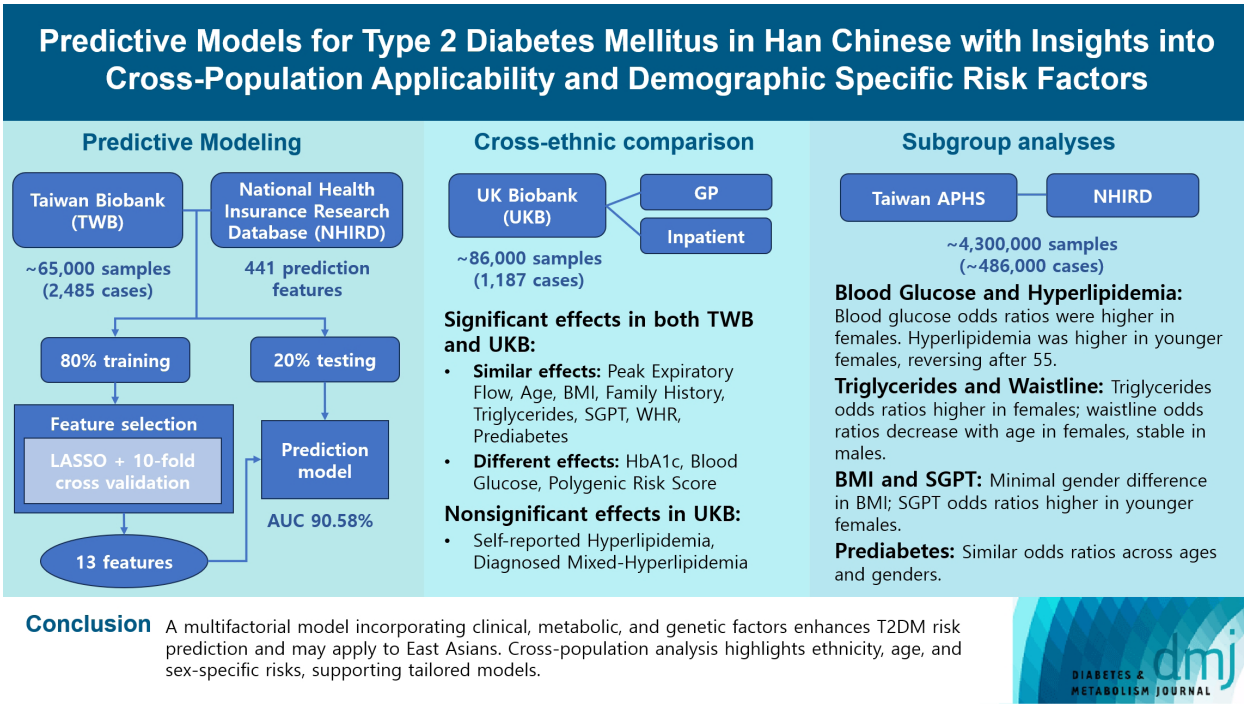
dmj
DIABETES & METABOLISM JOURNAL

Predictive Models for Type 2 Diabetes Mellitus in Han Chinese with Insights into Cross-Population Applicability and Demographic Specific Risk Factors

Ying-Erh Chen, Djeane Debora Onthoni, Shao-Yuan Chuang, Guo-Hung Li, Yong-Sheng Zhuang, Hung-Yi Chiou, Wayne Huey-Herng Sheu, Ren-Hua Chung

Diabetes Metab J 2025;49:1272-1286 | *https://doi.org/10.4093/dmj.2024.0319*



Highlights

- **A diabetes prediction model in Han Chinese achieved an AUC of 90.58%.**
- **Peak expiratory flow was identified as a novel predictor of T2DM risk.**
- **HbA1c and polygenic risk scores had distinct impacts in Han Chinese and Europeans.**
- **Sex- and age-specific risk factors varied significantly in Han Chinese.**
- **Findings emphasize personalized risk assessment and cross-population validation.**

How to cite this article:
Chen YE, Onthoni DD, Chuang SY, Li GH, Zhuang YS, Chiou HY, et al. Predictive Models for Type 2 Diabetes Mellitus in Han Chinese with Insights into Cross-Population Applicability and Demographic Specific Risk Factors. Diabetes Metab J 2025;49:1272-1286. https://doi.org/10.4093/dmj.2024.0319

Original Article
Metabolic Risk/Epidemiology

Diabetes Metab J 2025;49:1272-1286
https://doi.org/10.4093/dmj.2024.0319
pISSN 2233-6079 · eISSN 2233-6087

dmj
DIABETES & METABOLISM JOURNAL

Check for updates

Predictive Models for Type 2 Diabetes Mellitus in Han Chinese with Insights into Cross-Population Applicability and Demographic Specific Risk Factors

Ying-Erh Chen[1], Djeane Debora Onthoni[2], Shao-Yuan Chuang[2], Guo-Hung Li[2], Yong-Sheng Zhuang[2], Hung-Yi Chiou[2,3], Wayne Huey-Herng Sheu[4], Ren-Hua Chung[2]

[1]Department of Risk Management and Insurance, Tamkang University, New Taipei City, Taiwan
[2]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan
[3]School of Public Health, College of Public Health, Taipei Medical University, Taipei, Taiwan
[4]Institute of Molecular and Genomic Medicine, National Health Research Institutes, Zhunan, Taiwan

Background: The rising global incidence of type 2 diabetes mellitus (T2DM) underscores the need for predictive models that enhance early detection and prevention across diverse populations. This study aimed to identify predictors of incident T2DM within a Han Chinese population, assess their impact across various age and sex demographics, and explore their applicability to European populations.
Methods: Using data from about 65,000 participants in the Taiwan Biobank (TWB), we developed a predictive model, achieving an area under the receiver operating characteristic curve of 90.58%. Key predictors were identified through LASSO regression within the TWB cohort and validated using over 4 million records from Taiwan's Adult Preventive Healthcare Services (APHS) program and the UK Biobank (UKB).
Results: Our analysis highlighted 13 significant predictors, including established factors like glycosylated hemoglobin (HbA1c) and blood glucose levels, and less conventionally considered variables such as peak expiratory flow. Notable differences in the effects of HbA1c levels and polygenic risk scores between the TWB and UKB cohorts were observed. Additionally, age and sex-specific impacts of these predictors, detailed through APHS data, revealed significant variances; for instance, waist circumference and diagnosed mixed hyperlipidemia showed greater impacts in younger females than in males, while effects remained uniform across male age groups.
Conclusion: Our findings offer novel insights into the diagnosis and management of diabetes for the Han Chinese and potentially for broader East Asian populations, highlighting the importance of ethnic and demographic diversity in developing predictive models for early detection and personalized intervention strategies.

Keywords: Diabetes mellitus, type 2; East Asian peoples; Genetic risk score; Prediction algorithms; Risk factors

INTRODUCTION

The escalating global prevalence of type 2 diabetes mellitus (T2DM) has intensified the search for effective early detection and prevention strategies [1]. As a complex metabolic disorder influenced by a combination of genetic, environmental, and lifestyle factors, T2DM presents significant challenges in risk prediction and management [2]. Research has demonstrated the efficacy of lifestyle and pharmaceutical interventions in mitigating or postponing the onset of T2DM among high-risk groups, as demonstrated through randomized trials [3,4]. Nonetheless, while these interventions are crucial at a popula-

Corresponding author: Ren-Hua Chung https://orcid.org/0000-0002-9835-6333
Institute of Population Health Sciences, National Health Research Institutes, No. 35, Keyan Road, Zhunan, Miaoli 350, Taiwan
E-mail: rchung@nhri.edu.tw

Received: Jun. 20, 2024; Accept: Jan. 6, 2025

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2025 Korean Diabetes Association https://e-dmj.org

tion level, there remains a gap in reaching high-risk individuals who may need more intensive, resource-heavy intervention approaches. The success in pinpointing to these individuals depends on the precision of risk models to accurately classify them as high risk [5].

Recent advancements in risk prediction models have significantly improved our ability to identify both prevalent and incident T2DM. These models range from traditional regression-based approaches to more sophisticated, computationally intensive artificial intelligence techniques [6,7]. These models vary in their strategies of selecting risk factors: some rely solely on non-invasive predictors like family history, body mass index (BMI), age, and sex for a broader, more accessible screening strategy [8-10]. In contrast, others incorporate biomarkers such as fasting glucose, cholesterol levels, and polygenic risk scores (PRS) to refine and increase their predictive accuracy, aiming for a more precise identification of T2DM risk [11-13].

Despite the growth of T2DM risk prediction models, challenges remain in their generalizability and predictive performance across varied ethnic backgrounds. Research indicates that various ethnic groups have unique biological, environmental, and lifestyle risk profiles for T2DM development [14-16]. Moreover, within the same population, the impact of risk factors can vary across different age and sex groups. For instance, while the influence of metabolic risk factors such as obesity, prediabetes, and hyperlipidemia tends to decrease with age, lifestyle elements like sleep duration can exert more pronounced effects in older individuals [17]. Acknowledging these ethnic, age, and sex differences is crucial for crafting personalized approaches in the primary prevention and management of diabetes, underscoring the need for models that reflect the diversity of patient populations.

This study leverages data from the Taiwan Biobank (TWB), enriched with a broad range of predictive variables, to identify and incorporate risk factors into a comprehensive model for predicting the onset of T2DM. By comparing the effectiveness of this model with data from the UK Biobank (UKB), our research seeks to evaluate the consistency of predictive factors across different ethnic groups. Significantly, the influence of these selected risk factors was also examined in a separate dataset from the Adult Preventive Healthcare Services (APHS) program in Taiwan, which includes over 4 million individuals. These individuals were categorized into 10 age groups, ranging from 40 to over 85 years, with a 5-year increment, and analyzed separately for males and females. This extensive analysis of risk factor effects across different demographics offers novel insights into diabetes management for the Han Chinese and, by extension, potentially for East Asian populations.

METHODS

Study populations

Our study populations were derived from three significant sources: the TWB, the UKB, and the Taiwan Health Promotion Administration's APHS program. Initiated in 2008, TWB has enrolled over 150,000 individuals aged more than 20 years from Taiwan [18]. These participants provided comprehensive information through questionnaires detailing their health status and family disease history. They also underwent extensive health checks, including physical examinations and blood and urine analyses at the time of recruitment. Of these participants, approximately 115,000 underwent genotyping using genome-wide association study single nucleotide polymorphism arrays. Subsequently, their data were linked with their medical records from the National Health Insurance Research Database (NHIRD), spanning from 2009 to 2020. The NHIRD is a rich resource containing comprehensive medical records and expenditure claims from various healthcare services, including outpatient, inpatient, and ambulatory care, as well as pharmacy data for reimbursement purposes. We extracted data on diagnosis (using International Classification of Diseases, 9th Revision [ICD-9] and ICD-10 codes) related to outpatient, ambulatory, and inpatient services both 3 years prior to and following their recruitment into the study.

