International Journal of Diabetes in Developing Countries (July–September 2024) 44(3):528–538
https://doi.org/10.1007/s13410-023-01264-7

ORIGINAL ARTICLE

Multivariable prediction model of complications derived from diabetes mellitus using machine learning on scarce highly unbalanced data

Claudia C. Colmenares-Mejía[1] · Juan C. Rincón-Acuña[2,3] · Andrés Cely[1,4] · Abel E. González-Vélez[5] · Andrea Castillo[6] · Jossie Murcia[7] · Mario A. Isaza-Ruget[8]

Received: 22 November 2022 / Accepted: 27 September 2023 / Published online: 26 October 2023

Abstract

Background Diabetes mellitus (DM) increases the risk complications in addition to mortality. Quantifying the risk of complications using artificial intelligence could be a way to design comprehensive patient healthcare programs.

Objective Predicting the probability of macro and microvascular complications in patients with DM through Machine Learning

Methods Retrospective cohort study. Based on an outpatient follow-up program for diabetic patients, 64,081 records and 287 variables were identified, with highly unbalanced data. Predictive models for chronic kidney disease (CKD), lower extremity amputation (LEA), coronary heart disease (CHD), and early mortality (MOR) were developed. An exhaustive computational method was conducted to find the best combination between machine learning (ML) algorithms and sampling method.

Results The best model was determined by assessing its performance through the heuristics obtained from a comprehensive analysis of the accuracy and F1 values for ML, sampling, and dataset. Regarding each complication, 99.9% accuracy was obtained for LEA, 94.3% for CHD, 97.4% for MOR, and 98.8% for CKD. F1 was assessed to identify false positives, with 84.5% for CKD, 63.6% for MOR, 46.2% for LEA, and 44.8% for CHD.

Conclusions This ML model can be applied to predict CHD, CKD, and MOR. The success of ML predictions lies in the clinical definition of initial variables and their simplification for obtaining variables based on which the algorithms can identify patients that are likely to develop a complication. For clinical application of this system, it is necessary to assess the cross performance of metrics, as found here (accuracy higher 95% and F1-Score higher than 80%).

Keywords Complications · Diabetes mellitus · Machine learning · Predictive analytics · Risk predictions

Introduction

Diabetes mellitus (DM) is a carbohydrate metabolism disorder with high medical costs owing to its associated therapeutic implications, labor disability, acute and chronic complications, and early mortality. Poorly controlled DM increases the incidence of long-term complications, such as retinopathy, kidney disease, peripheral neuropathy, coronary heart disease (CHD), and peripheral vascular disease, all of which have a negative impact on patients, their families, and the society [1]. Around 12% of medical expenses worldwide are associated with treatment of patients with DM and associated complications [2]. Additionally, it has been reported that 366 million people will have DM by 2030, making complication predictions of this disease essential for preventing its health consequences [3, 4].

✉ Claudia C. Colmenares-Mejía
cccolmenaresm@unisanitas.edu.co

[1] Fundación Universitaria Sanitas, Bogotá D.C., Colombia

[2] Univerisdad de Santander, Campus Lagos del Cacique, Bucaramanga, Santander, Colombia

[3] Data Scientist. Corporate Data Management, Keralty, Bogotá D.C., Colombia

[4] Universidad Nacional de Colombia, Bogotá D.C., Colombia

[5] Preventive Medicine Service, University Hospital of Torrejón, Torrejón de Ardoz, Spain

[6] Dirección Gestión del Conocimiento, EPS Sanitas, Bogotá D.C, Colombia

[7] Instituto de Gerencia y Gestión Sanitaria, Fundación Universitaria Sanitas, Bogotá D.C., Colombia

[8] Research group INPAC, Fundación Universitaria Sanitas, Bogotá D.C., Colombia

Statistical models have traditionally been developed for estimating the probability of macro (CHD and peripheral vascular disease) and microvascular complications (retinopathy, kidney disease, and peripheral neuropathy) mainly in populations other than Hispanics living in high-income countries. Thus, in a Japanese population, 5-year predictive models showed high performance for CHD, mortality, and kidney disease (C-statistic of 0.725, 0.696, and 0.767, respectively), but a moderate performance for stroke and retinopathy evolution (C-statistic of 0.636 and 0.614, respectively) [5, 6]. These predictive models have not been validated for Latin American populations nor developed based on real-world data.

The application of machine learning (ML) and data mining to DM research is an imperative and innovative approach, as it enables optimal analysis of large volumes of available data, especially data collected during standard clinical care on patients (diagnosis, tests, samples, biomedical data, etc.) for gathering knowledge and support clinical treatment decision aimed at decreasing the possibility of future complications [7–9]. The MOSAIC project used this strategy to input data (random forest — RF) and balance working bases support vector machine (SVM) for obtaining a 3-, 5-, and 7-year prediction of retinopathy, kidney disease, and peripheral neuropathy disease development through logistic regression [10].

This article aims to generate knowledge based on biological data of patients with DM for predicting the probability of macro and microvascular complications through ML, a complementary tool capable of providing physicians with objective and timely information so that they can treat patients with DM according to their necessities and, eventually, improve the quality of life of patients, their families, and the society.

Materials and methods

Study Design

From an epidemiological perspective, this study employed a retrospective cohort design using data collected from patients enrolled in the "Chronic Diseases Healthcare Program" managed by a Colombian private health insurance company over a 5-year period between 2013 and 2018. All patients who have confirmed DM diagnosis and were admitted to the program in 2013 were included in this study and were followed up to 2018 in order to identify outcomes of interest.

Data sources

Data sources included were Electronic Health Records (Diabetes Registry and Ambulatory Consultation records), Business Intelligence–BI (Drugs and Procedures) systems, and high-cost disease (known as *Cuenta de Alto Costo*) registry issued to national health ministry annually.

