Developing Effective Questionnaire-Based Prediction Models for Type 2 Diabetes for Several Ethnicities

Kokkorakis, Michail; Folkertsma, Pytrik; van Dam, Sipko; Sirotin, Nicole; Taheri, Shahrad; Chagoury, Odette; Idaghdour, Youssef; Henning, Robert H.; Forte, Jose Castela; Mantzoros, Christos S.

DOI:
10.2139/ssrn.4476201

Publication date:
2023

Citation for published version (APA):
Kokkorakis, M., Folkertsma, P., van Dam, S., Sirotin, N., Taheri, S., Chagoury, O., Idaghdour, Y., Henning, R. H., Forte, J. C., Mantzoros, C. S., de Vries, D. H., & Wolffenbuttel, B. H. R. (2023). *Developing Effective Questionnaire-Based Prediction Models for Type 2 Diabetes for Several Ethnicities*. SSRN. https://doi.org/10.2139/ssrn.4476201

Developing Effective Questionnaire-based Prediction Models for Type 2 Diabetes for Several Ethnicities

Michail Kokkorakis[1,2], Pytrik Folkertsma[3,4], Sipko van Dam[3,4], Nicole Sirotin[5], Shahrad Taheri[6], Odette Chagoury[6], Youssef Idaghdour[7,8], Robert H. Henning[1], Jose Castela Forte[1,3], Christos S. Mantzoros[2], Dylan H. de Vries[3,4#], Bruce H.R. Wolffenbuttel[4#]

[1]Department of Clinical Pharmacy and Pharmacology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[2]Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

[3]Ancora Health B.V., Groningen, The Netherlands

[4]Department of Endocrinology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[5]Department of Preventive Medicine, Cleveland Clinic Abu Dhabi, Al Maryah Island, Abu Dhabi, United Arab Emirates

[6]National Obesity Treatment Centre, Qatar Metabolic Institute, Hamad Medical Corporation, Doha, Qatar; Weill Cornell Medicine-Qatar, Qatar Foundation, Doha, Qatar

[7]Program in Biology, Division of Science and Mathematics, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[8]Public Health Research Center, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

#These authors jointly supervised this work

Corresponding author: Michail Kokkorakis, mkokkora@bidmc.harvard.edu, +31 616 70 99 27

Keywords: prediction, type 2 diabetes, machine learning, prevalence, incidence, population screening, ethnic diversity, digital health

Word count: 4,805; Figures: 3; Table: 1

Abstract

Background

Type 2 diabetes disproportionately affects individuals of non-white ethnicity through a complex interaction of multiple factors. Early disease prediction and detection is therefore essential and requires tools that can be deployed at large scale. We aimed to tackle this problem by developing questionnaire-based prediction models for type 2 diabetes for multiple ethnicities.

Methods

Logistic regression models, using questionnaire-only features, were trained on the White population of the UK Biobank, and validated in five other ethnicities and externally in Lifelines. In total, 631,748 individuals were included for prevalence prediction and 67,083 individuals for the eight-year incidence prediction. Predictive accuracy was assessed and a detailed sensitivity analysis was conducted to assess potential clinical utility. Furthermore, we compared the questionnaire algorithms to clinical non-laboratory type 2 diabetes risk tools.

Findings

Our algorithms accurately predicted type 2 diabetes prevalence (AUC=0·901) and eight-year incidence (AUC=0·873) in the White UK Biobank population. Both models replicate well in Lifelines, with AUCs of 0·917 and 0·817 for prevalence and incidence. Both models performed consistently well across ethnicities, with AUCs of 0·855 to 0·894 for prevalence and from 0·819 to 0·883 for incidence. These models generally outperformed two clinically validated non-laboratory tools and correctly reclassified >3,000 type 2 diabetes cases. Model performance improved with the addition of blood biomarkers, but not with the addition of physical measurements.

Interpretation

Easy-to-implement, questionnaire-based models can predict prevalent and incident type 2 diabetes with high accuracy across all ethnicities, providing a highly-scalable solution for population-wide risk stratification.

Funding

This project was funded by the UMCG under project number: PPP-2019_023. The funder had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Introduction

The number of individuals living with type 2 diabetes mellitus (T2D) is rapidly increasing globally, driven by factors such as aging, urbanization, sedentarism, and the increasing prevalence of obesity (1). In 2019, diabetes accounted for 66·3 million disability-adjusted life years (DALYs) and 4·2 million deaths among adults worldwide (2), with disproportionately steep prevalence and complications among non-white ethnic minorities in low-income and middle-income countries (3).

Populations of non-white ethnic backgrounds are disproportionately affected by diabetes, with a three to five times higher prevalence of T2D than people of White-European background (4). South Asians, for instance, usually develop T2D five to ten years earlier and experience a two- to six-fold increased risk of developing T2D compared to White European individuals (5). Likewise, 23% of Black African-Caribbean individuals with T2D are diagnosed under the age of 40 years in comparison to only 9% of White Europeans (6). Among the predominantly Arab population of the Gulf Cooperation Council countries, T2D prevalence has been suggested to be as high as 25% to 36% when undiagnosed case estimates are included and occurs at a younger age (7). A previous study in the United Arab Emirates showed a prevalence rate of adult T2D and undiagnosed diabetes at 25% and 14·8%, respectively (8). Despite the greater incidence and prevalence of T2D and associated comorbidities in these populations, publicly available diabetic registries and, validated prediction models for screening or early diagnosis remain scarce (9). Existing risk prediction tools in these populations have shown only moderate sensitivity and specificity and are not widely used in clinical practice (10).

The clinical value of non-laboratory incident T2D prediction tools is well established; however, they lack extensive validation in a wide variety of ethnicities (11, 12). Data science and specifically Machine Learning (ML), has shown high potential to further improve risk stratification across a range of clinical applications, including early disease prediction in diabetes (13). More importantly, ML-based technologies can accommodate population-wide non-invasive screening, allowing for initial assessments and subsequent referrals (14). Large population cohorts, such as the UK Biobank (UKB) and Lifelines (LL), constitute a suitable platform for developing and validating data-driven population risk stratification algorithms. These biobanks comprise rich anthropometric, lifestyle, and medical information data, as well as long-term follow-up on disease outcomes of almost 700,000 individuals in total. Of the UKB participants, circa 82% self-identified as “White” and almost 18% self-identified as having a different ethnic background, henceforth referred to as “non-white”, such as “East Asian or South Asian” ancestry, “Black, African, Caribbean, or other Black” ancestries, “Mixed” ancestries, and “Other” ancestries.

In this context, we aimed to develop ML models to predict the prevalence and an eight-year incidence of T2D that could be easily and widely implemented for population screening across multiple ethnicities. We trained questionnaire-based algorithms in the White population of the UKB and validated them internally within the non-white ethnic groups and externally in LL. Finally, we assessed the algorithms’ potential clinical utility against two other ML-based models and two gold-standard clinical risk models. Herewith, we showcase significantly enhanced prediction models that can transform primary diabetes care.

Methods

Setting

The UKB is the largest longitudinal population-based cohort, consisting of 502,507 participants aged between 37–73 years old, recruited between 2006 and 2010 (15). For the UKB, ethical procedures are controlled by a dedicated Ethics and Guidance Council (http://www.ukbiobank.ac.uk/ethics). All participants provided written informed consent prior to enrollment. The validation cohort, LL, is a comprehensive and prospective White-European-based population cohort from the northern Netherlands. LL contains data from 168,205 participants collected between 2006 and 2013 (16). Similarly, all participants provided written informed consent prior to enrollment. For a complete overview of the collected data, please see https://www.ukbiobank.ac.uk/register-apply/ and https://catalogue.lifelines.nl/.

