Clustering Trajectories to Study Diabetic Kidney Disease

Veronica Distefano[1,2], Maria Mannone[1,3,4(✉)], Irene Poli[1], and Gert Mayer[5]

[1] European Centre for Living Technology (ECLT), Ca' Foscari University of Venice, Venice, Italy
{veronica.distefano,maria.mannone,irenpoli}@unive.it
[2] Department of Economic Sciences, Università del Salento, Lecce, Italy
[3] Department of Engineering, University of Palermo, Palermo, Italy
[4] Dipartimento di Scienze Molecolari e Nanosistemi (DSMN), Ca' Foscari University of Venice, Venice, Italy
[5] Department Internal Medicine IV (Nephrology and Hypertension), Innsbruck Medical University, Innsbruck, Austria
gert.mayer@i-med.ac.at

Abstract. Diabetic kidney disease (DKD) is a serious complication of type-2 diabetes, defined prominently by a reduction in estimated glomerular filtration rate (eGFR), a measure of renal waste excretion capacity. However DKD patients present high heterogeneity in disease trajectory and response to treatment, making the *one-model-fits-all* protocol for estimating prognosis and expected response to therapy as proposed by guidelines obsolete. As a solution, precision or stratified medicine aims to define subgroups of patients with similar pathophysiology and response to the therapy, allowing to select the best drug combinations for each subgroup. We focus on eGFR when aiming to identify eGFR decline trends by clustering patients according to their eGFR trajectory shape-similarity.

The study involved 256 DKD patients observed annually for four years. Using the Fréchet distance, we built clusters of patients according to the similarity of their eGFR trajectories to identify distinct clusters. We formalized the trajectory-clustering approach through category theory. Characteristics of patients within different progression clusters were compared at the baseline and over time.

We identified five clusters of eGFR progression over time. We noticed a bifurcation of eGFR mean trajectories and a switch between two other mean trajectories. This particular clustering approach identified different mean eGFR trajectories. Our findings suggest the existence of distinct dynamical behaviors in the disease progression.

Keywords: clustering · trajectory · precision medicine · category

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-57430-6_21.

M. Villani et al. (Eds.): WIVACE 2023, CCIS 1977, pp. 271–283, 2024.
https://doi.org/10.1007/978-3-031-57430-6_21

1 Introduction

Precision medicine [1,2] is a flourishing research area, which aims to find the best individualized treatment for patients according to their characteristics. In fact, the formula "one-model-fits-all" is unsatisfying when it comes to many diseases as far as progression and response to therapy is concerned. To find subgroups of similar patients, cluster analysis approach is a useful and informative tool, as witnessed by several studies [3–9].

Defining sub-groups of such an evolving population can help shed light on underlying common features in each sub-group, allowing physicians, if linked to pathophysiology and drug mode of action, to foster a more appropriate targeted treatment. This approach to medical research paves the way toward effective personalized or at least better stratified treatment. This approach to medical research paves the way toward effective individualized treatments. Of particular relevance is, for instance, the differentiation at the baseline, regarding different parameters. We focus on clusters of patients sharing the same disease-behavior across time, as instances of longitudinal studies. Longitudinal studies have been used also to address more general quality of life issues [10] and depression patterns across time [11], with statistical approaches such as growth mixture models.

In this article, we focus on patients with type-2 diabetes mellitus (T2DM) and its associated diabetic kidney disease (DKD) from the DC-ren dataset.[1] DKD is a serious public health problem and the main cause of end-stage renal disease (ESRD) in developed countries [26]. Longitudinal changes of renal function help inform on patients' clinical courses and if, identified by pathophysiologically relevant characteristics, help select individualized treatment according to patients' specific characteristics.

In this article, we will build clusters of patients' trajectories. This information can constitute a first step toward the development of a decision system to foster individualized strategies for DKD treatment [4,12]. We analyze trajectories of patients with respect to the dependent variable eGFR. The variation of eGFR provides an estimate of the severity of the disease and the response to treatment [8,16]. We build clusters of trajectories based on shape similarity and on eGFR range-similarity.

To group trajectories according to their shape similarity, we use the Fréchet distance, first proposed in the domain of calculus [13], and recently applied to medicine with the *kmlShape* clustering technique [14]. The Fréchet distance is evaluated upon the comparison between pairs of points following the profiles of the curves they belong to.

