MODELLING AND SIMULATION 2024

THE EUROPEAN SIMULATION

AND

MODELLING CONFERENCE

2024

ESM®'2024

EDITED BY

J. David Nuñez-Gonzalez

and

Manuel Graña

OCTOBER 23-25, 2024

SAN SEBASTIAN
SPAIN

A Publication of EUROSIS-ETI

The 38th Annual European Simulation and Modelling Conference 2024

SAN SEBASTIAN, SPAIN

OCTOBER 23-25, 2024

Organised by

ETI - The European Technology Institute

Sponsored by

EUROSIS - The European Simulation Society

AREA42

Co-Sponsored by

KULeuven BIOTEC

**University
of the Basque Country**

**University
of
Žilina**

**University
of
Skovde**

GODAN

OFFIS

Hosted by

University of the Basque Country
San Sebastian, Spain

EXECUTIVE EDITOR

**PHILIPPE GERIL
(BELGIUM)**

EDITORS

General Conference Chairs

Jose David Nuñez Gonzalez, UPV/EHU, Spain
Manuel Graña Romay, UPV/EHU, Spain

General Programme Chair

Philippe Geril, ETI Bv, Ostend, Belgium

Local Programme Committee

Jose David Nuñez Gonzalez, UPV/EHU, Spain
Manuel Graña Romay, UPV/EHU, Spain
Jose Manuel Lopez Guede, UPV/EHU, Spain
Josu Mayora Oira, UPV/EHU, Spain
Borja Fernandez Gauna, UPV/EHU, Spain
Oihana Aristondo Echeberria, UPV/EHU, Spain
Anna Kaminska Chuchmala, UPV/EHU, Spain
Asier Garmendia Orbegozo, Tecnalia, Spain
Miguel Angel Antón González, Tecnalia, Spain
Alain Ulazia Manterola, UPV/EHU, Spain
Andoni Elola Artano, UPV/EHU, Spain

Sebastián Ríos, University of Chile, Chile
Richard Duró Fernández, University of A Coruña, Spain
Maya Dimitrova, Bulgarian Academy of Science, Bulgaria
Lyuba Alboul, Sheffield Hallam University, Sheffield, United Kingdom
Anna Lekova, Bulgarian Academy of Science, Bulgaria
Roman Zaharev, Bulgarian Academy of Science, Bulgaria
Michal Wozniak, Wroclaw University of Science and Technology, Wroclaw, Poland
Emilia Barakova, Tue, Eindhoven, The Netherlands

Past Conference Chairs
Rob Vingerhoeds, ISAE-SUPAERO, Toulouse, France
Pierre de Saqui-Sannes, ISAE-SUPAERO, Toulouse, France

Journal Publication Chairs
Yan Luo, NIST, Gaithersburg, USA
Peter Lawrence, Swinburne University, Australia
Dr. Wan, International Islamic University Malaysia

ESM Conference Chair
António Carvalho Brito, FEUP - University of Porto, Porto, Portugal

INTERNATIONAL PROGRAMME COMMITTEE

Methodology and Tools
Claudia Krull, Otto-von-Guericke University, Magdeburg, Germany
Andrzej Najgebauer, Military University of Technology, Warsaw, Poland
Ian Oliver, Nokia Bell Labs, Finland
Jerzy Respondek, Silesian University of Technology, Gliwice, Poland

INTERNATIONAL PROGRAMME COMMITTEE

Object-Orientation and Re-Use
Ian Oliver, Nokia Bell Labs, Finland

Discrete Simulation Modeling Techniques and Tools
Helge Hagenauer, Universitaet Salzburg, Salzburg, Austria
Sophie Hennequin, ENIM, Metz Cedex, France
Konrad Polys, PAS, Institute of Applied Informatics, Gliwice, Poland