Outcomes

Outcomes of interest were chronic complications in patients with DM, specifically:

- Lower extremity amputation (LEA) was defined as any surgical proceedings performed to amputate an extremity in patients with a history of peripheral artery disease, nerve disease, or diabetic foot ulcer
- Chronic kidney disease (CKD) was defined as the presence of albuminuria, low glomerular filtration, or other signs of kidney damage according to the KDIGO guidelines [11].
- Coronary heart disease (CHD), defined as a history of CHD as per clinical records (ICD-10) or surgical proceedings (coronary bypass or stent placement due to coronary artery blockage)
- Mortality (MOR), defined as any cause specified in the death certificates

Predictors

A literature review was performed in order to identify the ideal set of variables (predictors) for each model. In general, predictors considered were data related to sociodemographic (sex, age, date of admission to the program, and city of residence), clinical (tobacco use, medical and surgical history, physical examination data, and prescription and/or use of antihypertensive drugs, NSAIDs, oral hypoglycemic drugs, and/or insulin), and laboratory (levels of glycemia, creatinine, albumin, albuminuria, creatinuria, lipids, hemoglobin, glycosylated hemoglobin (HbA1c), glomerular filtration rate (GFR), parathormone, and phosphorus) information. Minimal and ideal set of predictors for each model (LEA, CKD, CHD, MOR) is presented in Supplementary Information (Table S1). These predictors were derived from the literature review and were deemed most appropriate for each specific model.

Sample size

All patients from a private healthcare insurer network that met the eligibility criteria during 2013 and 2018 and had updated information were included.

Machine learning

For ML analysis, an incremental iterative method was applied, which was adapted based on the guidelines for predictive data mining in clinical medicine, derived from

CRISP Data Mining Methodology Extension for Medical Domain [12], and was structured into four steps: data preparation, data preprocessing, training, and validation (Supplementary information, Fig. S1).

Data preparation

Original records were organized in a semi-structured manner (Supplementary information, Table S2), and there was a wide heterogeneity among patients. For each patient, variables of interest were reported annually, ranging from unreported (no data) to up to 20 annual reports per patient, with an irregular time pattern. Included variables were selected based on scientific evidence and/or medical literature showing a correlation between the reports and the predicted complication.

Condensation of repeated-measure variables was used to consolidate multiple data collection procedures, such as blood pressure or laboratory data (Cholesterol or HbA1c), into one. For doing this, two summarization methods were applied: absolute variation of repeated measures (AVRM) and coefficient of variation (CV) calculation. Data imputation was developed by estimating the average value in a group of patients using the K-Means (Supplementary information, Fig. S2) technique and dividing the data into *n* clusters of equal variances. The value to be imputed was determined as the average of a subgroup, minimizing inertia or the sum of squares within the cluster. The Elbow method was used to determine the number of clusters to be used (Supplementary information, Fig. S3).

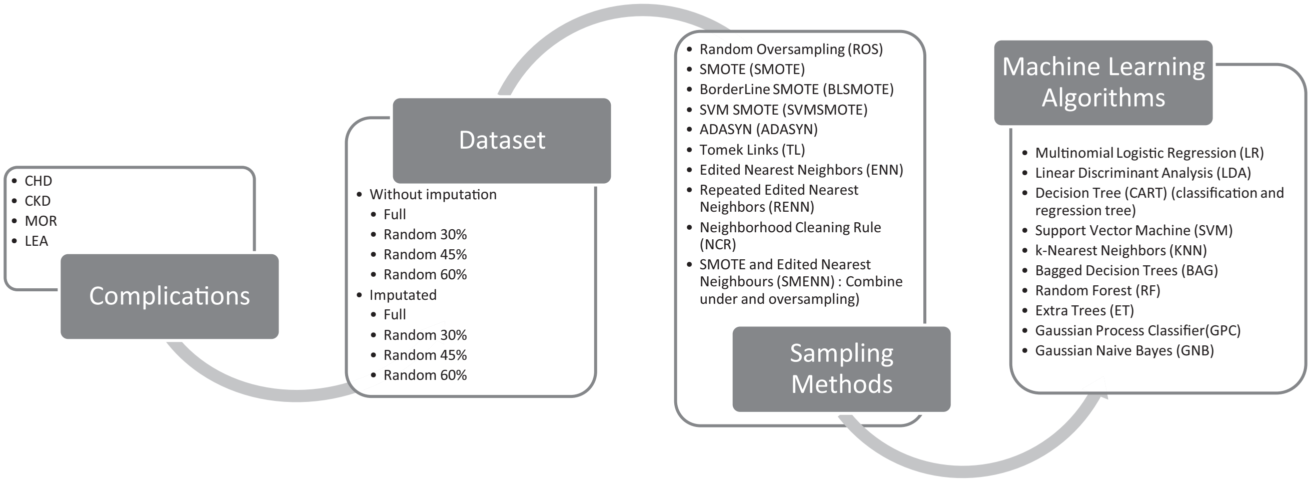
Pre-processing and predictor selection

For addressing this highly unbalanced data, different imputation sampling techniques were applied (over and undersampling) and complemented by a random selection of patients without complications to generate three additional datasets called "random 30–45–60%" from which 30%, 45%, or 60% of patients were randomly removed and mixed with patients that did develop a complication (Fig. 1).

After the values were summarized and imputed, variables with the highest influence over each type of complication were determined. Variables that did not contribute to the predictive process were eliminated through correlation analysis (Supplementary information, Fig. S5) Sequential forward floating selection technique (SFFS) in conjunction with a cross-validation scheme validated ML performance, based on K-Fold. Feature extraction is based on sequential engineering based on different techniques such as univariate analysis using chi-square, tree-like classifiers, and progressive sequencing with selection techniques.

