Journal of Biomedical Informatics 135 (2022) 104218

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Longitudinal deep learning clustering of Type 2 Diabetes Mellitus trajectories using routinely collected health records

Enrico Manzini [a,b,c,*,1], Bogdan Vlacho [d,1], Josep Franch-Nadal [d,e,f], Joan Escudero [g], Ana Génova [g], Elisenda Reixach [h], Erik Andrés [h], Israel Pizarro [i], José-Luis Portero [i], Dídac Mauricio [d,e,j,**,2], Alexandre Perera-Lluna [a,b,c,2]

[a] Universitat Politecnica de Catalunya, Barcelona, Spain
[b] Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine, Madrid, Spain
[c] Institut de Recerca Sant Joan de Deu, Barcelona, Spain
[d] DAP-Cat Group, Unitat de Suport a la Recerca, Fundaciò Institut Universitari per a la recerca a l'Atenciò Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barelona, Spain
[e] Center for Biomedical Research on Diabetes and Associated Metabolic Diseases (CIBERDEM CB15/00071), Instituto de Salud Carlos III, 28029, Madrid, Spain
[f] Primary Health Care Center Raval Sud, Institut Català de la Salut, Barcelona, Spain
[g] Grupo Pulso, Spain
[h] Fundació TIC Salut Social, Departament de Salut, Generalitat de Catalunya, Barcelona, Spain
[i] Novo Nordisk, Spain
[j] Department of Medicine, University of Vic - Central University of Catalonia, Vic, Spain

ARTICLE INFO

Keywords:
Type 2 diabetes
Deep learning
Longitudinal cluster
AutoEncoder
Diabetic complications
Electronic health records

ABSTRACT

Type 2 diabetes mellitus (T2DM) is a highly heterogeneous chronic disease with different pathophysiological and genetic characteristics affecting its progression, associated complications and response to therapies. The advances in deep learning (DL) techniques and the availability of a large amount of healthcare data allow us to investigate T2DM characteristics and evolution with a completely new approach, studying common disease trajectories rather than cross sectional values. We used an Kernelized-AutoEncoder algorithm to map 5 years of data of 11,028 subjects diagnosed with T2DM in a latent space that embedded similarities and differences between patients in terms of the evolution of the disease. Once we obtained the latent space, we used classical clustering algorithms to create longitudinal clusters representing different evolutions of the diabetic disease. Our unsupervised DL clustering algorithm suggested seven different longitudinal clusters. Different mean ages were observed among the clusters (ranging from 65.3±11.6 to 72.8±9.4). Subjects in clusters B (Hypercholesteraemic) and E (Hypertensive) had shorter diabetes duration (9.2±3.9 and 9.5±3.9 years respectively). Subjects in Cluster G (Metabolic) had the poorest glycaemic control (mean glycated hemoglobin 7.99±1.42%), while cluster E had the best one (mean glycated hemoglobin 7.04±1.11%). Obesity was observed mainly in clusters A (Neuropathic), C (Multiple Complications), F (Retinopathy) and G. A dashboard is available at dm2.b2slab.upc.edu to visualize the different trajectories corresponding to the 7 clusters.

1. Introduction

Type 2 diabetes mellitus(T2DM) is a chronic and highly prevalent disease ranking as eighth in terms of the overall burden of diseases measure developed by WHO [1]. According to the International Diabetes Federation, estimates of the disease's prevalence showed alarming increases since 2000, tripling the number of affected subjects in 2019 to 463 million [2]. Moreover, projections for the future indicate that the global impact of diabetes is likely to continue increasing considerably.

Nowadays, the evidence suggests that T2DM is a complex highly heterogeneous metabolic disease that encompasses different pathophysiological and genetic pathways [3]. The presentation and progression of the disease can vary between subjects leading to poor glycemic and metabolic control [4]. Indeed in Europe alone, the European Medicines

[*] Corresponding author at: Universitat Politecnica de Catalunya, Barcelona, Spain.
[**] Corresponding author.
E-mail addresses: enrico.manzini@upc.edu (E. Manzini), didacmauricio@gmail.com (D. Mauricio).
[1] These authors have contributed equally to this work and share first authorship.
[2] These authors have contributed equally to this work and share senior authorship.

https://doi.org/10.1016/j.jbi.2022.104218
Received 20 June 2022; Received in revised form 8 September 2022; Accepted 3 October 2022
Available online 8 October 2022

Agency (EMA), for the period between 2005 and 2017, approved 40 new drugs to treat diabetes [5]. In Catalonia (Spain), according to previously published data, 43.9% of the T2DM subjects did not reach an adequate glycemic control [6]. This issue might be since T2DM diagnosis is oversimplified via the assessment of blood levels of glucose only.

The availability of routinely collected data from Electronic Health Records (EHR) in recent years had a huge impact on clinical, pharmacoepidemiologic and health services research, but the accuracy, quality and heterogeneity of these data remain challenging to handle [7]. In the case of the diabetic disease, for example, they have been recently used to point out the complexity of T2DM and its pathophysiological phenotyping [3,8]. In 2018, a seminal study from Sweden, identified five novel clusters of subjects [3]. In a large Danish study of newly diagnosed T2DM, authors identified several distinct pathophysiological phenotypes according to beta cell function and insulin sensitivity [8]. However, these attempts to study different phenotype clusters in the diabetic population have been done with cross sectional studies, rather than focusing on the clinical evolution of the disease. Even if some attempts to study typical T2DM trajectories exists [9–11], these are limited to few factors that influence the evolution of the diseases, e.g. a limited subset of common comorbidities [9], or the evolution of few variables as the glycated hemoglobin (HbA1c) [10] or the body mass index (BMI) [11]. To the best of our knowledge, there are no studies focused on the whole evolution of the disease and that take into consideration several factors including comorbidities, medical treatments, HbA1c changes over time and other important variables for the description of the disease.

Deep learning (DL) has been proved to be very effective in the design of multivariate patient trajectories, especially for databases with many missings values, which is very common among data collected from EHRs, e.g. [12,13]. Even if the in recent years the number of DL models to extract temporal information from EHRs has notably increased, several challenges remain to be handled, with the major ones being irregularity, sparsity and heterogeneity of the data and the opacity of the models due to their black box nature [14].

The proper clusterization of T2DM subjects by their clinical characteristics is a major issue. Progress in this field should enable us to implement different strategies of precision medicine and include diagnostic algorithms for defining diabetes subtypes in order to implement the most appropriate clinical decisions [15].

The objective of this study is to exploit the potential of unsupervised deep learning algorithms to create different longitudinal phenotype clusters of T2DM patients using routinely collected data extracted from EHR. Hence, throughout this work, we make the following contributions: on one hand we adapt a novel DL model, the kernelized autoencoder (K-AE) [16] to work with EHR data, handling the problem of sparsity and irregularity of the data with a different kernel matrix and a different imputation process and trying to reduce the opacity of the model showing the results and the AE latent space in an interactive dashboard available at dm2.b2slab.upc.edu; on the other hand, we use this model to calculate longitudinal clusters of the T2DM disease, studying different typical trajectories.

