



Prevalence and predictive modeling of undiagnosed diabetes and impaired fasting glucose in Taiwan: a Taiwan Biobank study

Ren-Hua Chung ¹, Shao-Yuan Chuang ¹, Ying-Erh Chen,² Guo-Hung Li,¹ Chang-Hsun Hsieh,³ Hung-Yi Chiou,^{1,4} Chao A Hsiung¹

To cite: Chung R-H, Chuang S-Y, Chen Y-E, *et al.* Prevalence and predictive modeling of undiagnosed diabetes and impaired fasting glucose in Taiwan: a Taiwan Biobank study. *BMJ Open Diab Res Care* 2023;**11**:e003423. doi:10.1136/bmjdr-2023-003423

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjdr-2023-003423>).

Received 23 March 2023
Accepted 6 June 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Ren-Hua Chung;
rchung@nhri.edu.tw

ABSTRACT

Introduction We investigated the prevalence of undiagnosed diabetes and impaired fasting glucose (IFG) in individuals without known diabetes in Taiwan and developed a risk prediction model for identifying undiagnosed diabetes and IFG.

Research design and methods Using data from a large population-based Taiwan Biobank study linked with the National Health Insurance Research Database, we estimated the standardized prevalence of undiagnosed diabetes and IFG between 2012 and 2020. We used the forward continuation ratio model with the Lasso penalty, modeling undiagnosed diabetes, IFG, and healthy reference group (individuals without diabetes or IFG) as three ordinal outcomes, to identify the risk factors and construct the prediction model. Two models were created: Model 1 predicts undiagnosed diabetes, IFG_110 (ie, fasting glucose between 110 mg/dL and 125 mg/dL), and the healthy reference group, while Model 2 predicts undiagnosed diabetes, IFG_100 (ie, fasting glucose between 100 mg/dL and 125 mg/dL), and the healthy reference group.

Results The standardized prevalence of undiagnosed diabetes for 2012–2014, 2015–2016, 2017–2018, and 2019–2020 was 1.11%, 0.99%, 1.16%, and 0.99%, respectively. For these periods, the standardized prevalence of IFG_110 and IFG_100 was 4.49%, 3.73%, 4.30%, and 4.66% and 21.0%, 18.26%, 20.16%, and 21.08%, respectively. Significant risk prediction factors were age, body mass index, waist to hip ratio, education level, personal monthly income, betel nut chewing, self-reported hypertension, and family history of diabetes. The area under the curve (AUC) for predicting undiagnosed diabetes in Models 1 and 2 was 80.39% and 77.87%, respectively. The AUC for predicting undiagnosed diabetes or IFG in Models 1 and 2 was 78.25% and 74.39%, respectively.

Conclusions Our results showed the changes in the prevalence of undiagnosed diabetes and IFG. The identified risk factors and the prediction models could be helpful in identifying individuals with undiagnosed diabetes or individuals with a high risk of developing diabetes in Taiwan.

INTRODUCTION

Diabetes, characterized by elevated blood glucose levels, is often associated with complications such as kidney disease, eye damage,

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ It was estimated that among adults aged 20–79 years with diabetes globally, 44.7% are undiagnosed, and maximally 70% of individuals with pre-diabetes will develop diabetes. However, early detection of diabetes or pre-diabetes can prevent the development of further complications or the development of the disease.

WHAT THIS STUDY ADDS

⇒ The prevalence of undiagnosed diabetes has not been estimated in a large study in Taiwan, and a risk prediction model designed specifically for undiagnosed diabetes and pre-diabetes has not been developed for the Taiwan population. We estimated the prevalence of undiagnosed diabetes and impaired fasting glucose (IFG) in individuals without known diabetes in Taiwan and constructed a risk prediction model for identifying undiagnosed diabetes and IFG.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our results showed increased trends in the prevalence of undiagnosed diabetes and IFG in Taiwan. The identified risk factors and the prediction model could be helpful in identifying individuals with undiagnosed diabetes or individuals with a high risk of developing diabetes in Taiwan.

and heart and blood vessel diseases.¹ Globally, nearly 44.7% of adults aged 20–79 years with diabetes are unaware of their condition (ie, undiagnosed diabetes).² Further, approximately 35% of newly diagnosed patients with diabetes are discovered to already have developed complications, such as retinopathy, neuropathy and ischemic heart disease.³ Hence, early diagnosis and effective management of diabetes are important to prevent the development of further complications and to reduce the clinical and economic burden.⁴

Pre-diabetes, characterized by blood glucose levels higher than normal but lower

than the thresholds for diabetes,⁵ is variably defined by different professional organizations, such as the WHO, the American Diabetes Association (ADA), and the International Expert Committee.⁶ The WHO, for instance, defined pre-diabetes as fasting plasma glucose levels between 110 and 125 mg/dL (ie, impaired fasting glucose (IFG)) or 2-hour plasma glucose using the 75 g oral glucose tolerance test (OGTT) between 140 and 199 mg/dL (ie, impaired glucose tolerance (IGT)). Based on the WHO standard, the International Diabetes Federation (IDF) Diabetes Atlas (10th edition) reported that in 2021, the global prevalence of IFG was between 2.5% and 10% and the global prevalence of IGT was between 5.4% and 12.9%.⁷ It has been estimated that up to 70% of individuals with pre-diabetes will develop diabetes.⁵ However, lifestyle intervention can effectively delay or prevent disease progression,^{8–11} reinforcing the importance of early detection in the pre-diabetes stage.

In 2014, among the newly diagnosed cases of both type 1 and type 2 diabetes in Taiwan, type 2 diabetes constituted a significant majority (99.8%).¹² Furthermore, within the same year, the prevalence of type 2 diabetes exhibited a notable increase, escalating to 9.32% from 6.38% in 2008.¹² A separate study, which analyzed 1096 patients admitted to a specific hospital in Taiwan, found that between 24.5% and 50% of these patients, depending on the type of medical service received, were undiagnosed diabetes.¹³ The prevalence of undiagnosed diabetes based on large population-based studies in Taiwan has not been reported in the literature. The prevalence of IFG (defined as fasting glucose levels between 100 and 125 mg/dL) in Taiwan was estimated at 16.2%–35.5% for different age groups in the Nutrition and Health Survey of 2013–2016.¹⁴

Several prediction models have been constructed for detecting undiagnosed diabetes and pre-diabetes, such as the Danish Risk Score,¹⁵ the Leicester Risk Score,¹⁶ the Finnish Diabetes Risk Score (FINDRISC),¹⁷ the Indian Diabetes Risk Score,^{18 19} and the Taiwan Diabetes Risk Score (DRS).²⁰ These models, based on simple non-invasive questionnaires, achieved area under the curves (AUCs) between 67% and 80% for detecting diabetes in different populations. Common predictors used in these models included age, sex, body mass index (BMI), known hypertension, and family history of diabetes. Eleven existing diabetes risk scores, including the Danish Risk Score, FINDRISC, and Taiwan DRS, were evaluated in a Taiwanese cohort, and their AUCs for detecting diabetes ranged between 67% and 77%,^{20 21} suggesting that commonly used risk predictors for diabetes are also applicable to the Taiwan population.

Studies and national surveys have shown that the prevalence of undiagnosed diabetes in Asians is generally higher than that in other populations.^{7 22} For example, in the USA between 2017 and 2020, the crude prevalence of undiagnosed diabetes among non-Hispanic Asians was 5.4%, the highest among all ethnicities. This may be due to the lower average BMI in Asians, which leads

to less frequent screening for diabetes among Asians.²² Hence, a simple and fast diabetes screening tool with high accuracy would be particularly helpful for the Asian population.

In this study, we investigated the prevalence of undiagnosed diabetes and IFG and the trends in the prevalence over the years in the population without known diabetes in Taiwan, which is the population targeted for diabetes screening. Using a machine learning algorithm, we selected significant prediction variables and constructed risk prediction models for predicting undiagnosed diabetes, IFG, and healthy reference group (ie, individuals without diabetes or IFG) as three ordinal outcomes. This is contrary to the aforementioned models that considered binary outcomes (eg, undiagnosed diabetes vs non-diabetes).

