Clustering Cardiovascular Risk Trajectories of Patients with Type 2 Diabetes Using Process Mining

Joyce Pebesma[1], Antonio Martinez-Millana[2]*, Lucia Sacchi[3], Carlos Fernandez-Llatas[2], Pasquale De Cata[4], Luca Chiovato[4], Riccardo Bellazzi[3], Vicente Traver[2]

Abstract— **Patients with type 2 diabetes have a higher chance of developing cardiovascular diseases and an increased odds of mortality. Reliability of randomized clinical trials is continuously judged due to selection, attrition and reporting bias. Moreover, cardiovascular risk is frequently assessed in cross-sectional studies instead of observing the evolution of risk in longitudinal cohorts. In order to correctly assess the course of cardiovascular risk in patients with type 2 diabetes, we applied process mining techniques based on the principles of evidence-based medicine. Using a validated formulation of the cardiovascular risk, process mining allowed to cluster frequent risk pathways and produced 3 major trajectories related to risk management: high risk, medium risk and low risk. This enables the extraction of meaningful distributions, such as the gender of the patients per cluster in a human understandable manner, leading to more insights to improve the management of cardiovascular diseases in type 2 diabetes patients.**

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading causes of morbidity and mortality among people with Type 2 Diabetes Mellitus (T2DM) [1]. T2DM patients are in twice risk of developing Coronary Heart Disease and Hypertension compared to non-T2DM [2]. Clinical guidelines for the management of T2DM recommend the use of CVDs risk assessment scores using traditional predictors such as hypertension, dyslipidemia, body mass index, smoking habits and family history [3]. Multi-factorial intervention reduced CVD events and mortality in T2DM in the WHO Multinational Study of Vascular Disease in Diabetes, concluding that CVD risk estimation is important to plan both preventive and therapeutic programs including anti-lipid, anti-hypertensive and anti-platelet therapies [4].

The primary goal of T2DM management is improving glycemic control to prevent microvascular complications, while subsequently normalizing CVDs risk factors to reduce events and mortality. Several studies have reported on the benefits of using targeted drugs for CVDs risk reduction in patients with T2DM, such as Marso et al. [5]. All participants in these studies had T2DM, while more than 80% of them had a previous cardiovascular event. This made it difficult to determine the benefit of the targeted drugs on the natural course of the CVDs [6]. There are substantial gender differences in the risk of different CVD states in T2DM patients [7]. However, several trials and reviews reveal that there are different interpretations of the gender equality in the burden of CVD, whereby they want to improve clinical CVD outcomes [8].

Such limitations can be avoided by using the evidence-based medicine (EBM) paradigm [9]. EBM consists of a continuous learning process to provide high quality clinical care by extracting relevant information about prognosis and therapy. EBM conveys into a sequence of five steps: (1) analysis of clinical data; (2) trace the best procedures to achieve a clinical endpoint; (3) appraise the evidence for its closeness to the truth and clinical applicability; (4) apply the findings into routine clinical practice; and (5) assess the performance of the procedure. EBM is usually deployed through protocols based on scientific evidence. These protocols, also known as Clinical Pathways [10], consist of care and medication plans which define the explorations and therapies patients should follow to properly treat their disease. The exact clinical pathways can be extracted through the use of Electronic Health Records, showing the continuous process of care over time. Process Mining (PM) [11] allows to aggregate heterogeneous data and to infer the clinical pathways followed by patients over the years. PM can sort and order the activities in a sequential order, depicting and discovering processes. The activities are retrieved from an event log, which contains the records of the clinical acts. Event logs typically record the type of task, the time stamp their result.

Our approach is to study the clinical pathways of CVDs as an indicator for the Cardiovascular Disease Risk (CVR) progression in T2DM patients and discover differences due to the sex of the subjects. Thereby the CVR tool as defined in *projecto Cuore* [12] was used, to identify the course of patients and the multivariate factors affecting the evolution of CVDs in patients with T2DM. Different techniques enabled us to perform an analysis to discover similarities and differences on the patient management process. We used a dataset from Fondazione Salvatore Mauggeri (Pavia, Italy), which contains clinical and administrative data of 1020 subjects with T2DM. PM is employed to discover the common risk paths, as well as find exceptional and unexpected situations, contributing to the management of CVDs in T2DM patients.

