Cardiovascular Diabetology

RESEARCH

Open Access

Machine learning profiles of cardiovascular risk in patients with diabetes mellitus: the Silesia Diabetes-Heart Project

Hanna Kwiendacz[1*], Agata M. Wijata[2], Jakub Nalepa[3], Julia Piaśnik[4], Justyna Kulpa[4], Mikołaj Herba[4], Sylwia Boczek[4], Kamil Kegler[4], Mirela Hendel[4], Krzysztof Irlik[4], Janusz Gumprecht[1], Gregory Y. H. Lip[5,6] and Katarzyna Nabrdalik[1,5]

Abstract

Aims As cardiovascular disease (CVD) is a leading cause of death for patients with diabetes mellitus (DM), we aimed to find important factors that predict cardiovascular (CV) risk using a machine learning (ML) approach.

Methods and results We performed a single center, observational study in a cohort of 238 DM patients (mean age ± SD 52.15 ± 17.27 years, 54% female) as a part of the Silesia Diabetes-Heart Project. Having gathered patients' medical history, demographic data, laboratory test results, results from the Michigan Neuropathy Screening Instrument (assessing diabetic peripheral neuropathy) and Ewing's battery examination (determining the presence of cardiovascular autonomic neuropathy), we managed use a ML approach to predict the occurrence of overt CVD on the basis of five most discriminative predictors with the area under the receiver operating characteristic curve of 0.86 (95% CI 0.80–0.91). Those features included the presence of past or current foot ulceration, age, the treatment with beta-blocker (BB) and angiotensin converting enzyme inhibitor (ACEi). On the basis of the aforementioned parameters, unsupervised clustering identified different CV risk groups. The highest CV risk was determined for the eldest patients treated in large extent with ACEi but not BB and having current foot ulceration, and for slightly younger individuals treated extensively with both above-mentioned drugs, with relatively small percentage of diabetic ulceration.

Conclusions Using a ML approach in a prospective cohort of patients with DM, we identified important factors that predicted CV risk. If a patient was treated with ACEi or BB, is older and has/had a foot ulcer, this strongly predicts that he/she is at high risk of having overt CVD.

Keywords Diabetes mellitus, Machine learning, Cardiovascular disease, Prediction model, Michigan neuropathy screening instrument

*Correspondence:
Hanna Kwiendacz
hkwiendacz@sum.edu.pl
Full list of author information is available at the end of the article

Introduction

Every five seconds, one person in the world dies due to diabetes mellitus (DM) and its complications [1]. Cardiovascular disease (CVD) takes the greatest toll here, as it is a leading cause of death for both patients with type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM) [2]. DM itself doubles the risk of coronary artery disease, ischemic stroke, and vascular deaths, independently of other risk factors [3]. It is estimated that patients with DM and CVD and/or chronic kidney disease as well as those having at least three CVD risk factors or with DM duration over 20 years, are at a very high risk, with the 10-year risk of CVD death exceeding 10% [4]. Hence, it is clinically important to identify patients with the highest CVD risk to implement preventive procedures against cardiovascular (CV) events which may otherwise lead to death.

Although several calculators assessing CVD have been developed [5, 6], they are usually validated in a general population and do not accurately assess the risk among patients with DM [7, 8]. Therefore, dedicated risk models are proposed for this group [9], but yet it remains unclear which one is optimal [10]. Moreover, the risk assessment models are not personalized, hence they do not exploit patient-specific granular information that could be crucial in understating the risk of a particular person. Furthermore, they utilize classical statistical methods which commonly show only a potential association for the population studied but cannot necessarily predict the individual's risk [8]. That is why personalized prediction tools benefiting from data-driven machine learning (ML) approaches which can not only indicate the associations but also anticipate the future risk are of paramount practical importance and attract research much attention.

ML techniques can help uncover new clinical features and relationships between them that are pivotal while identifying high-risk patients, therefore, new risk factors that were not previously taken into consideration in traditional models can emerge, especially in multimorbid high risk patients [11–13]. Lately, unsupervised ML clustering has been successful in the detection of coronary artery atherosclerosis among T2DM patients, whereby on the basis of coronary computed tomography angiography, the algorithm was able to distinguish different plaque types and extents of coronary artery stenosis [14]. Moreover, predicting CVD events with ML models is more effective and offers more accurate estimations in comparison with traditional risk calculators [11, 12].

In our recent work, we demonstrated that ML can be utilized to identify new risk factors for predicting CVD in patients with metabolic-associated fatty liver disease [15], and for predicting cardiovascular events among patients with DM [16]. However, these experiments did not include any subgroup analysis of DM patients with associated diabetic peripheral neuropathy (DPN) or cardiovascular autonomic neuropathy (CAN) and those without these conditions which could be important factors in understanding a personal CVD risk, since those microvascular complications are associated with CVD [17, 18]. To tackle this research gap, we determine the risk of CVD among patients with DM in relation to the presence of diabetic neuropathy, with the use of ML approaches.

Significant efforts are put nowadays into designing precision medicine techniques, where accurate phenotyping of the patients might help in the diagnosis and prognosis of the disease, provide 'real time' risk assessment and improve management via implementation of tailored approaches for individuals [19, 20]. Following this research pathway, we aimed to precisely classify DM patients using prospectively collected granular individual data, with a special emphasis put on the DPN and CAN examination, according to the CV risk, and to profile their phenotypes which might be implemented in everyday clinical practice.

Patients and methods

We performed a single center, observational study in a cohort of T1DM and T2DM consecutive patients hospitalized in the Department of Internal Medicine and Diabetology in Zabrze, Poland, and patients from the Outpatient Diabetology Clinics in the Silesia Region, Poland, from October 6, 2021 to December 15, 2022. This is a part of the Silesia Diabetes-Heart Project (ClinicalTrials.gov Identifier: NCT05626413).

Inclusion and exclusion criteria

The inclusion criteria for the study were as follows: age $\geq$ 18 years and $\leq$ 85 years, T1DM for at least 5 years or T2DM of any duration. The exclusion criteria included: the lack of consent for participation in the study, other than T1DM and T2DM types of diabetes, any severe and acute illness, disabled and bedridden patients, solid organ transplant, other than diabetes previously diagnosed causes of neuropathy, pregnancy, alcoholism, severe hypoglycemia in the past 24 h, an estimated glomerular filtration rate (eGFR) < 30 ml/min/1.73 m^2, and the proliferative retinopathy.

Ethics committee consent

The study was performed in accordance with the ethical principles of the Declaration of Helsinki and approved by the Ethics Committee of the Medical University of Silesia (KNW/0022/KB1/10/17). A written informed consent to participate in the study was obtained from all patients enrolled into this study.

Medical history

Following written informed consent, we collected patients' demographic and clinical data including detailed documented medical history including pharmacotherapy. The presence of CVD was defined as at least one of the following: coronary artery disease, history of coronary revascularization, percutaneous cardiac intervention or coronary artery bypass grafting, atrial fibrillation, history of myocardial infraction or stroke/transient ischemic attack (TIA), carotid atherosclerosis (defined as carotid stenosis of at least 50% in diameter [21]) and/or lower limb atherosclerosis.