The approach to trajectory clustering is formalized within the framework of Category Theory [17,18]. It is an abstract branch of mathematics, initially developed to formalize the transformations between transformations, and to connect

[1] The project *Drug combinations for rewriting trajectories of renal pathologies in type II diabetes* (DC-ren), https://dc-ren.eu/, is funded by the Horizon 2020 research and innovation programme, Action RIA Research and Innovation action Call: H2020-SC1-BHC-2018-2020; Topic: SC1-BHC-02-2019.

different areas of mathematics between them. Applied category theory includes research in physics [19], chemistry [20], neuroscience [21]. A few applications also concern cluster analysis [22], with the formalization of a clustering method as a *functor*. A functor is a morphism between categories. A category is constituted by objects (points) and morphisms between them (arrows), whose composition is associative and has the identity element.

We aim to find subgroups of similar patients and build clusters of mean trajectories. We find cases of bifurcations and switch of trajectory clusters. To understand the possible pathophysiological reasons underlying patients exhibiting such a behavior, we analyze their medical and demographical variables. Coupled with drug mode of action, our results can be fed into a decision system, to find the best individualized treatments for future DKD patients. This article is the development of a first study where the Fréchet distance was applied to real data [23]. Here, we consider an extended dataset and a more refined computational approach. The novelty of our work is the use of categorical formalism for a medical real case study, and the application of a relatively-new statistical method, *kmlShape*, to a real data for a non-public dataset.

The article is organized as follows. After a review of some concepts of longitudinal cluster analysis and category theory (Sect. 2), we present a case of study with patients affected by DKD (Sect. 3), and we discuss our findings (Sect. 4).

2 Methodology

In this section, we present our trajectory, clustering approach using some formal tools of category theory; we then describe the *kmlShape* method, to investigate trajectory shape-similarity according to the Fréchet distance.

2.1 A Shape-Similarity Clustering of Longitudinal Data

Longitudinal data are measured repeatedly over time for the same individual. In this paper, we are interested in the evolution regarding the individual variation of estimated glomerular filtration rate (eGFR) in a small group of patients with type 2 diabetes and chronic kidney disease (DKD) at different stages. We used the *kmlShape* approach, that creates clusters of trajectories according to their evolution [14]. This approach is a variation of the longitudinal k-means [24] using a "shape-respecting distance" and a "shape-respecting mean."

The Fréchet distance [27] computes the shape similarity of two curves P_1 and P_2, based on the smallest of the maximum pairwise distances obtained with two respective reparametrizations, $\alpha : [0,1] \to [0,1]$ and $\beta : [0,1] \to [0,1]$, as follows:

$$F(P_1, P_2) = \inf_{\alpha,\beta \in R} \max_{t \in [0,1]} \{dist(P_1(\alpha(t)), P_2(\beta(t)))\}.$$

The approach of *kmlShape* considers the discrete version of the Fréchet distance, based on a sequence of pairs of points belonging to the two curves (represented as polylines). Since the two curves need not to have the same length, we have to

"walk through them" at different speeds. The ratio between the different speeds to move along the curves is the time scale λ, discussed later.

Since we are interested in assessing the trend of the disease rather than verifying its presence, we focused on this method. The same approach has recently been followed in another application of *kmlShape* to a medical dataset [27]. In fact, *kmlShape* quantifies the differences of trend between the eGFR trajectories. In addition, the *kmlShape* method presents a highest ARI index when compared with Traj and GMM method [28].

The Fréchet distance measures the longest link between the trajectories [14]. Its computation between two trajectories does not require the same number of time-points in each trajectory.

We consider a generalization of the Fréchet mean to n curves. To this aim, we implement the *kmlShape* with the *RandomAll* technique [14], with n patients randomly scattered through the leaves of a binary tree.

Genolini and co-authors [14] provided a generalized definition of the Fréchet distance including a time scale λ. Indeed, in the context of real data, there can be an issue of relative scale, because the variable of interest and the time variable are measured according to different unities. The change of time scale impacts the partitioning, and thus the resulting clusters. The meaning of the scale variation is the change of "travel speed" to go through a curve. The value of $\lambda = 0.5$ is empirically determined for each research problem. We run different tests before choosing this value. More details including the precise definition of the Fréchet's mean can be found in the article by Genolini and co-authors [14].