Training

Overall, as shown in Fig. 1, training was performed for each complication, with eight datasets (ALL, ALL imputed, and sampling 30%, 45%, and 60% imputed or not imputed), ten sampling methods: random oversampling (ROS); SMOTE; BorderLine SMOTE; BLSMOTE); SVM SMOTE (SVMSMOTE); ADASYN (ADASYN); Tomek Links (TL); edited nearest neighbors (ENN); repeated edited nearest neighbors (RENN); neighborhood cleaning rule (NCR); SMOTE and edited nearest neighbors (SMENN): combine under and oversampling), and ten ML algorithms: multinomial logistic regression (LR); linear discriminant analysis (LDA); decision tree (CART) (classification and regression tree); support vector machine (SVM); k-nearest neighbors (KNN); Bagged Decision Trees (BAG); random forest (RF); extra trees (ET); Gaussian process classifier (GPC); Gaussian naive Bayes (GNB), which resulted in a total of 3200



- CHD
- CKD
- MOR
- LEA

Complications

Dataset

- Without imputation
  - Full
  - Random 30%
  - Random 45%
  - Random 60%
- Imputated
  - Full
  - Random 30%
  - Random 45%
  - Random 60%

- Random Oversampling (ROS)
- SMOTE (SMOTE)
- BorderLine SMOTE (BLSMOTE)
- SVM SMOTE (SVMSMOTE)
- ADASYN (ADASYN)
- Tomek Links (TL)
- Edited Nearest Neighbors (ENN)
- Repeated Edited Nearest Neighbors (RENN)
- Neighborhood Cleaning Rule (NCR)
- SMOTE and Edited Nearest Neighbours (SMENN) : Combine under and oversampling)

Sampling Methods

Machine Learning Algorithms

- Multinomial Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- Decision Tree (CART) (classification and regression tree)
- Support Vector Machine (SVM)
- k-Nearest Neighbors (KNN)
- Bagged Decision Trees (BAG)
- Random Forest (RF)
- Extra Trees (ET)
- Gaussian Process Classifier(GPC)
- Gaussian Naive Bayes (GNB)

Fig. 1 General process followed to obtain relevant ML for predicting CHD, CKD, MOR, and LEA

trainings to be conducted with *each initial set of variables and patients*. Using this approach, it is possible to comprehensively review the behavior of a group of variables and avoid the elimination of sampling methods (over and undersampling), certain ML algorithms, or bias induced by statistical engineering of characteristics [13]. Blind assessment was omitted due to the novelty method to perform mathematical evaluation of the ML models outcome. This is a key factor to include in a long-term study, where an individual and detailed clinical follow-up will be performed, to evaluate if the ML model predicts a real case or false positive.

Validation

For addressing highly unbalanced data, over and undersampling were applied and supplemented by a random selection of patients with no complication. Moreover, to validate the performance of each model and its capacity for responding to new data or a generalization, the sample of each possible dataset was divided into 80% for training and 20% for testing (Fig. 2). The generalizing capacity of the models was validated using an error matrix.

An exhaustive, combinatorial algorithmic processing was performed to determine the combination (ML-sampling-biomarkers) with the best performance. Mass training was performed with different combinations of variables and processing techniques per patient group instead of choosing a given algorithm.

Accuracy was used to determine the models' general performance, which was supplemented by a recall analysis for identifying cases in which the algorithm successfully predicts a complication and a precision analysis for determining cases in which the algorithm does not predict a complication and the complication does occur. The aim was to find the best performance for predicting complications, especially for avoiding false negatives by assessing the F1-Score value as a harmonic mean of precision (specificity) and recall (sensitivity) values. A heuristic validation (HV) was chosen based on accuracy heuristics and F1 heuristics, which were obtained by a sigmoid function applied to each metric as a correction factor over the performance of three datasets (training, test, and cross-validation).

Ethics considerations

This study was reviewed and approvals obtained to use the data for analysis by the Ethics Committee of the Fundación Universitaria Sanitas (CEIFUS 320–18). The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Results

Participants and predictors

The initial dataset included 64,081 patients and 287 variables, which was restricted to patients with at least a year of follow-up between 2013 and 2018, resulting in a total of 28,828 patients (Supplementary information, Fig. S2). The 287 variables were reduced to 21 by applying a systematic grouping process over time defined in the preprocessing stage (Table 1; Fig. 3). The final set of variables included sex, age, socioeconomic status, body mass index (BMI), mean blood pressure, glycemia, HbA1c, total cholesterol, HDL cholesterol, LDL cholesterol, hemoglobin, GFR, CKD stage (KDIGO classification), kidney transplant history, hypertension, ACE inhibitor or ARB-2 prescription, insulin

Fig. 2 Cross-validation scheme for validating ML performance, based on K-fold

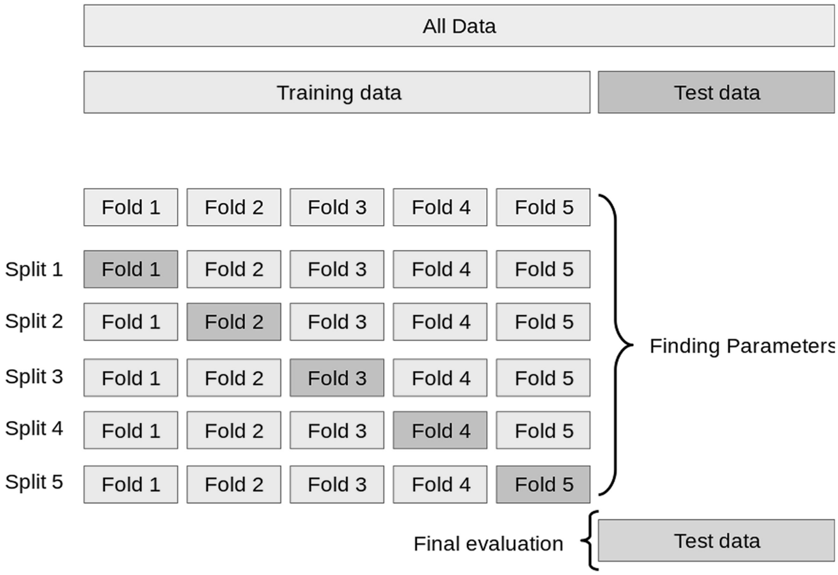


Table 1 Features names selection for each outcome

Outcome	Summarize shape of dataset	Target prediction VaR distribution	Feature names or variables per VaR amount

