

Original Paper

# Predicting Risk of Hypoglycemia in Patients With Type 2 Diabetes by Electronic Health Record–Based Machine Learning: Development and Validation

Hao Yang<sup>1\*</sup>, MSc; Jiayi Li<sup>2\*</sup>, MSc; Siru Liu<sup>3</sup>, PhD; Xiaoling Yang<sup>4</sup>, MSc; Jialin Liu<sup>1,5</sup>, MD

<sup>1</sup>Information Center, West China Hospital, Sichuan University, Chengdu, China

<sup>2</sup>Department of Clinical Laboratory Medicine, Jinniu Maternity and Child Health Hospital of Chengdu, Chengdu, China

<sup>3</sup>Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

<sup>4</sup>West China School of Nursing, Endocrinology and Metabolism Department, West China Hospital, Sichuan University, Chengdu, China

<sup>5</sup>Department of Medical Informatics, West China Medical School, Chengdu, China

\*these authors contributed equally

**Corresponding Author:**

Jialin Liu, MD

Information Center

West China Hospital

Sichuan University

No 37 Guoxue Road

Chengdu, 610041

China

Phone: 86 28 85422306

Fax: 86 28 85582944

Email: [dljl8@163.com](mailto:dljl8@163.com)

## Abstract

**Background:** Hypoglycemia is a common adverse event in the treatment of diabetes. To efficiently cope with hypoglycemia, effective hypoglycemia prediction models need to be developed.

**Objective:** The aim of this study was to develop and validate machine learning models to predict the risk of hypoglycemia in adult patients with type 2 diabetes.

**Methods:** We used the electronic health records of all adult patients with type 2 diabetes admitted to West China Hospital between November 2019 and December 2021. The prediction model was developed based on XGBoost and natural language processing. F1 score, area under the receiver operating characteristic curve (AUC), and decision curve analysis (DCA) were used as the main criteria to evaluate model performance.

**Results:** We included 29,843 patients with type 2 diabetes, of whom 2804 patients (9.4%) developed hypoglycemia. In this study, the embedding machine learning model (XGBoost3) showed the best performance among all the models. The AUC and the accuracy of XGBoost are 0.82 and 0.93, respectively. The XGBoost3 was also superior to other models in DCA.

**Conclusions:** The Paragraph Vector–Distributed Memory model can effectively extract features and improve the performance of the XGBoost model, which can then effectively predict hypoglycemia in patients with type 2 diabetes.

(*JMIR Med Inform* 2022;10(6):e36958) doi: [10.2196/36958](https://doi.org/10.2196/36958)

**KEYWORDS**

diabetes; type 2 diabetes; hypoglycemia; learning; machine learning model; EHR; electronic health record; XGBoost; natural language processing

## Introduction

Diabetes is a serious long-term disease. The global prevalence of diabetes in people aged 20-79 years is estimated to be 10.5%

(536.6 million) in 2021 and will rise to 12.2% (783.2 million) by 2045. Global health expenditures related to diabetes are estimated US \$966 billion in 2021 and projected to reach US

\$1054 billion by 2045 [1]. Diabetes continues to be a major clinical and public health concern [2].

Hypoglycemia (blood glucose < 3.9 mmol/L or 70 mg/dL) is a common adverse event of diabetes treatment. Hospital hypoglycemia occurs in 3%-18% of hospitalized diabetic patients [3]. Severe hypoglycemia usually causes potentially life-threatening complications and is associated with an increase length of stay and mortality [4,5]. Hypoglycemia is especially common in older patients with diabetes [5], and the risk doubles every decade after the age of 60 years [6]. Many factors can lead to a high risk of hypoglycemia in older patients, including physiological changes in drug metabolism, age-related decline in renal function, cognitive decline, an increase in comorbidity, and potential overtreatment [7,8]. Since there are many risk factors that induce hypoglycemia in patients with diabetes, and some risk factors may also change during hospitalization, it is a challenge to identify and prevent hypoglycemia in people with diabetes [9,10].

In recent years, machine learning has been widely used for hypoglycemia prediction. For example, Schroeder et al [11] employed the Cox prediction model for the 6-month risk of hypoglycemia. Karter et al [12] developed a tool to identify patients with type 2 diabetes at a high risk of hypoglycemia. Plis et al [13] described a support vector regression model for predicting hypoglycemic events. Furthermore, Jin et al [14] have integrated deep learning with natural language processing (NLP) to automatically detect hypoglycemic events from electronic health record (EHR) notes.

Although numerous hypoglycemia prediction models have been developed, there is still a need to improve the accuracy and

effectiveness of hypoglycemia prediction. In this study, we developed XGBoost ensembling NLP to predict the risk of hypoglycemia in hospitalized patients with type 2 diabetes, using data readily available in the EHRs.

## Methods

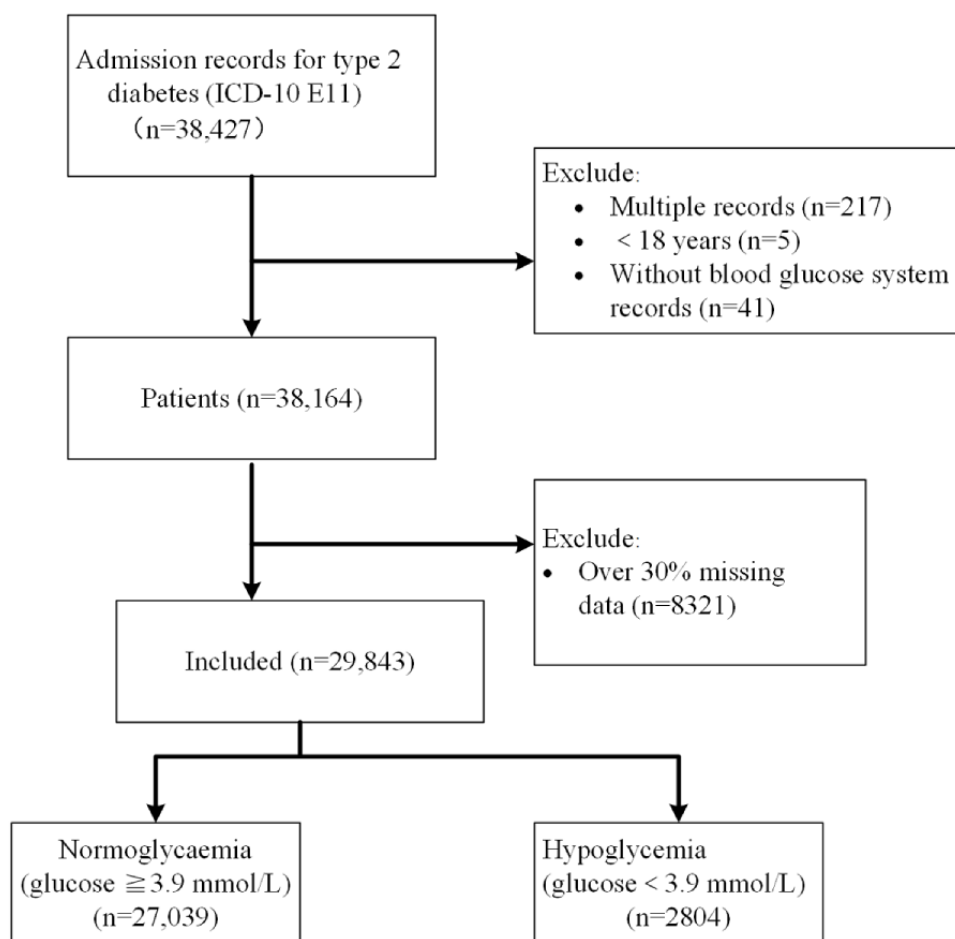
Our cohort included patients with type 2 diabetes from West China Hospital of Sichuan University. All patient data were obtained from the hospital's EHR system.

