scientific reports

Check for updates

OPEN

Development and validation of a type 2 diabetes mellitus prediction tool using a large Japanese regional insurance claims database

Tatsunori Satoh[1,2,3], Eiji Nakatani[3,4,5]✉, Hiroyuki Ariyasu[6], Chihiro Saito[3], Yuri Oshiro[3], Emi Ohata[3,7], Hiroshi Itoh[3,5,8], Kaori Hayashi[1] & Takeshi Usui[3,5]

With the increasing global prevalence of diabetes, predictive models are crucial for early interventions, including elderly participants who are often underrepresented in existing models. Existing models, however, are often derived from specific subpopulations or from individuals with prevalent diabetes, which limits their generalizability to the broader population. This study aimed to develop and validate a predictive model for type 2 diabetes mellitus (T2DM) onset using a large Japanese cohort including elderly participants. Data from the Shizuoka Kokuho Database, comprising over 2.5 million people, was used. The analysis included 463,248 adults aged 40 and above who underwent health checkups. Participants were split into derivation (308,832) and validation datasets (154,416) in a 2:1 ratio. Predictive factors, identified using Cox proportional hazards models, included demographics, clinical parameters, and lifestyle factors. During a median follow-up period of 5.17 years, 16.9% of the derivation group and 17.0% of the validation group developed T2DM. The model assigns weighted scores to factors like age, sex, BMI, blood pressure, lipid profiles, liver enzymes, kidney function, and lifestyle habits. The model achieved a Harrell's c-index of 0.656 (95% confidence interval, 0.652–0.659) in the validation dataset, indicating modest predictive performance. This model, based on routinely collected health check-up data, may facilitate risk stratification and guide preventive interventions.

Keywords Diabetes mellitus, Incidence, Predictive models, Risk factors

The rising prevalence of diabetes, particularly type 2 diabetes mellitus (T2DM), poses a significant public health challenge, affecting more than 400 million individuals worldwide[1]. Diabetes is a major contributor to a range of severe health issues including cardiovascular and cerebrovascular diseases, peripheral neuropathy, blindness, and renal impairment[2,3]. It also increases the risk for certain cancers, such as pancreatic cancer[4]. Diabetes significantly elevates health care costs, with the worldwide economic impact reaching approximately U.S. $1.3 trillion (95% CI 1.3–1.4) in 2015, accounting for over 1.8% of the global GDP[5]. The incidence of T2DM among the elderly continues to rise[6] highlighting its significance in this population.

In Japan, approximately 10 million people, or approximately one in every thirteen individuals, have been diagnosed with this condition[7,8]. The burden of T2DM in Japan is driven by unique factors such as rapid population aging and increased metabolic risk at lower body mass index thresholds compared with Western populations.

Despite these challenges, T2DM is a preventable disease[9]. Current best practices in T2DM management include the use of antidiabetic drugs, lifestyle modifications such as healthy eating and daily physical activity, and regular monitoring of arterial pressure and lipid levels[9]. Early preventive measures can yield substantial economic benefits[10,11] and are crucial for effective management, particularly for individuals with prediabetes.

[1]Division of Endocrinology, Metabolism and Nephrology, Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan. [2]Department of Gastroenterology, Shizuoka General Hospital, Shizuoka, Japan. [3]Graduate School of Public Health, Shizuoka Graduate University of Public Health, 4-27-2 Kitaando, Aoi-ku, Shizuoka 420-0881, Japan. [4]Department of Biostatistics and Health Data Science, Graduate School of Medical Science, Nagoya City University, Nagoya, Japan. [5]Research Support Center, Shizuoka General Hospital, Shizuoka, Japan. [6]Department of Diabetes and Endocrinology, Shizuoka General Hospital, Shizuoka, Japan. [7] Department of Academic Services, 4DIN Ltd, Tokyo, Japan. [8]Center for Preventive Medicine, Keio University, Tokyo, Japan. ✉email: nakatani.eiji.int@gmail.com

Furthermore, the nationwide annual health checkup system provides an unparalleled opportunity for large-scale risk stratification and early intervention.

The application of predictive models in disease management is increasingly acknowledged for its potential to delay T2DM onset[7,12–15], thereby extending healthy life years and reducing economic burdens. However, these models often face limitations given their reliance on hospital data and the scarcity of information from healthy populations. In addition, existing model developed in Japan was constructed using logistic regression frameworks[16] and restricted to selected subgroups such as participants with long-term follow-up, which may introduce selection bias and limit generalizability.

For predictive models to be effective, they require comprehensive datasets that include not only patients but also healthy and elderly individuals, ensuring both high accuracy and practical applicability to the general population.

To address these issues, the present study developed and validated a predictive model for the onset of T2DM using medical checkup data from the general population, applying a Cox proportional hazards model that incorporates time-to-event information and appropriately accounts for censoring.

Materials and methods

Data resource

Shizuoka prefecture, centrally located in Japan, has a population of nearly 3.7 million. The Shizuoka Kokuho Database (SKDB) is a regionally based longitudinal cohort comprising data from 2,571,418 individuals residing in this prefecture[17]. Widely utilized in various studies[18–20], the SKDB provides comprehensive, personally identifiable information by assigning a unique identifier to each participant. The SKDB is an administrative claims database for beneficiaries in Shizuoka Prefecture's municipal government insurance program and includes beneficiaries of the National Health Insurance and the Late-Stage Elderly Medical Care System. The dataset comprised basic subscriber information (e.g., sex, age, zip code, observation period, and reason for disenrollment, including death) and claims from public health insurance organizations (i.e., the National Health Insurance system for those below 75 years of age and the Late-Stage Elderly Medical Care System for those above 75 years of age). The Japanese Ministry of Health, Labour and Welfare recommends annual health checkups for insured persons aged 40 years and older; the content that can be gathered during medical check-ups in Japan is listed in Supplementary Table 1[17]. The health-check questionnaire is an obligatory component of Japan's annual health examination for insured adults aged ≥ 40 years. Participants complete the paper form on-site; trained public-health nurses review answers and request clarification if needed, after which a physician conducts the physical examination and finalises the record. The utility of the SKDB in real-world risk factor analysis is underscored by its comprehensive data on mortality and follow-up attrition derived from the Basic Resident Registration System. The Shizuoka Kokuho Database (SKDB) is updated annually. The latest release, SKDB2023, covers the period from April 1, 2012, to September 30, 2021.

Study design and participant population

In this retrospective cohort study, we used a population-based approach, analyzing data from the SKDB 2023 version covering the period from April 1, 2012, to September 30, 2021. We secured access to the SKDB on December 6, 2023, from which we extracted the dataset for our analysis. The index date was set as the first health checkup that occurred at least one year after cohort entry, with a one-year baseline period established as shown in Fig. 1. Demographic information and diabetes statuses were collected for one year preceding the index dates. The inclusion criteria were: (1) adults aged ≥ 40 years, (2) at least one annual health check-up during continuous insurance enrolment of ≥ 1 year, and (3) available baseline health check-up data and claims history during the 1-year baseline period. In the initial stages of our analysis, we excluded participants who were already diagnosed with diabetes or who were prescribed antidiabetic agents. To assess risk factors in a relatively healthy population, we applied the following exclusion criteria: (1) pre-existing diagnosis of type 2 diabetes mellitus or prescription of antidiabetic medications at baseline, (2) HbA1c ≥ 6.5% at baseline, (3) eGFR < 30 mL/min/1.73 m^2, and (4) history of any cancer (except non-melanoma skin cancer).

We then examined the association between various screening variables and the development of T2DM. Identification of diabetic patients was based on the prescription of antidiabetic agents (Supplementary Tables 2 and Supplementary Table 3), including insulin and insurance claims coded as T2DM (E11, or E14). To enhance the specificity of our T2DM identification, provisional diagnoses of T2DM were excluded.

We determined the presence of individual comorbidities using standard definitions from the Charlson–Elixhauser comorbidity index[21,22]. Our evaluation of specific comorbidities utilized data from a one-year period prior to the index health checkups. A comorbidity was confirmed present if documented in the insurance claims data.

Participants were followed from the date of their baseline health check-up until the earliest occurrence of death, disenrollment from the insurance system, or the end of the study period (September 30, 2021). Consequently, not all participants had a complete 8-year follow-up. Because this study was conducted using an administrative claims and health check-up database, participant refusals did not occur. Loss to follow-up was defined as disenrollment from the insurance system and was captured within the dataset.

