Original Article
Metabolic Risk/Epidemiology

Diabetes Metab J Published online Feb 25, 2021
https://doi.org/10.4093/dmj.2020.0081
pISSN 2233-6079 · eISSN 2233-6087

dmj
DIABETES & METABOLISM JOURNAL

Development and Validation of a Deep Learning Based Diabetes Prediction System Using a Nationwide Population-Based Cohort

Sang Youl Rhee[1,*], Ji Min Sung[2,*], Sunhee Kim[3], In-Jeong Cho[4], Sang-Eun Lee[5], Hyuk-Jae Chang[5]

[1]Department of Endocrinology and Metabolism, Kyung Hee University School of Medicine, Seoul,
[2]Integrative Research Center for Cerebrovascular and Cardiovascular diseases, Yonsei University Health System, Yonsei University College of Medicine, Seoul,
[3]Yonsei University College of Medicine, Yonsei University Health System, Seoul,
[4]Division of Cardiology, Ewha Womans University School of Medicine, Seoul,
[5]Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University Health System, Yonsei University College of Medicine, Seoul, Korea

Background: Previously developed prediction models for type 2 diabetes mellitus (T2DM) have limited performance. We developed a deep learning (DL) based model using a cohort representative of the Korean population.
Methods: This study was conducted on the basis of the National Health Insurance Service-Health Screening (NHIS-HEALS) cohort of Korea. Overall, 335,302 subjects without T2DM at baseline were included. We developed the model based on 80% of the subjects, and verified the power in the remainder. Predictive models for T2DM were constructed using the recurrent neural network long short-term memory (RNN-LSTM) network and the Cox longitudinal summary model. The performance of both models over a 10-year period was compared using a time dependent area under the curve.
Results: During a mean follow-up of 10.4±1.7 years, the mean frequency of periodic health check-ups was 2.9±1.0 per subject. During the observation period, T2DM was newly observed in 8.7% of the subjects. The annual performance of the model created using the RNN-LSTM network was superior to that of the Cox model, and the risk factors for T2DM, derived using the two models were similar; however, certain results differed.
Conclusion: The DL-based T2DM prediction model, constructed using a cohort representative of the population, performs better than the conventional model. After pilot tests, this model will be provided to all Korean national health screening recipients in the future.

Keywords: Diabetes mellitus, type 2; Mass screening; Prediabetic state; Prediction

INTRODUCTION

The rising global prevalence of diabetes mellitus (DM) and its related complications have increased the burden on the global health care system [1]. Recent reports suggest that one in 11 adults worldwide have DM, and it is considered to be one of the major causes of reduced life expectancy [2].

However, type 2 diabetes mellitus (T2DM) is a preventable disease. Early screening and appropriate interventions may prevent the onset and progress of T2DM. Previous clinical trials have demonstrated the efficacy of preventive interventions in subjects at high-risk of T2DM [3,4]. In addition to preventing the onset of T2DM, interventions may prevent the occurrence of long-term complications [5,6]; reports suggest that this approach is cost-effective [7].

The effective prevention and management of T2DM in the population necessitates the accurate identification of subjects who may develop T2DM. Several researchers have attempted

Corresponding author: Hyuk-Jae Chang https://orcid.org/0000-0002-6139-7545
Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
E-mail: hjchang@yuhs.ac

*Sang Youl Rhee and Ji Min Sung contributed equally to this study as first authors.

Received: Jul. 27, 2020; Accepted: Aug. 19, 2020

to develop models to predict the individual risk of T2DM [8-10]. However, the existing prediction model does not include various risk factors for T2DM, and its predictive power is limited [8,10-12].

Recently, with the development of artificial intelligence technology, efforts are being made to apply new techniques such as deep learning (DL) to existing disease models. DL is an algorithm used in the field of computer science, that identifies the patterns of large datasets and predicts the results [13,14].

It is known that the prediction accuracy of DL in various imaging types is comparable with that of skilled experts [15,16], and applications in clinical practice are also rising. Attempts have also been made to employ DL in the prediction of various chronic diseases including T2DM [17,18]. However, till date, the proposed model does not outperform conventional tools [19-22].

In this study, we constructed a DL-based T2DM prediction model using large scale longitudinal cohort data representative of the Korean population. We then compared this model to a conventional Cox regression based model and evaluated its performance and clinical utility.

METHODS

NHIS-Health Screening cohort

Except for 3% of 'medical protection' beneficiaries, 97% of the total Korean population is covered by a single health insurance system, namely, the National Health Insurance of Korea. Information on individual utilization of medical facilities, medications, and diagnostic codes, configured in the form of International Classification of Diseases, 10th revision (ICD-10) are archived in the National Health Insurance Service (NHIS) database [23]. In addition, the NHIS provides a biennial health check-up program for all beneficiaries over 40 years of age, that comprises evaluation of anthropometric parameters, a self-administered questionnaire on health related behavior, past medical history, family history, and laboratory tests.

The NHIS-Health Screening (NHIS-HEALS) cohort was established by including approximately 10% of the entire population of NHIS health check-ups between 2002 and 2003 [24]. The cohort comprises 514,795 individuals, and has been systemically sampled to represent the entire Korean population. The clinical course of the included subjects will be observed until follow-up is feasible, i.e., till death or immigration. It is currently possible to use the follow-up data until 2013 for research, after approval from the NHIS. All data for this cohort are provided to researchers after anonymization and de-identification.

Study subjects

From the overall cohort of 514,795 subjects, we excluded those with pre-existing type 1 DM or T2DM from the self-reported past medical history, and those with a fasting blood glucose (FBG) ≥126 mg/mL on baseline laboratory tests. We then excluded those with a diagnosis of DM (based on ICD-10 codes E10.x-E14.x, O24.x), or with prescriptions for anti-diabetic medication (oral hypoglycemic agents or insulin) in the health insurance claims database at the time of the baseline check-up. We also excluded those who died in 2002 to 2003, since serial clinical data on health check-ups or follow-ups were unavailable for determining the incidence of T2DM. Finally, 335,302 individuals were selected as candidates for the study, and only the latest health check-up data from the 2002 to 2006 period were included after the baseline date; 80% (268,241) of the subjects were randomly selected for inclusion during model development (Supplementary Fig. 1).