### Ethics Approval

The study was approved by the Medical Ethics Committee of West China Hospital Sichuan University (2020-608). West China Hospital is a large teaching hospital with 4300 beds and a leading medical center of western China [15].

### Patients

We performed a retrospective analysis of the available EHR of all patients with type 2 diabetes who were admitted to West China Hospital between November 2019 and December 2021. With the protection of patient privacy, only data related to the patient's hospitalization were retrieved, and the diagnosis was established based on International Classification of Diseases, 10th revision (ICD-10). The following inclusion criteria were used: (1) all patients with type 2 diabetes based on ICD-10, E11 (type 2 diabetes mellitus) with a hospital stay > 24 hours; (2) patients aged 18 years or older. Patients with more than 30% missing values were excluded from the analysis [16]. The patient selection process is shown in Figure 1.

**Figure 1.** The patient selection process.

### Variables Analyzed

The variables used to predict the risk of hypoglycemia in patients with type 2 diabetes included various demographic, laboratory, and clinical variables, as well as EHR notes. The

extraction of variables was based on experts' opinion and our research [16-20]. These variables were collected during the first 24 hours of admission. Through data preprocessing, we analyzed some missing values (Table 1). Random forest regression was used to handle all missing numerical variables.

**Table 1.** Statistics of missing values (N=29,843).

Features	Missing data, n (%)
Red blood cell count	1860 (6.2)
Hemoglobin	1858 (6.2)
Blood platelet count	1883 (6.3)
White blood cell count	1858 (6.2)
Total protein	1791 (6.0)
Albumin	1768 (5.9)
Globulin	1812 (6.1)
Urea	1755 (5.9)
Alanine aminotransferase	1821 (6.1)
Aspartate aminotransferase	1809 (6.1)
Cholesterol	2126 (7.1)
High-density lipoprotein	2128 (7.1)
Low-density lipoprotein	2131 (7.1)
Sodium	1516 (5.1)
Chlorine	1585 (5.3)
Thrombin time	3970 (13.3)
Creatinine	1749 (5.9)
Uric acid	1769 (5.9)
C-reactive protein	18,249 (61.1)
Procalcitonin	20,101 (67.3)
Glycosylated hemoglobin or HbA <sub>1c</sub> <sup>a</sup>	14,410 (48.3)
Prothrombin time	3725 (12.5)
Activated partial thromboplastin time	3779 (12.7)

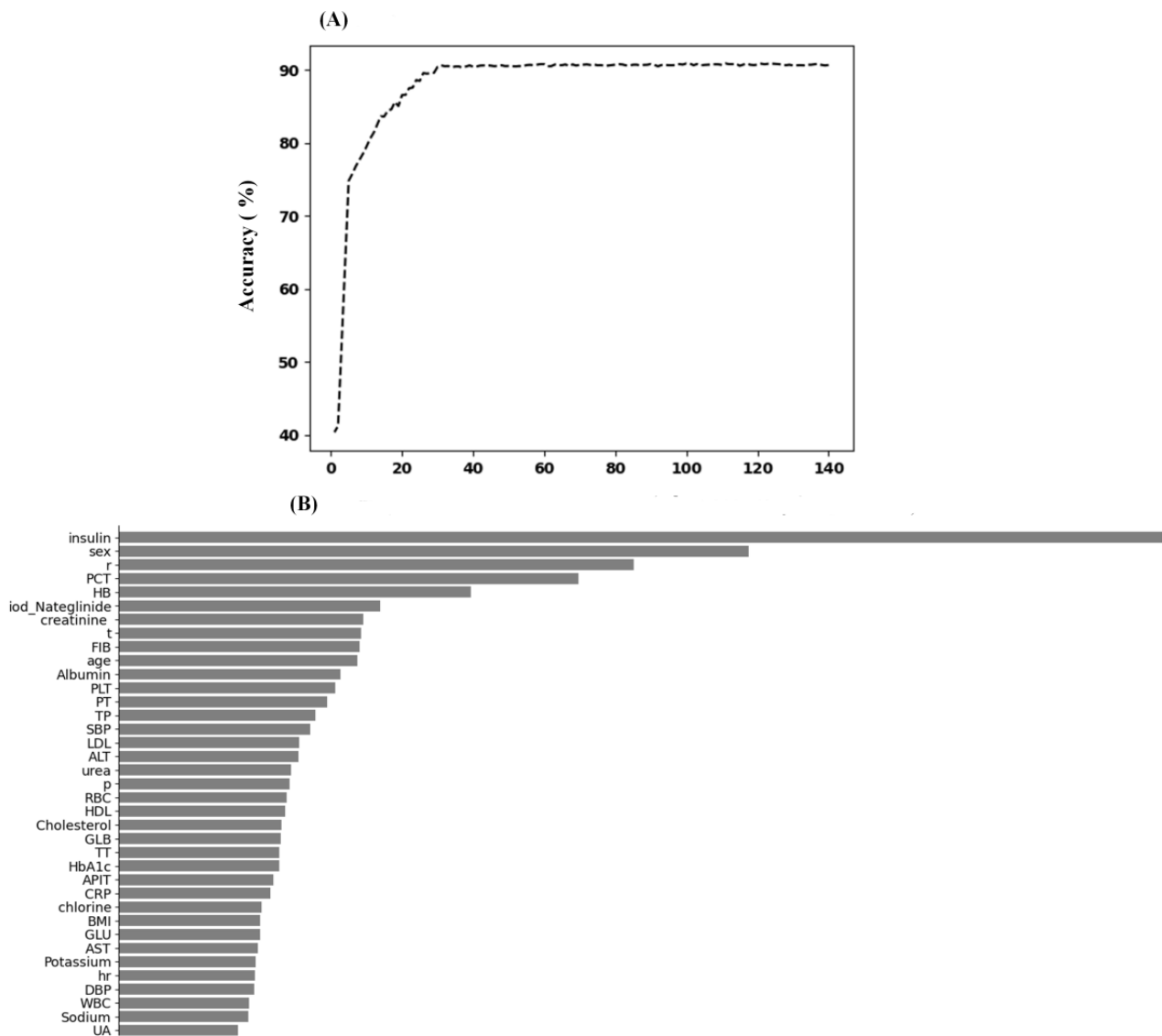
<sup>a</sup>HbA<sub>1c</sub>: glycated hemoglobin.