2. Material and methods

2.1. Data source and study population

Data from this study were extracted from the Information System for the Development of Research in Primary Care (SIDIAP) database [17] from January 1st 2013 to December 31th 2017. This database contains data from Electronic Health Records collected from approximately 5.6 million patients registered from 287 Primary Care Centers (PCC) in Catalonia. The SIDIAP database contains data on demographics, patient visits with health professionals, diagnoses, clinical variables, lifestyles information, medications prescriptions/dispensations and referrals to specialists, and laboratory test results, introduced by health professionals during routine health surveillance and health care.

We selected only subjects diagnosed with type 2 diabetes (ICD-10 diagnostic codes: E11 and E14 and their sub codes) [18]. In addition, subjects were not eligible if they were under 18 years old or they had other types of diabetes such as type 1 diabetes, secondary or gestational diabetes. For each eligible patient, clinical and analytical variables were available, such as glycated hemoglobin (HbA1c), body mass index (BMI), diastolic and systolic blood pressure (DBP and SBP), lipid profile and renal function. Furthermore, data on antidiabetic treatment (drugs in the A10 group of ATC classification) [19] and diagnosis of the most common comorbidities of diabetes were also available. ATC codes of the antidiabetic treatments grouped per pharmacological class and ICD-10 codes of the comorbidities are reported in Supplementary Tables S1 and S2 respectively. We included only patients that had more than 18 years at the moment of the T2Dm onset, with at least three measures for a minimum of two variables in the period under analysis, and who had at least one measure of HbA1C and BMI. The T2DM onset date was estimated for each subject as the minimum date between the date of the first T2DM diagnosis, the first prescription of the antidiabetic drugs, or the first measure of HbA1c>7% . The latter value was adopted as for an important part of the study subjects an initial HbA1c was determined before 2010, when HbA1c was standardized and accepted as a diagnostic criterion for diabetes. Thus, this algorithm was stricter but, at the same time, more precise in identifying the time of diagnosis of T2DM in our database.

2.2. Data preprocessing

Since routinely collected data extracted from EHR can contain errors arising from several different factors, we used clinical criteria to remove any extreme values that could potentially be measurement errors. Removed values are reported in Supplementary Tables S3. The study subjects diagram is presented in the Supplementary Figures S1. To mitigate the problem of irregularity of the data we represented each patient as a multivariate time series (MTS) $P_i \in \mathbb{R}^{n_{t,i} \times n_v}$ where:

- n_v is the number of variables considered for the analysis. We used the following variables: age, age at diagnosis of the subjects, sex, diabetes duration, five binary signals representing common diabetes secondary conditions (cardiovascular disease (CVD), chronic kidney disease (CKD), neuropathy, high blood pressure (HTN), and ophthalmological complications),nine pharmacological classes of antidiabetic drugs and 10 clinical and analytical variables (HbA1c, BMI, SBP, DBP, High-density lipoprotein (HDL), Low-density lipoprotein (LDL), Total cholesterol, triglycerides (TG), Creatinine and albumin to creatinine ratio (ACR)).
- $n_{t,i}$ is the number of time steps. Since every patient had an almost unique combination of number and frequency of measured variables, we standardized medical examinations dates, creating a temporal grid from January 1st, 2013 to December 31st, 2017, with a fixed frequency of six months. Then, we re-sampled the date of each examination according to this grid. In the case of two or more examinations inside the same time interval of the temporal grid, we used their mean value. Hence, we linearly interpolated these points. For this reason the maximum number of time steps is 10, but it can be lower, for example in the cases of patients who died before the end of the study.

The combination of measured variables and dates of the measurements changed from patient to patient, implying the possibility of missing values in the matrix Pi. We imputed such data using a two-step algorithm: firstly, we imputed missing values for a variable v_n when a patient had at least two existing measurements. For this we defined for each variable a mixed model, using the other variables as fixed effects and the patient as random effect. For each variable the best mixed model was chosen using the Bayesian Information Criterion (BIC). In the second step, we estimated values of v_n for patients having no, or just one, measures of that variable by using the Iterative Robust Model-Based Imputation (IRMI) [20] algorithm.

The Kernelized AutoEncoder To cluster patients considering the whole evolution of different variables during the five years of the follow-up period we used a Kernelized AutoEncoder (K-AE) [16], a deep learning model that can learn the representation of MTS mapping them in a latent space that embeds similarities and differences between patients. The MTS inputs that contain data extracted from EHRs of the patients are mapped in the latent space by an encoder built on a bidirectional stacked LSTM layer [21]. Then the decoder, composed by another stacked LSTM, maps the vectors in the latent space back in the original MTS. Furthermore, vectors in the latent space are aligned to a kernel matrix K built from the original MTS. Hence the model is trained with a cost function that is the weighted sum of the reconstruction error and the kernel alignment cost, namely:

$$L = \alpha_d \frac{1}{n_p} \sum_{i=1}^{n_p} MSE(P_i, \hat{P}_i) + \alpha_k \left\| \frac{ZZ^T}{\|ZZ^T\|_F} - \frac{K}{\|K\|_F} \right\|_F \tag{1}$$

Where:

- α_d and α_k are the weights of the reconstruction loss and kernel alignment cost functions respectively.
- $\hat{P}_i$ is the estimation of the MTS P_i made by the decoder.
- n_p is the number of patients.
- $MSE(P_i, \hat{P}_i)$ is the Mean Square Error between the reconstructed and the original MTS calculated for each patient for $n_{t,i}$ time steps available for that patient.
- $ZZ^T \in \mathbb{R}^{n_p \times n_p}$ is a matrix containing the dot products of the representations in the latent space of the MTS.
- $\| \cdot \|_F$ represents the Frobenius norm.
- K is the kernel matrix for the alignment.

2.3. Kernel matrix

In the original K-AE proposed by Bianchi et al. the kernel matrix was based on the time series cluster kernel [22]. Since this algorithm is based on some assumptions on the data (that are described as Gaussian Mixture Models with a time dependent mean and a constant covariance and missing values that must be Missing At Random), we decided to use a different kernel matrix based on the less restrictive Fast Global Alignment kernel (GAK) [23].

Practically, our kernel matrix K was obtained by iterating the original GAK algorithm several times using at each iteration a different set of variables to mitigate the possible effects of errors due to the imputation. At each iteration the unnormalized GAK was computed using only a random subset of variables chosen among the 10 clinical variables, the diagnosis and the pharmacological treatment (except for the first iteration in which all these variables were used) and it is added to the precedent GAK value. At the end of the iterations K was normalized to have 1 on the principal diagonal. This kernel alignment should help the Encoder create a latent space that better represents patients' evolutions and maintains distances between patients.

All these steps have been implemented in Python 3.8.10, using Tslearn 0.4.1 [24] for the creation of the kernel matrix and Tensorflow 2.4.3 [25] to create and define the deep learning model.