Outcome and candidate predictive variables

The primary outcome assessed in this study was the time to onset of T2DM (Fig. 1). A case was classified as T2DM if the patient was assigned a specific Japanese medical procedure code (Supplementary Table 3) following the initial health checkup. In addition, because ICD-10 code E14 (unspecified diabetes mellitus) is often used in Japan to indicate type 2 diabetes in both clinical and administrative data, we followed the approach of prior

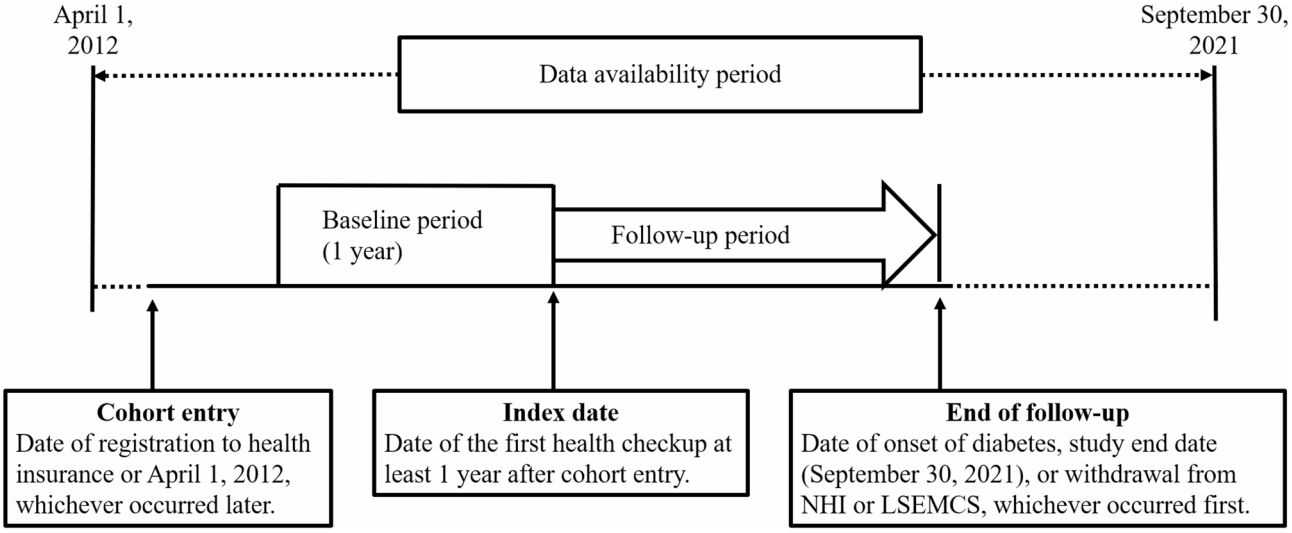


Fig. 1. Study design and participant population. The index date was defined as the first date of medical checkup during a period of continuous subscribership lasting at least one year following cohort entry. The baseline period was established as one year. NHI, national health insurance; LSEMCS, late-stage elderly medical care system; SKDB, Shizuoka Kokuho Database.

Japanese claims-based studies[23,24] and classified individuals with either E11 (type 2 diabetes mellitus) or E14 as having T2DM.

During the initial baseline period, comprehensive demographic data on patients were collected. Within the framework of this study, age, sex, and items based on health checkup-derived variables were identified as potential predictors.

These health checkups included various tests and measurements, including body mass index (BMI [kg/m^2]), systolic blood pressure (mmHg), diastolic blood pressure (mmHg), hematocrit (%), hemoglobin (g/dL), erythrocyte count (10^4/μL), triglyceride (mg/dL), HDL cholesterol (mg/dL), LDL cholesterol (mg/dL), AST (IU/L), ALT (IU/L), γ-GTP (IU/L), HbA1c (%), eGFR (mL/min/1.73 m^2), uric acid (mg/dL), and urinary protein (urine dipstick test).

Additional variables considered included prescribed medications for hypertension and hypercholesterolemia, weight gains of ≥ 10 kg since the age of 20, exercise habits, heavy drinking, and current habitual smoking. Heavy drinking was defined as a daily alcohol consumption exceeding 360 mL. Current smokers were defined as those who had smoked over 100 cigarettes or for at least 6 months and had continued smoking during the month preceding the study.

Statistical analysis

Continuous variables were presented as mean ± standard deviation, and categorical variables were summarized with frequencies and percentages. The Wilcoxon rank-sum test was used to compare continuous variables between two groups, while the chi-squared test was applied for categorical variables.

To develop a predictive scoring system for the onset of T2DM, two-thirds of all eligible individuals were randomly allocated to the derivation dataset and the remaining one-third to the validation dataset. Randomization was performed using a computer-generated random number sequence, ensuring that each individual had an equal probability of being assigned to either cohort. This sampling process was conducted once, and we subsequently compared baseline characteristics between the derivation and validation cohorts to confirm their consistency.

Univariable and multivariable cause-specific Cox proportional hazards regression analyses were executed using the derivation dataset to identify factors predictive of T2DM onset, treating deaths as censoring events. Cumulative incidence functions were estimated considering death as a competing risk, consistent with previous epidemiological studies[25]. These analyses produced HR, 95% confidence intervals (CIs), and *p*-values. Variables with p-values below 0.05 in univariable analysis were included in the multivariable model. In selecting variables for the multivariable analysis, if two variables exhibited a Spearman's correlation coefficient exceeding 0.4 in absolute value, we chose the variable with greater clinical relevance. Clinical relevance was defined based on both prior evidence of association with T2DM and practical interpretability for clinical and public health use. We excluded variables from the multivariable model if they had more than 5% missing values in the health checkup data. Variables significant in the multivariable analysis were identified as predictive factors.

Regression coefficients from the multivariable Cox proportional hazards model, representing the natural logarithm of the hazard ratios (HRs), were used to derive predictive index scores. These coefficients were standardized by multiplying them by a constant factor and then rounded to create integer-based weighted scores. The total predictive index was calculated as the sum of scores assigned to each predictive factor category. To evaluate the validity of this scoring system, we randomly divided the study population into derivation and

validation cohorts. Predictive performance was assessed in the validation cohort using Harrell's c-index[26] to measure model discrimination. In addition, we constructed cumulative incidence curves for T2DM onset across score categories, accounting for death as a competing risk. This internal validation demonstrated consistent predictive ability, with higher scores corresponding to increased risk of T2DM.

Given the non-random occurrence of missing covariates among participants, simple imputation of missing data was avoided. All *p*-values were two-tailed, and results were reported alongside 95% CIs. A *p*-value of less than 0.05 was considered statistically significant. All statistical analyses were conducted utilizing SAS version 9.4 (SAS Institute Inc., Cary, NC, USA) and EZR version 1.60 (Saitama Medical Center, Jichi Medical University, Saitama, Japan), which serves as a graphical user interface for R[27].

Results

Characteristics of patients

Figure 2 presents a detailed flow chart outlining the selection process of the study cohort from the extensive SKDB case dataset. Of the 2,571,418 individuals cataloged in the SKDB, 672,482 underwent health checkups and had a one-year baseline. The analytical cohort was refined by excluding patients already diagnosed with T2DM or on antidiabetic agents (58,642 cases), those with low eGFR (eGFR < 30 mL/min/1.73 m^2, 32,347 cases) or with a history of any cancer (118,245 cases). This exclusion process resulted in 463,248 patients eligible for the study.

These patients were then randomly assigned to either the derivation or validation datasets, used to develop and validate a predictive scoring system. The derivation dataset comprised 308,832 patients, while the validation dataset included 154,416 patients. Comparative analysis of these groups revealed no significant disparities in patient characteristics between the derivation and validation datasets, as detailed in Supplementary Table 4.

Predictive factors for type 2 diabetes mellitus

Table 1 presents the demographic and clinical characteristics of the patients in the derivation dataset. Over the observation period (median [max]: 5.17 [8.50] years), 52,152 of the 308,832 individuals (16.9%) were diagnosed with T2DM. The validation dataset ($n = 154,416$) had a similar observation period of 5.14 years (median) with a maximum of 8.50 years; during this time,17.0% of participants, or 26,279 individuals, received a T2DM diagnosis.

