment, clinical records, and others. Moreover, they will be stored in a MongoDB database type from where we can retrieve information either for clinical assistance or for analysis in research tasks. Having all this structured and accessible information allows healthcare professionals to exploit it at any level, to facilitate assistance, research, and teaching, while opening the doors to the use of AI with methods from ML or deep learning, for instance. These tools will be useful for healthcare system users, in a process-oriented manner, prioritizing prevention and based on the scientific evidence available at all time. After more than 10 years of working with the electronic clinical report, this is already a way of no return, which improves medical practice and the quality of care received by citizens.

### *Pragmatic Randomized Controlled Trials: A Way to Transform Data into Evidence*

Pragmatic randomized controlled trials (pRCTs), as compared to explanatory controlled trials, are meant to focus more on possible correlations between data of treatments and their outcomes in real-world healthcare than on investigating causal explanations for outcomes. Interestingly, this is not unlike the data science approaches to knowledge discovery as compared to more traditional statistical approaches. This similarity makes pRCT an interesting and barely frequented research field for data scientists. The latter should consider the concept of real-world data, also referred to as real-world evidence, described by Gibert et al. [16] as providing patient-level data gathered outside the conventional clinical trial setting, including pRCT.

Current medicine lacks enough comparative effectiveness research of marketed medicines. Although there are a number of scientific approaches to address this issue, the only one that could offer a cause-effect comparative assessment is the randomized controlled trial (RCT), and most specifically the pRCT. These are trials conducted resembling usual clinical practice, aiming to assess the effectiveness of marketed medicines that are prescribed and managed as in normal clinical practice, and that will provide results that will be generalizable to other settings. pRCTs that pose no or minimal incremental risk compared with usual clinical practice are named “low-risk pRCTs” [17] or “low-intervention trials” in the EU clinical trials regulation [18]. These trials are of critical relevance to public health since they provide ready-to-use information on the comparative effectiveness of the assessed medicines.

Research has shown that pRCTs with medicines are rarely conducted [19]. One of the main reasons for this is that pRCTs must follow all the ethical and administrative requirements that the EU regulation asks for all types of trials, such as pre-licensing RCTs [18]. One of the most relevant hurdles is the need to seek participants’ written informed consent. Recent developments regarding the informed consent hurdle could dramatically change the current situation.

The 2016 CIOMS (Council for International Organizations of Medical Sciences) ethical guidelines consider what type of human research could be conducted with a modification or waiver of participants’ informed consent [20]. Thus, any research fulfilling three provisions and that is approved by the relevant research ethics committee could be conducted without participants’ consent. The 3 provisions are as follows [20]: (a) the research

would not be feasible or practicable to carry out without the waiver or modification, (b) the research has important social value, and (c) the research poses no more than minimal risks to participants. It is clear that these provisions are applicable to low-risk pRCTs [21]. But not all low-risk pRCTs could be candidates to fulfil the CIOMS provisions: first trials must show a high degree of pragmatism that will ensure the generalizability of the results obtained [19, 21], something that is uncommon, even among those RCTs self-tagged as pragmatic [22]. The best way to assess the degree of pragmatism of an RCT is by using specific tools such as the PRECIS-2 tool [23], which considers nine domains – from the eligibility of participants to the statistical analysis, to be individually assessed.

To ascertain how many RCTs could be candidates for the modification or waiver of participants’ informed consent, a research was conducted on the EU-clinical trials register database of all phase 4 RCTs that were “ongoing” in July 2016 to June 2018. From 420 RCTs, only 21 could be candidates to fulfil the CIOMS provisions, and only 8 (out of 15 that responded to our questionnaires) fulfilled them following the assessment of the investigators and, with inconsistent results, of members of research ethics committees and patients [24]. This type of RCT will be eased if the EU Commission amends the regulation to allow the conduct of those low-risk pRCTs that fulfil the three CIOMS provisions [21].

pRCTs could pose a number of difficulties in their conduct due to the high number of participants that some of them need to answer the research questions. This also implies that the number of sites and investigators could be substantial, hence increasing its complexity from the operational perspective. An alternative to this type of trial is the conduct of an observational research by means of routinely collected data (RCD), such as that obtained from EHRs. Recent research shows, however, that 68% of RCD assessing the comparative effectiveness of interventions were previously compared in RCTs – many of them not assessing the effects in the real world – so they did not provide fundamentally novel research results [25]. Therefore, research conducted with RCD to add completely new information is not common. It would be reasonable to encourage the conduct of research based on RCD to answer those questions for which pRCTs would be unfeasible or impractical.