Heart Disease/CHD	101,960, 25	Class = 0, count = 61,333, percentage = 60.154%
		Class = 1, count = 40,627, percentage = 39.846%
			13: gender, age, SBP, Hba1c, cholesterol HDL, hemoglobin, GFR, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin
			12: gender, age, SBP, Hba1c, cholesterol HDL, hemoglobin, GFR, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin
			11: gender, age, SBP, Hba1c, cholesterol HDL, hemoglobin, GFR, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin
			15: gender, age, SBP, DBP, total cholesterol, cholesterol HDL, hemoglobin, GFR, history of kidney transplantation, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin
			16: gender, age, SBP, DBP, Hba1c, total cholesterol, cholesterol HDL, hemoglobin, GFR, history of kidney transplantation, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin

Chronic kidney disease	64,081, 25	Class = 0, count = 63,029, percentage = 98.358%
		Class = 1, count = 1052, percentage = 1.642%
		Distribution of class labels BEFORE resampling counter ({0: 63,029, 1: 1052})
		Distribution of class labels AFTER resampling counter ({0: 63,029, 1: 63,029})
			15: gender, age, SBP, DBP, total cholesterol, cholesterol HDL, hemoglobin, GFR, history of kidney transplantation, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin
			16: gender, age, SBP, DBP, Hba1c, total cholesterol, cholesterol HDL, hemoglobin, GFR, history of kidney transplantation, IECA, ARB-2, socioeconomic status, hypertension, insulin, metformin
			14: gender, age, SBP, DBP, total cholesterol, cholesterol HDL, hemoglobin, GFR, history of kidney transplantation, IECA, ARB-2, socioeconomic status, hypertension, insulin
			13: gender, age, SBP, DBP, total cholesterol, hemoglobin, GFR, history of kidney transplantation, IECA, ARB-2, socioeconomic status, hypertension, insulin

Lower extremity amputation	64,081, 25	Class = 0, count = 62,780, percentage = 97.970%
		Class = 1, count = 1301, percentage = 2.030%
		Distribution of class labels BEFORE resampling counter ({0: 62,780, 1: 1301})
		Distribution of class labels AFTER resampling counter ({0: 62,780, 1: 62,780})
			14: age, kidney disease progression, glycemia, creatinine, HbA1C, total cholesterol, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, insulin, metformin
			13: age, kidney disease progression, glycemia, creatinine, HbA1C, total cholesterol, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, insulin, metformin
			12: age, kidney disease progression, creatinine, HbA1c, total cholesterol, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, Insulin, metformin
			11: age, kidney disease progression, creatinine, total cholesterol, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, insulin, metformin
			10: age, kidney disease progression, creatinine, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, Insulin, metformin

Table 1 (continued)

Outcome	Summarize shape of dataset	Target prediction VaR distribution	Feature names or variables per VaR amount

Mortality	64,081, 25	Class = 0, count = 63,975, percentage = 99.835%
		Class = 1, count = 106, percentage = 0.165%
		Distribution of class labels BEFORE resampling counter ({0: 63,975, 1: 106})
		Distribution of class labels AFTER resampling counter ({0: 63,975, 1: 63,975})
			23: gender, age, BMI, kidney disease progression, SBP, DBP, glycemia, creatinine, Hba1c, total cholesterol, cholesterol HDL, hemoglobin, GFR, history of kidney transplantation, hypertension, IECA, ARB-2, CKD, insulin, metformin, DPP4 inhibitors
			20: gender, age, BMI, kidney disease progression, SBP, DBP, glycemia, creatinine, Hba1c, total cholesterol, cholesterol HDL, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, CKD, insulin, metformin
			21: gender, age, BMI, kidney disease progression, SBP, DBP, glycemia, creatinine, Hba1c, total cholesteroL, cholesterol HDL, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, CKD, socioeconomic status, insulin, metformin
			22: gender, age, BMI, kidney disease progression, SBP, DBP, glycemia, creatinine, Hba1c, total cholesterol, cholesterol HDL, cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, ARB-2, CKD, socioeconomic status, insulin, metformin
			24: gender, age, BMI, Kidney disease progression, SBP, DBP, glycemia, creatinine, Hba1c, Total Cholesterol, cholesterol HDL, Cholesterol LDL, hemoglobin, GFR, history of kidney transplantation, hypertension, IECA, ARB-2, CKD, socioeconomic status, insulin, Metformin, DPP4 inhibitors