2.4. Clusters definition and validation

Once we obtained the vectors in the latent space, we divided patients into different clusters testing different state of art algorithms for clustering (k-mean, clara, and hierarchical clustering) in this space [26]. The definition of the optimal number of clusters in a vector space can be a delicate issue. This is true especially in this case where a series of hyperparameters, such as the dimensions of the different layers and the importance given to the kernel alignment, can affect the structure of the latent space, changing the distribution of the clusters. Optimal cluster number and hyperparameters have been chosen maximizing mean silhouette width calculated not in the clustering space (i.e. the latent space of the K-AE) but in the space defined by the original data of the subjects. Once we obtained the clusters in the latent space using different configurations of the K-AE, clustering algorithm and number of clusters, we divided the study period into ten intervals of 6 months each. Then, for each interval we considered only those patients that had at least a measure of HbA1c and BMI. We calculated the silhouette width for the period considering, aside from the two variables just mentioned, also the pharmacological treatment, the comorbidities, the age and the diabetes duration in the corresponding period.

Clusters were defined separately for men and women. Once we obtained the clusters for both sexes, we matched them to pair each cluster of men with a cluster of women. In order to do so, we represented each cluster with a vector with the mean of all values of the patients in that cluster, and we calculated the euclidean distance between men and women clusters, pairing clusters according to the minimum distance.

To validate our clustering method, we also clusterized patients using the same variables used to create the latent space considering not their evolution but rather a baseline value, similarly to what was done in other works [3,8]. All results were also compared with a random cluster as a reference.

The clinical characteristics in each cluster were described for mean and standard deviation for continuous variables and frequencies and percentages for categorical variables. The evolution of different variables for each cluster has been tested using linear mixed models counting for sex and age of the patient as a random effect and studying the effect of each cluster on the variable evolution respect to the disease duration. The percentages of patients diagnosed with comorbidities or that received a given drug prescription for each cluster was tested with the two-proportion z-test ($\alpha = 0.05$). The statistical analyses were performed using R3.6.3 software [27].

3. Results

3.1. Clusters validation results

In order to better visualize results of this work we developed a web app available at dm2.b2slab.upc.edu.

After applying the inclusion criteria, we obtained 11,028 patients, of which 5226 women and 5802 men. Patients' age (calculated at 01.01.2013) ranged from 24 to 100 years, with a mean of 69.8±10.6 years. Age at T2DM onset ranged from 18 to 91 years, with a mean of 59.4±10.9 years. Of these patients, 8790 were still active at the end of the study, 1793 died during the analyzed period and the remaining 445 were transferred from the database. Mean age at death was 81.4±8.7 years.

Clusterization and validation have been carried out separately for sexes, obtaining in both cases an optimal number of clusters equal to 7. The relationship found between men and women clusters was bijective, i.e. each cluster of one sex was paired with exactly one cluster of the other sex and the other way around. For both sexes, the best configuration of parameters of the K-AE generated a latent space that can be used to cluster patients better with respect to using a baseline value for each variable, as reported in Fig. 1.

3.2. Clusters description

The main statistics of the seven clusters are reported in Table 1, while Fig. 2 contains the mean values of the different variables such as drug usage and comorbidities for the different clusters. Numerical values for these variables are reported in Supplementary Tables S4, S5 and S6.

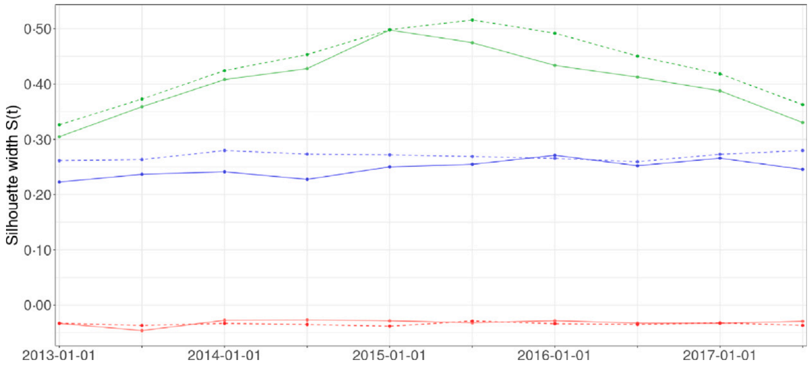

Fig. 1. Mean silhouette width evolution. Colors represent different clustering methods: green lines for clusters obtained with our K-AE model; blue lines for clusters obtained with a single basal value; and red lines for random clusterization. Dot and continuous lines represent respectively women and men clusters. K-AE used to cluster women had the following parameters: encoder dim. = 200; decoder dim. = 100; Latent space dim. = 200; α_k = 0.4. K-AE used to cluster men had the following parameters: encoder dim. = 150; decoder dim. = 200; Latent space dim. = 150; α_k = 0.3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

