DOCTORAL　THESIS

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

（5年間の2型糖尿病罹患リスクの予測モデルの開発および検証：多目的コホート糖尿病研究）

September, 2023

（2023年9月）

Juan Xu

徐　娟

Endocrinology and Metabolism

Yokohama City University Graduate School of Medicine

横浜市立大学 大学院医学研究科 医科学専攻 分子内分泌・糖尿病内科学

(Research Supervisor：Atsushi Goto, Professor)

ﾃﾞｰﾀｻｲｴﾝｽ研究科/医学研究科　公衆衛生学

（研究指導教員：後藤　温 教授）

(Doctoral Supervisor: Yasuo Terauchi, Professor)

（指導教員：寺内　康夫 教授）

Received May 13, 2022; accepted April 10, 2023; released online May 20, 2023

Original Article

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

Juan Xu[a], Atsushi Goto[b], Maki Konishi[c], Masayuki Kato[d], Tetsuya Mizoue[c], Yasuo Terauchi[a], Shoichiro Tsugane[e,f], Norie Sawada[e], Mitsuhiko Noda[g], for the JPHC Study Group[†]

[a]Department of Endocrinology and Metabolism, Graduate School of Medicine, Yokohama City University, Yokohama, Japan.

[b]Department of Health Data Science, Graduate School of Data Science, Yokohama City University, Yokohama, Japan.

[c]Department of Epidemiology and Prevention, Center for Clinical Sciences, National Center for Global Health and Medicine, Tokyo, Japan.

[d]Health Management Center and Diagnostic Imaging Center, Toranomon Hospital, Tokyo, Japan.

[e]Division of Cohort Research, National Cancer Center Institute for Cancer Control, Chuo-ku, Tokyo, Japan.

[f]National Institute of Health and Nutrition, National Institutes of Biomedical Innovation, Health and Nutrition, Shinjuku-ku, Tokyo, Japan.

[g]Department of Diabetes, Metabolism and Endocrinology, Ichikawa Hospital, International University of Health and Welfare, Ichikawa, Japan.

[†]Japan Members listed in http://epi.ncc.go.jp/en/jphc/781/3838.html.

Correspondence:

Atsushi Goto, MD, Ph.D., MPH

Department of Health Data Science,

Graduate School of Data Science,

Yokohama City University

22-2 Seto, Kanazawa-Ku,

Yokohama 236-0027, Japan

E-mail: agoto@yokohama-cu.ac.jp

Running Title:

Prediction models for incidence of type 2 diabetes

Numbers of Tables: 3

Numbers of Figures: 3

Numbers of Supplemental materials: 2

ABSTRACT

Background: This study aimed to develop models to predict the 5-year incidence of T2DM in a Japanese population and validate them externally in an independent Japanese population.

Methods: Data from 10,986 participants (aged 46–75 years) in the development cohort of the Japan Public Health Center-based Prospective Diabetes Study and 11,345 participants (aged 46–75 years) in the validation cohort of the Japan Epidemiology Collaboration on Occupational Health Study were used to develop and validate the risk scores in logistic regression models.

Results: We considered non-invasive (sex, body mass index, family history of diabetes mellitus, and diastolic blood pressure) and invasive (glycated hemoglobin [HbA1c] and fasting plasma glucose [FPG]) predictors to predict the 5-year probability of incident diabetes. The area under the receiver operating characteristic curve was 0.643 for the non-invasive risk model, 0.786 for the invasive risk model with HbA1c but not FPG, and 0.845 for the invasive risk model with HbA1c and FPG. The optimism for the performance of all models was small by internal validation. In the internal-external cross-validation, these models tended to show similar discriminative ability across different areas. The discriminative ability of each model was confirmed using external validation datasets. The invasive risk model with only HbA1c was well-calibrated in the validation cohort.

Conclusions: Our invasive risk models are expected to discriminate between high- and low-risk individuals with T2DM in a Japanese population.

Keywords: Diabetes, risk score, prediction model, Japanese population, Japan Public Health Center-based Prospective (JPHC) Study

1. Introduction

Diabetes mellitus (DM) is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both.[1] According to the International Diabetes Federation, the global prevalence of diabetes in 2021 was estimated to be 10.5% (537 million people) and was expected to rise to 12.2% (783 million) by 2045.[2] Diabetes is thought to be one of the top ten causes of adult death.[3] In Japan, because of its aging population, the absolute number of people with diabetes is expected to substantially increase in the coming decades.[4] Since several intervention studies in different ethnic populations have demonstrated that type 2 diabetes mellitus (T2DM) can be effectively prevented through diet and lifestyle modifications in high-risk individuals;[5-8] identifying high-risk individuals and having them make diet and lifestyle changes is important for preventing diabetes onset.

A disease risk score is a calculated number or score that estimates the probability or rate of disease occurrence, derived from the risk factors of the disease. At present, there are several diabetes risk scores.[9-13] However, the substantial differences in diabetes incidence among ethnic groups[14,15] impact the performance of each model.[16] Although there are at least six diabetes risk prediction models for the Japanese population,[17-22] none are based on a general population across multiple areas in Japan. Although invasive risk scores are likely to have better predictive performance, non-invasive risk scores may be useful because they are less expensive and more convenient than invasive risk scores in large-scale screening.

Therefore, we aimed to develop regression models that used non-invasive and invasive predictors to predict the 5-year incidence of diabetes in a Japanese population and validate them externally in an independent Japanese population.

2. METHODS/DESIGN

2.1 Study population

The Japan Public Health Center-based Prospective Study (JPHC Study), designed to collect evidence based on multipurpose cohort studies to benefit health maintenance and improvement approaches, was initiated in 1990 for Cohort I and in 1993 for Cohort II. It included residents of 11 public health center areas (Iwate, Akita, Nagano, Okinawa, and Tokyo prefectures for Cohort I; Ibaraki, Niigata, Kochi, Nagasaki, Okinawa, and Osaka prefectures for Cohort II), aged 40–69 years at each baseline survey. Participants in this analysis underwent annual health checkups, completed self-administered questionnaire surveys, and provided blood samples. Specific details of the study design have been published previously.[23]

The JPHC Diabetes Study started in 1998–1999 for Cohort II (residents of the Osaka prefectures were excluded because the health checkup schedule was different from those of the other areas) and in 2000–2001 for Cohort I. In the baseline surveys, participants in Cohort I were 51–70 years old and 46–75 years old in Cohort II. A self-administered questionnaire, given during health checkups, collected data regarding family history of diabetes, previous diabetes examination results, any diagnosis of diabetes by a physician, current diabetes medications, signs of diabetic complications, a brief history of changes in body weight, time spent walking, and childbirth history.[24] The 5-year follow-up survey was performed in the same way in 2003–2004 for Cohort II and in 2005–2006 for Cohort I.

Among 28,362 adults enrolled in the baseline survey of this study, 10,986 (39%) were included in the final analysis. As shown in **Figure 1**, participants with diabetes (n=2,776) and those whose diabetes status could not be determined (n=4) at the baseline survey were excluded. Then, participants who responded to the 6-year follow-up survey but not to the 5-year follow-up survey (n = 1,625) and those who did not respond to the 5-year follow-up survey (n = 12,964) were excluded. Finally, participants who could not be diagnosed as being either diabetic or non-diabetic (n=7) at the 5-year follow-up survey were excluded. The remaining 10,986 participants were included in the analysis to develop a prediction model.

The Japan Epidemiology Collaboration on Occupational Health (J-ECOH) Study is an ongoing multi-center epidemiologic study conducted on workers from 12 companies spanning various industries; details of the study design have been published elsewhere.[20] For the present external validation, we retrieved data from one participating company that provided health checkup data, including a family history of diabetes, and defined an analytic cohort comprising individuals who had received health checkups in the fiscal year 2013 (baseline). As described elsewhere[25], study participants in the J-ECOH study were asked to select up to three activities from a list of 20 activities and the frequency (times per month) and duration (minutes per occasion) for each activity. Leisure-time physical activity (minutes per month) was computed by summing up the duration of activities reported by each participant. A total of 19,827 participants aged 46–75 years underwent a baseline checkup and had no missing data necessary for the validation analysis. Of these, individuals with diabetes at baseline (n = 2,663) and non-attendants to the 5-year health checkup in the fiscal year 2018 (n = 5,819) were excluded. Finally, 11,345 (57%) were used to validate the prediction models (**Figure 1**).