RESEARCH DESIGN AND METHODS

Study participants

The Taiwan Biobank is a population-based study, which has recruited more than 150 000 individuals between the ages of 30 and 70 in Taiwan since 2008.²³ Each participant provided information through questionnaires such as self-reported disease status and family history of diseases, and underwent physical examinations and blood and urine tests including fasting glucose and glycated hemoglobin (HbA1c) at recruitment. Approximately 115 000 individuals were genotyped with genome-wide association study (GWAS) single-nucleotide polymorphism (SNP) arrays, which allowed us to perform standard GWAS quality control (QC) procedures, such as removing closely related individuals and those with discordance between self-reported gender and biological sex estimated from the genetic data. Detailed GWAS QC procedures are provided in the online supplemental materials. Individuals who fasted for less than 8 hours were excluded from the analysis.

Thereafter, the individuals were linked to their medical records in the National Health Insurance Research Database (NHIRD) between 2009 and 2020. The NHIRD contains enrollees' demographic data, medical records, and expenditure claims from outpatient, inpatient, and ambulatory care, and data associated with contracted pharmacies for reimbursement purposes.²⁴ The diagnosis codes (International Classification of Diseases, Ninth Revision (ICD-9) and International Classification of Diseases, Tenth Revision (ICD-10) codes) and prescription drug codes of the Taiwan Biobank samples' outpatient, ambulatory care, and inpatient services from 3 years before their recruitment were extracted. Based on the diagnosis and drug codes, we further excluded individuals who were using hypoglycemic drugs or who had been diagnosed with hyperthyroidism, Cushing's syndrome, or acromegaly—a common cause of hyperglycemia in Taiwan—before their recruitment.

Outcome definitions

Diabetes was classified into two categories for this study: known and undiagnosed. ‘Known diabetes’ referred to patients who had previously self-reported having type 1, type 2, or gestational diabetes in the questionnaires at recruitment. Additionally, this category also included those with a medical history of diabetes before they were recruited to the Taiwan Biobank. This was determined based on whether the patient had at least three clinical visits or at least one hospitalization within the past year, where diabetes was diagnosed according to ICD-9 codes (beginning with 250) or ICD-10 codes (beginning with E08-E13) from the NHIRD records. Individuals with known diabetes were excluded from our analysis.

On the other hand, “undiagnosed diabetes” referred to individuals who did not fall under the “known diabetes” category as defined above but showed elevated levels of both fasting glucose (≥ 126 mg/dL) and HbA1c ($\geq 6.5\%$) based on blood test results taken at the time of their recruitment. This threshold for defining undiagnosed diabetes is consistent with the criteria suggested by Selvin *et al*²⁵ when fasting glucose and HbA1c measurements are available for each study participant, which is the case with the Taiwan Biobank. The threshold has also been used in estimating the prevalence of undiagnosed diabetes in the USA.²⁶

IFG was defined as those who were not diabetic but had fasting glucose between 110 and 125 mg/dL, according to the WHO standard,²⁷ referred to as IFG_110. We also defined IFG based on the ADA standard (ie, fasting glucose between 100 and 125 mg/dL), referred to as IFG_100. Lastly, a healthy reference group referred to individuals who were not diabetic or had IFG.

Prevalence of undiagnosed diabetes and IFG

The crude prevalence of undiagnosed diabetes was calculated as the proportion of individuals with undiagnosed diabetes among those with undiagnosed diabetes, IFG, or healthy reference group. Similarly, the crude prevalence of IFG was calculated as the proportion of individuals with IFG among those with undiagnosed diabetes, IFG, or healthy reference group. Age-specific and sex-specific crude prevalence was calculated. A standardized prevalence was thereafter calculated using the direct method, based on the 2020 census data of the general Taiwan population as the standard. The census data were obtained from the Department of Household Registration of the Ministry of Interior in Taiwan. We partitioned the samples into four groups, each with a minimum of 10 000 participants, covering a 2- to 3-year span: 2012–2014 for the first group, 2015–2016 for the second, 2017–2018 for the third and 2019–2020 for the fourth group. We calculated each group’s prevalence, enabling us to investigate changes in the prevalence of undiagnosed diabetes and IFG over time.

Risk prediction models for undiagnosed diabetes and IFG

We compiled a list of 150 variables from the Taiwan Biobank survey data including basic information such as age, sex, and BMI; health behaviors such as drinking, smoking, and physical activity; female-specific variables such as age of menarche and number of pregnancies; and self-reported diseases (including first-degree relatives) such as hypertension, hyperlipidemia, and glaucoma, which can all be completed by self-assessment of an individual at home. Variables with missing rates $>10\%$ were excluded. The mice package in R,²⁸ which implements the multivariate imputation by chained equations, was used to impute the missing values in the remaining variables.

We randomly divided 80% of the samples into a training dataset and the remaining 20% as a testing dataset. The outcomes were considered as ordinal (healthy reference group, IFG, and undiagnosed diabetes). Two models were constructed: Model 1 considered the healthy reference group, IFG_110, and undiagnosed diabetes; Model 2 considered the healthy reference group, IFG_100, and undiagnosed diabetes. The forward continuation ratio model with the Lasso penalty implemented in the R package glmnet²⁹ was applied to the training dataset to select significant risk predictors. The hyperparameter in the Lasso penalty (the λ value) was selected based on the best-fitted model using the Bayesian Information Criterion. These procedures for model training were performed on the training dataset. The significant risk predictors were used to create the final prediction model. The probability of undiagnosed diabetes ($p1$) and the probability of undiagnosed diabetes or IFG ($p2$) were calculated from the final model using the R package VGAM.³⁰ The probability $p1$ was used to predict undiagnosed diabetes versus non-diabetes (including IFG and healthy reference group), and $p2$ was used to predict undiagnosed diabetes or IFG versus healthy reference group using the testing dataset.

AUC was calculated to evaluate the performance of the models. Optimal cut-off values to determine the sensitivities and specificities were selected using the Youden index based on the receiver operating characteristics (ROC) curves. Models 1 and 2 were also applied to predict undiagnosed diabetes or IFG+/HbA1c+ (fasting glucose between 110 and 125 mg/dL and HbA1c between 6.0% and 6.4% as defined by Washirasaksiri *et al*³¹) versus healthy reference group, allowing us to evaluate the performance of the trained models for different definitions of pre-diabetes. It is of interest to note that the IFG+/HbA1c+ subgroup has been shown to have a high risk of 5-year diabetes incidence, making it important to evaluate how our model performed for identifying this subgroup.

External validation analysis

We used the CardioVascular Disease risk FACTors Two-township Study (CVDFACTS), which is a community-based cohort study, to validate the prediction results.

CVDFACTS investigates risk factors for cardiovascular diseases in Taiwan.³² Approximately 6000 individuals were recruited between 1991 and 1993 in two towns, Chu-Dung and Pu-Tzu. Individuals who had a history of stroke, had fasted for less than 8 hours, or were not covered by the National Health Insurance were excluded from the baseline in the study. Several follow-up surveys and examinations were conducted, and this study used the fifth follow-up data, which was collected between 1999 and 2002 for the analysis. Data from individuals aged between 30 and 70 years old were extracted, which resulted in a total of 1481 samples for the analysis. The significant risk prediction variables selected from the Taiwan Biobank samples were extracted from the CVDFACTS survey and examination data. Known diabetes in CVDFACTS was defined as individuals who self-reported as having diabetes. HbA1c was not measured in CVDFACTS, hence, undiagnosed diabetes was defined as those who were not known having diabetes but had fasting glucose ≥ 126 mg/dL. The same definitions used for the Taiwan Biobank samples were applied to define IFG and healthy reference group. The risk prediction models constructed

using the Taiwan Biobank training dataset were applied to the CVDFACTS samples. Sensitivities and specificities were calculated using the optimal cut-off values selected for the testing dataset from the Taiwan Biobank.

RESULTS

Figure 1 shows our analysis flowchart. After the sample QC, 64 875 individuals remained for the analyses. Table 1 shows the characteristics of the four sample groups stratified by years. There were 12,572, 22,295, 15,990, and 14 018 individuals without known diabetes for the groups of 2012–2014, 2015–2016, 2017–2018, and 2019–2020, respectively. All groups comprised a majority of females (approximately 63%–66%). The mean BMI, fasting glucose level, and HbA1c and the proportion of self-reported hypertension were all higher in males compared with females.