### Variable Selection

After extracting all the variables, the parameter of feature importance in XGBoost was used to select and filter important

variables [21]. The parameters were set as follows: the number of estimators was 100 and max depth was set to 6. Ultimately, 37 predictive variables and their weights were selected from 176 variables (Figure 2).

**Figure 2.** The weights of variables importance. ALT: alanine aminotransferase; APIT: activated partial thromboplastin time; AST: aspartate aminotransferase; CRP: C-reactive protein; DBP: diastolic blood pressure; FIB: fibrinogen; GLB: globulin; GLU: glucose; HB: hemoglobin; HbA<sub>1c</sub>: glycated hemoglobin; HDL: high-density lipoprotein; hr: heart rate; iod-Nateglinide: iodine urea and Nateglinide; LDL: low-density lipoprotein; p: pulse; PCT: procalcitonin; PLT: blood platelet count; PT: prothrombin time; r: respiratory rate; RBC: red blood cell count; SBP: systolic pressure; t: body temperature; TP: total protein; TT: thrombin time; UA: uric acid; WBC: white blood cell count. (A) the curve between the number of features and accuracy. (B) the weights of variables importance (when accuracy is up to 90%).



### Data Imbalance

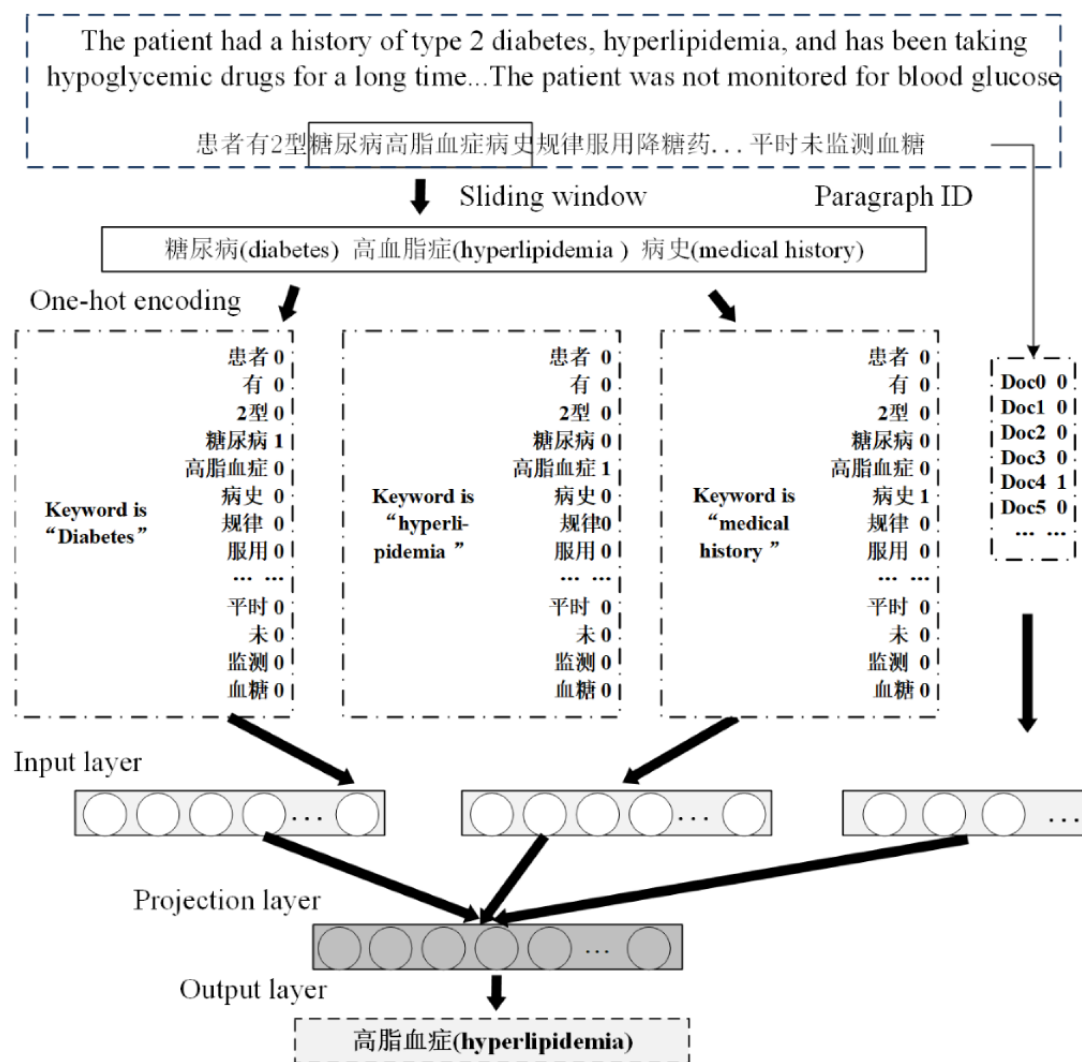
To overcome the data imbalance between the group with hypoglycemia and the normoglycemic control group, we used the Adaptive Synthetic (ADASYN) sampling method [22] to oversample the group with hypoglycemia to generate a portion of data that was comparable to the data from the normoglycemic group. The method for imbalanced learning was used to generate a low sample size to improve class imbalance. We used 5-fold cross-validation and sample balancing with ADASYN for each stratified training set. ADASYN was implemented using Imblearn in Python (Version 0.9.0; imbalanced-learn documentation) [23]. The sampling ratio was set to 1.