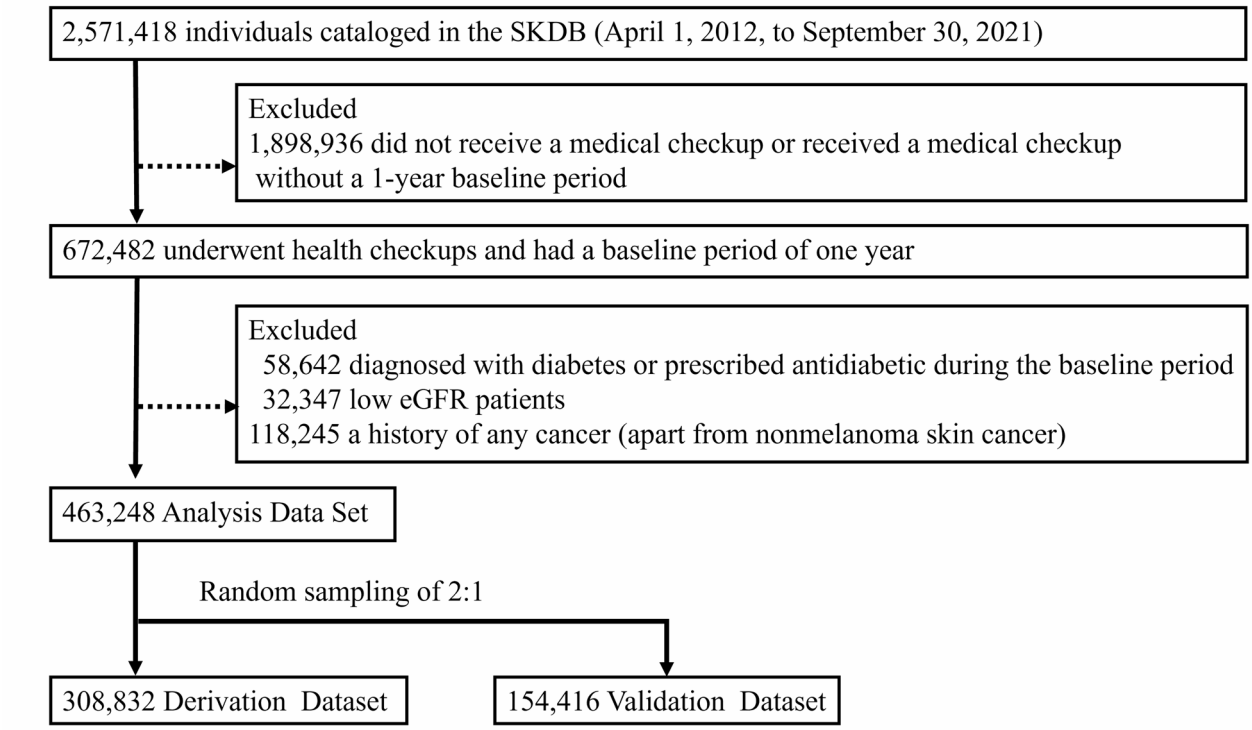


Fig. 2. Flow diagram of this study. Among 2,571,418 individuals registered in the Shizuoka Kokuho Database (SKDB) between April 1, 2012, and September 30, 2021, a total of 1,898,936 were excluded because they did not undergo a medical checkup or lacked a 1-year baseline period. Of the remaining 672,482 participants, those with pre-existing diabetes or antidiabetic prescriptions during the baseline period ($n = 58,642$), low estimated glomerular filtration rate (eGFR) ($n = 32,347$), or a history of cancer other than nonmelanoma skin cancer ($n = 118,245$) were excluded. The final analysis dataset included 463,248 individuals, which were randomly divided into a derivation dataset ($n = 308,832$) and a validation dataset ($n = 154,416$) in a 2:1 ratio. SKDB, Shizuoka Kokuho Database, eGFR: estimated glomerular filtration rate.

Variable	Category or unit	Without diabetes mellitus *n* = 256,680	With diabetes mellitus *n* = 52,152	*p*-value
Age	1 year	66.21 (11.71)	69.54 (9.62)	< 0.001
Age as categorical variable	< 50 years	29,134 (11.4)	1943 (3.7)	< 0.001
	50-<60 years	26,736 (10.4)	3696 (7.1)	
	60-<70 years	102,453 (39.9)	21,445 (41.1)	
	70-<80 years	67,406 (26.3)	16,931 (32.5)	
	≥ 80 years	30,951 (12.1)	8137 (15.6)	
Sex	Female	162,657 (63.4)	32,553 (62.4)	< 0.001
Health checkup				
BMI (kg/m^2)	< 18.5	26,492 (10.3)	4366 (8.4)	< 0.001
	18.5-<22	98,130 (38.3)	17,051 (32.7)	
	22-<25	83,113 (32.4)	17,710 (34.0)	
	25-<30	43,424 (16.9)	11,287 (21.7)	
	≥ 30	5375 (2.1)	1702 (3.3)	
Systolic blood pressure (mmHg)	< 120	73,880 (28.8)	11,482 (22.0)	< 0.001
	120-<130	63,647 (24.8)	12,733 (24.4)	
	130-<140	57,592 (22.4)	12,997 (24.9)	
	140-<150	32,593 (12.7)	7790 (14.9)	
	150-<160	15,966 (6.2)	3917 (7.5)	
	≥ 160	12,939 (5.0)	3212 (6.2)	
Diastolic blood pressure (mmHg)	< 70	75,990 (29.6)	14,140 (27.1)	< 0.001
	70-<80	89,990 (35.1)	18,220 (35.0)	
	80-<90	65,467 (25.5)	14,218 (27.3)	
	90-<100	19,578 (7.6)	4222 (8.1)	
	≥ 100	5581 (2.2)	1329 (2.5)	
Hematocrit (%)	≥ 50	4042 (1.9)	961 (2.2)	< 0.001
	40–50	138,036 (65.0)	27,692 (64.2)	
	30-<40	69,096 (32.6)	14,333 (33.2)	
	< 30	1100 (0.5)	177 (0.4)	
Hemoglobin (g/dL)	≥ 15.5	21,082 (9.9)	4381 (10.1)	0.504
	13.5-<15.5	90,784 (42.8)	18,344 (42.5)	
	11.5-<13.5	88,312 (41.6)	17,975 (41.6)	
	< 11.5	12,146 (5.7)	2475 (5.7)	
Erythrocyte count (10^4/μL)	≥ 550	2432 (1.1)	529 (1.2)	< 0.001
	450-<550	89,722 (42.2)	18,074 (41.9)	
	350-<450	116,610 (54.9)	23,687 (54.9)	
	< 350	3692 (0.7)	892 (2.1)	
Triglyceride (mg/dL)	< 100	142,793 (55.6)	25,661 (49.2)	< 0.001
	100-<150	67,987 (26.5)	14,980 (28.7)	
	≥ 150	45,861 (17.9)	11,503 (22.1)	
HDL cholesterol (mg/dL)	≥ 40	245,430 (95.6)	48,982 (93.9)	< 0.001
	< 40	11,208 (4.4)	3161 (6.1)	
LDL cholesterol (mg/dL)	< 120	110,372 (43.0)	22,476 (43.1)	0.022
	120-<140	64,818 (25.3)	12,893 (24.7)	
	≥ 140	81,408 (31.7)	16,774 (32.2)	
AST (IU/L)	< 30	223,423 (87.0)	43,504 (83.4)	< 0.001
	30-<50	29,187 (11.4)	7342 (14.1)	
	≥ 50	4067 (1.6)	1306 (2.5)	
ALT (IU/L)	< 30	229,118 (89.3)	44,912 (86.1)	< 0.001
	30-<50	21,246 (8.3)	5357 (10.3)	
	≥ 50	6313 (2.5)	1883 (3.6)	
γ-GTP (IU/L)	< 50	226,431 (88.2)	44,926 (86.2)	< 0.001
	50-<100	21,672 (8.4)	5022 (9.6)	
	≥ 100	8540 (3.3)	2194 (4.2)	

Continued

Variable	Category or unit	Without diabetes mellitus $n = 256,680$	With diabetes mellitus $n = 52,152$	p-value
Hemoglobin A1c (%)	< 5.5	104,516 (41.1)	13,647 (26.4)	< 0.001
	5.5-<6	129,659 (50.9)	26,029 (50.3)	
	6-<6.5	20,405 (8.0)	12,080 (23.3)	
eGFR (mL/min/1.73 m^2)	≥ 60 (G1-2)	199,850 (77.9)	38,319 (73.5)	< 0.001
	45-<60 (G3a)	48,945 (19.1)	11,501 (22.1)	
	30-<45 (G3b)	7885 (3.1)	2332 (4.5)	
Uric acid (mg/dL)	≥ 9	1371 (0.6)	381 (0.8)	< 0.001
	7–<9	21,154 (8.6)	5192 (10.3)	
	5–<7	105,720 (42.8)	23,143 (45.8)	
	< 5	118,741 (48.1)	21,850 (43.2)	
Urinary protein	–	222,284 (87.0)	44,040 (84.8)	< 0.001
	±	21,788 (8.5)	4706 (9.1)	
	≥+	11,291 (4.4)	3174 (6.1)	
Taking medication for common ailments				
Medication for hypertension	Presence	84,283 (32.8)	23,071 (44.2)	< 0.001
Medication for dyslipidemia	Presence	59,680 (23.3)	16,781 (32.2)	< 0.001
Lifestyle habits				
Habitual smoker[†]	Presence	31,753 (12.4)	5847 (11.2)	< 0.001
Exercise habits[‡]	Presence	116,220 (50.0)	23,708 (50.4)	0.096
≥ 10-kg weight gain since age 20	Presence	61,806 (26.6)	15,225 (32.3)	< 0.001
Heavy drinking[§]	Presence	11,750 (4.6)	2299 (4.4)	0.093