Type 2 Diabetes Classification

In the UKB, T2D diagnoses were assigned based on self-reported T2D, diabetes diagnosed by a doctor and T2D hospital record annotation based on the International Classification of Diseases (ICD-9 codes 250.X0, 250.X2, and ICD-10 codes E11.X). Supplementary Table S1A demonstrates the data fields associated with the age of diagnosis that were employed to calculate the years until diagnosis from the initial assessment. In cases where more than one age of diagnosis was reported, the lowest reported age was used. All cases diagnosed before their assessment center visit were then annotated as prevalent cases, while cases diagnosed after their assessment were annotated as incident cases.

In LL, prevalent and incident T2D were annotated based on self-reported T2D (Supplementary Table S1B). Ages of diagnosis were not asked for during follow-up, and T2D follow-up was only asked for some assessments (2A, 3A and 3B), while general diabetes follow-up was asked for all assessments (1B, 1C, 2A, 3A and 3B). Therefore, we estimated the age of T2D diagnosis for every incident case by taking the mean of the age the participant had at the assessment reporting a T2D diagnosis and the age at the previous assessment. To calculate more specific ages of T2D diagnosis, if an incident case had reported a general diabetes follow-up diagnosis before their T2D diagnosis, the mean of the age during that assessment and the previous assessment was used instead to determine the age of T2D diagnosis.

Both in the UKB and LL, all participants with glucose >7 mmol/L or HbA1c >48 mmol/L but without diagnosis were annotated as having undiagnosed T2D.

Input features

All categorical features were transformed to one-hot encoding, and the original categorical feature in numerical format was also kept. Due to the large number of candidate features in the questionnaire, we performed feature selection: we started with an initial list containing all features and sub-selected those with an absolute correlation greater than 0·02 to the target outcome. We then reduced this list to ten features by iteratively extracting the top correlated feature and regressing this feature from the rest of the features. To allow for external validation, we mapped the input features from the UKB to their associated or closest available LL feature (Supplementary Table S2). During feature selection, missing values were imputed using the mean. To investigate whether adding basic measurement and biomarker features improved model performance, we added these features to the questionnaire feature pool and performed feature selection and model training again (Supplementary Table S4).

Data preparation

For the prevalence analyses, everyone with glucose >7 mmol/L or HbA1c >48 mmol/L without a T2D diagnosis was removed from the dataset in an attempt to remove possible undiagnosed cases. For the incidence analyses, we first removed anyone with diagnosed T2D at baseline and participants with glucose >7 mmol/L or HbA1c >48 mmol/L. Additionally, we removed all incident T2D cases with more than eight years until diagnosis and all persons not developing T2D but not returning to the assessment center after eight years. Because the different inclusion criteria result in an under-representation of controls, we corrected the incidence in every ethnicity subset by oversampling the controls to obtain the incidence we observed when including remeasured participants only.

Model Training and Testing

We set out to predict prevalent and incident T2D across all ethnic groups of the UKB and in LL using questionnaire-based ML models. Self-reported ethnicity was extracted from the UKB, and participants were divided into six different ethnicity groups (Supplementary Table S3). We used Sklearn's LogisticRegression with default settings for model training on the White ethnic population group using ten-fold cross-validation (17). The model's performance was internally validated in the five other ethnicity categories of the UKB and externally validated in the independent LL cohort. All input features were normalized by fitting Sklearn's StandardScaler on the train set, then using this scaler to scale the features in both the train and test sets.

Moreover, we validated the non-laboratory clinical concise Finnish Diabetes Risk Score (FINDRISC) and the clinical Australian Type 2 Diabetes Risk Assessment Tool (AUSDRISK), which employ 9 and 13 features, respectively, spanning medical history, demographics, lifestyle, and anthropometrics, to predict incident T2D (11, 12).

Statistical Analysis and Risk Stratification

The Area Under the Receiver Operating Characteristics (AUC) values and associated CI were calculated using DeLong's method from the R pROC package (18). Additionally, AUC curves were compared to test for significant differences using the DeLong ROC test from the same package (18). To assess the potential clinical utility of the models across different populations, we took a two-step approach to risk stratification. First, we compared the ability of all models to identify individuals at high risk in the general population (including those with and without diabetes for prevalence, and those who did and did not develop diabetes for incidence). Youden's method was used to find the risk threshold yielding the best sensitivity/specificity balance. In addition to sensitivity and specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) and the respective Confidence Interval (CI) were calculated using the R epiR package (19). Then, we simulated another potential application of the incidence models across the different study populations. We stratified the population into three risk groups, each with exactly one-third of the incident T2D cases, aiming to identify the greatest number of individuals that eventually developed T2D during the follow-up period by screening the smallest possible population. Ultimately, to evaluate the improvement in risk prediction provided by our models compared to the abovementioned clinical tools we conducted reclassification analysis by calculating the categorical Net Reclassification Improvement (NRI) using the R Hmisc package (20). To ensure fair comparisons between models, we matched the sizes of the risk groups in the clinical models with our own risk groups, which were determined based on the maximum Youden's index.

Data and Resource Availability

Study data are available from UKB and LL but were used under license for the current study, which restricts their public availability. Data are, however, available from the authors upon reasonable request and when granted permission by the UKB and LL.

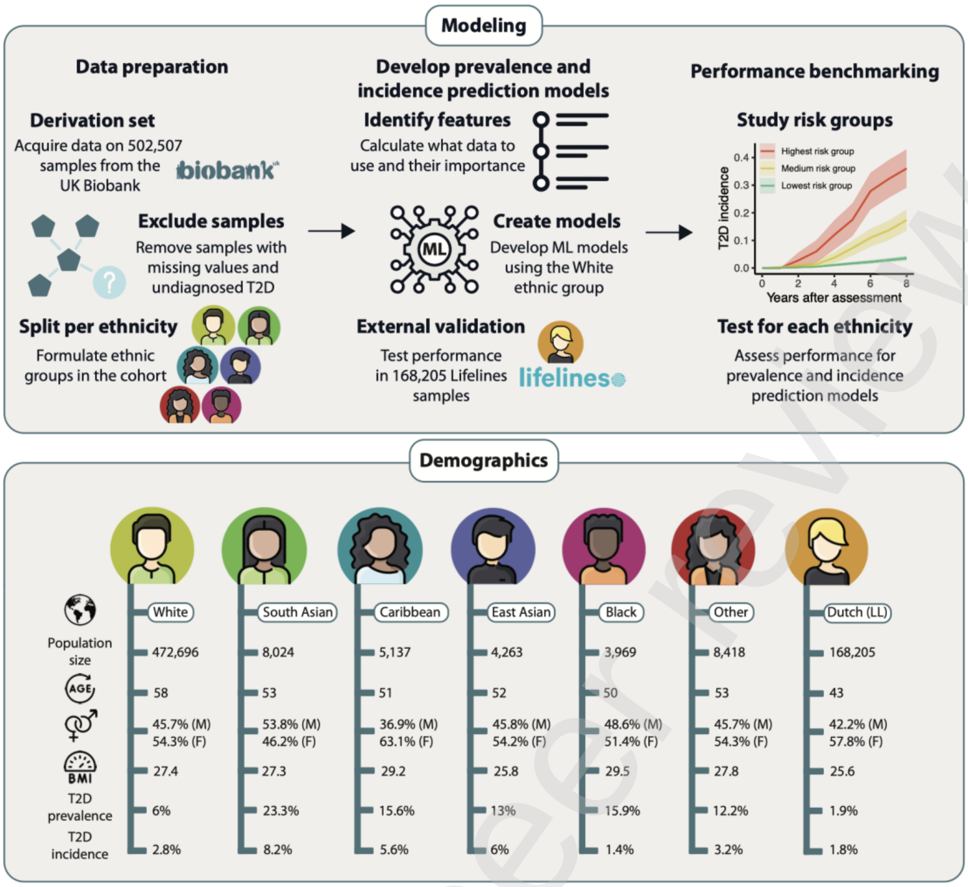
Code Availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Results

Baseline Characteristics

We set out to predict prevalent and incident T2D across all ethnic groups of the UKB and in LL using questionnaire-based ML models (Fig. 1). The included total group size for prevalent and incident T2D prediction models was 631,748 and 67,083 individuals, respectively. Baseline characteristics of the six ethnicity groups and LL are briefly presented in Figure 1 and in more detail in Supplementary table S4. Of note, the prevalence and incidence rates of T2D differed greatly between White and non-white populations, with non-white populations having between two- to almost four-fold higher prevalence (12·2-23·3%) and from half to as high as three-fold higher incidence (1·4-8·2%), than the White population of the UKB (6% and 2·8%, respectively). In contrast, LL had a lower prevalence (1·9%) and incidence (1·8%) of T2D compared to White UKB, in part explained by the age differences between these two populations.