Queueing Models
Izabella V. Lokshina, SUNY Oneonta, Oneonta NY, USA

Simulation and Artificial Intelligence
Martin Hruby, Brno University of Technology, Brno, Czech Republic
Vladimir Janousek, Brno University of Technology, Brno, Czech Republic
Kifaya Qaddoum, Digipen, Seattle, USA
Leon Rothkrantz, TU Delft, The Netherlands
Ali Shams Nateri, University of Guilan, Rasht, Iran

AI and Fuzzy Systems
Pilar Fuster-Parra, Universitat de les Illes Balears, Palma de Mallorca, Spain
Ali Shams Nateri, University of Guilan, Rasht, Iran

Agent Based Simulation
Kurt De Cock, Ghent University, Ghent, Belgium
Ioan Alfred Letia, TU Cluj Napoca, Romania
Pilar Fuster-Parra, Universitat de les Illes Balears, Palma de Mallorca, Spain
Isabel Praca, Ist. Superior do Porto, Portugal

Simulation and Optimization
José António Oliveira, Universidade do Minho, Campus de Gualtar, Braga, Portugal
Janos-Sebestyen Janosy, Hungarian Academy of Sciences, Budapest, Hungary

IOT and Smart Industry (Internet of Things & Industry 4.0)
Track Chair
Abderrazak Jemai, INSAT, University of Carthage, Tunis, Tunisia

Ahmed Chiheb Ammari, INSAT, Tunis, Tunisia
Mhamed Ghazel, CNRS, Lille, France
Konrad Polys, PAS, Institute of Applied Informatics, Gliwice, Poland
Habib Smei, Iset Rades, Tunis, Tunisia

High Performance Large Scale and Hybrid Computing
David Hill, Universite Clermont Auvergne, Clermont-Ferrand, France
Pierre Siron, ONERA, Toulouse, France
Jingjing Wang, SUNY Binghamton University, New York, USA

Simulation in Education and Graphics Visualization
Marco Roccetti, University of Bologna, Italy

Simulation in Environment, Ecology, Biology and Medicine
Joel Colloc, Université du Havre, Le Havre, France
Laurent Perochon, VetaGro-Sup, Lempdes, France

Analytical and Numerical Modelling Techniques
Ana M. Camacho, UNED, Madrid, Spain

Web and Cloud Based Simulation
Manuel Alfonseca, Universidad Autonoma de Madrid, Spain
Peter Kvasnica, Alexander Dubcek University of Trencin, Trencin, Slovak Republic
Yan Luo, NIST, Gaithersburg, USA
Jose Machado, University of Minho, Braga, Portugal

INTERNATIONAL PROGRAMME COMMITTEE

Physics Modelling and Cosmological Simulation
Philippe Geril, ETI BV, Ostend, Belgium

Cyber-Physical System Modelling
Eva Catarina Gomes Maia, Instituto Superior de Engenharia do Porto, Porto, Portugal
Frank Oppenheimer, OFFIS e.V., Oldenburg, Germany
Isabel Praça, Instituto Superior de Engenharia do Porto, Porto, Portugal

Cross Fertilisation between Simulation and Formal Methods
Silvano Dal Zilio, LAAS, Toulouse, France
Marc Pantel, INP Toulouse, Toulouse, France

Virtual Prototyping
Alexandre Nketsa, LAAS, Toulouse, France

Simulation in Energy and Power Systems
Janos-Sebestyen Janosy, KFKI Atomic Energy Research Institute, Budapest, Hungary

Renewable Energy Technologies
Cesar Garcia-Garcia, Universidad de Guadelajara, Guadelajara, Mexico

Simulation in Engineering Processes
Saad Mahmood Ali, Biomedical Engineering Department, University of Technology, Baghdad, Iraq
Chrissanti Angeli,Technological Institute of Piraeus, Athens, Greece
Ali Shams Nateri, University of Guilan, Rasht, Iran
Jan Studzinski, Polish Academy of Sciences, Warsaw, Poland