### *Data Storage*

Data generated by EMRs, including unstructured data, Internet of things, and imaging records, are estimated to

be about 150 exabytes ( $D_{GB} = D_{EB} \times 1,024^3$ ) and keep growing at exponential levels, quadrupling every 2–3 years [26], which entails the risk of overflowing the healthcare system. These big data must be saved for diagnostic activities, regulatory compliance, coverage, or clinical research, and up to 92% are stored on hard disk drives. Thus, data storage requirements are a key point, since we have to ensure that data are stored securely and efficiently and are quickly accessible. The cloud offers the only viable solution for the out-of-control growth of healthcare industry data. However, up to 80% of healthcare data are “dark” because they are spread across numerous single-point repositories and cannot be easily managed, searched, or exported for analysis (medical or administrative), regulatory request, or e-discovery [27]. In addition, most EHR storage is in the hands of private tech companies, such as, for instance, Microsoft Azure or Amazon Web Services, but there is an increasing interest in using public cloud systems that can scale rapidly and efficiently. IBM’s high-performance data and AI (HPDA) architecture based on storage and software-defined storage for healthcare has been developed for cloud-scale data management, multi-cloud workload orchestration, and converged high-performance computing with deep learning.

A fundamental need in storing clinical data is to make them accessible to those performing the analysis and, at the same time, guarantee patients’ privacy. For larger systematic analyses, cryptographic techniques are being developed. Alternatively, the so-called virtualization technologies allow scientists to submit their analytical tools to be available remotely on a server, enabling analyses to be performed without having to share the actual data.

In the field of dialysis, the Fresenius company has developed the Therapy Data Management System (TDMS) for supporting users in daily tasks and the Therapy Support Suite central clinical management for supporting users in carrying out analyses, creating and managing patient prescription, developing medication plans, and documenting treatment data and patient-related laboratory data. In addition, the TDMS provides enhanced reporting capabilities and allows for the professional management of a group of dialysis centres [28].

## Data Analysis

### *Data Analytics and Computational Tools for Kidney Diseases and Dialysis*

Haemodialysis therapy generates large-scale data from electronics, mechanics, physics, chemistry, or

physiology to assess the quality of the current treatment. Up-to-date data from the extracorporeal circuit (air detector, arterial and venous blood pump, substitution fluid pump, blood pressure, and blood volume) and the hydraulic circuit (dialysis fluid flow rate, temperature, conductivity, transmembrane pressure, blood leak detector, ultrafiltration rate, and dialysis dose) are recorded. For example, the DBB-EXA dialysis machine (Nikkiso Co.) generates 87 data values every 30 s and 41,756 clinically relevant data values over each 4-h-long dialysis session. A monitor connected via ethernet (network connection) sends the information in NoSQL (non-structured query language) to the data processing centre or a cloud computing platform. In addition, the introduction of potent computational platforms is able to link these data with other sources of data (e.g., health environmental records, biomedical research databases, genome sequencing databanks, pathology laboratories, or mobile Internet of things). To manage and analyse this enormous amount of data, we need to employ data science analytics, which provides theory and methods combining mathematics, statistics, AI, computer science, optimization, data visualization, and information science [29].

As described in the previous section, data collection is the key component of any data science project, since the quality of the knowledge extracted by a data mining process is dependent on the quality and suitability of data. Factors such as missing values and inconsistent, redundant, erroneous, or needless data have a high impact and lead to low-quality knowledge. This means that an essential stage to consider is data preprocessing. Thus, to extract information from data, we first need to process, examine, clean, verify for quality, and normalize each type of data before any further analysis. This process involved several steps: (1) deletion of any duplicate data that might appear within the data set, (2) resolution of any conflicting data, and (3) conversion of data into a format for further processing and analysis. However, the quality and accuracy of such data, and issues on how to harmonize data arriving from different sources so as to provide a comparable view from different studies are still in need of improvement. An overview of the use of big data in renal diseases has been recently reviewed in Saez-Rodriguez et al. [3].

