Diabetes Technology and Therapeutics

DOI: 10.1089/dia.2022.0210

Machine Learning for Predicting the Risk of Transition from Prediabetes to Diabetes

Authors

Thomas Zueger1,2* MD, Simon Schallmoser3* MSc, Mathias Kraus4 PhD, Maytal Saar-Tsechansky5 PhD, Stefan Feuerriegel3 PhD, Christoph Stettler1 MD

1 Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital, Bern, University Hospital, University of Bern, Bern, Switzerland

2 Department of Endocrinology and Metabolic Diseases, Kantonsspital Olten, Olten, Switzerland

3 Ludwig-Maximilian University, Munich, Germany

4 Friedrich-Alexander University, Erlangen-Nuremberg, Germany

5 The University of Texas at Austin, Austin, Texas, USA

* joint first authorship

Running title: Diabetes prediction using machine learning

Corresponding author

Thomas Zueger

Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital,

Bern University Hospital, University of Bern, Bern, Switzerland

Freiburgstrasse 15, 3010, Bern, Switzerland

E -Mail: thomas.zueger@insel.ch

Keywords: prediabetes, diabetes, machine learning, decision support

Abstract

Traditional risk scores for the prediction of type 2 diabetes (T2D) are typically designed for a general population and, thus, may underperform for people with prediabetes. Here, we developed machine learning (ML) models predicting the risk of T2D that are specifically tailored to people with prediabetes. We analyzed data of 13,943 individuals with prediabetes, and built a ML model to predict the risk of transition from prediabetes to T2D, integrating information about demographics, biomarkers, medications, and comorbidities defined by disease codes. Additionally, we developed a simplified ML model with only eight predictors, which can be easily integrated into clinical practice. For a forecast horizon of five years, the area under the receiver operating characteristic curve (AUROC) was 0.753 for our full ML model (79 predictors) and 0.752 for the simplified model. Our ML models allow for an early identification of people with prediabetes who are at risk of developing T2D.

Main Text

Introduction

Traditional risk scores for the prediction of type 2 diabetes (T2D), such as the American Diabetes Association (ADA) risk test (https://www.diabetes.org/risk-test), the Framingham Diabetes Risk Scoring Model (FDRSM)1, and other machine learning (ML) based risk scores2–6, are typically designed for a general population. Applying these on individuals who already were diagnosed with prediabetes may impede the prediction for this specific population. However, accurate predictions for these individuals are essential since the annual progression rate from prediabetes to T2D is estimated to be around 5–10%.7 Further, early detection of individuals at risk is crucial to allocate effective strategies that prevent disease progression, and, hence, reduce the onset of diabetes complications and associated public health burdens.

To the best of our knowledge, there exists only one prior work predicting the risk of transition from prediabetes to T2D using ML.8 Therein, the authors focus on a forecast horizon of 1 year and make use of historic patient data over several years. In contrast, information on comorbidities is missing. Hence, different from Cahn et al.,8 we aimed to develop ML models for forecast horizons up to 5 years, where we make use of complete electronic health records (EHRs), which include data on demographics, biomarkers, medications, and comorbidities defined by disease codes. Further, for our ML models, only one year of historic patient data is required. To facilitate clinical implementation, we additionally developed a simplified ML model with a reduced number of predictors that are easily accessible.

Materials and Methods

Data

In this retrospective analysis, we used anonymized EHRs from an Israeli health provider from 2003 until 2013. The EHRs contain information about demographics, biomarkers, and medications. Additionally, disease codes were recorded using the 9th edition of the International Classification of Diseases (ICD-9). The demographics data include sex, age,

body mass index (BMI), and blood pressure. From all biomarkers, we included the 50 most frequently recorded ones based on the complete data time frame. Details on data preprocessing are provided in Supplement 1. We grouped the 50 most often prescribed medications into 20 classes (Supplement 2). The class "antidiabetic medication" was not included as a predictor since it defines the outcome. Additionally, we included the 10 most frequently recorded comorbidities defined by ICD-9 disease codes. In total, this resulted in 84 predictors listed in Supplement 3 (Table S1), which were used to train our ML models. Table S1 further reports, which of the 84 predictors remained in the final ML models depending on the forecast horizon.