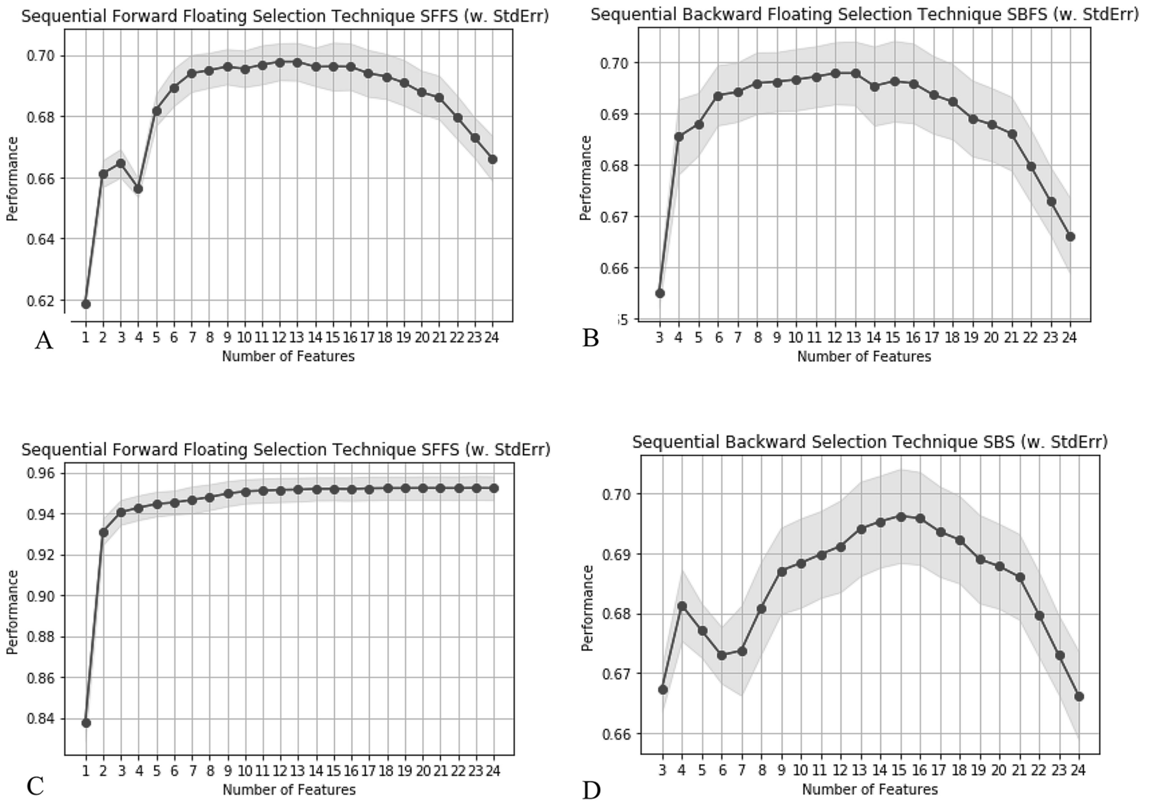


Fig. 3 Sequential forward floating selection technique (SFFS) per complication. **A** Heart Disease/ CHD; **B** Chronic Kidney Disease – CKD; **C** Lower Extremity Amputation – LEA; **D** Mortality - MOR

administration, metformin, and DPP4 inhibitor prescription (Table 2).

Model performance

The performance of dataset with and without imputation is presented for each complication in Supplementary information, Table S3. According to accuracy, there is no significant variation between the datasets (ALL — complete dataset) with missing records and the imputed ones (ACC > 94%). For Precision over class (PRE) measure, the imputed dataset (ALL — complete dataset) presents on average better performance for all outcomes (98.7% vs. 87.1%). If the average of all metrics is taken as a whole, the imputed dataset performs better (69.3% vs. 65.1%). This way of looking at the data is a starting point, but not decisive, since what really matter is to avoid false negatives and false positives.

Once the feasibility of using the selected datasets has been determined, it is necessary to evaluate which ML technique performs best, given the sampling method to be applied. The F1-Score metric is evaluated first, then precision and recall.

For CHD (Supplementary information, Table S4), the best result was obtained with the imputed dataset R30%_IMP, the BAG ML, and ROS sampling; HV was 58.3%, with a weighted accuracy of 77.5% and F1 of 39%. For LEA (Supplementary information, Table S5), the best result was obtained with the non-imputed dataset R30%, the LDA ML, and SVMSMOTE sampling or with no sampling (NONE); HV was 53.5%, with a weighted accuracy of 81.2% and F1 of 25.8%. For CKD (Supplementary information, Table S6), both RF and BAG algorithms yielded the same result; the first one was obtained with BorderLine SMOTE (BLSMOTE) or SVMSMOTE sampling and the second one with ROS sampling, with an imputed dataset R30%_IMP. HV was 71.8%, with a weighted accuracy of 80.8% and F1 of 62.7%. For MOR, RF was observed with ROS (67.2% HV) and SVMSMOTE (66.8% HV) samplings, with an imputed dataset R30%_IMP (Table 3).

The following maximum performance values were obtained per complication with cross heuristic evaluation: 77.5% for CHD, 81.2% for LEA, 80.9% for CKD, and 79.7% for MOR (Fig. 4). In general, accuracy was approximately 80%. The F1-Score Heuristic Validation (F1_ Heu) metric varied significantly per complication, CKD showing the highest performance, followed by MOR. For CHD and LEA, the results obtained were low for a system that is under production.