Simulation in Model Driven Engineering
Souvik Barat, Tata Consultancy Services, Pune, India
Cesar Garcia-Garcia, Universidad de Guadelajara, Guadelajara, Mexico
Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Castelo Branco, Portugal
Izabella V. Lokshina, SUNY Oneonta, Oneonta NY, USA
Frederic Mallet, Université Côte d'Azur, Cnrs, Inria, France
Fernando Mas Morate, University of Sevilla, Seville, Spain
Algirdas Pakštas, London Metropolitan University, London, United Kingdom

Simulation-based Evaluation of Interactive Systems
Cesar Garcia-Garcia, Universidad de Guadelajara, Guadelajara, Mexico
Andrzej Najgebauer, Military University of Technology, Warsaw, Poland
Algirdas Pakštas, London Metropolitan University, London, United Kingdom

Simulation and Modelling for Humanitarian/Emergency Operations
Mehdi Benhassine, Royal Military Academy, Brussels, Belgium
Cesar Garcia-Garcia, Universidad de Guadelajara, Guadelajara, Mexico
Algirdas Pakštas, London Metropolitan University, London, United Kingdom

Simulation in Hospital Logistics
Joel Colloc, Le Havre Normandy University, Le Havre, France
Jose Machado, University of Minho, Braga, Portugal
Jose Antonio V. Oliveira, University of Minho, Braga, Portugal

Simulation in Logistics
Remy Dupas, Université de Bordeaux , Bordeaux, France
Olivier Grunder, UTBM, Belfort, France
Marie-Ange Manier, UTBM, Belfort, France
Rosaldo Rossetti, University of Porto, Porto, Portugal
Pengjun Zheng, Ningbo University, Zhejiang, China P.R.

Supply Chain Simulation
Florina Covaci,"Babes-Bolyai" University, Cluj-Napoca, Romania
Eleni Mangina, University College Dublin (UCD), Dublin, Ireland

INTERNATIONAL PROGRAMME COMMITTEE

Intelligent Systems
Ying He, De Montfort University, Leicester, United Kingdom
José Machado, Universidade do Minho, Braga, Portugal
Manuel Filipe Santos, Universidade do Minho, Guimarães, Portugal

Real-Time GPS Simulation and Service Applications
Marwan Al-Akaidi, Chair SPC-IEEE UK&Ireland

Simulation with Petri Nets
Pascal Berruet, Universite Bretagne Sud, Lorient, France
Stefano Marrone, Seconda Universita degli Studi di Napoli, Naples, Italy
Alexandre Nketsa, LAAS-CNRS, Toulouse, France

Bond Graphs Simulation
Jesus Felez, Univ. Politecnica de Madrid, Spain
Andre Tavernier, BioSim, Brussels, Belgium

DEVS
Fernando Tricas, Universidad de Zaragoza, Spain

Fluid Flow Simulation
H. A. Nour Eldin, University of Wuppertal, Germany
Markus Fiedler, Blekinge Institute of Technology, Sweden

Emergency Risk Management Simulation
Mehdi Benhassine, Royal Military Academy, Brussels, Belgium
Joseph M. Saur, Georgia Tech Research Institute, Atlanta, USA,
Lode Vermeersch, Credendo, Brussels, Belgium

Learning Agents and Co-Simulation for CPES
Track Chair
Eric Veith, University of Oldenburg, Oldenburg, Germany

Stephan Balduin, Offis, Oldenburg, Germany
Alexandro Steinert, Offis, Oldenburg, Germany
Lasse Hammer, Offis, Oldenburg, Germany
Sharaf Aldin Alsharif, Offis, Oldenburg, Germany
Jan Sören Schwarz, Offis, Oldenburg, Germany
Eike Schulte, Offis, Oldenburg, Germany

Smart Cities Simulation
Track Chair
Istvan David, McMaster University, Hamilton, ON, Canada
Xinquan Chen, Anhui Polytechnic University, Wuhu, Anhui, China
Dahlan Abdul Ghani, Universiti Kuala Lumpur, Kuala Lumpur, Malyasia
Christina G. Georgantopoulou, Bahrain Polytechnic, Isa Town, Kingdom of Bahrain
Abderrazak Jemai, INSAT, Ecole Polytechnique de Tunis, Tunis, Tunisia
Eugen Pop, National University of Science and Technology POLITEHNICA of Bucharest, Romania

All articles published in these Proceedings have been peer reviewed.