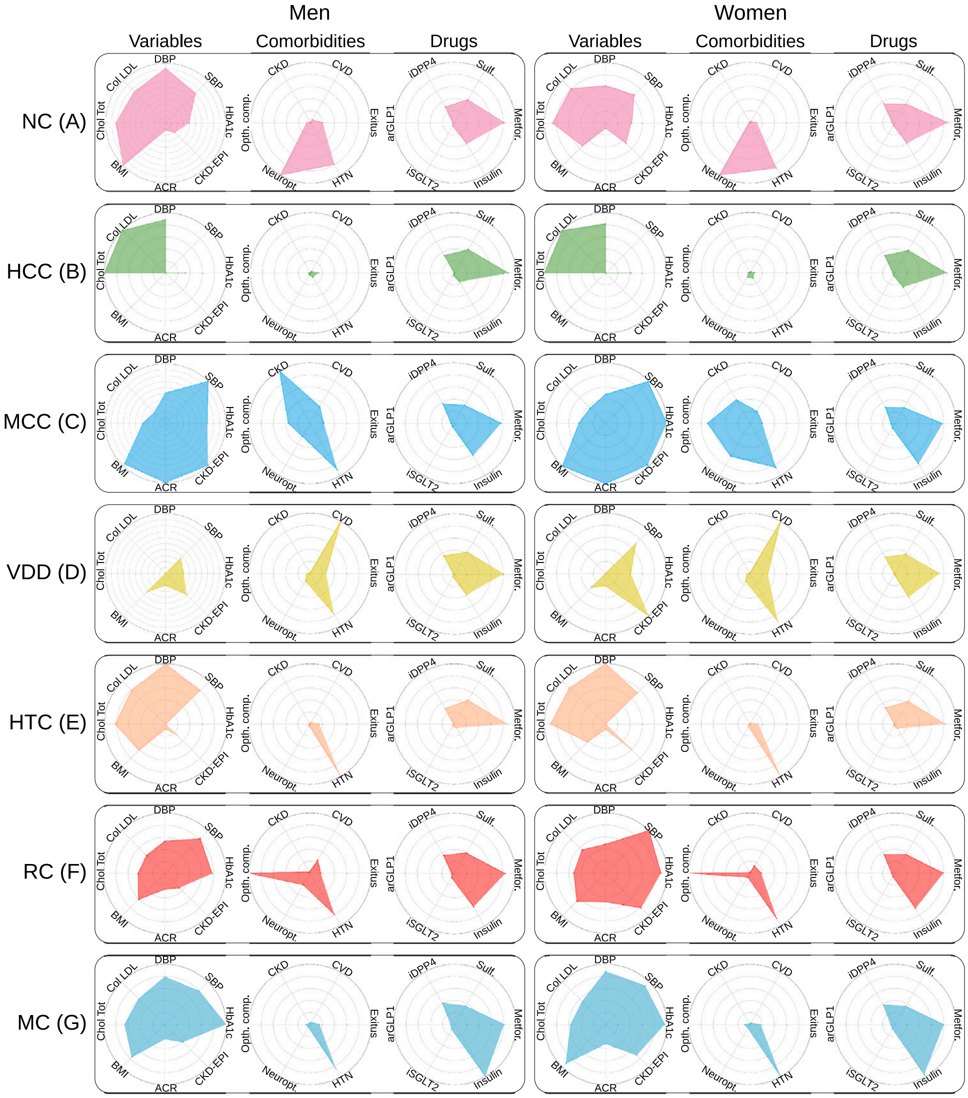

Fig. 2. Graphical representation of variables, comorbidities and drug usage for the seven clusters (Neuropatic Cluster-NC, Hypercholesteraemic Cluster-HCC, Multiple Complications Cluster-MCC, Vascular Disease Cluster-VDC, Hypertensive Cluster-HTC, Retinopathy Cluster-RC, and Metabolic Cluster-MC) for women and men. Variables plots represent the mean value of all the measures of the corresponding cluster over the analyzed period, comorbidities plots represent percentage of patients in the cluster that have been diagnosed for the corresponding diseases and drugs plots represent the percentage of patients in the cluster that used at least once one medication in the corresponding class. Note that CKD-EPI represents 100-eGFR in order to be coherent with other variables where higher values indicate worse values. (Used abbreviation: Chol. Tot.-Total Cholesterol, Opth.Comp. = Ophthalmological Complications; Neuropt. = Neuropathies; Sulf. = Sulfonylureas, Metfor. = Metformin.).

Table 1
Main statistics for the seven clusters (Neuropathic Cluster-NC, Hypercholesteraemic Cluster-HCC, Multiple Complications Cluster-MCC, Vascular Disease Cluster-VDC, Hypertensive Cluster-HTC, Retinopathy Cluster-RC, and Metabolic Cluster-MC).

		NC (A)	HCC (B)	MCC (C)	VDC (D)	HTC (E)	RC (F)	MC (G)
# patients	All	832	1472	562	1305	4135	1232	1490
	Male	315	831	250	893	2065	724	724
	Fem.	517	641	312	412	2070	508	766
Age[1] [y] Mean±SD[2]	All	69.4 ± 10.2	65.3 ± 11.6	69.3 ± 10.3	72.8 ± 9.4	71.1 ± 9.9	68.9 ± 10.5	69.2 ± 10.9
	Male	68 ± 10.4	64.7 ± 11.6	69 ± 9.5	71.4 ± 9.4	69.2 ± 9.9	66.8 ± 10.1	66.7 ± 11.1
	Fem.	70.3 ± 10.1	66.2 ± 11.5	69.6 ± 11	75.7 ± 8.9	73.0 ± 49.7	72 ± 10.3	71.5 ± 10.1
Age at diagnosis [y] Mean±SD	All	58.7 ± 10.5	56.1 ± 11.5	57.4 ± 11.6	62.2 ± 9.8	61.6 ± 10	56.5 ± 11.2	57.8 ± 11.2
	Male	57.6 ± 10.4	55.8 ± 11.4	58.1 ± 10.4	61 ± 9.6	60 ± 9.9	54.8 ± 10.5	55.8 ± 11.3
	Fem.	59.4 ± 10.5	56.6 ± 11.6	56.9 ± 12.5	64.7 ± 9.9	63.2 ± 9.8	59 ± 11.9	59.8 ± 10.8
T2DM duration [y] Mean±SD	All	10.7 ± 5.1	9.2 ± 3.9	11.9 ± 6.5	10.6 ± 4.8	9.5 ± 3.9	12.4 ± 6.7	11.3 ± 5.7
	Male	10.4 ± 4.5	8.9 ± 3.4	10.8 ± 5.4	10.4 ± 4.5	9.2 ± 3.6	12 ± 6.2	10.9 ± 5.4
	Fem.	10.9 ± 5.4	9.5 ± 4.4	12.7 ± 7.2	11 ± 5.3	9.8 ± 4.1	13 ± 7.2	11.7 ± 5.9
Death [%]	All	14.66	10.87	21	26.9	14.07	17.61	16.31
	Male	20.32	13.48	22.4	25.98	14.67	17.27	15.33
	Fem.	11.22	7.49	19.87	28.88	13.48	18.11	17.23

[1] Age calculated on January 1st, 2013.

[2] SD: Standard Deviation.

Similar mean ages were observed for clusters A, C, F, and G. Patients in cluster B were the youngest in terms of age and age at diagnosis, with mean values that were statistically lower than all other clusters. The mean age and age at diagnosis of clusters D and E were statistically higher than all other clusters but the difference between these two clusters was not statistically significant. The subjects in cluster F had a longer period of diabetes duration; however, this value was statistically similar to cluster C. In the clusters B and E, the subjects had a shorter period of diabetes duration, with values that are statistically similar (pairwise t-test with Benjamini–Hochberg adjustment, $\alpha = 0.05$ [28]). The percentage of deaths was highest in cluster D, followed by cluster C, while cluster B was the one with the lowest death percentage.

The clusters showed different clinical profiles, as reported in Figs. 2, 3 and in Supplementary Figures S2–S10 and Table S4 . Regarding glycaemic control, the mean HbA1c ranged from 7.0% to 8.0%. Cluster G had the worst glycaemic control, while cluster E the best glycaemic control. The variability of HbA1c was different among the clusters. In cluster C, we observed stable level of HbA1c, while in clusters A, B, D, E and F we observed upward trends, with markedly different slopes, that of cluster B being more than double that of other clusters. In cluster G, we observed a downward slope over the analyzed period. The majority of the subjects in the clusters A, C, E, F and G were obese (BMI>30), while in clusters B and D subjects were overweight (BMI>25). Cluster C had the highest, and cluster B the lowest, mean BMI. This variable tended to decrease over time in the clusters A, B, D, E and G; on the contrary, it tended to increase for clusters C and F.

Regarding the lipid profile, subjects in the cluster A, B and E were characterized by high levels for total cholesterol, LDL and HDL, being cluster B the one with the highest values, while triglycerides were higher in cluster C and G than the rest of the clusters. Conversely, cluster D was characterized by the lowest levels for HDL, LDL and total cholesterol. For all clusters LDL and total cholesterol tend to decrease over time, as well as triglycerides, even if in this case this trend is significant only for clusters B, C, E and G, with the slope of cluster G being more than double of the one of the other clusters. On the contrary, HDL remains stable over time, except for clusters B and G where it tends to increase. The renal profile among the clusters was also different. The mean glomerular filtration rate (eGFR) in all clusters ranged from 56.3 to 78.1 ml/min/1.73m^2. The lowest (eGFR) was observed in cluster C (which also had the highest ACR and creatinine levels); furthermore, in this cluster we observed a higher decrease of the eGFR. On the contrary, cluster B had the highest eGFR and the lowest ACR and Creatinine, with values that remained stable over time.