Modeling
Data preparation
Derivation set
Acquire data on 502,507 samples from the UK Biobank
biobank
Exclude samples
Remove samples with missing values and undiagnosed T2D
Split per ethnicity
Formulate ethnic groups in the cohort
Develop prevalence and incidence prediction models
Identify features
Calculate what data to use and their importance
Create models
ML
Develop ML models using the White ethnic group
External validation
Test performance in 168,205 Lifelines samples
lifelines
Performance benchmarking
Study risk groups
Highest risk group
Medium risk group
Lowest risk group
T2D incidence
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
Years after assessment
Test for each ethnicity
Assess performance for prevalence and incidence prediction models
Demographics

	White	South Asian	Caribbean	East Asian	Black	Other	Dutch (LL)
Population size	472,696	8,024	5,137	4,263	3,969	8,418	168,205
AGE	58	53	51	52	50	53	43
⚥	45.7% (M) 54.3% (F)	53.8% (M) 46.2% (F)	36.9% (M) 63.1% (F)	45.8% (M) 54.2% (F)	48.6% (M) 51.4% (F)	45.7% (M) 54.3% (F)	42.2% (M) 57.8% (F)
BMI	27.4	27.3	29.2	25.8	29.5	27.8	25.6
T2D prevalence	6%	23.3%	15.6%	13%	15.9%	12.2%	1.9%
T2D incidence	2.8%	8.2%	5.6%	6%	1.4%	3.2%	1.8%

Figure 1. Workflow showing the steps taken to prepare the data and to create questionnaire-based prediction models for prevalent and incident T2D. The lower panel shows the means of percentages of some essential demographic features for the ethnic populations within the UK Biobank and Lifelines (LL).

Contribution of Questionnaire Features

The correlation between different questionnaire features pertaining to nutrition, smoking, physical activity, medication, and medical history and prevalent or incident T2D for each population are presented in detail in Supplementary Figures S2A and S2B. The contribution of each feature to the prevalence and incidence model is shown in Fig. 2A and 2B. Both prevalence and incidence models put high importance on BMI and the number of medications taken, positioning them in the top three features of both models. Furthermore, incidence includes a feature representing to sedentarism (time spent watching television (TV)). We observe an evident performance saturation with five to six input variables, particularly for prevalence prediction.

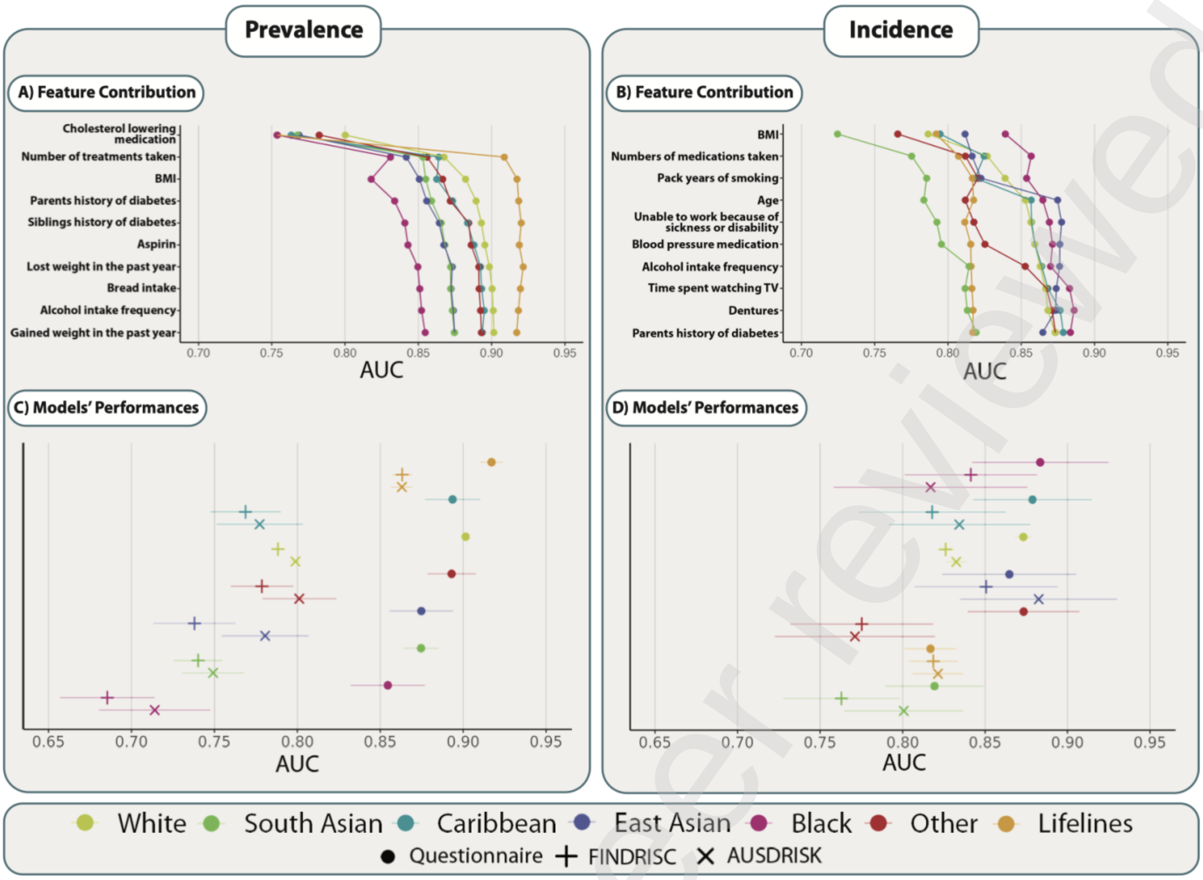


Figure 2. List of features in the prevalence (A) and incidence (B) prediction models and their contribution to the models' performance. Below, the performance of different models across populations for prevalence (C) and incidence (D) is shown. Each color-symbol combination refers to a specific model and population, explained in detail in the bottom panel. The AUC and 95% CI are shown for all models.

Performance of Type 2 Diabetes Prediction Models

With ten questionnaire features, the performance of prevalence prediction models measured by their AUC ranged from 0·855 to 0·901 (Fig. 2C and Supplementary Fig. 3A) within the UKB populations and an AUC of 0·917 in the independent validation cohort LL. For models predicting incident diabetes in the UKB, AUCs ranged from 0·819 to 0·883 (Fig. 2D and Supplementary Fig. 3B), while in LL the AUC was 0·817. The detailed performance metrics of the questionnaire-only models are shown in Supplementary Tables S5A and S5B.