Online supplemental figure S1 in the online supplemental materials shows the age-specific and sex-specific crude prevalence of undiagnosed diabetes in the four groups. Generally, males had higher undiagnosed rates

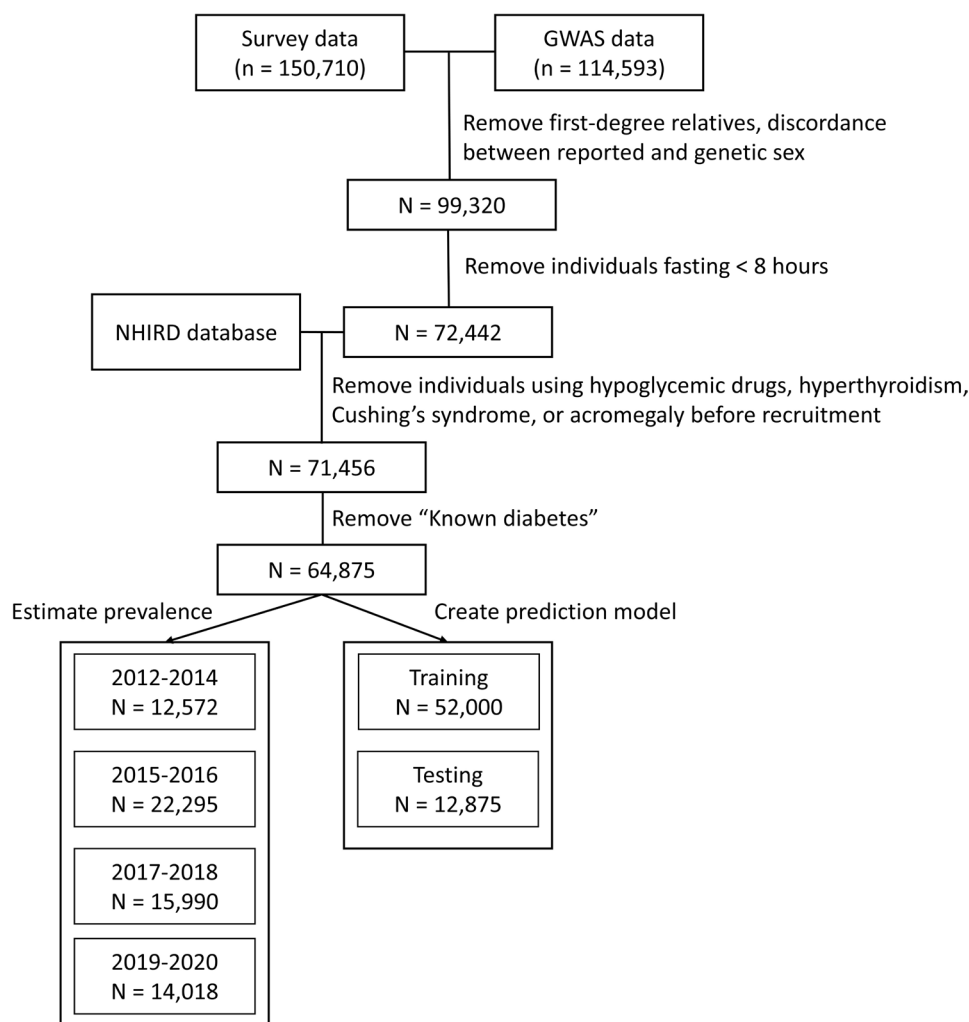


Figure 1 Flowchart of our analysis steps. GWAS, genome-wide association study; NHIRD, National Health Insurance Research Database.

Table 1 Characteristics of participant groups divided by the year of recruitment (2- to 3-year intervals)

	2012–2014		2015–2016		2017–2018		2019–2020	
	Male	Female	Male	Female	Male	Female	Male	Female
Sample size*	4314 (34.3%)	8258 (65.7%)	8064 (36.1%)	14 231 (63.9%)	5737 (35.8%)	10 253 (64.2%)	5175 (36.9%)	8843 (63.1%)
Age†	50.81 (10.94)	51.04 (9.99)	48.93 (11.14)	49.12 (10.46)	48.74 (11.48)	48.69 (10.80)	48.18 (11.31)	48.09 (10.69)
BMI†	24.97 (3.31)	23.44 (3.51)	25.22 (3.45)	23.35 (3.66)	25.40 (3.53)	23.49 (3.75)	25.47 (3.59)	23.48 (3.85)
Hypertension‡	15.22%	9.33%	13.95%	7.56%	12.39%	7.46%	12.63%	6.10%
Fasting glucose†	97.06 (15.29)	92.75 (13.14)	95.67 (13.87)	91.67 (12.80)	96.43 (15.22)	92.08 (13.02)	96.45 (14.33)	92.38 (13.61)
HbA1c†	5.67 (0.62)	5.62 (0.51)	5.65 (0.57)	5.59 (0.50)	5.77 (0.60)	5.70 (0.51)	5.66 (0.57)	5.59 (0.53)

*Sample number and its proportion.
 †Mean and SE.
 ‡Proportion of self-reported hypertension.
 BMI, body mass index; HbA1c, glycated hemoglobin.

than females within groups. The rates generally increased with age, except for males aged between 50 and 59 in the years 2012–2014 and 2019–2020, who showed the highest rates compared with other age groups. Online supplemental figures S2 and S3 in the online supplemental materials show the age-specific and sex-specific crude prevalence of IFG₁₁₀ and IFG₁₀₀ in the four groups. Males also had higher rates of IFG than females within groups, which also increased with age.

Figure 2 shows the standardized prevalence of undiagnosed diabetes and IFG. The standardized prevalence of undiagnosed diabetes was 1.11%, 0.99%, 1.16%, and 0.99% for 2012–2014, 2015–2016, 2017–2018, and 2019–2020, respectively. Moreover, the standardized prevalence of IFG₁₁₀ for the four groups was 4.49%, 3.73%, 4.30%, and 4.66%, while the standardized prevalence of IFG₁₀₀ was 21.0%, 18.26%, 20.16%, and 21.08%. No obvious increasing or decreasing trend over the years was observed for either the prevalence of undiagnosed diabetes or IFG.

For identifying risk factors, there were 140 variables remained after QC. Table 2 shows the significant variables selected from the Lasso regression based on 140 variables and the estimates of the effects of the significant variables based on the forward continuation ratio model for Models 1 and 2. Common variables included in other

prediction models such as age, BMI, waist to hip ratio (WHR), self-reported hypertension, and family history of diabetes were selected. Model 1 also included education levels and betel nut chewing, which may be specific to the Western Pacific or Taiwan population. On the other hand, Model 2 included alcohol consumption and the personal monthly income that was not included in Model 1. These additional variables provide further insight into the risk factors associated with undiagnosed diabetes and IFG.

