



Prediction of Cardiovascular Complication in Patients with Newly Diagnosed Type 2 Diabetes Using an XGBoost/GRU-ODE-Bayes-Based Machine-Learning Algorithm

Joonyub Lee¹, Yera Choi², Taehoon Ko³, Kanghyuck Lee^{3,4}, Juyoung Shin^{1,3,5}, Hun-Sung Kim^{1,3}

¹Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul; ²NAVER CLOVA AI Lab, Seongnam; Departments of ³Medical Informatics, ⁴Biomedicine and Health Sciences, College of Medicine, The Catholic University of Korea; ⁵Health Promotion Center, Seoul St. Mary's Hospital, Seoul, Korea

Background: Cardiovascular disease is life-threatening yet preventable for patients with type 2 diabetes mellitus (T2DM). Because each patient with T2DM has a different risk of developing cardiovascular complications, the accurate stratification of cardiovascular risk is critical. In this study, we proposed cardiovascular risk engines based on machine-learning algorithms for newly diagnosed T2DM patients in Korea.

Methods: To develop the machine-learning-based cardiovascular disease engines, we retrospectively analyzed 26,166 newly diagnosed T2DM patients who visited Seoul St. Mary's Hospital between July 2009 and April 2019. To accurately measure diabetes-related cardiovascular events, we designed a buffer (1 year), an observation (1 year), and an outcome period (5 years). The entire dataset was split into training and testing sets in an 8:2 ratio, and this procedure was repeated 100 times. The area under the receiver operating characteristic curve (AUROC) was calculated by 10-fold cross-validation on the training dataset.

Results: The machine-learning-based risk engines (AUROC XGBoost=0.781±0.014 and AUROC gated recurrent unit [GRU]-ordinary differential equation [ODE]-Bayes=0.812±0.016) outperformed the conventional regression-based model (AUROC=0.723±0.036).

Conclusion: GRU-ODE-Bayes-based cardiovascular risk engine is highly accurate, easily applicable, and can provide valuable information for the individualized treatment of Korean patients with newly diagnosed T2DM.

Keywords: Cardiovascular diseases; Diabetes mellitus, type 2; Korea; Machine learning

INTRODUCTION

Patients with type 2 diabetes mellitus (T2DM) are at an increased risk of developing vascular complications [1,2]. Sustained hyperglycemia, along with common concomitant T2DM

medical conditions (obesity, hypertension, dyslipidemia, and smoking), exert deleterious effects on endothelial cells throughout the body. Accordingly, patients with diabetes are prone to develop microvascular (retinopathy, nephropathy, and neuropathy) and macrovascular (stroke, myocardial infarction, periph-

Received: 16 May 2023, Revised: 22 July 2023, Accepted: 9 August 2023

Corresponding author: Hun-Sung Kim
Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea
Tel: +82-2-3147-8425, **Fax:** +82-504-292-9080, **E-mail:** 01cadiz@hanmail.net

Copyright © 2024 Korean Endocrine Society

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

eral artery disease, and aortic diseases) complications, significantly increasing the disease-related burden. In particular, cardiovascular complications are life-threatening and may result in life-long sequelae (neurologic symptoms after stroke and heart failure after myocardial infarction) even in those who survive. In the past few decades, remarkable advances have been made in radiologic intervention technologies and medications for managing patients with cardiovascular diseases [3]. These advancements have increased the average lifespan of T2DM patients. Paradoxically, as the number of T2DM patients increases, disease-related expenses substantially burden both the society and medical resources. More intensive and comprehensive therapies (blood glucose, blood pressure, low-density lipoprotein cholesterol [LDL-C] control, and lifestyle modification) have been shown to prevent cardiovascular diseases in patients with T2DM [4-7]. Moreover, cumulative evidence supports the idea that individual diabetes patients are at different risks of developing vascular complications [8]. Therefore, accurately stratifying the risk of developing cardiovascular complications and individualizing therapy in patients with T2DM early after diagnosis is important to efficiently utilize limited medical and socioeconomic resources.

Several risk engines (e.g., Framingham, UK Prospective Diabetes Study [UKPDS], and atherosclerotic cardiovascular disease [ASCVD]) have been devised to estimate the risk of developing cardiovascular complications in patients with T2DM [9-11]. However, concerns regarding the accuracy of these models are constantly raised [12-17]. Moreover, these conventional risk engines were not originally designed to estimate cardiovascular complications in Asian patients with diabetes. As the risk of developing cardiovascular complications varies between ethnicities, the accuracy of these engines may be significantly impaired when applied to Asian patients with T2DM [14]. Therefore, to estimate the risk of cardiovascular complications in Asian patients, developing an accurate risk engine based on the Asian population is critical for managing diabetes efficiently with limited medical resources.