Once a quality data set has been established, a data model needs to be identified. A data model is a representation of the data and their relationships, obtained by applying certain formal techniques, and can be understood at 3 levels: descriptive, predictive, and prescriptive. The

process of designing a data model begins with the identification of the necessary data and the relationships among components of that data. The structure of a data model can be described in different forms, such as a relational model, an object-oriented model, or a hierarchical structure. There are many data modelling techniques and also many open source tools that can be used to facilitate the design of a data model. The choice of the optimal algorithm depends on a variety of factors that include, but are not limited to, data type/learning approach (supervised or unsupervised learning), the importance of accuracy in the chosen model, the need for speed in data analysis, the data analysed, the size of the data set, the need for hierarchical output, or the need for categorical variables [30].

Complex computational methods based on ML techniques are also increasingly been applied to biomedical data to identify an optimal model, i.e., fully vetting the algorithms by building and testing multiple models for their appropriateness to the task. ML algorithms can be generally classified into three categories: (1) supervised learning, which uses training data to learn a function and which includes classification and regression approaches; (2) unsupervised learning, such as clustering algorithms; and (3) reinforcement learning. ML tools have been used as a screening tool to predict the progression of diabetic kidney disease [31] and to identify dialysis patients with high risk of death [32]. In addition, ML aids diagnosis for image processing in pathology, and laboratory medicine [33] and genomic data analysis [34].

Another data science approach to facilitate the interpretation of complex data models entails the use of interactive visual interfaces in a discipline known as visual analytics [35]. This approach could provide nephrologists with more effective ways to combine longitudinal clinical data with dialysis-generated health data to better understand patient progression. In addition, patients could be supported in understanding health plans and comparing their health measurements with other patients. Visual analytics systems have the potential to support intuitive analysis while masking the underlying complexity of the data. Using data science according to these ideas, nephrologists have the opportunity to develop novel therapeutic tools for a personalized medicine targeting dialysis patients.

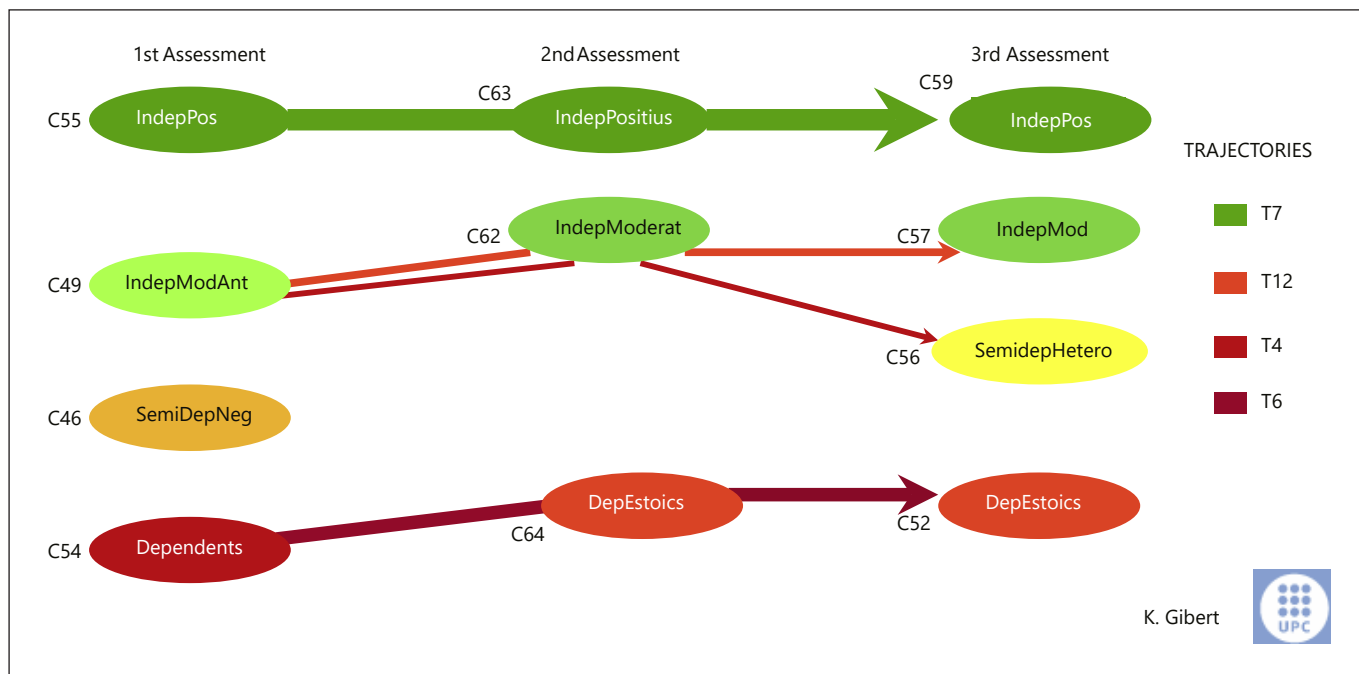
Whatever the methodology used to build up a clinical prediction model, this requires a validation protocol that guarantees generalization of the results in the application domain, taking into consideration that it can be different samples from a patient or a third-party