Table 3 shows the AUCs for predicting undiagnosed diabetes versus non-diabetes (including IFG and healthy reference group) and for predicting undiagnosed diabetes or IFG versus healthy reference group in the overall, male, and female samples based on the testing dataset using Models 1 and 2. Generally, predicting undiagnosed diabetes alone yielded higher AUCs than predicting undiagnosed diabetes or IFG, and prediction in females also had higher AUCs than prediction in males. Furthermore, Model 1, which defined IFG using a more stringent threshold, generally demonstrated higher AUCs compared with Model 2. We further applied Models 1 and 2 to predict undiagnosed diabetes or IFG+/HbA1c+ versus healthy reference group, and the results are also shown in table 3. The AUCs were all higher than those for predicting undiagnosed diabetes or IFG versus

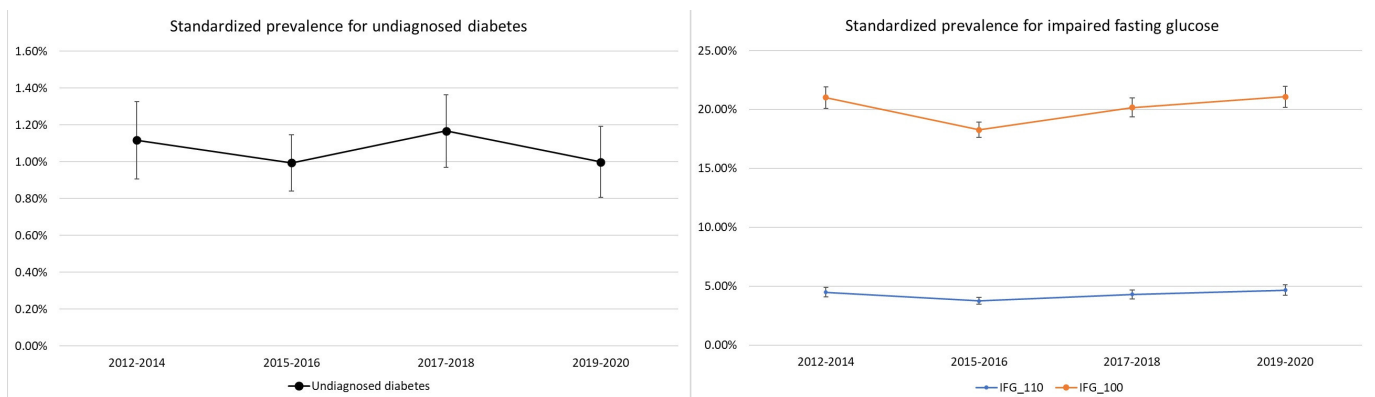


Figure 2 The standardized prevalence and its 95% CIs for undiagnosed diabetes, IFG₁₁₀ (fasting glucose between 110 and 125 mg/dL), and IFG₁₀₀ (fasting glucose between 100 and 125 mg/dL) over the years. IFG, impaired fasting glucose.

Table 2 Significant prediction variables selected by the Lasso regression for Models 1 and 2

Variable	Model 1			Model 2		
	Beta	SE	P value	Beta	SE	P value
Intercept 1	-12.935	0.406	0	-9.690	0.239	0
Intercept 2	-11.971	0.416	0	-11.808	0.249	0
Age (years)	0.046	0.002	5.32E-90	0.053	0.001	0
Sex	-0.215	0.047	5.15E-06	-0.297	0.036	7.89E-16
BMI (kg/m ²)	0.132	0.005	2.83E-128	0.100	0.007	1.24E-41
Waist to hip ratio	4.950	0.385	7.84E-38	3.256	0.227	1.77E-46
Education level*	-0.071	0.021	8.71E-04	-0.070	0.013	1.06E-07
Family history of diabetes†	0.593	0.041	1.13E-46	0.460	0.024	4.79E-79
Betel nut chewing‡	0.184	0.072	1.08E-02			
Self-reported hypertension§				0.139	0.034	4.59E-05
Body weight (kg)				0.008	0.002	1.03E-03
Alcohol consumption¶				0.138	0.021	2.16E-10
Personal monthly income**				0.011	0.003	3.64E-04

*Education level was coded as follows: 1: Never attended school/illiterate; 2: Self-taught/literate; 3: Elementary school; 4: Middle/Junior high school; 5: High school; 6: College; 7: Graduate school.

†1 was assigned for individuals whose parents and siblings had diabetes; 0 was assigned otherwise.

‡Betel nut chewing status was coded as: 1: Never or only chewed once or twice; 2: Frequent chewer.

§1 was assigned for individuals with self-reported hypertension; 0 was assigned otherwise.

¶Alcohol consumption was coded as follows: (1): No habit of drinking or only occasionally; (2): Already quit drinking; (3): Still drinking currently.

**Personal monthly income was coded into 22 levels, ranging from no income to over US\$6000. Each level represented an increase of US\$300.

healthy reference group, once again highlighting that AUCs were higher when the IFG definition was more stringent.

The optimal cut-off thresholds selected based on the Youden index from the ROC curves generated in the overall sample in the testing dataset were used to calculate the sensitivities and specificities. The cut-off threshold of $p1$ for predicting undiagnosed diabetes was 0.0065, and

the threshold of $p2$ for predicting undiagnosed diabetes or IFG was 0.0367 in Model 1. In Model 2, the thresholds for $p1$ and $p2$ were 0.0053 and 0.1773, respectively. **Figure 3** shows the results for the testing dataset from the Taiwan Biobank and the validation dataset from the CVDFACTS in Model 1. The overall sensitivity and specificity were 75.6% and 72.4% for the testing dataset, and 67.6% and 61.9% for the validation dataset, respectively,

Table 3 Area under the curves with their 95% CIs for predicting undiagnosed diabetes and IFG

	Model 1	Model 2
Undiagnosed diabetes versus non-diabetes		
Overall	80.39% (76.86%, 83.92%)	77.87% (74.22%, 81.51%)
Male	80.43% (75.01%, 85.85%)	77.50% (71.90%, 83.11%)
Female	81.10% (76.63%, 85.57%)	79.18% (74.63%, 83.72%)
Undiagnosed diabetes or IFG versus healthy reference group		
Overall	78.25% (76.52%, 79.98%)	74.39% (73.33%, 75.45%)
Male	75.14% (72.28%, 78.00%)	68.72% (66.97%, 70.46%)
Female	79.92% (77.71%, 82.11%)	75.86% (74.43%, 77.29%)
Undiagnosed diabetes or IFG+/HbA1c+ versus healthy reference group		
Overall	79.35% (77.56%, 81.14%)	78.32% (76.50%, 80.15%)
Male	77.11% (74.17%, 80.06%)	76.09% (73.12%, 79.06%)
Female	81.02% (78.79%, 83.25%)	80.49% (78.21%, 82.76%)

HbA1c, glycated hemoglobin; IFG, impaired fasting glucose.