Dealing with insufficient data to estimate sporadic medical events is the main challenge in developing accurate medical risk engines. In recent years, various machine-learning (ML) and deep learning methods have demonstrated greater strengths than conventional methods (e.g., logistic regression) in handling large-scale data. As the volume and availability of medical data increases, the ML-based methods are being widely utilized to predict medical events [18]. eXtreme Gradient Boosting (XGBoost) is a gradient-boosting decision tree-based ML algorithm

that combines several weak decision trees to generate a strong predictive model [19,20]. XGBoost evaluates and weighs the learning errors of weak predictive models and sequentially reflects them in the next learning model [19]. This algorithm is rapid, scalable, and robust to several issues, including multicollinearity, and is increasingly used to analyze high-dimensional medical data that often includes closely correlated features. The gated recurrent unit (GRU)-ordinary differential equation (ODE)-Bayes is a newly developed algorithm that integrates GRU-ODE and a Bayesian network [21]. The GRU-ODE-Bayes model was specifically designed to account for sporadic observations. Because the intervals between patient visits vary greatly and the clinical variables measured at visits tend to differ, sporadic observations are common in medical data and often cause problems in data processing and model selection [18]. Few attempts have been made to predict cardiovascular complications in patients with T2DM using ML algorithms [22-24]. In this study, we present risk engines developed using the XGBoost- and GRU-ODE-Bayes-based ML algorithm for predicting cardiovascular complications in Korean patients with T2DM.

METHODS

Population study

The present study is a retrospective cohort study based on an electronic medical record (EMR). The newly diagnosed T2DM patients who visited Seoul St. Mary's Hospital between July 2009 and April 2019 were enrolled in the study. The diagnosis of T2DM was confirmed by E11-14 (International Classification of Diseases 10th Revision [ICD-10]) codes encoded in the EMR. Cardiovascular diseases covered in this exam were stroke/transient ischemic attack (ICD-10 I60-66, I67.2, I67.8, I69, G45), coronary heart disease (ICD-10 I20-25), peripheral artery disease (ICD-10 I73), and aortic aneurysm (ICD-10 I70-72). Cardiovascular complication development was measured by operational definition, which mandates the diagnosis of cardiovascular disease (defined by the ICD-10 code) and related procedures during the observational period (diagnosis only for stroke/transient ischemic attack). Cardiovascular disease-related procedures included percutaneous coronary intervention, peripheral angiography, peripheral ballooning angioplasty, peripheral artery stent insertion, post-coronary bypass surgery angiography, and aortic intervention (endovascular aortic aneurysm and thoracic endovascular aortic repair).

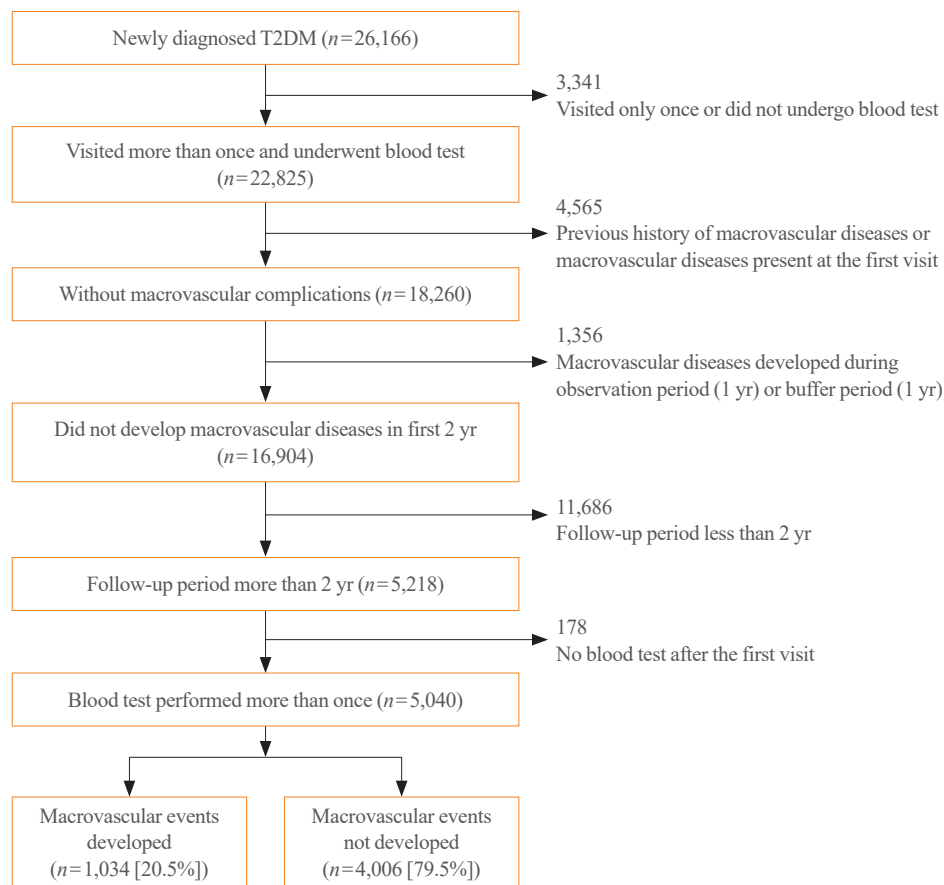


Fig. 1. Study scheme. The study enrolled 26,166 newly diagnosed type 2 diabetes mellitus (T2DM) patients. A total of 5,040 patients were analyzed to establish the model.