Clinical variables

All procedures of the national healthcare check-up were performed by experts according to standardized protocols [24]. The anthropometric parameters of systolic blood pressure (SBP), diastolic blood pressure (DBP), and body mass index (BMI) were used in this study. Among laboratory tests, FBG, total cholesterol (TC), aspartate aminotransferase, alanine aminotransferase, gamma-glutamyl transferase, and dip-stick based proteinuria tests were used. The personal behavior, past medical history, and family history of subjects were investigated using a questionnaire. For evaluating personal behavior, smoking, alcohol, and exercise habits were investigated. These three measures were classified as "yes" or "no" for current smoking status, "drinker" or "non-drinker" for alcohol consumption, and "yes" or "no" for exercise. The past medical and family history were examined for the presence of hypertension, heart disease, stroke, and other illnesses (including malignancy). The presence or absence of a condition was determined by the availability of a diagnosis from a doctor. Details of variables included in the analyses have been presented in Table 1. Additionally, in cases of missing data, multiple imputations were used under fully conditional specification [25] using the machine learning procedure [26].

Table 1. Baseline characteristics of the training set

Variable	All (*n*=268,241)	Missing, %	Incident T2DM			
			Yes (*n*=23,420)	Missing, %	No (*n*=244,821)	Missing, %
Age, yr	51.8±9.1	0.00	54.6±9.3	0.00	51.5±9.0	0.00
Male sex	149,723 (55.82)	0.00	14,306 (61.08)	0.00	135,417 (55.31)	0.00
BMI, kg/m^2	23.9±2.9	0.08	25.2±3.1	0.08	23.8±2.8	0.08
SBP, mm Hg	126.0±17.6	0.03	132.0±18.4	0.02	125.4±17.4	0.04
DBP, mm Hg	79.3±11.6	0.05	82.5±11.9	0.03	79.0±11.5	0.06
FBG, mg/dL	90.8±12.6	0.10	97.6±14.0	0.10	90.2±12.3	0.10
TC, mg/dL	199.6±37.3	0.14	208.3±39.7	0.12	198.8±36.9	0.14
Hemoglobin, g/dL	14.0±1.5	0.09	14.2±1.5	0.09	13.9±1.5	0.09
AST, IU/L	26.5±16.4	0.08	30.4±20.7	0.06	26.1±15.9	0.08
ALT, IU/L	25.5±20.4	0.08	32.4±25.2	0.07	24.8±19.8	0.08
GGT, IU/L	35.5±47.2	0.08	51.7±69.9	0.07	34.0±44.1	0.08
Proteinuria	4,048 (1.51)	0.25	594 (2.54)	0.30	3,454 (1.41)	0.24
Smoking	85,774 (31.98)	4.35	8,718 (37.22)	4.51	77,056 (31.47)	4.34
Alcohol	118,972 (44.35)	1.82	10,886 (46.48)	1.79	108,086 (44.15)	1.83
Exercise	113,809 (42.43)	3.01	9,450 (40.35)	3.18	104,359 (42.63)	2.99
Personal history						
Hypertension	17,365 (6.47)	0.00	2,849 (12.16)	0.00	14,516 (5.93)	0.00
Heart disease	2,709 (1.01)	0.00	428 (1.83)	0.00	2,281 (0.93)	0.00
Stroke	901 (0.34)	0.00	118 (0.50)	0.00	783 (0.32)	0.00
Others[a]	27,406 (10.22)	0.00	2,708 (11.56)	0.00	24,698 (10.09)	0.00
Family history						
Hypertension	22,306 (8.32)	11.58	2,056 (8.78)	11.99	20,250 (8.27)	11.54
Heart disease	7,910 (2.95)	12.07	650 (2.78)	12.40	7,260 (2.97)	12.04
Stroke	15,259 (5.69)	11.81	1,342 (5.73)	12.15	13,917 (5.68)	11.78
DM	14,778 (5.51)	11.83	1,689 (7.21)	12.00	13,089 (5.35)	11.81
Others[a]	39,946 (14.89)	11.65	3,031 (12.94)	12.03	36,915 (15.08)	11.61
Follow-up, yr	10.4±1.7	0.00	6.7±2.6	0.00	10.8±1.1	0.00
Check-up, *n*	2.9±1.0	0.00	2.8±1.0	0.00	2.9±1.0	0.00

Values are presented as mean±standard deviation or number (%).
T2DM, type 2 diabetes mellitus; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transferase.
[a]Malignancy.

Identification of new DM cases

The new onset of T2DM among the subjects was confirmed by their ICD-10 codes (E11.x-E14.x), prescriptions of anti-diabetic medications (oral anti-diabetic medications and/or insulin), and FBG levels. This definition is based on the consensus of relevant findings widely used in previous studies [27,28].

Construction of prediction models

Prediction models were constructed using records from baseline and follow-up visits (Supplementary Table 1). The intervals were defined as the periods from the first health check-up date to the date of diagnosis of T2DM, and to the end of the study for non-T2DM. The records used for analysis included health check-up data from the 2002 to 2006 period. For in-

stance, if a patient diagnosed with T2DM in 2005 had two health check-ups in 2002 and 2004, the analysis used both health check-up records, and if a patient with T2DM in 2009 had a health check-up four times a year between 2002 and 2008, only the health check-up data available between 2002 and 2006 was used. This controlled the difference in the amount of data that the subjects had in the check-up records, by adjusting their amount.

The Cox regression model was first employed using longitudinal data, with higher accuracy compared to a single measurement method; this method has been described earlier to compare DL with longitudinal data. The Cox regression model used the mean, standard deviations (SDs), minimum and maximum values for continuous variables, and the mean and SDs for categorical variables; these were computed from the periodic health check-up data. The detailed methods for this Cox regression model using longitudinal data, and its improved accuracy over single-measures, have been explained previously [29].

For the DL algorithm, a recurrent neural network-long short-term memory (RNN-LSTM) network was used [30]. The variables used in the DL algorithm were the same as those used in the Cox regression model, with longitudinal data. We designed the LSTM model using the following structures: to optimize algorithm, RMSProp was used to update parameters through back-propagation [31], and hyper-parameters at a learning rate of 0.01 were constructed with a dropout probability of 50%, and a mini-batch of 64. The exact answer was one-hot encoded to be used as cross entropy in a loss function; there were two classes. The particulars of DL and the model building process have been proven (Appendix 1).

Converting the output variables for longitudinal study

The use of a Machine Learning-LSTM to determine the occurrence of disease at a certain point in time needed to be examined. As in previous studies using vector variables, we converted binary into multi-class output variables, which are vector types [32-35]. We analyzed the case every year through output variable conversion to identify the specific point in time at which a disease occurred. In the output layer, each node expresses a time interval from 1 to 10 years, at intervals of 1 year. The value of each node is the survival probability for that point in time. The probability of survival after disease occurrence is 0, and the probability of a disease occurring after the disease-free survival time for censored cases are estimated by the Kaplan-Meier survival function. The predicted outputs are the probability of survival at each time [34].