This work was supported by European Commission Grant No 600914 (MOSAIC Project)

[1] University of Twente, Drienerlolaan 5, 7522NB Enschede, the Netherlands j.l.pebesma@alumnus.utwente.nl

[2] ITACA. Universitat Politècnica de València, Camino de vera sn 46022 Valencia, Spain anmarmil;carferll;vtraver@itaca.upv.es

[3] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy lucia.sacchi;riccardo.bellazzi@dei.unipv.it

[4] UO di Medicina Interna e Endocrinologia, ICS Maugeri, Pavia, Italy pasquale.decata;luca.chiovato@icsmaugeri.it

II. Materials and Methods

A. Process Mining on Type 2 Diabetes Mellitus

PM is a knowledge extraction technique which has the objective of providing human understandable schemes on how processes are performed [11]. PM is an emerging discipline in health informatics that is useful for obtaining new perspectives on how patients are managed and identifying the most commonly followed clinical pathways.

The management of T2DM is supported by clinical and non-clinical activities that vary per context, such as the management of patients in a public health care system and a private health care system is different [13].

Health care processes for managing T2DM are highly dynamic, complex and multidisciplinary in the way that each endocrinology unit has its own manner of implementing standard clinical guidelines [14]. Improving health care processes is not an easy task, even though it is clear that through its optimization, both the patients' quality of life and the use of clinical resources can be increased and optimized [15].

A Hospital Information System (HIS) contains any action or decision performed by medical or managing staff, which can be converted to a process log. Thereby the activities are stored as events, containing data about it's when, what, who and their result [16]. Each patient flow is considered a sample, which altogether make up the process log. PM techniques can provide a high realistic view on how the process was implemented, helping the involved stakeholders to obtain information about the sequential order of the activities, the role of each actor, bottlenecks and unexpected paths.

There are three main types of process mining approaches: process discovery, process conformance and enhancement.

- Discovery aims to discover the flow of activities and processes by analyzing the data from scratch, whereby its' algorithms infer graphical flows represent the real process that a patient (or group of patients) followed. Discovery algorithms can be event-based or activity-based, whereby activities also contain the result of the event.
- Conformance: aims to discover the database samples that match a given work flow.
- Enhancement: is a technique to provide visual information about the distribution of the load among the flows, within their nodes and transitions. This enables a visual way of identifying the most common flows and (potential) deadlocks and bottle necks.

Our approach for identifying CVD risk trajectories in T2DM population was based on a combination of process discovery and conformance. This method of conformance is elsewhere known as process clustering, in which we try to group similar cases in such a way that for every cluster it is possible to observe a simple model. Since no a priori number of clusters was known, we opted for quality threshold clustering. This type of clustering only requires a distance threshold within each cluster, defined as the similarity index.

B. Cardiovascular Risk

CVR was estimated using the algorithm proposed and validated in *Progetto Cuore* [12], which is described in Figure 1. This algorithm was developed by the Italian Ministry of Health to estimate the impact of CVDs in the general population through a board of indicators, such as prevalence, incidence and mortality rates. The study data set was collected containing all the input variables for CVR calculation in different time points.

The CVR model is a simple and objective way of assessing the likelihood of experiencing a first major cardiovascular event (myocardial infarction or stroke) over the following ten years. The model computes the value of six CVD risk factors: gender, history of diabetes, smoking, age, systolic blood pressure and total serum cholesterol. Using the this algorithm (Figure 1), we obtained CVR scores for each group of clinical measures in different time points, which allowed us to stratify our population into six categories from CVR I to CVR VI (Table I). The CVR category indicates how many persons out of 100 people, with the same characteristics, may develop CVDs over next 10 years.