Table 1. Demographic profiles of patients in the derivation dataset ($n = 308,832$). *BMI* Body mass index, *HDL* High-density lipoprotein, *LDL* Low-density lipoprotein, *AST* Aspartate aminotransferase, *ALT* Alanine aminotransferase, *GTP* Glutamyl transpeptidase. [†]Habitual smoker was defined as a person who has smoked > 100 cigarettes or has been a habitual smoker for ≥ 6 months and smoked for > 1 month. [‡]Exercise habit was defined a person who performed exercise that involved slight sweating for 30 min or more at least twice a week for over 1 year. [§]Heavy drinking was defined as a person who consumes alcohol daily with a daily intake exceeding 360 mL.

Table 2 outlines the results of both univariable and multivariable regression analyses conducted on the derivation dataset. The univariate analysis identified several statistically significant predictors, including age, gender, and 16 health checkup parameters, such as BMI, systolic blood pressure, and diastolic blood pressure. Medications, specifically antihypertensive and lipid-lowering drugs, were also significant predictors. Additionally, lifestyle factors, such as exercise habits, weight gain of ≥ 10 kg after the age of 20, and heavy drinking emerged as significant predictors in the analyses.

Among the pairs of variables exhibiting an absolute Spearman correlation coefficient greater than 0.4 (Supplementary Table 5), those with p-values < 0.05 were included in the multivariable model. Significant correlations were observed between hematocrit, hemoglobin, and erythrocyte count, as well as between hemoglobin and sex. Additionally, sex correlated with urinary acid levels, prompting its selection for further analysis. Both systolic and diastolic blood pressures exhibited correlations; however, given its recognized role as a risk factor for lifestyle-related diseases such as cardiovascular incidents across all age groups and sexes[28], systolic blood pressure was selected for inclusion. Despite ongoing debate regarding the roles of systolic versus diastolic blood pressure in pathophysiology, the former is deemed more crucial in blood pressure assessment. Furthermore, BMI and weight gains of ≥ 10 kg after age of 20 were correlated, leading to the selection of BMI for this study as its measured values are considered more reliable than self-reported weight. A correlation was noted between AST and ALT, with both identified as risk factors for lifestyle-related diseases, including T2DM[29–33]. Research has shown that elevated AST levels can independently serve as a risk factor for such diseases[31–33]. Consequently, we chose to focus on AST in our study.

The multivariable regression analysis indicated that increasing age, male sex, a BMI over 22, systolic blood pressure ≥ 130 mmHg, triglyceride levels above 100 mg/dL, HDL cholesterol below 40 mg/dL, LDL cholesterol above 140 mg/dL, AST levels over 30 IU/L, g-GTP over 50 IU/L, HbA1c levels exceeding 5.5%, reduced eGFR below 60 mL/min/1.73 m^2, and the presence of urinary protein at levels of [≥+], along with medication use for hypertension and dyslipidemia, exercise habits, and heavy drinking were all associated with an increased risk of developing T2DM.

Predictive scoring system for the onset of type 2 diabetes mellitus

By applying HRs derived from the multivariable Cox regression model, we developed a scoring system that predicts the onset of T2DM. This system assigns weighted scores to each predictive factor, as outlined in Table 3. The distribution of scores is shown in Supplementary Fig. 1a (derivation) and 1c (validation). In the derivation

Variable (reference)	Category or unit	Univariable model			Multivariable model		
		HR	95% CI	*p*-value	HR	95% CI	*p*-value
Age	1 year	1.02	1.02–1.05	<0.001			
Age (<50)	50-<60 years	1.67	1.58-1.77	<0.001	1.47	1.39-1.55	<0.001
	60-<70 years	2.26	2.16-2.37	<0.001	1.84	1.75-1.93	<0.001
	70-<80 years	2.70	2.57-2.83	<0.001	2.16	2.06-2.28	<0.001
	≥80 years	2.82	2.68-2.96	<0.001	2.19	2.07-2.32	<0.001
Sex (women)	Men	1.16	1.14-1.18	<0.001	1.17	1.15-1.19	<0.001
Health checkup							
BMI (<18.kg/m^2)	18.5-<22	1.03	1.00-1.05	0.091	0.98	0.94-1.01	0.190
	22-<25	1.25	1.21-1.29	<0.001	1.04	1.00-1.08	0.038
	25-<30	1.56	1.51-1.62	<0.001	1.15	1.11-1.20	<0.001
	≥30	2.04	1.92-2.15	<0.001	1.48	1.39-1.57	<0.001
Systolic blood pressure (<120 mmHg)	120–<130	1.22	1.19-1.25	<0.001	1.01	0.98-1.04	0.496
	130–<140	1.34	1.30-1.37	<0.001	1.04	1.01-1.07	<0.001
	140–<150	1.42	1.38-1.46	<0.001	1.06	1.03-1.10	0.010
	150–<160	1.49	1.44-1.55	<0.001	1.10	1.06-1.14	<0.001
	≥160	1.59	1.53-1.65	<0.001	1.17	1.12-1.22	<0.001
Diastolic blood pressure (<70 mmHg)	70-<80	1.05	1.03-1.08	<0.001			
	80-<90	1.13	1.10-1.15	<0.001			
	90-<100	1.17	1.13-1.21	<0.001			
	≥100	1.38	1.30-1.46	<0.001			
Hematocrit (≥50%)	40–<50	0.74	0.69-0.79	<0.001			
	30–<40	0.72	0.68-0.77	<0.001			
	<30	0.72	0.62-0.85	<0.001			
Hemoglobin (≥15.5 g/dL)	13.5–<15.5	0.90	0.87-0.93	<0.001			
	11.5–<13.5	0.85	0.83-0.88	<0.001			
	<11.5	0.92	0.87-0.96	<0.001			
Erythrocyte count (≥550 10^4/μL)	450–<550	0.82	0.75-0.89	<0.001			
	350–<450	0.76	0.70-0.83	<0.001			
	<350	0.99	0.89-1.11	0.884			
Triglyceride (<100 mg/dL)	100–<150	1.20	1.18-1.23	<0.001	1.03	1.00-1.05	0.029
	≥150	1.40	1.37-1.43	<0.001	1.09	1.06-1.11	<0.001
HDL cholesterol (≥40 mg/dL)	<40	1.43	1.38-1.49	<0.001	1.11	1.07-1.16	<0.001
LDL cholesterol (<120 mg/dL)	120–<140	0.95	0.93-0.97	<0.001	0.98	0.96-1.01	0.204
	140–	1.00	0.98-1.02	0.787	1.04	1.02-1.06	0.001
AST (<30 IU/L)	30–<50	1.29	1.25-1.32	<0.001	1.11	1.08-1.14	<0.001
	≥50	1.75	1.66-1.85	<0.001	1.37	1.29-1.46	<0.001
ALT (<30 IU/L)	30–<50	1.34	1.30-1.38	<0.001			
	≥50	1.70	1.62-1.78	<0.001			
γ–GTP (<50 IU/L)	50–<100	1.24	1.20-1.27	<0.001	1.07	1.04-1.11	<0.001
	≥100	1.41	1.35-1.47	<0.001	1.23	1.17-1.29	<0.001
HbA1c (<5.5%)	5.5-<6	1.48	1.45-1.51	<0.001	1.39	1.36-1.42	<0.001
	6–<6.5	4.13	4.03-4.24	<0.001	3.67	3.58-3.77	<0.001
eGFR (≥60 mL/min/1.73 m^2)	45–<60 G3a	1.20	1.17-1.22	<0.001	1.01	0.99-1.03	0.424
	30–<45 G3b	1.56	1.49-1.62	<0.001	1.16	1.11-1.22	<0.001
Uric acid (>9 mg/dL)	7–<9	0.81	0.73-0.90	<0.001			
	5–<7	0.68	0.62-0.76	<0.001			
	<5	0.56	0.50-0.62	<0.001			
Urinary protein (–)	±	1.13	1.10-1.17	<0.001	1.07	1.04-1.11	<0.001
	>+	1.50	1.44-1.55	<0.001	1.24	1.19-1.29	<0.001
Taking medication for common ailments							
Medication for hypertension (absence)	Presence	1.49	1.47-1.52	<0.001	1.17	1.15-1.20	<0.001
Medication for dyslipidemia (absence)	Presence	1.40	1.37-1.42	<0.001	1.13	1.10-1.15	<0.001
Lifestyle habits							
Habitual smoker (absence)	Presence	1.02	0.99-1.05	0.228			
Exercise habits	Presence	1.04	1.02-1.06	<0.001	1.05	1.03-1.07	<0.001