Solution to the problem of understanding classification decisions

In order to overcome problems that cannot explain the reason for classification, and to identify the effects of input variables, layer-wise relevance propagation (LRP) [36], one of the Explainable Artificial Intelligence (XAI) techniques used in artificial neural networks, were used [37,38].

The order of each variable was sorted in descending order by calculating the mean for the entire LRP output value for each input sample. The number of feature variables was n, the number of input samples was m, and the output value of the prediction model was $o=\{o_1 \dots o_m\}$, the ranking of feature variables was expressed as follows.

$$\mathrm{rank}(o)=\mathrm{desc}\left(\sum_{i=0}^{n}\sum_{j=0}^{m}\mathrm{lrp}_i(o_j)\right)$$

Using this technique, we demonstrated the influence of feature variables that were used for building the model.

Evaluation of prediction performance

The performance of the constructed model was evaluated in the validation dataset, which included 20% of the subjects. We evaluated the area under the curve (AUC) every year by comparing the survival probability based on Cox regression, and the probability of DL using the actual answer. Therefore, using calculation of time dependent AUC each year, we confirmed the predicted performance of our models, the Cox regression and DL [39,40]. The calibration was used to compare observed with predicted event probabilities.

Statistical tools

All statistical analyses were conducted using the SAS version 9.4 (SAS Inc., Cary, NC, USA) and R version 3.3.3 (www.R-project.org) statistical software packages.

Ethical statement

This study was approved by the Institutional Review Board (IRB) of the Yonsei University, Severance Hospital, Seoul, Korea (IRB no. 4-2016-0383). The requirement for informed consent was waived by the IRB as de-identified data was used for analyses.

RESULTS

Characteristics of the subjects

The mean age of subjects in the training set was 51.8±9.1 years, and 149,723 (55.82%) were male (Table 1). The mean BMI, SBP, and DBP were 23.9±2.9 kg/m^2, 126.0±17.6 mm Hg, and 79.3±11.6 mm Hg, respectively. Laboratory test results showed that mean FBG was 90.8±12.6 mg/dL, TC was 199.6±37.3 mg/dL, and proteinuria was present in 1.51% of the total subjects. Among the subjects, 85,774 (31.98%) were current smokers and 118,972 (44.35%) were regular alcohol drinkers; 113,809 (42.43%) exercised regularly. The mean duration of follow-up of the cohort was 10.4±1.7 years; 2.9±1.0 national health check-ups were performed during this period. The minimum and maximum health check-up frequencies per person were 2 and 5, respectively. In the selected individuals, 23,420 (8.7%) were diagnosed with T2DM during the follow-up period.

The characteristics of the validation and training sets were similar (Supplementary Table 2). In addition, the incidence of T2DM between the training and validation sets were also similar (Supplementary Table 3).

Hazard ratio for new onset T2DM in the Cox model

The Cox longitudinal summary model was used to estimate the hazard ratio (HR) and 95% confidence interval (CI) of clinical variables affecting new onset T2DM among subjects in the training set (Table 2). Various variables that significantly increased the HR for T2DM were identified. In particular, the HR of family history of DM (HR, 1.523; 95% CI, 1.462 to 1.586), age (HR, 1.369; 95% CI, 1.348 to 1.391), smoking (HR, 1.355; 95% CI, 1.308 to 1.405), personal history of heart disease (HR, 1.343; 95% CI, 1.254 to 1.439), and proteinuria (HR, 1.217; 95% CI, 1.090 to 1.359) were prominent among the variables. Conversely, the HR was significantly lower for male individuals (HR, 0.809; 95% CI, 0.773 to 0.846), alcohol drinkers (HR, 0.844; 95% CI, 0.816 to 0.873), and those who exercised (HR, 0.876; 95% CI, 0.852 to 0.901).

Clinical variables frequently observed in DL-based models

While constructing the DL-based model, the most frequently observed clinical variables in subjects with new T2DM were listed using the LRP algorithm (Table 3). Most of the variables were found to be similar to the risk factors of T2DM identified in the conventional model. However, the family or personal history related variables were not included in the 10 most frequently listed variables in the two methods.

Comparison of prediction models

The prediction performance of the Cox and DL-based prediction models was compared. The results demonstrated the performance of the DL-based model to be superior to that of the Cox model across all observation periods (Fig. 1).

The discriminative performances measured by AUC for 5 years were 0.842 (95% CI, 0.832 to 0.852) and 0.877 (95% CI, 0.869 to 0.885) in the Cox and DL models, respectively. In addition, the discriminative performances measured by AUC for 10 years were 0.807 (95% CI, 0.801 to 0.813) and 0.827 (95% CI, 0.821 to 0.833) in the Cox and DL models, respectively. Among the two predictive models, the DL-based model showed higher sensitivity for 5 years at 81.6% (95% CI, 79.8 to 83.4), and specificity, at 76.5% (95% CI, 76.2 to 76.8). This model also demonstrated higher sensitivity for 10 years, at 75.1% (95% CI, 73.9 to 76.2) and specificity, at 74.0% (95% CI, 73.7 to 74.4). The detailed analysis results of these two models have been summarized separately (Supplementary Table 4). The calibration results of both the models have also been summarized separately (Supplementary Fig. 2).

DISCUSSION

Effective screening of high-risk subjects in the population, and evidence-based interventions will help in improving public health, and will reduce the burden of T2DM on the national health care system [3,4]. Establishing public health system based interventions in countries or regions known to be at high risk of T2DM, including Korea, are expected to provide considerable benefits to the population. It is essential to develop an accurate model for predicting T2DM for achieving these goals.

However, many of the previous studies were not based on subjects that were representative of the general population, and their accuracy using conventional methodology was not satisfactory. In addition, since various factors influence the occurrence and exacerbation of T2DM, predictive models constructed using few variables have low power, while models including an excess of variables are complex and cumbersome, and are unsuitable for use in the clinic [41]. Most large studies have included individuals with specific ethnic or national backgrounds, and their findings are not generalizable to other populations [42]. Therefore, the existing DM prediction model

Table 2. Hazard ratios for T2DM risk factors in the Cox longitudinal summary model of the training set