The UKB is a large-scale biomedical database and research resource, representing one of the most comprehensive health initiatives undertaken in the UK. Established in 2006, it encompasses detailed genetic and health information from over 500,000 volunteer participants aged between 40 and 69 years at the time of recruitment. Its goal is to facilitate research into the roles of genetic, environmental, and lifestyle factors in disease development. Participants have provided a wealth of data, including blood samples, genetic profiles, detailed medical histories, and regular updates on their health status [19].

Lastly, the Taiwan Health Promotion Administration offers an APHS program, targeting individuals aged 40 to 65 for a health checkup every 3 years, and annually for those over 65. The services include questionnaires, physical examinations, and laboratory tests. From 2013 to 2018, we obtained APHS data on over 4 million individuals, ensuring there was no over-

lap for the same individual across the studied years. This dataset was subsequently linked to the participants' medical records in the NHIRD, from which diagnosis information was extracted for 3 years before and another 3 years after their enrollment into the APHS program, mirroring the approach taken with the TWB cohort. More details about the quality control procedures in the three study cohorts are provided in the Supplementary Methods.

Outcome definitions

To ensure our study accurately identified new cases of T2DM, we excluded individuals who already had diabetes at the time of recruitment from all study cohorts. This exclusion applied to those with self-reported diabetes, a medical history indicating diabetes, or elevated glucose levels from initial blood tests, effectively removing cases of prevalent T2DM and allowing for the focus on incident T2DM. Detailed specifications regarding these exclusion parameters are elaborated in the Supplementary Methods.

For the TWB cohort, we defined the incident T2DM within a 3-year post-recruitment window, requiring a minimum of three outpatient diabetes diagnoses or at least one diabetes-related hospitalization within a 365-day period, according to their ICD records [20]. This method ensured accurate identification of incident T2DM cases, a protocol also applied to the APHS dataset to maintain outcome assessment consistency.

Regarding the UKB cohort, given the variability in the duration of ICD records accessible from general practitioner (GP) and inpatient data sources, we similarly defined the incidence of T2DM based on ICD codes within 3 years following participant recruitment. Additionally, data from three follow-up questionnaires were utilized to ascertain incident diabetes cases, thereby offering a comprehensive approach to identifying new onset T2DM within the study timeframe. More details about the definitions of incident T2DM in the TWB and UKB are also provided in the Supplementary Methods.

Prediction model construction in the Taiwan Biobank

The TWB, with its comprehensive collection of clinical features including data from questionnaires, physical, and laboratory measurements, provided a more detailed dataset than the APHS dataset. This extensive dataset enabled the identification of significant predictors for the incidence of T2DM in the Han Chinese population. In addition, we incorporated the PRS for T2DM, calculated using the PRS-CSx algorithm [21]. To tackle the issue of missing data among our variables, we employed the mice package in R version 4.3.2 (R Foundation for Statistical Computing, Vienna, Austria) [22], which executes multivariate imputation by chained equations, ensuring a thorough approach to imputing missing data. More detailed explanations for compiling the prediction features from the TWB are provided in the Supplementary Methods.

In constructing our prediction model, we allocated 80% of the samples to the training dataset and the remaining 20% to the testing dataset. This split resulted in 52,089 samples for training and 13,023 samples for testing. To pinpoint the most relevant features, we applied LASSO regression, selecting the optimal hyperparameter lambda through 10-fold cross-validation using the glmnet package in R. We adopted the lambda.1se criterion, which selects the most regularized model with the cross-validated error within one standard error of the minimum, for feature selection. The features identified by LASSO regression were further evaluated using logistic regression within the training dataset. Only those features that surpassed the rigorous Bonferroni correction threshold were included in the final prediction model. We assessed the performance of our prediction model using the area under the receiver operating characteristic curve (AUC), which allowed us to measure the model's accuracy in differentiating between individuals who would develop T2DM and those who would not.

Effects of the important features across different ethnicities and age and sex groups

We conducted a comparative analysis of the effects for key features identified in the TWB against those derived from the UKB. This comparison aimed to explore the variations in the impact of these features on the incidence of T2DM across different ethnic groups. The effect sizes were calculated using logistic regression, incorporating the final set of features included in our prediction model. Furthermore, the large sample size of the APHS dataset facilitated a detailed analysis of the effects for our predictive features across diverse age and sex demographics. The APHS cohort was divided by sex into male and female groups. Within each of these groups, we further categorized samples based on age, with 5-year increments starting at age 40. Logistic regression analyses were then conducted within each subgroup, employing the same set of features used in our primary prediction model, allowing for a nuanced understanding of how these factors influence T2DM risk across diverse demographic groups.

Prediction model construction in the UK Biobank

We also used data from the UKB data to select significant features and construct a prediction model, allowing us to compare the selected features and model performance with those derived from the TWB. Detailed explanations for compiling the prediction features from the UKB are provided in the Supplementary Methods. Similar to the process used for the TWB model, the UKB data were split into 80% of training and 20% of testing, followed by feature selection using LASSO and model construction via logistic regression.

Ethics approval

Written informed consent for participants in the Taiwan Biobank and UK Biobank studies was obtained from all participants. Research in this study was approved by the Institutional Review Board of the National Health Research Institutes in Taiwan (reference number: EC1091202-E).

RESULTS

Table 1 presents the demographic characteristics and health metrics of the samples from both the TWB and UKB, including those stratified by age. The mean age at recruitment in the TWB (48.93 years) was notably younger than in the UKB (57.86 years), attributed to the TWB's broader recruitment age range of 20 to 70 years, compared to the UKB's focus on individuals aged 40 to 70 years. The incidence rate of T2DM within 3 years post-recruitment was significantly higher in the TWB than in the UKB across the overall sample (3.81% vs. 1.37%), a pattern consistent across different age categories.

Moreover, while the TWB recorded fasting glucose levels, it showed higher average blood glucose and glycosylated hemoglobin (HbA1c) levels compared to the UKB, which measured random glucose levels. This difference may contribute to the observed higher incidence of diabetes in the TWB. In contrast,

Table 1. Characteristics of participant groups across all ages and by age stratification in the Taiwan and UK Biobanks

Characteristic	All		<40 years	40–49 years		50–59 years		60–70 years	
	TWB	UKB	TWB	TWB	UKB	TWB	UKB	TWB	UKB
Sample size[a]	65,112 (35.8)	86,484 (43.7)	19,409 (31.2)	15,225 (34.8)	15,534 (38.9)	17,391 (31.45)	27,508 (41.6)	13,087 (39.9)	43,442 (46.8)
Incident T2DM cases	2,485 (3.81)	1,187 (1.37)	301 (1.55)	402 (2.64)	121 (0.77)	885 (5.08)	325 (1.18)	897 (6.85)	741 (1.70)
Age, yr	48.93±10.77	57.86±7.73	34.77±2.87	44.93±2.60	45.20±2.70	55.01±2.55	54.97±2.88	63.80±2.80	64.21±2.84
BMI, kg/m^2	24.02±3.68	27.55±4.67	24.01±4.26	24.12±3.72	27.32±5.12	23.85±3.35	27.69±4.88	24.02±3.21	27.54±4.36
WHR	0.86±0.06	0.87±0.09	0.83±0.06	0.85±0.06	0.85±0.09	0.86±0.06	0.87±0.09	0.88±0.06	0.88±0.09
Hypertension (self-reported)	6,068 (9.31)	26,761 (30.94)	434 (2.22)	813 (5.33)	2,402 (15.46)	2,091 (12.02)	7,676 (27.90)	2,730 (20.86)	16,683 (38.40)
Hyperlipidemia/High cholesterol (self-reported)[b]	3,596 (5.52)	12,354 (14.28)	313 (1.61)	486 (3.19)	609 (3.92)	1,295 (7.44)	2,864 (10.41)	1,502 (11.47)	8,881 (20.44)
Essential hypertension (ICD)[c]	3,141 (4.82)	6,028 (6.97)	222 (1.14)	363 (2.38)	444 (2.85)	1,112 (6.39)	1,559 (5.66)	1,444 (11.03)	4,025 (9.26)
Mixed hyperlipidemia (ICD)[c]	972 (1.49)	448 (0.51)	80 (0.41)	135 (0.88)	34 (0.21)	352 (2.02)	123 (0.44)	405 (3.09)	291 (0.66)
Fasting/random blood glucose, mg/dL[d]	93.04±12.09	90.03±13.24	89.50±9.79	92.08±11.49	87.26±12.35	94.95±13.99	89.37±12.62	96.10±11.09	91.44±13.72
HbA1c, %	5.62±0.48	5.40±2.60	5.46±0.41	5.56±0.45	5.20±2.50	5.71±0.52	5.30±2.50	5.75±0.45	5.40±2.60
SBP, mm Hg	118.51±17.74	139.15±19.27	110.82±14.02	115.17±15.98	129.15±16.43	120.99±17.25	136.83±18.17	129.16±18.77	144.09±19.22
DBP, mm Hg	73.16±11.06	82.65±10.45	70.08±10.37	72.91±11.39	81.35±10.76	74.61±10.99	83.42±10.47	75.26±10.58	82.63±10.28
Family history of diabetes	21,640 (33.23)	16,727 (19.34)	5,401 (27.8)	5,509 (36.18)	3,380 (21.75)	6,579 (37.82)	5,870 (21.33)	4,151 (31.71)	7,477 (17.21)
Total cholesterol, mg/dL	196.52±35.44	221.80±43.90	185.17±33.25	193.28±33.80	214.50±39.50	205.24±35.66	226.30±42.50	203.24±35.38	221.50±45.90
Triglycerides, mg/dL	111.96±87.33	155.10±87.90	100.60±91.84	112.57±86.45	145.70±95.90	119.17±92.27	156.50±91.00	115.03±71.05	157.60±82.50
LDL-C, mg/dL	122.08±31.60	138.60±33.50	114.67±30.63	120.10±31.04	134.20±31.00	127.71±31.69	142.00±32.60	126.35±31.22	138.10±34.60
HDL-C, mg/dL	54.93±13.38	56.30±14.60	54.27±13.01	54.45±13.20	54.60±13.80	55.90±13.74	56.80±14.80	55.15±13.50	56.70±14.70
PEF, L/sec	4.25±2.29	2.73±0.77	4.63±2.32	4.52±2.31	3.12±0.79	3.99±2.22	2.82±0.75	3.73±2.16	2.53±0.72

Values are presented as number (%) or mean±standard error.
TWB, Taiwan Biobank; UKB, UK Biobank; T2DM, type 2 diabetes mellitus; BMI, body mass index; WHR, waist-to-hip ratio; ICD, International Classification of Diseases; HbA1c, glycosylated hemoglobin; SBP, systolic blood pressure; DBP, diastolic blood pressure; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; PEF, peak expiratory flow.
[a]Total number and percentage of male participants, [b]Hyperlipidemia for the TWB and high cholesterol for the UKB, [c]Diseases identified based on ICD codes in medical records, [d]Fasting blood glucose for the TWB and random blood glucose for the UKB.