2.2 Category Theory for Trajectory Clustering

Patients with similar characteristics over time can be computationally and graphically grouped together in the same cluster [15]. The comparison between processes over time can be contextualized in the framework of category theory [25]. Here, we use its diagrammatic language to describe patients' grouping according to their trajectory similarity. We also discuss the transition from a patient-based representation to a state-based representation. First, we briefly summarize the basic definitions of category theory.

A *category* is constituted by objects (points) and morphisms (arrows) between them. The composition of morphisms must be associative, and there should exist the identity morphism. A *functor* is a generalization of a function. More precisely, it is a mapping between categories (mapping objects and morphisms of a category into objects and morphisms of another category, preserving structures), and a *natural transformation* is a mapping between functors.

According to Spivak [20], category theory constitutes a powerful (i.e., precise) communication tool of ideas tool between different fields of mathematics. It can be used to compare structures and methods of different disciplines. Category theory starts being applied to several domains of science to acquire an abstract and thus general overview [20]. According to [29], category theory can also be used for medical dataset. Here, we use category theory as a bridge between clinical practice as defined by physicians, real data of patients, and information theory. We

use this formalism to make more precise the comparison between each patient at different time-points, and between different patients at the same time-point. In addition, connecting the case study with the categorical framework allows one to recover all theorems and methods defined in abstract mathematics, which have the potential to make possible further applications and developments.

Let us consider a dataset composed of n patients characterized by p observable variables at four time points t_0, t_1, t_2, t_3. Each patient is characterized as a triplet $(\mathbf{x}_i(t_k), \mathbf{D}(t_k), y_i(t_{k+1}))$, where i is the individual (the patient); t_k is the time point $k = 0, 1, 2, 3$; $\mathbf{x}_i(t_k) = x_i^1(t_k), ..., x_i^p(t_k)$ is a set of values of variables, characterizing the individual; $\mathbf{D}(t_k) = D_1(t_k), D_2(t_k), D_3(t_k), D_4(t_k)$ stands for the given drug combination; $y_i(t_{k+1})$ is the value of the response variable Y at t_{k+1}, measured after one year of treatment. The response variable is evaluated as the variation of the dependent variable, that is, the estimated glomerular filtration rate (eGFR); we thus indicate it as E in the following. The trajectory over time of the i-th patient (p_i) with respect to the eGFR (E) is: $p_i^E(t_0) \rightarrow p_i^E(t_1) \rightarrow p_i^E(t_2)$. For the i'-th patient we have: $p_{i'}^E(t_0) \rightarrow p_{i'}^E(t_1) \rightarrow p_{i'}^E(t_2)$. We can evaluate the distance of a patient with respect of herself/himself through time, or the distance between different patients at the same time. We indicate the distance between patients i, i' with respect to the variable E and time t_k as $d_{i,i'}^E(t_k)$, and the distance between values observed at times $t_k, t_{k'}$ of the variable E for the same patient i as $d_i^E(t_k, t_{k'})$, see diagram (1). In such a patient-based representation, each point is a patient at a time-point. This representation is dual to the state-based representation, which will be useful to create the *state map* (Fig. 1).

$$\begin{array}{ccc} p_i^E(t_0) & \xrightarrow{d_{ii'}^E(t_0,t_0)} & p_{i'}^E(t_0) \\ \downarrow{\scriptstyle d_{ii}^E(t_0,t_1)} & & \downarrow{\scriptstyle d_{i'i'}^E(t_0,t_1)} \\ p_i^E(t_1) & \xrightarrow{d_{ii'}^E(t_1,t_1)} & p_{i'}^E(t_1) \\ \downarrow{\scriptstyle d_{ii}^E(t_1,t_2)} & & \downarrow{\scriptstyle d_{i'i'}^E(t_1,t_2)} \\ p_i^E(t_2) & \xrightarrow{d_{ii'}^E(t_2,t_2)} & p_{i'}^E(t_2) \end{array} \tag{1}$$

In the language of categories, the construction of diagram (1), with observations and distances, can be described as an enriched *double category* with metrics in $\mathbb{R}$ [18], whose objects are the values of variable E, and whose morphisms are vertical and horizontal distances $d_{i,i'}^E(t_k)$, $d_i^E(t_k, t_{k'})$. The comparison between trajectories of different patients involves both of these distances.