EUROSIS-ETI Publications are WoS, IET and the German National Library of Science and Technology referenced.
ESM® Proceedings are indexed on WoS, IET, SCOPUS and Elsevier Engineering Village.

A CIP Catalogue record for this book is available from the Royal Library of Belgium under nr.12620 – (D/2011/12.620/1).

For permission to publish a complete paper write EUROSIS, c/o Philippe Geril, ETI Executive Director, Torhoutsesteenweg 162 Box 04.02., B-8400 Ostend, Belgium.

EUROSIS is a Division of ETI Bv, The European Technology Institute, Torhoutsesteenweg 162, Box 04.02, B-8400 Ostend, Belgium

Printed in Belgium by Reproduct NV, Ghent, Belgium
Cover Design by C-Design Bvba, Wielsbeke, Belgium
Cover pictures of San Sebastian are licensed under the Creative Commons Attribution-Share.
Publisher reference number: 9789492859

ESM® is a European registered trademark of the **European Technology Institute** under nr: 002433290
ESM® is a UK registered trademark of the **European Technology Institute** under nr: UK00902433290

EUROSIS-ETI Publication

ISBN: 978-9-492-859-33-4
EAN: 9789492859334

DATA-DRIVEN MODEL FOR CHRONIC KIDNEY DISEASE PROGRESSION: A WORK IN PROGRESS

Candelaria Alvarez, Remo Suppi
Universidad Autónoma de Barcelona, Spain
email: AnaCandelaria.Alvarez@uab.cat

Jose Ibeas
Interventional and Computational Nephrology, I3PT
Parc Tauli University Hospital, Spain

Javier Balladini
Universidad Nacional del Comahue, Argentina

KEYWORDS

Data-driven model, simulation, chronic kidney disease.

ABSTRACT

In medicine, data-driven models can be used to simulate disease progression and generate clinical decision support systems (CDSS). While artificial intelligence (AI) and machine learning (ML) models are common, their lack of traceability poses challenges in medical contexts, where transparency is crucial. This study aims to create a traceable data-driven model for Chronic Kidney Disease (CKD) progression without using AI or ML. The study will develop and validate a simulator based on this model with real-world data. This paper describes the development of the model, the processing of CKD patient data, the implementation of the simulator, and the validation of results.

INTRODUCTION

In medicine, data-driven models can simulate disease progression, supporting the development of clinical decision support systems (CDSS). These models, when developed using artificial intelligence (AI) or machine learning (ML) techniques Solomatine et al. (2008), often lack traceability, making it challenging to understand their internal processes and decision-making. This is problematic, as the ability to interpret and justify model outputs is crucial for clinical decision-making and gaining the trust of healthcare professionals.

Numerous studies in medicine use data-driven models, such as Frisch et al. (2021). However, there is a notable gap in research on modelling Chronic Kidney Disease (CKD), a growing public health issue defined by kidney damage lasting three or more months. CKD is assessed using glomerular filtration rate (GFR) and albuminuria and is classified into six stages based on GFR ranges (measured in mL/min/1.73 m2): G1 (> 90), G2 (60-89), G3a (45-59), G3b (30-44), G4 (15-29), and G5 (<15). The GFR value is typically estimated using equations; hence, it is referred to as the estimated GFR (eGFR). These equations are defined separately for men and women, and they use the individual's age in years and creatinine level in mg/dL KDIGO (2012).

The main aim of this study is to design a data-driven model for CKD progression with full traceability, avoiding AI or ML techniques. This article details the definition of the model using statistical and data analysis methods, the adaptation of a dataset for predictions, the development of a simulator based on this model using real-world data, and the validation of results with an independent data subset. Finally, the results obtained will be discussed.