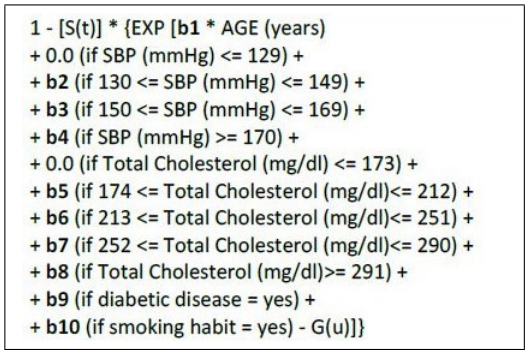


Fig. 1: Algorithm for cardiovascular risk computation. Intercept (S(t)), coefficients (bn) and G(u) are parameters for male and female subjects

C. Data set descriptive analysis

From a total of 1,020 subjects, subjects without CVR observations were excluded, obtaining a data set of 930 subjects with 9,227 records of CVR events in the period from December 1996 to February 2015. The biomedical ethics committee of the hospital approved the design of the study. The data set was composed of 49% females and 51% males with an age of 69 ± 10 years old. Table I depicts for each CVR state the total number of records, the number of unique subjects and the duration of the observation in days for both males and females. Due to the non-parametric distribution of the duration, we tested if the sex was influencing the duration with a Wilcoxon rank-sum test (*p* row in Table I). None of the distributions showed a statistically significant difference. The unique number of observed CVR states was taken by counting and summing each unique state per subject. The result can be seen in Table I. Hereby it can be seen that the higher states are observed most often, in contrary to state I being observed least often.

TABLE I: DESCRIPTION OF THE STUDY DATA SET

State	Number of records	Unique subjects	Duration of stage (days) female *mean (sd)*	male *mean (sd)*
I	1127	179	251.7 (360.7)	300.5 (561.6)
II	1830	291	217.0 (298.4)	224.6 (343.9)
III	1358	300	235.7 (330.3)	231.1 (358.1)
IV	962	259	220.6 (323.5)	233.1 (332.6)
V	1534	335	238.9 (353.9)	230.1 (307.5)
VI	2415	335	240.5 (332.4)	218.7 (297.9)

III. RESULTS

The process mining algorithms were executed for clustering the CVR trajectories, using different thresholds for the topological similarity of the trajectories. Figure 2 shows the number of discovered clusters (y axis) depending on the used similarity index (x axis). Beyond the number of found cluster it is important to know the number of subjects included in the clusters and the number of subjects who are not included in these clusters. In this picture the size of the point indicates the mean number of subjects per cluster for each similarity index. We can see for instance that the clusters computed with a similarity index below 40% have less subjects than the clusters with a similarity index over 65%. Moreover, the filling color shows the number of subjects not included in any cluster (outliers). This provides meaningful information, as we can see that the number of outliers decreases with the increase of the similarity index.

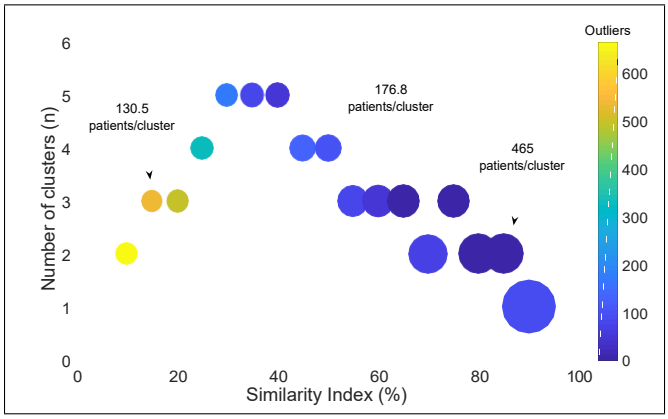
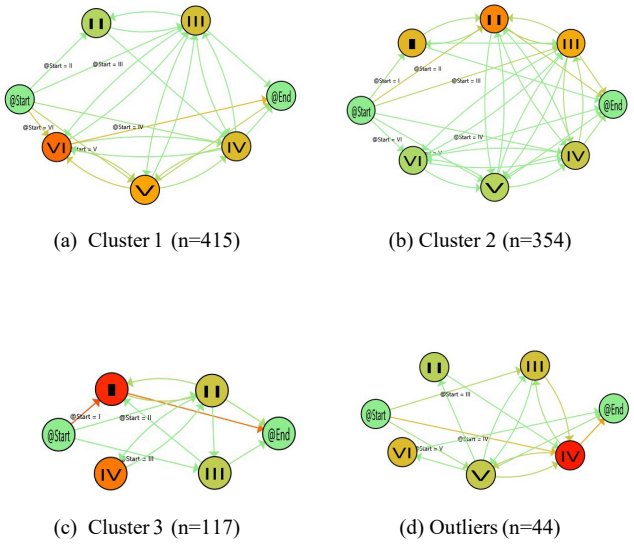