Variable		HR	95% CI	*P* value
Age, /10 yr		1.369	1.348–1.391	<0.0001
Male sex		0.809	0.773–0.846	<0.0001
BMI, kg/m^2	Mean	1.105	1.100–1.110	<0.0001
	SD	0.986	0.973–0.999	0.041
SBP, mm Hg	Mean	1.007	1.006–1.009	<0.0001
	SD	1.002	1.000–1.004	0.0342
DBP, mm Hg	Mean	1.001	0.999–1.004	0.3908
	SD	1.001	0.999–1.004	0.3334
FBG, mg/dL	Mean	1.059	1.058–1.060	<0.0001
	SD	0.970	0.969–0.970	<0.0001
TC, mg/dL	Mean	1.003	1.003–1.003	<0.0001
	SD	1.003	1.002–1.004	<0.0001
Hemoglobin, g/dL	Mean	1.082	1.066–1.098	<0.0001
	SD	1.102	1.075–1.131	<0.0001
AST, IU/L	Mean	0.991	0.989–0.993	<0.0001
	SD	1.006	1.004–1.008	<0.0001
ALT, IU/L	Mean	1.019	1.018–1.020	<0.0001
	SD	0.989	0.987–0.990	<0.0001
GGT, IU/L	Mean	1.002	1.002–1.002	<0.0001
	SD	1.000	1.000–1.000	0.9787
Proteinuria	Yes[a]	1.217	1.090–1.359	0.0005
	SD	1.230	1.071–1.413	0.0035
Smoking	Yes[a]	1.355	1.308–1.405	<0.0001
	SD	0.938	0.885–0.994	0.0306
Alcohol	Yes[a]	0.844	0.816–0.873	<0.0001
	SD	1.180	1.118–1.244	<0.0001
Exercise	Yes[a]	0.876	0.852–0.901	<0.0001
	SD	1.069	1.022–1.118	0.004
Personal history	Hypertension	1.192	1.152–1.233	<0.0001
	Heart disease	1.343	1.254–1.439	<0.0001
	Stroke	1.156	1.027–1.302	0.0162
	Others[b]	1.106	1.072–1.141	<0.0001
Family history	Hypertension	0.937	0.903–0.973	0.0007
	Heart disease	0.876	0.822–0.933	<0.0001
	Stroke	0.954	0.912–0.997	0.0382
	DM	1.523	1.462–1.586	<0.0001
	Others[b]	0.937	0.907–0.967	<0.0001

T2DM, type 2 diabetes mellitus; HR, hazard ratio; CI, confidence interval; BMI, body mass index; SD, standard deviation; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transferase.

[a]'Yes' means that the mean of a categorical variable (flexible variables every check-up) consisting of 0 or 1 is ≥0.5, [b]Malignancy.

Table 3. Rank of risk factors in deep learning model

Rank	Sum of ranks[a]	Feature name	Mean of values[b]	Feature name
1	225139	FBG	0.540717434	FBG
2	390825	Age	0.225780582	Age
3	415586	Sex[c]	0.198429918	ALT
4	452213	ALT	0.183752354	BMI
5	474756	BMI	0.155880525	GGT
6	506423	GGT	0.131993265	SBP
7	528990	SBP	0.11983712	TC
8	590453	TC	0.099701395	Sex
9	686835	AST	0.080821141	Alcohol
10	696361	Alcohol[c]	0.068915813	Exercise

FBG, fasting blood glucose; ALT, alanine aminotransferase; BMI, body mass index; GGT, gamma-glutamyl transferase; SBP, systolic blood pressure; TC, total cholesterol; AST, aspartate aminotransferase. [a]Ranking each sample by absolute value of layer-wise relevance propagation (LRP), then ascending order by summing the ranks by variables in all samples, [b]Calculate the mean for the absolute value of LRP by variable in all samples and sort in descending order, [c]Sex is specified as male or female, alcohol as yes or no.

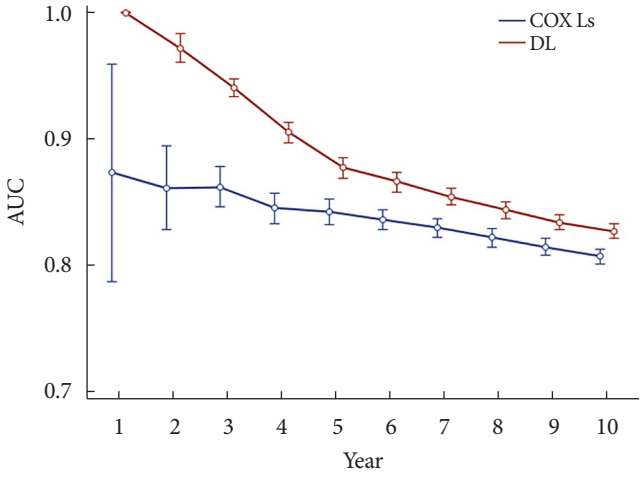


Fig. 1. Area under the curve (AUC) by year for the Cox longitudinal summary model (Cox Ls) and deep learning (DL) model.

did not provide fully satisfactory accuracy in the Korean population [11,12]. One recent study showed that the C-statistics for the models for DM risk at 10 years were 0.71 (95% CI, 0.70 to 0.73) for the men and 0.76 (95% CI, 0.75 to 0.78) for the women [12]. The use of artificial intelligence based technologies for disease prediction has facilitated the introduction of DL-based diabetes prediction models in recent years [19-22]. These results show that the performance of the DL-based prediction model for T2DM is favorable; however, compared with the existing model, the advantages are not very remarkable. A study using data from the Korean National Health and Nutrition Examination Survey found that the performance of DL-based prediction for T2DM was AUC 80.11 [21]. This result is an accuracy of about 80%, which is similar to the previous model. As a result of a study conducted based on electronic medical records of 8,454 subjects, the risk of DM for 5 years was similar to that of the traditional model [19].

We had conducted this study to address the limitations of previous studies. This study is particularly remarkable in that is has been based on a large cohort representative of the population of Korea. Various variables such as anthropometric parameters, personal behavior, past medical history, family history, and laboratory tests were utilized in model development. Additionally, long-term follow-up data for approximately 10 years were available for outcome evaluation. Moreover, careful statistical analysis facilitated the presentation of time dependent AUCs of the two models, and the clinical variables affecting the occurrence of T2DM in the DL-model. Consequently, both models provided reliable results. In particular, the DL-based model performed better performance than the conventional Cox model. The short-term predictive power of the DL-based model also demonstrated excellent performance, with an AUC as high as 0.877 in 5 years. The most important implication of this study lies in the development of a highly accurate DL-based prediction model using a model that is universally applicable to Korean adults aged over 40 years. This provides the considerable advantage of being able to easily and accurately assess the future yearly risk of developing diabetes in the nationwide population. In recent years, there has been a free ongoing trial service in Korea that offers predictive tests using this model for the future risk of diabetes in health checkup recipients who agree to undergo testing. In the future, if its feasibility is established, the service will be provided free of charge to all Koreans.