Data-driven model for CKD progression

The proposed data-driven model predicts CKD progression in new patients using historical data. It includes variables such as eGFR, CKD stage (G1-G5), age, sex, and laboratory parameters from blood tests, including creatinine, albumin, sodium, potassium, glucose, urea nitrogen, platelet count, and alkaline phosphatase. To reflect the progression of the disease, the model considers that patients may have multiple records at each stage corresponding to different blood tests, and also in different stages. This initial model definition does not incorporate the time variable. This is because the primary objective is to determine the values of the variables at each stage of the disease, temporarily setting aside the rate at which it progresses. Future expansions of the model will include additional information such as progression rate and medications.

The primary feature of the model lies in its application of data analysis and statistics to the dataset for predicting the progression of CKD in a new patient. This approach ensures that the obtained results are traceable and explainable. The procedure involves determining the patient's stage and subsequently predicting their progression through subsequent stages. This second part is conducted following a series of tasks. In this explanation, a patient P in stage G2 is considered as the input, and its progression to G5 is regarded as output. Figure 1 shows the process for predicting the progression from G2 to G3a. The words in blue indicate the parameters that can be modified by the user, and they are: *similarity criteria* (Euclidean or Cosine), *similarity value* (value $\in R$), *similarity variables* (list of variables), *age window filter* (value $\in R$), *new age option* (CDF filtered or Similar) and *new age filter* (value $\in R$). The words in purple are the new variable values for the patient in G3a.

The first task obtains records of G2 patients similar to P using the *similarity criteria* (cosine similarity or Euclidean distance) and the *similarity value* (percentage or distance) applied to the variables specified by *similarity variables.* The internal process of this task compares P with a subset of real G2 records that have been previously filtered by P's sex and its age plus the *age window filter* parameter. For cosine similarity, a record is included if its comparison value meets or exceeds the specified percentage. For Euclidean distance, a record is included if the comparison value is less than or equal to the specified distance. If this set, known as the P window in G2, is empty, progression cannot be estimated.

Once the records within the window have been computed, their real progression towards G3a is obtained. If no records show progression to G3a, the procedure cannot continue. In such cases, an attempt should be made to estimate progression to another stage like G3b or G4 from the same window of P in G2. If records are available, the age for P in G3a is determined according to the *new age option* parameter. If 'Similar' is selected, the age for P in G3a matches the age of the most similar patient in the window who progressed to G3a. If 'CDF filtered' is chosen, the cumulative distribution function (CDF) table t f is calculated for the age values in the progression records of real patients in G3a. Next, a uniformly distributed random decimal r between 0 and 1 is generated. Then the variable value v from table t whose CDF value is the closest to r is used as the value of AGE_YEARS.

Subsequently, data from the G3a subset is filtered using the age calculated plus the *new age filter* parameter. Laboratory variables are then calculated from this filtered data, using the same procedure of CDF table and random value. Finally, the eGFR value is obtained and checked to ensure it falls within the defined range for G3a. If affirmative, the progression of P from G2 to G3a is finished. If negative, no progression to G3a is achieved. This scenario may occur if there are limited progression records, resulting in age and creatinine values that, when applied to eGFR equations with patient sex, yield out-of-range results. At this point, the process of generating new variables could be repeated, but we have chosen not to do so.

Following this, the entire procedure is repeated based on the data calculated for P in G3a: calculating its window of similar real patients in G3a, obtaining the evolution of these records towards G3b, and calculating the values of the new variables for P in G3b using the evolution data. By repeating this sequence, the progression of P towards G4 and subsequently towards G5 can be estimated.