Definitions for prediabetes and diabetes were based on laboratory measurements of HbA1c, recorded ICD-9 codes, and medications. Onset of prediabetes was defined by either a single measurement of HbA1c between 5.7% – 6.4% (39 mmol/mol – 47 mmol/mol) or the record of an ICD-9 code corresponding to prediabetes (790.21, 790.22, or 790.29). T2D was defined by either two measurements of HbA1c ≥ 6.5% (48 mmol/mol), where the onset was set to the year of the first measurement, the record of an ICD-9 code corresponding to diabetes (249.x or 250.x), or if any prescription of antidiabetic medication and/or device for self-measurement of blood glucose (SMBG) was recorded (the distribution of diagnosis criteria is given in Supplement 4). The antidiabetic medications and SMBGs are listed in Supplement 5 and their use among individuals with diabetes in Supplement 6 (Figure S1).

Patient Selection

We selected individuals who were considered as having prediabetes in 2008 and for whom data until 2013 was available. This ensured that each selected individual can be included for all forecast horizons. Further, we only considered people with known age and sex.

Model Training and Validation

For our ML model, we chose a gradient boosting model implemented in the CatBoost package (version 1.0.5)9 using Python 3.6. Gradient boosting is a machine learning technique in which a sequence of weak learners (here: decision trees) are sequentially optimized to minimize the prediction errors of the previous weak learners. The final

gradient boosting model consists of an ensemble of weak learners, which are highly effective in modeling complex relationships that generalize well to unseen observations.9

In addition to the full ML model, which is trained on all 84 predictors, we built a second, simplified ML model containing only a subset of broadly available predictors, which are well known risk factors of T2D: age, BMI, glucose, HbA1c, triglycerides, high-density lipoprotein (HDL), alanine transaminase (ALT), and serum creatinine measurements. Such a simplified ML model has the advantage that it can be more easily integrated into clinical practice.

We used nested cross-validation, where four inner folds were used to tune the ML models (see Supplement 1) and five outer folds to measure the corresponding out-of-sample performance. This ensures that our ML models generalize well to unseen individuals. We trained separate models for forecast horizons of 1 to 5 years. We compared our ML models to a logistic regression with L2 regularization (implemented using scikit-learn10) including (1) all 84 predictors and (2) only the eight predictors from the simplified ML model. Additionally, we compared our ML models to the FDRSM, which estimates the individual risk of developing T2D within a 7-year forecast horizon using six predictors, namely glucose, BMI, HDL, parental history of T2D, triglyceride level, and blood pressure.1 We calibrated the FDRSM to our data (i.e., people with prediabetes) by training a logistic regression using the aforementioned six predictors for all five forecast horizons. Model performance was primarily assessed based on the area under the receiver operating characteristic curve (AUROC). Additional performance metrics are reported in Supplement 7 (Table S2).

Model Explainability

To identify the most important predictors of our full ML model, we calculated SHapley Additive exPlanations (SHAP) values. SHAP values are a unified approach for estimating the individual contribution of a predictor to the overall model output and, hence, provide a ranking of the most important predictors.11 Additionally, they inform whether larger (smaller) values of a predictor are attributed with an increased risk of transition from prediabetes to T2D.

We report the MI-CLAIM checklist,12 which was developed to improve transparent reporting of ML in medicine, in Supplement 8 (Table S3).

Results

Our final sample consists of 13,943 individuals with prediabetes out of which 2,102 (15.1%) transitioned to T2D diabetes within five years. Patient characteristics are summarized in Table 1. The table shows that more female individuals transition to T2D and that well known risk factors for T2D and comorbidities such as BMI, dyslipidemia, and hypertension are increased in individuals with T2D.

Model Performance

Our ML model can identify people who transition from prediabetes to T2D with AUROCs of 0.773 (1 year) and 0.753 (5 years). From the 84 predictors which were inserted into the ML model, 73–80 remained in the final ML model depending on the forecast horizon (the predictors after feature selection through the ML model can be found in Supplement 3 Table S1). The AUROCs for the simplified ML model, encompassing only 8 predictors, are 0.779 (1 year) and 0.752 (5 years).