Baseline characteristics of the subjects

An ML algorithm-based risk engine was developed for predicting macrovascular diseases in T2DM patients. For this study, patients newly diagnosed with T2DM, who visited Seoul St. Mary's Hospital between July 2009 and April 2019, were enrolled. Patients less than 18 years old were excluded, and a total of 26,166 patients were recruited. Patients who visited only once or did not undergo a blood test ($n=3,341$) were also excluded. Furthermore, patients with a history of cardiovascular diseases or who presented with cardiovascular diseases at the first visit were excluded ($n=4,565$). A 1-year observation period and a 1-year buffer period were included in the study to maximize the predictability of cardiovascular diseases and minimize non-diabetes-related cardiovascular events. Patients who developed cardiovascular events during this period ($n=1,356$) were excluded. Out of the 5,040 patients, subjects whose follow-up period was less than 2 years ($n=11,686$) and those who did not undergo a blood test after the first visit ($n=178$) were excluded. Thus, the final sample for analysis included 5,040 patients, among which

1,034 patients (20.5%) developed cardiovascular complications (Fig. 1). When baseline characteristics were compared, body mass index (BMI; 24.30 ± 3.34 kg/cm² vs. 24.55 ± 3.74 kg/cm²), aspartate aminotransferase (AST; 24.96 ± 13.48 U/L vs. 26.59 ± 25.13 U/L), alanine aminotransferases (ALT; 28.88 ± 19.50 U/L vs. 31.63 ± 34.01 U/L), and triglycerides (150.23 ± 131.15 mg/dL vs. 144.65 ± 110.63 mg/dL) were comparable between patients who developed cardiovascular complications and those who did not develop cardiovascular complications (Table 1).

Compared to patients without cardiovascular complications, those who developed cardiovascular complications were aged (over 60 years; 61.50% vs. 44.18%), had a higher proportion of females (50.87% vs. 43.66% male), systolic blood pressure (135.92 ± 21.41 mm Hg vs. 129.29 ± 18.43 mm Hg), diastolic blood pressure (79.60 ± 11.39 mm Hg vs. 77.12 ± 10.58 mm Hg), and hemoglobin A1C (HbA1c) levels ($7.78\% \pm 1.80\%$ vs. $7.45\% \pm 1.79\%$). Patients who developed cardiovascular complications had lower (78.13 ± 27.62 mL/min/1.73 m² vs. 84.22 ± 22.91 mL/min/1.73 m²) total cholesterol (169.42 ± 43.76 mg/dL

Table 1. Baseline Characteristics of Subjects

Characteristic	Macrovascular complications			Developed prediction model	
	Developed	Not developed	<i>P</i> value	Training model	Test model
Number	1,034 (20.5)	4,006 (79.5)		3,528 (70.0)	1,512 (30.0)
Age				59 (52–66)	59 (51–66)
18–59 yr, %	38.5	55.82		51.9	51.8
≥60 yr, %	61.5	44.18		48.4	48.2
Male sex	43.7	50.87		46.9	47.0
Height, cm	162.7±8.9	162.20±9.08	0.388	162 (155–170)	163.0 (156.0–169.0)
Weight, kg	64.6±11.6	64.89±12.80	<0.050	63 (56–72)	63.0 (56.0–71.0)
BMI, kg/m ²	24.3±3.3	24.55±3.74	0.215	24.1 (22.1–26.2)	24.0 (22.0–26.2)
SBP, mm Hg	136±21	129±18	<0.050	130 (120–140)	130 (120–140)
DBP, mm Hg	80±11	77±11	<0.050	80 (70–82)	79 (70–82)
HbA1c, %	7.8±1.8	7.5±1.8	<0.050	6.9 (6.3–7.8)	6.8 (6.3–7.8)
BUN, mg/dL	19.6±13.4	17.0±7.1	<0.050	15.5 (12.7–19.3)	15.8 (12.9–19.2)
Creatinine, mg/dL	1.1±1.2	0.9±0.6	<0.050	0.8 (0.7–1.0)	0.8 (0.7–1.0)
GFR, mL/min/1.73 m ²	78.1±27.6	84.2±22.9	<0.050	84.3 (71.0–97.5)	83.77 (69.8–96.4)
AST, U/L	25±14	27±25	0.116	22 (18–28)	22 (18–27)
ALT, U/L	29±20	32±34	0.058	24 (17–35)	23.0 (17–34)
Total cholesterol, mg/dL	169±44	176±42	<0.050	168.0 (144.0–195.0)	168 (144–196)
Triglyceride, mg/dL	150±132	145±111	0.457	119.0 (80.0–175.0)	118 (80–167)
HDL-C, mg/dL	44±12	46±13	<0.050	45.0 (38.0–53.0)	45 (39–53)
LDL-C, mg/dL	95±33	100±35	<0.050	93.0 (73.0–116.0)	92 (73–116)
CV complications, %	-	-		20.4	19.4

Values are expressed as number (%) or mean±standard deviation unless otherwise indicated. Student's *t* test was used for statistical analysis, and a *P*<0.05 was regarded as statistically significant. The baseline characteristics of patients who developed macrovascular complications and those who did not develop macrovascular complications were compared (2nd–4th column). The data were expressed after the missing values (Supplemental Table S1) were discarded. The patient data for training and testing were distributed in a 7:3 ratio with comparable baseline characteristics (5th–6th columns).

BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; HbA1c, hemoglobin A1C; BUN, blood urea nitrogen; GFR, glomerular filtration rate; AST, aspartate aminotransferase; ALT, alanine aminotransferase; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; CV, cardiovascular.

vs. 175.51±42.30 mg/dL), high-density lipoprotein cholesterol (HDL-C; 44.39±11.76 mg/dL vs. 46.16±12.69 mg/dL), and LDL-C (94.56±33.10 mg/dL vs. 99.50±34.97 mg/dL) compared to those with no cardiovascular complications (Table 1). For the study, 5,040 newly diagnosed T2DM patients were enrolled whose baseline characteristics were older, female dominant, higher blood pressure and HbA1c, and a lower estimated glomerular filtration rate (eGFR) in the patients who developed cardiovascular complications compared to those without cardiovascular complications.

Study design

The study was designed to cover three periods (Fig. 2). During the observation period, the input features from the baseline in-

formation were measured. Age, sex, medication history, height, weight, blood pressure, HbA1c, blood urea nitrogen (BUN), creatinine, AST, ALT, total cholesterol, triglycerides, HDL-C, and LDL-C were also measured. We included a buffer period to screen out non-diabetes-related cardiovascular events to minimize the overestimation of cardiovascular events. The buffer period was followed by an observation period during which cardiovascular complications could possibly occur. To allocate an appropriate duration for each period, we examined the area under the receiver operating characteristic curve (AUROC) of the risk engines by changing the duration of each period in the dataset. Based on these data, 1 year each was allocated for the observation and buffer periods, respectively, and 5 years for the outcome period.

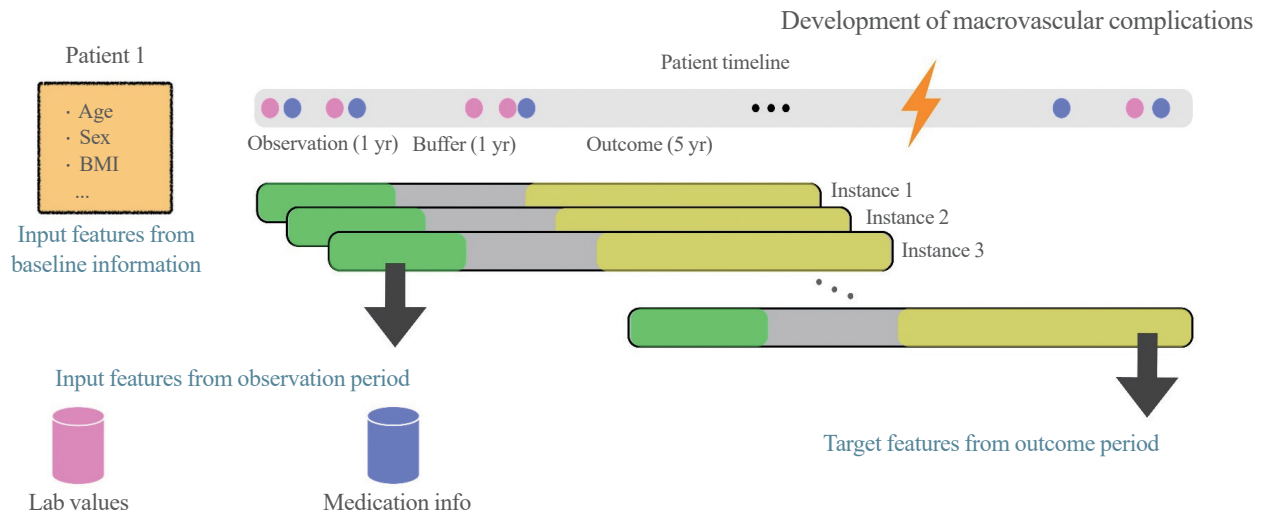


Fig. 2. Study design. To minimize non-diabetes-related cardiovascular events and maximize the accuracy of the model, observation (1 year), buffer (1 year), and outcome (5 years) periods were included in the study design. Variables for the model were acquired during the observation period and cardiovascular events were measured during the outcome period. Cardiovascular events that developed during the buffer period were excluded. BMI, body mass index.

XGBoost/GRU-ODE-Bayes-based ML analysis

The target outcome for both the models was defined as a binary variable indicating the occurrence of macrovascular complications during the outcome period. We examined 72 hyperparameter sets for GRU-ODE-Bayes and XGBoost analyses. The provided data were randomly split into training and testing sets in an 8:2 ratio and repeated for 100 times. We performed a 10-fold cross-validation on the training dataset to compute the mean and standard deviations of AUROC and areas under the precision-recall curve (AUPRC). Considering that models with high AUROC also exhibited good AUPRC performance, the hyperparameter set that yielded the highest average AUROC for GRU-ODE-Bayes and XGBoost was selected as the final set (Supplemental Materials S1, S2). Using the previously selected optimal hyperparameter set, we trained the GRU-ODE-Bayes and XGBoost-based models with the training dataset and evaluated their performance using the test set. To assess the efficacy and robustness of the model, AUROC, AUPRC, positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, F1 score, and accuracy were measured. We set the threshold for PPV, NPV, sensitivity, specificity, F1 score, and accuracy at 0.5.