Similar trajectories can be grouped within the same cluster of trajectories. The clustering is described as a functor [22] from the category of dataset to the category of the partitioned dataset. This concept can be applied to the trajectory-clustering, from the category of trajectories to the category of clusters of trajectories (Fig. 2). In the language of categories, the comparison between similar clustering methods corresponds to an arrow between arrows, that is, a

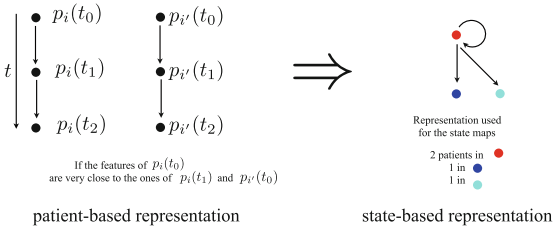

Fig. 1. Patient-based representation and state-based representation. The representation on the left is typical of categories. $p_i(t)$ indicates the clinical characteristics of the i-th patient at time t, and $p_{i'}(t)$ refers to the i'-th patient. Time flows vertically. The representation on the right neglects the detail on single-patient and time, in favor of a description of the clinical states where one or more patients can stay or return (loop arrow). The second representation can be built from the overall analysis of single-patients longitudinal data.

natural transformation. The comparison between the clusters that are obtained with slightly different methods is formalized as an arrow (morphism) in the category of clusters of trajectories. Thus, one can shift the attention from the natural transformation (comparison between clustering methods) to a morphism (comparison of clusters obtained with slightly different methods). Natural transformations (arrows between arrows) formalize the comparison between different transformations. Trajectory clustering processes, despite their differences, can be seen as processes from trajectories to clusters of trajectories, and thus we can use the language of categories to compare them.

2.3 Study Population

We considered 256 DKD patients observed during annual visits in a time span of four years. 48.4% were male, the mean age was 67 years. The characteristics of patients at the baseline are presented in Table 1.

The variability in eGFR decline was analyzed with cluster analysis. The eGFR is defined in the Appendix. In clinical routine the eGFR trajectory is used to judge the response to the therapeutic treatment: the *controlled disease* corresponds to an increase of eGFR or a decrease not exceeding 5% of the baseline value (the value at t_0), while the *uncontrolled disease* corresponds to an eGFR decrease higher than 10% of the baseline value [30].

The mean eGFR at t_0 ranges from 31 and 90 ml/min/1.73 m^2; at t_3 it is comprised between 19 and 120 ml/min/1.73 m^2, denoting an overall decrease of kidney efficiency through time. The descriptive statistics of eGFR, with mean and standard deviation at each time-point, are presented in Table 2. The mean value of eGFR decreases with time, indicating an overall worsening of the disease in the group.

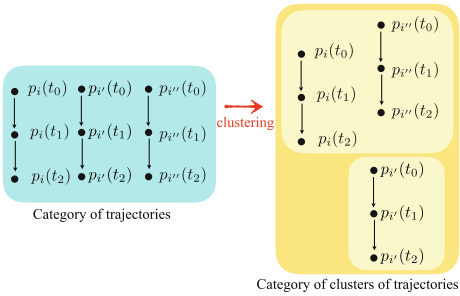

Fig. 2. Clustering as a functor from the category of trajectories to the category of clusters of trajectories. In category theory, a functor is a generalization of a function, mapping point and arrows from a category to another one. Here, we consider a mapping from the category of trajectories to the category of clusters of trajectories. The points are the patients at given time points, and the arrows are the comparisons of their clinical values. Trajectories are given by the comparisons of patients with themselves over time. In the second category, we group patients presenting similar trajectories inside the same cluster.

We derive the profile of patients of this population considering a set of the most relevant variables describing their characteristics. The variables are measured at the baseline (time t_0) and at three follow-ups (t_1, t_2, t_3). At t_0 (Table 1), the 256 patients have a mean eGFR of 64 ± 16. Their mean values of systolic blood pressure and diastolic blood pressure are, respectively, 138 ± 16 and 78 ± 10 mmHg, and serum triglycerides (172 ± 106 mg/dl); these values are moderately high. The mean values of blood glucose (144 ± 46 mg/dl) and HbA1c ($7.2 \pm 1.2\%$) are also elevated. The mean values of total cholesterol (181 ± 44 mg/dl) and serum potassium (4.5 ± 0.5 mmol/l) are in the normal range. The mean value of UACR is moderately elevated (78.94 ± 283.85 mg/g Creatinine). The large standard deviation takes into account the great variability of UACR values across patients. In the following, we examine trajectory clusters obtained with the *kml-Shape* method.