Additionally, we performed an exploratory analysis of the potential added benefit of two other types of models: one including basic physical measurements and one including blood biomarkers (Supplementary Fig. S4A, S4B, S5A, S5B, S7A, S7B, S8A, S8B). For prevalence prediction, including basic measurements significantly improved the performance of questionnaire-only models for all UKB populations, except for Other, yet lowered the AUC of LL (Supplementary Table S8A, Supplementary Fig. S10). In contrast, for incidence prediction, adding basic measurements significantly increased the performance of only two populations, UKB White and LL, though all populations showed higher AUCs. Including biomarkers led to a significant improvement in all instances except for incidence prediction among the Black population, where the Questionnaire-only models seem to yield a marginally higher performance (Supplementary Fig. S10 and Supplementary Tables S8A, S8B). The feature importance of these models is shown in Supplementary Fig. S4A, S4B, S7A, S7B.

Comparison with non-laboratory clinical risk models

We then also compared the questionnaire-only models to two clinically validated non-laboratory risk scores. First, we tested the performance of the concise FINDRISC, developed as a simple screening tool for individuals at high-risk of developing T2D. We observed that the questionnaire-based models significantly outperformed FINDRISC for prevalence prediction in all populations, and they significantly outperformed FINDRISC in four out of seven populations for predicting incidence (Fig. 2C, 2D, and Supplementary Tables S9A, S9B). Similarly, the questionnaire-based models significantly outperformed the AUSDRISK models in all prevalence predictions as well as in three out of seven populations for incidence prediction (Fig. 2C, 2D, and Supplementary Tables S9A, S9B). In all other instances, there were no significant differences, however our models yielded overall higher AUCs.

Sensitivity analysis and clinical utility of risk stratification

Finally, we conducted an in-depth sensitivity analysis of the risk stratification for all models to assess their potential clinical utility (Supplementary Tables S5A, S5B, S6A, S6B, S7A, and

S7B). Based on the thresholds provided by the Youden index, the questionnaire-only models obtained very high sensitivity-specificity balance, PPV, and NPV. Both sensitivity and specificity were consistently high (above 74% and 83% for prevalence, and 75% and 68% for incidence, respectively) for all populations. The corresponding NPVs for all models were above 93% and 98% for prevalence and incidence, respectively. For the models including biomarkers, further improvement in the sensitivity-specificity balance was seen, with a lower proportion of individuals identified as high risk also translating to higher PPV across the populations for prevalence and incidence. All corresponding NPVs were above 97% and 99% for prevalence and incidence, respectively.

In the second step of the analysis, we observed that the questionnaire-only models can identify small groups of very high risk individuals who eventually developed diabetes during follow-up (Fig. 3). By screening as little as 0·47% to 7·6% of different populations, the questionnaire-only models identified 33% of all individuals who developed T2D. In these high-risk groups, the average incidence of T2D was at least ten-fold higher compared to the lowest-risk group. The models also identify 66% of all individuals who developed T2D while screening only between 11·5% to 23·1% of all individuals across different populations. These slightly larger groups also show at least a six-fold higher risk across all populations, compared to lowest risk population. For the two other types of models (with additional physical measurements and the ones with the addition of biomarkers), the highest risk groups generally showed even higher average incidence despite the similar size (Supplementary Fig. S6 and S9). For all ethnicities, 66% of incident T2D cases could be identified by screening less than 10% of each population using the model, including biomarkers.

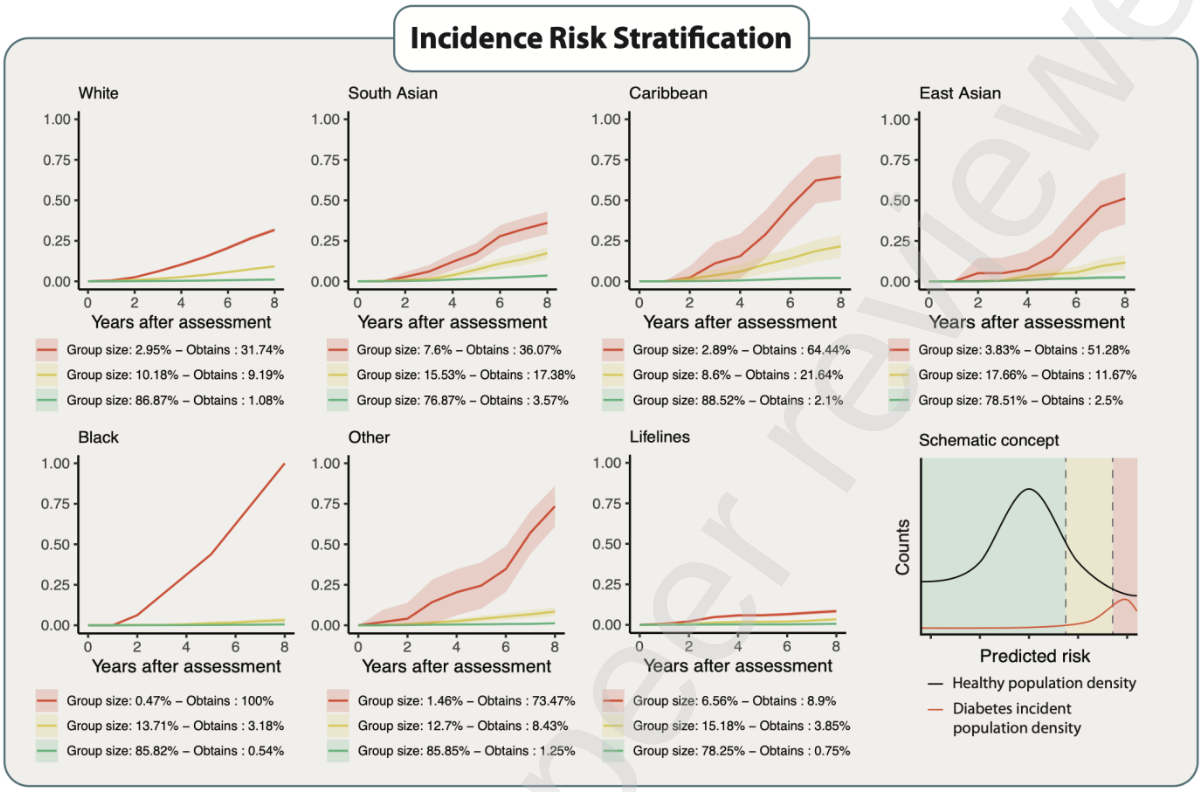


Figure 3. Risk identification for developing T2D. The x-axis represents the interval of years between the biobank entry and the moment of receiving a diagnosis of T2D. The y-axis represents the incidence of T2D. The stronger-

colored lines represent the group sizes, and the lighter-colored lines show the 95% CI. The bottom-right panel conceptualizes the risk groups (green, yellow, and red areas), while each group contains 33% of all T2D incident cases (area under the orange curve).

Reclassification Analysis

Ultimately, the reclassification analysis demonstrates that in almost all cases our models correctly reclassify more cases than the clinical tools FINDRISC and AUSDRISK. Notably, for the White, Caribbean, Other, and South Asian populations our models correctly reclassify more events reaching statistical significance compared to FINDRISC. Compared to AUSDRISK, our models reach statistical significance among the Whit and Other populations in correctly reclassifying T2D cases, along with statistically significant NRI values (Table 1, Supplementary Table S10A). The addition of physical measurements overall reclassifies more events correctly and seems to perform better in LL, compared to the Questionnaire Models (Supplementary Table S10B). The models also including biomarkers, outperform the clinical tools and reach clinical significance in almost all instances (Supplementary Table S10C). The high/low risk reclassifications, along with NRIs, and reclassification of non-event percentages are demonstrated in detail in the Supplementary Tables 10A-C.

Table 1. Reclassification analysis comparing our questionnaire-based models to FINDRISC and AUSDRISK. Positive reclassification events indicate that our models correctly reclassify more cases than the other two models, whereas negative events indicate the opposite. Reclassification percentages (%) are represented along with the CI, as well as the reclassification of events per 10,000 individuals with CI.