population. In addition, whether the tool will be used to make treatment decisions should be evaluated as an intervention.

#### *Data Science and Interpretation-Oriented Tools for an Easy Follow-Up of Long-Term Treatments*

Data science is often characterized as a triad that, beyond data collection and data modelling and analysis, also has to consider the design and use of interpretation-oriented tools to extract actionable knowledge for medical decision-making. Here, a data science methodology useful to identify dynamic understandable patterns from data is presented. Clustering based on rules by states (ClbRxS) [36] is a general data science approach to extract relevant decisional knowledge from data useful to support complex decision-making in real problems that evolve along time. When several waves of data are available, the ClbRxS proposes a multiview-like clustering approach [37], where each view corresponds to one data wave and local clustering is performed inside each of the waves. This produces several classifications of the objects (patients) regarding their situations in the several data waves. The method includes the use of several interpretation-oriented tools to support a further process of conceptualization of the clusters (such as the class panel graph [16] or traffic lights panels [38]) that constitutes a final step of post-processing [39], devoted to guarantee the understandability of the clusters from the clinical point of view, which is critical to bridge the gap between the results of the data mining step and the effective decision-making layer. The final outcome of the proposed methodology is named trajectories map and consists of a visual diagram with most frequent paths of objects (patients) through the classes of successive data waves, as seen in Figure 1. This map is very intuitive and provides an easily interpretable perspective of how patients evolve across time, allowing the experts to both follow-up each single patient and analyse main evolution patterns to associate treatments or actions to each of them. The results in Figure 1 come from a real application to understand the emotional and functional evolution of spinal cord injury patients from Hospital Guttman (a referral centre of neurological lesion in southern Europe) after discharge [40].

The trajectories map is also useful to associate the path followed across time to each patient and to analyse the relationship of this path with other variables not included in the clustering process, which are called illustrative variables. These can help enrich interpretation or identify the differences between patients following one path or an-



**Fig. 1** Trajectories map of social inclusion of spinal cord injury patients. This is an automatic software-generated visual representation. Each column of nodes represents a data wave. Each node represents one of the clusters found in that wave, which is labelled according to the patient's profile represented by the cluster, after a conceptualization process made by experts on the basis of cluster information automatically provided by the software. In this case, labelling regards the functionality, wellness, and sometimes evolution of the lesion. Colours can be associated with a latent target concept, like, for example, more (in red) or less (in green) global impairment of the patient. This is represented with the vertical position of the nodes in the graph. Edges represent the paths followed by patients along time, and thickness of the edges is associ-

ated with observed prevalence of each path. *IndepPos*, functionally independents, with assistive technologies required, and feel wellness; *IndepModAnt*, functionally independent with moderate distress and old lesions; *SemiDepNeg*, very distressed and require help for some specific daily life activities such as moving from bed to wheelchair or going to bathroom; dependents, dependent and psychologically heterogeneous; *IndepPositius*, independent and feel wellness; *IndepModerat*, independent and moderate wellness; *DepEstoics*, dependents, but feel wellness; *IndepMod*, independents and moderate wellness; and *SemidepHetero*, with dependence for some specific daily life activities such as moving from bed to wheelchair or going to bathroom, psychologically heterogeneous.

other, thus identifying possible predictors of the path, to be used by medical doctors to determine treatments. In the particular application presented, 4 main patterns were discovered and several hypotheses were elaborated about the reasons for psychological distress or decreases in quality of life of patients over time (which resulted to be related with the evolution time of the lesion and the dysfunctional families). The use of the proposed methodology permits the synthesis of a dynamic structure of a complex domain in an easy representation that can be understood by health professionals without technical skills in data science and the association of the discovered paths with potential predictors, useful to associate actions or decisions to each pattern. Needless to say, this methodology has great potential to assess the follow-up of patients in the area of nephrology.