Anthropometric measurements

Every participant had anthropometric measurements performed with the use of the weight with height gauge SECA 799, which included the measurement of height (in meters) and body weight (in kilograms). The body mass index (BMI) was calculated by dividing the weight in kilograms by height in meters squared (kg/m^2).

Laboratory test results

On the day of the site visit, fasting venous blood was drawn, and morning urine spots from previous 3 days were collected (on the day of the informed consent signing, each patient has been instructed about the proper management of urine void). The following blood and urine biochemical parameters were assessed: hemoglobin A1c (HbA1c), serum creatinine concentration, lipid profile, and urinary albumin to creatinine ratio (UACR). HbA1c was measured using a high-performance liquid chromatography method (HPLC), and the results were expressed in the National Glycohemoglobin Standardization Program/Diabetes Control and Complications trial units [22]. Serum creatinine concentration was measured using the Jaffe's method [23], and estimated glomerular filtration rate was calculated on the basis of the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula [24]. Enzymatic methods were used to measure cholesterol and triglycerides concentration, whereas the concentration of the low-density lipoprotein cholesterol was calculated through the Friedewald formula [25]. The UACR was estimated using immunoturbidimetric methods and expressed in mg/g creatinine [26].

The diagnosis of diabetic peripheral neuropathy

We used The Michigan Neuropathy Screening Instrument (MNSI) [27] which is designed for patients with diabetes to assess the presence of DPN. It was used for the first time in 1994 and since then it has been widely adopted. The MNSI consists of two separate parts: a 15-item questionnaire and a lower extremity examination. The 15-item questionnaire was completed independently by each patient participating in the study. Positive responses to all questions 1–15 (except for questions 7 and 13) count as 1 point, while to questions 7 and 13 negative responses count as 1 point. All the points are summed up to obtain the final score, and the score ≥ 7 points indicate DPN. The second component of MNSI is a foot examination performed by healthcare professionals. Each foot is inspected for appearance (0–normal; 1–abnormal), ulcerations (0–absent; 1–present), Achilles tendon reflex (0–present; 0.5–present with reinforcement; 1–absent), and vibration sensation tests with a 128-Hz tuning-fork (0–correct; 0.5–reduced; 1–absent). Each foot is assessed separately and the final score is a sum of all the examined aspects. A score ≥ 2.5 is considered abnormal [27]. If a patient had ulceration of foot, we used the term the diabetic foot syndrome [28, 29] and we divided it into the current ulceration (during the MNSI examination) and a healed one (medical history collection).

The diagnosis of cardiovascular autonomic neuropathy

For the diagnosis of CAN, we used DiCAN (Diabetic Cardiac Autonomic Neuropathy, Medicore) which exploits an Ewing battery. In the diagnosis of CAN, the American Diabetes Association recommends the use of Ewing's tests (the so-called Ewing battery developed in the 1970s) [30], including five non-invasive cardiovascular reflex tests, i.e., (1) Heart rate (HR) response to deep breathing; (2) HR response to standing up; (3) Blood pressure response to standing up; (4) Valsalva maneuver; (5) Blood pressure response to sustained handgrip, to assess autonomic functions [30, 31]. The first two tests measure parasympathetic function (primarily the ability of the vagus nerve to slow down HR during heart rate-increasing procedures), while the third and fifth tests measure sympathetic function (blood pressure fluctuations) using baroreceptors. In contrast, the Valsalva maneuver has both parasympathetic and sympathetic components [30]. Subsequently, on the basis of the abovementioned test results, the DiCAN device suggests a diagnosis (normal, early involvement, severe involvement, definite involvement, atypical pattern).

Quality of life

Since diabetic neuropathy may influence quality of life, we tested the patients with the SF-36 questionnaire [32] which is the RAND (research and development) Health Survey (Version 1.0) consisting of 36 items and covering 8 concepts of health: physical functioning, role limitations due to physical health problems, role limitations due to personal or emotional problems, energy/fatigue, emotional well-being, social functioning, bodily pain,

general health. In addition, it also includes a separate item—the health change which indicates a perceived change in health status [32] (Table 1).

Predicting cardiovascular disease using machine learning

The prediction of the occurrence of a CVD for a patient with diabetes was based on demographic data (2 parameters), clinical data (diabetes-related: 3 parameters and concomitant diseases: 5 parameters), laboratory data (5 parameters), data regarding medications (15 parameters), Ewing's battery test results (9 parameters), MNSI results (9 parameters) and quality of life (SF-36) questionnaire (11 parameters). In total, 59 features were evaluated (Table 1), with the average number of missing values of 1.42%. Prior to the implementation of feature selection and CVD prediction mechanisms, the missing data was imputed using factorial analysis [33]. The most discriminative predictors were selected using a χ^2 test following a Monte Carlo approach with 1000 repetitions to ensure the stability of the selected features. In each Monte Carlo iteration, we randomly sampled 80% of patients (with overlaps) for whom the most discriminative predictors were selected by picking the features with $p < 0.05$ obtained by the χ^2 test. Finally, five of the most frequently selected predictors (within 1000 independent repetitions) were considered the most discriminative. It is of note that exploiting the four most frequently picked features leads to statistically significantly worse classification results obtained by the multiple logistic regression (MLR) model (Wilcoxon matched-pairs signed rank test, $p < 0.0001$), whereas using the six most frequently selected predictors leads to over-fitting of MLR to the dataset.

Then, the MLR model was fitted using the selected discriminators, and the optimal cut-point value was extracted from the receiver operating characteristic curve (ROC) using the Index of Union approach [34]. We also performed unsupervised hierarchical clustering of all patients based on the selected features [35]—the number of groups (clusters) was determined in two ways: (*i*) by considering the binary division into two clusters (we hypothesize that the patients can be grouped into those at high- and low- risk of having CVD), and (*ii*) by determining the optimal number of clusters using the Calinski-Harabasz qualitative criterion [36]. In (*ii*), the cohort of patients may be clustered into a larger number of groups (depending on the patients' characteristics), potentially corresponding to different patient profiles. To evaluate the classification performance, we report sensitivity, specificity, and the percentage of correctly classified (CC) high- and low-risk patients, i.e., with and without CVD. For the MLR model, the ROC curves, alongside the area under those curves (AUC) were determined. Decision curve analysis (DCA) was used to assess the clinical utility of the model.

The MATLAB R2022b environment was used for feature selection, hierarchical clustering and visualization of results, whereas GraphPad Prism 9.4.1 was exploited for MLR. The visualization of multidimensional feature spaces was realized using t-distributed stochastic neighbor embedding (t-SNE) [37].

Results

A group of 249 patients formed the primary eligible population (Fig. 1), of which 238 (mean age ± SD 52.15 ± 17.27 years, 54% women) were qualified for medical examinations and were included in this study. The reasons for non-completion are presented in Fig. 1. Of the patients, 31% had T1DM and 69% had T2DM. CVD was reported in 53/238 (22%) patients.