The Korean Diabetes Prevention Study is currently being conducted in Korea to evaluate the clinical utility of preventive interventions for high-risk patients with diabetes [43]. If the results are conducive, and independent evidence for the prevention of diabetes is established based on national screening projects, Korea will be able to provide a system for systematic screening of high-risk populations and preventive interventions. The provision of diabetes prediction systems to the en-

tire national population based on artificial intelligence, and efforts for the dissemination of evidence-based interventions are rarely observed worldwide. Therefore, we believe that it is necessary to introduce the Korean model to global researchers, and to discuss the future impact on public health.

This study has certain limitations. First, the accuracy of the long-term prediction close to 10 years is lower than that of the short-term prediction of 5 years or less. Second, the inaccuracy of claims-based research may be debated. Third, since all subjects do not necessarily undergo national health checkups, certain errors may have been introduced. Most of the currently available variables have been included in the model; however, the adequacy of the type and numbers of the variables are difficult to estimate. For instance, the detailed classification of personal behavior and family history of chronic disease was difficult; it is possible that the influence of this variable was not accurately calculated. Additionally, the HR was significantly higher for the family history of DM in the conventional than the DL-based model. This is a notable limitation since no mechanisms were available to explain these results based on the current DL based model. Therefore, the results of this study did not completely shift the existing paradigm. We hope that these limitations may be addressed by determining outcomes for longer terms, more detailed clinical phenotyping, application of better analytical methodologies and reflecting the variables that have recently been updated. In particular, we speculate that the addition of individual genomic, microbiomic, and pertinent biomarker data will maximize its predictive power.

Despite these limitations, we successfully constructed a DL-based prediction model based on a representative nationwide cohort, which may easily and accurately predict the risk of T2DM in all members of the general population; we also demonstrated its good performance. This prediction model has already been used among some national health screening examinees in Korea. To the best of our knowledge, this is the first global instance of implementation of a DL-based diabetes prediction system for the entire national population. It is possible that the considerable burden of diabetes may be eventually reduced in Korea if evidence-based personalized preventive interventions are realized in future.

SUPPLEMENTARY MATERIALS

Supplementary materials related to this article can be found online at https://doi.org/10.4093/dmj.2020.0081.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

AUTHOR CONTRIBUTIONS

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

ORCID

Sang Youl Rhee *https://orcid.org/0000-0003-0119-5818*
Ji Min Sung *https://orcid.org/0000-0003-1958-7596*
Hyuk-Jae Chang *https://orcid.org/0000-0002-6139-7545*

FUNDING

None

ACKNOWLEDGMENTS

The authors would like to thank professor Jeong-Taek Woo of Kyung Hee University for his exceptional discourse and inspiration, which encouraged us to conduct the present study.

REFERENCES

1. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018;14:88-98.
2. International Diabetes Federation. IDF Diabetes Atlas. 8th ed. Brussels: International Diabetes Federation; 2017.
3. Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 2001;344:1343-50.
4. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002;346:393-403.
5. Lindstrom J, Peltonen M, Eriksson JG, Ilanne-Parikka P, Aunola S, Keinanen-Kiukaanniemi S, et al. Improved lifestyle and

decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish Diabetes Prevention Study (DPS). Diabetologia 2013;56:284-93.
6. Diabetes Prevention Program Research Group. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. Lancet Diabetes Endocrinol 2015;3:866-75.
7. Diabetes Prevention Program Research Group. The 10-year cost-effectiveness of lifestyle intervention or metformin for diabetes prevention: an intent-to-treat analysis of the DPP/DPPOS. Diabetes Care 2012;35:723-30.
8. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med 2009;151:775-83.
9. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ 2009;338:b880.
10. Lindstrom J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 2003;26:725-31.
11. Lee YH, Bang H, Kim HC, Kim HM, Park SW, Kim DJ. A simple screening score for diabetes for the Korean population: development, validation, and comparison with other scores. Diabetes Care 2012;35:1723-30.
12. Ha KH, Lee YH, Song SO, Lee JW, Kim DW, Cho KH, et al. Development and validation of the Korean diabetes risk score: a 10-year national cohort study. Diabetes Metab J 2018;42:402-14.
13. Deo RC. Machine learning in medicine. Circulation 2015;132:1920-30.
14. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. Am J Gastroenterol 2010;105:1224-6.
15. Chen JH, Asch SM. Machine learning and prediction in medicine: beyond the peak of inflated expectations. N Engl J Med 2017;376:2507-9.
16. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-67.
17. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12:e0174944.
18. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018;20:e10775.
19. Choi BG, Rha SW, Kim SW, Kang JH, Park JY, Noh YK. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. Yonsei Med J 2019;60:191-9.
20. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:515.
21. Ryu KS, Lee SW, Batbaatar E, Lee JW, Choi KS, Cha HS. A deep learning model for estimation of patients with undiagnosed diabetes. Appl Sci 2020;10:421.
22. Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do TTT, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput Methods Programs Biomed 2019;182:105055.
23. Shin DW, Cho B, Guallar E. Korean National Health Insurance database. JAMA Intern Med 2016;176:138.
24. Seong SC, Kim YY, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. BMJ Open 2017;7:e016640.
25. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16:219-42.
26. SAS Institute Inc.: SAS/STAT® 14.1 User's Guide. Available from: https://support.sas.com/documentation/onlinedoc/stat/examples/141/index.html (cited 2021 Jan 4).
27. Lee YH, Han K, Ko SH, Ko KS, Lee KU; Taskforce Team of Diabetes Fact Sheet of the Korean Diabetes Association. Data analytic process of a nationwide population-based study using national health information database established by National Health Insurance Service. Diabetes Metab J 2016;40:79-82.
28. Ko SH, Han K, Lee YH, Noh J, Park CY, Kim DJ, et al. Past and current status of adult type 2 diabetes mellitus management in Korea: a National Health Insurance Service database analysis. Diabetes Metab J 2018;42:93-100.
29. Cho IJ, Sung JM, Chang HJ, Chung N, Kim HC. Incremental value of repeated risk factor measurements for cardiovascular disease prediction in middle-aged Korean adults: results from the NHIS-HEALS (National Health Insurance System-National Health Screening Cohort). Circ Cardiovasc Qual Outcomes 2017;10:e004197.
30. Hochreiter S, Schmidhuber J. Long short-term memory. Neu-