Continued

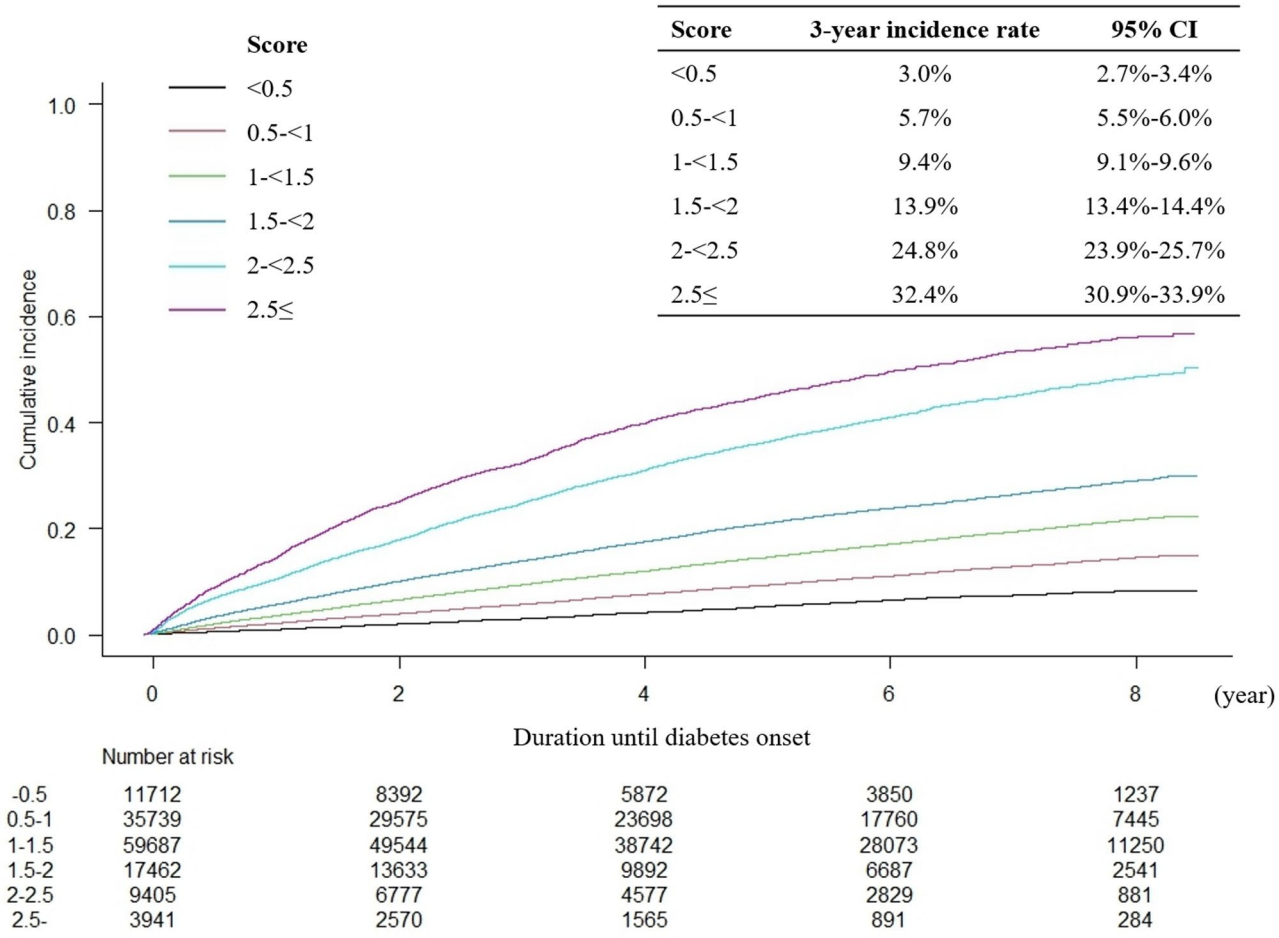
Variable (reference)	Category or unit	Univariable model			Multivariable model		
		HR	95% CI	*p*-value	HR	95% CI	*p*-value
≥10-kg weight gain since age 20	Presence	1.37	1.35-1.40	<0.001			
Heavy drinking (absence)	Presence	1.05	1.01-2.09	0.029	1.08	1.04-1.14	0.001

Table 2. Results from univariable and multivariable Cox proportional hazards regression analyses for type 2 diabetes mellitus. *HR* Hazard ratio, *CI* Confidence interval, *BMI* Body mass index, *HDL* High-density lipoprotein, *LDL* Low-density lipoprotein, *AST* Aspartate aminotransferase, *ALT* Alanine aminotransferase, *GTP* Glutamyl transpeptidase.

Variable (unit)	Category	Score
Age	40-<50 years	0
	50-<60 years	0.384
	60-<70 years	0.609
	70-<80 years	0.772
	≥ 80 years	0.783
Sex	Men	0.156
Health checkup		
BMI (kg/m^2)	18.5-<22	−0.024
	22-<25	0.039
	25-<30	0.14
	≥30	0.39
Systolic blood pressure (mmHg)	120–<130	0.009
	130–<140	0.036
	140–<150	0.059
	150–<160	0.093
	≥ 160	0.16
Triglyceride (mg/dL)	100–<150	0.025
	≥ 150	0.083
HDL cholesterol (mg/dL)	< 40	0.107
LDL cholesterol (mg/dL)	120–<140	−0.015
	140–	0.037
AST (IU/L)	30–<50	0.101
	≥ 50	0.314
γ–GTP (IU/L)	50–<100	0.07
	≥ 100	0.205
Hemoglobin A1c (%)	5.5-<6	0.329
	6–<6.5	1.301
eGFR (mL/min/1.73 m^2)	45–<60 G3a	0.01
	30–<45 G3b	0.149
Urine protein	±	0.069
	≥+	0.212
Therapeutic medication		
Antihypertensive medication	Presence	0.157
Dyslipidemia medication	Presence	0.120
Lifestyle habits		
Heavy drinking	Presence	0.081
Exercise habits	Presence	0.046

Table 3. Predictive scores for the onset of type 2 diabetes mellitus. *BMI* Body mass index, *HDL* High-density lipoprotein, *LDL* Low-density lipoprotein, *AST* Aspartate aminotransferase, *ALT* Alanine aminotransferase, *GTP* Glutamyl transpeptidase.

cohort, the median predictive index was 1.163 (range − 0.039 to 4.039), whereas in the validation cohort, it was 1.165 (range − 0.039 to 3.595). The distribution of these predictive scores across individuals in the derivation and validation datasets is shown in Supplementary Fig. 1. In both datasets, an increase in the predictive score correlates with a higher proportion of T2DM diagnoses (Supplementary Fig. 1b and 1 d), demonstrating good



Score	3-year incidence rate	95% CI
<0.5	3.0%	2.7%-3.4%
0.5-<1	5.7%	5.5%-6.0%
1-<1.5	9.4%	9.1%-9.6%
1.5-<2	13.9%	13.4%-14.4%
2-<2.5	24.8%	23.9%-25.7%
2.5≤	32.4%	30.9%-33.9%

Fig. 3. Cumulative incidence curves for the onset of diabetes mellitus classified by score category from validation dataset. Patients were classified into the following score categories: less than 0.5, 0.5 to less than 1, 1 to less than 1.5, 1.5 to less than 2, 2 to less than 2.5, and 2.5 or higher.

calibration between predicted and observed risks. For the derivation dataset, the predictive score yielded an HR of 2.64 (95% CI, 2.60–2.68) and a c-index of 0.652 (95% CI, 0.650–0.654). Additionally, the cumulative incidence curves for the onset of T2DM, segmented by predictive score, are presented in Supplementary Fig. 2, demonstrating a consistent rise in incidence with increasing scores.