Fig. 2: Clusters of the CVR trajectories. The size of the point indicates the mean number of subjects on each cluster and the filling color indicates the number of subjects which did not fit in any of the discovered clusters

These results show that similarity indexes under 30% and over 65% are not well suited for inferring CVR trajectories, as two situations may occur:

- 1-2 clusters with a lot of outliers and a very simple topology (for example CVR II → CVR III),
- 1-2 clusters with few outliers, but a very complex topology (for example CVR II → III → IV → V → IV).

Therefore we chose a compromise solution of selecting a similarity index of 60% and three different clusters.



(a) Cluster 1 (n=415) (b) Cluster 2 (n=354)

(c) Cluster 3 (n=117) (d) Outliers (n=44)

Fig. 3: The three found clusters with similarity= 60%

TABLE II: MOST COMMON TRAJECTORY PER CLUSTER, GENDER DISTRIBUTION AND TOTAL NUMBER OF SUBJECTS.S=START;E=END

Cluster	Trajectory	Gender (F / M)	Number of subjects
1	S→VI→E	(16% / 84%)	415 (44.6%)
2	S→II→III→E	(57% / 43%)	354 (38.1%)
3	S→I→E	(83% / 17%)	117 (12.6%)
Outliers	S→IV→E	(59% / 41%)	44 (4.8%)

Figure 3 shows the extracted flows for the selected configuration of clusters (including the outliers). Hereby it is visible that in every flow, the most crowded state is different. The first cluster (Figure 3-a) describes the course of high risk subjects, which are the majority of the time in risk states V and VI. The second cluster (Figure 3-b) describes the course of medium risk subjects, who are most of the time in states II-III. The third cluster (Figure 3-c) describes the behavior of low risk patients who are firstly in low risk state (I) and then evolve for a short time to higher risk states. The last group of patients (outliers in Figure 3-d) contains individual trajectories which did not fit into the previous clusters. These are mainly composed of patients who are the majority of the time in risk state IV.

The most common trajectory followed per cluster can be seen table II. Aside from the trajectory, it shows the gender distribution per cluster. The first and third cluster have a significant difference between the number of males

and females. Whereas the first cluster mostly contains males, the third mostly contains females. This might be an indicator that females often are in a lower risk states.

IV. Discussion

The application of Process Mining technologies allowed us to identify three different trajectories in the course of cardiovascular risk in patient with type 2 diabetes: high risk, medium risk and low risk.

Patient trajectories were clustered using a heuristic model with a similarity index. The index was chosen based upon a compromise between the number of clusters and outliers (Figure 2). This way of quality clustering was chosen, since we had no a priori knowledge about the trajectories of possible groups. We propose to implement efficient clustering techniques (for example, k-means) to obtain groups based on a pre-defined number of clusters. The advantage of having a pre-specified number of clusters, thus quantitative clustering, is that it works efficiently when there is previous knowledge about the possible variations in trajectories.

The complexity of health care data, the heterogeneity of diseases and the (lack of) data quality in clinical routine collected data makes hard to extract process models in a clear and understandable way [17]. Finally, another point to evaluate is the ability to model the time. The representation of time is crucial in health care, as it can represent back and forth clinical situations (response to treatments, effect of follow-up, etc). Clinical pathway models should embrace the temporal dimension, as changes may occur silently [18]. Our approach of using PM described the clinical course of CVR effectively and provided clusters grouping the variety of trajectories which can be way forward for the definition of personalized preventive and therapeutic plans in an efficient way and based on evidence.

References

[1] A. D. Shah, C. Langenberg, and et al, "Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 19 million people," *The Lancet Diabetes & Endocrinology*, vol. 3, no. 2, pp. 105 – 113, 2015.

[2] "9. cardiovascular disease and risk management," *Diabetes Care*, vol. 40, no. Supplement 1, pp. S75–S87, jan 2017.