ral Comput 1997;9:1735-80.
31. Tieleman T, Hinton G. 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning. Available from: https://www.coursera.org/learn/neural-networks-deep-learning (cited 2020 Jan 4).
32. Street WN. A neural network model for prognostic prediction. Proceedings of the Fifteenth International Conference on Machine Learning; 1998 Jul 24-27; Madison, WI. San Francisco: Morgan Kaufmann Publishers; 1998. pp. 540-6.
33. Baesens B, Van Gestel T, Stepanova M, Van den Poel D, Vanthienen J. Neural network survival analysis for personal loan data. J Oper Res Soc 2005;56:1089-98.
34. Chi CL, Street WN, Wolberg WH. Application of artificial neural network-based survival analysis on two breast cancer datasets. AMIA Annu Symp Proc 2007;2007:130-4.
35. Dezfouli HN, Bakar MRA, Dezfouli HN. Feed forward neural networks models for survival analysis. 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE); 2012 Sep 10-12; Langkawi, MY. IEEE; 2012. pp. 1-5.
36. Bach S, Binder A, Montavon G, Klauschen F, Muller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 2015;10:e0130140.
37. Escalante HJ, Escalera S, Guyon I, Baro X, Gucluturk Y, Guclu U, et al. Explainable and interpretable models in computer vision and machine learning. Cham: Springer; 2018. Chapter, Explanation methods in deep learning: users, values, concerns and challenges; pp.19-36.
38. Arras L, Montavon G, Muller KR, Samek W. Explaining recurrent neural network predictions in sentiment analysis. Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2017 Sep 8; Copenhagen, DM. Association for Computational Linguistics; 2017. pp. 159-68.
39. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337-44.
40. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61:92-105.
41. Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. Endocrinol Metab (Seoul) 2016;31:38-44.
42. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343:d7163.
43. Rhee SY, Chon S, Ahn KJ, Woo JT; Korean Diabetes Prevention Study Investigators. Hospital-based Korean diabetes prevention study: a prospective, multi-center, randomized, open-label controlled study. Diabetes Metab J 2019;43:49-58.

Appendix 1. Model building and training in the recurrent neural network

Recurrent neural network long short-term memory (RNN-LSTM) was developed to solve long-range dependency and vanishing gradient problems seen in RNN, which degraded performance as event length increased. Our proposed LSTM model is designed with the following structure. Our proposed LSTM model is designed with the following structure. For the optimization of the algorithm, RMSProp [31] was used to update parameters through back-propagation. The learning sample set S is consisted of ordered pairs (x, y) with inputs and correct answers. Input x is an element of input set X, whereas correct answer y is an element of correct answers set Y. Here, x is consisted of serial data in the form of an individual's information. The T in input $x=(x_1,\ldots, x_t, \ldots, x_T)$ represents the length of a sample (i.e., the number of events) and varies individually. A health examination record is consisted of an event; and, all events $x_1, \ldots, x_T$ for a person are in chronological order. x_t represents an event (examination record) at a specific time and is used as a vector with features. The correct answer $y=(y_1, \ldots, y_K)$ has a Boolean value indicating whether or not type 2 diabetes mellitus (T2DM) occurred in the past. Hyper-parameters at a learning rate of 0.01 were configured, a dropout probability of 50%, and a mini-batch of 64. The correct answer is one-hot encoded to be used as cross entropy in a loss function. K—the number of classes—is set as 2. Assuming the output of prediction model is $\hat{y}=(\hat{y}_1, \ldots, \hat{y}_K)$, cross entropy as shown in Equation 1 is used for our loss function.

$$E(y,\hat{y}) = -\sum_{i=1}^{K}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad \text{(Equation 1)}$$

The output sequence for LSTM was $o=(o_1, \ldots o_t, \ldots, o_T)$ where T was the event length being the same as the event length of x. The prediction result of an input sample provided the probability of T2DM in the near future taken the occurrence of past event into account.

Only the last output o_T among $o_1, \ldots o_t, \ldots, o_T$ was used and reflected to an output z as shown in Equation 2. Here, $W \in R^{K \times H}$ was the parameter to be optimized and H was the number of hidden nodes in the last hidden layer. To calculate the probability of $\hat{y}_i$ from z, the softmax function was used as shown in Equation 3.

$$z = W\mathbf{o}_T \quad \text{(Equation 2)}$$

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad \text{for } i = 1, \ldots, K \quad \text{(Equation 3)}$$

LSTM had a memory cell with input, forget and output gates. Each LSTM unit uses the equations in Equation 4 which are commonly used in LSTM.

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o) \\ h_t &= o_t \circ \tanh(c_t) \end{aligned} \quad \text{(Equation 4)}$$

σ was the logistic sigmoid function and i, f, o, and c were respectively the input gate, forget gate, output gate, and cell. The subscripts in weight matrix above have an obvious meaning. For example, W_{hi} is the hidden-input gate matrix while W*xo* being the input-output gate matrix. The bs are bias terms which are added for i, f, o, and c equations. Let N be the number of LSTM blocks and M the number of inputs, W_i, W_f, W_c, $W_o \in R^{N \times M}$.

Supplementary Table 1. Variables used in each prediction model

Model	Variable
Cox Ls	Age at baseline, sex
	BMI, SBP, DBP, FBG, TC, hemoglobin, AST, ALT, GGT, proteinuria
	Smoking, alcohol, exercise
	Personal history of hypertension, heart disease, stroke, and others (including malignancy), family history of hypertension, heart disease, stroke, DM, and others (including malignancy)
DL	Date of each health examination
	Age at baseline, sex
	BMI, SBP, DBP, FBG, TC, hemoglobin, AST, ALT, GGT, proteinuria
	Smoking, alcohol, exercise
	Personal history of hypertension, heart disease, stroke, and others (including malignancy), family history of hypertension, heart disease, stroke, DM, and others (including malignancy)

Cox Ls, Cox longitudinal summary; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transferase; DM, diabetes mellitus; DL, deep learning.