Feature selection led to obtaining the five most discriminative predictors: beta-blocker (BB) (selected in 965/1000 Monte Carlo iterations), age (868/1000), current left foot ulceration (627/1000), angiotensin converting enzyme inhibitors (ACEi) (567/1000) and healed foot ulceration (558/1000) (Table 1). The classification performance of the MLR model for determining high-risk patients is quantified in Table 2.

The obtained results indicate that 44/53 (83.02%) high- and 137/185 (73.51%) low-risk patients were correctly identified, thus 181/238 (76.05%) of all patients were correctly classified. For patients with and without diagnosed neuropathy, 16/19 (84.21%) and 28/34 (82.35%) high-risk were respectively correctly identified, with 23/37 (62.16%) and 63/148 (42.57%) low-risk patients were correctly determined, indicating a significantly larger number of false positive high-risk detections in the latter group. The predictive performance of the MLR model is further reflected in its sensitivity and specificity, amounting to 0.83 and 0.74, respectively. The area under the ROC curve for this classifier operating on the five most discriminative features reaches AUC: 0.86 (95% CI 0.80–0.91) (Fig. 2a). The clinical utility of MLR was also determined using DCA (Fig. 2b), which shows that above the 7% probability threshold and below 48%, the model had a higher utility in terms of net benefit than alternative treatment strategies, i.e., treating none or all patients. It is of note that exploiting all (n = 59) predictors in the MLR led to over-fitting, hence to memorizing the dataset, thus the ML model was unable to generalize.

Hierarchical clustering of all patients based on the most discriminative patient parameters was performed in two ways. In the first case, patients were divided into two groups (Fig. 3a), as we hypothesize that the patients can be clustered into the high- and low-risk ones. In this case, as for the MLR model, sensitivity,

Table 1 Clinical patient characteristics

Parameter	Patients without CVD (n = 185)	Patients with CVD (n = 53)	*p*-value
Demographic parameters			
Men, n%	88 (47.57%)	22 (41.51%)	0.435
Age (years)	**48.32 ± 17.26**	**65.51 ± 8.51**	**< 0.0001**
Clinical parameters			
Diabetes-related			
BMI (kg/m^2)	28.48 ± 6.17	30.86 ± 5.29	0.003
Duration of diabetes (years)	10.76 ± 8.48	13.79 ± 10.93	0.078
Type of diabetes (% of type 1)	1.63 ± 0.48	1.91 ± 0.30	< 0.001
Concomitant diseases			
Arterial hypertension	92 (49.73%)	48 (90.57%)	< 0.001
Chronic kidney disease	78 (42.16%)	32 (60.38%)	0.019
Healed foot ulceration	**4 (2.16%)**	**3 (5.66%)**	**0.184**
Diabetic peripheral neuropathy	11 (5.95%)	3 (5.66%)	0.938
Diabetic retinopathy	16 (8.65%)	7 (13.21%)	0.322
Laboratory parameters			
eGFR (ml/min/1.73m^2)	91.52 ± 22.48	74.50 ± 19.56	< 0.0001
HbA1c (%)	8.66 ± 2.19	8.61 ± 2.06	0.958
High density lipoprotein	1.50 ± 0.49	1.40 ± 0.47	0.196
Total cholesterol (mmol/l)	4.72 ± 1.17	4.84 ± 1.51	0.989
Triglycerides (mmol/l)	1.62 ± 1.08	1.78 ± 0.91	0.117
Pharmacotherapy			
ACEi	**49 (26.49%)**	**31 (58.49%)**	**< 0.001**
Alpha-lipoic acid	1 (0.54%)	2 (3.77%)	0.063
Antidepressants	10 (5.41%)	6 (11.32%)	0.129
Antiepileptic drugs	7 (3.78%)	5 (9.43%)	0.097
ARB	23 (12.43%)	9 (16.98%)	0.392
ASA	22 (11.89%)	34 (64.15%)	< 0.0001
Beta-blocker	**46 (24.86%)**	**37 (69.81%)**	**< 0.0001**
GLP-1 RA	10 (5.41%)	6 (11.32%)	0.129
Insulin	122 (65.95%)	35 (66.04%)	0.990
Metformin	96 (51.89%)	36 (67.92%)	0.038
NOAC	0 (0.00%)	3 (5.66%)	0.001
SGLT-2i	31 (16.76%)	12 (22.64%)	0.326
Statin	53 (28.65%)	37 (69.81%)	< 0.0001
Type SGLT-2i	0.37 ± 0.86	0.49 ± 0.97	0.334
VKA	0 (0.00%)	2 (3.77%)	0.008
Diabetic cardiovascular autonomic neuropathy			
Blood pressure analysis in the standing position	0.54 ± 0.72	0.90 ± 0.85	0.005
Blood pressure analysis in the standing position after 1 min	0.50 ± 0.72	0.55 ± 0.78	0.829
Blood pressure analysis in the standing position after 3 min	0.45 ± 0.71	0.55 ± 0.76	0.334
Handgrip test	0.54 ± 0.81	0.50 ± 0.79	0.812
Heart rate variation with deep breathing	0.82 ± 0.93	1.32 ± 0.89	0.001
Test 30:15	0.36 ± 0.65	0.47 ± 0.67	0.166
Valsalva test after 1 minutre	0.34 ± 0.54	0.51 ± 0.71	0.197
Valsalva test after 20 s	0.36 ± 0.69	0.83 ± 0.96	0.002
Ewing's battery test result	1.79 ± 1.72	2.38 ± 1.50	0.017
Diagnostic parameters of peripheral neuropathy			
Achilles tendon reflex in the left foot	0.03 ± 0.15	0.02 ± 0.14	0.615
Achilles tendon reflex in the right foot	0.02 ± 0.12	0.02 ± 0.14	0.913

Table 1 (continued)

Parameter	Patients without CVD (n = 185)	Patients with CVD (n = 53)	*p*-value
Left foot appearance	66 (35.68%)	31 (58.49%)	0.003
Current left foot ulceration	**2 (1.08%)**	**2 (3.77%)**	**0.179**
Right foot appearance	69 (37.30%)	31 (58.49%)	0.006
Current right foot ulceration	4 (2.16%)	3 (5.66%)	0.184
Vibration perception in the left foot	0.05 ± 0.20	0.12 ± 0.29	0.028
Vibration perception in the right foot	0.04 ± 0.17	0.10 ± 0.25	0.013
Peripheral neuropathy score	14 (7.57%)	11 (20.75%)	0.006
Quality of life questionnaire			
SF36 Emotional well being	60.37 ± 16.60	59.55 ± 16.37	0.629
SF36 Energy fatigue	55.00 ± 17.81	52.74 ± 16.69	0.260
SF36 General health	51.80 ± 17.25	46.04 ± 20.29	0.032
SF36 Health change	41.94 ± 23.29	38.68 ± 22.77	0.235
SF36 Pain	71.16 ± 29.17	58.87 ± 28.81	0.004
SF36 Physical functioning	80.55 ± 23.74	59.43 ± 25.58	< 0.0001
SF36 Role limitation due to emotional problems	65.76 ± 39.45	62.26 ± 41.36	0.592
SF36 Role limitation due to physical health	54.37 ± 41.70	42.45 ± 39.40	0.070
SF36 Social functioning	75.34 ± 27.72	75.00 ± 29.32	0.941
MNSI score	3.84 ± 2.86	5.19 ± 3.21	0.007
PDN diagnosis according to MNSI score	37 (20.00%)	19 (35.85%)	0.016