Risk model	**Ethnicity**	**Reclassification events %**	**Reclassification events N per 10,000**	**P-value**
FINDRISC	White	6·4 (5·2 – 7·6)	637 (519 – 756)	<0·001
FINDRISC	Black	2·2 (-5·2 – 9·5)	217 (-518 – 953)	0·6
FINDRISC	Caribbean	12·6 (3·7 – 21·5)	1,264 (374 – 2,154)	0·005
FINDRISC	East Asian	9·8 (-2·8 – 22·4)	984 (-278 – 2,245)	0·1
FINDRISC	Other	14·8 (6·4 – 23·3)	1,481 (637 – 2,326)	<0·001
FINDRISC	South Asian	12·7 (6·1 – 19·3)	1,269 (610 – 1,928)	<0·001
FINDRISC	Lifelines	-2·8 (-6·3 – 0·7)	-279 (-627 – 69)	0·1
AUSDRISK	White	5·9 (4·4 – 7·4)	591 (441 – 741)	<0·001
AUSDRISK	Black	3·4 (-8·2 – 15·1)	345 (-819 – 1,509)	0·6
AUSDRISK	Caribbean	5·7 (-3·9 – 15·3)	571 (-389 – 1,532)	0·2

AUSDRISK	East Asian	0 (-16·6 – 16·6)	0 (-1,656 – 1,656)	1
AUSDRISK	Other	25·6 (14·7 – 36·6)	2,564 (1,472 – 3,656)	<0·001
AUSDRISK	South Asian	7·8 (-0·9 – 16·4)	776 (-91 – 1,642)	0·08
AUSDRISK	Lifelines	0·4 (-3·7 – 4·4)	38 (-365 – 441)	0·9

Discussion

In this study of over 600,000 individuals, we showed for the first time that questionnaire-based ML models can accurately predict T2D prevalence and eight-year incidence across all ethnicities present within the UKB, as well as the LL external validation cohort. For almost all ethnicities, these models outperformed two established clinically validated T2D risk assessment tools. Despite the improvement in performance verified with the addition of blood biomarkers, the questionnaire-only models showed clinical utility for the detection of prevalent and incident T2D.

Previous research on the performance of prediction models for incident T2D has shown substantial differences across ethnicities. A re-estimation of the Atherosclerosis Risk in Communities (ARIC) model for the prediction of five-year diabetes risk in the Coronary Artery Risk Development Study in Young Adults (CARDIA) cohort showed significant differences in performance between White and African Americans (AUC 0·902 vs 0·816) (21). Another study of 12,043 Black and White individuals focusing on T2D prediction using anthropometric features and lipid levels reported an AUC of 0·79 (22). In this study, we observed less variation in the model performances between White and Black individuals for both prevalent and incident T2D prediction. The models developed herein outperform what has been previously demonstrated in Black populations, even without glucose as an input feature, and contradict the results of previous analyses that suggested that risk scores trained in European-descent population are not applicable to other ethnic groups (22, 23). Additionally, our questionnaire-based models significantly outperformed FINDRISC and AUSDRISK across all seven populations for prevalent T2D detection. For incidence, our models outperformed the above-mentioned tools in four populations compared to FINDRISC and three populations compared to AUSDRISK. This is especially relevant since both FINDRISC and AUSDRISK have been shown to perform only moderately well in several non-white populations (24, 25), despite AUSDRISK including ethnicity as an input feature and being intended to be used in the ethnically diverse Australian population (26). As expected, the addition of blood biomarkers to the models resulted in further improvements in predictive performance with AUCs generally above 0·90, mainly due to high correlations conferred by these features (Supplementary Fig. S7A, S7B, S10). Despite being significant, these improvements in AUC were not substantial enough to unequivocally justify their deployment over the questionnaire-only models considering the practical challenges discussed further in detail below.

As such, the goal of population-level risk stratification is not merely to predict individual risk accurately but to clearly distinguish groups with different levels of risk (27). To assess the potential stratification utility of our models, we first optimized their sensitivity-specificity balance with the Youden index. We found that all models achieved high to very high sensitivity and specificity for both prevalence and incidence prediction across all ethnicities. Given the low prevalence and incidence of T2D in White populations, a high specificity and NPV were expected for the White UKB population and LL. However, specificity and NPV remained high even in other ethnicities with higher prevalence and incidence rates (Supplementary Tables S5A, S5B, S6A, S6B, S7A, and S7B). The main difference with the addition of biomarkers was the increase in PPV, stemming from the lower number of individuals identified as high risk (between 20% and 29% for questionnaire-only predictions and generally around 18% when biomarkers were included). However, we also aimed to assess the usefulness of the models in settings where resources are limited, or population health data is lacking and where it is essential to accurately identify as many high-risk individuals as possible while minimizing the number of screened individuals. In such instances, screening more than a quarter of the population might be prohibitive from a cost and logistics perspective, hampering the model's clinical utility. Herein, we demonstrated that all models can also be applied to identify smaller groups of individuals at very high risk and that 33% and 66% of all incident diabetes cases can be identified by screening less than 10% and 23% of the population using the questionnaire-only models, respectively.

The data from these two simulated scenarios suggests that while there is a benefit from including additional measurements in risk stratification models, questionnaire-only models predict prevalent and incident diabetes with high accuracy and clinical utility. By not being subject to the practical limitations associated with collecting physical measurements or biomarkers, a questionnaire-based tool comprises the first step towards identifying an initial high-risk population that could be referred for subsequent diagnostic or prognostic assessment in a primary care setting. At a sensitivity and specificity as high as 80%, we see that questionnaire-only models applied to the largest population we studied, with almost 180,000 White individuals in the UKB training set, would recommend follow-up for less than 40 thousand individuals based on their eight-year risk, and around 65,000 of the more than 300,000 individuals potentially undiagnosed with T2D. In the context of population health prevention programs, deploying more selective models brings about two advantages. On the one hand, it requires considerably fewer individuals to be screened to detect a substantial portion of high-risk individuals. On the other hand, in line with previous research, it has been shown that such programs are most effective when targeted at a specific outcome, such as T2D risk reduction, and when including high-risk individuals, as opposed to a non-stratified population (28). Based on our reclassification analyses, all models developed herein, can correctly reclassify predicted T2D cases and in many instances outperform the currently available models. Of note, our models have demonstrated significantly better net reclassification improvements and correctly reclassify more events when compared to available clinical tools. Specifically, when compared to FINDRISC, there is an additional 3,387 positive cases that are correctly reclassified using our models, per 10,000 events, reaching statistical significance. Likewise, for the comparisons with AUSDRISK the respective amount of positive cases that are correctly and significantly reclassified using our models is 3,155 per 10,000 cases.

Eventually, translating the models presented in this study into population health risk stratification tools for primary diabetes care is not without challenges. In fact, most digital health innovations fail to advance into clinical practice or fall short of their anticipated impact (29). This lack of adoption is often the result of a poor understanding of end-user needs and inability to integrate the solution into current care frameworks (29). We built questionnaire-only models with the intent that individuals could complete them, potentially digitally, without requiring invasive biomarker collection or a visit to primary care facilities. While not replacing a trained clinician's evaluation, a patient-centered tool would facilitate timely screening and reach a larger audience by eliminating the need for primary care visits in the first phase. Policymakers have been encouraged to focus on prevention and innovating to enable large-scale diabetes awareness programs (30).