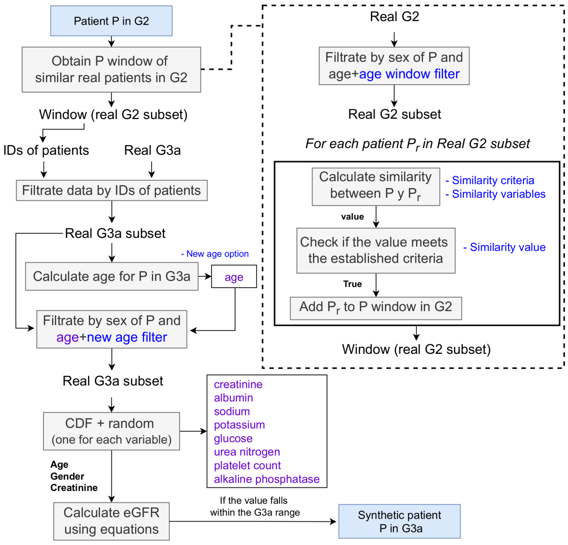
Patient P in G2
Obtain P window of similar real patients in G2
Window (real G2 subset)
IDs of patients
Real G3a
Filtrate data by IDs of patients
Real G3a subset
- New age option
Calculate age for P in G3a
age
Filtrate by sex of P and age+new age filter
Real G3a subset
CDF + random (one for each variable)
creatinine
albumin
sodium
potassium
glucose
urea nitrogen
platelet count
alkaline phosphatase
Age
Gender
Creatinine
Calculate eGFR using equations
If the value falls within the G3a range
Synthetic patient P in G3a
Real G2
Filtrate by sex of P and age+age window filter
Real G2 subset
For each patient P_r in Real G2 subset
Calculate similarity between P y P_r
- Similarity criteria
- Similarity variables
value
Check if the value meets the established criteria
- Similarity value
True
Add P_r to P window in G2
Window (real G2 subset)

Figure 1: Process for predicting CKD progression.

Data preprocessing

This study uses the MIMIC-III 1.4 dataset Johnson et al. (2016), which comprises clinical data for over forty thousand patients admitted to critical care units at Beth Israel Deaconess Medical Center (Boston, USA) from 2001 to 2012. Specifically, the selected data includes general patient information, details of admissions and results of laboratory tests conducted during each stay. Before using this information in the simulation process, it was necessary to perform several tasks in the dataset: conditioning, filtering and reduction.

The **conditioning** process aimed to reorganise the tables into a comprehensive format, with each row containing complete information about the patient, the admission, and one laboratory result. Thus, for each patient admission, there will be as many rows as laboratory tests conducted. In order to do this, the tables for patients, admissions, diagnoses, and laboratory results were processed and merged. Finally, two new columns were added: one for eGFR, calculated using the equations mentioned previously, and another for the CKD stage. The **filtering** task involved selecting rows relevant to CKD patients by using 28 diagnostic codes for CKD and related conditions, along with 69 laboratory item codes from blood and urine, corresponding to the analytical variables defined in the model. The conditioned data were filtered by these codes, resulting in a table of CKD patients and their relevant laboratory parameters. Lastly, during the **reduction** task, the 69 columns of laboratory items were reduced: for each analytical variable defined in the model, the column with the greatest amount of data was selected.

In summary, each row in the final table contains the results of one blood test for a CKD patient during a hospital admission: PATIENT_ID, HADM_ID (Admission identifier), eGFR (Estimated GFR, mL/min), AGE-YEARS, FEMALE (Female (1) or Male (0)), 50912 (Creatinine in mg/dL), 50862 (Albumin in mg/dL), 50863 (Alkaline phosphatase in IU/L), 50931 (Glucose in mg/dL), 51006 (Urea nitrogen in mg/dL), 51265 (Platelets in K/uL), 50983 (Sodium in mEq/L), 50971 (Potassium in mEq/L). The final table contains 6273 rows for 591 patients in G1, 15035 rows for 1702 patients in G2, 17977 rows for 2440 patients in G3a, 27084 rows for 3033 patients in G3b, 35783 rows for 3094 patients in G4 and 36873 rows for 2178 patients in G5.