Table 2 Patients' description according to presence of complications

Variable	All	No complications	LEA	CHD	CKD	MOR
	N = 28.828	*n* = 26.077	*n* = 44	*n* = 1520	*n* = 641	*n* = 773
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Sex						
Male	12,733 (44.2)	11,144 (42.7)	28 (63.6)	918 (60.4)	415 (64.7)	381 (49.3)
Female	16,095 (55.8)	14,933 (57.3)	16 (36.4)	602 (39.6)	226 (35.3)	392 (50.7)
Age[a] (years)	67 (59–74)	66 (58–74)	67 (58–73)	68 (62–73.5)	64 (55–71)	77 (70–83)
Socioeconomic status						
0	2580 (8.9)	2312 (8.9)	5 (11.4)	148 (9.7)	81 (12.6)	64 (8.3)
1	17,797 (61.7)	16,049 (61.5)	23 (52.3)	1.034 (68.0)	361 (56.3)	466 (60.3)
2	1153 (4.0)	1027 (3.9)	3 (6.8)	78 (5.1)	30 (4.7)	29 (3.7)
3	3004 (10.4)	2755 (10.6)	5 (11.4)	103 (6.8)	64 (9.9)	89 (11.5)
4	604 (2.1)	541 (2.1)	2 (4.6)	33 (2.2)	19 (2.9)	16 (2.1)
5	3690 (12.8)	3393 (13.0)	6 (13.6)	124 (8.2)	86 (13.5)	109 (14.1)
History of hypertension	26,006 (90.2)	23,495 (90.1)	38 (86.4)	1426 (93.8)	528 (82.4)	722 (93.4)
ACE inhibitor consumption	7109 (24.7)	6363 (24.4)	15 (34.1)	454 (29.9)	93 (14.5)	229 (29.6)
ARB consumption	14,572 (50.5)	13,038 (50.0)	23 (52.3)	896 (58.9)	328 (51.2)	419 (54.2)
Metformin consumption	22,908 (79.5)	21,064 (80.8)	28 (63.6)	1216 (80.0)	193 (30.1)	490 (63.4)
DPPIV inhibitor consumption	9758 (66.2)	8803 (33.8)	12 (27.3)	586 (38.6)	210 (32.8)	231 (29.8)
Insulin use	9032 (31.3)	7770 (29.8)	33 (75.0)	637 (41.9)	403 (62.9)	350 (45.2)
History of CKD	9655 (33.5)	8257 (31.7)	19 (43.2)	636 (41.8)	-	403 (52.1)
BMI[a]	26.7 (24.1–29.8)	26.8 (24.2–29.9)	24.6 (22.6–26.8)	27.0 (24.5–29.9)	25.3 (22.9–28.4)	24.8 (22.2–27.6)
MBP[a]	91.3 (88.6–94.1)	91.3 (88.7–94.2)	89.3 (87.0–94.2)	90.7 (88.1–93.8)	92.5 (88.9–96.5)	89.4 (86.5–92.9)
Glycemia[a]	121.2 (106.5–144.9)	121.1 (106.5–144.5)	131.1 (104.3–173.7)	123.2 (107.6–147.8)	122.5 (101.0–158.2)	120.3 (105.1–147.0)
HbA1c[a]	6.7 (6.2–7.5)	6.7 (6.2–7.4)	7.0 (6.1–8.3)	6.7 (6.2–7.6)	6.8 (6.0–7.7)	6.8 (6.2–7.6)
Total cholesterol[a]	179.4 (157.1–202.2)	180.5 (158.8–203.0)	174.8 (134.2–206.7)	164.7 (141.3–186.9)	172.0 (145.5–200.9)	170.7 (144.1–195.0)
HDL cholesterol[a]	45.0 (38.1–53.7)	45.2 (38.4–53.9)	44.1 (35.6–60.1)	42.2 (35.9–50.6)	40.4 (34.7–49.4)	46.3 (38.3–54.8)
LDL cholesterol[a]	97.7 (78.9–117.5)	98.7 (80.2–118.4)	83.3 (69.4–115.2)	85.1 (67.3–105.0)	93.1 (71.9–115.2)	90.8 (70.4–111.2)
Hemoglobin[a]	14.4 (13.2–15.5)	14.5 (13.3–15.5)	11.7 (10.3–14.3)	14.0 (12.6–15.2)	11.6 (10.8–12.7)	111.2 (11.4–14.2)
GFR[a]	75.5 (61.2–87.7)	76.6 (63.1–88.5)	67.0 (37.9–87.8)	67.8 (54.3–81.3)	16.2 (8.4–43.8)	59.6 (41.1–74.7)

HT arterial hypertension, *ACE* angiotensin converter enzyme, *ARB* angiotensin II receptor blockers, *DPPIV* dipeptidyl peptidase IV, *CKD* chronic kidney disease, *BMI* body mass index; *MBP* median blood pressure, *HbA1c* glycated hemoglobin, *HDL* high-density lipoproteins, *LDL* low-density lipoproteins, *GFR* glomerular filtration rate

[a]Reported as median and interquartile range

Table 3 Performance consolidated per ML, sampling, and complication

Complication	ML	Sampling	Accuracy	F1-Score	Precision	Recall
AEI	LDA	BLSMOTE	99.9%	46.2%	100.0%	30.0%
ECC	BAG	NONE	94.3%	44.8%	87.0%	33.6%
MOR	RF	ROS	97.4%	63.6%	98.0%	50.9%
ERC	RF	NONE	98.8%	84.5%	94.2%	81.8%

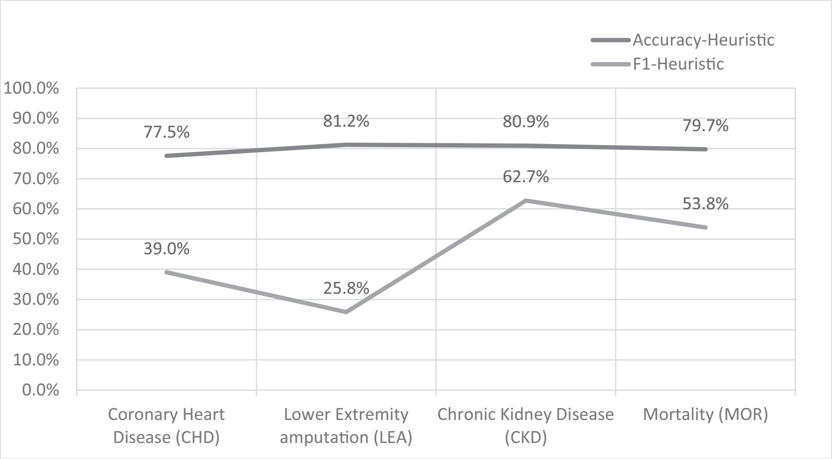


Fig. 4 Performance of the prediction model (ML) for complications associated with diabetes mellitus

Discussion

This paper demonstrates how data mining and computational methods can provide efficient insights for clinical practice through personalized models using individualized and real-time information for each patient for predicting an outcome of interest. Data mining predictive methods can be applied to the development of decision models for procedures like prognosis, diagnosis, and treatment planning. After these have been evaluated and verified, they can be included into automatic and real-time systems of clinical data.