### OSS for Data Science in the Medical Domain

As stated in the Introduction, part of the challenges faced by the healthcare sector have become computer science-related problems that can be addressed through data science. Let us put forward 2 aspects of this that are transversal to the complete data science domain: the use of OSS in medical practice and the potential benefit of endowing health practitioners with computer programming skills.

The last decade has witnessed how OSS has overtaken proprietary software in data science, not only amongst core experts but also in industry and business applications in general. In practice, this means that the advantages of OSS have proved to be more than those of the proprietary counterparts. It has been suggested [41] that

in the medical informatics domain, software vendors have failed to provide stable enough technical partnerships and cost-effective products. This is compounded by the limited standardization that proprietary products provide due to incompatibilities between software products of different vendors. OSS eliminates licensing costs, promotes compatibility, and allows customizing the software tools to the medical client needs and requirements [42]. Furthermore, by its own nature, OSS encourages and nurtures innovation and collaboration, shortening software development cycles [43, 44]. The OSS movement in medical informatics is promoted by international organizations such as the Open Source Working Group of the AMIA [45], the Open Source Health Informatics Working Group of the IMIA [46], or the Libre/Free and OSS Working Group of the EFMI [47], to name a few.

To this day, there is very limited work on the development of OSS solutions specifically tailored to the nephrology domain, and from the data science perspective; they mostly concern data collection, storage, and management. A pioneering experience was PatientView [48], a National Health Service-run UK-wide online system letting nephrology patients see test results almost in real time. It also allows patients to add supplementary information and can be seen, overall, as a data collection and management system. In fact, most OSS available systems focus on this part of data science, including kidney diseases clinical database construction [49] and, increasingly, EMR/EHR data processing. The latter has been applied, for instance, to haemodialysis clinical workflow modelling [50]. The extension to systems that cover and integrate also data analytics is still fairly speculative, although some attempts have been made, such as, for instance, the definition of an architecture for clinical decision support that integrates openEHR specifications and Bayesian networks methods [51], and, very recently and with direct application to nephrology, an architecture of OSS tools that combines textual information extraction from EHRs, faceted search, and information visualization that also involves, even if as a prototype, the use of ML methods [52].

Even if the question may sound far-fetched and, given the computerization and *datafication* of clinical practice, should medical practitioners learn computer programming skills? It stands to reason that the best way to engage healthcare professionals is by providing them with some level of control over these technologies; this means at least familiarity with programming skills, in a way that such control is not fully left in the hands of computer scientists, thus guaranteeing that computer technologies fully re-

spond to healthcare needs and requirements. There have been pilot studies and proofs of concept that demonstrate that transferring such skills is viable [53, 54] if support from the medical institutions in the form of, for instance, help desks and regulatory affairs is provided. Given the current accessibility of OSS and the increasing usability of programming tools (in the form, for instance, of mobile or web applications), technical barriers are low. An example of this is the R programming language, developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand [55], oriented towards statistics and widely used by medical statisticians. It is a GNU project [56], so it guarantees end users the freedom to run, study, share, and modify the software. Most of such users work in research, especially in the fields of biomedicine and epidemiology, but also economics, sociology, or political science, to name a few. For those who want to start in R, the following resources, among others, could be useful: Quick R is a reference guide, including good examples with codes for common tasks on data management, graphics, and basic and some advanced statistical techniques [57]. The Cookbook for R [58] has good recipes for data analysis and specifically about data visualization in R; all recipes include codes ready to be copied and pasted into R. It must also be noted that R has a steep learning curve; this is in part because the way to analyse data is not as intuitive as, for example, in Statistical Package for the Social Sciences (SPSS), a proprietary software (now part of IBM) traditionally favoured in the life sciences. Arguably, that intuitive feeling could be seen as a mirage or a simplified view, while with R, the user takes full control of the analysis.

It is also true, though, that no standard for teaching coding to non-programmers such as healthcare workers exists so far and that software quality standard issues are far from a small hurdle to overcome. A more gradual approach would entail adding programming skills to current medical students' curricula, ensuring that new generations of medical practitioners can "both practice medicine and engage in the development of useful, innovative technologies to increase efficiency and adapt to the modern medical world" [59].

## Conclusions

1. Healthcare is quickly becoming data dependent, and data science is a discipline that holds the promise of contributing to the development of personalized medicine, although nephrology still lags behind in this process.