All participants provided written informed consent. The JPHC Study was approved by the ethics committees of Yokohama City University and the National Cancer Center, Japan, and was also approved by the ethics committee of the National Center for Global Health and Medicine, Japan. The J-ECOH study was approved by the Ethics Committee of the National Center for Global Health and Medicine, Japan.

2.2 Predictors

Based on previous literature, we selected 16 potential diabetes predictors (non-invasive predictors: age, sex, body mass index [BMI], time spent walking, family history of DM, systolic blood pressure [SBP], and diastolic blood pressure [DBP]; levels of invasive predictors: alanine aminotransferase [ALT], aspartate aminotransferase [AST], γ-glutamyl transferase [GGT], high-density lipoprotein [HDL], total cholesterol [TC], triglyceride [TG], estimated glomerular filtration rate [eGFR], fasting plasma glucose [FPG], and glycated hemoglobin [HbA1c]). All these factors were associated with the development of T2DM in previous studies.[26–34]

Data on age, height, weight, time spent walking, and family history of DM were acquired from the questionnaire; BMI was calculated as the weight in kilograms divided by the squared height in meters. The participants were classified into four levels based on the time spent walking: walking time < 0.5, 0.5–1, 1–2, or > 2 hours per day. A family history of diabetes was defined as the presence of diabetes in first-degree relatives. Blood pressure measurements were recorded during the health checkups.

When collecting blood samples, participants were not required to fast. Since fasting status has a great influence on TG levels, this parameter was excluded from our analysis. eGFR (mL/min/1.73 m^2) was calculated using the formula: $= 194 \times \text{serum creatinine}^{-1.094} \times \text{age}^{-0.287} \times 0.739$ (if female).[35] The recorded HbA1c level (expressed per the Japan Diabetes Society [JDS]) was converted to the National Glycohemoglobin Standardization Program (NGSP) equivalent using the following formula: HbA1c (%) =1.02 × HbA1c (JDS) (%) + 0.25%.[36]

2.3 Primary outcome measures

The diagnostic criteria for DM were as follows: (1) HbA1c value $\geq 6.5\%$, (2) FPG value ≥ 126 mg/dL, (3) random plasma glucose level ≥ 200 mg/dL, (4) physician-diagnosed DM (self-reported), or (5) undergoing any kind of diabetes treatment, including diet or exercise interventions (self-reported). These diagnostic criteria were used to exclude patients with diabetes at baseline and to confirm the number of patients diagnosed with diabetes at the 5-year follow-up in both the JPHC and J-ECOH studies. It was previously shown that 94% of self-reported diabetes cases were confirmed by medical reports in a subsample of the JPHC Study participants.[37]

2.4 Statistical analysis

After the multiple imputations as described later, logistic regression models were used to develop prediction models for diabetes incidence and to estimate β coefficients, odds ratios (ORs), and 95% confidence intervals (CIs). First, we examined all variables in the univariate regression model. We used a multiple logistic regression model with backward variable selection (fastbw function from the rms package) to determine significant variables in each multiple imputed dataset and in each JPHC Diabetes Study area. Predictors selected in more than 50% of the multiple imputed datasets among >50% of the areas were included in the final models[38]. Model 1 considered all non-invasive risk factors as potential predictors; Model 2 considered all non-invasive and invasive predictors, except FPG; and Model 3 considered all variables. Because the proportion of available FPG values was low, a model with FPG could produce unstable estimates because of missing data. Therefore, we developed Models 2 and 3 separately, although we imputed the FPG values using the multiple imputation method.

We used the rcorr function from the Hmisc package to assess multicollinearity, which suggested that the predictors did not strongly correlate with each other. We also examined missing values for several predictors. Assuming that the probability of missing data is determined only by the observed data (i.e., missing at the random condition), we used the multiple imputations by chained equations (MICE) algorithm[39] to impute the missing data. One hundred datasets were created based on the known information to obtain different imputed values.

Among the continuous predictors, age, DBP, eGFR, and TC levels tended to be linearly associated, whereas the remaining variables were more likely to be non-linearly associated with diabetes incidence (predictors selected in the final model are shown in **Supplemental Figure 1**), after assessing non-linearity using restricted cubic splines (rcs function from the rms package) and Akaike's information criterion (AIC function from the stats package). The rcs function was used to fit the nonlinear regression models by setting up special attributes (such as knots and nonlinear term indicators). The AIC evaluates how well a model fits the data (a smaller value of AIC is better).[40] Pooled β coefficients were estimated over the imputed datasets (fit.mult.impute function from the Hmisc package). All analyses were performed using R, version 4.2.0 (R Foundation for Statistical Computing, Vienna, Austria).[41]

2.5 Model validation

The final models were developed in the entire sample (eight areas) and evaluated via an internal validation of the JPHC Study dataset. The J-ECOH Study dataset was used for external validation. For the internal validation, we assessed the discrimination of the prediction models by calculating the area under the receiver operator characteristic (ROC) curve (AUC; also known as C-statistic) [40,42] using the roc function from the pROC package. Bootstrapping was used to quantify the optimism of our prediction models and to obtain optimism-corrected performance estimates (the number of bootstrap iterations was 1000). Optimism-corrected performance was calculated as optimism-corrected performance = apparent performance in the original sample − optimism, where optimism = bootstrap performance − test performance).[42] An AUC value of 0.5 indicates that the model is no better than random chance, while a value of 1 indicates that the model perfectly distinguishes cases and non-cases. We assessed the calibration (the agreement of observed outcomes with the predicted risk) of the prediction models by creating calibration plots using the val.prob.ci.2 function from the CalibrationCurves package. Apparent AUCs and calibration plots were estimated using a stacked dataset that stacks the 100 imputed data sets into a single data set.[42] Optimism-corrected AUCs were estimated within each imputed data set and averaged over 100 imputed data sets to obtain summary results.[42]

In the absence of a sufficiently large sample size, a random split sample approach or a non-random split sample approach is likely to provide unstable validation results. Therefore, to validate prediction models in different settings, we performed the internal-external cross-validation in the JPHC Diabetes Study (**Supplemental Figure 2**), as recommended by Steyerberg and Harrell.[42,43] For the internal-external cross-validation, the model development was performed in 7 areas by sequentially dropping one area at a time. Then, the models were validated in the omitted area by calculating AUC using the roc function from the pROC package.

For external validation, the discrimination and calibration performances of the developed models also used AUCs (roc function from the pROC package) and calibration plots (val.prob.ci.2 function from the CalibrationCurves package). In addition, to adjust the predicted risks for the validation cohort, we estimated the correction factor by using the function odds_adjust from the predtools package.

All analyses for model validation were conducted in each imputed dataset, and validation parameters were averaged to obtain pooled results.

To understand the impact on participants who did not participate in the follow-up survey, sensitivity analyses were also performed for the JPHC Diabetes Study and the J-ECOH Study.

Sensitivity analyses included all participants without diabetes at baseline. MICE was also used to impute missing data and 100 datasets were created based on known information to obtain different imputed values. Since people who did not participate in the 5-year follow-up survey could not determine whether they had diabetes, we counted the status of the patients in 100 datasets after imputation. If they were considered to have diabetes in more than 50 datasets, they were diagnosed with diabetes, otherwise, they were not. The average of probability was used to create the calibration plot.

2.6 Model presentation

The models were presented as formula based on the logistic regression coefficients. Thereafter, the risk score was calculated using an Excel spreadsheet (Microsoft; Redmond, WA, USA) created according to the formula (**eMaterial**: DM_model_calculations.xlsx). In addition, the study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[44] to improve the transparency and quality of reporting of these prediction models.

3. Results

The characteristics of the JPHC Study participants are presented in **Table 1 and Supplemental Table 1.** At the 5-year follow-up, 707 (6.4%) new diabetes cases were recorded. The median age was 63 years, and the number of women was 7377 (67.1%). People tended to exercise more than 2 hours a day (43.7%) rather than less than half an hour (12.6%). Approximately 11.2% of the participants had a family history of diabetes. Missing values were observed for 12 predictors in the derivation cohort. FPG was the variable with the most missing values in the data set, 7131 (64.9%). The mice package was used to perform multiple imputations for the missing values. In total, 8896 of the required 164,790 values (5.4%) were needed to impute for the final analysis.