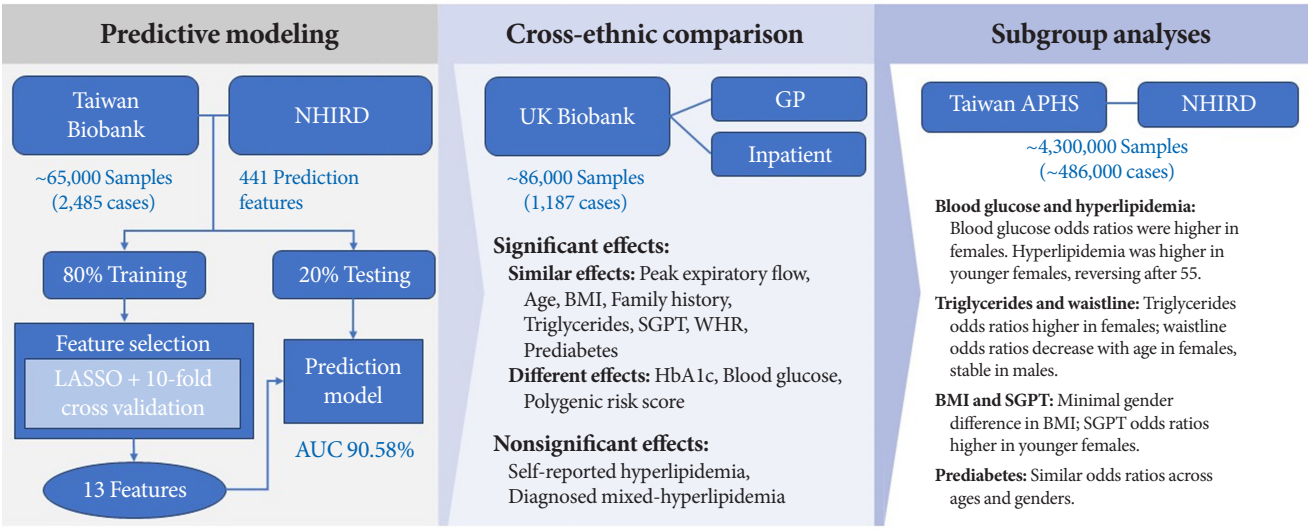


Fig. 1. Analysis flowchart and summary of findings. Predictive modeling linked the Taiwan Biobank to the National Health Insurance Research Database (NHIRD), analyzing approximately 65,000 samples across 441 features. Data was divided into training (80%) and testing (20%) datasets. LASSO regression with 10-fold cross-validation identified 13 significant features from the training dataset, achieving an area under the receiver operating characteristic curve (AUC) of 90.58%. Cross-ethnic comparisons were conducted using data from the UK Biobank, linked with general practitioner (GP) and hospital inpatient records, covering about 86,000 samples. Of the 13 features, 11 were significant in the UK Biobank, with eight showing similar effects and three showing different effects; two features were nonsignificant. Subgroup analysis utilized data from Taiwan's Adult Preventive Healthcare Services (APHS) program, involving over 4.3 million samples, to assess feature effects across various age and sex groups. BMI, body mass index; SGPT, serum glutamate pyruvate transaminase; WHR, waist-to-hip ratio; HbA1c, glycosylated hemoglobin.

mean BMI readings in the UKB exceeded those in the TWB across both the total and age-stratified samples, despite similar mean waist-to-hip ratio (WHR) values between the two cohorts. Furthermore, the UKB reported higher percentages of self-reported hypertension and clinically diagnosed hypertension than the TWB. A greater proportion of UKB participants also reported high cholesterol levels compared to those in the TWB reporting hyperlipidemia, though the prevalence of clinically diagnosed hyperlipidemia was greater in the TWB. Notably, a higher percentage of TWB participants had a family history of diabetes than those in the UKB. However, mean blood pressure and lipid levels were generally higher in the UKB than in the TWB, indicating variances in cardiovascular risk factors between the two populations.

Fig. 1 presents our analysis flowchart and summarizes the findings. For predictive modeling of T2DM, the TWB samples were used for feature selection and model training. The LASSO regression analysis conducted on the training data from the TWB successfully identified 16 features selected from 441 features as potential predictors of incident T2DM over a 3-year period. Upon further analysis using logistic regression to jointly evaluate these 16 features, 13 were found to have *P* values less than the adjusted threshold of 3.125E-03 (accounting for the multiple testing correction for 16 features). The beta estimates along with their 95% confidence intervals (CI) for these 13 statistically significant features are depicted in Fig. 2.

Fig. 2 further explores the effects of the 13 features identified, comparing these with estimates from the UKB as outlined in the cross-ethnic comparison section of Fig. 1. Among these features, 11 consistently emerged as significant predictors in the UKB as well, with their effect sizes distinctly differing from zero. This uniform significance underscores their potential as reliable indicators of T2DM risk across diverse populations. However, exceptions were noted for self-reported hyperlipidemia and clinically diagnosed mixed hyperlipidemia, which were not significant in the UKB. Remarkably, the effect size attributed to HbA1c levels presented a notable discrepancy between the cohorts, being significantly higher in the UKB (3.14) as compared to the TWB (2.55). Conversely, the effect size for blood glucose was higher in the TWB (0.61) than in the UKB (0.18). This difference likely reflects the measurement methods used: fasting glucose in the TWB, a more direct indicator of

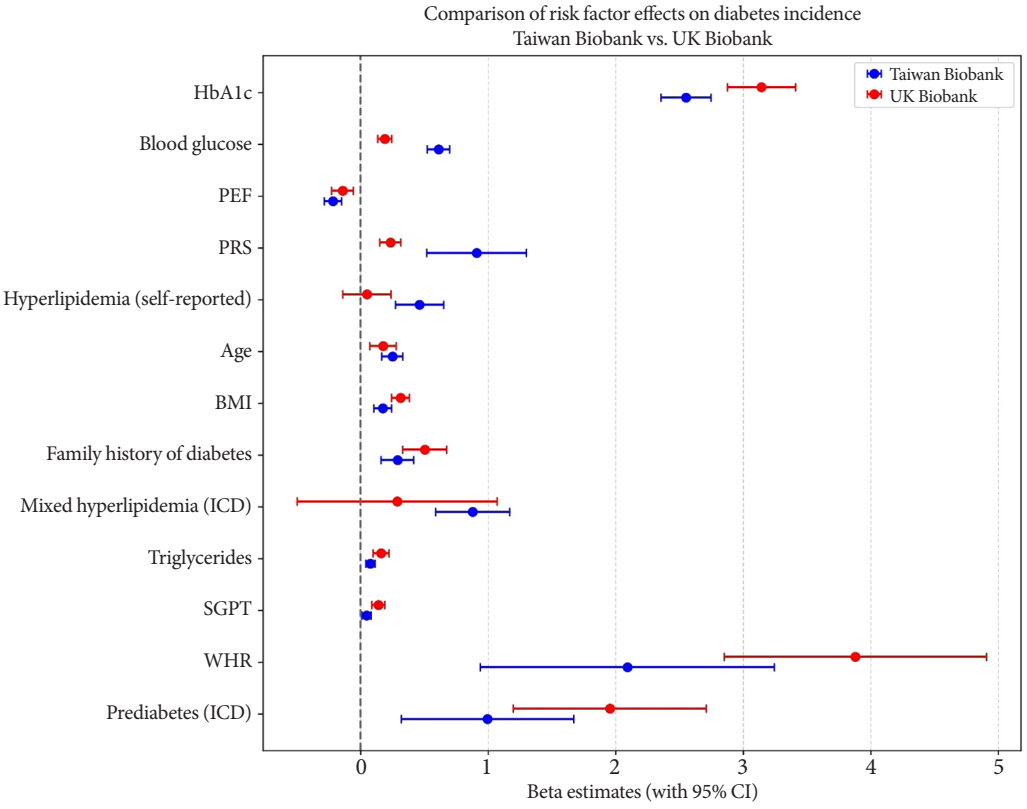


Fig. 2. Comparison of the effect sizes of risk factors for type 2 diabetes mellitus incidence between the Taiwan and UK Biobanks. Note that variables such as blood glucose, peak expiratory flow (PEF), polygenic risk score (PRS), age, body mass index (BMI), triglycerides, and serum glutamate pyruvate transaminase (SGPT) have been standardized. HbA1c, glycosylated hemoglobin; ICD, International Classification of Diseases; WHR, waist-to-hip ratio; CI, confidence interval.

T2DM, versus random glucose in the UKB. Moreover, the PRS for T2DM exhibited a significantly larger effect size in the TWB (0.91) than in the UKB (0.23), suggesting differential genetic influences on diabetes risk between the two populations. This nuanced analysis underscores the importance of considering both genetic and clinical factors in predicting diabetes risk, while also highlighting the potential variability in their impact across different ethnic groups.

Fig. 3 demonstrates the stepwise enhancement in the AUC values achieved by progressively integrating the 13 significant features identified from the TWB. These features were added to the predictive model in order of their significance, as determined by association *P* values from the TWB training dataset, with their prediction performance evaluated in both the TWB testing dataset and the entire UKB datasets. HbA1c levels and blood glucose concentrations emerged as the most significant predictors, standing first and second in terms of significance. Their inclusion in the predictive model yielded AUC percentages of 88.00% for the TWB and 82.61% for the UKB, indicating robust predictive capabilities. The addition of peak expiratory flow (PEF) as the third most significant predictor further enhanced model performance, elevating the AUC to 88.99% in the TWB, though it offered a negligible improvement in the UKB, slightly enhancing the AUC to 82.62%. Interestingly, the PRS was identified as the fourth most significant feature in the TWB, outranking other traditional predictors such as age, WHR, and family history of diabetes. This underscores the growing importance of genetic factors in predicting diabetes risk. However, it's noteworthy that the incremental AUC gains attributed to PRS and subsequent features were relatively modest in both the TWB and UKB datasets.