2. Data will guide medical decisions based on individual patient characteristics rather than on averages over a whole population usually based on RCTs that excluded kidney disease patients.
3. CWSs are moving towards specific processes and pathologies, where information is displayed intelligently, accompanying the needs of professionals. The ARGOS institutional project in e-health is an example of such a system for the standardization of clinical practice.
4. There is increasing interest in obtaining data concerning the effectiveness of available treatments in current patient care based on pragmatic clinical trials focused on correlation between treatments and outcomes.

### Acknowledgements

We acknowledge our invited speakers at the “3rd Meeting of Science and Dialysis” for their generous support to this initiative; medical doctors, nurses, and patients of the dialysis units for their interest in the technology and future of dialysis devices; the Direction of Hospital Universitari Bellvitge for their technical assistance in the organization of the meeting; and Rosa Perez-Garzón for her excellent administrative work. We thank CERCA Program/Generalitat de Catalunya for institutional support. M.H. is member of the Big Data and Artificial Intelligence working group (BigSEN) from the Spanish Society of Nephrology (SEN).

### References

- 1 Hueso M, Vellido A, Montero N, Barbieri C, Ramos R, Angoso M, et al. Artificial intelligence for the artificial kidney: pointers to the future of a personalized hemodialysis therapy. *Kidney Dis*. 2018;4(1):1-9.
- 2 Vellido A. Societal issues concerning the application of artificial intelligence in medicine. *Kidney Dis*. 2019;5(1):11-7.
- 3 Saez-Rodriguez J, Rinschen MM, Floege J, Kramann R. Big science and big data in nephrology. *Kidney Int*. 2019;95(6):1326-37.
- 4 Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc*. 2004;11(2):104-12.
- 5 Hoff T. Deskillling and adaptation among primary care physicians using two work innovations. *Health Care Manage Rev*. 2011;36(4):338-48.
- 6 Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-9.
- 7 Barbieri C, Molina M, Ponce P, Tothova M, Cattinelli I, Ion Titapiccolo J, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int*. 2016;90(2):422-9.
- 8 Barbieri C, Cattinelli I, Neri L, Mari F, Ramos R, Brancaccio D, et al. Development of an artificial intelligence model to guide the management of blood pressure, fluid volume, and dialysis dose in end-stage kidney disease patients: proof of concept and first clinical assessment. *Kidney Dis*. 2019;5(1):28-33.
- 9 Déziel C, Bouchard J, Zellweger M, Madore F. Impact of hemocontrol on hypertension, nursing interventions, and quality of life: a randomized, controlled trial. *Clin J Am Soc Nephrol*. 2007;2(4):661-8.
- 10 Ronco C. Data acquisition and management. In: Ronco C, editor. *Hemodialysis technology*. Vicenza: Karger Medical and Scientific Publishers; 2002. p. 449.
- 11 Usvyat L, Dalrymple LS, Maddux FW. Using technology to inform and deliver precise personalized care to patients with end-stage kidney disease. *Semin Nephrol*. 2018;38(4):418-25.
- 12 Stevenson JK, Campbell ZC, Webster AC, Chow CK, Tong A, Craig JC, et al. eHealth interventions for people with chronic kidney disease. *Cochrane Database Syst Rev*. 2016 Oct;8(10):CD012379.
- 13 Rosner MH, Lew SQ, Conway P, Ehrlich J, Jarrin R, Patel UD, et al. Perspectives from the kidney health initiative on advancing technologies to facilitate remote monitoring of patient self-care in RRT. *Clin J Am Soc Nephrol*. 2017;12(11):1900-9.
- 14 Joana JM, Gracia R, García AL, Bolart J. *Gestión con Éxito de Proyectos de Transformación: el Caso ICS*. Barcelona: Editorial PROFIT; 2011.
- 15 González Cuyàs M, Figuerola Batista M, Galdà Azofra J, Gracia Escoriza R, de Haro Martín L. La transformación de una red hospitalaria pública: el Institut Català de la Salut. In: Temes Montes J, Gengibar Torres M, editors. *Gestión hospitalaria, 5a ed*. MC Graw Hill; 2011.

### Disclosure Statement

C.T. and R.D.R. received honoraria from Palex for their participation in this meeting.

### Funding Sources

Palex provided funds for the celebration of the “3rd Meeting of Science and Dialysis.”