Table 1. The mean values and their standard deviations for the 256 patients at the baseline. SBP is the systolic blood pressure, DBP diastolic blood pressure, SCR the serum creatinine, TOTCHOL is the total cholesterol, BG the blood glucose, STRIG is the serum triglycerides, SPOT the serum potassium, HB the hemoglobin, UACR the ratio of albumine to serum creatinine.

	unit	N	mean	std	min	max
eGFR	ml/min/1.73 m^2	256	64	16	31	90
SBP	mm Hg	256	138	16	100	180
DBP	mm Hg	256	78	10	44	101
BG	mg/dl	256	144	46	47	326
HbA1c	%	256	7.2	1.2	4.9	11.8
SCR	mg/dl	256	1.08	0.30	0.66	2.12
TOTCHOL	mg/dl	256	181	44	85	363
STRIG	mg/dl	256	172	106	44	859
SPOT	mmol/l	256	4.5	0.5	3.2	6.1
HB	g/dl	256	13.5	1.5	9.7	17.6
UACR	mg/g	256	78.94	283.85	0.0	2777.14

Table 2. Values of the mean eGFR for the 256 patients at each time-point.

time	min	max	mean	std
baseline (t_0)	31	90	64.0	16.3
follow-up 1 (t_1)	23	122	63.3	18.9
follow-up 2 (t_2)	15	105	60.6	18.3
follow-up 3 (t_3)	19	120	58.6	18.2

3 Results of the Longitudinal Clustering

In this section, we present the clusters of longitudinal data obtained with the *kmlShape* method (Fig. 3). We chose 5 as the number of clusters following clinical practice to analyze heterogeneity in eGFR patients' trends, taking into account that the eGFR is computed from different variables [30]. Such a choice is also motivated by the analysis of eGFR trajectories in different follow-ups, whose mean presents a slow decline. The choice of 5 classes allowed us to highlight behaviors such as crossing and bifurcations[2] having a medical interest.

[2] Each cluster of trajectories contains the same patients. However, we noticed that there are crossings, that we called "switches", between the mean values of eGFR of specific clusters. It means that, for instance, the patients in a specific cluster had an improvement over time, while the patients belonging to a cluster of initial good values of eGFR, had later a worsening of their condition. Or, in another case, we notice that two initially-close clusters of trajectories are then moving apart (the bifurcation). This is interesting from a medical point of view, because the patients who are initially quite close, then can have a different disease behavior.

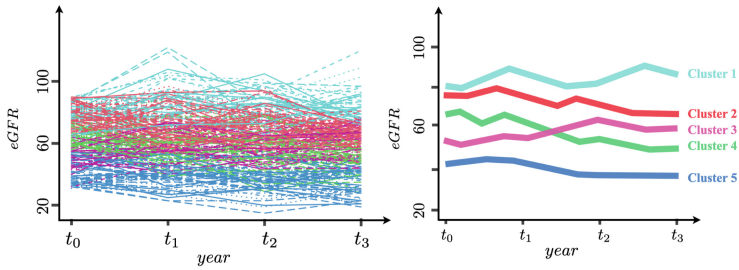

Fig. 3. Patients' eGFR trajectories (left), and mean eGFR trajectories obtained with *kmlShape* (right).