Table 2 details the AUC values for the overall sample as well as for samples stratified by age groups in the TWB. For the overall sample, the model achieved an AUC of 90.58%. Notably, the AUC values displayed a trend of decline with increasing age. This trend was highlighted by the highest AUC of

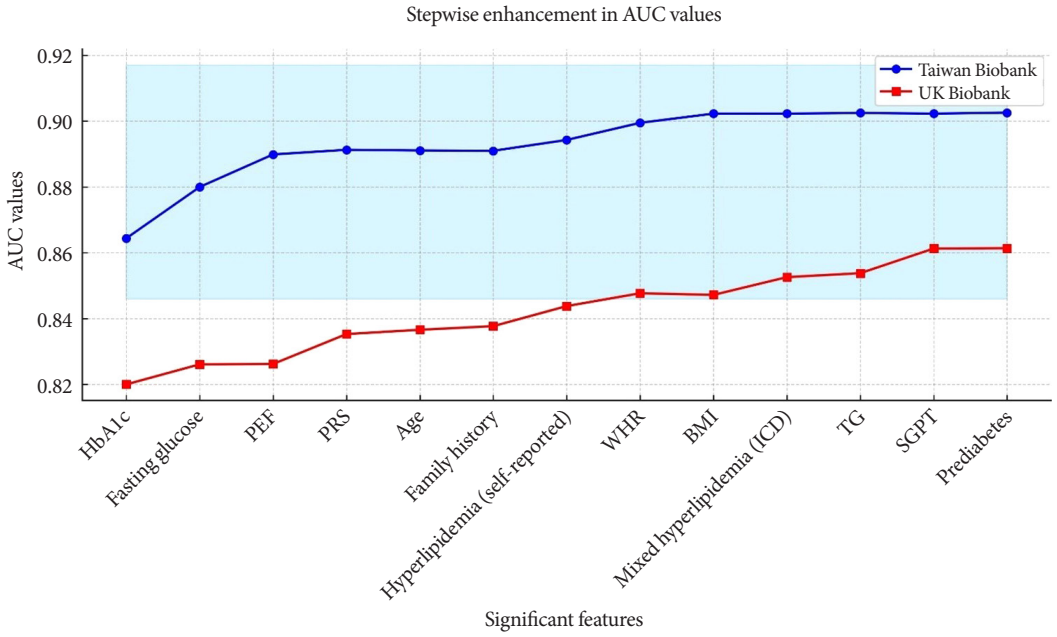


Fig. 3. Stepwise enhancement in area under the receiver operating characteristic curve (AUC) values by progressively adding the significant features to the prediction models in the Taiwan and UK Biobanks. HbA1c, glycosylated hemoglobin; PEF, peak expiratory flow; PRS, polygenic risk score; WHR, waist-to-hip ratio; BMI, body mass index; ICD, International Classification of Diseases; TG, triglyceride; SGPT, serum glutamate pyruvate transaminase.

Table 2. AUC values, sensitivity, and specificity of the prediction model in the Taiwan Biobank testing dataset

Variable	AUC	Sensitivity	Specificity
Overall	90.58 (89.15–92.02)	78.34 (71.65–86.61)	85.88 (77.08–91.20)
<40 years	92.11 (87.13–97.09)	82.60 (71.74–95.65)	90.68 (78.34–98.17)
40–49 years	91.12 (87.53–94.70)	86.45 (76.02–93.75)	84.82 (81.80–93.10)
50–59 years	89.40 (86.95–91.84)	77.61 (68.66–83.08)	84.89 (83.30–91.78)
60–70 years	83.62 (79.92–87.31)	66.66 (61.21–79.39)	88.03 (78.35–90.32)

Values are presented as estimate (95% confidence interval).
AUC, area under the receiver operating characteristic curve.

92.11% observed in individuals under 40 years of age, and the lowest AUC of 83.62% found in the age group of 60 to 70 years. This diminishing AUC with advancing age can primarily be attributed to a decrease in sensitivity, which dropped from 82.60% in the <40 age group to 66.66% in the 60 to 70 age group. In contrast, specificity remained relatively stable across the age groups, ranging from 84.82% to 90.68%. This pattern suggests that while the model's ability to correctly identify those without T2DM (specificity) remains high across ages, its ability to correctly identify those with T2DM (sensitivity) decreases as age increases.

For predictive modeling of T2DM in the UKB dataset, LASSO regression was also applied to the training data, successfully narrowing down 47 features from an initial pool of 18,300 as potential predictors for incident T2DM over a 3-year period. These selected features were further examined using logistic regression, which revealed that 12 of them had *P* values below the adjusted threshold of 1.063E-03, accounting for multiple testing corrections. The beta estimates and their 95% CI for these 11 statistically significant features are illustrated in Supplementary Fig. 1. Among the features identified, several overlapped with those found using the TWB dataset, including HbA1c, glucose, PRS, family history of diabetes, BMI, serum glutamate pyruvate transaminase (SGPT), WHR, and predia-

betes. Additionally, the UKB analysis highlighted other population-specific features, such as high-density lipoprotein cholesterol (HDL-C) and reticulocyte count. The model's performance in the testing dataset achieved an AUC of 88.69% (95% CI, 85.43% to 91.12%), which is comparable to the 90.58% AUC reported using the TWB dataset.

Supplementary Table 1 presents the demographic characteristics and health metrics of the APHS samples, stratified by age and sex. Notably, the APHS samples displayed a consistently higher T2DM incidence rate within the same age group compared to those from the TWB (e.g., 5.25% in the APHS vs. 2.64% in the TWB for the 40 to 49 age group). A contributing factor to this discrepancy could be the APHS's methodology, which did not exclude individuals with undiagnosed diabetes—defined as those lacking a medical history of diabetes but presenting elevated fasting glucose and HbA1c levels at baseline—due to the unavailability of fasting glucose and HbA1c measurements at enrollment. This approach differs from that of the TWB and likely influenced the higher T2DM incidence observed in the APHS. Similar patterns were observed in the rates of self-reported and clinically diagnosed hypertension. Additionally, average triglyceride (TG) levels were elevated in the APHS cohort relative to the TWB. Conversely, the APHS exhibited lower rates of self-reported and clinically diagnosed hyperlipidemia, as well as reduced mean low-density lipoprotein cholesterol levels, when compared to the TWB.

We proceeded to analyze the odds ratios for eight out of the 13 significant features to understand their impact on the incidence of T2DM across different age and sex groups in the APHS (as outlined in the subgroup analysis section in Fig. 1). Given the dataset's constraints, specifically the unavailability of HbA1c, PEF, PRS, and family history of diabetes in the APHS data, these factors were excluded from this comparison. Additionally, the analysis of age was omitted as it had been stratified by age groups. Instead of WHR, which could not be assessed due to the absence of hip circumference measurements, waistline measurements were considered. Our findings revealed distinct patterns in how these six features influenced the incidence of T2DM across various demographics, with the findings detailed in Fig. 4.

Notably, odds ratios for blood glucose were generally higher in females than in males across most age groups, aligning closely in those over 80 years old. For self-reported hyperlipidemia, odds ratios were initially higher in females than in males in those under 55, but this trend reversed between the ages of 55 and 80, with males exhibiting higher odds ratios; the rates then equalized in individuals older than 80. Furthermore, the impact of clinically diagnosed mixed hyperlipidemia was predominantly significant in females, with most male odds ratios approximating 1, indicating a negligible association. Similarly, the odds ratios for TG were consistently higher in females than in males, with significant positive effects in males only observed in those aged over 70.

In the case of waistline measurements, the odds ratios showed a decrease with advancing age in females, whereas in males, they remained consistent across all age groups. BMI-related odds ratios exhibited minimal variation between males and females across different ages, with a noted decrease in the odds ratios up until the age of 55, after which they stabilized. SGPT levels maintained general consistency across different age groups for both sexes, except in females under 55, where the odds ratios were elevated compared to those over 55. Lastly, for clinically diagnosed prediabetes, similar odds ratios were observed between males and females, with consistent results across various age groups.

DISCUSSION

In this study, leveraging the comprehensive clinical dataset provided by the TWB, we meticulously constructed a predictive model for incident T2DM over a forecast period of 3 years. Through rigorous selection and analysis, 13 significant predictors of T2DM were identified, covering a range of metrics from blood and body measurements to detailed disease histories. Furthermore, we undertook a comparative analysis with the UKB, examining the effect sizes of these predictors across the two cohorts. Remarkably, most identified predictors showed consistent effect sizes between the TWB and UKB, suggesting a degree of universal applicability in these metrics as indicators of T2DM risk across diverse populations. Furthermore, using the large APHS dataset, the effects of these predictors across different age and sex groups were examined in detail.

In our analysis, a significantly higher incidence rate of T2DM was observed in the TWB compared to the UKB. One possible reason is the difference in how incident T2DM was defined between the two cohorts. In the TWB, the NHIRD records of outpatient and inpatient visits fully covered the 3-year follow-up period for each participant, and the disease was defined solely using ICD codes. In contrast, in the UKB, GP records had different end dates for coverage across Scotland (May 2017), Wales

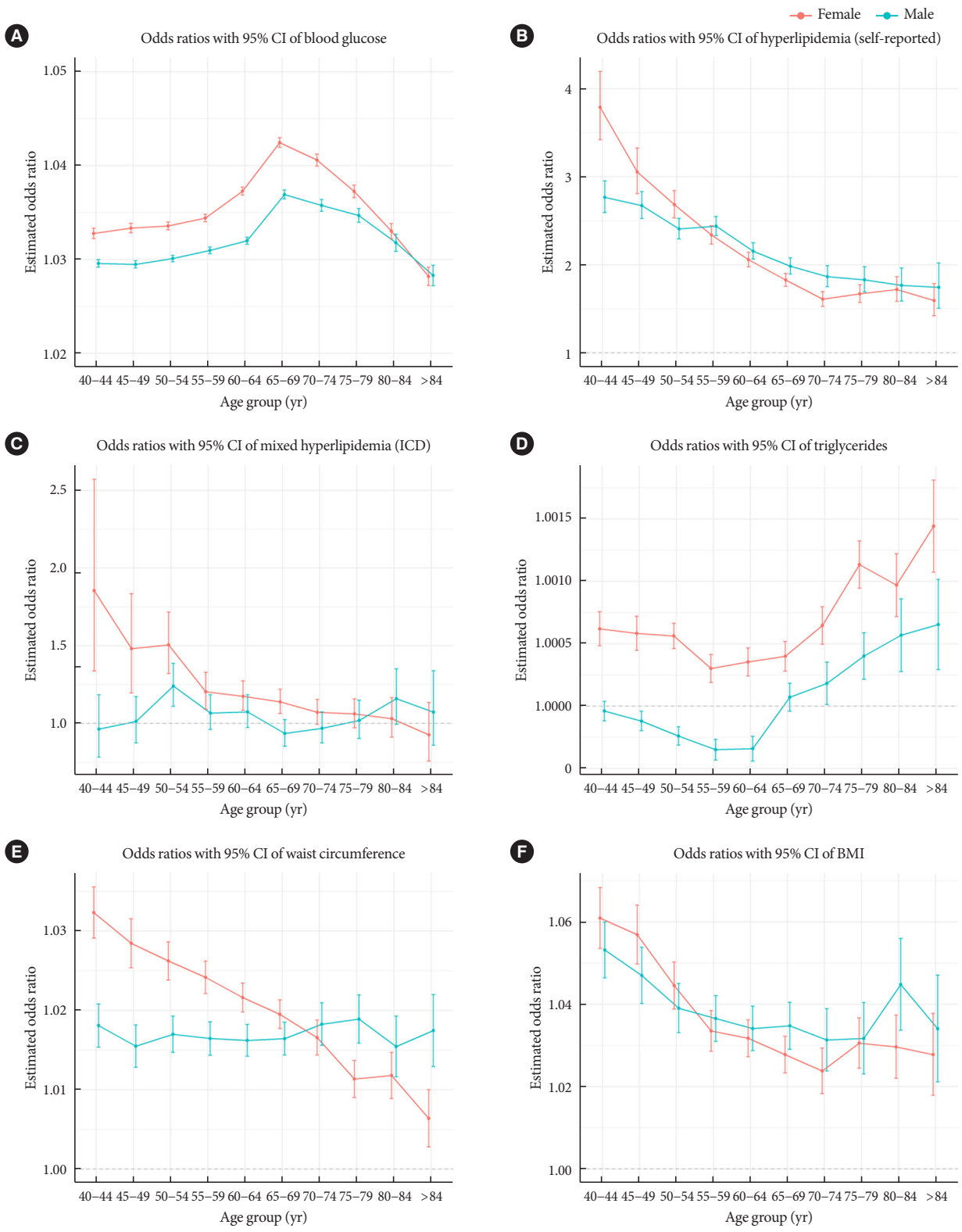


Fig. 4. Odds ratios and their 95% confidence intervals (CIs) of risk predictors in different age and sex groups. (A) Blood glucose, (B) hyperlipidemia, (C) mixed hyperlipidemia (International Classification of Diseases [ICD]), (D) triglyceride, (E) waist circumference, (F) body mass index (BMI), (G) serum glutamate pyruvate transaminase (SGPT), and (H) prediabetes. (*Continued to the next page*)