The most relevant variables of our models, according to their potential capacity as early markers for the development of complications, vary per each one of them. The same set of variables was used, but the less predictive were eliminated: For CHD outcome, the correlated variables LDL cholesterol, HDL cholesterol, consumption of ARA2 drugs, GFR, and HbA1c; in the case of chronic kidney disease (CKD), the variables LDL cholesterol, HDL cholesterol, consumption of ACE inhibitors, glycemia, and hemoglobin were eliminated; for lower extremity amputation (LEA), LDL cholesterol, HDL cholesterol, glycemia, and hemoglobin were eliminated; for the mortality (MOR) case, LDL cholesterol, HDL cholesterol, consumption of ACEI drugs, glycemia, and hemoglobin were eliminated.

The best approach to be applied for algorithm assessment needs to be determined. First, a general segmentation of the dataset, comprising imputed and non-imputed data and data applicable to different sampling techniques and ML algorithms, is recommended. This is only possible by performing an exhaustive computational processing, where multiple models are trained based on different sampling techniques and datasets. In this study, more than 3200 experiments were performed, allowing elimination of general hypotheses, such as the one stating that algorithms work better with imputed datasets. For LEA, the best performance was obtained with the non-imputed dataset and without any sampling methods. Based on the tests performed, it could be stated that imputation tends to improve performance (CHD, CKD, and MOR work better); however, this is not a general rule, as in some health predictions (such as LEA), class balancing (patients without complications vs. with complications) and correlation between variables and complications to be predicted are more important.

The algorithm's evaluation depends on its intended use, i.e., it is determined based on the relevance of false positives and false negatives. Accuracy, as a general assessment method, should be revised and used carefully to avoid false negatives. For this reason, the sole evaluation of accuracy is highly limited, and metrics, such as F1-Score, need to be assessed. The F1-Score was the most important metric in this case as the aim was to avoid false negatives or justify the lack of complication when the patient was going to develop it. Based on the obtained results, the CKD case can be used in a clinical environment given its performance in all indicators (80.9% ACC_Heu and 62.7% F1_Heu). For MOR prognosis, its use should be limited to the prognosis of true positives (precision) and interpretation of false positives. For CHD and LEA, its use in clinical environments should be evaluated given its low performance in recall, precision, and F1, even though its performance is close to 80% for accuracy metric heuristic validation (ACC_Heu).

It is noteworthy that by evaluating applicability of the models, the best result was obtained for cases associated with kidney complications. The other models and complications may yield better results if their observation window is broadened, or if the variables are supplemented by unequivocal clinical elements, such as creatinine for CKD. At first, this is particularly interesting because it evidences that an algorithm can help physicians identify patients needing special healthcare by calculating future risk level.

These results are novel as they show alternative options for the treatment of variables, which yield higher metrics in the prediction of CKD (without HV heuristics), such as 81.2% accuracy, 84.5% F1-Score, 94.2% precision, and 81.8% recall. This is evidenced by comparing the aforementioned metrics with the ones reported by Casanova et al. (75% accuracy, 74% recall, and 75% precision) [14], Rau et al. (75% recall and 87.3% F1) [15], Chen et al. (88.6% accuracy) [16], Huang et al. (65.2% accuracy, 63.2% recall, and 67.2% precision) [17], and Chu Su et al. (87% accuracy, 88% precision, and 83% recall) [18].

The success of ML algorithm predictions lies in correct definition, exploration, and assessment of variables based on which algorithms can effectively distinguish patients who may develop a certain complication. All technical efforts should be focused on improving the models, without overfitting and assessing the metrics directly related to prediction of a future disease or the correct prediction of the class showing a complication. Moreover, it is vital to conduct this type of study with a clinical proposal on the correlation between the variables and the element to be predicted; although statistics is useful, assumptions require scientific validation. Variable management over time is essential, and it has been evidenced that synthesizing variables over time using new methods (VAMR or CV) results in insights or key elements based on which machines can make correct predictions on the risk of a complication. Variables forming a dataset must be imputed, and their result must be validated. An exhaustive computational search on the performance of an algorithm should be performed, using an imputed or non-imputed dataset, in addition to accurate application of sampling methods for predicting a given complication.

Some strengths need to be highlighted. This paper describes the application of a modern data mining pipeline, resulting in significant benefits: (1) it applies an exhaustive training process in an iterative manner, exploiting the advantages of significant modern computational services to combine different approaches; in the healthcare field, this strategy results in the utilization of clinical data and development of a trained model acting as the brain of a calculator for the risk of complications in diabetic patients, and (2) it provides a multivariate index of patients' conditions. AI-based strategies were used to input missing data (K-means) and address class unbalance. Models were created considering different prediction approaches and validated through last generation data science principles. Final models demonstrate asymmetry in the predictive performance of each studied complication, suggesting that the variables to be used should be reviewed in detail and differentiated based on each complication to be predicted. A single variable set has limited performance for the prediction of all complications, making it necessary to create pertinent variable groups for each complication.

Also, some limitations need to be recognized. Working with unbalanced classes, which are common in the clinical environment, is a limitation for this type of study because patients presenting a certain condition (incident cases) are in minority. This constitutes a challenge because if a dataset cannot be created with predicting variables based on which ML can differentiate patients, the results obtained could not be used in a clinical environment. Many clinical variables were found to directly correlate with other variables and affect ML performance when included in the model. Variable clearance becomes necessary for avoiding their correlation. Another limitation, and related to retrospective design, is the presence of missing data given that working databases were real-world data collected with other purposes rather than research; although it can be imputed, its clinical feasibility must be demonstrated. Prospective cohorts would be ideal to validate these predictive models.

A key feature is the summarization into variables with repeated measures over time without losing their predictive capacity. Using the CV method, it is possible to synthesize a set of indexed variables to a measurement period preserving their predominance. Before synthesizing repeated measures, data or mistaken clinical measurements need to be cleared and missing values need to be imputed through an adequate segmentation of patients based on variables that have previously been clinically validated as affecting or causing the variable being imputed. Algorithm performance can be assessed with reliable, homogeneously treated, and clinically validated data. The development of predictive models for complications in patients with DM may help assess the correlation between individual factors and a specific complication's onset to consequently stratify the patient for a healthcare center based on that risk and develop tools to support informed clinical decisions regarding treatment.