Figure 1A shows the performance of our ML models in comparison to the logistic regression, the simplified logistic regression, and the calibrated FDRSM. To evaluate the performance differences between models, we conducted Mann-Whitney U tests.13 Both ML models significantly outperform the calibrated FDRSM ($p<0.001$ for both models across all forecast horizons). Additionally, both ML models demonstrate superior predictive power in comparison to the logistic regression ($p<0.05$ for both models across all forecast horizons except for the simplified ML model and a forecast horizon of 5 years where $p=0.08$). Further, both ML models outperform the simplified logistic regression ($p<0.05$ across all forecast horizons except for the full ML model and a forecast horizon of 1 year where $p=0.48$). The difference in prediction performance between the full and the simplified ML model was not statistically significant at the 5% significance level.

Model Explainability

Figure 1B lists the ten most important predictors according to their SHAP values for all five forecast horizons. Glucose, HbA1c, age, and triglycerides are important predictors for all forecast horizons. Further important predictors were BMI, serum creatinine, and ALT.

We performed several additional analyses to check the robustness of our ML model (see Supplement 9).

Discussion

We developed a ML model that predicts the risk for transition from prediabetes to T2D over forecast horizons of 1 to 5 years. Its AUROCs ranged from 0.753 to 0.780, thereby outperforming the logistic regression ($p<0.05$), the simplified logistic regression ($p<0.05$ except for a forecast horizon of 1 year where $p=0.48$) and the FDRSM ($p<0.001$), a well-established diabetes risk score that we have calibrated to our data. Prediction performance was largely robust, and, moreover, inter-year variability was small.

We calculated SHAP values to identify the most important predictors of our ML model. The two most important predictors across all forecast horizons were glucose and HbA1c measurements. This can be expected, since these two measures are disease defining markers,14 which are known to be predictive.15 Age and BMI are well known risk factors for T2D,1 and both were accordingly among the most important predictors in our ML model. The biomarkers triglyceride and serum creatinine were also identified as important predictors. The former has been already included in the FDRSM as a risk factor for T2D, and several studies indicate an increased diabetes risk for individuals with higher triglyceride levels.1,15 Previous studies have also shown that low serum creatinine levels are a risk factor for T2D.16,17 Further important predictors are discussed in Supplement 10.

ML models with a substantial number of predictors can be incorporated into clinical decision support systems, where data can be directly retrieved from EHRs. However, we are aware that a model with 84 predictors might be impractical in a clinical setting, where predictors must be manually inserted into a software. Furthermore, some of these predictors may not be broadly available or vary across health providers (e.g., disease

codes). Hence, we developed a second, simplified ML model with only eight easily accessible predictors, namely age, BMI, glucose, HbA1c, triglycerides, HDL, ALT, and serum creatinine measurements. These were also identified by the SHAP values as important predictors in the full ML model (e.g., age, glucose, HbA1c were among the most important predictors across all forecast horizons, HDL on the other hand only for a forecast horizon of 1 year). Despite the parsimonious structure, the simplified ML model still works well and outperforms the calibrated FDRSM ($p<0.001$). Thereby, it might provide a valuable alternative in clinical practice.

Our ML models have several benefits. First, they were specifically designed for people with prediabetes. This is important as traditional risk scores and ML models are typically designed for a general population,1–6 for which the prediction performance may not generalize to individuals with prediabetes. Second, it is possible to model non-linear relationships and interactions between predictors, thereby increasing the accuracy in identifying those at highest risk. Current guidelines recommend yearly diabetes screening for people with prediabetes.18 Our ML approach may allow for a more differentiated approach since it enables a personalized risk stratification over a five-year horizon and thus may allow for an individualized screening procedure. Furthermore, by maximizing model specificity, resources could be specifically allocated to individuals at highest risk for the transition from prediabetes to diabetes. This would ease the burden on both the healthcare system and patients. Overall, a strength of our ML models is that they are specifically designed for people with prediabetes and that they include a larger number of predictors, which adds to an improved prediction performance.

When we compare our ML models to those from Cahn et al.,8 AUROCs do not differ substantially (0.779 [range over 10 seeds: (0.777, 0.783)] for our simplified model vs. 0.782 [95% CI: (0.778, 0.788)] for the model from Cahn et al.8 using only 1 year of historic patient data) on the 1-year forecast horizon. Unfortunately, Cahn et al.8 do not provide the performance of their model using 1 year of historic patient data for forecast horizons larger than 1 year, making it impossible to directly compare our models to theirs on these forecast horizons.