Metformin was the most frequently prescribed drug in all of the clusters, except in cluster G subjects in whom the most frequently prescribed drug was insulin. In cluster B, 86.6% of the subjects had a metformin prescription at some time during the analysis period. This percentage is significantly higher than in all other clusters. In contrast, cluster C was the one with the lower percentage of patients treated with metformin (77.4%). Sulphonylureas were mainly prescribed in cluster E and B (44,6% and 44.4%) with percentages that are statistically similar to each other, and higher than in all other clusters. In cluster G, the highest percentages of prescriptions of all other classes of antidiabetic drugs (SGLT-2i, DPP-4i and arGLP1) occurred. Subjects in this cluster were also more frequently insulin-treated (97,5%), followed by cluster C (70.28%), while subjects in cluster E were the least frequently treated with insulin (8.6%). In all cases, the percentages of insulin-treated patients were significantly different among all clusters. The results of the usage of antidiabetic drugs among the clusters are presented in Supplementary Table S5.

The distribution of common comorbidities among the clusters was also different. Cardiovascular complications were almost entirely represented in cluster D (the oldest), significantly higher than in all other clusters. For instance, cardiovascular disease was only found in 2.2% of the subjects from clusters B and E. Neuropathy occurred mostly in cluster A, while in cluster E, the complication only occurred in 4.1% of the subjects. Renal complications were primarily present in cluster C (69.4%), while the percentage in other clusters was much lower, ranging from 1.88% in cluster E to 0.2% in cluster B. High blood pressure was very common in all clusters, except for cluster B: in cluster E it was present in all patients, while in other clusters it ranged from 91% to 81%, except for cluster B where it was 9.24%. In contrast, ophthalmological complications were primarily prevalent in cluster F, followed by cluster C, while in other clusters, this complication occurred less frequently.

Due to differential clinical factors, we named cluster A as Neuropathic Cluster (NC), cluster B as Hypercholesteraemic Cluster (HCC), cluster C as Multiple Complications Cluster (MCC), cluster D as Vascular Disease Cluster (VDC), cluster E as Hypertensive Cluster (HTC), cluster F as Retinopathy Cluster (RC) and cluster G as Metabolic Cluster (MC).

4. Discussion

Results of our unsupervised deep learning cluster analysis suggest seven longitudinal phenotype clusters of T2DM persons with different clinical evolutions.

Health care professionals often need to identify, quantify, and interpret relationships among different health variables [29]. Access to big real-world clinical data and the advances in artificial intelligence applications in recent years has made it possible to develop algorithms that generally are more accurate for the prediction and classification of patients [29]. Increasingly, deep learning is changing the analysis

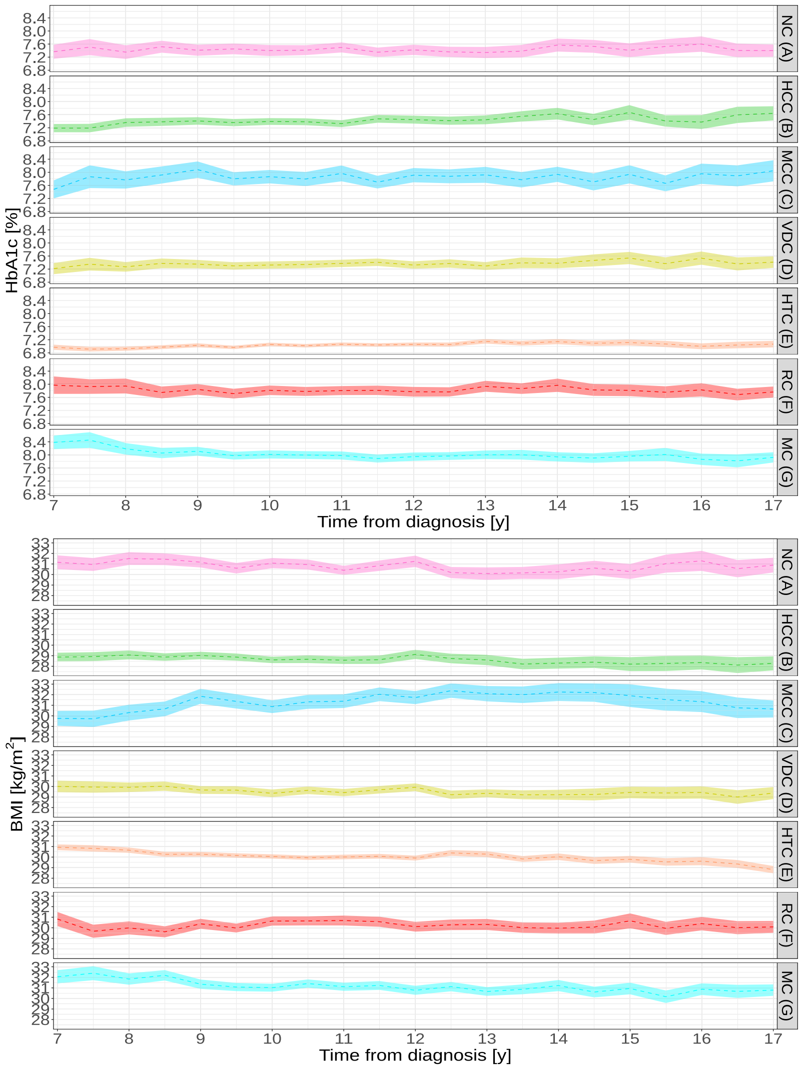

Fig. 3. HbA1c and BMI evolution calculated over a period of 10 years for the seven clusters (Neuropathic Cluster-NC, Hypercholesteraemic Cluster-HCC, Multiple Complications Cluster-MCC, Vascular Disease Cluster-VDC, Hypertensive Cluster-HTC, Retinopathy Cluster-RC, and Metabolic Cluster-MC). Dot lines represent the mean values while intervals represent lower and upper quantiles computed with 0.95 confidence intervals.

of databases extracted from EHRs. It requires less manual feature engineering, while the high volume and coverage of healthcare datasets enable successful training of complex deep learning models [30].

In general, architectures based on AEs have proven to be optimal when dealing with data extracted from EHR. In DeepPatient [13], the authors suggest that a vector of features learned without supervision can be used effectively for several tasks, from disease prediction to clustering. In [12], the authors showed that an architecture based on a variational AE can be used to cluster longitudinal data from patients, resulting in purer clusters than in methods not based on unsupervised deep learning. Similar results were achieved in [31], where a denoising AE was used to derive patients representations to be used for prediction and study of patient similarity. As for these studies, also our results suggest that patient representation obtained through our K-AE model can be used to obtain better clusters with respect to classical techniques not based on DL.