XGBoost

The XGBoost model was trained using the official XGBoost binary package version 1.5.0. Hyperparameters of XGBoost, such as the learning rate, number of trees, maximum tree depth, L2

regularization term, minimum split loss, subsample ratio of records, and columns, were optimized by the grid search method. The final XGBoost model was trained with the following hyperparameters: learning rate=0.03; number of trees=100; maximum tree depth=3; L2 regularization term=0.01; subsample ratio of records=0.8; and subsample ratio of columns (for each tree)=0.6.

GRU-ODE-Bayes

In the GRU-ODE-Bayes model, variables that dynamically change over the observation period (laboratory values and medication information, including HbA1c, BUN, creatinine, AST, ALT, eGFR, total cholesterol, triglyceride, HDL-C, LDL-C, prescribing sulfonylurea, metformin, sulfonylurea/metformin combination, meglitinide, alpha-glucosidase inhibitor/metformin combination, thiazolidindione, dipeptidyl peptidase-4 inhibitors, metformin/dipeptidyl peptidase-4 inhibitor combination sodium glucose transporter 2 inhibitors, insulin, lixisenatide, exenatide, dulaglutide, statins, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, calcium channel blockers, and angiotensin receptor blocker compounds) were treated as input variables, whereas a set of baseline demographic/clinical variables (including age, sex, height, weight, BMI, systolic blood pressure, diastolic blood pressure, HbA1c, BUN, creatinine, AST, ALT, eGFR, total cholesterol, triglycerides, HDL-C, and LDL-C), which were assumed to remain stable over time, were included as static covariate to determine the initial hidden state

of the model. In addition to the dynamic input variables, a corresponding observation mask was created to indicate the observed/missing values (Supplemental Table S1) and fed into the model. Missing static covariates values were replaced with the respective feature-wise median values.

The official PyTorch repository for GRU-ODE-Bayes was adapted for the current dataset. The final GRU-ODE-Bayes model was trained using the following settings and hyperparameters: solver=Dopri5; epochs=100 (early stopping if the loss does not decrease below 0.001); optimizer=Adam; learning rate=0.001; hidden siz=100; no dropouts, lambda (weighting between classification and mean squared error loss)=0.0001 (more weight on mean squared error loss); delta t (time step)=0.1; and T (total time)=100.

Statistical analysis

The baseline characteristics of patients in the groups that developed and did not develop cardiovascular events were compared using the Student's *t* test. A receiver operating characteristic curve was drawn, and the AUROC was measured to assess the validity of the models. The accuracy of the models was compared using non-parametric analysis [25]. Multivariate logistic regression analysis was used to develop a conventional risk engine, which was compared to ML-based risk engines. A $P < 0.05$ was considered statistically significant. Machine learning models were developed in Python version 3.10.4, and Pytorch version 1.12.1, XGBoost version 1.7.0, and Scikit-learn version 1.1.2 packages were used. For GRU-ODE-Bayes, we used the code from the Github repository (https://github.com/edebrouwer/gru_ode_bayes). All statistical analyses were performed using the R software version 4.2.0 (R Foundation for Statistical Computing, Vienna, Austria).

Data protection and privacy

The data were stored as encrypted files on a password-secured computer and were accessible only to the principal investigator.

All the data were anonymized and did not contain personally identifiable information [26]. Subsequently, the data were passed on to analysts or model developers. Because this was a retrospective analysis, the participants were not at risk of physical or mental harm; therefore, consent was not required. This study was approved by the Institutional Review Board of Seoul St. Mary's Hospital, Catholic University of Korea (KC20RISI0844).

RESULTS

XGBoost-/GRU-ODE-Bayes-based ML algorithm predict the cardiovascular complications in patients with T2DM

The purpose of this study was to predict the risk of macrovascular diseases in newly diagnosed patients with T2DM using readily obtainable patient medical data (such as height, weight, systolic/diastolic blood pressures, and blood test results from separate visits) through the implementation of an ML algorithm. To maximize the accuracy of the risk engine, we designed a 1-year observation period that collected patients' information during hospital visit and over a buffer period of 1 year to minimize non-T2DM-related cardiovascular events. During the next 5 years of the observation period, the development of cardiovascular events was measured. The input data and development of cardiovascular events were respectively applied for training by the XGBoost- and GRU-ODE-Bayes-based ML algorithms. We also performed logistic regression analysis with the current dataset to compare the accuracy of ML-based risk engine with that of conventional methods.