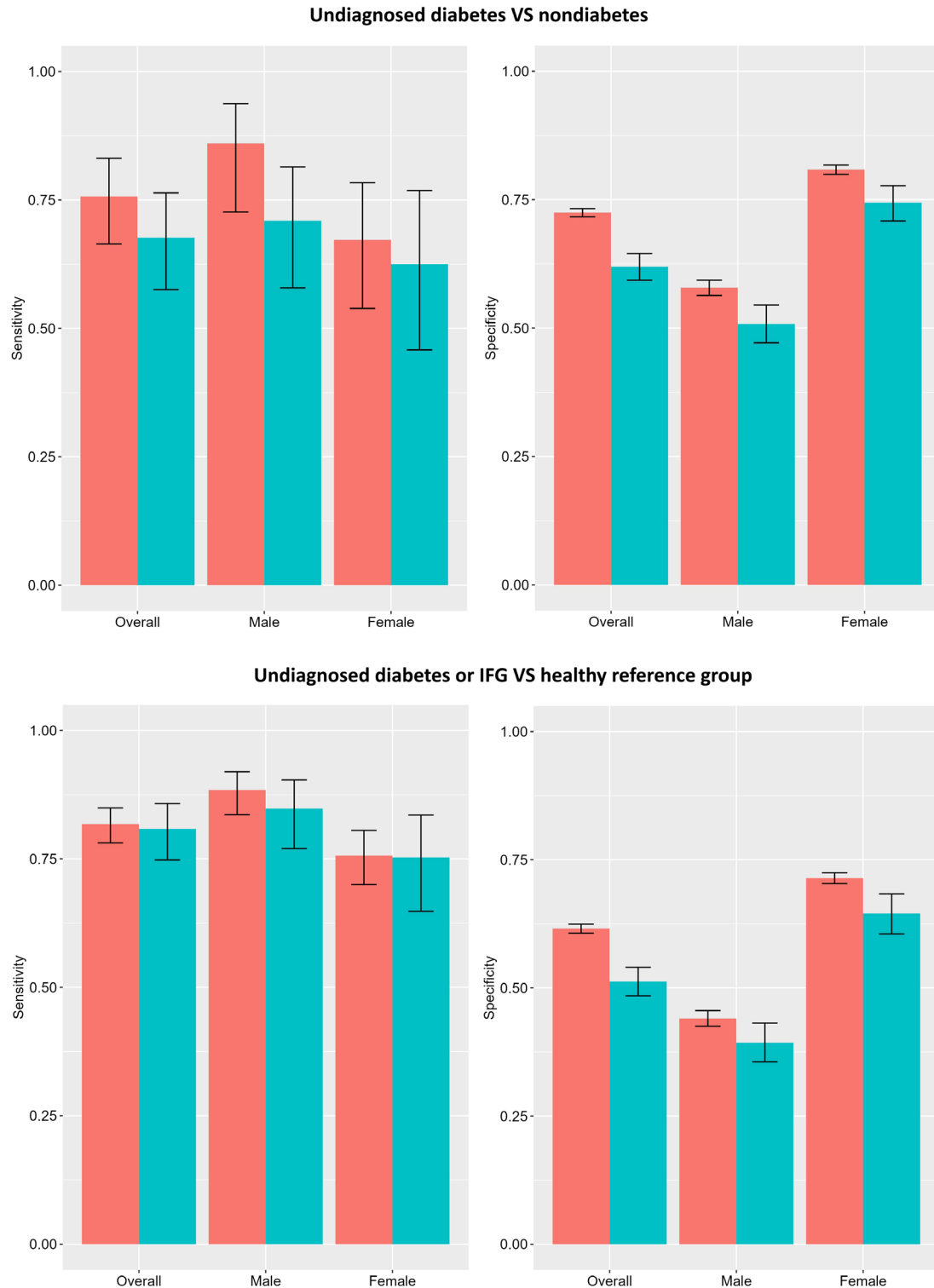


Figure 3 Sensitivities and specificities for Model 1. The upper section of the figure illustrates the sensitivities (left) and specificities (right) of the model in predicting undiagnosed diabetes versus non-diabetes. The lower section of the figure shows the sensitivities (left) and specificities (right) for predicting undiagnosed diabetes or IFG₁₁₀ (fasting glucose between 110 and 125 mg/dL) versus healthy reference group. The sensitivities and specificities were calculated in the overall, male, and female samples in the Taiwan Biobank testing dataset (TWB) and the external CardioVascular Disease risk FACTors Two-township Study (CVDFACTS) validation dataset. The 95% CIs are depicted as error bars in the figure. The results for TWB and CVDFACTS are represented in orange and green bars, respectively.

for predicting undiagnosed diabetes in Model 1. For predicting undiagnosed diabetes or IFG, the overall sensitivity was higher (81.7%), but specificity was lower

(61.5%) than those for predicting undiagnosed diabetes. The same trend was observed for the validation dataset. In Model 2, the sensitivity was higher for predicting

undiagnosed diabetes compared with predicting undiagnosed diabetes or IFG, while the specificities were similar, as shown in online supplemental figure S4 in the online supplemental materials. The estimates from the CVDFACTS were generally lower than the estimates from the Taiwan Biobank testing dataset. This is not surprising since the cut-off thresholds were optimized based on the Taiwan Biobank testing dataset. We also calculated the sensitivity and specificity using only male or female samples. As shown in figure 3 and online supplemental figure S4, the sensitivities were higher in males than in females, while the specificities were higher in females than in males.

DISCUSSION

To our knowledge, the prevalence of undiagnosed diabetes and IFG in Taiwan using the WHO's definition has not been reported in the literature. For example, in the IDF Diabetes Atlas (10th edition) report, the undiagnosed rates of diabetes and prevalence of IFG in Taiwan were extrapolated from data in nearby countries with similar ethnicity, language, and World Bank income classification. Our results filled this gap as it provides useful information that will improve the estimates of both prevalence in the Western Pacific region and globally.

Our results revealed a minor variation in the prevalence of undiagnosed diabetes in Taiwan, ranging from 0.99% to 1.16% during 2012–2020, without apparent increasing or decreasing trend. The estimates were close to the recent estimates in the USA (from 1.10% to 1.23%) using the same definition of undiagnosed diabetes.²⁶ However, the prevalence estimates from our study were lower than those calculated for the population without known diabetes in Japan (2.9%–5.6%).^{33 34} This is expected, as their definition for undiagnosed diabetes was broader, requiring either fasting glucose ≥ 126 mg/dL or HbA1c $\geq 6.5\%$, while our definition required both criteria to be met. Furthermore, the prevalence of IFG based on the WHO definition in our study was estimated between 3.73% and 4.66% from 2012 to 2020. This estimate aligns closely with the extrapolated prevalence of 4.5% in Taiwan in 2021, as reported in the IDF Diabetes Atlas (10th edition). On the other hand, the prevalence of IFG based on the ADA definition in Taiwan was estimated to be between 18.26% and 21.08% from 2012 to 2020, which is higher than 16% found in the Thai population.³⁵

Our variable selection procedure identified both BMI and WHR as significant predictors. This is consistent with the finding based on an Indian population that a composite measure of BMI and waist circumference resulted in a better predictor for type 2 diabetes than either BMI or waist circumference alone.¹⁹ More interestingly, our variable selection procedure identified education level and betel nut chewing in Model 1 and personal monthly income in Model 2 as risk factors not commonly included in predicting undiagnosed diabetes

in literatures. For example, in the review paper of risk prediction models for incident or undiagnosed diabetes by Collins *et al*,³⁶ education level was included as a risk predictor for incident diabetes in only one model that is also based on a Taiwan population,³⁷ and none of the models reviewed in the paper included betel nut chewing. A previous study in Taiwan reported that education level was negatively associated with 5-year diabetes incidence,³⁷ while another showed that patients with diabetes with higher education levels had better knowledge of diabetes.³⁸ As discussed in Hill-Briggs *et al*,³⁹ while education level and personal income are correlated, they have distinct implications for health outcomes. In addition, a higher prevalence of diabetes has been found in populations with lower income in studies from the USA and Canada.^{39 40} A study from Taiwan also suggests that poverty is associated with not only diabetes incidence but also inequality of diabetes care.⁴¹ Hence, as suggested by Sun *et al*,³⁷ considering social deprivation in diabetes prevention is important in reducing health inequalities. Furthermore, a study showed that the prevalence of betel nut chewing is high in Taiwan, Mainland China, Malaysia, Indonesia, Nepal, and Sri Lanka.⁴² In Taiwan, this prevalence was approximately 7% in 2018.⁴³ Betel nut chewing has been found to be associated with current and incident diabetes.^{44 45} Our result provides information for identifying risk factors and developing predicting models for undiagnosed diabetes and IFG in countries with a high prevalence of betel nut chewing.

The performance of our model in predicting undiagnosed diabetes (with an overall AUC of 80.39% for Model 1 and 77.87% for Model 2) is comparable to the models in the literature using simple questionnaires. For example, AUCs of 72%–80% were reported for predicting undiagnosed diabetes.^{15–17 33 46 47} Our model also has a higher AUC than the AUC of the Taiwan DRS (reported as 76% by Li *et al*²⁰). Moreover, our model resulted in the highest AUC for predicting undiagnosed diabetes or IFG (overall AUC of 78.25% for Model 1 and 74.39% for Model 2) compared with previous studies which reported AUCs of 67%–72%.^{16 48–50} This could be because we specifically considered the healthy reference group, IFG, and undiagnosed diabetes as three ordinal outcomes in the same model. Interestingly, when our models were applied to identify the high-risk pre-diabetes subgroup (IFG+/HbA1c+) or undiagnosed diabetes, the AUCs further increased to 79.35% and 78.32% for Models 1 and 2, respectively.

There were several strengths in our study. Large-scale biobanks usually contain a high proportion of related samples. The samples included in this study all had SNP array data, which allowed us to perform stringent sample QC and identify unrelated individuals. Furthermore, linking the Taiwan Biobank with NHIRD allowed us to use the survey and blood test results from the Taiwan Biobank and medical records from NHIRD to robustly define known and undiagnosed diabetes. Finally, the large sample size from the Taiwan Biobank allowed us

to estimate the prevalence and train and fine-tune the prediction model. A limitation of our study is that the 2-hour plasma glucose based on OGTT was not measured in the Taiwan Biobank. Hence, the prevalence of IGT and pre-diabetes, defined using both IFG and IGT, could not be estimated. Furthermore, in our study, undiagnosed diabetes was identified based on a single measurement of fasting glucose and HbA1c at recruitment, in contrast to the clinical practice of using repeated measurements for diabetes diagnosis. However, Selvin *et al*²⁵ have shown that using both fasting glucose and HbA1c measurements in one sample can yield a high positive predictive value for subsequent diagnosis, which effectively reduces the potential drawback of relying on a one-time measurement.