The diagnosis of T2DM is made on chronically elevated blood glucose concentrations and on exclusion of other forms of diabetes (autoimmunity, pregnancy, pancreatic disease…) [32]. The fact that this diagnosis process is based only on one consequence of the diseases (high HbA1c) and on exclusion criteria should admonish us to the lack of understanding we have of the diseases causes and on the urge to develop precision medicine for diabetes [32]. One step further is this field is represented by the stratification of T2DM patients in different groups on the basis of different clinical characteristics [33]. To date, several studies have been published regarding the possible phenotyping of T2DM, although none of them clustered patients based on the evolution of the disease. They rather stratified them based on observations at a single time (e.g. at diagnosis), possibly analyzing the evolution of the diseases for a limited period of time after the observations, but only once the patients were stratified. This limits the affidability of the clusters and also their applicability in clinical practice, since these studies are often based on the measures of variables that are not routinely collected in clinical practice [15]. To the best of our knowledge, this is the first time that the evolution of T2DM has been investigated through longitudinal clusters. Our algorithm, in fact, used a K-AE to map 5 years of routinely collected data from EHR in a latent space that embedded the different characteristics and trajectories of patient phenotypes.

One of the most recent works related to T2DM phenotype clustering was realized by Ahlqvist et al. [3]. Five different clusters were identified based on six different variables measured at diagnosis, these being glutamic acid decarboxylase (GAD) antibodies, age at diagnosis, BMI, HbA1c, and homoeostasis model estimates of β-cell function (HOMA-B) and insulin resistance (HOMA-IR). According to clustering of subjects based on these phenotypic variables, five subgroups were defined: SAID severe autoimmune diabetes; SIDD, severe insulin-deficient diabetes; SIRD, severe insulin-resistant diabetes; MOD, mild obesity-related diabetes; MARD, mild age-related diabetes. Even if this study has been confirmed by subsequent works, although with some differences due to the variables used [34] or the population [35], a recent study suggested that in five years from diagnosis, approximately a quarter of the patients changed cluster with respect to their initial classification [36]. This result suggests that T2DM is a complex disease that shows different phenotype trajectories that can hardly be described by a cross sectional study.

Our deep learning algorithm suggested 7 typical trajectories inside the diabetic population with different evolutions of the disease. The Multiple Complications and Metabolic clusters have the highest BMI, a high percentage of insulin drugs use (70.3% and 97.5%, respectively) and a percentage of Metformin usage lower than in other clusters. Moreover, in the Multiple Complications Cluster, we observed higher percentages of renal complications (69.4%) and the highest levels of ACR, with signs of macroalbuminuria. In the Metabolic Cluster the percentage of renal complications is much lower but still it is the second highest one. This is also the only cluster in with both BMI and HbA1c tend to decrease over time. It is interesting to note that those two clusters share some similarities with the SIRD cluster (high BMI, renal complications,..) and that they could represent two possible evolutions of patients that at diagnosis were assigned to this cluster, with the Metabolic Cluster characterized by a more intense pharmacological treatment. The Vascular Disease and Hypertensive clusters were composed by older subjects (compared with the rest of the clusters at the moment of diagnosis) and both of them showed better glycaemic control with respect to other clusters (mean HbA1c 7.3%±1.3 and 7.0%±1.1, respectively). Also in this case, these two clusters share similarities with the MARD cluster and they show different treatments, with the Vascular Disease cluster characterized by a higher use of insulin. The Hypercholesteraemic cluster was characterized by an early age at diagnosis of T2DM, and relatively low BMI, that tends to decrease over time, sharing some similarities with Ahlqvist SAID and SIDD clusters. Their SIDD cluster can also be compared with our Retinopathy Cluster, characterized by a similar age at diagnosis but a higher BMI that slightly increases over time with respect to the HC. Moreover, both SIDD and our RC clusters had the highest percentage of ophthalmological complications. Finally, the Neuropathic cluster is similar to the RC in terms of Hba1C and BMI but with a higher age at diagnosis and a shorter diabetes duration.

Recent guidelines on treatment of diabetes highlighted the importance of personalized treatments for risk management of patients with T2DM [37]. From this point of view, subclassification of T2DM patients could be a powerful tool for the personalization of diabetic management [38]. The treatment of different risk factors and complications (such as CVD, obesity, CKD, and others) is in fact often decided separately, without discussions on the whole clinical phenotypic evolution of diabetes in each individual patient [38]. Our unsupervised algorithm suggested 7 different longitudinal clusters of patients that showed different clinical profiles, but that also had different treatments and comorbidities. As mentioned before, for instance, the Metabolic and the Multiple Complication Clusters share common phenotypic characteristics such as age at diagnosis, diabetes duration and high levels of HbA1c. However, the pharmacological treatment in the Metabolic Cluster is more intensive, while the rate of patients developing comorbidities (such as CVD and ophthalmological) were much lower. Further studies will be necessary to evaluate the future clinical applications of the proposed stratification. The proposed approach, once validated and improved, might be a useful instrument to support the clinical management of type 2 diabetes.

The present study has some limitations. Firstly, we only have access to a limited amount of health care data and for a limited period of time (5 years). Therefore, the clustering would be more precise if possible linkage with genetic data or specific laboratory parameters related to β-cell function. Secondly, data came from the routine clinical practice of primary care attended T2DM subjects, hence there was missing data. To minimize the effect of this, we only filtered subjects who had a minimum of two variables measured during the analyzed period and had at least one measure of HbA1C and BMI. It should also be noted that recent studies have questioned the utility and existence of different clusters in the T2DM population [39,40]. Besides the clusters described in current work, the main advantage of its technique is to map patient data into a single vector of the latent space. Distances between different vectors embed differences between the patients they represent in terms of the evolution of the disease, obtaining a metric to evaluate patients' similarities. Such representations could represent a further step toward precision medicine in diabetes but also in the management of all the chronic diseases, since it relies only on routinely collected data and can potentially be applied to the study of all chronic diseases.

Another limitation of the methodology is the interpretability. Deep learning models can produce accurate predictions; however, the presence of the black-box models decreases the interpretability and transparency of their inner workings. This may be an important issue since clinicians often are unwilling to accept machine recommendations which lack clarity as to their underlying reasoning [30]. In order to make our results clearer, we developed an interactive dashboard showing the evolution of the clusters but also a representation of the latent space obtained by the K-AE.

In conclusion, using unsupervised deep learning algorithms on routinely collected health care data from primary care centers in Catalonia, seven different clusters were obtained in terms of evolution of clinical characteristics, comorbidities, and important outcomes. Although our finding confirms the heterogeneity of the T2DM subjects, further studies are needed to provide more clinical and genetically enriched data to perform more precise type 2 diabetes subtype clustering. The development of unsupervised deep machine learning might be a valuable tool for future applicability in decision-support to help clinicians in the management of T2DM or other chronic diseases. Ultimately these innovative methodologies will help develop precise strategies to implement and improve disease management, using a more individualized target approach.