### Embedding Models

We used a Python implementation of the Paragraph Vector model available at Gensim [24] and trained 100-dimensional vectors on our corpus. Due to the large computing time of training these corpora and because they are unlabeled corpora, we trained the distributed memory model of paragraph vectors or Paragraph Vector–Distributed Memory (PV-DM) [25] based on textual data from patients with diabetes (Figure 3).

The doc2vec model was used to train and feature map the chief complaints (CCs), history of present illness (HPI), and family history (FH) in the EHR. The results were feature fused into XGBoost model [21] to generate XGBoost1 (XGBoost+CC), XGBoost2 (XGBoost+CC+HPI), and XGBoost3 (XGBoost+CC+HPI+FH).

Figure 3. Training process of Paragraph Vector–Distributed Memory (PV-DM) model.



### Statistical Analysis

As clinical indicators, categorical variables were shown as counts and percentages, and continuous variables were shown as means and SDs. Comparisons between groups were analyzed by a 2-tailed *t* test for continuous variables and chi-square test for categorical variables. All statistical analyses were carried out in R software (version 4.1.2; R Core Team). The statistical significance was considered as  $P < .05$ . The computing environment of this study includes central processing unit i7-7800x; memory 16 GB; operating system Windows 11, build 22598.200; and Python programming language.

## Results

### Participant Characteristics

The cohort included 29,843 patients with type 2 diabetes, of whom 2804 (9.4%) patients developed hypoglycemia. Among

the 29,843 patients, the proportion of female patients in the group with hypoglycemia ( $n=1065$ , 38.0%) was higher than that in the normoglycemia group ( $n=9479$ , 35.1%;  $P=.002$ ). The BMI of patients in the hypoglycemia and normoglycemia groups were 23.6 (SD 5.24) and 24.3 (SD 4.26), respectively. Statistically, the BMI of patients in the normoglycemia group was significantly higher than that of patients in the hypoglycemia group ( $P < .001$ ). The proportion of insulin use in patients in hypoglycemia group ( $n=1575$ , 56.2%) was much higher than that for patients in the normoglycemia group ( $n=7306$ , 27.0%). In addition, the proportion of patients taking sulfonylureas or Nateglinide in the hypoglycemia group ( $n=1382$ , 49.3%) was also higher than that in the normoglycemia group ( $n=9273$ , 34.3%), which was a statistically significant difference ( $P < .001$ ). The demographics of patients in normoglycemia and hypoglycemia groups are shown in Table 2.

**Table 2.** Demographics of patients with diabetes (N=29,843).

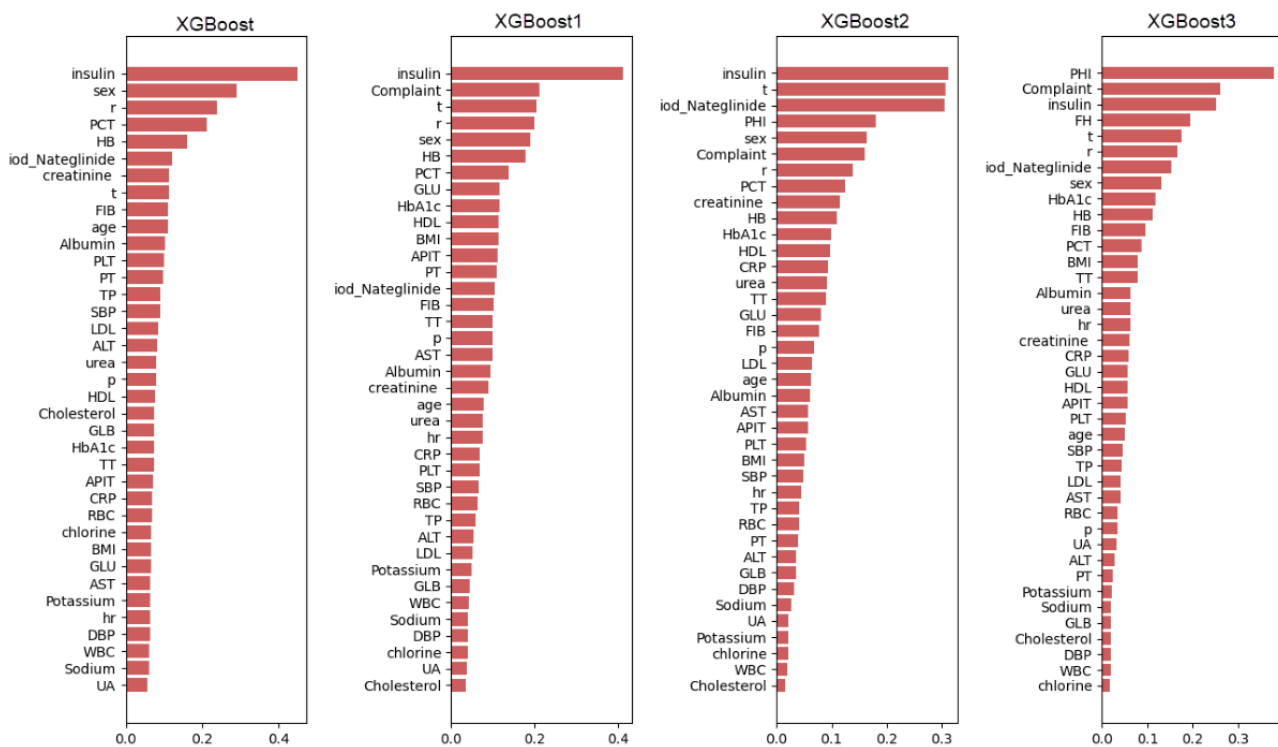
Variables	Normoglycemia (blood glucose>3.9 mmol/L; n=27,039)	Hypoglycemia (blood glucose<3.9 mmol/L; n=2804)	P values
<b>Sex, n (%)</b>			.002
Female	9479 (35.1)	1065 (38)	
Male	17,560 (64.9)	1739 (62)	
Age (years), mean (SD; range)	64.2 (12.3; 18-104)	64.8 (12.6; 19-98)	.03
BMI, mean (SD)	24.3 (4.26)	23.6 (5.24)	<.001
<b>Insulin, n (%)</b>			<.001
No	19,733 (73)	1229 (43.8)	
Yes	7306 (27)	1575 (56.2)	
<b>Sulfonylureas or Nateglinide, n (%)</b>			<.001
No	17,766 (65.7)	1422 (50.7)	
Yes	9273 (34.3)	1382 (49.3)	

**Feature Selection**

We applied XGBoost and its ensemble models for feature selection to discard noninformative features and retain important features (Figure 4). Finally, 37 features were selected from 176

features. In the XGBoost model, insulin was the most important predictor variable among all the predictor variables, followed by sex, respiratory rate, Procalcitonin, and hemoglobin (Figure 4). However, these variables had different weights in XGBoost1, XGBoost2, and XGBoost3 (Figure 4).