The AUROC of the GRU-ODE-Bayes- and XGBoost-based engines was 0.812 ± 0.016 and 0.781 ± 0.014 , respectively (Table 2, Fig. 3), while that of the conventional logistic regression analysis-based risk engine was 0.723 ± 0.036 . The AUROC of GRU-ODE-Bayes showed a statistically significant difference compared to that of XGBoost ($P=0.013$) and logistic regression ($P < 0.001$) (Table 3). The ML-based cardiovascular risk engines

Table 2. The Performance of XGBoost/GRU-ODE-Bayes-Based Machine-Learning Algorithm for Predicting the Cardiovascular Complications in Patients with Type 2 Diabetes Mellitus

	AUROC	AUPRC	PPV	NPV	Sensitivity	Specificity	F1	Accuracy
GRU-ODE-Bayes	0.812 ± 0.016	0.607 ± 0.033	0.889 ± 0.041	0.868 ± 0.004	0.251 ± 0.027	0.994 ± 0.003	0.39 ± 0.033	0.869 ± 0.005
XGBoost	0.78 ± 0.014	0.569 ± 0.029	0.842 ± 0.065	0.851 ± 0.006	0.233 ± 0.036	0.993 ± 0.004	0.364 ± 0.047	0.85 ± 0.008

Values are expressed as mean \pm standard deviation.

XGBoost, eXtreme Gradient Boosting; GRU, gated recurrent unit; ODE, ordinary differential equation; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; PPV, positive predictive value; NPV, negative predictive value.

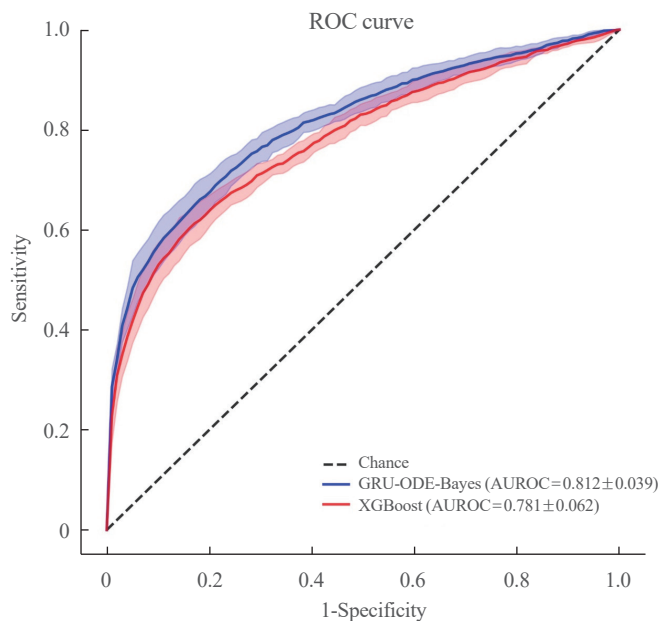


Fig. 3. The eXtreme Gradient Boosting (XGBoost)/gated recurrent unit (GRU)-ordinary differential equation (ODE)-Bayes-based machine-learning algorithm predicts cardiovascular complications in patients with type 2 diabetes mellitus. A receiver operating characteristic curve was drawn based on the machine-learning algorithms (GRU-ODE-Bayes [blue] and XGBoost [red]). The area under the receiver operating characteristic curve (AUROC) was measured (right bottom). The ± 1 standard deviation was indicated as grey area in the curve. ROC, receiver operating characteristic.

predicted cardiovascular events in newly diagnosed T2DM patients with high accuracy. The accuracy of these ML-based risk engines was superior to that of conventional logistic regression methods, and the GRU-ODE-Bayes-based engines exhibited the best performance.

DISCUSSION

Patients with T2DM are at different risks of developing cardiovascular complications that are life-threatening but preventable when managed appropriately [4-7]. A precise cardiovascular risk engine can enable individualized therapy to prevent cardiovascular complications in T2DM patients. In this study, we developed a highly functional cardiovascular risk engine using XGBoost- and GRU-ODE-Bayes-based ML algorithms for newly diagnosed T2DM patients. These engines estimated the risk of developing cardiovascular complications with high accuracy (AUROC of GRU-ODE-Bayes and XGBoost was 0.812 ± 0.016 and 0.781 ± 0.014 , respectively) and were superior to the conventional logistic regression method (AUROC = 0.723 ± 0.036).

Table 3. Comparison of AUROCs of Models

	GRU-ODE-Bayes	XGBoost	Logistic regression
GRU-ODE-Bayes	-	0.013	<0.001
XGBoost	0.013	-	<0.001
Logistic regression	<0.001	<0.001	-

The AUROC of each macrovascular risk engine was compared by paired *t* test. All *P* values were calculated by paired *t* test for the results of 100 replicates.

AUROC, area under the receiver operating characteristic curve; GRU, gated recurrent unit; ODE, ordinary differential equation; XGBoost, eXtreme Gradient Boosting.