Overall, our study has several strengths and certain inherent limitations. First, this study represents the largest hitherto reporting on the performance and potential clinical utility of a questionnaire-based risk stratification model for prevalent and incident T2D in two biobanks and across multiple ethnicities. From a modeling perspective, this minimizes the chances of overfitting and provides evidence of the model's validity. Second, we applied strict inclusion and exclusion criteria, thereby minimizing the risk of including individuals with undiagnosed T2D. Third, we validated two widely non-laboratory clinical tools, FINDRISC and AUSDRISK, in all ethnic groups of the UKB and externally in LL, which provides a comprehensive benchmark for the performance of our models. On the other hand, as with all self-reported biobank data, ethnicity data may only be partially accurate. Specifically, self-reported ethnic background can be influenced by individual perceptions, cultural and social factors, and may not always accurately reflect an individual's ancestry and levels of admixture. Additionally, the categories used to describe ethnicity can differ between countries, making it difficult to compare results across studies. Lastly, due to the observational nature of this study, we cannot identify causal relationships between the features included in the models and the predicted outcomes.

In conclusion, questionnaire-based ML models predict prevalent and incident T2D in multiple ethnicities with high accuracy and have the potential to enhance early diagnosis if deployed for population health screening in primary diabetes care. While biomarker-based models achieved enhanced performance, the questionnaire-only models produced significantly high and clinically useful predictions to be considered a valid alternative to these models and the challenges their large-scale deployment can pose. This is particularly important for populations of non-white ethnicity who are disproportionately impacted by T2D and for regions with limited resources and access to primary diabetes care.

Conflict of interest

MK, NS, ST, OC, YI, and RHH have no conflict of interest to declare. PF, SvD, JCF, and DdV are employed by Ancora Health B.V. All employees own shares of Ancora Health B.V. BHRW

sits on the medical advisory board of Ancora Health B.V. CSM has been a shareholder of and reports grants through his institution and personal consulting fees from Coherus Inc., AltrixBio, grants through his institution from Merck, and grants through his institution personal consulting fees from Novo Nordisk, reports personal consulting fees and support with research reagents from Ansh Inc., reports personal consulting fees from Genfit, Lumos, Amgen, Corcept, Intercept, 89Bio, AstraZeneca and Regeneron, reports support (educational activity meals at and through his institution) from Amarin, Novo Nordisk and travel support and fees from TMIOA, Elsevier, the California Walnut Commission, College Internationale Research Servier, and the Cardio Metabolic Health Conference; none of which is related to the work presented herein.

Acknowledgments

All authors critically reviewed the manuscript for important intellectual content. The authors thank the UKB data access granted through application 55495 and data access to the Lifelines data through application OV20_00020.

Author Contributions

MK, JCF, DdV, and BHRW conceived and designed the study. MK was the lead author, accessed the data, interpreted the analyses, and wrote the manuscript. PF conducted data cleaning and the statistical analyses. MK, SvD, JCF, and DdV checked the statistical analyses. PF, SvD, JCF, and DdV contributed to drafting the manuscript. DdV and BHRW worked in supervisory capacities. All other co-authors read the manuscript and provided constructive feedback. The lead author MK has full access to all the data in the study and had final responsibility for the decision to submit for publication.

References

1. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 2004;27(5):1047-53.
2. GBD. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396(10258):1204-22.
3. Gujral UP, Narayan KMV. Diabetes in Normal-Weight Individuals: High Susceptibility in Nonwhite Populations. Diabetes Care. 2019;42(12):2164-6.
4. Goff LM, Ladwa M, Hakim O, Bello O. Ethnic distinctions in the pathophysiology of type 2 diabetes: a focus on black African-Caribbean populations. Proceedings of the Nutrition Society. 2020;79(2):184-93.
5. Banerjee AT, Shah BR. Differences in prevalence of diabetes among immigrants to Canada from South Asian countries. Diabet Med. 2018;35(7):937-43.
6. Paul SK, Owusu Adjah ES, Samanta M, Patel K, Bellary S, Hanif W, et al. Comparison of body mass index at diagnosis of diabetes in a multi-ethnic population: A case-control study with matched non-diabetic controls. Diabetes Obes Metab. 2017;19(7):1014-23.
7. Meo SA, Usmani AM, Qalbani E. Prevalence of type 2 diabetes in the Arab world: impact of GDP and energy consumption. Eur Rev Med Pharmacol Sci. 2017;21(6):1303-12.
8. Sulaiman N, Mahmoud I, Hussein A, Elbadawi S, Abusnana S, Zimmet P, et al. Diabetes risk score in the United Arab Emirates: a screening tool for the early detection of type 2 diabetes mellitus. BMJ Open Diabetes Res Care. 2018;6(1):e000489.
9. Davies MJ, Aroda VR, Collins BS, Gabbay RA, Green J, Maruthur NM, et al. Management of Hyperglycemia in Type 2 Diabetes, 2022. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care. 2022;45(11):2753-86.
10. Ng SW, Zaghloul S, Ali HI, Harrison G, Popkin BM. The prevalence and trends of overweight, obesity and nutrition-related non-communicable diseases in the Arabian Gulf States. Obes Rev. 2011;12(1):1-13.
11. Lindstrom J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26(3):725-31.
12. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010;192(4):197-202.
13. Rigla M, Garcia-Saez G, Pons B, Hernando ME. Artificial Intelligence Methodologies and Their Application to Diabetes. J Diabetes Sci Technol. 2018;12(2):303-10.
14. Morgenstern JD, Buajitti E, O'Neill M, Piggott T, Goel V, Fridman D, et al. Predicting population health with machine learning: a scoping review. BMJ Open. 2020;10(10):e037860.
15. Collins R. What makes UK Biobank special? Lancet. 2012;379(9822):1173-4.
16. Klijs B, Scholtens S, Mandemakers JJ, Snieder H, Stolk RP, Smidt N. Representativeness of the LifeLines Cohort Study. PLoS One. 2015;10(9):e0137203.
17. Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage. 2009;45(1 Suppl):S199-209.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.
19. Ramos-Louro P, Arellano Perez Vertti RD, Reyes AL, Martinez-Nava GA, Espinosa R, Pineda C, et al. mtDNA haplogroup A enhances the effect of obesity on the risk of knee OA in a Mexican population. Sci Rep. 2022;12(1):5173.
20. Jr FEH. Hmisc: Harrell Miscellaneous. 2023.

21. Lacy ME, Wellenius GA, Carnethon MR, Loucks EB, Carson AP, Luo X, et al. Racial Differences in the Performance of Existing Risk Prediction Models for Incident Type 2 Diabetes: The CARDIA Study. Diabetes Care. 2016;39(2):285-91.

22. Wilkinson L, Yi N, Mehta T, Judd S, Garvey WT. Development and validation of a model for predicting incident type 2 diabetes using quantitative clinical data and a Bayesian logistic model: A nationwide cohort and modeling study. PLoS Med. 2020;17(8):e1003232.

23. Glumer C, Vistisen D, Borch-Johnsen K, Colagiuri S, Collaboration D-. Risk scores for type 2 diabetes can be applied in some populations but not all. Diabetes Care. 2006;29(2):410-4.

24. Dugee O, Janchiv O, Jousilahti P, Sakhiya A, Palam E, Nuorti JP, et al. Adapting existing diabetes risk scores for an Asian population: a risk score for detecting undiagnosed diabetes in the Mongolian population. BMC Public Health. 2015;15:938.

25. Rokhman MR, Arifin B, Zulkarnain Z, Satibi S, Perwitasari DA, Boersma C, et al. Translation and performance of the Finnish Diabetes Risk Score for detecting undiagnosed diabetes and dysglycaemia in the Indonesian population. PLoS One. 2022;17(7):e0269853.