ACEi Angiotensin Converting Enzyme Inhibitor, *ARB* Angiotensin II Receptor Blocker, *ASA* Acetylsalicylic Acid, *BMI* Body Mass Index, *GLP-1 RA* Glucagon-like Peptide-1, *HbA1c* Hemoglobin A1c, *HDL-C* High Density Lipoprotein Cholesterol, *PDN* Peripheral Diabetic Neuropathy, *SGLT-2i* Sodium-Glucose Cotransporter-2, *NOAC* Novel Oral Anticoagulants, *VKA* Vitamin K Anticoagulant

For each parameter (if applicable), we report its mean ± standard deviation (SD), whereas for each binary parameter, the total number of ones and the percentage of ones are given. The p-values were calculated using either Mann–Whitney *U*-test or χ^2 test as appropriate.

The most discriminative features are rendered in bold

specificity, and the percentages of CC high- and low-risk patients were determined (Table 2). The obtained results indicate high sensitivity of this solution (1.00) with lower specificity compared to the MLR model (0.38 vs. 0.74).

From the clinical assessment point of view, it is less costly to classify a low-risk patient as a high-risk one, rather than to miss a patient at a high risk of CVD—here, hierarchical clustering led to correctly classifying all high-risk patients, with an increased number of false positive classifications (i.e., low-risk patients incorrectly classified as being high-risk). The t-SNE visualization of the two-group clustering allows for identifying the group of patients in which there is no risk of CVD (Fig. 3a, cluster 1), and the group at high risk of CVD. Clustering was also carried out for the optimal number of clusters (5 clusters) determined based on the Calinski-Harabasz criterion (Fig. 3b). The cohort, based on the 5 most discriminating features, was clustered into 5 groups for which we observe an increased risk of CVD, reflected in the increasing number of high-risk patients in the following clusters (Fig. 4). The values of the discriminative patients' parameters for all clusters (in two- and five-group clustering) are summarized in Table 3 and presented in Fig. 4.

Discussion

The key findings of our study are that we managed to determine five out of 59 most discriminative patients' parameters (the presence of past or current foot ulceration, age and the treatment with BB and ACEi) which enabled us to identify patients at risk of CVD. On its basis, we clustered individuals with similar phenotypes in order to stratify their CV risk, showing good predictive value (AUC > 0.8) and clinical usefulness on decision curve analysis.

All of the determined parameters are easy to obtain and interpret, therefore they might be used in everyday practice as they are gathered just based on medical history collection and simple foot visual examination. This finding is of utmost importance as dividing individual patients into high- and low-risk personalized strata enables to tailor a proper treatment pathway for an individual, which stays in line with precision medicine principles [38].

Patient's age was one of the parameters selected by the model, in accordance with existing knowledge

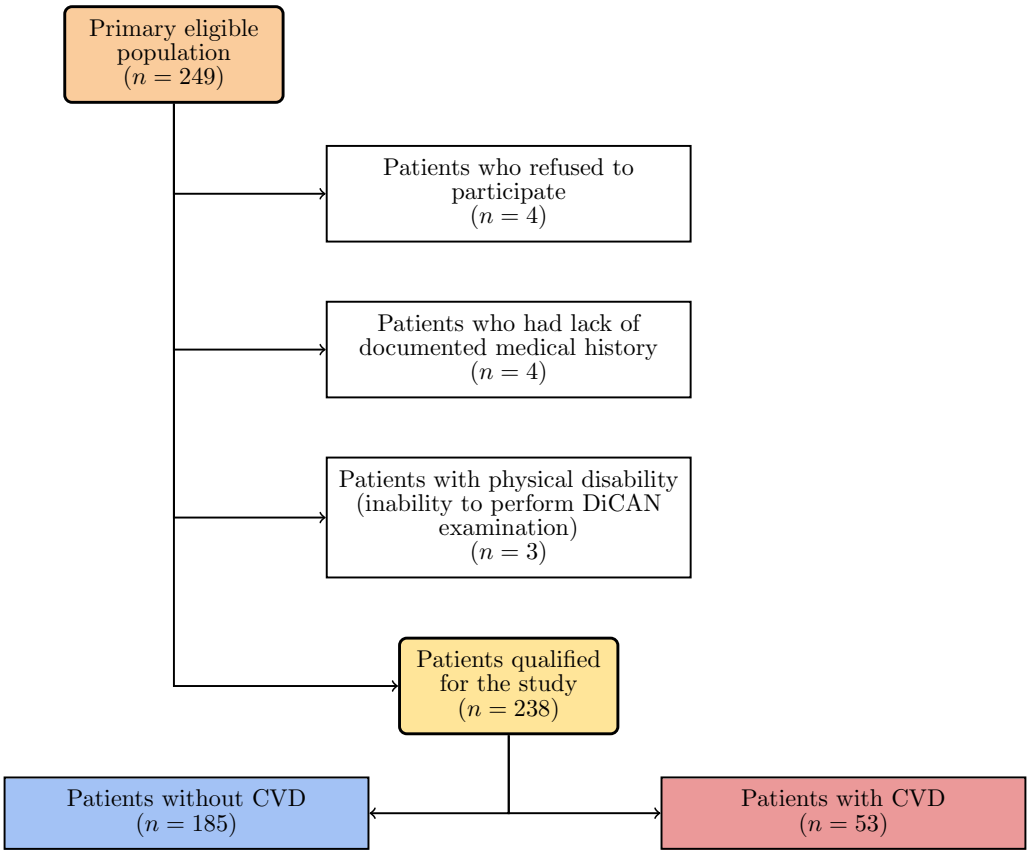

Fig. 1 Patient flowchart

Table 2 Cardiovascular event prediction results based on the most discriminative [5] features using the multiple logistic regression model and hierarchical clustering

ML approach	Sensitivity	Specificity	CC with event, %	CC without event, %	CC All, %
Multiple logistic regression	0.83	**0.74**	83.02	**73.51**	**75.63**
Hierarchical clustering	**1.00**	0.38	**100.00**	37.84	51.68

The best metrics are boldfaced

[39], given that older the patient becomes, the higher the risk of CVD is. Subsequent features were related to foot ulceration. For many years, the presence of the diabetic foot syndrome increases the mortality rate more than twice in comparison with DM patients without ulcerations [40], and results in diminished 5-year survival (43%) compared to non-DM patients with ulcerations (56%) [41]. Indeed, life expectancy for patients after amputation is comparable to advanced congestive heart failure of aggressive neoplasm disease [42]. Therefore, diabetic foot syndrome is believed to be a proxy for CVD [43]. Additionally, foot ulcerations are known to take part in the development of atherosclerosis leading to coronary artery disease and exacerbations of CVDs [44, 45].