Supplementary Table 2. Baseline characteristics of the validation set

Variable	All (*n*=67,061)	Incident T2DM	
		Yes (*n*=5,736)	No (*n*=61,325)
Age, yr	51.8±9.1	54.5±9.2	51.5±9.1
Male sex	37,637 (56.12)	3,558 (62.03)	34,079 (55.57)
BMI, kg/m^2	23.9±2.9	25.1±3.1	23.7±2.9
SBP, mm Hg	125.9±17.5	131.8±18.4	125.3±17.3
DBP, mm Hg	79.2±11.5	82.5±11.9	78.9±11.4
FBG, mg/dL	90.9±12.6	98.0±13.9	90.2±12.2
TC, mg/dL	199.6±37.1	208.7±40.1	198.8±36.7
Hemoglobin, g/dL	14.0±1.5	14.2±1.5	13.9±1.5
AST, IU/L	26.6±17.2	30.7±20.7	26.2±16.7
ALT, IU/L	25.5±20.3	32.8±25.5	24.9±19.6
GGT, IU/L	35.9±49.1	53.1±74.8	34.2±45.6
Proteinuria	947 (1.41)	136 (2.37)	811 (1.32)
Smoking	21,441 (31.97)	2,186 (38.11)	19,255 (31.40)
Alcohol	29,731 (44.33)	2,641 (46.04)	27,090 (44.17)
Exercise	28,685 (42.77)	2,363 (41.20)	26,322 (42.92)
Personal history			
Hypertension	4,238 (6.32)	694 (12.10)	3,544 (5.78)
Heart disease	663 (0.99)	97 (1.69)	566 (0.92)
Stroke	204 (0.30)	31 (0.54)	173 (0.28)
Others[a]	6,962 (10.38)	660 (11.51)	6,302 (10.28)
Family history			
Hypertension	5,474 (8.16)	490 (8.54)	4,984 (8.13)
Heart disease	2,015 (3.00)	158 (2.75)	1,857 (3.03)
Stroke	3,714 (5.54)	315 (5.49)	3,399 (5.54)
DM	3,720 (5.55)	451 (7.86)	3,269 (5.33)
Others[a]	9,859 (14.70)	741 (12.92)	9,118 (14.87)
Follow-up, yr	10.4±1.7	6.7±2.6	10.8±1.1
Check-up, *n*	2.9±1.0	2.8±1.0	2.9±1.0

Values are presented as mean±standard deviation or number (%).
T2DM, type 2 diabetes mellitus; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; TC, total cholesterol; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transferase; DM, diabetes mellitus.
[a]Malignancy.

Supplementary Table 3. Incidence of type 2 diabetes mellitus of the subjects

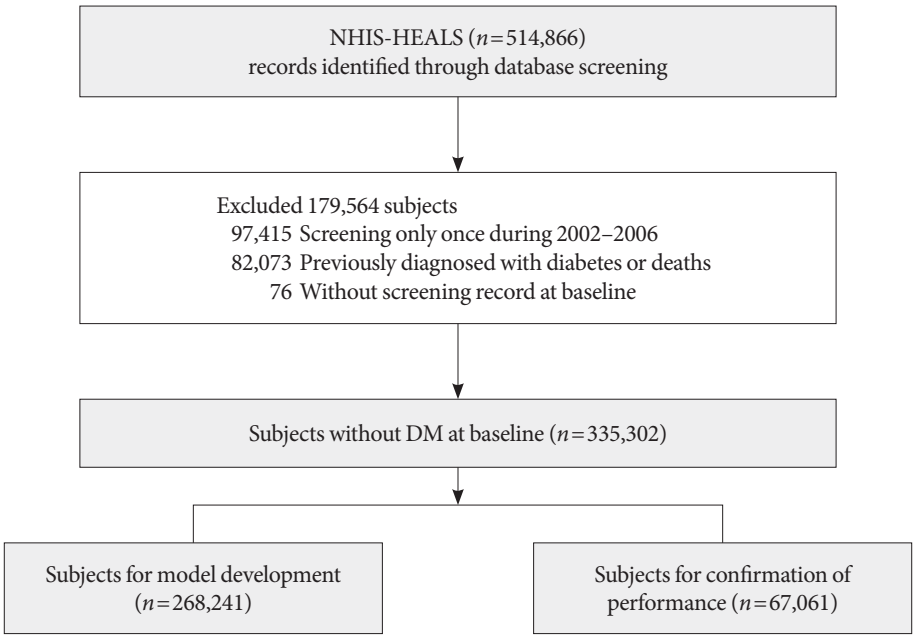
Variable	Year										
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Training set											
Event rate	0.04	0.22	0.51	0.76	0.87	0.94	1.06	1.07	1.15	1.07	1.05
Incidence rate, /100,000 PY	40.64	222.60	514.24	771.08	895.56	981.79	1,117.47	1,151.61	1,251.85	1,185.50	1,183.96
Validation set											
Event rate	0.03	0.20	0.56	0.77	0.87	0.96	0.95	1.05	1.10	1.04	1.01
Incidence rate, /100,000 PY	31.32	204.66	560.16	780.61	895.77	1,005.71	1,009.28	1,132.40	1,202.02	1,158.59	1,138.75

PY, person-year.

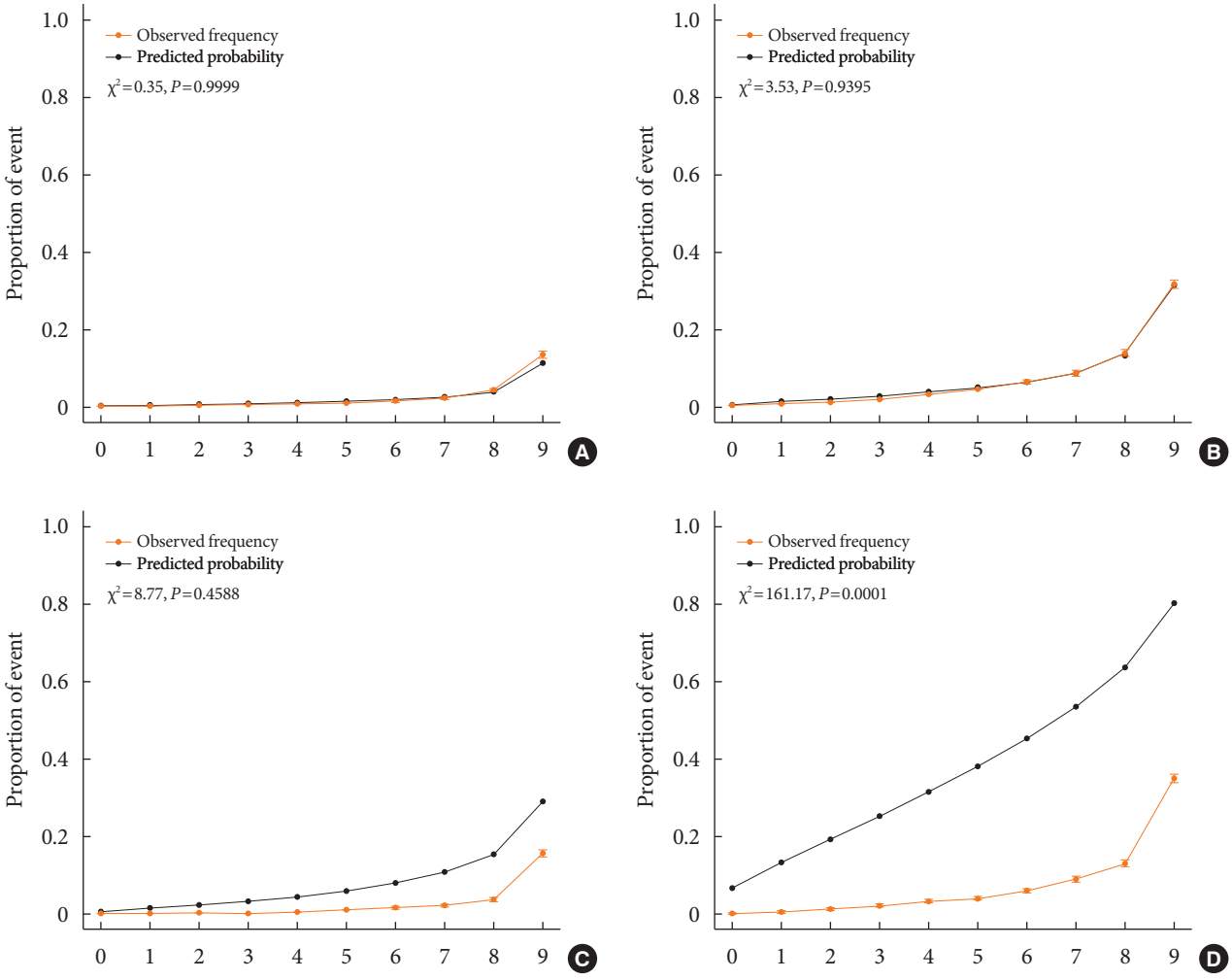
Supplementary Table 4. Statistics of the Cox Ls and DL model on diabetes prediction by year