Characteristics of the J-ECOH Study participants are presented in **Table 1**. There were fewer women (15.6 %), and approximately 17.6% of participants had a family history of diabetes in the J-ECOH study. There were 673 (5.9%) new diabetes cases at the 5-year follow-up. We also compared the baseline characteristics of participants who were not included in the final analysis of the JPHC Diabetes Study and the J-ECOH Study and found that they had similar characteristics to the analyzed participants (**Supplemental Table 3**).

Table 2 shows the differences in parameters between participants with and without diabetes and the relationship between risk factors and type 2 diabetes risk. There was little difference in age between participants with and without incident diabetes; however, there was a higher proportion of men among those with incident diabetes than among those without it. The risk of diabetes decreased with increased walking time. In addition, participants with incident type 2 diabetes had a family history of diabetes more frequently. For continuous variables (BMI, SBP, DBP, and the levels of ALT, AST, GGT, TC, FPG, and HbA1c), the median values were higher in the diabetes group than in the non-diabetes group. In contrast, HDL levels tended to be lower in those with incident diabetes than in those without diabetes.

Finally, sex, BMI, family history of DM, and DBP were selected for Model 1, family history of DM and HbA1c for Model 2, and family history of DM, HbA1c, and FPG for Model 3. For internal-external cross-validation, the AUCs of Model 1 ranged from 0.532 to 0.723, the AUCs of Model 2 ranged from 0.742 to 0.851, and the AUCs of Model 3 ranged from 0.807 to 0.895 (**Figure 2**). For the internal validation of the final models, the model performance is shown in **Figure 2**. The AUC of Model 1 was 0.643, that of Model 2 yielded an AUC of 0.786, and that of Model 3 had an AUC of 0.845. After bootstrap optimism correction, the AUCs slightly decreased to 0.639, 0.785, and 0.844, respectively. The discriminative ability of each model was confirmed in the J-ECOH Study; the AUCs were 0.692, 0.831, and 0.874 in Models 1,2, and 3, respectively.

The calibration curves (**Figure 3**) indicated that the predicted and empirical probabilities were close to each other, indicating that the prediction models fitted the data well in the development cohort. As shown in **Figure 3**, the probability of diabetes in high-risk participants was overestimated in Models 1 and 3 in the validation cohort. The extent of agreement between the observed outcomes and predicted risk in Model 2 was better than that in Models 1 and 3 in the validation cohort.

The predictive performance did not materially change when a family history of diabetes was defined as the presence of diabetes in a family member, regardless of the degree of the relationship (**Supplemental Table 2; Supplemental Figure 4**). In addition, the calibration plots in the validation cohort remained unchanged after the intercept adjustments (**Supplemental Figure 5**). After a sensitivity analysis that included participants who did not participate in the follow-up survey, the AUCs in the JPHC Diabetes Study changed to 0.631, 0.764, and 0.848, and those in the J-ECOH Study changed to 0.676, 0.834, and 0.874 in models 1, 2, and 3, respectively (**Supplemental Figure 3**). The calibration performance did not improve in the sensitivity analysis, as shown in **Supplemental Figure 6**.

Table 3 shows the content of the Excel spreadsheet used to obtain approximate predictions for the individuals. Using the medians for continuous predictors and the category with more participants for categorical variables, we calculated the average risk probability of DM to be 3.94% in Model 1, 3.32% in Model 2, and 1. 54% in Model 3. Here, we provide an example using Model 2 to show how to obtain DM risk probability. A male with a family history of diabetes demonstrated a BMI of 25 kg/m^2, a diastolic blood pressure of 80 mmHg, and an HbA1c of 6%. By entering these data into Excel, the risk of DM was estimated to be 23.89%.

4. Discussion

In this study, we developed three models to predict the risk of DM. All models showed good discrimination and calibration in internal validations. The internal-external cross-validation indicated that these models showed similar discriminative ability across eight areas. To the best of our knowledge, this is the first diabetes risk score developed and validated using a nationwide population in Japan to predict the 5-year incidence of type 2 diabetes. For the non-invasive model, sex, BMI, family history of diabetes, and DBP were used to create a non-invasive prediction model that showed good predictive ability (AUC=0.643) for the 5-year incidence of type 2 diabetes. The risk models that included HbA1c showed better predictive ability, with an AUC of 0.786, and the predictive model performed best when both FPG and HbA1c levels were included (AUC=0.845), consistent with previous studies.[18–21] Although the AUC values decreased after optimism correction, all remained reliable, as also observed in the internal-external cross-validation and external validation cohort. The AUC values were higher in the J-ECOH Study than in the JPHC Diabetes Study, indicating that the developed models were generally good at discrimination. For the calibration performance, however, calibration plots of Models 1 and 3 were poor in the validation cohort. This indicates that the predicted probabilities overestimated the observed probabilities in the validation cohort. In comparison, Model 2 was well-calibrated in the J-ECOH Study. Since Model 2 tended to underestimate the observed probability in the highest decile of the predicted probability in the J-ECOH Study, the model should be used with caution, especially for those with a high predicted probability.

Several earlier studies developed diabetes prediction models for Japanese populations,[17-22] including the earliest known diabetes risk score model that was published in 2008 for residents of the Ibaraki prefecture.[17] The model included BMI, blood glucose level, SBP, treatment for hypertension, TG levels, and smoking habits as predictors; however, it did not provide the AUC value. The Hisayama study included 1935 participants in the development model and 1147 in the validation model. However, all the participants were residents of a rural town, suggesting limited study generalizability.[18] Two risk models were established in the Hisayama Study. Age, sex, family history of diabetes, abdominal circumference, BMI, hypertension, regular exercise, and current smoking were included in the noninvasive risk model, with an AUC of 0.700, which increased to 0.772 when FPG levels were added. The participants in the Toranomon Hospital Health Management Center Study 6 mainly involved apparently healthy Japanese government employees[19]; it included four risk scores. The AUC of the model that included age, sex, family history of diabetes, current smoking, and BMI was 0.708, which increased to 0.836 when the FPG level was added, 0.837 when HbA1c was included, and 0.887 when both FPG and HbA1c levels were added. In the Japan Epidemiology Collaboration on Occupational Health Study (J-ECOH Study),[20, 21] most participants were workers in large companies, and the risk predictors did not include a family history of diabetes. 3- and 7-year predicted probabilities of DM were created using age, sex, smoking status, abdominal obesity, BMI, and hypertension status in the basic model or by adding FPG or HbA1c levels or adding both FPG and HbA1c levels. The AUC values ranged from 0.717 to 0.893 for the 3-year incidence of DM and from 0.73 to 0.89 for the 7-year incidence of DM. The Aizawa Hospital Study[22] included individuals who underwent general health examinations at the Health Center of Aizawa Hospital (development cohort, 2080 individuals; validation cohort, 2079 individuals).

Compared with these previous studies, we developed the model based on a population across multiple areas in Japan. Our models provided AUCs (unlike the Ibaraki Prefectural Health Study), included a family history of DM (unlike the J-ECOH Study), and were not limited to one region or occupation (unlike all the studies mentioned before). Therefore, we believe that our models are more representative of a Japanese population. We confirmed the validity of our prediction models with internal validation using bootstrapping and internal-external cross-validation in the JPHC Diabetes Study. These procedures are recommended by Steyerberg and Harrell.[42,43] In addition, we fully utilized the information of continuous variables such as HbA1c or FPG using the cubic spline function to model potential nonlinear relations between variables and to avoid information loss. Finally, our models showed good performance in distinguishing between individuals with and without the risk of developing diabetes.

There are several possible explanations as to why the population of the J-ECOH study did not present good calibration performance. As shown in Table 1, the study participants of the J-ECOH study were younger (median age: 51 vs. 63) and tended to have lower SBP (median: 122 vs. 130) than those in the JPHC Diabetes Study. These factors are established risk factors for type 2 diabetes and these were not included in our prediction models, which may have affected the calibration performance.