**Figure 4.** Weight of the variables in the different models. ALT: alanine aminotransferase; APIT: activated partial thromboplastin time; AST: aspartate aminotransferase; CRP: C-reactive protein; DBP: diastolic blood pressure; FIB: fibrinogen; GLB: globulin; GLU: glucose; HB: hemoglobin; HDL: high-density lipoprotein; hr: heart rate; iod-Nateglinide: Iodine urea and Nateglinide; LDL: low-density lipoprotein; p: pulse; PCT: procalcitonin; PLT: blood platelet count; PT: prothrombin time; r: respiratory rate; RBC: red blood cell count; SBP: systolic pressure; t: body temperature; TP: total protein; TT: thrombin time; UA: uric acid; WBC: white blood cell count.



**Model Performance**

Table 3 shows the results of the 4 machine learning methods after 5-fold cross-validation. The area under the receiver operating characteristic curve (AUC=0.822) and accuracy

(0.934) of the XGBoost3 were higher than all other models. The XGboost3 was superior to other models in terms of model performance, which was evaluated using AUC and decision curve analysis [26] (Figure 5).

The oversample may have an impact on the accuracy of the test set. After completing the model training, we sampled 138 adult patients with type 2 diabetes (hypoglycemia=28, nonhypoglycemia=110) in West China Hospital from January

to March 2022 for validation. The results showed that the prediction accuracy rate reaches 89.86%. The confusion matrix is shown in Figure 6.

**Table 3.** Accuracy and area under the receiver operating characteristic curve (AUC) of different models.

Model	Embedding method	AUC, mean (SD)	Accuracy, mean (SD)	P value
XGBoost	XGBoost	0.718 (0.0014)	0.892 (0.002)	• N/A <sup>a</sup>
XGBoost1	XGBoost+CC <sup>b</sup>	0.785 (0.0012)	0.919 (0.002)	• <.001 vs XG-Boost
XGBoost2	XGBoost+CC+HPI <sup>c</sup>	0.817 (0.0023)	0.928 (0.001)	• <.001 vs XG-Boost • <.001 vs XG-Boost1
XGBoost3	XGBoost+CC+HPI+FH <sup>d</sup>	0.822 (0.0024)	0.934 (0.002)	• <.001 vs XG-Boost • <.001 vs XG-Boost1 • <.001 vs XG-Boost2

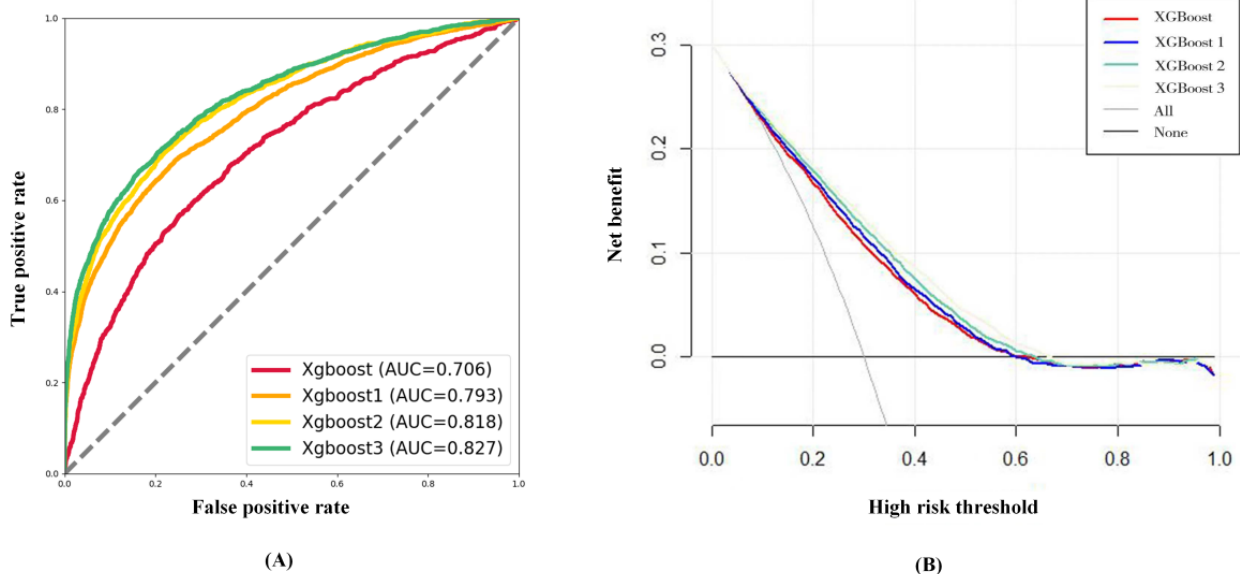
<sup>a</sup>N/A: not applicable.

<sup>b</sup>CC: chief complaints.

