medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Prediction of type 2 diabetes mellitus onset using logistic regression-based scoreboards

Yochai Edlitz[1,2], Eran Segal[1,2, †]

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

[2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.

[†] Corresponding author

Abstract

Type 2 diabetes mellitus (T2DM) accounts for ~90% of all cases of diabetes which are estimated with an annual world death rate of 1.6 million in 2016. Early detection of T2D high-risk patients can reduce the incidence of the disease through a change in lifestyle, diet, or medication. Since populations of lower socio-demographic status are more susceptible to T2D and might have limited resources for laboratory testing, there is a need for accurate yet accessible prediction models based on non-laboratory parameters. This paper introduces one non-laboratory model which is highly accessible to the general population and one highly precise yet simple laboratory model. Both models are provided as an accessible scoreboard form and also as a logistic regression model. We based the models on data from 44,879 non-diabetic, UK Biobank participants, aged 40-65, predicting the risk of T2D onset within the next 7.3 years (SD 2.3). The non-laboratory prediction model for T2DM onset probability incorporated the following covariates: sex, age, weight, height, waist, hips-circumferences, waist-to-hip Ratio (WHR) and Body-Mass Index (BMI). This logistic regression model achieved an Area Under the Receiver Operating Curve (auROC) of 0.82 (0.79-0.85 95% CI) and an odds ratio (OR) between the upper and lower prevalence deciles of x77 (28-98). We further analysed the contribution of laboratory-based parameters and devised a blood-test model based on just five blood tests. In this model, we included age, sex, Glycated Hemoglobin (HbA1c%), reticulocyte count, Gamma Glutamyl-Transferase, Triglycerides, and HDL cholesterol to predict T2D onset. This logistic-regression model achieved an auROC of 0.89 (0.86-0.91) and a deciles' OR of x87 (27-152). Using the scoreboard results, the Anthropometrics model classified three risk groups, a group with 1%(1-2%); a group with 9% (7-11%) probability, and a group with a 15% (7-23%) risk of developing T2D. The Five blood tests scoreboard model, further classified into four risk groups: 0.9% (0.7%-1%); 8%(6-11%); 18%(14-22%) and a high risk group of 38%(23-54%) of developing T2D. We analysed several more comprehensive models which included genotyping data and other environmental factors and found that it did not provide cost efficient benefits over the five blood tests model. The Five blood tests and anthropometric models, both in their logistic regression form and scoreboard form, outperform the commonly used non-laboratory models, the Finnish Diabetes Risk Score (FINDRISC) and the German Diabetes Risk Score (GDRS). When trained using our data, the FINDRISC achieved an auROC of 0.75 (0.71-0.78), and the GDRS auROC resulted in 0.58 (0.54-0.62), respectively.

1.Introduction

Diabetes mellitus is defined as a group of diseases characterised by symptoms of chronic hyperglycemia. It is becoming one of the world's most challenging epidemics. The prevalence of T2D has increased from 4.7% in 1980 to 8.5% in 2014. An estimated 1.6 million deaths were directly caused by diabetes in 2016. T2D is generally characterised by insulin resistance, resulting in hyperglycemia, and it accounts for ~90% of all diabetes cases [1,2].

In recent years, the prevalence of diabetes has been rising more rapidly in low and middle-income countries (LMICs) than in high-income countries[3]. In 2014 Beagley et al. estimated that 45.8% or 174.8 million of all diabetes cases in

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

adults are undiagnosed. 83.8% of all cases of undiagnosed diabetes are in LMICs [4], where laboratory diagnosis testing is limited for some of the populations in these countries[5].

According to several studies, a healthy diet, regular physical activity, maintaining normal body weight and avoiding tobacco use can prevent or delay T2D onset [3,6,7,8,9]. A screening tool that can identify individuals at risk will enable a lifestyle or medication intervention. Ideally, such a screening tool should be accurate, simple and low cost. It should also be easily available, allowing populations who have difficulty accessing laboratories to be screened by other means.

Several such tools are in use today [10,11,12]. The Finnish Diabetes Risk Score (FINDRISC), a commonly used, non-invasive T2D risk-score model, estimates patients aged between 35 and 64 developing T2D within the next ten years. FINDRISC was created based on a prospective cohort of 4,746 and 4,615 individuals in Finland in 1987 and 1992. The FINDRISC model uses gender, age, Body Mass Index (BMI), blood pressure medications, a history of high blood glucose, physical activity, daily consumption of fruits, berries, or vegetables and family history of Diabetes as the parameters for the model. The FINDRISC might be used as a scoreboard model or a logistic regression (LR) model [13,14,15].

Another commonly used prediction model is the German Diabetes Risk Score (GDRS), which estimates a five-year risk for developing T2D. The GDRS is based on 9,729 men and 15,438 women aged 35-65 years from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study [16]. The GDRS is a Cox regression model based on age, height, waist circumference, the prevalence of hypertension (yes/no), smoking behaviour, physical activity, moderate alcohol consumption, coffee consumption, intake of whole-grain bread, intake of red meat, and parent and sibling history of T2D [17,18]. This model, too, can be assigned as an accessible scoreboard model.