Our study had several limitations. First, approximately 51% (12964/25582) of the participants without diabetes in the JPHC Diabetes Study and 34% (5819/17164) of the participants without diabetes in the J-ECOH Study participated in the baseline survey but did not visit the 5-year follow-up survey, potentially causing selection bias. However, when we included those who did not complete the 5-year follow-up survey and imputed the outcomes using the MICE, the results did not materially change (**Supplemental Figure 3**). Second, we did not conduct oral glucose tolerance tests to define the incidence of type 2 diabetes, possibly underestimating the incidence.[24] Furthermore, although our internal validation via bootstrapping did not indicate any severe optimism, some optimism may exist because our bootstrapping procedure could not incorporate the uncertainty of the model selection and variable selection. In addition, we used the dataset from 20 years ago to create the prediction model, which may not be as accurate as data collected more recently. Finally, although our previous findings[45] suggested that adding a genetic risk score might provide incremental model predictive performance, we did not include the genetic risk score in this study.

In conclusion, 5-year models for predicting the incidence of type 2 diabetes, with high discrimination and calibration, were developed and validated in this population-based study among a Japanese population. The invasive risk model with only HbA1c provides a tool for the targeted selection of patients with the greatest need for intervention.

Acknowledgments

The authors thank Dr. Masanori Arai (Department of Endocrinology and Metabolism, Graduate School of Medicine, Yokohama City University, Yokohama, Japan) for useful advice, and Misa Katayama and Mari Takaki (Yokohama City University) for secretarial assistance. The authors also express gratitude to the study participants and staff members involved in each study area.

Fundings Sources

The JPHC Study was supported by Health Sciences Research Grants (Research on Health Services H10-074, Medical Frontier Strategy Research H13-008, Clinical Research for Evidence-based Medicine H14-008 and H15-006, and Comprehensive Research on Life-Style Related Diseases including Cardiovascular Diseases and Diabetes Mellitus H16-019, H17-019, H18-028, H19-016, and H25-016) and a grant-in-aid for Cancer Research, and a grant-in-aid for the Third Term Comprehensive Ten-Year Strategy for Cancer Control from the Ministry of Health, Labour and Welfare of Japan. The JPHC Study was also supported by the National Cancer Center for Research and Development Fund (since 2011) (23-A-31[toku], 26-A-2, 29-A-4, and 2020-J-4). The J-ECOH Study was supported by a grant from the Industrial Health Foundation and the Grant of National Center for Global Health and Medicine (22A1008). The funders did not have any involvement in the study design, collection, analysis, and interpretation of the data, preparation of the manuscript, or decision to publish. JX was supported by fellowship grants from the Rotary Yoneyama Memorial Foundation.

Conflict of interest

The authors declare that they have no conflicts of interest with respect to this research study and paper.

Data Availability: Data analyzed in the present study are not publicly available because permission has not been obtained from the ethical board, but the information on how to access to JPHC data is available by following instructions at https://epi.ncc.go.jp/en/jphc/805/8155.html. J-ECOH Study data are available at the National Center for Global Health and Medicine and can be shared upon request by academic researchers for non-commercial research. Inquiries and applications can be made to the Department of Epidemiology and Prevention, Center for Clinical Sciences, National Center for Global Health and Medicine, Tokyo, Japan (Dr. Mizoue, mizoue@ri.ncgm.go.jp).

References

[1] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2014;37:S81–90. https://doi.org/10.2337/dc14-S081.

[2] Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract. 2022 Jan;183:109119. https://doi.org/10.1016/j.diabres.2021.109119.

[3] Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396:1204–22. https://doi.org/10.1016/S0140-6736(20)30925-9.

[4] Goto A, Noda M, Inoue M, Goto M, Charvat H. Increasing number of people with diabetes in Japan: Is this trend real? Intern Med 2016;55:1827–30. https://doi.org/10.2169/internalmedicine.55.6475.

[5] Sherwin RS, Anderson RM, Buse JB, Chin MH, Eddy D, Fradkin J, et al. The prevention or delay of type 2 diabetes. Diabetes Care 2002;25:742–9. https://doi.org/10.2337/diacare.25.4.742.

[6] Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen 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. https://doi.org/10.1056/NEJM200105033441801.

[7] 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. https://doi.org/10.1056/NEJMoa012512.

[8] Pan XR, Li GW, Hu YH, Wang JX, Yang WY, An ZX, et al. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and diabetes study. Diabetes Care 1997;20:537–44. https://doi.org/10.2337/diacare.20.4.537.

[9] Lindström J, Tuomilehto J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. Diabetes Care 2003;26:725–31. https://doi.org/10.2337/diacare.26.3.725.

[10] Glümer C, Carstensen B, Sandbæk A, Lauritzen T, Jørgensen T, Borch-Johnsen K. A Danish diabetes risk score for targeted screening - The Inter99 study. Diabetes Care 2004;27:727–33. https://doi.org/10.2337/diacare.27.3.727.

[11] Aekplakorn W, Bunnag P, Woodward M, Sritara P, Cheepudomwit S, Yamwong S, et al. A risk score for predicting incident diabetes in the Thai population. Diabetes Care 2006;29:1872–7. https://doi.org/10.2337/dc05-2141.

[12] 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. https://doi.org/10.1136/bmj.b880.

[13] Sun F, Tao Q, Zhan S. An accurate risk score for estimation 5-year risk of type 2 diabetes based on a health screening population in Taiwan. Diabetes Res Clin Pract 2009;85:228–34. https://doi.org/10.1016/j.diabres.2009.05.005.

[14] McBean AM, Li SL, Gilbertson DT, Collins AJ. Differences in diabetes prevalence, incidence, and mortality among the elderly of four racial/ethnic groups: Whites,

blacks, Hispanics, and Asians. Diabetes Care 2004;27:2317–24. https://doi.org/10.2337/diacare.27.10.2317.

[15] Oldroyd J, Banerjee M, Heald A, Cruickshank K. Diabetes and ethnic minorities. Postgrad Med J 2005;81:486–90. https://doi.org/10.1136/pgmj.2004.029124.

[16] He S, Chen X, Cui K, Peng Y, Liu K, Lv Z, et al. Validity evaluation of recently published diabetes risk scoring models in a general Chinese population. Diabetes Res Clin Pract 2012;95:291–8. https://doi.org/10.1016/j.diabres.2011.10.039.

[17] Sasai H, Sairenchi T, Irie F, Iso H, Tanaka K, Ota H. Development of a diabetes risk prediction sheet for specific health guidance. Nihon Koshu Eisei Zasshi 2008;55:287–94 [article in Japanese].

[18] Doi Y, Ninomiya T, Hata J, Hirakawa Y, Mukai N, Iwase M, et al. Two risk score models for predicting incident Type 2 diabetes in Japan. Diabet Med 2012;29:107–14. https://doi.org/10.1111/j.1464-5491.2011.03376.x.

[19] Heianza Y, Arase Y, Hsieh SD, Saito K, Tsuji H, Kodama S, 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 2012;55:3213–23. https://doi.org/10.1007/s00125-012-2712-0.

[20] Nanri A, Nakagawa T, Kuwahara K, Yamamoto S, Honda T, Okazaki H, et al. Development of risk score for predicting 3-year incidence of Type 2 diabetes: Japan Epidemiology Collaboration on Occupational Health Study. PLOS ONE 2015;10:e0142779. https://doi.org/10.1371/journal.pone.0142779.

[21] Hu H, Nakagawa T, Yamamoto S, Honda T, Okazaki H, Uehara A, et al. Development and validation of risk models to predict the 7-year risk of type 2 diabetes: The Japan

Epidemiology Collaboration on Occupational Health Study. J Diabetes Investig 2018;9:1052–9. https://doi.org/10.1111/jdi.12809.

[22] Miyakoshi T, Oka R, Nakasone Y, Sato Y, Yamauchi K, Hashikura R, et al. Development of new diabetes risk scores on the basis of the current definition of diabetes in Japanese subjects. Endocr J 2016;63:857–65. https://doi.org/10.1507/endocrj.EJ16-0340.

[23] Tsugane S, Sawada N. The JPHC study: design and some findings on the typical Japanese diet. Jpn J Clin Oncol. 2014;44:777–82. https://doi.org/10.1093/jjco/hyu09 6.

[24] Noda M, Kato M, Takahashi Y, Matsushita Y, Mizoue T, Inoue M, et al. Fasting plasma glucose and 5-year incidence of diabetes in the JPHC diabetes study—Suggestion for the threshold for impaired fasting glucose among Japanese. Endocr J 2010;57:629–37. https://doi.org/10.1507/endocrj.k10e-010.