This study has limitations. First, it only covers an Israeli population, and, hence, its generalizability to other populations may be limited. However, our ML model could be extended to other populations in the future. Second, in this retrospective analysis, diagnosis of prediabetes and diabetes was solely based on HbA1c measurements, ICD-9 codes, and prescribed antidiabetic medications, since we did not have information on time-point of glucose sampling (i.e., fasting glucose). However, this can also be beneficial as it ensures direct applicability of our ML models to EHRs. Third, the missing time-point of glucose sampling might limit its use as a predictor. However, the use of random glucose as a predictor also offers advantages, since random glucose is frequently measured in clinical practice, whereas assessment of fasting glucose is more cumbersome or not available. This makes our ML models more broadly applicable. Fourth, the comparison of our ML models to the FDRSM might be unfair since the latter was developed for a forecast horizon of seven years. Even though we calibrated it to our data (i.e., people with prediabetes) and our forecast horizons, the predictors included in the FDRSM might have been specifically chosen for a forecast horizon of seven years. However, the results of our robustness check, where we trained our ML models and the FDRSM on a forecast horizon of seven years, revealed that our ML models are still significantly superior (AUROCs: 0.754 [full] / 0.751 [simplified] vs. 0.708 [FDRSM]; $p<0.001$).

Conclusion

Our ML models allow for an early identification of individuals with prediabetes who are at risk of developing T2D. This may help to allocate effective treatment to those at highest risk, thereby preventing disease progression, and, translating into reduced diabetes complications and the associated burden on individuals with prediabetes, the health care system and society.

Author Contributions

T.Z. and S.S. share first authorship. T.Z., S.S., M.K., S.F., and M.S.-T. designed the study. T.Z., S.S., and S.F. wrote the manuscript. S.S. and M.K. analysed the data. M.S.-T. and C.S. critically reviewed the manuscript and contributed to the interpretation of the results. S.F. is the guarantor of this work and, as such, had full access to all the data in the study and

takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final draft of the manuscript.

Authors Disclosure Statement

The authors have nothing to disclose. No competing financial interests exist.

Funding Information

This was an investigator-initiated study (no funding).

References

1. Wilson PWF, Meigs JB, Sullivan L, et al.: Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 2007; 167(10):1068–1074.

2. Yu W, Liu T, Valdez R, et al.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010; 10:16.

3. Wu H, Yang S, Huang Z, et al.: Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 2018; 10:100–107.

4. Zou Q, Qu K, Luo Y, et al.: Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet 2018; 9:515.

5. Anderson JP, Parikh JR, Shenfeld DK, et al.: Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. J Diabetes Sci Technol 2015; 10(1):6–18.

6. Kopitar L, Kocbek P, Cilar L, et al.: Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 2020; 10(1):11981.

7. Tabák AG, Herder C, Rathmann W, et al.: Prediabetes: a high-risk state for diabetes development. Lancet 2012; 379(9833):2279–2290.

8. Cahn A, Shoshan A, Sagiv T, et al.: Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. Diabetes/Metab Res Rev 2020; 36(2):e3252.

9. Prokhorenkova L, Gusev G, Vorobev A, et al.: CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018; 31.

10. Pedregosa F, Varoquaux G, Gramfort A, et al.: Scikit-learn: Machine learning in Python. J Mach Learn Res 2011(12):2825-2830.

11. Lundberg SM, Lee SI: A unified approach to interpreting model predictions. Adv Neural Inf Process Syst; 2017:4768–4777.

12. Norgeot B, Quer G, Beaulieu-Jones BK, et al.: Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020; 26(9):1320–1324.

13. Mann HB, Whitney DR: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat 1947; 18(1):50–60.

14. American Diabetes Association: 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021 2021; 44(Suppl 1):S15-S33.

15. Abbasi A, Sahlqvist A-S, Lotta L, et al.: A Systematic Review of Biomarkers and Risk of Incident Type 2 Diabetes: An Overview of Epidemiological, Prediction and Aetiological Research Literature. PLoS One 2016; 11(10):e0163721.

16. Hjelmesæth J, Røislien J, Nordstrand N, et al.: Low serum creatinine is associated with type 2 diabetes in morbidly obese women and men: a cross-sectional study. BMC Endocr Disord 2010; 10:6.

17. Song DK, Hong YS, Sung Y-A, Lee H: Association of serum creatinine levels and risk of type 2 diabetes mellitus in Korea: a case control study. BMC Endocr Disord 2022; 22(1):4.