Two out of five most discriminative patients' parameters are related to pharmacotherapy, as patients with a higher CV risk more often used BB and ACEi than individuals less prone to CVD. Treatment with those drugs, most probably, does not increase the CV risk per se, but reflect the presence of other comorbidities. The two above-mentioned groups of drugs are commonly used, as the first line therapy, in coronary artery disease, heart failure, and hypertension [46], which demonstrate their utility in CV prediction.

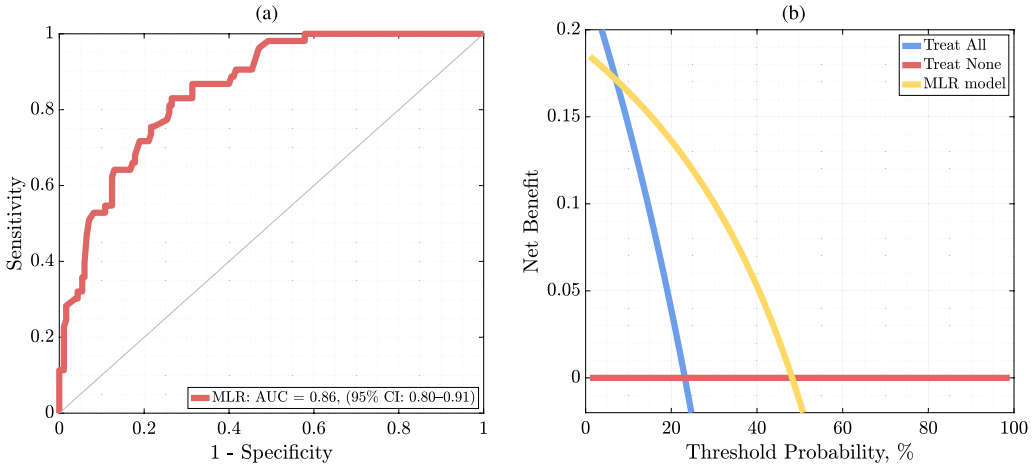

Fig. 2 ROC curve and Decision Curve Analysis. **a** The ROC curve obtained using the multiple logistic regression model fitted over the most discriminative [5] patient's parameters, together with **b** the decision curve analysis presenting clinical utility of the application. In the case of the ROC curve, the 45° curve through the origin shows the classifier's discriminatory ability no better than random sampling

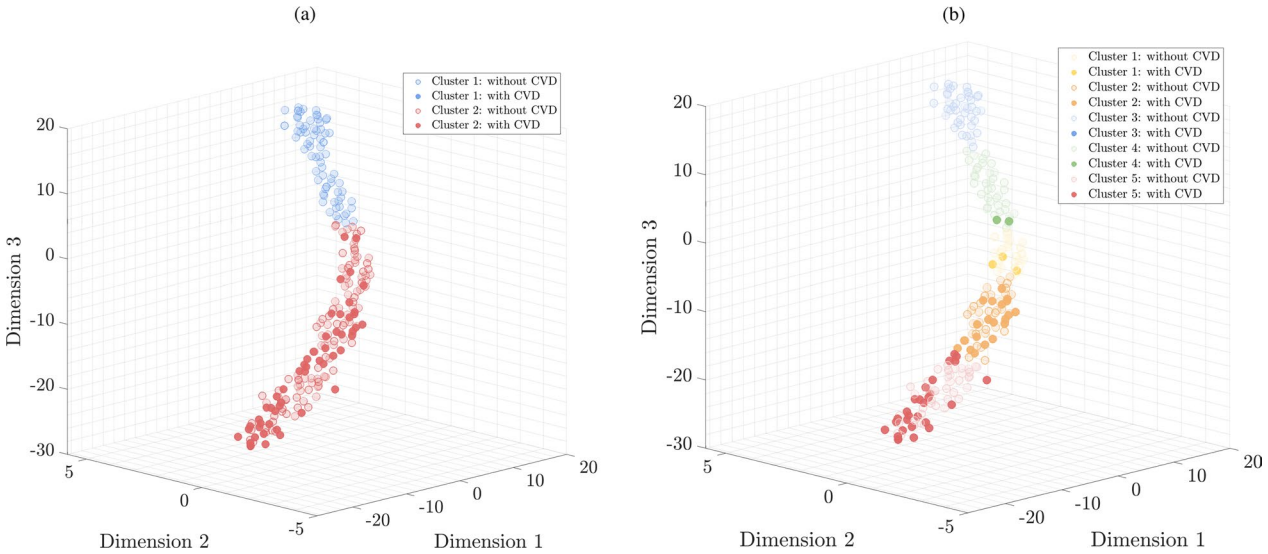

Fig. 3 The t-SNE visualization of two hierarchical clustering results obtained for the most discriminative [5] patient parameters with different numbers of clusters: **a** 2 and **b** 5, respectively

Apart from determining the features increasing the CV risk for DM patients, our analysis clustered patients into high and low-risk strata. Previously, cluster analysis was used to interpret data obtained in the Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS [47]) study and Exenatide Study of Cardiovascular Event Lowering (EXCEL [48]) trial where four distinct phenotypes of patients (differing in terms of CV outcomes) were identified. The highest incidence rate of composite CV outcome was noticed among patients who had the highest mean age (confirmed also in our study), and were predominantly Caucasian males, with the highest median UACR and the lowest eGFR with a prior history of heart failure [49].

In our study, patients were divided into two and five clusters. The first division (2 clusters) was based on the hypothesis that we can infer two clusters of high- and low-risk patients, while the second one (5 clusters) was the automatically determined optimal number of clusters for the analyzed cohort of patients. The two-group clustering revealed that the group of high-risk patients contained the individuals who were older, treated with ACEi and BB, and had past or current foot ulceration. This grouping enabled to correctly classify all of the