The objective of the present research was to develop clinically usable models which are easy to use and highly predictive of T2D onset. We developed two simple models and compared their predictive power to the highly esteemed FINDRISC and GDRS as our baseline. We trained all models on a training data set and tested them on the holdout data set, taken from the UK Biobank (UKB) observational study cohort. We based one of the models on easily accessible anthropometric measures and the other on an invasive laboratory test using only five blood samples.As our models were trained and evaluated using the UKB database, they are most relevant for the U.K. population aged 40-65. Still, they can also be used for people similar to our research cohort (as presented in Table 1) and might be adapted to additional populations. Both models are given both in their logistic regression form and as accessible scoreboards.

2.Results

We analysed the data of 20,346 participants from the UK Biobank's (UKB) cohort who revisited the UKB assessment centers during 2012-2013 and 48,705 participants who revisited the centers from 2014 onwards (see Figure 1 and Methods). During the screening process of our cohort, we kept the data of the participants who returned for a second or third visit, tested negative for T2D and were not treated for T2D. The final cohort sample included data of 44,873 participants, of whom 2.16% developed T2D during a follow-up period of 7.3±2.3 years (see Table 1, Figure 1A and Methods).

Before training the models, we partitioned our data into training, validation, and holdout test sets to avoid overfitting. The training dataset consisted of 25,122 participants, and the validation dataset included 10,757 participants. We explored the training and validation datasets to select the optimal features for our models. We used the holdout test set, which included 8,994 participants, to report the final models' results (see Figure 1S and Methods).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2.1 Anthropometric based model

To provide an accessible, simple, non-laboratory and non-invasive T2D prediction model, we built anthropometric based ascoreboard model where a patient can easily mark its result in each of the scoreboard questions, consisting of the following eight parameters: age, sex, weight, height, hip circumference, waist circumference, body mass index (BMI) and the waist-to-hip ratio (WHR) (Figure 2A). The patient then sums up its final score which relates to one of three risk groups first group score range [1-70] has a 1% [1-2% 95% CI] of developing T2D; Second group, score range 71-83 predicts a 9%[7-11% 95%CI] of developing T2D; Third group 84-92 15% [7-23% 95%CI] of developing T2D (Figure 2C).

We also provide a similar model in its logistic regression form for more accurate computer aided results. Testing this model using the holdout test set, the logistic regression form of this model achieved an area under the receiver operating curve (auROC) of 0.82 (0.79-0.85) and an average precision score (APS) of 0.12 (0.09-0.16) at 95% CI). Using the model in its scoreboard form, we achieved an auROC of 0.81 (0.78-0.84) and an APS of 0.09 (0.07-0.12). Both model's forms outperformed the two models which we used as a reference: the FINDRISC model, which has an auROC of 0.75 (0.71-0.78) and an APS of 0.07 (0.05-0.09), and the GDRS model, which has an auROC of 0.58 (0.54-0.62) and APS of 0.03 (0.02-0.04), see Figure 3A-B and Methods. With the cohort's baseline prevalence of 2.17%, the LR anthropometric model achieved deciles' OR of x77 (27.7-98.1), and its scoreboard form achieved deciles OR of x61 (17.7-101) compared to the FINDRISC x23 (6.80-70.4) and the x4.1 (1.75-9.24), see Figure 3C and Table 2.

Analysing the models' feature importance, the WHR and the BMI have the highest predictability in the anthropometric model due to their highest logs-odds-ratio (β) (Figure 3D). These two body habitus measures are commonly mentioned in the literature as indicators associated with chronic illness [19,20,21,22].

2.2five blood tests based model

In addition, for those cases where laboratory testing will be available, we developed a more accurate tool for predicting T2D onset. This tool uses five blood test scores as an input to a logistic regression model which we also simplified to a scoreboard model (Figure 2B,D). Using the Five blood tests scoreboard (Figure 2B), we bin the resulted scores into four groups: first group score range [1-114] has a 0.9% [0.7-1%] 95% CI probability of developing T2D; Second group, score range 115-125 predicts an 8% [6-11%] 95% CI probability of developing T2D; Third group score range 126-145 18% [14-22%] 95% CI of developing T2D; The fourth group score range is 146-156, participants in these score range has 35% [24-46%] 95% CI of developing T2D.

To derive this model, we started by a feature selection process from a full-feature GBDT model, using only the training and validation data sets. We clustered the features of this model into 13 categories such as lifestyle, diet, and anthropometrics. Based on this process, we concluded that the blood tests have higher predictability than the other clusters assessed. We thus trained a full blood test model using 59 blood tests that are available in the training dataset. Applying a recursive feature elimination process on the top 10 predictive features, we established the features of our final model, which is based on five blood tests (see Methods).