[25] Yamamoto, S, Inoue, Y, Kuwahara, K, Miki T, Nakagawa T, Honda T. et al. Leisure-time, occupational, and commuting physical activity and the risk of chronic kidney disease in a working population. Sci Rep 2021;11:12308. https://doi.org/10.1038/s41598-021-91525-4.

[26] Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. Nature 2001;414:782–7. https://doi.org/10.1038/414782a.

[27] Pan XR, Yang WY, Li GW, Liu J. Prevalence of diabetes and its risk factors in China, 1994. Diabetes Care 1997;20:1664–9. https://doi.org/10.2337/diacare.20.11.1664.

[28] Gale EAM, Gillespie KM. Diabetes and gender. Diabetologia 2001;44:3–15. https://doi.org/10.1007/s001250051573.

[29] Harita N, Hayashi T, Sato KK, Nakamura Y, Yoneda T, Endo G, et al. Lower serum creatinine is a new risk factor of Type 2 diabetes: the Kansai healthcare study. Diabetes Care 2009;32:424–6. https://doi.org/10.2337/dc08-1265.

[30] Boffetta P, McLerran D, Chen Y, Inoue M, Sinha R, He J, et al. Body mass index and diabetes in Asia: A cross-sectional pooled analysis of 900,000 individuals in the Asia Cohort Consortium. PLOS ONE 2011;6:e19930. https://doi.org/10.1371/journal.pone.0019930.

[31] Harrison TA, Hindorff LA, Kim H, Wines RCM, Bowen DJ, McGrath BB, et al. Family history of diabetes as a potential public health tool. Am J Prev Med 2003;24:152–9. https://doi.org/10.1016/s0749-3797(02)00588-3.

[32] Gress TW, Nieto FJ, Shahar E, Wofford MR, Brancati FL. Hypertension and antihypertensive therapy as risk factors for type 2 diabetes mellitus. N Engl J Med 2000;342:905–12. https://doi.org/10.1056/NEJM200003303421301.

[33] Zhao J, Zhang Y, Wei FJ, Song JN, Cao Z, Chen C, et al. Triglyceride is an independent predictor of type 2 diabetes among middle-aged and older adults: A prospective study with 8-year follow-ups in two cohorts. J Transl Med 2019;17:403. https://doi.org/10.1186/s12967-019-02156-3.

[34] Fraser A, Harris R, Sattar N, Ebrahim S, Davey Smith GD, Lawlor DA. Alanine aminotransferase, gamma-glutamyltransferase, and incident diabetes: The British Women's Heart and Health Study and meta-analysis. Diabetes Care 2009;32:741–50. https://doi.org/10.2337/dc08-1870.

[35] Matsuo S, Imai E, Horio M, Yasuda Y, Tomita K, Nitta K, et al. Revised equations for estimated GFR from serum creatinine in Japan. Am J Kidney Dis 2009;53:982–92. https://doi.org/10.1053/j.ajkd.2008.12.034.

[36] Kashiwagi A, Kasuga M, Araki E, Oka Y, Hanafusa T, Ito H, et al. International clinical harmonization of glycated hemoglobin in Japan: From Japan Diabetes Society to National Glycohemoglobin Standardization Program values. J Diabetes Investig 2012;3:39–40. https://doi.org/10.1111/j.2040-1124.2012.00207.x.

[37] Waki K, Noda M, Sasaki S, et al. Alcohol consumption and other risk factors for self-reported diabetes among middle-aged Japanese: a population-based prospective study in the JPHC study cohort I. Diabet Med. 2005;22: 323-31. https://doi.org/10.1111/j.1464-5491.2004.01403.x

[38] Gupta RK, Harrison EM, Ho A, Docherty AB, Knight SR, van Smeden M, et al. Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study. Lancet Respir Med 2021;9:349-59. https://doi.org/10.1016/S2213-2600(20)30559-2.

[39] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45:1-67.

[40] Harrell FE. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York: Springer; 2021.

[41] R Core Team. R: A language and environment for statistical computing, http://www.R-project.org/index.html; 2020. Vienna, Austria: R Foundation for Statistical Computing.

[42] Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. 2nd ed. Springer; 2019.

[43] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69: 245-247. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[44] Collins, GS, Reitsma, JB, Altman, DG, Moons, KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55-U103. https://doi.org/10.1002/bjs.9736.

[45] Goto A, Noda M, Goto M, Yasuda K, Mizoue T, Yamaji T, et al. Predictive performance of a genetic risk score using 11 susceptibility alleles for the incidence of Type 2 diabetes in a general Japanese population: A nested case-control study. Diabet Med 2018;35:602-611. https://doi.org/10.1111/dme.13602.

Table 1. Characteristics of participants in the JPHC Diabetes Study and the J-ECOH Study[a]

Characteristic[a]	JPHC Diabetes Study (n= 10986)		Characteristic[a]	J-ECOH Study (n= 11,345)	
	Value [b]	Missing values, n (%)		Value[b]	Missing values, n (%)
Age (years)	63 (57–67)	0	Age (years)	51 (48–54)	0
Women	7377 (67.1%)	0	Women	1,773 (15.6 %)	0
BMI (kg/m^2)	23.5 (21.5–25. 6)	23 (0.2)	BMI (kg/m^2)	23.2 (21.4–25.3)	0
Walking time (hours per day)			Leisure-time physical activity (minutes per month)	0 (0–84)	391 (3.4)
≤0.5 hours	1379 (12. 6%)	130 (1.2)			

0.5–1 hour	2322 (21.1%)				
1–2 hours	2349 (21.4%)				
≥2 hours	4806 (43.7%)				
Family history of diabetes	1225 (11.2%)	0	Family history of diabetes	1,996 (17.6%)	0
SBP (mmHg)	130 (119–140)	6 (0.1)	SBP (mmHg)	122 (113–130)	0
DBP (mmHg)	78 (70–84)	6 (0. 1)	DBP (mmHg)	79 (72–84)	0
HDL (mg/dL)	57 (48–67)	1 (0.0)	HDL (mg/dL)	55 (46–65)	0
TC (mg/dL)	207 (186–230)	1 (0.0)	TC (mg/dL)	201 (181–221)	16 (0.1)
FPG (mg/dL)	93 (88–100)	7131 (64.9)	FPG (mg/dL)	98 (92–105)	0
HbA1c (%)	5.5 (5.1–5.7)	34 (0.3)	HbA1c (%)	5.5 (5.3–5.7)	0

ALT (IU/L)	18 (15–24)	7 (0. 1)	ALT (IU/L)	21 (16–29)	0
AST (IU/L)	22 (19–27)	1 (0.0)	AST (IU/L)	21 (18–26)	0
GGT (IU/L)	21 (15–33)	7 (0. 1)	GGT (IU/L)	30 (20–51)	0
eGFR (mL/min/1.73 m^2)	73.8 (63.4–82.5)	1549 (14.1)	eGFR (mL/min/1.73 m^2)	78.8 (69.7–89.4)	5549 (48.9)
5-year outcome	707 (6.4%)	0	5-year outcome	673 (5.9%)	0

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; FPG, fasting plasma glucose; GGT, γ-glutamyl transferase; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; SBP, systolic blood pressure; TC, total cholesterol.

[a]Characteristics were collected at baseline.

[b]Continuous variables are medians (interquartile ranges) and categorical variables are numbers (percentages).

Table 2. Distribution of study variables by DM status in the JPHC Diabetes Study.