In the validation dataset, which comprised 154,416 individuals, the predictive score attained a c-index of 0.656 (95% CI, 0.652–0.659) with an HR of 2.70 (95% CI, 2.64–2.76). The associated cumulative incidence curves are shown in Fig. 3. Cumulative incidence rates of T2DM up to three years, classified by predictive score, were as follows: 3.0% (95% CI, 2.7–3.4) for scores below 0.5, 5.7% (95% CI, 5.5–6.0) for scores ranging from 0.5 to below 1, 9.4% (95% CI, 9.1–9.6) for scores from 1 to below 1.5, 13.9% (95% CI, 13.4–14.4) for scores from 1.5 to below 2, 24.8% (95% CI, 23.9–25.7) for scores from 2 to below 2.5, and 32.4% (95% CI, 30.9–33.9) for scores 2.5 and above. Incidence rates after one and five years are presented in Supplementary Table 6. As shown in Fig. 3, higher predictive scores were consistently associated with greater cumulative incidence, demonstrating clear risk stratification across categories.

Discussion

Our study has developed a comprehensive predictive tool through meticulous analysis of a large dataset comprising 463,248 patients from 2012 to 2021. This tool aims to identify individuals at imminent risk of developing T2DM. Compared with previous Japanese studies[7,16,29], the originality of our work lies in its broader and more heterogeneous cohort, inclusion of older adults, and the use of Cox regression analysis to appropriately account for censoring. By employing categorical predictors and excluding individuals with baseline HbA1c ≥ 6.5%, we targeted true incident cases and enhanced applicability to both clinical and public health practice. The resulting score-based model is simple, provides absolute 1-, 3-, and 5-year risks, and can be used without specialized computational tools.

Previous studies have identified multiple risk factors for the onset of T2DM, such as male sex, older age, elevated BMI, high SBP, increased liver enzymes, elevated HbA1c levels, urinary protein, dyslipidemia, hypertension, and exercise habits[7,29,30,34–39]. and our results are considered reasonable. These factors are

well-documented across numerous studies and are in line with the risk factors listed in the Japanese diabetes guidelines[40]. However, a significant concern in Japan is the lack of comprehensive assessments of individual risks for T2DM, despite known ethnic and regional variations in T2DM susceptibility[41]. Our study addresses this gap by reassessing these risk factors and incidence rates within a large-scale cohort in Japan, assigning weights to each factor based on their contribution to T2DM onset.

Although several studies have previously developed predictive models for T2DM, they often involved smaller sample sizes, focused on specific subgroups, or had limited generalizability[7,12,29,42,43]. In particular, Nanri et al. and Sasagawa et al. reported models based on occupational health screening data[7,29], whereas our study utilized health check-up data from the general population, thereby including a broader age distribution and older individuals. Furthermore, we applied stricter exclusion criteria (e.g., pre-existing T2DM or HbA1c ≥ 6.5%), which may have influenced the baseline risk distribution and contributed to differences in model performance. Another distinction is that while prior studies incorporated lifestyle questionnaire data, our model was constructed using only routinely collected health check-up variables, enhancing its practical applicability but potentially reducing discriminatory power. Finally, our analytic framework employed cause-specific Cox proportional hazards models with competing risk methodology, in contrast to conventional Cox models used previously. These methodological and population-level differences likely account for the observed variation in discriminatory ability. Furthermore, we evaluated the calibration of the proposed predictive model. The calibration plots demonstrated good agreement between predicted and observed risks, thereby supporting the external validity of the model. Taken together, these findings highlight that direct comparisons across models should be interpreted cautiously, but they also underscore the robustness and real-world relevance of our approach.

In addition to these strengths, it is important to consider how our study differs from recently published models, such as that of Xu et al.[16]. While Xu et al. applied logistic regression and restricted the analysis to participants with at least five years of follow-up, our study employed a Cox proportional hazards model to fully utilize time-to-event information and include all eligible participants at baseline. The former approach may introduce selection bias by excluding individuals with shorter follow-up, whereas our survival-based framework appropriately handles censoring and provides incidence-based risk prediction. These methodological differences limit direct comparison of C-index values between studies. Nonetheless, both Xu et al. and our study identified similar predictors, including age, BMI, and HbA1c, underscoring the robustness of these risk factors across different analytic strategies.

T2DM is known as a preventable disease[9], and the benefits of prevention or early intervention have been well-documented[44]. These considerations extend beyond the non-elderly population, as the prevalence of T2DM among the elderly is on the rise globally[6]. In this context, prevention strategies for older adults are particularly valuable. A key aspect of our study is the inclusion of older adults in the development of predictive models. The benefits of predicting and preventing T2DM among the elderly are as follows: First, timely interventions can mitigate the progression of T2DM and its complications, such as cardiovascular disease and renal impairment, thereby enhancing quality of life for elderly individuals[45]. Second, early prediction of T2DM can identify at-risk individuals, facilitating interventions that prevent the onset of frailty—a crucial factor in maintaining independence and alleviating healthcare burdens[46]. Third, by preventing the complications associated with poorly managed T2DM, a considerable reduction in healthcare costs can be achieved[47]. Therefore, the development of predictive models that incorporate older adults not only enhances preventive care but also helps in circumventing the costly treatments required for advanced T2DM complications.

Recent literature has advocated for the development of prediction models[48,49]. We employed a random splitting method to construct our prediction model. Collins et al. have highlighted potential drawbacks of split-sample method, including reduced sample sizes and the risk of cherry-picking through repeated random splits to achieve favorable outcomes[48]. However, our study mitigates these concerns with a large dataset and a single instance of random sampling. Furthermore, Dhiman et al., underscore that the split-sample method continues to be widely used in developing prediction models[49], reinforcing the validity of our methodology.

In Japan, two reports from the same cohort have presented data on T2DM prediction models. Nanri et al. achieved C-indexes of 0.717 with a non-invasive model and 0.893 with an invasive model[7], while Sasagawa et al. reported a C-index of 0.872[29]. Notably, Nanri et al. did not include survival time at the three-year mark in their model, whereas Sasagawa et al. employed survival analysis in their approach. However, the methods used to calculate the C-index for evaluating predictive performance in these studies are not described, making direct comparisons with our work challenging. Our model encompasses the entire study period and utilizes Harrell's C-index for assessing predictive performance. Given these differences, a direct comparison of predictive accuracy between our model and those from the previous studies is not feasible.

Our predictive model differs from previous studies by Nanri et al.[7] and Sasagawa et al.[29] in key aspects. While it demonstrated lower discrimination than those models, this likely reflects differences in study populations, inclusion criteria, available predictors, and analytic methods. Although direct comparisons are difficult, a notable strength of our study is that it is based on a general population including older adults, making it broadly applicable at the population level. We excluded individuals with HbA1c ≥ 6.5% at baseline to focus on those without diagnosed diabetes, enhancing early detection and intervention. Unlike prior studies targeting younger occupational cohorts, our study included a general population aged 40 and above, with many elderly participants, making our model applicable to a broader demographic. Additionally, our model utilizes data readily available from routine health check-ups without requiring complex calculations, increasing its practicality for widespread clinical use.

Machine learning and deep learning are increasingly utilized in the development of predictive models[12,29,42,43]. Sasagawa et al. reported improved prediction accuracy with deep learning compared with traditional Cox proportional hazards models[29]. Specifically, Cox proportional hazards-based models yielded an area under the

receiver-operating characteristic curve (AUC) of 0.872 (95% CI; 0.858–0.886) for predicting T2DM mellitus. In contrast, DeepSurv-based models achieved a slightly higher AUC of 0.878 (95% CI; 0.864–0.892), indicating marginally better reclassification performance. However, the minimal difference in AUC suggests that the benefits of using DeepSurv over Cox proportional hazards models may be limited, suggesting that the latter could be adequately effective for constructing predictive models.

In developing the prediction model, we also considered constructing sex-specific models. However, stratification by sex substantially reduced the sample size in each subgroup, which did not improve the C-index and instead widened its standard error. Given our objective of creating a simple and broadly applicable screening tool for the general population, we prioritized a unified model. Importantly, sex was included as a predictor in this unified model, ensuring that sex-related differences in risk were appropriately accounted for.

As this is an observational study, conventional a priori power calculations are not directly applicable. Following contemporary recommendations[50], we justified the adequacy of the sample size not by relying on arbitrary thresholds, but primarily through the precision of model performance estimates. Specifically, the narrow 95% confidence intervals around the c-index in both the derivation and validation datasets indicate that the estimates are stable and reproducible. In addition, we conducted post hoc power analyses (Supplementary Table 7), which demonstrated that strong to moderate associations remain well detectable even at low exposure prevalences, whereas weak associations become more difficult to detect in extremely rare exposures. Taken together, these findings support that the study population is sufficiently large and appropriate to ensure robust and reliable conclusions.