Validation

The simulator based on the model was implemented in Python. The validation process was configured using two new parameters: *validation stages* and *validation percentage*. The former specifies the stages (e.g., G2 to G5) for patient data used in validation. From the pool of identifiers meeting this criteria, a random sample is selected based on the *validation percentage*. This sample is used for the validation, and the other is used for the data-driven model in the simulator. The parameters for the validation process were the following: Euclidean (similarity criteria), 0.5 (similarity value), [eGFR, Creatinine, Female, Age] (similarity variables), 0.5 (age window filter), CDF filtered (new age option), 0.5 (new age filter), 20% (validation percentage) and G2-G5 (validation stages). Patients in G1 were not considered for validation as they exhibit greater disparities due to the absence of an upper limit for the eGFR value.

Firstly, a simulation was carried out for each patient from G2 to G5. Subsequently, simulations were performed starting from other states of the validation records, in this case G3a, G3b, and G4. These results will enable an analysis of the accuracy of simulations starting from real patient data in more advanced stages, thus requiring fewer simulation steps. With this configuration, 2,975 rows corresponding to 45 different patient identifiers were used as validation data. The validation process resulted in a table with a total of 3,950 rows; this value is greater than the number of validation rows because there are simulations starting from different initial stages for the same patient. In the following subsections, the analysis of the results is explained.

Patient results

An analysis was performed to verify the performance of the model and simulator by visualising the differences between the real and simulated mean eGFR for an individual patient. The results obtained for the combinations of each CKD stage and initial stage were analysed to determine the influence of successive simulation steps. eGFR was chosen for this analysis because it is the principal indicator of CKD progression.

For the initial analysis, the mean eGFR was calculated for the validation data grouped by patient and CKD stage. Subsequently, the mean eGFR was calculated for the simulation results grouped by patient, CKD stage, and initial stage. Figure 2 illustrates these results for a specific patient, where the blue bars depict the real progression, and the other bars show the simulated progression for each initial stage. The graphic begins at G3a because the first simulated result is derived from G2 to G3a. In this case, it is observed that the simulated value in G3a (starting from G2) is 3.4 units higher than the real value. Moving to G3b, there are two simulated values: in orange from the initial stage G2, and in green from the initial stage G3a. Both are close to the real value, with differences of +1.17 and -1.65, respectively. In G4, the simulated values from G2 and G3b show very small differences from the real value (-0.52 and +0.11), while the green value is slightly higher (+1.66). Finally, in G5, the orange value from simulation starting from G2 closely matches the real value (+0.1502), whereas the others differ by approximately -2.5 units.

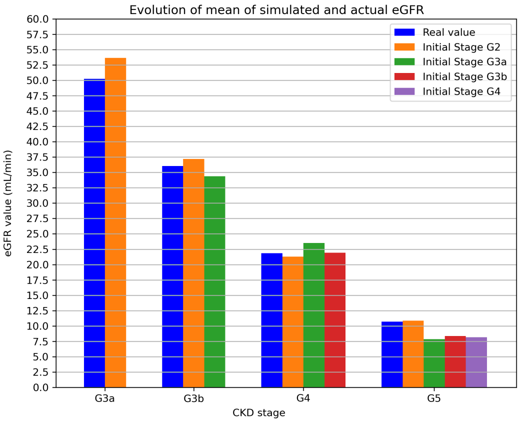

Figure 2: Simulated and real mean of eGFR for one patient.

For the second analysis, the results were grouped according to patient identifier and CKD stage (G3b, G4 or G5). For each group containing more than one row, the initial state that generated the smallest difference between the mean simulated eGFR and the mean real eGFR was identified and counted. These results are presented in the third column in Table 1. It is noteworthy that in no case did starting from the previous state yield the best results. In the case of simulations in G5, approximately 60% of the best results were obtained from simulations that began in G2 and G3a (involving 4 and 3 simulation steps, respectively).

General results

To assess the quality of predictions overall, metrics based on the eGFR variable were calculated. Simulated eGFR values served as the estimated values, while the mean eGFR of the validation records (computed in the

Table 1: Frequency of the smallest difference between the mean simulated eGFR and the mean real eGFR, grouped by CKD stage and the initial stage.