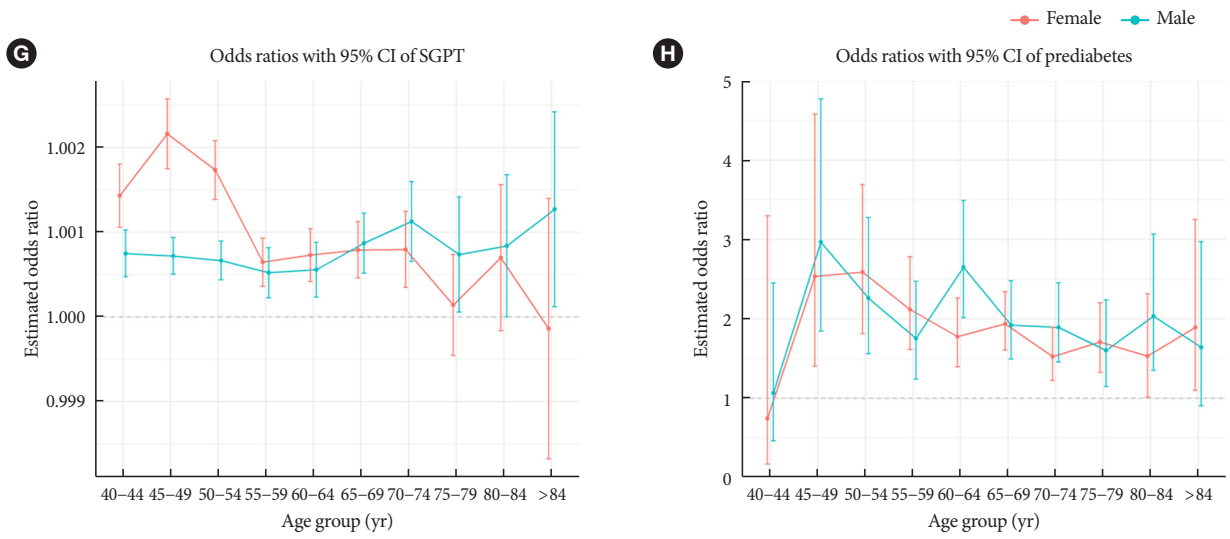


Fig. 4. Continued.

(September 2017), and England (ranging from August 2016 to June/July 2017), while the end dates of hospital inpatient data were August 2022, May 2022, and October 2022 for Scotland, Wales, and England, respectively. Hence, in addition to using ICD records, self-reported diabetes from follow-up questionnaires was also considered to define incident T2DM in the UKB.

Additionally, the declining trend in the incidence of clinically diagnosed T2DM in the UK, particularly after 2013, may contribute to the observed difference. A retrospective cohort study from the UK reported that the incidence rate of new clinical diagnoses of T2DM decreased by a third from its peak in 2013 [23], a trend also confirmed by two other studies [24,25]. This reduction may be partially explained by the shift towards diagnosing more cases of prediabetes rather than T2DM, as evidenced by the tripling of prediabetes diagnoses during the same period. The reliance on HbA1c alone, which may have lower sensitivity in detecting diabetes compared to fasting glucose, particularly in older adults and certain ethnic groups, could also lead to an underestimation of diabetes incidence in the UK [23], further explaining the lower observed rates compared to the TWB.

Conversely, the incidence of clinically diagnosed T2DM showed an increasing trend from 2005 to 2014 in Taiwan, likely due to the increasing prevalence of obesity and metabolic syndrome [26]. Notably, the incidence rate of clinically diagnosed T2DM among the 60 to 79 age group in the UK during 2009 to 2018 ranged from 12.49 to 13.69 per 1,000 person-years in males and 8.85 to 11.01 in females [23], whereas significantly higher rates were observed in Taiwan (25.52 to 26.91 per 1,000 person-years in males and 25.26 to 26.78 in females) [26]. This observation aligns with the fact that the proportion of diagnosed diabetes in the UK for the years 2021 to 2022 was about 6.4% [27], whereas it reached approximately 9.32% in Taiwan in 2014 [28].

The younger mean recruitment age in TWB hints at an earlier onset or identification of diabetes within the Taiwan population, which may reflect distinct genetic susceptibilities or lifestyle patterns predisposing to this condition. Higher levels of blood glucose and HbA1c in TWB participants point to a greater tendency towards glucose intolerance, suggesting that, despite lower average BMI values, this group may have a distinct metabolic profile that raises their risk of diabetes. Interestingly, while traditional risk factors such as BMI, hypertension, and hyperlipidemia presented more prominently in the UKB cohort, the higher incidence of diabetes in the TWB, particularly among those with a family history of the disease, accentuates the potential contribution of genetic predisposition. The disparities in healthcare practices, accessibility, and preventive measures between the two regions may also play a crucial role in these observed differences in the incidence rates.

The predictive model developed in this study demonstrates several key advantages over traditional methods like glucose or HbA1c measurements, particularly in the areas of early detection and personalized medical strategies. Unlike methods that rely solely on glucose levels or HbA1c, this model incorporates a broad spectrum of predictors including genetic factors, PEF, and detailed metabolic profiles, allowing for a more nuanced

understanding of diabetes risk. This comprehensive approach enables the detection of T2DM risk at earlier, potentially asymptomatic stages. Moreover, the inclusion of non-traditional predictors like PEF and various metabolic indicators provides a more holistic assessment of an individual's risk, beyond what can be ascertained through traditional blood tests alone. By leveraging these diverse predictors, the model facilitates more tailored and potentially more effective intervention strategies, enhancing the ability to manage or even prevent diabetes in a targeted manner.

In our analysis, we identified a combination of modifiable and non-modifiable risk factors as significant predictors for incident T2DM, each with distinct clinical implications. Modifiable risk factors, including elevated HbA1c, blood glucose, hyperlipidemia, high TG, elevated SGPT, increased BMI, and WHR, as well as the presence of prediabetes, underscore the importance of lifestyle and medical interventions in diabetes prevention. These factors can be effectively managed through dietary changes, increased physical activity, weight loss, and pharmacological treatments, which have been shown to significantly reduce the risk of progressing to T2DM [29-31]. On the other hand, non-modifiable factors such as age, family history of diabetes, and PRS highlight the need for early screening and personalized risk assessments in individuals predisposed to diabetes. By categorizing these risk factors, our findings support a dual approach in clinical practice: aggressively managing modifiable factors while closely monitoring those with non-modifiable risks to mitigate the overall incidence of T2DM.

Our analysis identified HbA1c and blood glucose as the top 2 significant predictors in the TWB, leading to an AUC of 87.36% in TWB and 82.61% in UKB. These two predictors have been consistently included in other prediction models based on Asian populations [9] and were also selected in the prediction model trained using the UKB data in our and other studies [12,32]. These results underscore the importance of HbA1c and blood glucose as independent predictors of risk, highlighting their critical role in assessing the risk of T2DM. Their combined use alone demonstrated substantial accuracy in prediction. In fact, in the 'Standards of Care in Diabetes: 2024' published by the American Diabetes Association [33], HbA1c was modified as the top of the testing priority for the diagnosis of diabetes followed by fasting glucose, supporting our analysis findings. However, a recent study based on pooled global data showed that relying solely on either fasting plasma glucose or HbA1c might significantly underestimate the true burden of diabetes, particularly in low- and middle-income countries where diagnosis rates are currently low [34]. This finding further supports our advocacy for including both glycemic indicators in regular health examinations, such as the APHS program in Taiwan, to enhance the risk prediction of T2DM.

In addition to HbA1c and blood glucose, our analysis highlighted lung function, measured by PEF, as the third most significant predictor for T2DM in TWB, with its significance also confirmed in the UKB. This aligns with prior research indicating that measures of lung function, including forced expiratory volume in 1 second (FEV_1), forced vital capacity (FVC), and the FEV_1/FVC ratio, are associated with the risk of developing T2DM [35,36]. Despite such associations, lung function parameters have seldom been integrated into T2DM risk prediction models. This oversight may stem from the historical focus on metabolic factors to the exclusion of respiratory health indicators. Our findings, therefore, advocate for a broader, multidisciplinary approach in constructing T2DM risk models, suggesting that incorporating lung function measures could enhance predictive accuracy and offer new insights into the systemic nature of T2DM pathophysiology.

To further address the potential impact of smoking and lung function on our findings, we conducted additional analyses within the TWB and UKB datasets. In TWB, smoking was significantly associated with T2DM incidence when only age was included as a covariate (odds ratio of 1.40 for current smokers vs. others, $P=5.14E-124$). However, smoking was not selected by the LASSO regression model, likely due to its correlations with other factors such as WHR, PEF, BMI, TG, and SGPT, suggesting its influence on T2DM risk may be indirectly captured through these variables, particularly lung function. Similarly, our analysis in UKB demonstrated that PEF is influenced by a wide range of factors, including age, sex, BMI, smoking, metabolic markers like HbA1c, and environmental exposures such as PM2.5 (Supplementary Table 2). These findings indicate that PEF reflects the cumulative effects of age, lifestyle, and metabolic factors, rather than serving as a direct mechanistic link to T2DM. Given the variability and multifactorial influences on PEF, its relationship with T2DM risk likely reflects broader systemic changes. Smoking, for example, induces pulmonary and systemic immunological changes that contribute to lung dysregulation and reduced lung function, while age-related declines in PEF and obesity-related changes in lung mechanics further highlight its complex associations with metabolic risk [37]. By framing PEF as an integrative marker of

health influenced by diverse factors, rather than focusing on direct mechanisms, our findings underscore the importance of considering lung function and related variables in T2DM risk prediction models. Future research using longitudinal designs and comprehensive pulmonary function testing is warranted to better characterize these interactions and enhance the predictive utility of PEF for metabolic risk assessment.