Characteristics[b]	Participants without incident DM[a] (n = 10279)	Participants with incident DM[a] (n = 707)	Odds ratio (95% CI) [c,d]			
			Univariate	Model 1	Model 2	Model 3
Age[e] (years)	63 (57–67)	64(59–68)	1.23 (1.09–1.38)	–	–	–
Sex (%)						
Female	6980 (95%)	397 (5%)	1 (ref.)	1 (ref.)	–	–
Male	3299 (91%)	310 (9%)	1.65 (1.42–1.93)	1.74 (1.49–2.04)	–	–
BMI (kg/m^2)	23.5 (21.5–25.5)	24.5 (22.4–26.7)	1.78 (1.45–2.18)	1.73 (1.41–2.13)	–	–
Walking time[e] (hours per day)						

≤0.5 hour	1278 (93%)	101 (7%)	1.22 (0.97–1.54)	–	–	–
0.5–1 hour	2164 (93%)	158 (7%)	1.13 (0.92–1.38)	–	–	–
1–2 hours	2196 (93%)	153 (7%)	1.08 (0.88–1.32)	–	–	–
≥2 hours	4514 (94%)	292 (6%)	1 (ref.)	–	–	–
Family history of diabetes (%)						
Yes	1082 (88%)	143 (12%)	2.16 (1.78–2.62)	2.26 (1.86–2.75)	1.64 (1.33–2.03)	1.56 (1.23–1.98)
No	9197 (94%)	564 (6%)	1 (ref.)	1 (ref.)	1 (ref.)	1 (ref.)
SBP[e] (mmHg)	130 (118–140)	134 (124–144)	1.44 (1.29–1.60)		–	–
DBP (mmHg)	78 (70–84)	80 (70–86)	1.19 (1.08–1.32)	1.04 (0.94–1.16)	–	–

HDL[e] (mg/dL)	57 (48–68)	53 (45–64)	0.60 (0.49–0.74)	–	–	–
TC[e](mg/dL)	207 (186–229)	211 (188–232)	1.13 (1.02–1.25)	–	–	–
FPG (mg/dL)	93 (88–99)	106 (97–115)	4.16 (2.83–6.10)	–	–	2.95 (1.98–4.39)
HbA1c (%)	5.4 (5.1–5.7)	5.9 (5.6–6.1)	3.50 (2.91–4.22)	–	3.44 (2.86–4.13)	2.63 (2.17–3.19)
ALT[e] (IU/L)	18 (14–24)	21 (16–28)	1.58 (1.37–1.83)	–	–	–
AST[e] (IU/L)	22 (19–26)	24 (20–29)	1.54 (1.32–1.79)	–	–	–
GGT[e] (IU/L)	21 (15–32)	26 (18–43)	2.07 (1.77–2.42)	–	–	–
eGFR[e] (mL/min/1.73 m^2)	73.8 (63.4–82.5)	73.5 (63.4–83.0)	0.98 (0.92–1.06)	–	–	–

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; CI, confidence interval; DBP, diastolic blood pressure; DM, diabetes mellitus; eGFR, estimated glomerular filtration rate; FPG, fasting plasma glucose; GGT, γ-glutamyl transferase; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; ref., reference; SBP, systolic blood pressure; TC, total cholesterol.

[a] Continuous variables are shown as medians (interquartile ranges) and categorical variables as numbers (percentages) unless otherwise indicated.

[b] A backward stepwise variable selection method was used to select the variables to be included in the prediction model.

[c] Odds ratios were estimated using logistic regression models after multiple imputations. Model 1 included sex, BMI, family history of DM, and DBP. Model 2 included a family history of DM, and HbA1c. Model 3 included a family history of DM, FPG level and HbA1c.

[d] Interquartile range (0.75 vs. 0.25 quantile) odds ratios are shown for continuous variables. For example, odds ratio for age compares the 3rd quartile with the 1st quartile of age. Odds ratios for categorical predictors were compared between each group and the reference group (the smallest group).

[e] Not included in each model after the backward stepwise variable selection method.

Table 3. Prediction Model and Calculation Table.

Predictors[a]	Variables	Units	Coefficient			Average values[c]	Coefficient × Average values			Your patient (Example using Model 2)[d]	
			Model 1	Model 2	Model 3		Model 1	Model 2	Model 3		
Constant	Intercept	–	-1.47	-7.96	-4.11	1	-1.47	-7.96	-4.11	1	-7. 96
Sex	Female	0/1	-0.56			1	-0.56	–	–	**0**	–
BMI	BMI	kg/m2	-0.08	–	–	23.50	-1.83	–	–	**25**	–
	$(\text{BMI-}19.0)^{3}+$		0.00	–	–	91.13	0.45	–	–	216.00	–
	$(\text{BMI-}22.4)^{3}+$		-0.01	–	–	1.33	-0.02	–	–	17.58	–

	$(BMI-24.7)^3+$		0.01	–	–	0.00	0.00	–	–	0.03	–
	$(BMI-28.9)^3+$		-0.00	–	–	0.00	0.00	–	–	0.00	–
Family history of DM	Family history of DM	0/1	0.82	0.50	0.45	0	0.00	0.00	0.00	**1**	0.50
DBP[b]	DBP	mm Hg	0.00	–	–	78	0.24	–	–	**80**	–
HbA1c[b]	HbA1c	%	–	0.77	0.44	5.5	–	4.24	2.43	**6.0**	4.63
	$(HbA1c-4.9)^3+$		–	1.59	1.44	0.2	–	0.34	0.31	1.33	2.11

	(HbA1c-5.5)^{3}+		–	-3.49	-3.17	0.0	–	0.00	0.00	0.13	-0.44
	(HbA1c-6.0)^{3}+		–	1.90	1.73	0.0	–	0.00	0.00	0.00	0.00
FPG[b]	FPG	mg/dl	–	–	-0.03	93	–	–	-3.02	**100**	–
	(FPG -81)^{3}+		–	–	0.00	1728.0	–	–	0.07	6859.00	–
	(FPG -88)^{3}+		–	–	0.00	125.00	–	–	0.16	1728.00	–
	(FPG -93)^{3}+		–	–	-0.00	0.00	–	–	0.00	343.00	–

(FPG - 99) ^{3}+	–	–	0.00	0.00	–	–	0.00	1.00	–
(FPG - 112) ^{3}+	–	–	-0.00	0.00	–	–	0.00	0.00	–
				Probability	3.94%	3.32%	1.54%		23.89%

Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; DM, diabetes mellitus; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin.

[a]Variables were selected using the backward stepwise method, and multiple imputations by chained equations (MICE) method was used to handle missing data.

[b]Knots were placed at the 10th, 50th, and 90th percentiles for HbA1c; at the 5th, 35th, 65th, and 95th percentiles for BMI, and at the 5th, 27.5th, 50th, 72.5th, and 95th percentiles for FPG.

[c]To calculate the average risk probability of the DM. The medians were used for continuous predictors. The category with more participants were used for categorical variables.

[d]An example is provided. A male with a family history of diabetes, diastolic blood pressure of 80 mmHg, BMI of 25 kg/m^2, and HbA1c level of 6.0%.

After pooling the coefficients in the final multivariable model, the formula for the five-year incidence of type 2 diabetes can be summarized as 1/[1+exp (−L)],

where L in Model 1 = −1. 4677114 − 0.55636706 × [Sex = "female"] − 0.077979787 × BMI + 0.0048939561 × (BMI − 19.0)^3 − 0.014293364 × (BMI − 22.4)^3 + 0.010584929 × (BMI − 24.7)^3 − 0.0011855209 × (BMI − 28.9)^3 + 0.81638492 × [Family history of diabetes = "YES"] + 0.0030199043 × DBP;

where L in Model 2 = −7.9560656 + 0.49588037 × [Family history of diabetes = "YES"] + 0.77107227 × HbA1c + 1.5861765 × (HbA1c − 4.9)^3 − 3. 4895883 × (HbA1c − 5.5)^3 + 1.9034118 × (HbA1c – 6.0)^3;

where L in Model 3 = −4. 1097962 + 0. 44533254 × [Family history of diabetes = "YES"] + 0.44201803 × HbA1c + 1.4426444 × (HbA1c − 4.9)^3 − 3.1738177 × (HbA1c − 5.5)^3 + 1.7311733 × (HbA1c – 6.0)^3 − 0.032485574 × FPG + 0.000040103209 × (FPG − 81)^3 + 0.0012713229 × (FPG − 88)^3 − 0.0028839757 × (FPG − 93)^3 + 0.001772353 × (FPG − 99)^3 − 0.00019980342 × (FPG − 112)^3.

Notes: in L,

1. Square brackets [c] = 1 if the participant falls into category c; [c] = 0 otherwise.

2. Round brackets indicate $(x)_+ = x$ if $x > 0$, and $(x)_+ = 0$ otherwise.

3. Measurement units: BMI (kg/m^2), DBP (mmHg), HbA1c (%), and FPG (mg/dL).

4. Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin.