CRediT authorship contribution statement

Enrico Manzini: Conceptualization, Data curation, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Bogdan Vlacho:** Data curation, Methodology, Writing – original draft, Writing – review & editing. **Josep Franch-Nadal:** Data curation, Methodology, Writing – review & editing. **Joan Escudero:** Writing – review & editing, Resources. **Ana Génova:** Writing – review & editing, Resources. **Elisenda Reixach:** Conceptualization, Resources. **Erik Andrés:** Conceptualization, Resources. **Israel Pizarro:** Conceptualization, Resources. **José-Luis Portero:** Conceptualization, Resources. **Dídac Mauricio:** Data curation, Methodology, Writing – review & editing. **Alexandre Perera-Lluna:** Conceptualization, Resources, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Enrico Manzini reports was provided by Pulso.

Enrico Manzini, Alexandre Perera-Lluna reports a relationship with Spanish Ministry of Economy and Competitiveness that includes: funding grants.

Enrico Manzini, Alexandre Perera-Lluna reports a relationship with Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine that includes: funding grants.

Enrico Manzini, Alexandre-Perera-Lluna reports a relationship with Instituto de Investigación Carlos III that includes: funding grants.

Enrico Manzini, Alexandre Perera-Lluna reports a relationship with Share4Rare project that includes: funding grants. Josep Franch-Nadal, Didac Mauricio reports a relationship with AstraZeneca Pharmaceuticals LP that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with Ascensia Diabetes Care pain L that includes: speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with Boehringer Ingelheim GmbH that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with G K that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with Lilly pain that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with M D that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with Novartis Pharmaceuticals Corporation that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with Novo Nordisk Inc that includes: funding grants and speaking and lecture fees.

Josep Franch-Nadal, Didac Mauricio reports a relationship with Sanofi that includes: funding grants and speaking and lecture fees.

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness (www.mineco.gob.es) TEC2014-60337-R, DPI2017-89827-R, Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Spain, initiatives of Instituto de Investigación Carlos III (ISCIII), and Share4Rare project (Grant Agreement 780262). This study was possible thanks to the commitment of physicians and nurses working in the Catalan Health Institute to provide optimal care to patients with diabetes. CIBER of Diabetes and Associated Metabolic Diseases (CIBERDEM) is an initiative from Instituto de Salud Carlos III, Madrid, Spain.

This analysis is part of the DiaCare Project of Novo Nordisk and the Fundació TicSalut (Departament de Salut, Generalitat de Catalunya), in collaboration with Grupo Pulso, for the benefit of people with type 2 diabetes.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104218.

References

[1] R. Busse, M. Blümel, D. Scheller-Kreinsen, A. Zentner, Tackling chronic disease in Europe. Strategies, interventions and challenges, Political Science, Vol. 2009, 2010.

[2] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A.A. Motala, K. Ogurtsova, J.E. Shaw, D. Bright, R. Williams, Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition, Diabetes Res. Clin. Pract. 157 (2019) http://dx.doi.org/10.1016/j.diabres.2019.107843.

[3] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R.B. Prasad, D.M. Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spégel, H. Mulder, E. Lindholm, O. Melander, O. Hansson, U. Malmqvist, A. Lernmark, K. Lahti, T. Forsén, T. Tuomi, A.H. Rosengren, L. Groop, Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables, Lancet Diabetes Endocrinol. 6 (5) (2018) 361–369, http://dx.doi.org/10.1016/S2213-8587(18)30051-2.

[4] A. Bonnefond, P. Froguel, Clustering for a better prediction of type 2 diabetes mellitus, Nat. Rev. Endocrinol. 17 (4) (2021) http://dx.doi.org/10.1038/s41574-021-00475-4.

[5] E. Blind, H. Janssen, K. Dunder, P. Graeff, The European Medicines Agency's approval of new medicines for type 2 diabetes, Diabetes Obes. Metab. 20 (2017) http://dx.doi.org/10.1111/dom.13349.

[6] M. Mata-Cases, D. Mauricio, I. Vinagre, R. Morros, E. Hermosilla, F. Fina, M. Rosell Murphy, C. Castell, J. Franch-Nadal, B. Bolíbar, Treatment of hyperglycaemia in type 2 diabetic patients in a primary care population database in a Mediterranean Area (Catalonia, Spain), Diabetes Metab. 5 (2014) 338, http://dx.doi.org/10.4172/2155-6156.1000338.

[7] S.G. Nicholls, S.M. Langan, E.I. Benchimol, Routinely collected data: The importance of high-quality diagnostic coding to research, CMAJ 189 (33) (2017) E1054–E1055, http://dx.doi.org/10.1503/cmaj.170807, URL https://www.cmaj.ca/content/189/33/E1054.

[8] J.V. Stidsen, J.E. Henriksen, M.H. Olsen, R.W. Thomsen, J.S. Nielsen, J. Rungby, S.P. Ulrichsen, K. Berencsi, J.A. Kahlert, S.G. Friborg, I. Brandslund, A.A. Nielsen, J.S. Christiansen, H.T. Sørensen, T.B. Olesen, H. Beck-Nielsen, Pathophysiology-based phenotyping in type 2 diabetes: A clinical classification tool, Diabetes/Metab. Res. Rev. 34 (5) (2018) e3005, http://dx.doi.org/10.1002/dmrr.3005, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/dmrr.3005, e3005 DMRR-17-RES-100.R2.

[9] W. Oh, E. Kim, M.R. Castro, P.J. Caraballo, V. Kumar, M.S. Steinbach, G.J. Simon, Type 2 diabetes mellitus trajectories and associated risks, Big Data 4 (1) (2016) 25–30, http://dx.doi.org/10.1089/big.2015.0029.

[10] I. Walraven, M.R. Mast, T. Hoekstra, A. Jansen, A.A. van der Heijden, S.P. Rauh, F. Rutters, E. van't Riet, P.J. Elders, A.C. Moll, et al., Distinct HbA1c trajectories in a type 2 diabetes cohort, Acta Diabetol. 52 (2) (2015) 267–275, http://dx.doi.org/10.1007/s00592-014-0633-8.

[11] R.A. Hubbard, J. Xu, R. Siegel, Y. Chen, I. Eneli, Studying pediatric health outcomes with electronic health records using Bayesian clustering and trajectory analysis, J. Biomed. Inform. 113 (2021) 103654, http://dx.doi.org/10.1016/j.jbi.2020.103654.

[12] J. de Jong, M.A. Emon, P. Wu, R. Karki, M. Sood, P. Godard, A. Ahmad, H. Vrooman, M. Hofmann-Apitius, H. Fröhlich, Deep learning for clustering of multivariate clinical patient trajectories with missing values, GigaScience 8 (11) (2019) 1–14, http://dx.doi.org/10.1093/gigascience/giz134.

[13] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: An unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (May) (2016) 1–10, http://dx.doi.org/10.1038/srep26094.

[14] F. Xie, H. Yuan, Y. Ning, M.E.H. Ong, M. Feng, W. Hsu, B. Chakraborty, N. Liu, Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies, J. Biomed. Inform. (2021) 103980, http://dx.doi.org/10.1016/j.jbi.2021.103980.