Our analysis highlights the contrasting impact of HbA1c levels and the PRS for T2DM incidence between the TWB and the UKB. The pronounced effect of HbA1c in the UKB compared to the TWB suggests possible differences in metabolic profiles or diabetes progression stages at baseline. On the other hand, the greater impact of PRS in the TWB reflects the potential for genetic risk factors to exhibit variable influence depending on the population's genetic architecture. Additionally, research has shown that PRS yields a higher odds ratio per standard deviation change for prevalent T2DM in TWB samples compared to European samples [38]. Such findings underscore the importance of considering ethnic and genetic diversity in T2DM research and highlight the need for further studies to explore these variations in risk factors across different populations.

The feature selection procedure using the UKB identified several population-specific variables that were not selected in the TWB, including HDL-C, reticulocyte count, and vascular/heart problems. These findings are consistent with previous studies using the UKB, which have also identified HDL-C and reticulocyte count as important predictors for future T2DM [39,40]. The identification of these additional features in the UKB highlights the influence of population-specific factors in diabetes risk prediction. This suggests that while there are core predictors common across populations, such as HbA1c, glucose, and BMI, the inclusion of population-specific variables can enhance the predictive accuracy of models tailored to different ethnic groups.

The stratification of our predictive model's accuracy by age groups underscores the nuanced dynamics between age, sex, and diabetes risk factors, revealing an intricate landscape of risk prediction across demographics. Notably, younger individuals exhibit higher predictive accuracy, primarily due to their increased sensitivity compared to older age groups. This enhancement in predictive accuracy among younger age groups may stem from the relative influence of other environmental and lifestyle factors that become more significant in older populations. Further exploration into the impact of various risk factors across different age and sex categories reveals that factors like self-reported hyperlipidemia and BMI have more pronounced effects in the younger demographic. Intriguingly, specific factors, such as waist circumference and clinically diagnosed mixed hyperlipidemia, show greater impacts uniquely in younger females, whereas their effects remain uniform across male age groups. This suggests that gender-specific physiological factors, possibly influenced by hormonal variations, significantly affect metabolic risk factors [41]. Additionally, the decline in the predictive power of waist circumference with advancing age in females could be attributed to postmenopausal changes, which alter body fat distribution and metabolic function. These insights advocate for the development of tailored predictive models that adapt to the unique risk profiles of different demographic segments, thereby enhancing the precision of diabetes risk prediction and offering potential pathways for more effective intervention strategies.

This study's findings must be interpreted within the context of its methodological limitations, particularly those stemming from the diversity in cohort characteristics. The TWB, UKB, and APHS program each have unique participant profiles, diagnostic criteria, and follow-up procedures. Specifically, the definition of diabetes and the diagnostic tools employed (e.g., fasting glucose vs. HbA1c measurements) differed notably between the cohorts. Such variations can lead to differential classification of diabetes cases, potentially skewing the incidence rates and the perceived effectiveness of the predictors used in our models. Additionally, the cohorts' follow-up intervals and the timing of assessments were not uniform, which may impact the longitudinal analysis of risk factor emergence and the progression of T2DM. Environmental factors, such as regional differences in diet, healthcare access, and lifestyle, which significantly influence diabetes risk, were also challenging to control across the international cohorts. These discrepancies can affect both the baseline risk and the progression of diabetes, complicating cross-population comparisons and the applicability of findings to other settings. Moreover, the retrospective design of our study introduces potential biases associated with the reliance on historical medical records and self-reported data. Such data sources are susceptible to inaccuracies due to recall bias or incomplete documentation, which could affect the results' validity. Despite these challenges, our study contributes to the understanding of diabetes predictors in diverse populations and underscores the need for context-specific diabetes risk assessments and predictive modeling.

In conclusion, our comprehensive analysis within the TWB and comparative insights with the UKB have illuminated critical aspects of T2DM risk prediction and the multifaceted nature of its determinants. By integrating a wide range of clinical, metabolic, and genetic predictors, we have developed a robust model that underscores the importance of a multifactorial approach to understanding and predicting T2DM. Our findings highlight not only the significant predictors such as HbA1c, blood glucose, and lung function (PEF) but also the intricate variations in their effects across different populations, age groups, and sexes. This research emphasizes the critical need for personalized and population-specific risk assessment models that consider the unique genetic, environmental, and lifestyle factors contributing to T2DM risk. Moving forward, our work advocates for the inclusion of a broader spectrum of risk factors in predictive models and underscores the potential benefits of incorporating such models into preventive healthcare strategies. By doing so, we aim to enable earlier identification and intervention for individuals at high risk of T2DM, ultimately contributing to better health outcomes and reducing the global burden of this chronic disease.

SUPPLEMENTARY MATERIALS

Supplementary materials related to this article can be found online at https://doi.org/10.4093/dmj.2024.0319.

CONFLICTS OF INTEREST

Wayne Huey-Herng Sheu has been a International Editorial Board Member of the *Diabetes & Metabolism Journal* since 2024. He was not involved in the review process of this article. Otherwise, there was no conflict of interest

AUTHOR CONTRIBUTIONS

Conception or design: Y.E.C., S.Y.C., H.Y.C., W.H.H.S., R.H.C.
Acquisition, analysis, or interpretation of data: Y.E.C., D.D.O., G.H.L., Y.S.Z.
Drafting the work or revising: all authors.
Final approval of the manuscript: H.Y.C., W.H.H.S., R.H.C.

ORCID

Ying-Erh Chen *https://orcid.org/0000-0001-8735-0385*
Ren-Hua Chung *https://orcid.org/0000-0002-9835-6333*

FUNDING

This study was supported by grants PH-112-GP-04 and PH-112-PP-10 from the National Health Research Institutes and MOST 110-2314-B-400-023 from the National Science and Technology Council in Taiwan.

ACKNOWLEDGMENTS

We thank the participants from the Taiwan Biobank, UK Biobank, and the Adult Preventive Healthcare Services program in Taiwan.

REFERENCES

1. Ogurtsova K, Guariguata L, Barengo NC, Ruiz PL, Sacre JW, Karuranga S, et al. IDF diabetes atlas: global estimates of undiagnosed diabetes in adults for 2021. Diabetes Res Clin Pract 2022;183:109118.
2. Forbes JM, Cooper ME. Mechanisms of diabetic complications. Physiol Rev 2013;93:137-88.
3. Sheng Z, Cao JY, Pang YC, Xu HC, Chen JW, Yuan JH, et al. Effects of lifestyle modification and anti-diabetic medicine on prediabetes progress: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 2019;10:455.
4. Tuomilehto J, Uusitupa M, Gregg EW, Lindstrom J. Type 2 diabetes prevention programs-from proof-of-concept trials to national intervention and beyond. J Clin Med 2023;12:1876.
5. Seah JY, Yao J, Hong Y, Lim CG, Sabanayagam C, Nusinovici S, et al. Risk prediction models for type 2 diabetes using either fasting plasma glucose or HbA1c in Chinese, Malay, and Indians: results from three multi-ethnic Singapore cohorts. Diabetes Res Clin Pract 2023;203:110878.
6. Mohsen F, Al-Absi HR, Yousri NA, El Hajj N, Shah Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. NPJ Digit Med 2023;6:197.
7. Asgari S, Khalili D, Hosseinpanah F, Hadaegh F. Prediction models for type 2 diabetes risk in the general population: a systematic review of observational studies. Int J Endocrinol Metab 2021;19:e109206.
8. Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of

existing models. Lancet Diabetes Endocrinol 2014;2:19-29.
9. Xu S, Coleman RL, Wan Q, Gu Y, Meng G, Song K, et al. Risk prediction models for incident type 2 diabetes in Chinese people with intermediate hyperglycemia: a systematic literature review and external validation study. Cardiovasc Diabetol 2022; 21:182.
10. Kokkorakis M, Folkertsma P, van Dam S, Sirotin N, Taheri S, Chagoury O, et al. Effective questionnaire-based prediction models for type 2 diabetes across several ethnicities: a model development and validation study. EClinicalMedicine 2023; 64:102235.
11. Rout M, Wander GS, Ralhan S, Singh JR, Aston CE, Blackett PR, et al. Assessing the prediction of type 2 diabetes risk using polygenic and clinical risk scores in South Asian study populations. Ther Adv Endocrinol Metab 2023;14:20420188231220120.
12. Lugner M, Rawshani A, Helleryd E, Eliasson B. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. Sci Rep 2024;14:2102.
13. Tsai SF, Yang CT, Liu WJ, Lee CL. Development and validation of an insulin resistance model for a population without diabetes mellitus and its clinical implication: a prospective cohort study. EClinicalMedicine 2023;58:101934.
14. Seah JY, Sim X, Khoo CM, Tai ES, van Dam RM. Differences in type 2 diabetes risk between East, South, and Southeast Asians living in Singapore: the multi-ethnic cohort. BMJ Open Diabetes Res Care 2023;11:e003385.
15. Cronje HT, Katsiferis A, Elsenburg LK, Andersen TO, Rod NH, Nguyen TL, et al. Assessing racial bias in type 2 diabetes risk prediction algorithms. PLOS Glob Public Health 2023;3: e0001556.
16. Goff LM, Ladwa M, Hakim O, Bello O. Ethnic distinctions in the pathophysiology of type 2 diabetes: a focus on black African-Caribbean populations. Proc Nutr Soc 2020;79:184-93.
17. Wang T, Zhao Z, Wang G, Li Q, Xu Y, Li M, et al. Age-related disparities in diabetes risk attributable to modifiable risk factor profiles in Chinese adults: a nationwide, population-based, cohort study. Lancet Healthy Longev 2021;2:e618-28.
18. Feng YA, Chen CY, Chen TT, Kuo PH, Hsu YH, Yang HI, et al. Taiwan Biobank: a rich biomedical research database of the Taiwanese population. Cell Genom 2022;2:100197.
19. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779.
20. Chung RH, Chuang SY, Chen YE, Li GH, Hsieh CH, Chiou HY, et al. Prevalence and predictive modeling of undiagnosed diabetes and impaired fasting glucose in Taiwan: a Taiwan Biobank study. BMJ Open Diabetes Res Care 2023;11:e003423.
21. Ruan Y, Lin YF, Feng YA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet 2022;54:573-80.
22. Van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1-67.
23. Pal K, Horsfall L, Sharma M, Nazareth I, Petersen I. Time trends in the incidence of clinically diagnosed type 2 diabetes and pre-diabetes in the UK 2009-2018: a retrospective cohort study. BMJ Open Diabetes Res Care 2021;9:e001989.
24. Sharma M, Nazareth I, Petersen I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. BMJ Open 2016;6:e010210.
25. Holden SH, Barnett AH, Peters JR, Jenkins-Jones S, Poole CD, Morgan CL, et al. The incidence of type 2 diabetes in the United Kingdom from 1991 to 2010. Diabetes Obes Metab 2013;15: 844-52.
26. Sheen YJ, Hsu CC, Jiang YD, Huang CN, Liu JS, Sheu WH. Trends in prevalence and incidence of diabetes mellitus from 2005 to 2014 in Taiwan. J Formos Med Assoc 2019;118 Suppl 2: S66-73.
27. Wise J. Diabetes cases in UK reach all time high, charity warns. BMJ 2023;381:848.
28. Tu ST. Taiwan diabetes atlas 2019: type 2 diabetes. Taipei: Taiwanese Association of Diabetes Educators; 2019.
29. Ley SH, Hamdy O, Mohan V, Hu FB. Prevention and management of type 2 diabetes: dietary components and nutritional strategies. Lancet 2014;383:1999-2007.
30. Arsenault BJ, Despres JP. Physical activity for type 2 diabetes prevention: some is better than none, more is better, and earliest is best. Diabetes Care 2023;46:1132-4.
31. Majety P, Lozada Orquera FA, Edem D, Hamdy O. Pharmacological approaches to the prevention of type 2 diabetes mellitus. Front Endocrinol (Lausanne) 2023;14:1118848.
32. Helmink MA, Peters SA, Westerink J, Harris K, Tillmann T, Woodward M, et al. Development and validation of a lifetime prediction model for incident type 2 diabetes in patients with established cardiovascular disease: the CVD2DM model. Eur J Prev Cardiol 2024;31:1671-8.
33. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: standards of care