	AUC (time dependent)	Sensitivity	Specificity	Accuracy	PPV	NPV
Cox Ls						
1 yr	0.873 (0.787–0.959)	1.000 (1.000–1.000)	0.788 (0.785–0.791)	0.788 (0.788–0.788)	0.000 (0.000–0.001)	1.000 (1.000–1.000)
2 yr	0.861 (0.828–0.894)	0.773 (0.704–0.842)	0.843 (0.840–0.846)	0.843 (0.843–0.843)	0.010 (0.008–0.012)	0.999 (0.999–1.000)
3 yr	0.862 (0.846–0.878)	0.799 (0.767–0.832)	0.787 (0.784–0.790)	0.787 (0.787–0.787)	0.032 (0.029–0.035)	0.998 (0.997–0.998)
4 yr	0.845 (0.833–0.857)	0.757 (0.732–0.783)	0.791 (0.788–0.794)	0.791 (0.791–0.791)	0.057 (0.053–0.061)	0.995 (0.994–0.996)
5 yr	0.842 (0.832–0.852)	0.778 (0.758–0.798)	0.759 (0.755–0.762)	0.759 (0.759–0.759)	0.077 (0.073–0.081)	0.992 (0.992–0.993)
6 yr	0.836 (0.828–0.844)	0.761 (0.743–0.778)	0.761 (0.758–0.764)	0.761 (0.761–0.761)	0.102 (0.098–0.107)	0.989 (0.988–0.990)
7 yr	0.830 (0.822–0.837)	0.744 (0.729–0.760)	0.763 (0.760–0.766)	0.762 (0.762–0.762)	0.127 (0.122–0.132)	0.985 (0.984–0.986)
8 yr	0.822 (0.814–0.829)	0.722 (0.708–0.737)	0.767 (0.764–0.771)	0.765 (0.765–0.765)	0.153 (0.148–0.158)	0.979 (0.978–0.981)
9 yr	0.814 (0.808–0.821)	0.755 (0.742–0.767)	0.718 (0.715–0.722)	0.720 (0.720–0.720)	0.159 (0.154–0.164)	0.976 (0.975–0.978)
10 yr	0.807 (0.801–0.813)	0.739 (0.727–0.751)	0.722 (0.719–0.726)	0.723 (0.723–0.723)	0.180 (0.175–0.186)	0.971 (0.969–0.973)
DL						
1 yr	0.999 (0.999–1.000)	1.000 (1.000–1.000)	0.998 (0.997–0.998)	0.998 (0.997–0.998)	0.026 (0.001–0.051)	1.000 (1.000–1.000)
2 yr	0.971 (0.960–0.983)	0.901 (0.851–0.950)	0.939 (0.937–0.941)	0.939 (0.937–0.941)	0.030 (0.025–0.035)	1.000 (1.000–1.000)
3 yr	0.940 (0.933–0.947)	0.879 (0.853–0.906)	0.843 (0.840–0.846)	0.843 (0.840–0.846)	0.047 (0.043–0.051)	0.999 (0.998–0.999)
4 yr	0.905 (0.897–0.913)	0.815 (0.792–0.838)	0.829 (0.826–0.832)	0.829 (0.826–0.832)	0.074 (0.069–0.078)	0.996 (0.996–0.997)
5 yr	0.877 (0.869–0.885)	0.816 (0.798–0.834)	0.765 (0.762–0.768)	0.766 (0.763–0.770)	0.083 (0.079–0.087)	0.994 (0.993–0.994)
6 yr	0.866 (0.858–0.873)	0.768 (0.751–0.785)	0.798 (0.794–0.801)	0.797 (0.793–0.800)	0.120 (0.115–0.125)	0.990 (0.989–0.991)
7 yr	0.854 (0.848–0.861)	0.779 (0.764–0.794)	0.761 (0.758–0.764)	0.762 (0.758–0.765)	0.132 (0.127–0.136)	0.987 (0.986–0.988)
8 yr	0.844 (0.837–0.850)	0.772 (0.758–0.785)	0.752 (0.749–0.756)	0.753 (0.750–0.757)	0.153 (0.148–0.159)	0.983 (0.982–0.984)
9 yr	0.834 (0.828–0.840)	0.741 (0.728–0.754)	0.763 (0.759–0.766)	0.761 (0.758–0.764)	0.181 (0.175–0.187)	0.977 (0.975–0.978)
10 yr	0.827 (0.821–0.833)	0.751 (0.739–0.762)	0.740 (0.737–0.744)	0.741 (0.738–0.744)	0.193 (0.187–0.198)	0.973 (0.971–0.974)

Cox Ls, Cox longitudinal summary model; DL, deep learning; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value.



Supplementary Fig. 1. Study progression. NHIS-HEALS, National Health Insurance Service-Health Screening; DM, diabetes mellitus.



Supplementary Fig. 2. Calibration of two predictive models. Calibration and Hosmer-Lemeshow test for the Cox longitudinal summary model in (A) 5 years and (B) 10 years. Calibration and Hosmer-Lemeshow test for the deep learning model in (C) 5 years and (D) 10 years.