Our study presents 1-year, 3-year, and 5-year incidence rates for each predictive score category, calculated using cumulative incidence functions that account for the competing risk of death. Figure 3 illustrates the cumulative incidence of T2DM by predictive score category, showing clear stratification of risk. This finding highlights the clinical utility of our model by providing easily interpretable probabilities that reflect different levels of diabetes risk. Such visual representation reinforces the practical application of our risk score, supporting clinicians in stratifying patients according to their risk and enabling individuals to better recognize their own risk level. By presenting incidence rates in an accessible manner, our results may encourage more proactive health management and lifestyle modifications.

In our study, we also evaluated the potential benefit of constructing sex-specific models. While men showed a slightly higher C-index (0.673, 95% CI: 0.669–0.677) compared with women (0.647, 95% CI: 0.643–0.651), these values were largely comparable to that of the unified model (0.652, 95% CI: 0.650–0.654). Importantly, the confidence intervals overlapped substantially, indicating no statistically or clinically meaningful improvement in discrimination. Moreover, stratification by sex reduced the sample size of each subgroup, resulting in wider confidence intervals and decreased stability of the estimates. Consistent with established principles of prediction model development, subgroup-specific models are generally recommended only when they provide clear and reproducible improvements in predictive performance[51]. Given our aim of developing a simple, broadly applicable risk scoring system, we therefore prioritized a unified model while incorporating sex as a predictor to account for sex-related differences in risk.

In this study, we included adults aged 40 and above to develop a predictive model applicable to a broad population, encompassing both middle-aged and elderly individuals. While stratifying the analysis by variables such as age groups or sex could highlight subgroup-specific risk factors, it would reduce the sample size within each subgroup, potentially compromising the predictive performance and robustness of the model. While a modified stratification strategy may need to be considered based on the method proposed by Yong et al.[52], we believe that the current unified model is the most appropriate approach, balancing simplicity and predictive performance. Therefore, we opted for a unstratified model that includes all eligible participants, enhancing the practical applicability of our findings in diverse clinical settings.

Family history is a well-recognized risk factor for T2DM[53–56]. The absence of this variable in our study may have reduced the model's discriminative ability, as genetic predisposition could not be accounted for. Nevertheless, our model maintains broad applicability by relying exclusively on routinely collected health check-up data, which facilitates its feasibility and scalability for large populations. Future studies integrating family history with health check-up information may further improve prediction performance and enhance personalized risk stratification.

While our study offers valuable insights, it is important to recognize its limitations. First, the geographical scope of our study was limited, as the database only included a subset of residents from Shizuoka prefecture. Given that susceptibility to T2DM varies by ethnicity[36,57], these results should be applied with caution to populations other than Japanese. Additionally, our participant pool was limited to individuals enrolled in NHI and LSEMCS, all of whom were undergoing health screenings. Consequently, generalizing these findings to a broader population may not be advisable. Second, since our analysis was restricted to individuals aged 40 and above, which may limit the applicability of our results to younger demographics. Third, we did not include family history of T2DM in our current risk model because this information was not collected in the database. However, it is noted that adults with a family history of T2DM tend to have higher BMI, waist circumference, blood pressure, and glucose levels compared with those without[53–56]. Nanri et al. also noted that the predictive power of risk models for T2DM that exclude family history was comparable to those that include it[7], suggesting that this omission may not compromise the model's effectiveness. Furthermore, our dataset lacked information on pregnancy, infertility treatments, and oral contraceptive pill (OCP) use because these are not covered by public health insurance in Japan. However, since our participants were all aged 40 and above, with many being elderly, the impact of these factors is likely minimal. Additionally, although steroid use can increase the risk of T2DM, we couldn't accurately identify steroid use from the claims data due to lack of detailed information. Despite this, we found that diseases commonly treated with steroids, such as collagen diseases, were linked to a higher risk of T2DM, indirectly accounting for thetential effects of steroid use. Fourth, the 95% confidence

intervals for T2DM incidence rates presented in Fig. 3, Supplementary Fig. 2, and Supplementary Table 6 are intended for estimation rather than prediction. Therefore, when applying the predictive model in clinical settings, this distinction must be considered. Fifth, the present scoring system is based on coefficients with decimal values, which may be cumbersome for manual calculation in clinical practice. While conversion to integer scores could improve usability, it may also reduce model performance; therefore, we did not adopt this approach. This remains a limitation of our study and warrants consideration for future refinements aimed at clinical applicability. Despite these limitations, this study suggests that predicting individual risks of T2DM onset for health screening participants could transform health awareness and facilitate early intervention in T2DM management.

Conclusions

The predictive model developed in this study serves as a valuable tool for the early identification of individuals at high risk for T2DM. Although the model demonstrates reliable predictive accuracy suitable for practical use, future research should aim to broaden its demographic applicability and improve its predictive capacity by incorporating more risk factors. Early identification of at-risk individuals can facilitate timely preventive measures, which could substantially reduce the incidence and burden of T2DM.

Data availability

The datasets generated and analyzed during the current study are not publicly available due to data use restrictions imposed by Shizuoka Prefecture, but are available from the corresponding author on reasonable request and with permission of Shizuoka Prefecture.

Received: 13 January 2025; Accepted: 24 September 2025
Published online: 30 October 2025

References

1. Cho, N. et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **138**, 271–281 (2018).
2. Henning, R. J. Type-2 diabetes mellitus and cardiovascular disease. *Future Cardiol.* **14**, 491–509 (2018).
3. Beckman, J. A., Creager, M. A. & Libby, P. Diabetes and atherosclerosis: epidemiology, pathophysiology, and management. *JAMA* **287**, 2570–2581 (2002).
4. Song, S. et al. Long-Term diabetes mellitus is associated with an increased risk of pancreatic cancer: A Meta-Analysis. *PLoS One*. **10**, e0134321 (2015).
5. Bommer, C. et al. Global economic burden of diabetes in adults: projections from 2015 to 2030. *Diabetes Care*. **41**, 963–970 (2018).
6. Kalyani, R. R., Golden, S. H. & Cefalu, W. T. Diabetes and aging: unique considerations and goals of care. *Diabetes Care*. **40**, 440–443 (2017).
7. Nanri, A. et al. Development of risk score for predicting 3-Year incidence of type 2 diabetes: Japan epidemiology collaboration on occupational health study. *PLoS One*. **10**, e0142779 (2015).
8. The National Health and Nutrition Survey Japan. *Ministry of Health, Labour and Walfare.* (2019). https://www.who.int/news-room/fact-sheets/detail/diabetes (2019).
9. American Diabetes Association. Standards of medical care in Diabetes-2019 abridged for primary care providers. *Clin. Diabetes*. **37**, 11–34 (2019).
10. Leal, J. et al. Benchmarking the Cost-Effectiveness of interventions delaying diabetes: A simulation study based on NAVIGATOR data. *Diabetes Care*. **43**, 2485–2492 (2020).
11. Bhanpuri, N. H. et al. Estimated reduction in medication cost during first year of a continuous care intervention for treatment of type 2 diabetes. *Value Health*. **21**, S73 (2018).
12. Lai, H., Huang, H., Keshavjee, K., Guergachi, A. & Gao, X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr. Disord*. **19**, 101 (2019).
13. Vettoretti, M. et al. Addressing practical issues of predictive models translation into everyday practice and public health management: a combined model to predict the risk of type 2 diabetes improves incidence prediction and reduces the prevalence of missing risk predictions. *BMJ Open. Diabetes Res. Care* **8**, e001223 (2020).
14. Yokota, N. et al. Predictive models for conversion of prediabetes to diabetes. *J. Diabetes Complications*. **31**, 1266–1271 (2017).
15. Wang, H. et al. A retrospective population study to develop a predictive model of prediabetes and incident type 2 diabetes mellitus from a hospital database in Japan between 2004 and 2015. *Med. Sci. Monit.* **26**, e920880 (2020).
16. Xu, J. et al. Development and validation of prediction models for the 5-year risk of type 2 diabetes in a Japanese population: Japan public health Center-based prospective (JPHC) diabetes study. *J. Epidemiol.* **34**, 170–179 (2024).
17. Nakatani, E., Tabara, Y., Sato, Y., Tsuchiya, A. & Miyachi, Y. Data resource profile of shizuoka kokuho database (SKDB) using integrated health- and care-insurance claims and health checkups: the shizuoka study. *J. Epidemiol. advpub* https://doi.org/10.2188/jea.JE20200480 (2021).
18. Shoji-Asahina, A. et al. Risk factors, treatment and survival rates of late-onset acquired haemophilia A: A cohort study from the Shizuoka Kokuho database. *Haemophilia* **29**, 799–808 (2023).
19. Hashizume, H., Nakatani, E., Sasaki, H. & Miyachi, Y. Hydrochlorothiazide increases risk of nonmelanoma skin cancer in an elderly Japanese cohort with hypertension: the Shizuoka study. *JAAD Int.* **12**, 49–57 (2023).
20. Ubukata, N., Nakatani, E., Hashizume, H., Sasaki, H. & Miyachi, Y. Risk factors and drugs that trigger the onset of Stevens–Johnson syndrome and toxic epidermal necrolysis: A population-based cohort study using the Shizuoka Kokuho database. *JAAD Int.* **11**, 24–32 (2023).
21. Quan, H. et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care*. **43**, 1130–1139 (2005).
22. Elixhauser, A., Steiner, C., Harris, D. R. & Coffey, R. M. Comorbidity measures for use with administrative data. *Med. Care*. **36**, 8–27 (1998).
23. Ono, Y., Taneda, Y., Takeshima, T., Iwasaki, K. & Yasui, A. Validity of claims diagnosis codes for cardiovascular diseases in diabetes patients in Japanese administrative database. *Clin. Epidemiol.* **12**, 367–375 (2020).
24. Nagai, Y., Kazumori, K., Takeshima, T., Iwasaki, K. & Tanaka, Y. A claims database analysis of dose-dependency of Metformin and incidence of lactic acidosis in Japanese patients with type 2 diabetes. *Diabetes Ther.* **12**, 1129–1141 (2021).
25. Sato, S. et al. High mean corpuscular volume as a predictor of esophageal cancer: A cohort study based on the Japanese Shizuoka Kokuho database. *PLoS One*. **20**, e0318791 (2025).