Figure Legends

Figure 1: Participant selection flow diagram for the development and validation cohorts.

Figure 2. Receiver operating characteristic curves for the development and validation cohorts.

Abbreviations: AUC, the area under the receiver operating characteristic (ROC) curve; BMI, body mass index; DBP, diastolic blood pressure; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin.

Model 1: included sex, BMI, a family history of DM, and DBP.

Model 2: included a family history of DM and HbA1c

Model 3: included a family history of DM, HbA1c, and FPG

C-statistic (AUC): in the JPHC Diabetes Study, Model 1 = 0.643, Model 2 = 0.786, and Model 3 = 0.845; after optimism correction, the AUCs decreased to 0.639, 0.785, and 0.844, respectively. The number of bootstrap iterations was 1000. After internal-external cross-validation, the AUCs of each area in Model 1 = 0.629, 0.688, 0.634, 0.723, 0.633, 0.532, 0.595, and 0.686, respectively; the AUCs of each area in Model 2 = 0.823, 0.772, 0.754, 0.846, 0.851, 0.806, 0.742, and 0.798, respectively; the AUCs of each area in Model 3 = 0.855, 0.853, 0.817, 0.895, 0.884, 0.807, 0.809, and 0.868, respectively. The AUCs in the J-ECOH Study were 0.692, 0.831, and 0.874 in Models 1, 2, and 3, respectively.

Figure 3. Calibration plots to show relations between predicted and observed probabilities in the development and validation cohorts.

Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin.

Model 1: included sex, BMI, a family history of DM, and DBP.

Model 2: included a family history of DM and HbA1c

Model 3: included a family history of DM, HbA1c, and FPG

Calibration plots were created to graphically assess the agreement of the mean observed risk with the mean predicted risk according to the deciles of the predicted risk. Ideal: ideal line for the prediction model. Flexible calibration (RCS): "RCS" generates a flexible calibration curve based on restricted cubic splines. CL flexible: 95% confidence limits for the flexible calibration curve with dashed lines. Grouped observations: mean predicted probability and observed proportion of diabetes incidence in each of the deciles (ten groups of equal size).

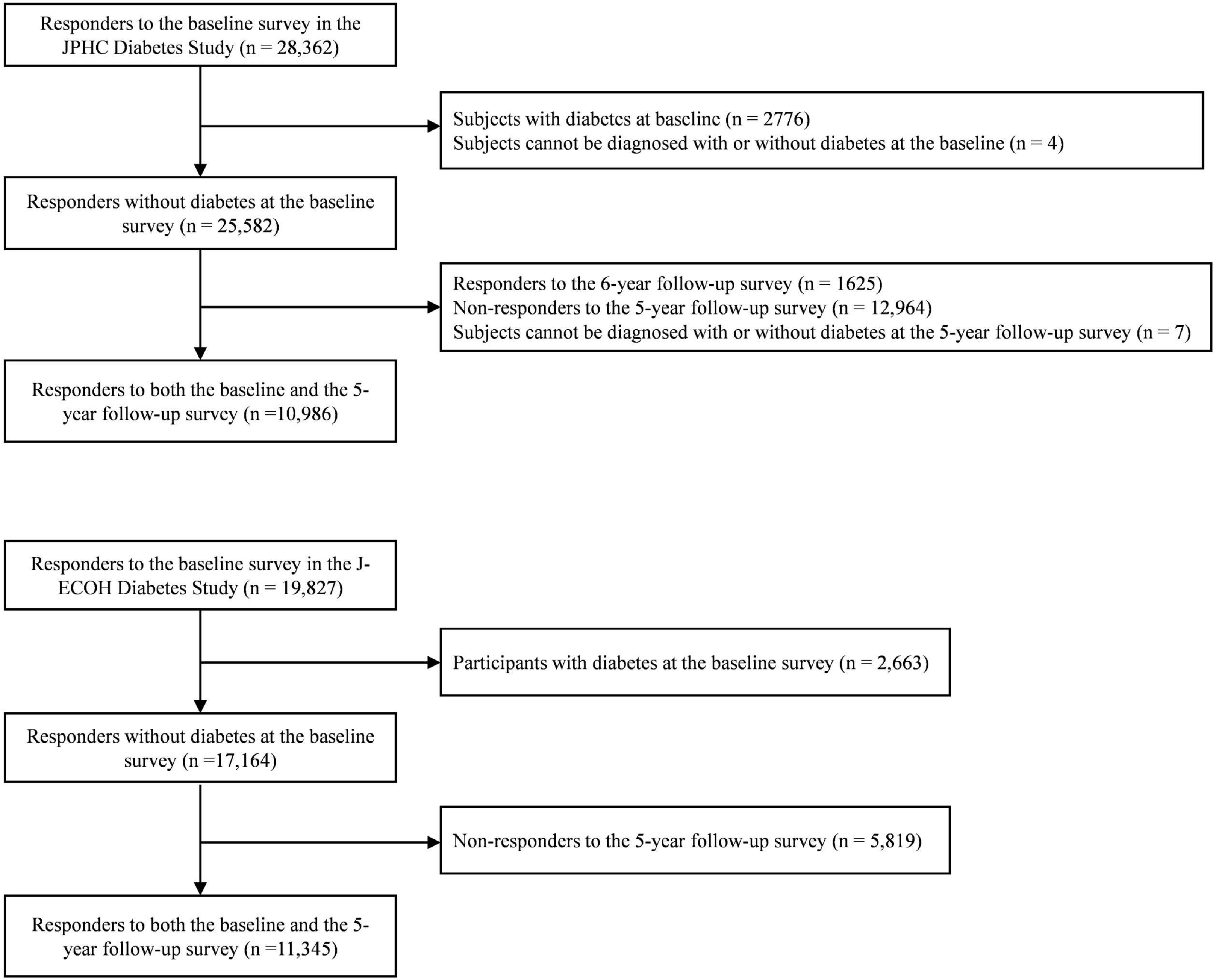
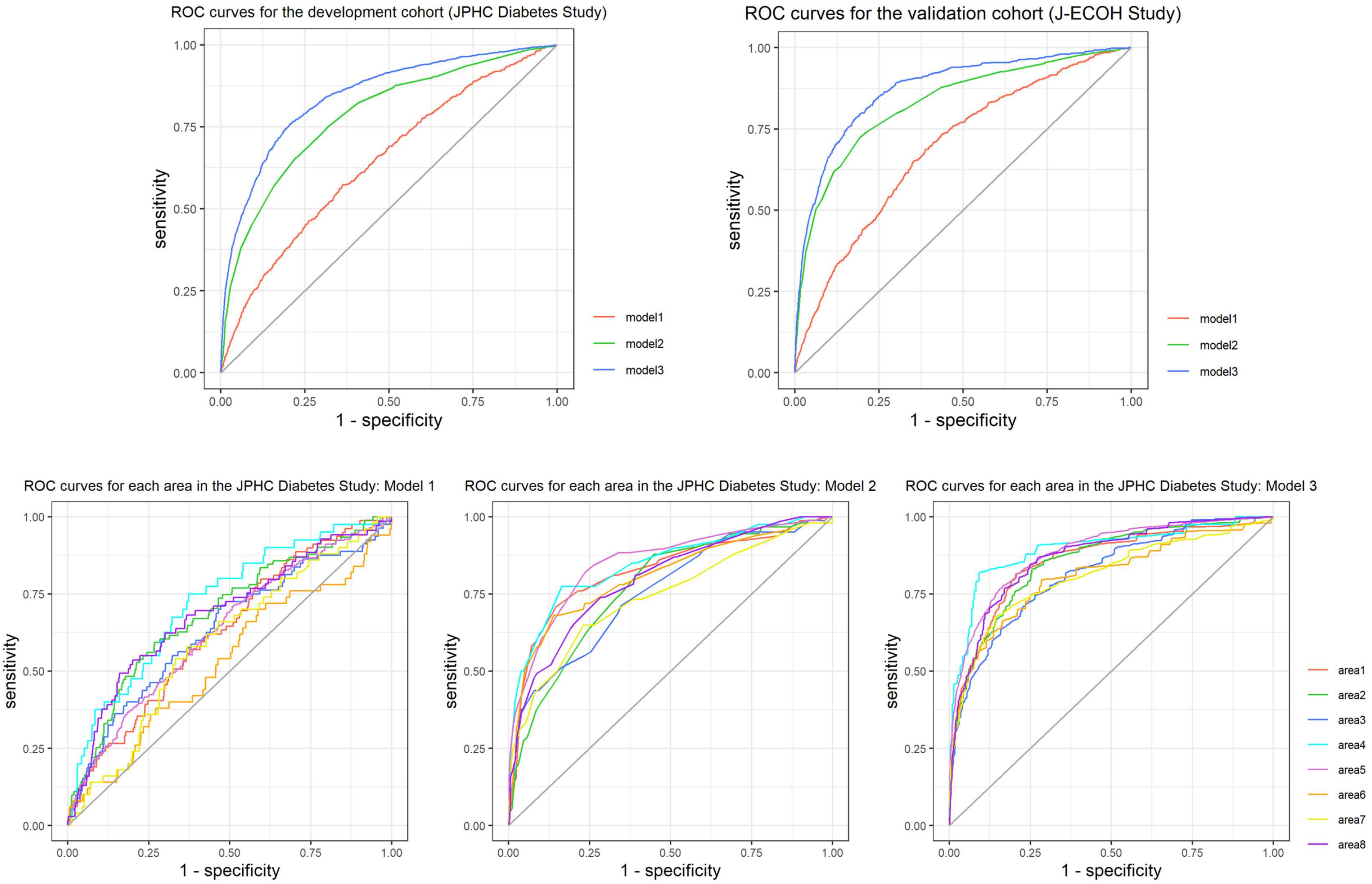