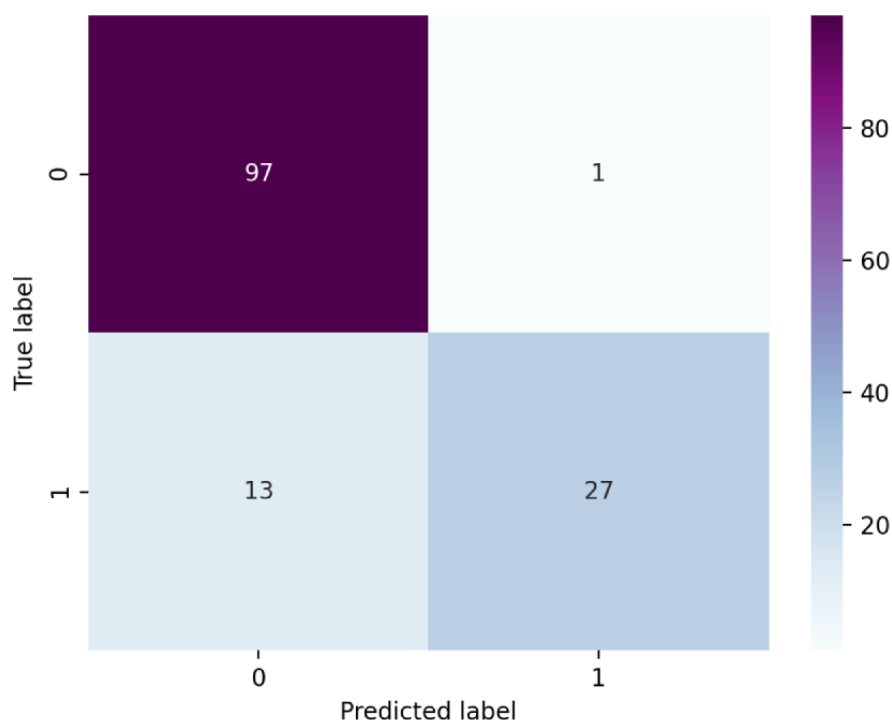
<sup>c</sup>HPI: history of present illness.

<sup>d</sup>FH: family history.

**Figure 5.** Comparison between the change detection algorithm (CDA) and receiver operating characteristic (ROC) curve of different models. (A) The ROC curve of the 4 models. (B) The DCA curve of the 4 models.





**Figure 6.** The confusion matrix of XGBoost3.

## Discussion

### Principal Findings

In this study, we used only some common types of features within 24 hours of patient admission to develop a hypoglycemia prediction model, because the earlier hypoglycemia is predicted or detected, the better we can avoid it. This study found that in women, patients at an older age, patients with low BMI, and those using insulin or various hypoglycemic drugs, the risk of hypoglycemia developing in type 2 diabetes increases. There was a statistical difference ( $P < .001$ ). Some studies have shown that these factors increase significantly in the incidence of hypoglycemia in patients with type 2 diabetes [27-29]. This may be related to the higher risk of sulfonylurea-related hypoglycemia in women compared to men [30]. One possible reason for this is the pharmacokinetics and pharmacodynamics of sulfonylureas in women [31]. In patients with type 2 diabetes, low BMI may be associated with reduced insulin resistance [32]. Patients with obesity can benefit from the same type of antidiabetic drugs that patients with low or normal weight use [33]. This phenomenon is known as the “obesity paradox,” but the mechanism is unknown [34]. This indicates that a standard BMI or overweight are key determinants in reducing the risk of severe hypoglycemia in patients with type 2 diabetes [35].

We developed a hypoglycemia risk prediction model based on XGBoost integrated PV-DM, which can be applied to patients with type 2 diabetes. The result showed that XGBoost3 has the largest AUC and highest accuracy to predict hypoglycemia. There is a significant difference between this model and other models ( $P < .001$ ). Consistent with previous research [36],

combining numerical variables with textual data from EHR can effectively improve the predictive performance of the model. Applying this model to clinical practice could help physicians adjust hypoglycemic drugs based on patient characteristics and hypoglycemia risk factors. This study demonstrates that the inclusion of EHR increases the prognostic accuracy of hypoglycemia in patients with diabetes, providing a more comprehensive and optimized method for predicting hypoglycemic events.

This study also has some limitations. First, the study was carried out in a single institution, and the performance of the model and the distribution of covariates may differ when applied to a sample from a different institution. Second, this study involved Chinese patients. Due to ethnic differences, the results of this study need to be further verified in other ethnic groups.

### Conclusions

We developed a multivariate risk prediction model to predict the occurrence of hypoglycemia in patients with type 2 diabetes. In this prediction model, the PV-DM model can effectively extract the EHR notes and improve the performance of the XGBoost model.

The predictive model can help predict the occurrence of hypoglycemia in patients with type 2 diabetes and provide clinicians with an effective way to prevent hypoglycemia in patients with diabetes. In future research, we will focus on external validation of this model in a larger cohort of patients with type 2 diabetes and explore combining state-of-the-art methods in NLP with deep learning to enhance the model’s predictive power.

## Authors' Contributions

J Liu, HY, and J Li conceptualized the study. J Liu, HY, J Li, SL, and XY carried out the collection and analysis of the literature and data, and drafted the manuscript. Both HY and J Li acted as first authors for this study. All authors reviewed and approved the final version of the manuscript.

## Conflicts of Interest

None declared.