In conclusion, our study documented the current trends in the prevalence of undiagnosed diabetes and IFG in Taiwan. We also identified risk factors that are important for predicting undiagnosed diabetes and IFG. The prediction model will be useful in identifying individuals with undiagnosed diabetes or individuals with a high risk of developing diabetes in Taiwan.

Author affiliations

¹Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

²Department of Risk Management and Insurance, Tamkang University, Taipei, Taiwan

³Division of Endocrinology and Metabolism, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

⁴School of Public Health, College of Public Health, Taipei Medical University, Taipei, Taiwan

Acknowledgements We thank the participants from the Taiwan Biobank and CVDFACTS studies.

Contributors R-HC, Y-EC, C-HH, H-YC, and CAH designed the study. R-HC and G-HL performed the analyses. SYC performed the validation analysis. R-HC is the guarantor. All authors helped interpret the analysis results and approved the final manuscript.

Funding This study was supported by grants PH-111-GP-04 and PH-111-PP-10 from the National Health Research Institutes, and MOST 110-2314-B-400-023 from the National Science and Technology Council in Taiwan.

Competing interests None declared.

Ethics approval This study involves human participants and was approved by the institutional review board of National Health Research Institutes (reference no: EC1091202-E). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data may be obtained from a third party and are not publicly available. The Taiwan Biobank data can be applied through the Taiwan Biobank.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which

permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Ren-Hua Chung <http://orcid.org/0000-0002-9835-6333>

Shao-Yuan Chuang <http://orcid.org/0000-0003-2138-4771>

REFERENCES

- Forbes JM, Cooper ME. Mechanisms of diabetic complications. *Physiol Rev* 2013;93:137–88.
- Ogurtsova K, Guariguata L, Barengo NC, *et al*. IDF diabetes Atlas: global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res Clin Pract* 2022;183:109118.
- Gedebjerg A, Almdal TP, Berencsi K, *et al*. Prevalence of Micro- and Macrovascular diabetes complications at time of type 2 diabetes diagnosis and associated clinical characteristics: a cross-sectional baseline study of 6958 patients in the Danish DD2 cohort. *J Diabetes Complications* 2018;32:34–40.
- Evans M, Morgan AR, Patel D, *et al*. Risk prediction of the diabetes missing million: identifying individuals at high risk of diabetes and related complications. *Diabetes Ther* 2021;12:87–105.
- Tabák AG, Herder C, Rathmann W, *et al*. Prediabetes: a high-risk state for diabetes development. *Lancet* 2012;379:2279–90.
- Echouffo-Tcheugui JB, Selvin E. Prediabetes and what it means: the epidemiological evidence. *Annu Rev Public Health* 2021;42:59–77.
- Federation ID. *IDF diabetes atlas 10th edition*. 2021: 141.
- Knowler WC, Barrett-Connor E, Fowler SE, *et al*. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393–403.
- Lindström J, Louheranta A, Manninen M, *et al*. The Finnish diabetes prevention study (DPS): lifestyle intervention and 3-year results on diet and physical activity. *Diabetes Care* 2003;26:3230–6.
- Tuso P. Prediabetes and lifestyle modification: time to prevent a preventable disease. *Perm J* 2014;18:88–93.
- Li G, Zhang P, Wang J, *et al*. The long-term effect of lifestyle interventions to prevent diabetes in the China DA Qing diabetes prevention study: a 20-year follow-up study. *Lancet* 2008;371:1783–9.
- Hsu C-C, Tu S-T, Sheu W-H. Diabetes Atlas: achievements and challenges in diabetes care in Taiwan. *J Formos Med Assoc* 2019;118 Suppl 2:S130–4.
- Lee C-C, Wang C-H, Tsan K-W. Undiagnosed diabetes among hospitalized patients - analysis of 1096 hospitalized patients in a single center, MMH. *J Intern Med Taiwan* 2003;14.
- Pan W-H. *Nutrition and health survey in Taiwan (NAHSIT) report 2013-2016*. Taiwan, 2016.
- Glümer C, Carstensen B, Sandbæk A, *et al*. A Danish diabetes risk score for targeted screening: the Inter99 study. *Diabetes Care* 2004;27:727–33.
- Gray LJ, Taub NA, Khunti K, *et al*. The Leicester risk assessment score for detecting Undiagnosed type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet Med* 2010;27:887–95.
- Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* 2003;26:725–31.
- Mohan V, Deepa R, Deepa M, *et al*. A simplified Indian diabetes risk score for screening for Undiagnosed diabetic subjects. *J Assoc Physicians India* 2005;53:759–63.
- Venkatrao M, Nagarathna R, Patil SS, *et al*. A composite of BMI and waist circumference may be a better obesity metric in Indians with high risk for type 2 diabetes: an analysis of NMB-2017, a nationwide cross-sectional study. *Diabetes Res Clin Pract* 2020;161:108037.
- Li HY, Chang YC, Wei JN, *et al*. Validation of diabetes risk scores for predicting diabetes diagnosed by oral glucose tolerance test. *Diabetes Care* 2010;33:e26.
- Lin J-W, Chang Y-C, Li H-Y, *et al*. Cross-sectional validation of diabetes risk scores for predicting diabetes, metabolic syndrome, and chronic kidney disease in Taiwanese. *Diabetes Care* 2009;32:2294–6.
- Menke A, Casagrande S, Geiss L, *et al*. Prevalence of and trends in diabetes among adults in the United States, 1988-2012. *JAMA* 2015;314:1021–9.
- Fan CT, Lin JC, Lee CH. Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical Island. *Pharmacogenomics* 2008;9:235–46.
- Hsieh C-Y, Su C-C, Shao S-C, *et al*. Taiwan's national health insurance research database: past and future. *Clin Epidemiol* 2019;11:349–58.