in diabetes-2024. Diabetes Care 2024;47(Suppl 1):S20-42.
34. NCD Risk Factor Collaboration (NCD-RisC). Global variation in diabetes diagnosis and prevalence based on fasting glucose and hemoglobin A1c. Nat Med 2023;29:2885-901.
35. Lee HY, Shin J, Kim H, Lee SH, Cho JH, Lee SY, et al. Association between lung function and new-onset diabetes mellitus in healthy individuals after a 6-year follow-up. Endocrinol Metab (Seoul) 2021;36:1254-67.
36. Peng Y, Zhong GC, Wang L, Guan L, Wang A, Hu K, et al. Chronic obstructive pulmonary disease, lung function and risk of type 2 diabetes: a systematic review and meta-analysis of cohort studies. BMC Pulm Med 2020;20:137.
37. Tian T, Jiang X, Qin R, Ding Y, Yu C, Xu X, et al. Effect of smoking on lung function decline in a retrospective study of a health examination population in Chinese males. Front Med (Lausanne) 2023;9:843162.
38. Ge T, Irvin MR, Patki A, Srinivasasainagendra V, Lin YF, Tiwari HK, et al. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. Genome Med 2022;14:70.
39. Edlitz Y, Segal E. Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards. Elife 2022;11:e71862.
40. Bi Y, Yang Y, Yuan X, Wang J, Wang T, Liu Z, et al. Association between liver enzymes and type 2 diabetes: a real-world study. Front Endocrinol (Lausanne) 2024;15:1340604.
41. DeFronzo RA. Pathogenesis of type 2 diabetes mellitus. Med Clin North Am 2004;88:787-835.

SUPPLEMENTARY METHODS

Data quality control for study cohorts

Taiwan Biobank

Since its establishment in 2008, the Taiwan Biobank (TWB) has gathered samples from over 150,000 individuals aged 20 and older. About 115,000 of these participants have been genotyped using the TWB1.0 and TWB2.0 genome-wide association study (GWAS) single nucleotide polymorphism (SNP) arrays. The TWB1.0 chip contains about 653,000 SNPs designed on the Axiom Genome-Wide Array Plate system, and the TWB2.0 chip expands this design to approximately 750,000 SNPs. We followed standard GWAS quality control (QC) procedures, which involved removing SNPs not in Hardy-Weinberg Equilibrium, removal of duplicate samples, elimination of closely related participants, and exclusion of individuals whose genetic sex did not match their reported gender. After QC, the samples were imputed using the Michigan Imputation Server with the 1000 Genomes Project Phase 3 as the reference, focusing on SNPs with an Rsq greater than 0.3 and a minor allele frequency (MAF) over 1%. Subsequent analyses included over 7.3 million SNPs. Each participant's data in the TWB were linked to their medical records from the National Health Insurance Research Database (NHIRD), which covers the period from 2009 to 2020. The NHIRD provides extensive records and expenditure claims for various healthcare services, including outpatient, inpatient, ambulatory care, and pharmacy services for reimbursement purposes. We extracted data related to diagnoses (using International Classification of Diseases, 9th Revision [ICD-9] and ICD-10 codes) for outpatient, ambulatory, and inpatient services from 3 years before and after each participant's recruitment into the study.

UK Biobank

The UK Biobank (UKB) is an extensive, prospective study that initially enrolled more than 500,000 participants aged 40 to 69 during 2006–2010. Like the TWB, these individuals were genotyped using GWAS SNP arrays. Approximately 230,000 individuals were linked to their general practitioner (GP) records, and about 450,000 were linked to their hospital inpatient data. QC was first performed on the GWAS data according to the protocols established by Bycroft et al. [1]. This included the exclusion of heterozygosity outliers, individuals with atypical sex chromosome configurations not classified as XX or XY, and non-relatives. Our analysis primarily focused on the white British population, which reduced our sample size to 339,455 participants. Consistent with the TWB cohort procedures, we included individuals who had GP or hospital inpatient records for at least 3 years before and after their recruitment dates, narrowing down the analysis dataset to 86,484 individuals. For the GP records, Read V2 and V3 codes were converted to Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) using the lookup table provided by the NHS Digital Technology Reference data Update Distribution (TRUD) [2]. These SNOMED-CT codes were then mapped to ICD-9 using both one-to-one and one-to-many mapping tables from the Unified Medical Language System (UMLS) [3]. The imputed genotypes were utilized to calculate the polygenic risk scores (PRS), adhering to methods detailed by Bycroft et al. [1] and considering only those with imputation quality scores above 0.3.

Adult Preventive Healthcare Services program data

The Taiwan Health Promotion Administration offers an Adult Preventive Healthcare Services (APHS) program, targeting individuals aged 40 to 65 for a health checkup every 3 years, and annually for those above 55, indigenous people over 55, polio survivors over 35, and all individuals over 65. The program provides a subsidy of 520 Taiwan dollars (TWD) per case (an additional 200 TWD for those eligible for hepatitis B and C screening). The services include questionnaires (medical, family, and medication history, health behaviors, depression screening), physical examinations (general checkup, height, weight, blood pressure, body mass index [BMI], waist circumference), laboratory tests (urinalysis for protein, estimated glomerular filtration rate calculation, blood biochemistry tests including glutamate oxaloacetate transaminase, glutamate pyruvate transaminase, creatinine, blood sugar, lipids), and hepatitis B and C screening for those aged 45 to 79 (40 to 79 for indigenous people). It also offers health consultations on smoking cessation, alcohol moderation, betel nut avoidance, regular exercise, maintaining healthy weight, diet, accident prevention, and oral health care. We obtained APHS data on over 4 million individuals from 2013 to 2018. Given that individuals may participate in the APHS program multiple times, we retained only the first enrollment record for each individual when multiple records were present to maintain consistency and prevent duplication in our analyses.

Definitions of existing and incident type 2 diabetes mellitus

In the TWB, existing diabetes was defined as individuals who either self-reported having diabetes, had NHIRD records with ICD-9 codes beginning with the initial three digits of 250 or ICD-10 codes starting with E08–E13 within 3 years prior to the baseline, or had abnormal glucose levels (fasting glucose ≥126 mg/dL or glycosylated hemoglobin [HbA1c] ≥6.5%). Similarly, for the UKB, existing diabetes was determined for individuals who self-reported having diabetes, had GP or hospital records with ICD-9 codes starting with 250 or ICD-10 codes starting with E08–E13, or had abnormal glucose levels (HbA1c ≥6.5%). It is important to note that fasting glucose levels were not used as diagnostic criteria here because the UKB measured only random glucose levels. Finally, in the APHS program data, existing diabetes was identified in individuals who self-reported as having diabetes or who had NHIRD records with ICD-9 codes starting with 250 or ICD-10 codes starting with E08–E13 within 3 years prior to their enrollment. Blood glucose and HbA1c were not used as criteria since only random glucose was measured and HbA1c testing was not conducted. Individuals identified with existing diabetes were excluded from further analyses to focus on incident cases.

In both the TWB and the APHS data, incident type 2 diabetes mellitus (T2DM) was defined as those who had a minimum of three outpatient diabetes diagnoses or at least one diabetes-related hospitalization within a 365-day period, according to their ICD records in the NHIRD. The diagnosis was based on the same ICD codes as those used for existing diabetes. For the UKB, incident T2DM was defined as those who had a minimum of two outpatient diabetes diagnoses or at least one diabetes-related hospitalization within a 365-day period, according to their ICD records in the GP or hospital records. Furthermore, participants were considered to have incident T2DM if they self-reported diabetes in at least one of the three follow-up questionnaires within 3 years after the baseline.

Calculation of polygenic risk scores in the Taiwan and UK Biobanks

We used PRS-CSx [4] to compute the posterior effect sizes of the GWAS SNPs for both European and East Asian populations. PRS-CSx utilizes a Bayesian regression framework featuring a continuous shrinkage prior to estimate the effect sizes. It incorporates GWAS summary statistics from various ancestries into a unified shrinkage prior, adjusting for ancestry-specific allele frequencies and linkage disequilibrium (LD) patterns. We used publicly available T2DM summary statistics from extensive GWAS studies, including 74,124 T2DM cases and 824,006 controls from the European DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) consortium [5], and 77,418 T2DM cases and 356,122 controls from the East Asian Asian Genetic Epidemiology Network (AGEN) study [6] for PRS-CSx. The LD reference panels, which PRS-CSx constructed from the 1000 Genomes Project Phase 3 samples, were employed. These panels facilitated the calculation of posterior effect sizes for SNPs that correlate to HapMap3 SNPs with MAF greater than 1%. PRS-CSx generated both ancestry-specific and cross-ancestry posterior effect sizes. For further analyses, we utilized the ancestry-specific posterior effect sizes calculated for Europeans and East Asians.

We computed the population-specific PRS using the samples from the UKB and TWB. For each individual i, the PRS calculation in the was performed as follows:

$$PRS_i = \sum_{k=1}^{N} w_k d_{ki}$$

In this equation, w_k represents the posterior effect size of SNP k, derived from the N SNPs analyzed using PRS-CSx. The term d_{ki} denotes the dosage of SNP k in individual i. We utilized the '—score' option in PLINK2 to compute the PRS.