18. Draznin B, Aroda VR, Bakris G, et al.: 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2022 2022; 45(Suppl 1):S17-S38.

Diabetes Technology and Therapeutics

Machine Learning for Predicting the Risk of Transition from Prediabetes to Diabetes (DOI: 10.1089/dia.2022.0210)

Table 1: Patient characteristics.

	Characteristics of all people with prediabetes at baseline	Baseline characteristics of people who transition to T2D within T years					Characteristics at diabetes diagnosis depending on diagnosis criterion		
							ICD-9 codes or anti-diabetic medication	HbA1c	All
Forecast horizon [years]		T=1	T=2	T=3	T=4	T=5			
Number of samples	13943	534	947	1377	1805	2102	1768	573	2102
Incidence [%]	-	3.8	6.8	9.9	12.9	15.1	12.7	4.1	15.1
Demographic data									
Sex									
Male	6523 (46.8)	223 (41.8)	400 (42.2)	587 (42.6)	761 (42.2)	898 (42.7)	743 (42.4)	264 (46.1)	898 (42.7)

Diabetes Technology and Therapeutics
Machine Learning for Predicting the Risk of Transition from Prediabetes to Diabetes (DOI: 10.1089/dia.2022.0210)

Female	7420 (53.2)	311 (58.2)	547 (57.8)	790 (57.4)	1044 (57.8)	1204 (57.3)	1025 (58.5)	309 (53.9)	1204 (57.3)
Age [years]	51.2 (8.7)	53.1 (9.1)	53.5 (8.8)	53.5 (8.6)	53.6 (8.6)	53.6 (8.5)	56.3 (8.8)	57.5 (7.6)	56.5 (8.6)
BMI [kg/m2]	30.0 (5.7)	31.4 (5.8)	31.3 (5.9)	31.4 (6.0)	31.3 (5.8)	31.2 (5.7)	32.1 (6.2)	32.9 (6.5)	32.1 (6.2)
Biomarkers									
Random glucose [mg/dL]	97.5 (9.8)	104.9 (9.9)	104.6 (9.4)	103.7 (9.7)	103.3 (9.7)	103.0 (9.7)	108.3 (11.4)	110.5 (11.2)	108.3 (11.2)
HbA1c [%]	5.84 (0.26)	6.06 (0.26)	6.06 (0.25)	6.04 (0.24)	6.02 (0.25)	6.0 (0.25)	6.17 (0.38)	6.6 (0.24)	6.24 (0.39)
HbA1c [mmol/mol]	40.4 (2.7)	42.5 (2.6)	42.6 (2.5)	42.4 (2.5)	42.2 (2.5)	42.0 (2.5)	43.8 (4.2)	48.8 (2.7)	44.7 (4.3)
Comorbidities									

Dyslipidemia	9096 (65.2)	383 (71.7)	689 (72.8)	991 (72.0)	1307 (72.4)	1517 (72.2)	1454 (83.0)	467 (81.5)	1712 (81.4)
Hypertension	5337 (38.3)	256 (47.9)	465 (49.1)	689 (50.0)	902 (50.0)	1037 (49.3)	994 (56.8)	362 (63.2)	1201 (57.1)

Age, BMI, and biomarkers are reported as mean (standard deviation). Sex and comorbidities are reported as counts (percentage).

Diabetes Technology and Therapeutics
Machine Learning for Predicting the Risk of Transition from Prediabetes to Diabetes (DOI: 10.1089/dia.2022.0210)