## References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022 Jan;183:109119. [doi: [10.1016/j.diabres.2021.109119](https://doi.org/10.1016/j.diabres.2021.109119)] [Medline: [34879977](https://pubmed.ncbi.nlm.nih.gov/34879977/)]
2. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of type 2 diabetes - global burden of disease and forecasted trends. *J Epidemiol Glob Health* 2020 Mar;10(1):107-111 [FREE Full text] [doi: [10.2991/jegh.k.191028.001](https://doi.org/10.2991/jegh.k.191028.001)] [Medline: [32175717](https://pubmed.ncbi.nlm.nih.gov/32175717/)]
3. Ruan Y, Tan GD, Lumb A, Rea RD. Importance of inpatient hypoglycaemia: impact, prediction and prevention. *Diabet Med* 2019 Apr;36(4):434-443. [doi: [10.1111/dme.13897](https://doi.org/10.1111/dme.13897)] [Medline: [30653706](https://pubmed.ncbi.nlm.nih.gov/30653706/)]
4. Yun J, Ko S. Avoiding or coping with severe hypoglycemia in patients with type 2 diabetes. *Korean J Intern Med* 2015 Jan;30(1):6-16 [FREE Full text] [doi: [10.3904/kjim.2015.30.1.6](https://doi.org/10.3904/kjim.2015.30.1.6)] [Medline: [25589828](https://pubmed.ncbi.nlm.nih.gov/25589828/)]
5. Pathak RD, Schroeder EB, Seaquist ER, Zeng C, Lafata JE, Thomas A, SUPREME-DM Study Group. Severe hypoglycemia requiring medical intervention in a large cohort of adults with diabetes receiving care in U.S. integrated health care delivery systems: 2005-2011. *Diabetes Care* 2016 Mar;39(3):363-370 [FREE Full text] [doi: [10.2337/dc15-0858](https://doi.org/10.2337/dc15-0858)] [Medline: [26681726](https://pubmed.ncbi.nlm.nih.gov/26681726/)]
6. Huang ES, Laiteerapong N, Liu JY, John PM, Moffet HH, Karter AJ. Rates of complications and mortality in older patients with diabetes mellitus: the diabetes and aging study. *JAMA Intern Med* 2014 Feb 01;174(2):251-258 [FREE Full text] [doi: [10.1001/jamainternmed.2013.12956](https://doi.org/10.1001/jamainternmed.2013.12956)] [Medline: [24322595](https://pubmed.ncbi.nlm.nih.gov/24322595/)]
7. Ligthelm RJ, Kaiser M, Vora J, Yale J. Insulin use in elderly adults: risk of hypoglycemia and strategies for care. *J Am Geriatr Soc* 2012 Aug;60(8):1564-1570. [doi: [10.1111/j.1532-5415.2012.04055.x](https://doi.org/10.1111/j.1532-5415.2012.04055.x)] [Medline: [22881394](https://pubmed.ncbi.nlm.nih.gov/22881394/)]
8. American Diabetes Association. 12. Older Adults: standards of Medical Care in Diabetes-2020. *Diabetes Care* 2020 Jan;43(Suppl 1):S152-S162. [doi: [10.2337/dc20-S012](https://doi.org/10.2337/dc20-S012)] [Medline: [31862755](https://pubmed.ncbi.nlm.nih.gov/31862755/)]
9. Hulkower RD, Pollack RM, Zonszein J. Understanding hypoglycemia in hospitalized patients. *Diabetes Manag (Lond)* 2014 Mar;4(2):165-176 [FREE Full text] [doi: [10.2217/DMT.13.73](https://doi.org/10.2217/DMT.13.73)] [Medline: [25197322](https://pubmed.ncbi.nlm.nih.gov/25197322/)]
10. Mathioudakis NN, Abusamaan MS, Shakarchi AF, Sokolinsky S, Fayzullin S, McGready J, et al. Development and Validation of a Machine Learning Model to Predict Near-Term Risk of Iatrogenic Hypoglycemia in Hospitalized Patients. *JAMA Netw Open* 2021 Jan 04;4(1):e2030913 [FREE Full text] [doi: [10.1001/jamanetworkopen.2020.30913](https://doi.org/10.1001/jamanetworkopen.2020.30913)] [Medline: [33416883](https://pubmed.ncbi.nlm.nih.gov/33416883/)]
11. Schroeder EB, Xu S, Goodrich GK, Nichols GA, O'Connor PJ, Steiner JF. Predicting the 6-month risk of severe hypoglycemia among adults with diabetes: development and external validation of a prediction model. *J Diabetes Complications* 2017 Jul;31(7):1158-1163 [FREE Full text] [doi: [10.1016/j.jdiacomp.2017.04.004](https://doi.org/10.1016/j.jdiacomp.2017.04.004)] [Medline: [28462891](https://pubmed.ncbi.nlm.nih.gov/28462891/)]
12. Karter AJ, Warton EM, Lipska KJ, Ralston JD, Moffet HH, Jackson GG, et al. Development and validation of a tool to identify patients with type 2 diabetes at high risk of hypoglycemia-related emergency department or hospital use. *JAMA Intern Med* 2017 Oct 01;177(10):1461-1470 [FREE Full text] [doi: [10.1001/jamainternmed.2017.3844](https://doi.org/10.1001/jamainternmed.2017.3844)] [Medline: [28828479](https://pubmed.ncbi.nlm.nih.gov/28828479/)]
13. Plis K, Bunescu R, Marling C, Shubrook J, Schwartz F. A machine learning approach to predicting blood glucose levels for diabetes management. *AAAI Workshop: Modern Artificial Intelligence for Health Analytics* 2014 Jun 18 [FREE Full text]
14. Jin Y, Li F, Vimalananda VG, Yu H. Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: empirical study. *JMIR Med Inform* 2019 Nov 08;7(4):e14340 [FREE Full text] [doi: [10.2196/14340](https://doi.org/10.2196/14340)] [Medline: [31702562](https://pubmed.ncbi.nlm.nih.gov/31702562/)]
15. West China Hospital of Sichuan University. URL: <http://www.wchscu.cn/Home.html> [accessed 2022-03-25]
16. Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *PLoS One* 2021;16(2):e0246306 [FREE Full text] [doi: [10.1371/journal.pone.0246306](https://doi.org/10.1371/journal.pone.0246306)] [Medline: [33539390](https://pubmed.ncbi.nlm.nih.gov/33539390/)]
17. Silbert R, Salcido-Montenegro A, Rodriguez-Gutierrez R, Katabi A, McCoy RG. Hypoglycemia among patients with type 2 diabetes: epidemiology, risk factors, and prevention strategies. *Curr Diab Rep* 2018 Jun 21;18(8):53 [FREE Full text] [doi: [10.1007/s11892-018-1018-0](https://doi.org/10.1007/s11892-018-1018-0)] [Medline: [29931579](https://pubmed.ncbi.nlm.nih.gov/29931579/)]
18. 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 Feb 12;4(1):24 [FREE Full text] [doi: [10.1038/s41746-021-00394-8](https://doi.org/10.1038/s41746-021-00394-8)] [Medline: [33580109](https://pubmed.ncbi.nlm.nih.gov/33580109/)]
19. Arvind V, Kim JS, Oermann EK, Kaji D, Cho SK. Predicting surgical complications in adult patients undergoing anterior cervical discectomy and fusion using machine learning. *Neurospine* 2018 Dec;15(4):329-337 [FREE Full text] [doi: [10.14245/ns.1836248.124](https://doi.org/10.14245/ns.1836248.124)] [Medline: [30554505](https://pubmed.ncbi.nlm.nih.gov/30554505/)]