Previous studies attempted to estimate the risk of cardiovascular disease in patients with T2DM [9-11]. However, questions regarding the accuracy of these models are constantly raised [12-17]. The Framingham risk engine underestimates the risk of cardiovascular disease in patients with diabetes [12,16], while the UKPDS engine overestimates the risk [13-15]. The most recently developed ASCVD risk engines also overestimate the risk of cardiovascular disease [17]. Recent advancements in ML algorithms have provided researchers with powerful tools for developing more accurate cardiovascular risk engines. Longato et al. [23] developed an recurrent neural network-based cardiovascular risk engine. This model was developed based on 214,676 patients with diabetes, including those with cardiovascular disease, to estimate the 4-point major adverse cardiovascular events (death, heart failure, infarction, and stroke) [23]. The model was developed for administrative purposes, and the AUROC ranged from 0.792 to 0.812 [23]. Ravaut et al. [24] demonstrated an adverse outcome (including micro and macrovascular complications) prediction model for patients with diabetes. This model was based on 1,029,366 patients from Ontario, Canada, to predict the three-year risk of adverse outcomes [24]. The model was developed for administrative purposes, and the AUROC ranged from 0.794 to 0.796 for cardiovascular disease and 0.689 to 0.692 for amputation [24]; the researchers adopted XGBoost, which was superior to GRU-D or logistic regression (AUROC = 0.725 to 0.786) [24]. Alaa et al. [22] proposed a ML-based cardiovascular disease risk prediction model for the general population of United Kingdom. This model was developed based on 423,604 individuals from the general population without cardiovascular disease as the baseline [22]. AutoPrognosis, which provides the best ML algorithm for analysis, was used for the model. The AutoPrognosis framework includes XGBoost but not GRU-ODE-Bayes [22]. The AUROC of this model reached

0.774, whereas that of the conventional regression model reached 0.758 [22].

The purpose of this study was to develop a cardiovascular complication risk engine that can provide physicians with accurate information on cardiovascular risks in newly diagnosed T2DM patients in a routine clinical setting. From this perspective, our model has several advantages over those reported previously. First, our study was based solely on patients with newly diagnosed T2DM without a history of cardiovascular disease. Previous studies were based on the general population or on diabetes patients with or without a cardiovascular history [22-24]. Second, our data are based on real-world clinical EMR data, whereas previous studies were mostly based on health administrative data developed for public healthcare purposes. Third, our study was thoroughly designed for the Korean population, whereas the previous risk engine researchers underrepresented the Asian population in their studies, which would substantially affect the accuracy of prediction in clinical applications [22-24]. Fourth, to maximize the quality of data and accuracy of the model, we designed a buffer period and excluded patients who developed cardiovascular events during the buffer and observation periods. A similar study scheme was demonstrated for an engine developed by Ravaut et al. [24] but not in others. The risk engine developed by Ravaut et al. [24] adopted 2 years of observation and a 3-year buffer period. Finally, our models estimate the risk of cardiovascular events in T2DM patients with high accuracy. The accuracy of this model (AUROC of GRU-ODE-Bayes and XGBoost was 0.812 ± 0.016 and 0.781 ± 0.014 , respectively) was comparable to that of the Italian model (AUROC=0.792 to 0.812) and superior to that of the Canadian model (AUROC=0.794 to 0.796 for cardiovascular disease and 0.689 to 0.692 for amputation) and United Kingdom model (AUROC=0.774). Importantly, some of these models adopted the XGBoost ML algorithm but not GRU-ODE-Bayes. To our knowledge, this is the first cardiovascular complication risk stratification model to utilize the GRU-ODE-Bayes algorithm. Because GRU-ODE-Bayes outperformed XGBoost in estimating cardiovascular events on our data, we believe that this research will encourage other researchers to adopt the GRU-ODE-Bayes algorithm to estimate other medical events. Previous studies have consistently demonstrated the superiority of ML-based algorithms over conventional regression models. In agreement with these previous observations, our study shows the superiority of ML algorithms (AUROC of GRU-ODE-Bayes and XGBoost was 0.812 ± 0.016 and 0.781 ± 0.014 , respectively) over conventional regression models (AUROC= 0.723 ± 0.036)

in predicting cardiovascular complications [22-24]. Despite these advantages, this study has certain limitations. First, this was a single-center retrospective analysis of a relatively small number of patients. However, as this was a well-refined, relatively well-controlled dataset, we speculate that the quality of the data partly compensates for the low quantity [26]. Owing to the retrospective nature of the study, the possibility remains that factors (e.g., smoking), which can potentially influence cardiovascular complications [27-29], were omitted despite our best effort to incorporate maximum number of risk factors. Second, we were unable to include the cardiovascular events that occurred outside the hospital. The definition of a cardiovascular event in our study requires a diagnosis of cardiovascular disease and related procedures during the observational period. Consequently, events that occurred outside the hospital were not considered. This oversight may have led to an underestimation of the actual number of cardiovascular events. Future studies based on larger numbers of patients in multiple centers from other populations in Asia will increase the applicability and accuracy of the model. Organ-oriented risk engines (such as brain, heart, aorta, and peripheral vessels) or engines that stratify the risk of other diabetes-related medical issues (such as hypoglycemia and microvascular complications) should be studied in the future.

In summary, we developed a cardiovascular risk stratification model using an ML algorithm for patients with newly diagnosed T2DM. Our data suggest that ML-based algorithms, particularly GRU-ODE-Bayes, can predict cardiovascular events more accurately than conventional logistic regression-based methods. The GRU-ODE-Bayes-based model is highly accurate and easily applicable because it only requires variables that are readily obtained in routine clinical care settings. Future real-world validation of this model by measuring the occurrence of cardiovascular events and the long-term effect of individualized therapy based on the GRU-ODE-Bayes-based engine will confirm the value of this study.