26. Harrell, F. E. Jr, Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
27. Kanda, Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transpl.* **48**, 452–458 (2012).
28. Kannel, W. B. Historic perspectives on the relative contributions of diastolic and systolic blood pressure elevation to cardiovascular risk profile. *Am. Heart J.* **138**, S205–S210 (1999).
29. Sasagawa, Y. et al. Application of deep neural survival networks to the development of risk prediction models for diabetes mellitus, hypertension, and dyslipidemia. *J. Hypertens.* **42**, 506–514 (2024).
30. Ahn, H. R. et al. The association between liver enzymes and risk of type 2 diabetes: the Namwon study. *Diabetol. Metab. Syndr.* **6**, 14 (2014).
31. Ruban, A. et al. Liver enzymes and risk of stroke: the atherosclerosis risk in communities (ARIC) study. *J. Stroke Cerebrovasc. Dis.* **22**, 357–368 (2020).
32. Li, J. et al. Liver enzymes, alcohol consumption and the risk of diabetes: the suita study. *Acta Diabetol.* **59**, 1531–1537 (2022).
33. Ndrepepa, G. Aspartate aminotransferase and cardiovascular disease—a narrative review. *J. Lab. Precis Med.* **6**, 6–6 (2021).
34. Collins, G. S., Mallett, S., Omar, O. & Yu, L. M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* **9**, 103 (2011).
35. Aekplakorn, W. et al. A risk score for predicting incident diabetes in the Thai population. *Diabetes Care.* **29**, 1872–1877 (2006).
36. Schulze, M. B. et al. Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam study. *Diabetes Care.* **32**, 2116–2119 (2009).
37. Doi, Y. et al. Two risk score models for predicting incident type 2 diabetes in Japan. *Diabet. Med.* **29**, 107–114 (2012).
38. Heianza, Y. et al. Development of a new scoring system for predicting the 5 year incidence of type 2 diabetes in japan: the Toranomon hospital health management center study 6 (TOPICS 6). *Diabetologia* **55**, 3213–3223 (2012).
39. Gress, T. W., Nieto, F. J., Shahar, E., Wofford, M. R. & Brancati, F. L. Hypertension and antihypertensive therapy as risk factors for type 2 diabetes mellitus. Atherosclerosis risk in communities study. *N Engl. J. Med.* **342**, 905–912 (2000).
40. Araki, E. et al. Japanese clinical practice guideline for diabetes 2019. *J. Diabetes Investig.* **11**, 1020–1076 (2020).
41. Walker, R. J., Williams, S., Egede, L. E. & J. & Influence of Race, ethnicity and social determinants of health on diabetes outcomes. *Am. J. Med. Sci.* **351**, 366–373 (2016).
42. Choi, B. G. et al. 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.* **60**, 191–199 (2019).
43. Farran, B. et al. Use of Non-invasive parameters and Machine-Learning algorithms for predicting future risk of type 2 diabetes: A retrospective cohort study of health data from Kuwait. *Front. Endocrinol.* **10**, 624 (2019).
44. Twigg, S. M. et al. Prediabetes: a position statement from the Australian diabetes society and Australian diabetes educators association. *Med. J. Aust.* **186**, 461–465 (2007).
45. Twito, O., Frankel, M. & Nabriski, D. Impact of glucose level on morbidity and mortality in elderly with diabetes and pre-diabetes. *World J. Diabetes.* **6**, 345–351 (2015).
46. García-Esquinas, E. et al. Diabetes and risk of frailty and its potential mechanisms: A prospective cohort study of older adults. *J. Am. Med. Dir. Assoc.* **16**, 748–754 (2015).
47. Forbes, A., Murrells, T. & Sinclair, A. J. Examining factors associated with excess mortality in older people (age ≥ 70 years) with diabetes - a 10-year cohort study of older people with and without diabetes. *Diabet. Med.* **34**, 387–395 (2017).
48. Collins, G. S. et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384**, e074819 (2024).
49. Dhiman, P. et al. Prediction model protocols indicate better adherence to recommended guidelines for study conduct and reporting. *J Clin. Epidemiol* **169**, 111287 (2024).
50. Riley, R. D. et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit. Health.* **7**, 100857 (2025).
51. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach To Development, Validation, and Updating,* Springer Nature, (2019).
52. Yong, F. H., Tian, L., Yu, S., Cai, T. & Wei, L. J. Optimal stratification in outcome prediction using baseline information. *Biometrika* **103**, 817–828 (2016).
53. Bianco, A. et al. The surprising influence of family history to type 2 diabetes on anaerobic performance of young male élite athletes. *Springerplus* **3**, 224 (2014).
54. Prasad, D. S., Kabir, Z., Dash, A. K. & Das, B. C. Prevalence and risk factors for diabetes and impaired glucose tolerance in Asian indians: a community survey from urban Eastern India. *Diabetes Metab. Syndr.* **6**, 96–101 (2012).
55. Pomara, F., Russo, G. & Gravante, G. Influence of family history to type 2 diabetes on the body composition and homeostasis model assessment: a comparison between young active and sedentary men. *Minerva Med.* **97**, 379–383 (2006).
56. Zamora-Ginez, I. et al. Risk factors for diabetes, but not for cardiovascular disease, are associated with family history of type 2 diabetes in subjects from central Mexico. *Ann. Hum. Biol.* **39**, 102–107 (2012).
57. Ali, O. Genetics of type 2 diabetes. *World J. Diabetes.* **4**, 114–123 (2013).

Acknowledgements

A database from the Japan Pharmaceutical Information Center was used for the drug code search. We thank Phoebe Chi, MD, from Edanz (https://jp.edanz.com/ac), for editing a draft of this manuscript.

Author contributions

Study conception and design: T.S., E.N., and T.U. Acquisition of data: E.N. Analysis and interpretation of data: T.S. and E.N. Drafting of the work: T.S. Critical revision of the manuscript: T.S., E.N., H.A., C.S., Y.O., E.O., H.I., K.H., and T.U. Final approval of the manuscript: T.S., E.N., H.A., C.S., Y.O., E.O., H.I., K.H., and T.U.

Declarations

Competing interests

The authors declare no competing interests.

Ethics

The study protocol received approval from the ethics committee at Shizuoka Graduate University of Public Health (#SGUPH2021_001_065 – 2), ensuring compliance with all applicable ethical standards. Furthermore, we adhered strictly to the guidelines and regulations approved for this research. Due to the retrospective nature

of the study, the ethics committee at Shizuoka Graduate University of Public Health (#SGUPH2021_001_065 – 2) waived the need of obtaining informed consent. Before our access, all patient data were thoroughly anonymized, ensuring the privacy and confidentiality of the subjects[1].
The authors declare no conflicts of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-21831-8.

Correspondence and requests for materials should be addressed to E.N.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025