### Author Contributions

This report comes from the 3rd Science of Dialysis Meeting that took place in Hospital Universitari Bellvitge on September 20, 2019. The contributions of each author are as follows: M.H. wrote the sections *Introduction*, *Data Storage*, and *Data Analytics and Computational Tools for Kidney Diseases and Dialysis*; and A.V. wrote the *Introduction* and *OSS for Data Science in the Medical Domain*. As Chairmen, they oversaw the consistency of the manuscript edition. Other sections were written by the following participants: L.d.H. wrote *The ARGOS Project in E-Health: A Path towards Standardization of Clinical Practice*; J.C. wrote *E-Health and Remote Health Models for Dialysis Units: The Nephrologist's Point of View*, C.T. wrote about *The R Project for Statistical Computing* included in *OSS for Data Science in the Medical Domain*, K.G. wrote *Data Science and Interpretation-Oriented Tools for an Easy Follow-Up of Long-Term Treatments*, R.D.R. wrote *Pragmatic Randomized Controlled Trials: A Way to Transform Data into Evidence*, and J.M.C. helped drafting the manuscript and provided intellectual content of critical importance to the work. *Conclusions* were written by all participants.

- 16 Gibert K, García-Rudolph A, Rodríguez-Silva G. The role of KDD support-interpretation tools in the conceptualization of medical profiles: an application to neurorehabilitation. *Acta Informatica Med.* 2008;16(4):178–82.
- 17 Lantos JD, Wendler D, Septimus E, Wahba S, Madigan R, Bliss G. Considerations in the evaluation and determination of minimal risk in pragmatic clinical trials. *Clin Trials.* 2015; 12(5):485–93.
- 18 Regulation (EU) No. 536/2014 of the European Parliament and of the Council of 16 of April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC. Official Journal of the European Union L 158/1-76; 27 May 2014. Available from: [https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg\\_2014\\_536/reg\\_2014\\_536\\_en.pdf](https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf).
- 19 Dal-Ré R, Janiaud P, Ioannidis JPA. Real-world evidence: how pragmatic are randomized controlled trials labeled as pragmatic? *BMC Med.* 2018;16(1):49.
- 20 Council for International Organizations of Medical Sciences (CIOMS). *International ethical guidelines for health-related research involving humans*. Geneva; 2016. Available from: <https://cioms.ch/wp-content/uploads/2017/01/WEBCIOMS-EthicalGuidelines.pdf>.
- 21 Dal-Ré R, Avedaño-Solà C, Bloechl-Daum B, de Boer A, Eriksson S, Fuhr U, et al. Low risk pragmatic trials do not always require participants' informed consent. *BMJ.* 2019; 364:l1092.
- 22 Janiaud P, Dal-Ré R, Ioannidis JPA. Assessment of pragmatism in recently published randomized clinical trials. *JAMA Intern Med.* 2018;178(9):1278–80.
- 23 Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ.* 2015;350:h2147.
- 24 Dal-Ré R, Avedaño-Solà C, de Boer A, James SK, Rosendaal FR, Stephens R, et al. A limited number of medicines pragmatic trials had potential for waived informed consent following the 2016 CIOMS ethical guidelines. *J Clin Epidemiol.* 2019;114:60–71.
- 25 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Current use of routinely collected health data to complement randomized controlled trials: a meta-epidemiological survey. *CMAJ Open.* 2016;4(2):E132–40.
- 26 Data Never Sleeps 7.0. Available from: <https://www.domo.com/learn/data-never-sleeps-7>.
- 27 Improving Healthcare Data Management. Available from: <https://www.healthitoutcomes.com/doc/where-should-healthcare-data-be-stored-in-and-beyond-0001>.
- 28 Therapy Support Suite (TSS). Available from: <https://www.freseniusmedicalcare.com/en/healthcare-professionals/renal-it/therapy-support-suite-tss>.
- 29 Gibert K, Horsburgh JS, Athanasiadis IN, Holmes G. Environmental data science. *Environ Model Softw.* 2018;106:4–12.
- 30 Aktolun C. Artificial intelligence and radiomics in nuclear medicine: potentials and challenges. *Eur J Nucl Med Mol Imaging.* 2019;46(13):2731–6.