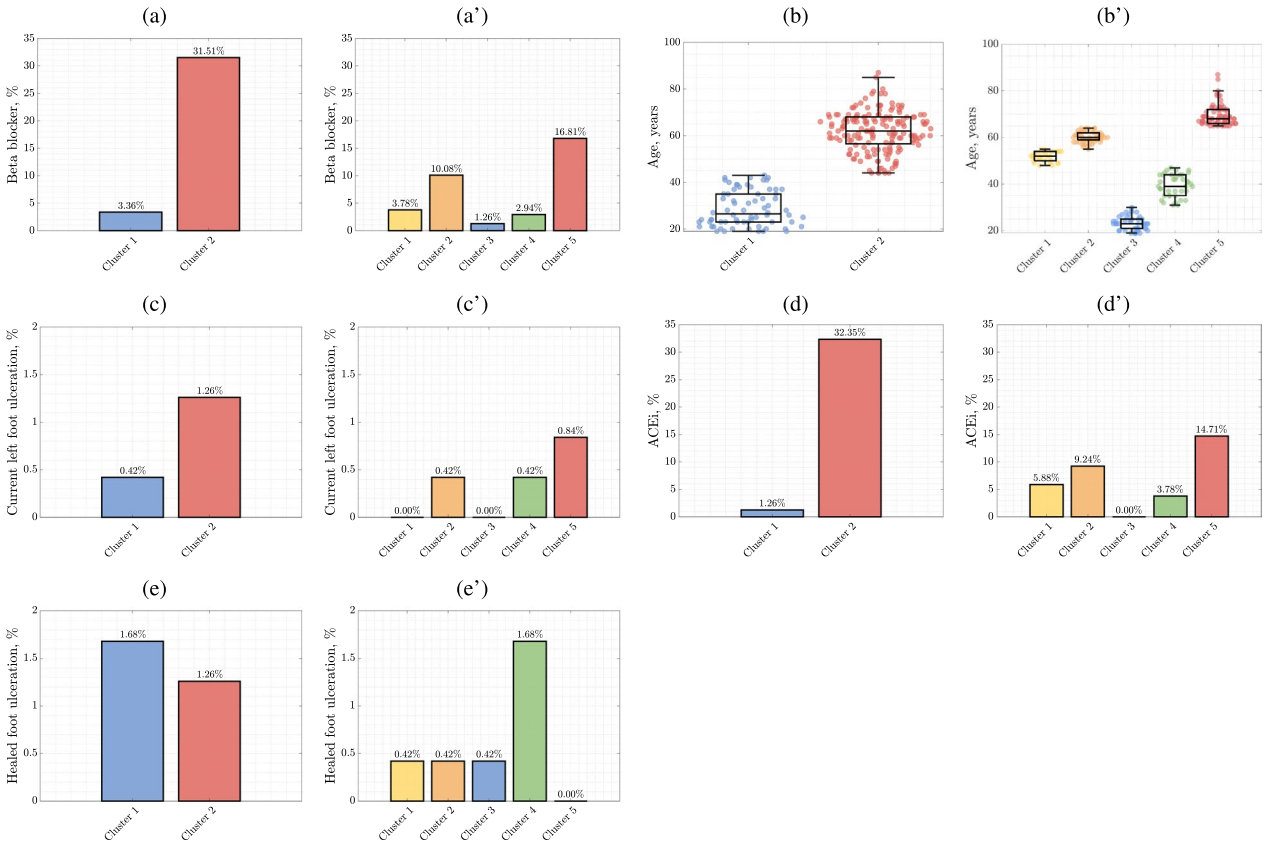

Fig. 4 The most discriminative features for individual clusters in (*i*) the binary (two-cluster) approach **a**–**e** and (*ii*) with the optimal number [5] of clusters **a**–**e**' determined using the Calinski-Harabasz qualitative criterion

Table 3 Patient's parameter values (% yes) for the most discriminative [5] predictors obtained for patients grouped using hierarchical clustering into 2 and 5 clusters

Parameter	Two-group clustering			Five-group clustering					
	Cluster 1 (n = 70)	Cluster 2 (n = 168)	*p*-value	Cluster 1 (n = 28)	Cluster 2 (n = 56)	Cluster 3 (n = 42)	Cluster 4 (n = 39)	Cluster 5 (n = 73)	*p*-value
Beta blocker	8 (11.43%)	75 (44.64%)	< 0.0001	9 (32.15%)	24 (42.86%)	3 (7.14%)	7 (17.95%)	2 (2.74)	< 0.0001
Age	28.79 ± 7.57	61.89 ± 8.70	< 0.0001	51.75 ± 2.15	60.05 ± 2.28	23.26 ± 3.04	39.31 ± 4.80	69.72 ± 4.60	< 0.0001
Current left foot ulceration	1 (1.43%)	3 (1.79%)	0.845	0 (0.00%)	1 (1.79%)	0 (0.00%)	1 (2.56%)	2 (2.74%)	0.759
ACEi	3 (4.28%)	77 (45.83%)	< 0.0001	14 (50.00%)	22 (39.39%)	0 (0.00%)	9 (23.08%)	35 (47.95%)	< 0.0001
Healed foot ulceration	4 (5.71%)	3 (1.78%)	0.102	1 (3.57%)	1 (1.78%)	1 (2.38%)	4 (10.26%)	0 (0.00%)	0.043
CV event	0 (0.00%)	53 (31.54%)	< 0.0001	3 (10.71%)	20 (35.71%)	0 (0.00%)	2 (5.13%)	28 (38.36%)	< 0.0001

The p-values were calculated using Mann–Whitney U-test, χ^2 test, or Kruskal–Wallis tests, where appropriate

patients with CVD, which is its great strength, as none of the high-risk patients remained undiagnosed. On the other hand, in a group of individuals without overt CVD, there were some false positive indications, which might be costly considering health care expenditures.

Taking into consideration the five-group clustering, two groups (i.e., clusters 2 and 5) appeared to be of high CV risk—the oldest patients were treated in large extent with ACEi but not BB and having current foot ulceration (cluster 5 in Fig. 3), and slightly younger individuals treated with both of the above-mentioned

drugs, with a relatively small percentage of diabetic foot syndrome (cluster 2 in Fig. 3). Using both ACEi and BB probably indicates multimorbidity of the examined patients, while the sole use of ACEi might be used due to its nephroprotective effect as its recommendation in the DM management guidelines [50]. On the basis of our results, we can profile patients' CV risk and predict that the highest probability of having CVD is associated with clusters 2 or 5. On the other hand, the youngest patients mostly without the concomitant treatment and low percentage of diabetic foot syndrome (cluster 3 in Fig. 3) might be classified as relatively safe from the CV point of view.

Limitations

We are aware of the limitations of our study. This was a single center study, therefore, its outcomes should be validated over a larger and more heterogeneous population of patients with diabetes to further robustify our findings. Moreover, we were unable to analyze UACR due to high number of missing values (14.29% of all patients did not have the UACR parameter calculated), which was one of the parameters important in cluster analysis of patients from EXCEL and TECOS trials [49]. Utilizing other algorithms for imputing missing values, also for those parameters with relatively large percentage of missing data points, may indeed constitute an interesting research pathway to allow for including such parameters in the predictive ML models [51, 52]. Although we exploited the MLR models for determining high-risk patients, utilizing other data-driven techniques, built upon both classic [53–55] and deep [56] machine learning approaches, may not only help improve the classification performance of the system, but also enhance its robustness against missing and noisy data [57].

Conclusion

Using a ML approach in a prospective cohort of patients with DM, we identified important factors that predicted CV risk. If a patient is treated with ACEi or BB, is older and has/had a foot ulcer, this strongly predicts that she/he is at high risk of having overt CVD.