20. Li K, Shi Q, Liu S, Xie Y, Liu J. Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. *Medicine (Baltimore)* 2021 May 14;100(19):e25813 [FREE Full text] [doi: [10.1097/MD.00000000000025813](https://doi.org/10.1097/MD.00000000000025813)] [Medline: [34106618](https://pubmed.ncbi.nlm.nih.gov/34106618/)]
21. 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 Presented at: KDD '16; August 13-17; San Francisco, CA p. 785-794.
22. Haibo H, Yang B, Garcia E. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008 Presented at: IEEE; 1-8 June; Hong Kong p. 1322-1328.
23. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *JMLR* 2017 Jan;18(1):559-563 [FREE Full text]
24. Řehůřek R, Sojka P. Gensim-statistical semantics in Python. 2011 Presented at: EuroScipy; 25-28 August; Paris URL: <https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf>
25. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *PMLR*. 2014 Presented at: 31st International Conference on Machine Learning; June 21-26; Beijing, China p. 1188-1196 URL: <http://proceedings.mlr.press/v32/le14.pdf>
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26(6):565-574 [FREE Full text] [doi: [10.1177/0272989X06295361](https://doi.org/10.1177/0272989X06295361)] [Medline: [17099194](https://pubmed.ncbi.nlm.nih.gov/17099194/)]
27. Han K, Yun J, Park Y, Ahn Y, Cho J, Cha S, et al. Development and validation of a risk prediction model for severe hypoglycemia in adult patients with type 2 diabetes: a nationwide population-based cohort study. *Clin Epidemiol* 2018;10:1545-1559 [FREE Full text] [doi: [10.2147/CLEP.S169835](https://doi.org/10.2147/CLEP.S169835)] [Medline: [30425585](https://pubmed.ncbi.nlm.nih.gov/30425585/)]
28. Lee SE, Kim K, Son KJ, Song SO, Park KH, Park SH, et al. Trends and risk factors in severe hypoglycemia among individuals with type 2 diabetes in Korea. *Diabetes Res Clin Pract* 2021 Aug;178:108946 [FREE Full text] [doi: [10.1016/j.diabres.2021.108946](https://doi.org/10.1016/j.diabres.2021.108946)] [Medline: [34252506](https://pubmed.ncbi.nlm.nih.gov/34252506/)]
29. Gonzalez C, Monti C, Pinzon A, Monsanto H, Ejzykowicz F, Argentinean Recap Group. Prevalence of hypoglycemia among a sample of sulfonylurea-treated patients with type 2 diabetes mellitus in Argentina: The real-life effectiveness and care patterns of diabetes management (RECAP-DM) study. *Endocrinol Diabetes Nutr (Engl Ed)* 2018 Dec;65(10):592-602. [doi: [10.1016/j.endinu.2018.05.014](https://doi.org/10.1016/j.endinu.2018.05.014)] [Medline: [30076124](https://pubmed.ncbi.nlm.nih.gov/30076124/)]
30. Kajiwara A, Kita A, Saruwatari J, Oniki K, Morita K, Yamamura M, et al. Higher risk of sulfonylurea-associated hypoglycemic symptoms in women with type 2 diabetes mellitus. *Clin Drug Investig* 2015 Sep;35(9):593-600. [doi: [10.1007/s40261-015-0314-6](https://doi.org/10.1007/s40261-015-0314-6)] [Medline: [26293520](https://pubmed.ncbi.nlm.nih.gov/26293520/)]
31. Soldin OP, Mattison DR. Sex differences in pharmacokinetics and pharmacodynamics. *Clin Pharmacokinet* 2009;48(3):143-157 [FREE Full text] [doi: [10.2165/00003088-200948030-00001](https://doi.org/10.2165/00003088-200948030-00001)] [Medline: [19385708](https://pubmed.ncbi.nlm.nih.gov/19385708/)]
32. Tsai T, Lee C, Cheng B, Kung C, Chen F, Shen F, et al. Body Mass Index-mortality relationship in severe hypoglycemic patients with type 2 diabetes. *Am J Med Sci* 2015 Mar;349(3):192-198 [FREE Full text] [doi: [10.1097/MAJ.0000000000000382](https://doi.org/10.1097/MAJ.0000000000000382)] [Medline: [25526505](https://pubmed.ncbi.nlm.nih.gov/25526505/)]
33. Cai X, Yang W, Gao X, Zhou L, Han X, Ji L. Baseline Body Mass Index and the efficacy of hypoglycemic treatment in type 2 diabetes: a meta-analysis. *PLoS One* 2016;11(12):e0166625 [FREE Full text] [doi: [10.1371/journal.pone.0166625](https://doi.org/10.1371/journal.pone.0166625)] [Medline: [27935975](https://pubmed.ncbi.nlm.nih.gov/27935975/)]
34. Gravina G, Ferrari F, Nebbiai G. The obesity paradox and diabetes. *Eat Weight Disord* 2021 May;26(4):1057-1068. [doi: [10.1007/s40519-020-01015-1](https://doi.org/10.1007/s40519-020-01015-1)] [Medline: [32954485](https://pubmed.ncbi.nlm.nih.gov/32954485/)]
35. Plečko D, Bennett N, Mårtensson J, Bellomo R. The obesity paradox and hypoglycemia in critically ill patients. *Crit Care* 2021 Nov 01;25(1):378 [FREE Full text] [doi: [10.1186/s13054-021-03795-z](https://doi.org/10.1186/s13054-021-03795-z)] [Medline: [34724956](https://pubmed.ncbi.nlm.nih.gov/34724956/)]
36. Arnaud E, Elbattah M, Gignon M, Dequen G. Deep learning to predict hospitalization at triage: integration of structured data and unstructured text. In: *IEEE*. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13; Atlanta, GA p. 4836-4841.

## Abbreviations

- ADASYN:** adaptive synthetic
- AUC:** area under the receiver operating characteristic curve
- CC:** chief complaints
- DCA:** decision curve analysis
- EHR:** electronic health record
- FH:** family history
- HPI:** history of present illness
- ICD-10:** International Classification of Diseases, 10th revision
- NLP:** natural language processing
- PV-DM:** Paragraph Vector–Distributed Memory

*Edited by C Lovis; submitted 31.01.22; peer-reviewed by M Elbattah, A Staffini, E Sükei; comments to author 27.04.22; revised version received 08.05.22; accepted 31.05.22; published 16.06.22*

*Please cite as:*

*Yang H, Li J, Liu S, Yang X, Liu J*

*Predicting Risk of Hypoglycemia in Patients With Type 2 Diabetes by Electronic Health Record–Based Machine Learning: Development and Validation*

*JMIR Med Inform 2022;10(6):e36958*

URL: <https://medinform.jmir.org/2022/6/e36958>

doi: [10.2196/36958](https://doi.org/10.2196/36958)

PMID:

©Hao Yang, Jiayi Li, Siru Liu, Xiaoling Yang, Jialin Liu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 16.06.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.