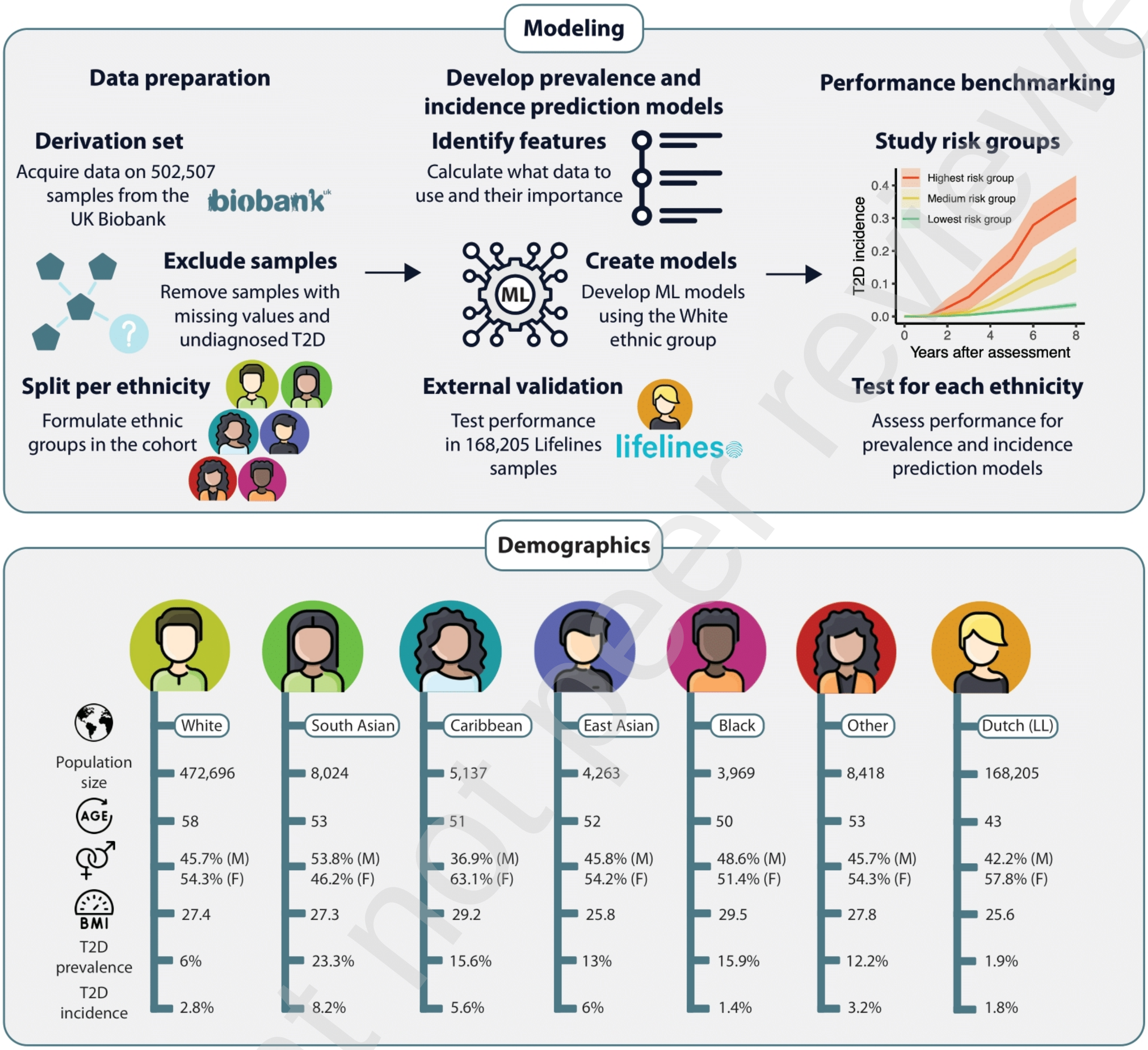
26. Wu J, Hou X, Chen L, Chen P, Wei L, Jiang F, et al. Development and validation of a non-invasive assessment tool for screening prevalent undiagnosed diabetes in middle-aged and elderly Chinese. Prev Med. 2019;119:145-52.

27. Smith GD, Spiegelhalter D. Shielding from covid-19 should be stratified by risk. BMJ. 2020;369:m2063.

28. Zhou X, Siegel KR, Ng BP, Jawanda S, Proia KK, Zhang X, et al. Cost-effectiveness of Diabetes Prevention Interventions Targeting High-risk Individuals and Whole Populations: A Systematic Review. Diabetes Care. 2020;43(7):1593-616.

29. Mantena S, Celi LA, Keshavjee S, Beratarrechea A. Improving community health-care screenings with smartphone-based AI technologies. Lancet Digit Health. 2021;3(5):e280-e2.