- 31 Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci Rep.* 2019;9(1):11862–9.
- 32 Akbilgic O, Obi Y, Potukuchi PK, Karabayir I, Nguyen DV, Soohoo M, et al. Machine learning to identify dialysis patients at high death risk. *Kidney Int Rep.* 2019;4(9):1219–29.
- 33 Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Acad Pathol.* 2019;6: 2374289519873088.
- 34 Niel O, Bastard P. Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. *Am J Kidney Dis.* 2019; 74(6):803–10.
- 35 Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput Appl.* Forthcoming.
- 36 Gibert K, Rodríguez-Silva G, Rodríguez-Roda I. Knowledge discovery with clustering based on rules by states: a water treatment application. *Environ Model Softw.* 2010;25(6):712–3.
- 37 Bickel S, Scheffer T. Multi-view clustering. *Proceedings of the IEEE International Conference on Data Mining*; 2004. Vol. 4; p. 19–26.
- 38 Gibert K, Conti D. aTLP, C: preface. *AI Commun.* 2015 Jan 1;28(1):1–3.
- 39 Sierra K, Rodríguez-Silva G, Annicchiarico R. Post-processing: bridging the gap between modelling and effective decision-support. The profile assessment grid in human behaviour. *Math Computer Model.* 2013;57(7–8): 1633–9.
- 40 KGibert, García-Rudolph A, Curcoll L, Soler D, Pla L, Tormos JM. Knowledge discovery about quality of life changes of spinal cord injury patients: clustering based on rules by states. *Stud Health Technol Inform.* 2009;150: 579–83.
- 41 Caban JJ, Joshi A, Nagy P. Rapid development of medical imaging tools with open-source libraries. *J Digit Imaging.* 2007;20(Suppl 1): 83–93.
- 42 Karopka T, Schmuhl H, Demski H. Free/libre open source software in health care: a review. *Healthc Inform Res.* 2014;20(1):11–22.
- 43 Ratib O, Rosset A, Heuberger J. Open source software and social networks: disruptive alternatives for medical imaging. *Eur J Radiol.* 2011;78(2):259–65.
- 44 Aminpour F, Sadoughi F, Ahmadi M. Utilization of open source electronic health record around the world: a systematic review. *J Res Med Sci.* 2014;19(1):57–64.
- 45 Open Source Resources. Available from: <https://www.amia.org/programs/working-groups/open-source>.
- 46 Open Source Health Informatics Working Group. Available from: <https://imtia-medinfo.org/wp/open-source-health-informatics>.
- 47 European Federation for Medical Informatics. Available from: <https://www.efmi.org/workinggroups/lifoss-librefree-and-open-source-software>.
- 48 Hudson C, Darking M, Cox J. Understanding the value of patientview for enabling self-care practice in chronic kidney disease. *J Ren Care.* 2019;46(1):13–24.
- 49 Singh SK, Malik A, Firoz A, Jha V. CDKD: a clinical database of kidney diseases. *BMC Nephrol.* 2012;13(1):23.
- 50 Yu B, Wijesekera D. Building dialysis workflows into EMRs. *Proced Technol.* 2013;9: 985–95.
- 51 Arikan SS. An experimental study and evaluation of a new architecture for clinical decision support-integrating the openEHR specifications for the Electronic Health Record with Bayesian Networks. Doctoral [dissertation]. University College London; 2016.
- 52 Sonntag D, Profitlich HJ. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artif Intell Med.* 2019; 93:13–28.
- 53 Kubben P, Looije P, Scherpbier A, Merode F. Teaching computer programming to medical doctors, nurses and hospital staff: a pilot study. *OAJN.* 2017;4(2):555632. .
- 54 Zhang MW, Tsang T, Cheow E, Ho CS, Yeong NB, Ho RC. Enabling psychiatrists to be mobile phone app developers: insights into app development methodologies. *JMIR mHealth uHealth.* 2014;2(4):e53.
- 55 Ihaka RR. *Past and future history*. Auckland, New Zealand: Statistics Department. The University of Auckland; 1998. Technical report.
- 56 The GNU project. Available from: [https://en.wikipedia.org/wiki/GNU\\_Project](https://en.wikipedia.org/wiki/GNU_Project).
- 57 Quick-R. Available from: <https://www.stat-methods.net>.
- 58 The Cookbook for R. Available from: <http://www.cookbook-r.com>.
- 59 Morton CE, Smith SF, Lwin T, George M, Williams M. Computer programming: should medical students be learning it? *JMIR Med Educ.* 2019;5(1):e11940.