Figure Legend

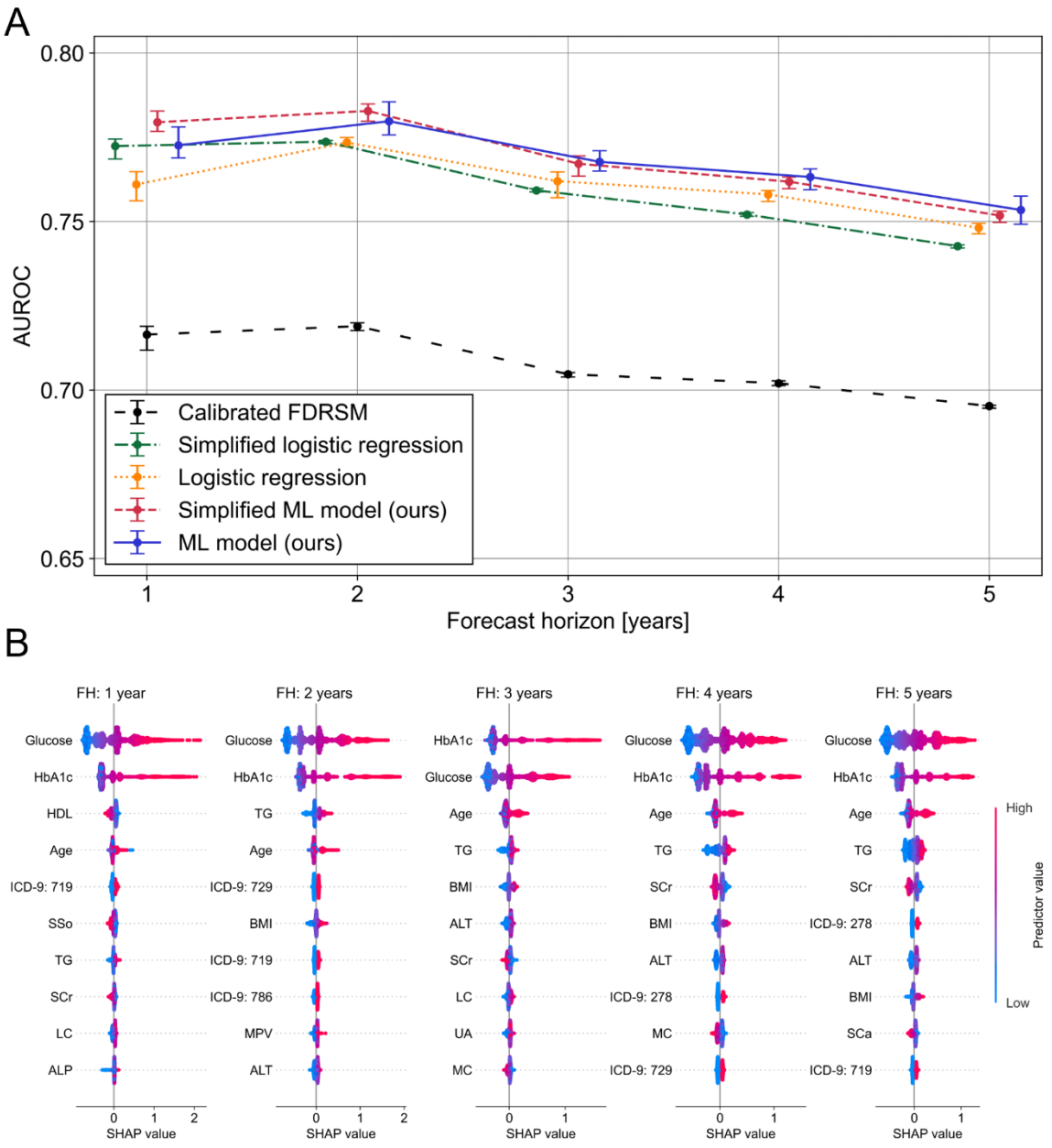

Figure 1: (A) Prediction performances for our ML model, our simplified ML model, the logistic regression, and the calibrated FDRSM for different forecast horizons. Error bars indicate the range of different performance scores from 10 runs with different random seeds. We performed statistical tests to assess whether the performance differences between our ML models and the logistic regression/calibrated FDRSM are statistically significant. This is shown by: * $p<0.05$ and *** $p<0.001$ across all forecast horizons. (B) SHAP plots for all five forecast horizons. The ranking of the predictors is based on their

importance listed in descending order. Each dot represents one patient. The position of the dot on the x-axis denotes its SHAP value. Elements with a positive (negative) SHAP value pull the prediction towards (non-)transition to type 2 diabetes. The color of each dot indicates the corresponding predictor value.

Abbreviations: FH, forecast horizon; HDL, high-density lipoprotein; ICD-9: 719, other and unspecified disorders of joint; BMI, body mass index; TG, triglycerides; SCr, serum creatinine; SSo, serum sodium; ALT, alanine transaminase; ICD-9: 729, other disorders of soft tissues; ICD-9 786, symptoms involving respiratory system and other chest symptoms; P, phosphorus; LC, lymphocytes; UA, uric acid; SCa, serum calcium; MC, monocytes; ICD-9: 278, obesity and other hyperalimentation.