30. Polyzos SA, Mantzoros CS. Diabetes mellitus: 100 years since the discovery of insulin. Metabolism. 2021;118:154737.



Modeling

Data preparation

Derivation set

Acquire data on 502,507 samples from the UK Biobank

Exclude samples

Remove samples with missing values and undiagnosed T2D

Split per ethnicity

Formulate ethnic groups in the cohort

Develop prevalence and incidence prediction models

Identify features

Calculate what data to use and their importance

Create models

Develop ML models using the White ethnic group

External validation

Test performance in 168,205 Lifelines samples

Performance benchmarking

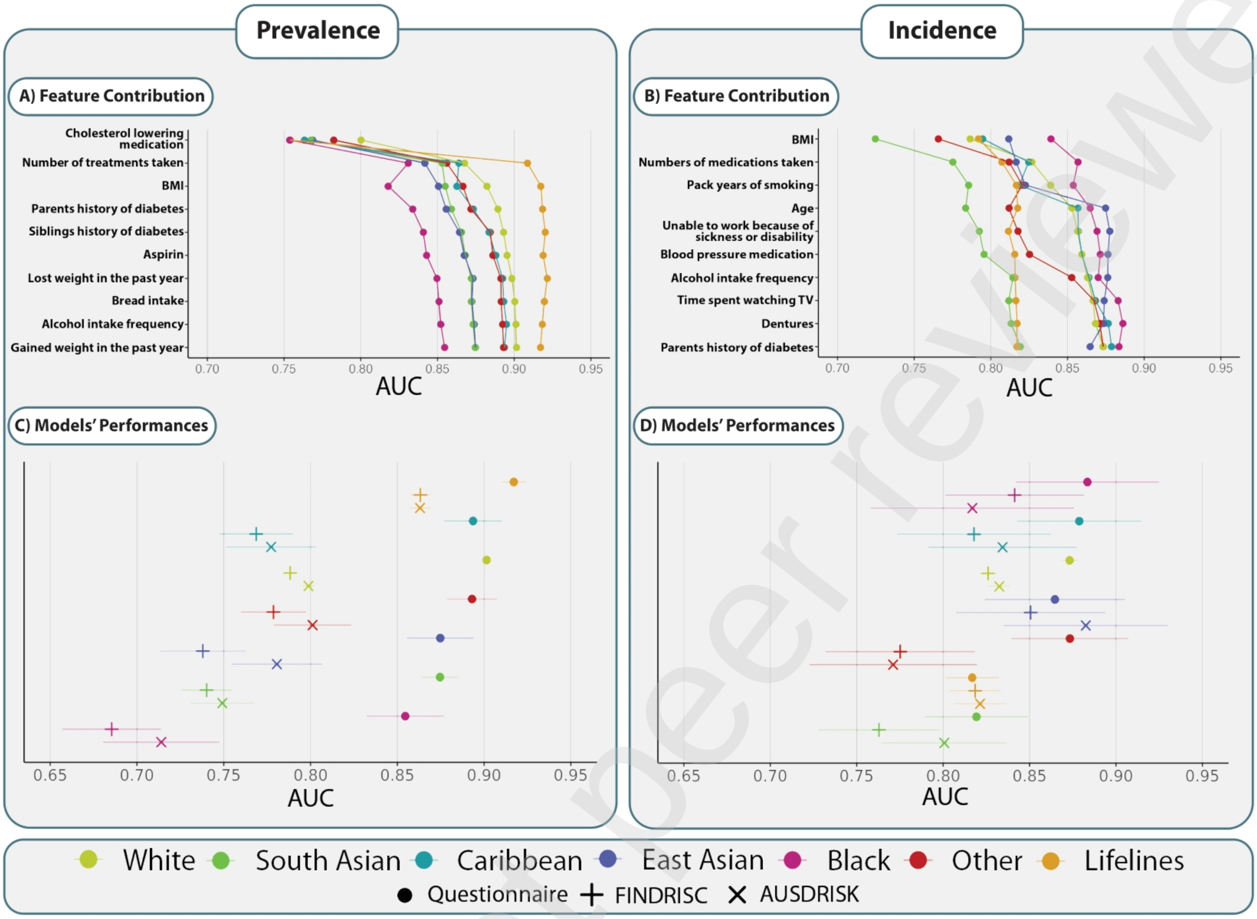
Study risk groups

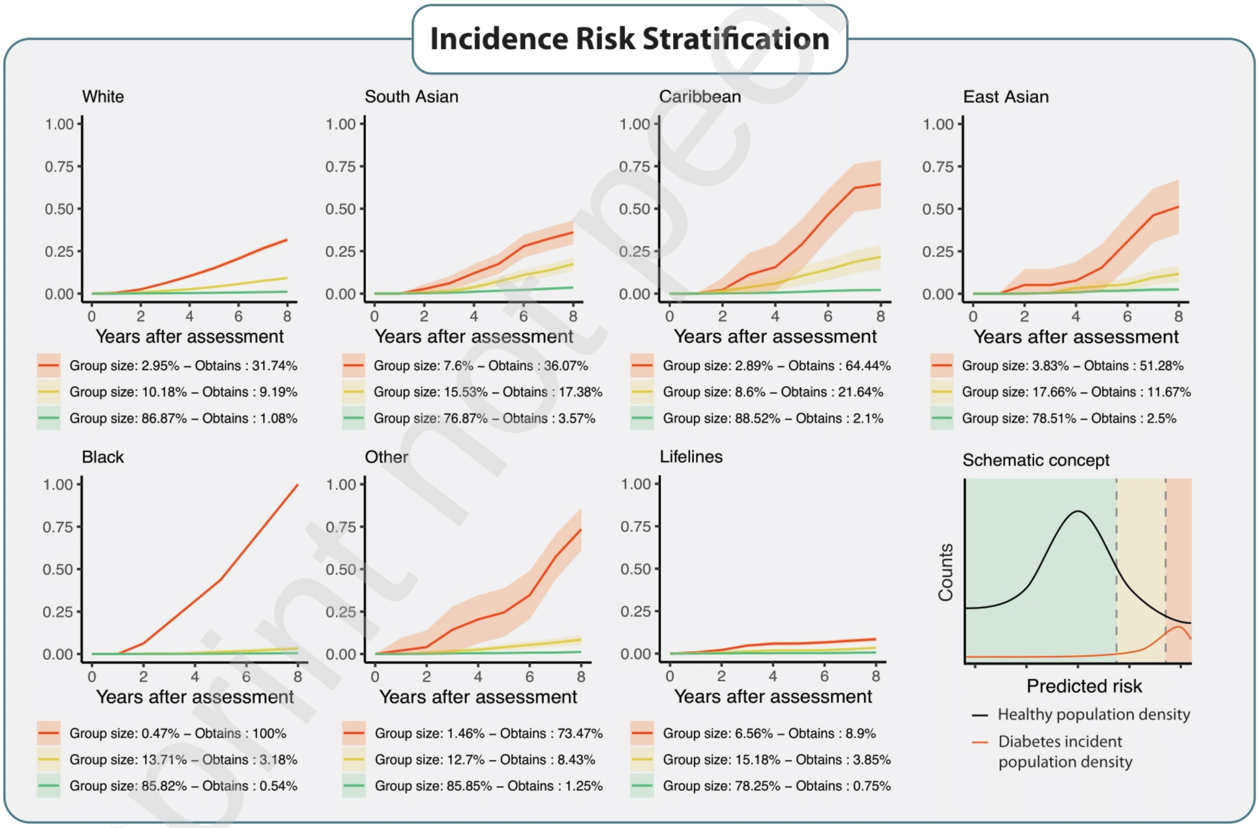
Test for each ethnicity

Assess performance for prevalence and incidence prediction models

Demographics

	White	South Asian	Caribbean	East Asian	Black	Other	Dutch (LL)
Population size	472,696	8,024	5,137	4,263	3,969	8,418	168,205
AGE	58	53	51	52	50	53	43
⚥	45.7% (M) 54.3% (F)	53.8% (M) 46.2% (F)	36.9% (M) 63.1% (F)	45.8% (M) 54.2% (F)	48.6% (M) 51.4% (F)	45.7% (M) 54.3% (F)	42.2% (M) 57.8% (F)
BMI	27.4	27.3	29.2	25.8	29.5	27.8	25.6
T2D prevalence	6%	23.3%	15.6%	13%	15.9%	12.2%	1.9%
T2D incidence	2.8%	8.2%	5.6%	6%	1.4%	3.2%	1.8%


Prevalence
Incidence
A) Feature Contribution
Cholesterol lowering medication
Number of treatments taken
BMI
Parents history of diabetes
Siblings history of diabetes
Aspirin
Lost weight in the past year
Bread intake
Alcohol intake frequency
Gained weight in the past year
0.70
0.75
0.80
0.85
0.90
0.95
AUC
B) Feature Contribution
BMI
Numbers of medications taken
Pack years of smoking
Age
Unable to work because of sickness or disability
Blood pressure medication
Alcohol intake frequency
Time spent watching TV
Dentures
Parents history of diabetes
0.70
0.75
0.80
0.85
0.90
0.95
AUC
C) Models' Performances
0.65
0.70
0.75
0.80
0.85
0.90
0.95
AUC
D) Models' Performances
0.65
0.70
0.75
0.80
0.85
0.90
0.95
AUC
● White ● South Asian ● Caribbean ● East Asian ● Black ● Other ● Lifelines
● Questionnaire + FINDRISC × AUSDRISK



Incidence Risk Stratification
White
1.00
0.75
0.50
0.25
0.00
0 2 4 6 8
Years after assessment
Group size: 2.95% – Obtains : 31.74%
Group size: 10.18% – Obtains : 9.19%
Group size: 86.87% – Obtains : 1.08%
South Asian
Years after assessment
Group size: 7.6% – Obtains : 36.07%
Group size: 15.53% – Obtains : 17.38%
Group size: 76.87% – Obtains : 3.57%
Caribbean
Years after assessment
Group size: 2.89% – Obtains : 64.44%
Group size: 8.6% – Obtains : 21.64%
Group size: 88.52% – Obtains : 2.1%
East Asian
Years after assessment
Group size: 3.83% – Obtains : 51.28%
Group size: 17.66% – Obtains : 11.67%
Group size: 78.51% – Obtains : 2.5%
Black
Years after assessment
Group size: 0.47% – Obtains : 100%
Group size: 13.71% – Obtains : 3.18%
Group size: 85.82% – Obtains : 0.54%
Other
Years after assessment
Group size: 1.46% – Obtains : 73.47%
Group size: 12.7% – Obtains : 8.43%
Group size: 85.85% – Obtains : 1.25%
Lifelines
Years after assessment
Group size: 6.56% – Obtains : 8.9%
Group size: 15.18% – Obtains : 3.85%
Group size: 78.25% – Obtains : 0.75%
Schematic concept
Counts
Predicted risk
Healthy population density
Diabetes incident population density