Abbreviations

ACEi	Angiotensin converting enzyme inhibitor
AUC	Area under the curve
BB	Beta blocker
BMI	Body mass index
CAN	Cardiovascular autonomic neuropathy
CC	Correctly classified
CI	Confidence interval
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
CV	Cardiovascular
DCA	Decision curve analysis
DM	Diabetes mellitus
DiCAN	Diabetic Cardiac Autonomic Neuropathy
DPN	Diabetic peripheral neuropathy
eGFR	Estimated glomerular filtration rate
EXCEL	Exenatide Study of Cardiovascular Event Lowering
HbA1c	Hemoglobin A1c
HR	Heart rate
ML	Machine learning
MLR	Multiple logistic regression
MNSI	Michigan Neuropathy Screening Instrument
RAND	Research and development
ROC	Receiver operating characteristic
SF-36	The 36-Item Shirt Form Survey
TECOS	Trial Evaluating Cardiovascular Outcomes with Sitagliptin
TIA	Transient ischemic attack
T1DM	Type 1 diabetes mellitus
T2DM	Type 2 diabetes mellitus
t-SNE	T-distributed stochastic neighbor embedding
UACR	Urinary albumin to creatin ratio

Acknowledgements

None.

Author contributions

HK, KN, JG, GYHL—conceptualization; HK, KN—funding acquisition, HK, KN, JG—methodology; HK, KN—project administration; HK, KN—resources; HK, KN, JP, JK, MH, SB, KK, MH, KI—investigation; HK, AMW—prepared the data set for statistical analysis; AMW, JN—designed the machine learning algorithms; AMW—implemented and verified the machine learning algorithms; AMW, JN—performed the computational experiments; AMW, JN—performed the data analysis; AMW, JN—prepared tables and figures; HK, KN, JG, GYHL—supervision; HK, KN, AMW, JN—writing—original draft; HK, KN, AMW, JN, JG, GYHL—writing—review and editing.

Funding

This work was supported by the Medical University of Silesia, grant number PCN-1-085/N/2/K. JN and AMW were supported by the Silesian University of Technology funds through the Excellence Initiative—Research University program (Grant 02/080/SDU/10-21-01). AMW was partially supported by the Silesian University of Technology funds through the grant for maintaining and developing research potential.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was performed in accordance with the ethical principles of the Declaration of Helsinki and approved by the Ethics Committee of the Medical University of Silesia (KNW/0022/KB1/10/17). A written informed consent to participate in the study was obtained from all patients enrolled into this study.

Consent for publication

All the authors have read the manuscript, approved its contents and its publication. The manuscript has not been published or submitted for publication elsewhere.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author details

[1]Department of Internal Medicine, Diabetology and Nephrology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland. [2]Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland. [3]Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland. [4]Students' Scientific Association by the Department of Internal Medicine, Diabetology and Nephrology

in Zabrze, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland. [5]Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart & Chest Hospital, Liverpool, UK. [6]Danish Center for Health Services Research, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark.

Received: 25 May 2023 Accepted: 24 July 2023
Published online: 24 August 2023

References

1. IDF Diabetes atlas. 10th edition.
2. Rawshani A, Rawshani A, Franzén S, Eliasson B, Svensson A-M, Miftaraj M, et al. Mortality and cardiovascular disease in type 1 and type 2 diabetes. N Engl J Med. 2017;376(15):1407–18.
3. Sarwar N, Gao P, Kondapally Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet (London, England). 2010;375(9733):2215–22.
4. Rawshani A, Sattar N, Franzén S, Rawshani A, Hattersley AT, Svensson A-M, et al. Excess mortality and cardiovascular disease in young adults with type 1 diabetes in relation to age at onset: a nationwide, register-based cohort study. Lancet (London, England). 2018;392(10146):477–86.
5. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987–1003.
6. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357: j2099.
7. Echouffo-Tcheugui JB, Kengne AP. Comparative performance of diabetes-specific and general population-based cardiovascular risk assessment models in people with diabetes mellitus. Diabetes Metab. 2013;39(5):389–96.
8. Kengne AP, Patel A, Colagiuri S, Heller S, Hamet P, Marre M, et al. The Framingham and UK Prospective Diabetes Study (UKPDS) risk equations do not reliably estimate the probability of cardiovascular events in a large ethnically diverse sample of patients with diabetes: the Action in Diabetes and Vascular Disease: Pretera. Diabetologia. 2010;53(5):821–31.
9. Chamnan P, Simmons RK, Sharp SJ, Griffin SJ, Wareham NJ. Cardiovascular risk assessment scores for people with diabetes: a systematic review. Diabetologia. 2009;52(10):2001–14.
10. Sofogianni A, Stalikas N, Antza C, Tziomalos K. Cardiovascular risk prediction models and scores in the era of personalized medicine. J Pers Med. 2022;12(7):1180.
11. Mora D, Nieto JA, Mateo J, Bikdeli B, Barco S, Trujillo-Santos J, et al. Machine learning to predict outcomes in patients with acute pulmonary embolism who prematurely discontinued anticoagulant therapy. Thromb Haemost. 2022;122(4):570–7.
12. Lip GYH, Genaidy A, Tran G, Marroquin P, Estes C, Sloop S. Improving stroke risk prediction in the general population: a comparative assessment of common clinical rules, a new multimorbid index, and machine-learning-based algorithms. Thromb Haemost. 2022;122(1):142–50.
13. Nopp S, Spielvogel CP, Schmaldienst S, Klauser-Braun R, Lorenz M, Bauer BN, et al. Bleeding risk assessment in end-stage kidney disease: validation of existing risk scores and evaluation of a machine learning-based approach. Thromb Haemost. 2022;122(9):1558.
14. Jiang Y, Yang ZG, Wang J, Shi R, Han PL, Qian WL, et al. Unsupervised machine learning based on clinical factors for the detection of coronary artery atherosclerosis in type 2 diabetes mellitus. Cardiovasc Diabetol. 2022;21(1):1–10.
15. Drożdż K, Nabrdalik K, Kwiendacz H, Hendel M, Olejarz A, Tomasik A, et al. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. Cardiovasc Diabetol. 2022;21(1):240.
16. Nabrdalik K, Kwiendacz H, Drożdż K, Irlik K, Hendel M, Wijata AM, et al. Machine learning predicts cardiovascular events in patients with diabetes: the Silesia diabetes-heart project. Curr Probl Cardiol. 2023;48:101694.
17. Chowdhury M, Nevitt S, Eleftheriadou A, Kanagala P, Esa H, Cuthbertson DJ, et al. Cardiac autonomic neuropathy and risk of cardiovascular disease and mortality in type 1 and type 2 diabetes: a meta-analysis. BMJ Open Diab Res Care. 2021;9:2480.
18. Le Dinh T, Phi Thi Nguyen N, Thanh Thi Tran H, Luong Cong T, Ho Thi Nguyen L, Do Nhu B, et al. Diabetic peripheral neuropathy associated with cardiovascular risk factors and glucagon-like peptide-1 concentrations among newly diagnosed patients with type 2 diabetes mellitus. Diabetes Metab Syndr Obes. 2022. https://doi.org/10.2147/DMSO.S344532.
19. Ashley EA. The precision medicine initiative: a new national effort. JAMA. 2015;313(21):2119–20.
20. Guo Y. A new paradigm of "real-time" stroke risk prediction and integrated care management in the digital health era: innovations using machine learning and artificial intelligence approaches. Thromb Haemost. 2022;122:5–7.
21. Kleindorfer DO, Towfighi A, Chaturvedi S, Cockroft KM, Gutierrez J, Lombardi-Hill D, et al. 2021 guideline for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline from the American Heart Association/American Stroke Association. Stroke. 2021;52(7):e364-467.
22. Little RR. Glycated hemoglobin standardization–National glycohemoglobin standardization program (NGSP) perspective. Clin Chem Lab Med. 2003;41(9):1191–8.
23. Moore JF, Sharer JD. Methods for quantitative creatinine determination. Curr Protoc Hum Genet. 2017;93:A-3O.
24. Pugliese G, Solini A, Bonora E, Orsi E, Zerbini G, Giorgino F, et al. The chronic kidney disease epidemiology collaboration (CKD-EPI) equation provides a better definition of cardiovascular burden associated with CKD than the modification of diet in renal disease (MDRD) study formula in subjects with type 2 diabetes. Atherosclerosis. 2011;218(1):194–9.
25. Sampson M, Wolska A, Cole J, Zubirán R, Otvos JD, Meeusen JW, et al. Accuracy and clinical impact of estimating low-density lipoprotein-cholesterol at high and low levels by different equations. Biomedicines. 2022;10(12):3156.
26. Committee ADAPP. 11 Chronic kidney disease and risk management: standards of medical care in diabetes-2022. Diabetes Care. 2022;45(1):S175-84.
27. Feldman EL, Stevens MJ, Thomas PK, Brown MB, Canal N, Greene DA. A practical two-step quantitative clinical and electrophysiological assessment for the diagnosis and staging of diabetic neuropathy. Diabetes Care. 1994;17(11):1281–9.
28. Edmonds M, Manu C, Vas P. The current burden of diabetic foot disease. J Clin Orthop trauma. 2021;17:88–93.
29. Armstrong DG, Boulton AJM, Bus SA. Diabetic foot ulcers and their recurrence. N Engl J Med. 2017;376(24):2367–75.
30. Vinik AI, Maser RE, Mitchell BD, Freeman R. Diabetic autonomic neuropathy. Diabetes Care. 2003;26(5):1553–79.
31. Ewing DJ, Campbell IW, Clarke BF. Assessment of cardiovascular effects in diabetic autonomic neuropathy and prognostic implications. Ann Intern Med. 1980;92(2II):308–11.
32. 36-Item short form survey (SF-36) Scoring instructions. RAND.
33. Audigier V, Husson F, Josse J. A principal components method to impute missing values for mixed data. Adv Data Anal Classif. 2016;10(1):5–26.
34. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. Comput Math Methods Med. 2017. https://doi.org/10.1155/2017/3762651.
35. Fernández A, Gómez S. Versatile linkage: a family of space-conserving strategies for agglomerative hierarchical clustering. J Classif. 2020;37(3):584–97.
36. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
37. Oliveira FHM, Machado ARP, Andrade AO. On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with Parkinson's disease. Comput Math Methods Med. 2018;2018:8019232.
38. Sethi Y, Patel N, Kaka N, Kaiwan O, Kar J, Moinuddin A, et al. Precision medicine and the future of cardiovascular diseases: a clinically oriented comprehensive review. J Clin Med. 2023;12(5):1799.