Risk prediction features for type 2 diabetes mellitus

In the TWB, significant risk predictors for T2DM were identified by compiling a comprehensive list of variables from survey data, body measurements, and laboratory tests. The survey data included basic demographic information such as age, sex, and socioeconomic status, along with health behaviors like alcohol consumption, smoking habits, and physical activity levels. It also incorporated self-reported conditions involving first-degree relatives, such as hypertension, hyperlipidemia, and glaucoma. Body measurements taken into account included BMI, blood pressure, and lung function tests. From the laboratory tests, variables such as fasting glucose, HbA1c, red blood cell counts, creatinine, and microalbumin levels were analyzed. Missing values were imputed using the "mice" package in R version 4.3.2 (R Foundation for Statistical Computing, Vienna, Austria) [7], which applies multivariate imputation by chained equations. Additionally, features related to T2DM co-

morbidities were extracted from the medical records in the NHIRD covering the 3 years before participants' recruitment. These comorbidities included:

· Obesity (ICD-9 code 278)
· Various heart diseases (ICD-9 codes 410–414, 420–425, 427, 429)
· Chronic heart failure (ICD-9 code 428)
· Ischemic stroke (ICD-9 code 434)
· Sleep apnea (ICD-9 code 780.57)
· Gestational diabetes (ICD-9 code 648.8)
· Polycystic ovary syndrome (ICD-9 code 256.4)
· Frozen shoulder (ICD-9 code 726)
· Hemochromatosis (ICD-9 code 275.03)
· Kidney failure (ICD-9 codes 584, 585)
· Dementia (ICD-9 codes 290, 294)
· Acanthosis nigricans (ICD-9 code 701.2)
· Blindness (ICD-9 code 369)
· Prediabetes (ICD-9 code 790.29)
· Essential hypertension (ICD-9 codes 401, 401.0, 401.1, 401.9)
· Mixed hyperlipidemia (ICD-9 code 272.2)

Altogether, 441 variables were considered in this study.

For the UKB, we compiled a similar list of potential predictive features from survey questionnaires, body measurements, and laboratory tests. Features with more than 20% missing data were excluded, and missing values were imputed using the 'mice' package in R. Features related to T2DM comorbidities, as mentioned above, were defined using ICD-9 and ICD-10 codes from hospital inpatient and GP data. A total of 18,300 features were compiled.

SUPPLEMENTARY REFERENCES

1. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203-9.
2. National Health Service England. TRUD: Technology reference update distribution. Available from: https://isd.digital.nhs.uk/trud/users/guest/filters/0/home (cited 2025 Mar 24).
3. National Library of Medicine. Unified Medical Language System (UMLS). Available from: https://www.nlm.nih.gov/research/umls/index.html (cited 2025 Mar 24).
4. Ruan Y, Lin YF, Feng YA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet 2022;54:573-80.
5. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 2018;50:1505-13.
6. Spracklen CN, Horikoshi M, Kim YJ, Lin K, Bragg F, Moon S, et al. Identification of type 2 diabetes loci in 433,540 East Asian individuals. Nature 2020;582:240-5.
7. Van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1-67.

Supplementary Table 1. Characteristics of participant groups across all ages and by age stratification in the Adult Preventive Healthcare Services data

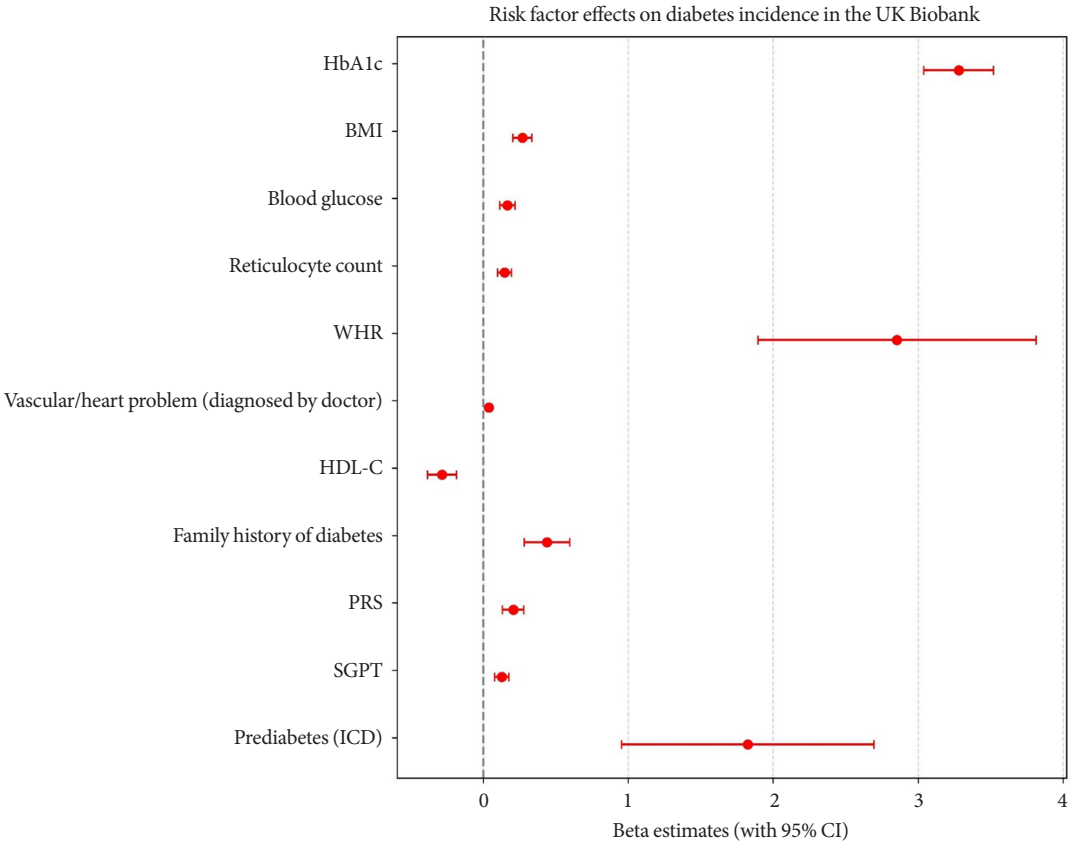
Characteristic	All	40–49 years	50–59 years	60–69 years	70–79 years	>79 years
Sample size[a]	4,368,325 (43)	1,222,837 (42.43)	1,349,041 (42.2)	1,074,167 (43.5)	509,516 (43.95)	212,764 (46.68)
Incident T2DM cases	486,020 (11.13)	64,173 (5.25)	128,292 (9.51)	174,929 (16.29)	86,643 (17.0)	31,968 (15.03)
Age, yr	57.71±11.67	44.38±3.0	54.46±2.86	64.0±2.71	74.02±2.8	84.11±3.62
BMI, kg/m^2	24.69±3.93	24.65±4.28	24.72±3.84	24.89±3.73	24.67±3.71	23.85±3.74
Waistline, cm	83.34±10.65	82.16±11.38	82.81±10.43	84.21±10.13	85.03±10.12	85.1±10.38
Hypertension (self-reported)	908,673 (20.8)	115,022 (9.41)	230,835 (17.11)	282,604 (26.31)	188,674 (37.03)	91,516 (43.01)
Hyperlipidemia/High cholesterol (self-reported)	216,078 (4.95)	35,564 (2.91)	66,022 (4.89)	70,960 (6.61)	32,417 (6.36)	11,105 (5.22)
Essential hypertension (ICD)[b]	513,921 (11.76)	36,104 (2.95)	101,727 (7.54)	149,016 (13.87)	152,189 (29.87)	74,882 (35.19)
Mixed hyperlipidemia (ICD)[b]	73,842 (1.69)	6,633 (0.54)	16,910 (1.25)	23,814 (2.22)	19,526 (3.83)	6,959 (3.27)
Random blood glucose, mg/dL	104.2±39.1	99.16±35.74	104.56±40.59	107.8±41.36	106.86±38.09	106.28±35.52
SBP, mm Hg	129.48±19.28	123.35±17.9	128.18±18.6	132.91±18.88	136.68±19.28	138.43±20.41
DBP, mm Hg	78.73±12.12	77.95±12.76	79.5±12.14	79.43±11.55	78.11±11.43	76.3±12.02
Total cholesterol, mg/dL	198.96±39.96	194.74±38.71	204.25±40.33	201.55±40.19	194.82±39.29	186.57±38.65
Triglycerides, mg/dL	141.59±132.84	146.55±161.37	147.38±140.28	138.56±109.44	128.83±91.17	122.17±79.63
LDL-C, mg/dL	114.55±38.22	110.3±37.2	118.04±39.15	117.29±38.49	112.76±37.11	107.28±35.69
HDL-C, mg/dL	55.1±17.39	54.73±16.95	55.74±17.75	55.22±17.38	54.82±17.44	53.34±17.34

Values are presented as number (%) or mean±standard error.
T2DM, type 2 diabetes mellitus; BMI, body mass index; ICD, International Classification of Diseases; SBP, systolic blood pressure; DBP, diastolic blood pressure; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol.
[a]Total number and percentage of male participants, [b]Diseases identified based on ICD codes in medical records.

Supplementary Table 2. Associations between peak expiratory flow and type 2 diabetes mellitus risk factors in the UK Biobank

Variable	Estimate	SE	*P* value
Age	–0.061	0.001	0
Sex[a]	–2.760	0.021	0
BMI	0.012	0.002	2.89E-10
Smoking (current)[b]	–0.606	0.026	3.34E-122
Smoking (ever)[b]	0.024	0.015	1.22E-01
NO_2	0.001	0.019	9.68E-01
NOx	–0.028	0.018	1.18E-01
PM2.5	–0.045	0.015	2.22E-03
PM10	–0.007	0.008	4.24E-01
HbA1c	–0.258	0.023	1.71E-28
Glucose	0.003	0.001	8.07E-07
Reticulocyte count	–0.539	0.199	6.76E-03
WHR	–1.603	0.128	4.47E-36
Vascular/heart problem	–0.329	0.047	1.71E-12
HDL-C	–0.001	0.001	1.35E-01
Family history of diabetes	0.065	0.018	2.47E-04
Polygenic risk score	–0.048	0.059	4.18E-01
SGPT	0.001	0.001	2.83E-01
Prediabetes	–0.362	0.221	1.01E-01

SE, standard error; BMI, body mass index; HbA1c, glycosylated hemoglobin; WHR, waist-to-hip ratio; HDL-C, high-density lipoprotein cholesterol; SGPT, serum glutamate pyruvate transaminase.
[a]Male is the baseline, [b]Never-smoking is the baseline.



Supplementary Fig. 1. The effect sizes and their 95% confidence intervals (CIs) for the risk factors identified in the UK Biobank. Note that variables such as blood glucose, reticulocyte count, high-density lipoprotein cholesterol (HDL-C), polygenic risk score (PRS), and serum glutamate pyruvate transaminase (SGPT) have been standardized. HbA1c, glycosylated hemoglobin; BMI, body mass index; WHR, waist-to-hip ratio; ICD, International Classification of Diseases.