- 25 Selvin E, Wang D, Matsushita K, *et al.* Prognostic implications of single-sample Confirmatory testing for Undiagnosed diabetes: a prospective cohort study. *Ann Intern Med* 2018;169:156–64.
- 26 Fang M, Wang D, Coresh J, *et al.* Undiagnosed diabetes in U.S. adults: prevalence and trends. *Diabetes Care* 2022;45:1994–2002.
- 27 World Health Organization, IDF. *Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. Report of a WHO/IDF consultation.* Geneva, 2006: 50.
- 28 van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by Chained equations in R. *J Stat Softw* 2011;45:1–67.
- 29 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
- 30 Yee TW. Quantile regression via vector generalized additive models. *Stat Med* 2004;23:2295–315.
- 31 Washirasaksiri C, Srivanichakorn W, Borrisut N, *et al.* Fasting plasma glucose and Hba1C levels predict the risk of type 2 diabetes and diabetic retinopathy in a Thai high-risk population with Prediabetes. *Front Pharmacol* 2022;13:950225.
- 32 Chuang SY, Bai CH, Chen WH, *et al.* Fibrinogen independently predicts the development of ischemic stroke in a Taiwanese population: CVDFACTS study. *Stroke* 2009;40:1578–84.
- 33 Heianza Y, Arase Y, Saito K, *et al.* Development of a screening score for Undiagnosed diabetes and its application in estimating absolute risk of future type 2 diabetes in Japan: Toranomon hospital health management center study 10 (TOPICS 10). *J Clin Endocrinol Metab* 2013;98:1051–60.
- 34 Bando Y, Kanehara H, Aoki K, *et al.* Characteristics of undiagnosed diabetes mellitus in a population undergoing health screening in Japan: target populations for efficient screening. *Diabetes Res Clin Pract* 2009;83:341–6.
- 35 Washirasaksiri C, Srivanichakorn W, Godsland IF, *et al.* Increasing Glycaemia is associated with a significant decline in HDL cholesterol in women with prediabetes in two national populations. *Sci Rep* 2021;11:12194.
- 36 Collins GS, Mallett S, Omar O, *et al.* Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103.
- 37 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.
- 38 Chen CC, Chen CL, Ko Y. The misconceptions and determinants of diabetes knowledge in patients with diabetes in Taiwan. *J Diabetes Res* 2020;2020:2953521.
- 39 Hill-Briggs F, Adler NE, Berkowitz SA, *et al.* Social determinants of health and diabetes: a scientific review. *Diabetes Care* 2020;44:258–79.
- 40 Bird Y, Lemstra M, Rogers M, *et al.* The relationship between socioeconomic status/income and prevalence of diabetes and associated conditions: a cross-sectional population-based study in Saskatchewan, Canada. *Int J Equity Health* 2015;14:93.
- 41 Hsu C-C, Lee C-H, Wahlqvist ML, *et al.* Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. *Diabetes Care* 2012;35:2286–92.
- 42 Lee C-H, Ko A-S, Warnakulasuriya S, *et al.* Intercountry prevalences and practices of Betel-Quid use in South, Southeast and Eastern Asia regions and associated oral preneoplastic disorders: an international collaborative study by Asian Betel-Quid consortium of South and East Asia. *Int J Cancer* 2011;129:1741–51.
- 43 Yang YH, Warnakulasuriya S, Yang HF, *et al.* Public health measures to reduce Areca nut and Betel Quid use for control of oral cancer in Taiwan. *Oral Oncol* 2020;108:104915.
- 44 Tseng CH. Betel nut chewing and incidence of newly diagnosed type 2 diabetes mellitus in Taiwan. *BMC Res Notes* 2010;3:228.
- 45 Tung T-H, Chiu Y-H, Chen L-S, *et al.* A population-based study of the association between Areca nut chewing and type 2 diabetes mellitus in men (Keelung community-based integrated screening programme No.2). *Diabetologia* 2004;47:1776–81.
- 46 Ryu KS, Lee SW, Batbaatar E, *et al.* A deep learning model for estimation of patients with Undiagnosed diabetes. *Applied Sciences* 2020;10:421.
- 47 Cho E, Min D, Lee HS. Development and validation of an Undiagnosed diabetes screening tool: based on the Korean national health and nutrition examination survey. *Healthcare* 2010;9:1138.
- 48 Franciosi M, De Berardis G, Rossi MCE, *et al.* Use of the diabetes risk score for opportunistic screening of Undiagnosed diabetes and impaired glucose tolerance: the IGL00 (impaired glucose tolerance and long-term outcomes observational) study. *Diabetes Care* 2005;28:1187–94.
- 49 Barengo NC, Tamayo DC, Tono T, *et al.* A Colombian diabetes risk score for detecting Undiagnosed diabetes and impaired glucose regulation. *Prim Care Diabetes* 2017;11:86–93.
- 50 Mao T, Chen J, Guo H, *et al.* The efficacy of new Chinese diabetes risk score in screening Undiagnosed type 2 diabetes and Prediabetes: a community-based cross-sectional study in Eastern China. *J Diabetes Res* 2020;2020:7463082.

Supplemental Methods

GWAS QC procedures

The TWB chip was used to genotype samples from the Taiwan Biobank. This chip is a customized Affymetrix Axiom Genome-Wide Array, encompassing approximately 653,000 single nucleotide polymorphisms (SNPs) [1]. A rigorous quality control (QC) process was conducted at both the SNP and individual levels. SNPs exhibiting call rates less than 95% and possessing Hardy-Weinberg Equilibrium p-values $< 10^{-4}$ were eliminated from the analysis. Similarly, samples demonstrating call rates below 95%, those that were duplicates (PLINK [2] pi_hat statistics exceeding 0.9), or those that failed to pass the PLINK sex check (specifically, inconsistencies between self-reported and actual biological sex) were also excluded. Additional criteria for exclusion included potential sample contamination, which was identified when the median of the PLINK pi_hat statistics of a sample with all other samples > 0.05 . Finally, PRIMUS software [3] was employed to identify and retain the largest possible set of unrelated individuals by utilizing a PLINK pi_hat threshold of 0.1, thereby eliminating first-degree relatives.

Supplemental Figures

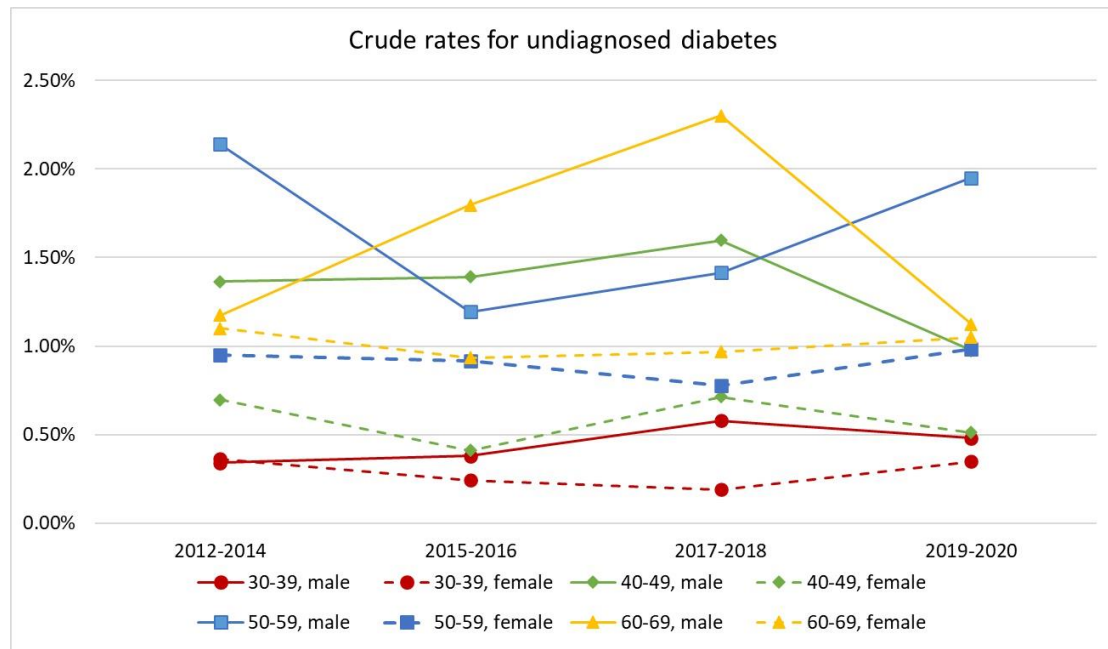


Figure S1. The age- and sex-specific crude rates for undiagnosed diabetes over years

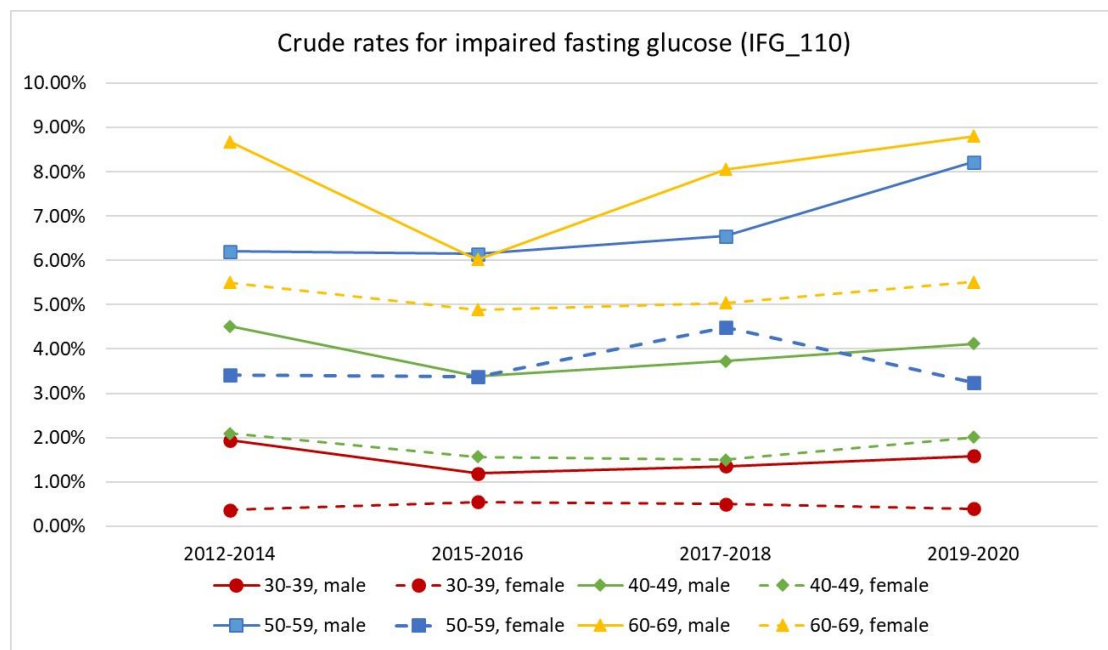


Figure S2. The age- and sex-specific crude rates for IFG_110 (fasting glucose between 110 and 125 mg/dl) over years