39. Tromp J, Paniagua SMA, Lau ES, Allen NB, Blaha MJ, Gansevoort RT, et al. Age dependent associations of risk factors with heart failure: pooled population based cohort study. BMJ. 2021;372: n461.
40. Boyko EJ, Ahroni JH, Smith DG, Davignon D. Increased mortality associated with diabetic foot ulcer. Diabet Med. 1996;13(11):967–72.
41. Chammas NK, Hill RLR, Edmonds ME. Increased mortality in diabetic foot ulcer patients: the significance of ulcer type. J Diabetes Res. 2016;2016:2879809.
42. Morbach S, Furchert H, Gröblinghoff U, Hoffmeier H, Kersten K, Klauke G-T, et al. Long-term prognosis of diabetic foot patients and their limbs: amputation and death over the course of a decade. Diabetes Care. 2012;35(10):2021–7.
43. Dietrich I, Braga GA, de Melo FG, da Costa AC. The diabetic foot as a proxy for cardiovascular events and mortality review. Curr Atheroscler Rep. 2017;19(11):44.
44. Meloni M, Bellia A, Giurato L, Lauro D, Uccioli L. Below-the-ankle arterial disease: a new marker of coronary artery disease in patients with diabetes and foot ulcers. Acta Diabetol. 2022;59(10):1331–8.
45. Balasubramanian GV, Chockalingam N, Naemi R. The role of cutaneous microcirculatory responses in tissue injury, inflammation and repair at the foot in diabetes. Front Bioeng Biotechnol. 2021;9: 732753.
46. Jensen J, Poulsen MK, Petersen PW, Gerdes B, Rossing K, Schou M. Prevalence of heart failure phenotypes and current use of therapies in primary care: results from a nationwide study. ESC Hear Fail. 2023. https://doi.org/10.1002/ehf2.14324.
47. Green JB, Bethel MA, Armstrong PW, Buse JB, Engel SS, Garg J, et al. Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes. N Engl J Med. 2015;373(3):232–42.
48. Stone GW, Kappetein AP, Sabik JF, Pocock SJ, Morice M-C, Puskas J, et al. Five-year outcomes after PCI or CABG for left main coronary disease. N Engl J Med. 2019;381(19):1820–30.
49. Sharma A, Zheng Y, Ezekowitz JA, Westerhout CM, Udell JA, Goodman SG, et al. Cluster analysis of cardiovascular phenotypes in patients with type 2 diabetes and established atherosclerotic cardiovascular disease: a potential approach to precision medicine. Diabetes Care. 2021;45(1):204–12. https://doi.org/10.2337/dc20-2806.
50. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 11. Chronic kidney disease and risk management: standards of care in diabetes-2023. Diabetes Care. 2023;46(1):S191-202.
51. Goretzko D, Heumann C, Bühner M. Investigating parallel analysis in the context of missing data: a simulation study comparing six missing data methods. Educ Psychol Meas. 2020;80(4):756–74.
52. Slade E, Naylor MG. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. Stat Med. 2020;39(8):1156–66.
53. Anand V, Downs SM. Probabilistic asthma case finding: a noisy or reformulation. AMIA Annu Symp Proc. 2008;2008:6–10.
54. Kotowski K, Kucharski D, Machura B, Adamski S, Gutierrez Becker B, Krason A, et al. Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features. Comput Biol Med. 2023;152: 106378.
55. Salman I, Vomlel J. Learning the structure of Bayesian networks from incomplete data using a mixture model. Informatica. 2023. https://doi.org/10.31449/inf.v47i1.4497.
56. Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-Keridis LA, Upadhyay TK, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front Med. 2023;10:1150933.
57. Dubel R, Wijata AM, Nalepa J. On the impact of noisy labels on supervised classification models. In: Mikyška J, de Mulatier C, Paszynski M, Krzhizhanovskaya VV, Dongarra JJ, Sloot PMA, editors. BT, computational science, ICCS 2023. Cham: Springer; 2023. p. 111–9.

Publisher's Note