Table 3. Characteristics of patients in each of the five clusters of trajectories.

	unit	cluster 1 ($N = 160$)		cluster 2 ($N = 292$)		cluster 3 ($N = 152$)		cluster 4 ($N = 216$)		cluster 5 ($N = 204$)		p-value
		mean	std	mean	std	mean	std	mean	std	mean	std	
eGFR	mg/dl	84	11	72	10	57	9	57	11	38	9	0.000
age	years	62	10	66	7	69	8	69	8	73	9	0.000
BMI	kg	30.90	5.18	30.54	4.65	31.53	4.97	31.38	4.65	30.31	5.16	0.050
BG	mg/dl	148	50	149	48	144	56	152	54	155	74	0.363
HbA1c	%	7.4	1.1	7.2	1.2	7.3	1.3	7.4	1.2	7.3	1.4	0.182
TOTCHOL	mg/dl	176	41	184	44	186	49	178	53	172	39	0.020
STRIG	mg/dl	159	88	169	102	169	117	203	150	194	131	0.000
SPOT	mmol/l	4.5	0.5	4.4	0.5	4.4	0.5	4.5	0.5	4.9	0.6	0.000
HB	g/dl	14.3	1.4	14.0	1.3	13.5	1.5	13.2	1.4	12.7	1.5	0.000
CRP	mg/l	0.68	2.13	0.49	1.46	0.56	1.15	0.49	0.99	0.86	2.32	0.035
UACR	mg/g	34.91	102.07	43.53	157.58	54.68	153.43	88.09	281.23	122.91	324.18	0.000

Examining the resulting mean trajectories, we notice a bifurcation between cluster 1 and cluster 2, and a switch between cluster 3 and cluster 4. In the following, we analyze how the relevant variables describe the profile of patients, trying to understand the pattern of their dynamic behavior.

The mean values of the selected variables for patients in each cluster are shown in Table 3. They are clusters of patients, grouped according to the shape similarity of their eGFR trajectories. Patients are distributed along five different levels of eGFR, ranging from a mean value of 84 ± 11 in Cluster 1, and 30 ± 9 in Cluster 5. For the patients in each cluster, we also computed the mean values of the variables which presented a statistical significance (with the p-value test): age, body-mass index, blood glucose, HbA1c, total cholesterol, serum triglycerides, serum potassium, hemoglobin, and UACR. The information provided by these other variables can shed light on unexpected behaviors of eGFR mean trajectories.

Observing Fig. 3, we notice that the eGFR trajectories for patients in Clusters 1 and 2 start from close values of mean eGFR around 80 mg/dl, then have a different trend. Both Clusters 1 and 2 present higher eGFR values (84 ± 11 and 72 ± 10, respectively), UACR under control (34.91 ± 102.07 and 43.04 ± 87.38), but mean values of STRIG higher for patients in Cluster 2 (169 ± 102 against

159 ± 88 of Cluster 1). The mean age of patients in Cluster 2 is also higher than the mean age of patients in Cluster 1 (66 ± 7 against 62 ± 10 of Cluster 1). We then notice that eGFR trajectories of patients in Clusters 3 and 4 present a switch after the second time-point, that is, the first follow-up (t_1). Patients in Clusters 3 and 4 present trajectories of eGFR slightly lower, that is, 57 ± 9 and 57 ± 11, respectively. The main difference with respect to the other variables is constituted by the mean values of CRP, that is, the C-reactive protein: 0.56 ± 1.15 in Cluster 3, and 0.49 ± 0.99 in Cluster 4. Moreover, we notice the difference of the mean value of STRIG (169 ± 117 in Cluster 3, 203 ± 150 in Cluster 4) and of UACR (54.68 ± 153.43 in Cluster 3, 88.09 ± 281.23 in Cluster 4). The improvement of mean eGFR across time for patients in Cluster 4 can be due to the effect of drug treatment, more effective for patients belonging to this specific subgroup. Patients in Cluster 5 present the lowest values of mean eGFR (38 ± 9), and they are characterized by critical values of HB and SPOT (12.7 ± 1.5 and 4.9 ± 0.6, respectively).

4 Discussion

Diabetic kidney disease is a devastating complication of type-2 diabetes mellitus that reduces quality and quantity of life of affected patients and puts an enormous burden on healthcare budget. In addition to the optimization of lifestyle, the selection of the optimum drug combination for therapy is crucial to prevent the incidence and progression of DKD. Once thought to be a uniform disease, it is now evident that there is massive inter-individual and longitudinal intra-individual heterogeneity in disease pathophysiology, clinical presentation, and response to therapy. Linking the characteristics of each patient with the features of a specific subgroup of patients can give hints about the possible effective drug combination.