CONFLICTS OF INTEREST

This study was supported by the Daewoong Pharmaceutical company. The opinions expressed in this paper are those of the authors and do not necessarily represent those of Daewoong Pharmaceutical company. While this study received technical guidance from NAVER CLOVA AI Lab for the development of an AI prediction model, it had no effect on the research outcomes.

AUTHOR CONTRIBUTIONS

Conception or design: H.S.K. Acquisition, analysis, or interpretation of data: J.L., Y.C., T.K., K.L., J.S., H.S.K. Drafting the work or revising: J.L., H.S.K. Final approval of the manuscript: J.L., Y.C., T.K., K.L., J.S., H.S.K.

ORCID

Joonyub Lee <https://orcid.org/0000-0003-0533-9786>

Hun-Sung Kim <https://orcid.org/0000-0002-7002-7300>

REFERENCES

- Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* 2000;321:405-12.
- Emerging Risk Factors Collaboration; Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* 2010;375:2215-22.
- Gregg EW, Sattar N, Ali MK. The changing face of diabetes complications. *Lancet Diabetes Endocrinol* 2016;4:537-47.
- Gaede P, Lund-Andersen H, Parving HH, Pedersen O. Effect of a multifactorial intervention on mortality in type 2 diabetes. *N Engl J Med* 2008;358:580-91.
- UK Prospective Diabetes Study Group. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. *BMJ* 1998;317:703-13.
- Pan A, Wang Y, Talaie M, Hu FB. Relation of smoking with total mortality and cardiovascular events among patients with diabetes mellitus: a meta-analysis and systematic review. *Circulation* 2015;132:1795-804.
- Colhoun HM, Betteridge DJ, Durrington PN, Hitman GA, Neil HA, Livingstone SJ, et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. *Lancet* 2004;364:685-96.
- Ahlqvist E, Storm P, Karajamaki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* 2018;6:361-9.
- Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB Sr, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. *J Am Coll Cardiol* 2014;63(25 Pt B):2935-59.
- Stevens RJ, Kothari V, Adler AI, Stratton IM; United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56). *Clin Sci (Lond)* 2001;101:671-9.
- Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837-47.
- Guzder RN, Gatling W, Mullee MA, Mehta RL, Byrne CD. Prognostic value of the Framingham cardiovascular risk equation and the UKPDS risk engine for coronary heart disease in newly diagnosed type 2 diabetes: results from a United Kingdom study. *Diabet Med* 2005;22:554-62.
- van der Heijden AA, Ortegón MM, Niessen LW, Nijpels G, Dekker JM. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: the Hoorn Study. *Diabetes Care* 2009;32:2094-8.
- Yang F, Ye J, Pomerantz K, Stewart M. Potential modification of the UKPDS risk engine and evaluation of macrovascular event rates in controlled clinical trials. *Diabetes Metab Syndr Obes* 2013;6:247-56.
- 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: Preterax and Diamicron-MR Controlled Evaluation (ADVANCE) Study. *Diabetologia* 2010;53:821-31.
- McEwan P, Williams JE, Griffiths JD, Bagust A, Peters JR, Hopkinson P, et al. Evaluating the performance of the Framingham risk equations in a population with diabetes. *Diabet Med* 2004;21:318-23.
- DeFilippis AP, Young R, Carrubba CJ, McEvoy JW, Budoff MJ, Blumenthal RS, et al. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a

- modern multiethnic cohort. *Ann Intern Med* 2015;162:266-75.
18. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317-8.
 19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016 Aug 13-17; San Francisco, CA. New York: Association for Computing Machinery; 2016. p. 785-94.
 20. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29:1189-232.
 21. De Brouwer E, Simm J, Arany A, Moreau Y. GRU-ODE-Bayes: continuous modeling of sporadically-observed time series. In: Wallach H, Larochelle H, Beygelzimer A, d'Alche-Buc F, Fox E, Garnett R. *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*; 2019 Dec 8-14; Vancouver. San Diego: Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2020. p. 7347-58.
 22. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One* 2019;14:e0213653.
 23. Longato E, Fadini GP, Sparacino G, Avogaro A, Tramontan L, Di Camillo B. A deep learning approach to predict diabetes' cardiovascular complications from administrative claims. *IEEE J Biomed Health Inform* 2021;25:3608-17.
 24. Ravaut M, Sadeghi H, Leung KK, Volkovs M, Kornas K, Harish V, et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *NPJ Digit Med* 2021;4:24.
 25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837-45.
 26. Shin SY, Kim HS. Data pseudonymization in a range that does not affect data quality: correlation with the degree of participation of clinicians. *J Korean Med Sci* 2021;36:e299.
 27. Kim HS, Kim DJ, Yoon KH. Medical big data is not yet available: why we need realism rather than exaggeration. *Endocrinol Metab (Seoul)* 2019;34:349-54.
 28. Kyoung DS, Kim HS. Understanding and utilizing claim data from the Korean National Health Insurance Service (NHIS) and Health Insurance Review & Assessment (HIRA) database for research. *J Lipid Atheroscler* 2022;11:103-10.
 29. Lee S, Kim HS. Prospect of artificial intelligence based on electronic medical record. *J Lipid Atheroscler* 2021;10:282-90.