[3] *Recommendations For Managing Type 2 Diabetes In Primary Care*. International Diabetes Federation, 2017.

[4] E. W. Gregg, N. Sattar, and M. K. Ali, "The changing face of diabetes complications," *The Lancet Diabetes & Endocrinology*, vol. 4, no. 6, pp. 537 – 547, 2016.

[5] S. P. Marso, G. H. Daniels, K. Brown-Frandsen, P. Kristensen, J. F. Mann, M. A. Nauck, S. E. Nissen, S. Pocock, N. R. Poulter, L. S. Ravn *et al.*, "Liraglutide and cardiovascular outcomes in type 2 diabetes," *New England Journal of Medicine*, vol. 375, no. 4, pp. 311–322, 2016.

[6] M. Abdul-Ghani, R. A. DeFronzo, S. D. Prato, R. Chilton, R. Singh, and R. E. Ryder, "Cardiovascular disease and type 2 diabetes: Has the dawn of a new era arrived?" *Diabetes Care*, vol. 40, no. 7, pp. 813–820, jun 2017.

[7] A. Juutilainen and S. e. a. Kortelainen, "Gender difference in the impact of type 2 diabetes on coronary heart disease risk," *Diabetes care*, vol. 27, no. 12, pp. 2898–2904, 2004.

[8] L. Mosca, E. Barrett-Connor, and N. Kass Wenger, "Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes," *Circulation*, vol. 124, no. 19, pp. 2145–2154, 2011.

[9] D. L. Sackett, "Evidence-based medicine," *Seminars in Perinatology*, vol. 21, no. 1, pp. 3 – 5, 1997, fatal and Neonatal Hematology for the 21st Century.

[10] J. Fox, E. Black, I. Chronakis, R. Dunlop, V. Patkar, M. South, and R. Thomson, "From guidelines to careflows: modelling and supporting complex clinical processes," *Studies in health technology and informatics*, vol. 139, p. 4462, 2008.

[11] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, J. Buijs *et al.*, "Process mining manifesto," in *International Conference on Business Process Management*. Springer, 2011, pp. 169–194.

[12] S. Giampaoli, L. Palmieri, C. Donfrancesco, C. L. Noce, L. Pilotto, and D. Vanuzzo, "Cardiovascular health in italy. ten-year surveillance of cardiovascular diseases and risk factors: Osservatorio epidemiologico cardiovascolare/health examination survey 1998–2012," *European Journal of Preventive Cardiology*, vol. 22, no. 2_suppl, pp. 9–37, jul 2015.

[13] M. N. Munshi, H. Florez, E. S. Huang, R. R. Kalyani, M. Mupanomunda, N. Pandya, C. S. Swift, T. H. Taveira, and L. B. Haas, "Management of diabetes in long-term care and skilled nursing facilities: A position statement of the american diabetes association," *Diabetes Care*, vol. 39, no. 2, pp. 308–318, 2016.

[14] UK Hypoglycaemia Study Group, "Risk of hypoglycaemia in types 1 and 2 diabetes: effects of treatment modalities and their duration," *Diabetologia*, vol. 50, no. 6, pp. 1140–1147, Jun 2007.

[15] P. Depablos-Velasco and S.-C. et al, "Quality of life and satisfaction with treatment in subjects with type 2 diabetes: results in spain of the panorama study," *Endocrinologia y nutricion : organo de la Sociedad Espanola de Endocrinologia y Nutricion*, vol. 61, no. 1, p. 1826, January 2014.

[16] A. Dagliati, L. Sacchi, M. Bucalo, and D. S. et al, "A data gathering framework to collect type 2 diabetes patients data," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, June 2014, pp. 244–247.

[17] C. Fernandez-Llatas, A. Martinez-Millana, A. Martinez-Romero, J. M. Benedi, and V. Traver, "Diabetes care related process modelling using process mining techniques. lessons learned in the application of interactive pattern recognition: Coping with the spaghetti effect," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 2127–2130.

[18] A. Dagliati, L. Sacchi, C. Cerra, P. Leporati, P. De Cata, L. Chiovato, J. H. Holmes, and R. Bellazzi, "Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in type 2 diabetes patients," in *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*. IEEE, 2014, pp. 240–243.