This is why, starting from a DKD dataset, we built subgroups of similar patients. In particular, we noticed indeed the bifurcation and a switch between mean trajectories. From a theoretical point of view related with basic definitions of category theory, we connected the comparison between clustering methods with the comparison between their results as clusters. A clustering method can be seen as a transformation, and the comparison between clusters is a natural transformation. Here, we estimated similarities and differences of the two methods (and thus, the natural transformation between them) in terms of their effect on a given dataset.

From our analysis of the results, we emphasize that patients with similar levels of eGFR at the baseline can then present a different disease evolution. This fact can be explained with different characteristics of the other variables at each time-point. This result is found using a shape-similarity method, the *kmlShape*, using the Fréchet distance.

The considered patients were given, at each time-point, one of four possible combinations of drugs, described in the Appendix: RASI + GLP1a, RASi + SGLT2i, RASi + MCRa, and RASi only. Analyzing the eGFR mean trajectories

shown in Fig. 3, and comparing them with the respective treatment received by the patients, we notice that the patients in Cluster 1 mostly received RASi only; patients in Clusters 2 and 3 were given RASi + MCRa; patients in Cluster 4 mostly received SGLT2i. On the other hand, patients in Clusters 3 and 5 did not receive GLP1a, independently by the level of eGFR.

The information achieved with trajectory clustering can be fed into a decision system, to predict the disease evolution of patient, according to their baseline clinical overview. Thus, our study can lead to a machine-learning application to help physicians deal with new cases of DKD disease. We highlighted here a connection between abstract mathematics, medical practice, and patients' real data, with a potential for further technological applications.

This research may help foster new strategies to improve DKD patients' lives.

Funding and Data Availability. Funding. This publication was supported by the European Union's Horizon 2020 research and innovation program under grant agreement Nr. 848011. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. **Institutional review.** DC-ren approval number of the Ethics Committee of the Medical University Innsbruck: EK Nr. 1188/2020, 19.06.2020. **Informed consent.** The DC-ren cohort consists of patients from PROVALID and informed consent was obtained from all patients. A: Ethical approval from the Ethics Committee of the Medical University Innsbruck AN4959 322/4.5 370/5.9 (4012a); 29.01.2013 and approval of the Ethics Committee of Upper Austria, Study Nr. I-1-11; 30.12.2010. H: Approval from Semmelweis University, Regional and Institutional Committee Of Science And Research Ethics; No.12656-0/2011-EKU (421/PV11.); 17.06.2011. UK: Approval from WoSRES, NHS; Rec. Reference: 12/WS/0005 (13.01.2012). NL: Approval of the Medical Ethical Committee of the University Medical Center Groningen, ABRnr. NL35350.042.11. **Data availability.** The dataset is not publicly available.

This publication was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 848011

References

1. Mayer, G., Heerspink, H.J.L., Aschauer, C., Heinzel, A., Heinze, G., Kainz, A., et al.: Systems biology-derived biomarkers to predict progression of renal function decline in type 2 diabetes. Diabetes Care **40**, 391–397 (2017)
2. Park, S., Xu, H., Zhao, H.: Integrating multidimensional data for clustering analysis with applications to cancer patient data. J. Am. Stat. Assoc. **116**(533), 14–26 (2021)
3. Liu, L., Lin, L.: Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data. Comput. Stat. Data Anal. **138**, 239–259 (2019)
4. Philipson, L.H.: Harnessing heterogeneity in type 2 diabetes mellitus. Nature **16**(79), 80 (2019)