[15] W. Chung, K. Erion, J. Florez, A. Hattersley, m.-f. Hivert, C. Lee, M. McCarthy, J. Nolan, J. Norris, E. Pearson, L. Philipson, A. McElvaine, W. Cefalu, S. Rich, P. Franks, Precision medicine in diabetes: A consensus report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD), Diabetes Care 43 (7) (2020) 1617–1635, http://dx.doi.org/10.2337/dci20-0022.

[16] F.M. Bianchi, L. Livi, K. Mikalsen, M. Kampffmeyer, R. Jenssen, Learning representations of multivariate time series with missing data, Pattern Recognit. 96 (2019) 106973, http://dx.doi.org/10.1016/j.patcog.2019.106973.

[17] B. Bolíbar, F. Fina Avilés, R. Morros, M.d.M. Garcia-Gil, E. Hermosilla, R. Ramos, M. Rosell, J. Rodríguez, M. Medina, S. Calero, D. Prieto-Alhambra, Grupo SIDIAP, [SIDIAP database: Electronic clinical records in primary care as a source of information for epidemiologic research], Med. Clin. 138 (14) (2012) 617–621, http://dx.doi.org/10.1016/j.medcli.2012.01.020, URL http://www.ncbi.nlm.nih.gov/pubmed/22444996.

[18] G. Brämer, International statistical classification of diseases and related health problems. Tenth revision, World Health Stat. Q. Rapport Trimestriel de Statistiques Sanitaires Mondiales 41 (1) (1988) 32—36, URL http://europepmc.org/abstract/MED/3376487.

[19] ATC classification index with DDDs, 2021, WHO Collaborating Centre for Drug Statistics Methodology, Oslo, Norway.

[20] M. Templ, A. Kowarik, P. Filzmoser, Iterative stepwise regression imputation using standard and robust methods, Comput. Statist. Data Anal. 55 (10) (2011) 2793–2806, http://dx.doi.org/10.1016/j.csda.2011.04.012.

[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[22] K.Ø. Mikalsen, F.M. Bianchi, C. Soguero-Ruiz, R. Jenssen, Time series cluster kernel for learning similarities between multivariate time series with missing data, Pattern Recognit. 76 (2018) 569–581, http://dx.doi.org/10.1016/j.patcog.2017.11.030.

[23] M. Cuturi, Fast global alignment kernels, in: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, 2011.

[24] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, E. Woods, Tslearn, A machine learning toolkit for time series data, J. Mach. Learn. Res. 21 (118) (2020) 1–6, URL http://jmlr.org/papers/v21/20-091.html.

[25] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, URL https://www.tensorflow.org/, Software available from tensorflow.org.

[26] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, Cluster: Cluster analysis basics and extensions, 2021, URL https://CRAN.R-project.org/package=cluster, R package version 2.1.2.

[27] R. Core Team, R: A language and environment for statistical computing, 2020, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

[28] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1) (1995) http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[29] K. Johnson, J. Soto, B. Glicksberg, S. Khader, R. Miotto, M. Ali, E. Ashley, Artificial intelligence in cardiology, J. Am. Coll. Cardiol. 71 (2018) 2668–2679, http://dx.doi.org/10.1016/j.jacc.2018.03.521.

[30] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1419–1428, http://dx.doi.org/10.1093/jamia/ocy068.

[31] L. Lei, Y. Zhou, J. Zhai, L. Zhang, Z. Fang, P. He, J. Gao, An effective patient representation learning for time-series prediction tasks based on EHRs, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, 2019, http://dx.doi.org/10.1109/BIBM.2018.8621542.

[32] H. Fitipaldi, M.I. McCarthy, J.C. Florez, P.W. Franks, A global overview of precision medicine in type 2 diabetes, Diabetes 67 (10) (2018) 1911–1922, http://dx.doi.org/10.2337/dbi17-0045.

[33] A.L. Gloyn, D.J. Drucker, Precision medicine in the management of type 2 diabetes, Lancet Diabetes Endocrinol. 6 (11) (2018) 891–900, http://dx.doi.org/10.1016/S2213-8587(18)30052-4.

[34] R.C. Slieker, L.A. Donnelly, H. Fitipaldi, G.A. Bouland, G.N. Giordano, M. Åkerlund, M.J. Gerl, E. Ahlqvist, A. Ali, I. Dragan, et al., Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: An IMI-RHAPSODY study, Diabetologia 64 (9) (2021) 1982–1989, http://dx.doi.org/10.1007/s00125-021-05490-8.

[35] X. fen Xiong, Y. Yang, L. Wei, Y. Xiao, L. Li, L. Sun, Identification of two novel subgroups in patients with diabetes mellitus and their association with clinical outcomes: A two-step cluster analysis, J. Diabetes Investig. 7 (2021) http://dx.doi.org/10.1126/scitranslmed.aaa9364.

[36] O.P. Zaharia, K. Strassburger, A. Strom, G.J. Bönhof, Y. Karusheva, S. Antoniou, K. Bódis, D.F. Markgraf, V. Burkart, K. Müssig, J.H. Hwang, O. Asplund, L. Groop, E. Ahlqvist, J. Seissler, P. Nawroth, S. Kopf, S.M. Schmid, M. Stumvoll, A.F. Pfeiffer, S. Kabisch, S. Tselmin, H.U. Häring, D. Ziegler, O. Kuss, J. Szendroedi, M. Roden, B.F. Belgardt, A. Buyken, J. Eckel, G. Geerling, H. Al-Hasani, C. Herder, A. Icks, J. Kotzka, E. Lammert, D. Markgraf, W. Rathmann, Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: A 5-year follow-up study, Lancet Diabetes Endocrinol. 7 (9) (2019) 684–694, http://dx.doi.org/10.1016/S2213-8587(19)30187-1.

[37] J.B. Buse, D.J. Wexler, A. Tsapas, P. Rossing, G. Mingrone, C. Mathieu, D.A. D'Alessio, M.J. Davies, 2019 Update to: management of hyperglycemia in type 2 diabetes, 2018. a consensus report by the American diabetes association (ADA) and the European Association for the Study of Diabetes (EASD), Diabetes Care 43 (2) (2020) 487–493.

[38] H. Tanabe, H. Masuzaki, M. Shimabukuro, Novel strategies for glycaemic control and preventing diabetic complications applying the clustering-based classification of adult-onset diabetes mellitus: A perspective, Diabetes Res. Clin. Pract. 180 (2021) 109067, http://dx.doi.org/10.1016/j.diabres.2021.109067.

[39] M. Lugner, S. Gudbjörnsdottir, N. Sattar, A. Svensson, M. Miftaraj, K. Eeg-Olofsson, B. Eliasson, S. Franzén, Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: Nationwide observational study, Diabetologia (2021) 1–9, http://dx.doi.org/10.1007/s00125-021-05485-5.

[40] J.M. Dennis, B.M. Shields, W.E. Henley, A.G. Jones, A.T. Hattersley, Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: An analysis using clinical trial data, Lancet Diabetes Endocrinol. 7 (6) (2019) 442–451, http://dx.doi.org/10.1016/S2213-8587(19)30087-7.