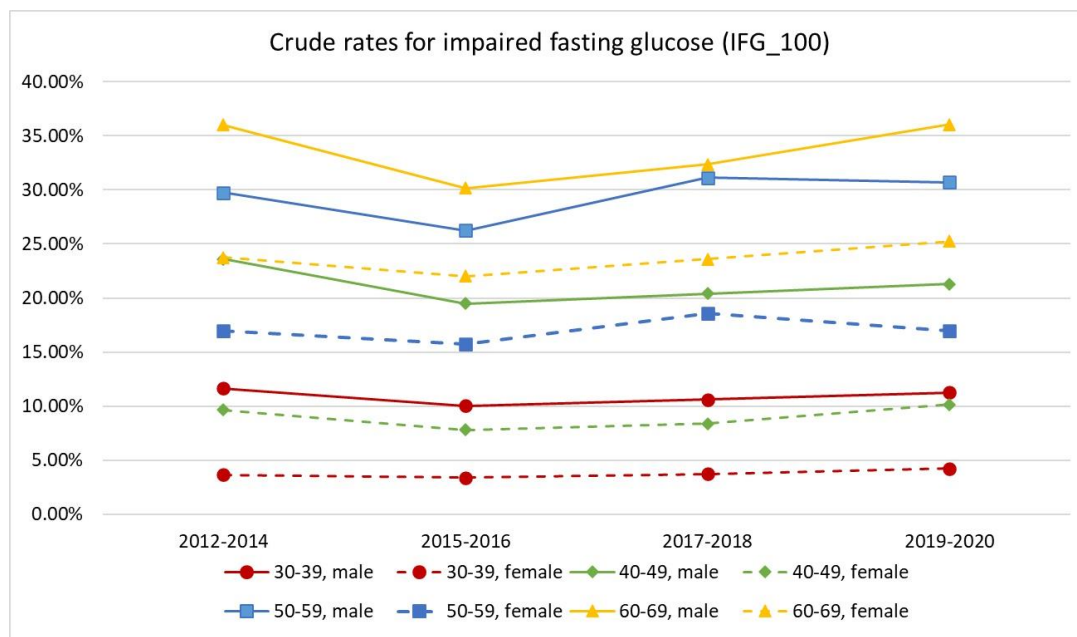


Figure S3. The age- and sex-specific crude rate for IFG_100 (fasting glucose between 100 and 125 mg/dl) over years

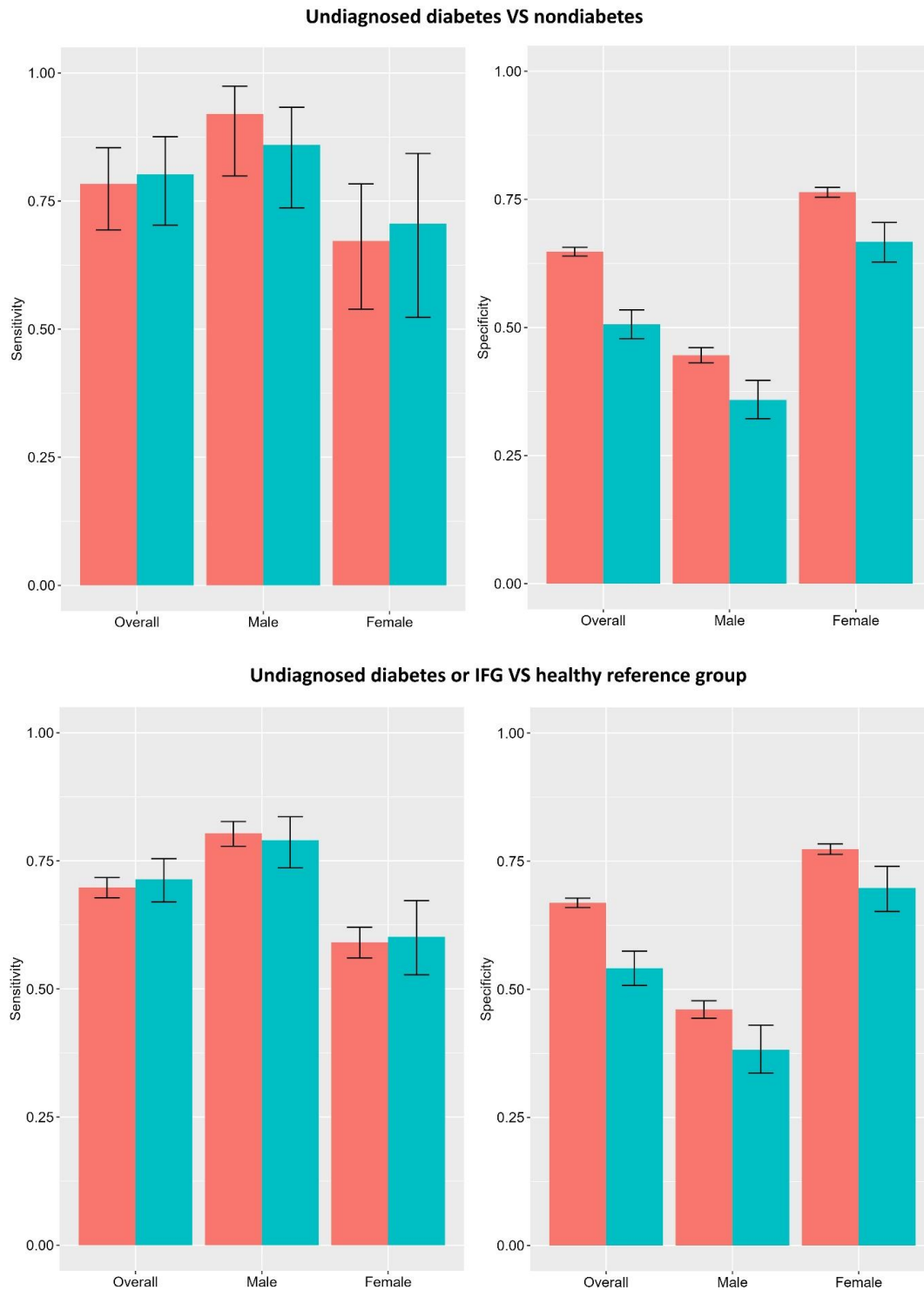


Figure S4. Sensitivities and specificities for Model 2. The upper section of the figure illustrates the sensitivities (left) and specificities (right) of the model in predicting undiagnosed diabetes versus nondiabetes. The lower section of the figure shows the

sensitivities (left) and specificities (right) for predicting undiagnosed diabetes or IFG_100 (fasting glucose between 100 and 125 mg/dl) versus healthy reference group. The sensitivities and specificities were calculated in the overall, male, and female samples in the Taiwan Biobank testing dataset (TWB) and the external CVDFACTS validation dataset. The 95% confidence intervals are depicted as error bars in the figure. The results for TWB and CVDFACTS are represented in orange and green bars, respectively.

References

- [1] Chen CH, Yang JH, Chiang CWK, et al. (2016) Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Human molecular genetics* 25(24): 5321-5331. [10.1093/hmg/ddw346](https://doi.org/10.1093/hmg/ddw346)
- [2] Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3): 559-575. [10.1086/519795](https://doi.org/10.1086/519795)
- [3] Staples J, Nickerson DA, Below JE (2013) Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genetic epidemiology* 37(2): 136-141. [10.1002/gepi.21684](https://doi.org/10.1002/gepi.21684)