5. Fuchs, S., Di Lascio, M., Durante, F.: Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. Comput. Stat. Data Anal. **159**, 107201 (2021)
6. Amiri, S., Clarke, B.S., Clarke, J.L.: Clustering categorical data via ensembling dissimilarity matrices. J. Comput. Graph Statist. **27**(1), 195–208 (2017)
7. Boucquemont, J., Loubère, L., Metzger, M., Combe, C., Stengel, B., Leffondre, K.: Identifying subgroups of renal function trajectories. Nephrol. Dial Transpl. **32**, ii185–ii193 (2017)
8. Kerschbaum, J., et al.: Intra-individual variability of eGFR trajectories in early diabetic kidney disease and lack of performance of prognostic biomarkers. Nat. Sci. Rep. **10**, 1973 (2020)
9. Karpati, T., Leventer-Roberts, M., Feldman, B., Cohen-Stavi, C., Raz, I., Balicer, R.: Patient clusters based on HbA1c trajectories: a step toward individualized medicine in type 2 diabetes. PLoS ONE **13**(11), e0207096 (2018)
10. Park, S.: Examining trajectories of early adolescents' life satisfaction in South Korea using a growth mixture model. Appl. Res. Qual. Life **17**, 149–168 (2022). https://doi.org/10.1007/s11482-020-09884-5
11. Liu, C., Wei, Y., Ling, Y., et al.: Identifying trajectories of Chinese high school students' depressive symptoms: an application of latent growth mixture modeling. Appl. Res. Qual. Life **15**, 775–789 (2020). https://doi.org/10.1007/s11482-018-9703-3
12. Perco, P., Mayer, G.: Molecular, histological, and clinical phenotyping of diabetic nephropathy: valuable complementary information? Kidney Int. **93**, 308–310 (2018)
13. Fréchet, M.: Sur quelques points du calcul fonctionnel. Rendiconti Circolo Matematico Palermo **22**, 1–72 (1906)
14. Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., Subtil, F.: kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. PLoS ONE **11**(6), e0150738 (2016)
15. Pinaire, J., Aze, J., Bringay, S., Poncelet, P., Genolini, C., Landais, P.: Hospital healthcare flows: a longitudinal clustering approach of acute coronary syndrome in women over 45years. Health Inform. J. 1–17 (2021). https://doi.org/10.1177/14604582211033020
16. Mayer, B.: Using systems biology to evaluate targets and mechanism of action of drugs for diabetes comorbidities. Diabetologia **59**, 2503–2506 (2016)
17. Mac, L.S.: Categories for the Working Mathematicians. Cambridge University Press, New York (1978)
18. Grandis, M.: Higher Category Theory. World Scientific, Singapore (2020)
19. Baez, J., Lauda, A.: A prehistory of n-categorical physics. In: Halvorson, H. (ed.) Deep Beauty: Understanding the Quantum World Through Mathematical Innovation. Cambridge University Press (2011)
20. Spivak, D.: Category Theory for the Sciences. MIT Press, Cambridge (2014)
21. Ehresmann, A., Góomez-Ramirez, E.: Conciliating neuroscience and phenomenology via Category Theory. Progr. Biophys. Mol. Biol. (PBMB) **119**, 347–359 (2015)
22. Carlsson, G., Mémoli, F.: Classifying clustering schemes. Found. Comput. Math. **13**, 221–252 (2013)
23. Mannone, M., Distefano, V., Silvestri, C., Poli, I.: Clustering longitudinal data with category theory for diabetic kidney disease. In: CLADAG 2021, Book of Abstract (2021, to appear)
24. Genolini, C., Falissard, B.: KmL: K-means for longitudinal data. Comput. Stat. **25**(2), 1–34 (2010)

25. Tran, C.S., Nicolau, D., Nayak, R., Verhoeven, P.: Modeling credit risk: a category theory perspective. J. Risk Financ. Manage. **14**(298), 1–21 (2021)
26. Alicic, R.Z., Rooney, M.T., Tuttle, K.R.: Diabetic kidney disease: challenges, progress, and possibilities. Clin. J. Am. Soc. Nephrolol. **12**, 2032–2045 (2017)
27. Pinaire, J., Aze, J., Bringay, S., Poncelet, P., Genolini, C., Landais, P.: Hospital healthcare flows: a longitudinal clustering approach of acute coronary syndrome in women over 45 years. Health Inform. J. **27**(3) (2021)
28. Verboon, P., Pat-El, R.: Clustering longitudinal data using R: A Monte Carlo study. Eur. J. Res. Methods Behav. Soc. Sci. **18**, 144–163 (2022)
29. Varoutas, P.-C., Rizand, P., Livartowski, A.: Using category theory as a basis for a heterogeneous data source search meta-engine: the Prométhée framework. In: Johnson, M., Vene, V. (eds.) AMAST 2006. LNCS, vol. 4019, pp. 381–387. Springer, Heidelberg (2006). https://doi.org/10.1007/11784180_30
30. Thöni, S., Keller, F., Denicolò, S., Buchwinkler, L., Mayer, G.: Biological variation and reference change value of the estimated glomerular filtration rate in humans: a systematic review and meta-analysis. Front. Med. (Lausanne) **6**(9), 1009358 (2022)