Future studies should include variables with information on patient lifestyle, such as GPS tracking of the patient's movements, the places visited. Although this may raise controversy in terms of privacy vs. predictive effectiveness, it will enable effective modeling of patient lifestyle and result in highly personalized and effective predictions in addition to clinical data.

Conclusions

ML is a key technology for transforming patient's variables into clinically valuable information through rules developed by medical and engineering experts. This study highlights the ability to predict DM complications, providing valuable support for clinical diagnosis and treatment decisions; ML algorithms can analyze vast amounts of patient data, enabling healthcare professionals to make more informed decisions tailored to individual patients' needs.

The findings of this study emphasize the importance of integrating ML into the field of medicine, as it has the capacity to transform patient data into actionable insights.

By harnessing the power of ML, healthcare professionals can improve patient outcomes, optimize treatment plans, and ultimately enhance the overall quality of care for individuals affected by DM.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13410-023-01264-7.

Acknowledgment The authors want to thank the funding institutions: MINCIENCIAS, EPS SANITAS Colombia, as well as University of Santander (UDES) and University Foundation SANITAS for all the support in this process.

Funding This study was funded by Ministry of Science Technology and Innovation of the Republic of Colombia (a.k.a. MINCIENCIAS) from call 811 of 2018, under code C160I000000011758-19, contract number 433 of 2019, and contingent recovery number 80740-433-2019.

Data Availability Data is available through the corresponding author upon justified request.

Declarations

Competing Interests The authors have no financial or non-financial interests to disclose.

References

1. Situación de la enfermedad renal crónica, la hipertensión arterial y la diabetes mellitus en Colombia 2020 | Cuenta de Alto Costo n.d. https://cuentadealtocosto.org/site/erc/situacion-de-la-enfermedad-renal-cronica-la-hipertension-arterial-y-la-diabetes-mellitus-en-colombia-2020/. Accessed April 22, 2022.
2. Dall TM, Yang W, Gillespie K, Mocarski M, Byrne E, Cintina I, et al. The economic burden of elevated blood glucose levels in 2017: diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. Diabetes Care. 2019;42:1661–8. https://doi.org/10.2337/DC18-1226.
3. Zimmet P, Alberti KG, Magliano DJ, Bennett PH. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. Nat Rev Endocrinol. 2016;12:616–22. https://doi.org/10.1038/nrendo.2016.105.
4. Forbes JM, Cooper ME. Mechanisms of diabetic complications. Physiol Rev. 2013;93:137–88. https://doi.org/10.1152/physrev.00045.2011.
5. Tanaka S, Tanaka S, Iimuro S. Predicting macro- and microvascular complications in type 2 diabetes. Diabetes Care. 2013;36:1193–9. https://doi.org/10.2337/dc12-0958.
6. Laxy M, Schöning VM, Kurz C, Holle R, Peters A, Meisinger C, et al. Performance of the UKPDS outcomes model 2 for predicting death and cardiovascular events in patients with type 2 diabetes mellitus from a German population-based cohort. Pharmacoeconomics. 2019;37:1485–94. https://doi.org/10.1007/S40273-019-00822-4/TABLES/5.
7. Sim J, Kim YA, Kim JH, Lee JM, Kim MS, Shim YM, et al. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Sci Rep. 2020;10:1–12. https://doi.org/10.1038/s41598-020-67604-3.
8. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10: e1379. https://doi.org/10.1002/WIDM.1379.
9. Shamout F, Zhu T, Clifton DA. Machine learning for clinical outcome prediction. IEEE Rev Biomed Eng. 2021;14:116–26. https://doi.org/10.1109/RBME.2020.3007816.
10. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. J Diabetes Sci Technol. 2018;12:295–302. https://doi.org/10.1177/1932296817706375.
11 Levin A, Stevens PE, Bilous RW, Coresh J, De Francisco ALM, De Jong PE, et al. Kidney disease: improving global outcomes (KDIGO) CKD work group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int Suppl. 2011;2013(3):1–150. https://doi.org/10.1038/KISUP.2012.73.
12. Niaksu O. CRISP data mining methodology extension for medical domain. Balt J Mod Comput 2015;3(2):92–109.
13. Abhari S, Kalhori SRN, Ebrahimi M, Hasannejadasl H, Garavand A. Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. Healthc Inform Res. 2019;25:248. https://doi.org/10.4258/HIR.2019.25.4.248.
14. Casanova R, Saldana S, Simpson SL, Lacy ME, Subauste AR, Blackshear C, et al. Prediction of incident diabetes in the Jackson heart study using high-dimensional machine learning. PLoS One. 2016;11:e0163942. https://doi.org/10.1371/journal.pone.0163942.
15. Rau HH, Hsu CY, Lin YA, Atique S, Fuad A, Wei LM, et al. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Comput Methods Programs Biomed. 2016;125:58–65. https://doi.org/10.1016/j.cmpb.2015.11.009.
16. Chen J, Tang H, Huang H, Lv L, Wang Y, Liu X et al (2015) Development and validation of new glomerular filtration rate predicting models for Chinese patients with type 2 diabetes. J Transl Med13.https://doi.org/10.1186/s12967-015-0674-y.
17. Huang GM, Huang KY, Lee TY, Weng JTY (2015) An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. BMC Bioinformatics 16. https://doi.org/10.1186/1471-2105-16-S1-S5.
18. Chu-Su Y, Liu CS, Chen RS, Lin CW. Artificial neural networks for estimating glomerular filtration rate by urinary dipstick for type 2 diabetic patients. Biomed Eng (Singapore). 2016;28:1650016. https://doi.org/10.4015/S1016237216500162.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.