CKD stage	Initial stage	Frequency
G3b	G2	26 (60.46%)
	G3a	17 (39.54%)
G4	G2	11 (25%)
	G3a	19 (43.18%)
	G3b	14 (31.82%)
G5	G2	12 (27.27%)
	G3a	15 (34.09%)
	G3b	11 (25%)
	G4	6 (13.64%)

Table 2: Metrics for the comparison between real and simulated eGFR.

CKD stage	Initial stage	MSE	RMSE	MAE
G3a	G2	15.93	3.99	3.32
G3b	G2	18.22	4.27	3.61
	G3a	21.49	4.64	3.89
G4	G2	20.24	4.5	3.84
	G3a	18.07	4.25	3.61
	G3b	17.64	4.2	3.53
G5	G2	18.65	4.32	3.51
	G3a	16.75	4.09	3.28
	G3b	15.51	3.94	3.21
	G4	19.2	4.38	3.58

previous analysis) served as the actual values. The comparison was conducted for each initial stage, and the data were grouped by CKD stage. The results obtained are presented in Table 2. The values in columns MSE (mean squared error), RMSE (root mean square error) and MAE (mean absolute error) are expressed in the unit of measure of eGFR (mL/min).

In general terms, the average RMSE is 4.258 and the average MAE is 3.538. This indicates that simulated eGFR values differ from real values by between 3.5 and 4.3 mL/min. In each row, the RMSE and MAE metrics show a slight variation, differing by less than 1 unit. Specifically, MAE consistently yields lower values as it penalises errors less severely. However, given the minimal discrepancy between RMSE and MAE, it can be concluded that there are no significant discrepancies between real and simulated values. Within each CKD stage, we observe slight variations in RMSE and MAE values depending on the initial stage. Similar to patient-level analysis, it is notable that in cases like G5, conducting more simulation steps (from G2, G3a, or G3b) results in smaller errors compared to starting from the previous stage (G4).

CONCLUSION AND FUTURE WORK

After conducting this research, it can be concluded that the proposed objectives have been met. An initial data-driven model for CKD progression was defined. Then, a dataset of real clinical data was adapted for this work. A simulator based on this model and dataset was developed, and simulations validated the model. Finally, the quality of predictions was assessed against real data. It is crucial to highlight the importance of providing explainable tools in the medical field. Traceability remains a significant advantage of the data-driven model over other ML or AI techniques.

After the validation, it was observed that the best results were not obtained in a single simulation step. This finding is particularly significant, as it might initially be assumed that starting from the previous state would provide the most accurate results, involving a single simulation step based on the real data of the validation patient. Therefore, it is concluded that the methodology presented improves result quality by utilising the patient's evolution history and similar real patients.

Future work will focus on analysing correlations among the variables defining the patient, with the aim of incorporating any detected correlations into the model's predictions. Integrating the time variable into both the model and simulator is also a key objective. Once this is achieved, the accuracy of the new results will be compared to those from a temporal series algorithm, such as LSTM. Long-term goals include validating the model with medical professionals and implementing a Clinical Decision Support System (CDSS) to offer treatment recommendations for new CKD patients.

ACKNOWLEDGEMENTS

This publication is supported under contract PID2020-112496GB-I00/ AEI / 10.13039/501100011033

REFERENCES

Frisch, H., Sprau, A., McElroy, V., Turner, J., Becher, L., Nevala, W., Leontovich, A., and Markovic, S. (2021). Cancer immune control dynamics: a clinical data driven model of systemic immunity in patients with metastatic melanoma. *BMC Bioinformatics*, 22.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035.

KDIGO (2012). Clinical practice guideline for the evaluation and management of chronic kidney disease.

Solomatine, D., See, L., and Abrahart, R. (2008). *Data-Driven Modelling: Concepts, Approaches and Experiences*, pages 17–30. Springer Berlin Heidelberg, Berlin, Heidelberg.