Responders to the baseline survey in the JPHC Diabetes Study (n = 28,362)
Subjects with diabetes at baseline (n = 2776)
Subjects cannot be diagnosed with or without diabetes at the baseline (n = 4)
Responders without diabetes at the baseline survey (n = 25,582)
Responders to the 6-year follow-up survey (n = 1625)
Non-responders to the 5-year follow-up survey (n = 12,964)
Subjects cannot be diagnosed with or without diabetes at the 5-year follow-up survey (n = 7)
Responders to both the baseline and the 5-year follow-up survey (n =10,986)
Responders to the baseline survey in the J-ECOH Diabetes Study (n = 19,827)
Participants with diabetes at the baseline survey (n = 2,663)
Responders without diabetes at the baseline survey (n =17,164)
Non-responders to the 5-year follow-up survey (n = 5,819)
Responders to both the baseline and the 5-year follow-up survey (n =11,345)

Figure 1.



ROC curves for the development cohort (JPHC Diabetes Study)

ROC curves for the validation cohort (J-ECOH Study)

ROC curves for each area in the JPHC Diabetes Study: Model 1

ROC curves for each area in the JPHC Diabetes Study: Model 2

ROC curves for each area in the JPHC Diabetes Study: Model 3

Figure 2.

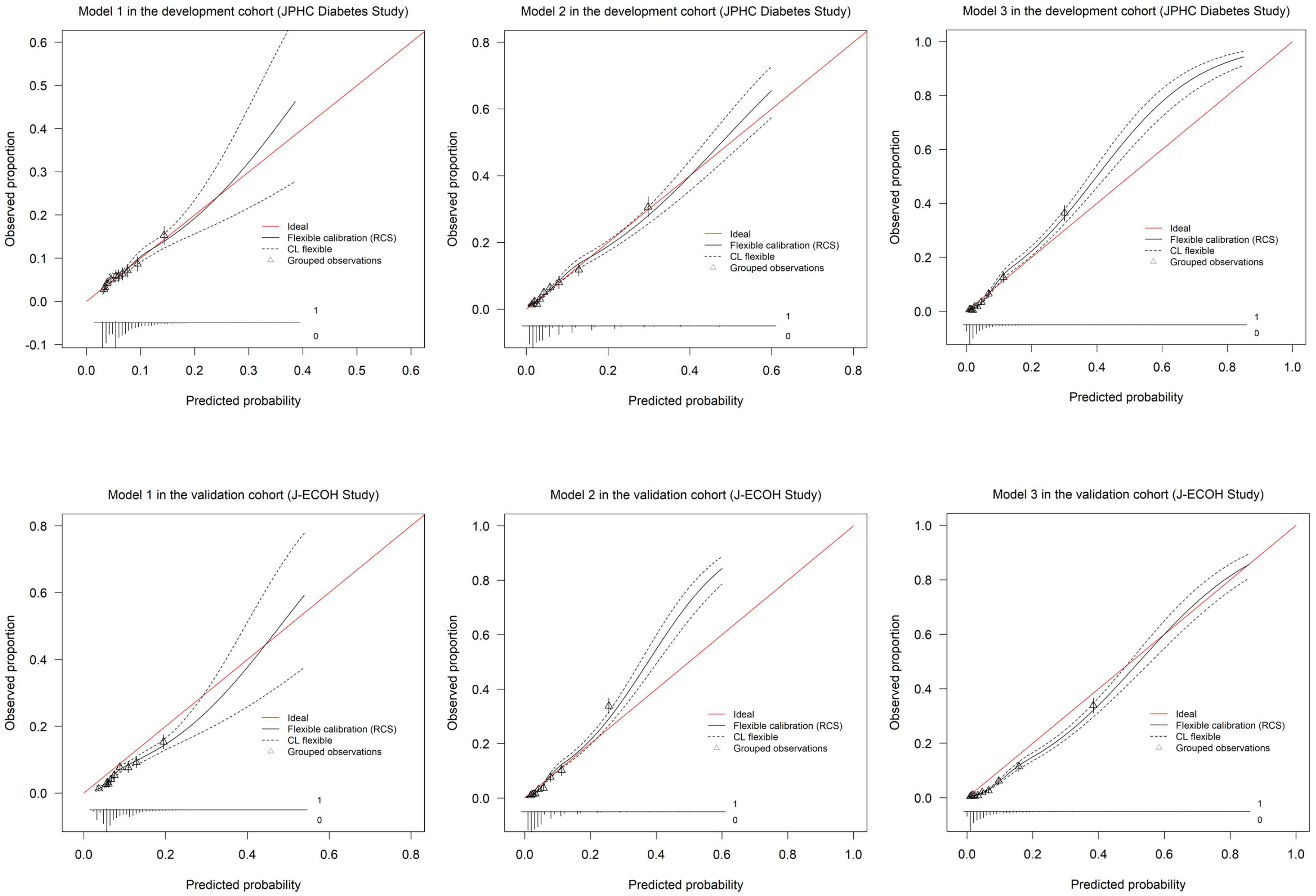

Model 1 in the development cohort (JPHC Diabetes Study)
Model 2 in the development cohort (JPHC Diabetes Study)
Model 3 in the development cohort (JPHC Diabetes Study)
Model 1 in the validation cohort (J-ECOH Study)
Model 2 in the validation cohort (J-ECOH Study)
Model 3 in the validation cohort (J-ECOH Study)
Observed proportion
Predicted probability
Ideal
Flexible calibration (RCS)
CL flexible
Grouped observations
1
0
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0

Figure 3.

Publication List

I. 主 論 文（本人を筆頭とする原著論文）

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

Xu, J., Goto, A., Konishi, M., Kato, M., Mizoue, T., Terauchi, Y., Tsugane, S., Sawada, N., Noda, M., for the JPHC Study Group[†]

([†]Japan Members listed in http://epi.ncc.go.jp/en/jphc/781/3838.html.)

Journal of Epidemiology. JE20220329. Doi: 10.2188/jea.JE20220329. Advanced online publication.2023

II. 参 考 論 文（主論文の内容以外の論文）

1. Prediction models for neutralization activity against emerging SARS-CoV-2 variants: A cross-sectional study
   Goto, A., Miyakawa, K., Nakayama, I., Yagome, S., **Xu, J.**, Kaneko, M., Ohtake, N., Kato, H., and Ryo A.:
   Frontiers in Microbiology. Vol 14, pp. 1～10, 2023

2. Usual source and better quality of primary care are associated with lower loneliness scores: a cross-sectional study
   Kaneko, M., Shinoda, S., Nakayama, I., **Xu, J.**, Yagome, S., Goto, A.:
   Family Practice. cmad049, pp. 1～9, 2023