Using the five blood tests logistic regression model we achieved the following results for the test set: an auROC of 0.89 (0.86-0.91), an APS of 0.26 (0.2-0.33), and a deciles' OR of x87 (27-152). When using the scoreboard model. we achieve an auROC of 0.88 (0.85-0.9), an APS of 0.18 (0.14-0.23), and a deciles' OR of x78 (23-139) (Figure 3 A-C, Table 2). The five blood tests model results are superior to those of our non-laboratory anthropometric model, as well as those of the highly-esteemed FINDRISC and GDRS models (Figure 3 A-C, Table 2).

We then compared these results to those of a 59 blood tests input features of logistic regression model and to those of a GBDT model including 13 feature clusters, which consisted of 279 individual features available in the dataset and

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

genetics data. These two models achieved an auROC of 0.91 (0.88-0.93) and 0.92 (0.9-0.94); an APS of 0.32 (0.25-0.39) and 0.34 (0.28-0.42); and a deciles' OR of x117 (37-163) and x133(45-167), respectively.

The five blood tests that we used are the following: the Glycated Haemoglobin test(HbA1c%), which measures the average blood sugar for the past two to three months and which is one of the means to diagnose Diabetes; the Reticulocyte Count; the Gamma-Glutamyl Transferase Test (GGT); the HDL Cholesterol Test, and the Triglycerides Test. We also included the time to prediction (time between visits); gender; age at the repeated visit; and a bias term which is related to the prevalence in the population. We computed the values of these features' associated coefficients with their 95%CI to enable a reconstruction of the models (Figure 3E).

As expected, the HbA1c% feature had the highest predictive power since it is one of the criteria for T2D diagnosis. The second-highest predictive feature was the high-light-scatter-reticulocytes-count, which reflects the number of new red blood cells in the body[23]. HDL cholesterol, which is known to be beneficial for health, especially in the context of cardiovascular diseases and T2D [15,24,25], was found to be inversely correlated to the predicted probability of T2D onset. Interestingly, age and sex had a very low OR value, meaning that they hardly contributed to the model, probably because the T2D relevant information of these features latent within the blood-tests' data. Using the five blood tests scoreboard model, we removed the use of age, as we found that it did not contribute to the final score of the model.

Table 1 Cohort statistical data

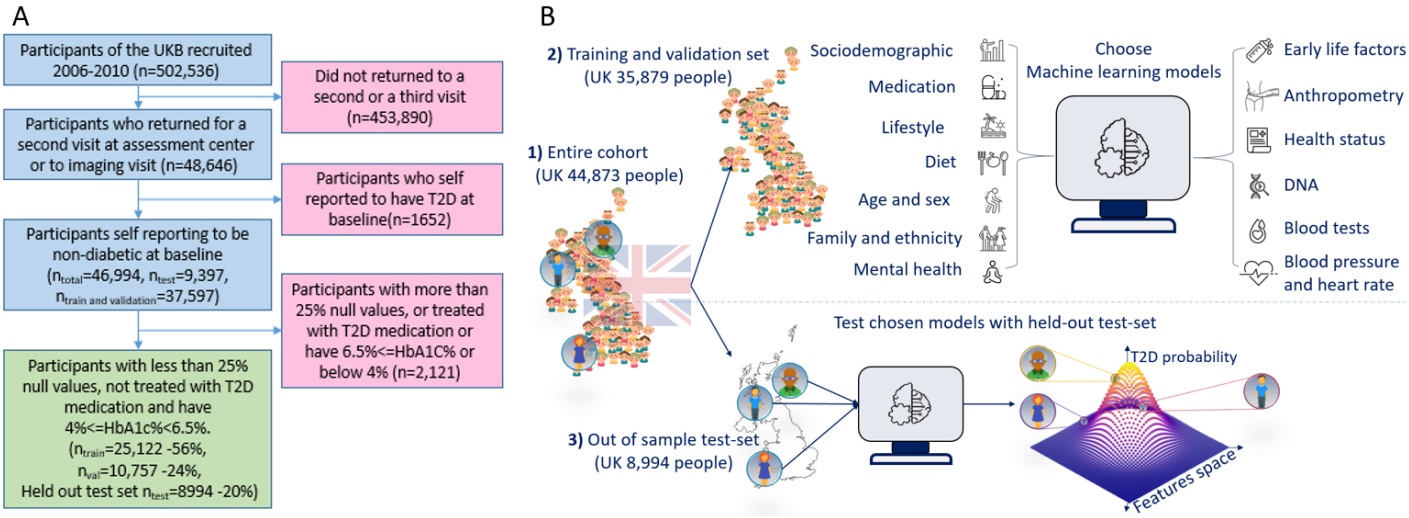
	UKB population	Train, validation and test sets	Train set	Validation set	Test set
Number of participants	502536	44873	25122	10757	8994
The time between visits [Years]	-	7.3±2.3	7.3±2.3	7.3±2.3	7.3±2.3
Males in the population [%]	45.6	47.9	48	47.6	48
Diabetic at first visit [%]	4.8	0	0	0	0
Diabetic at last visit [%]	-	2.16	2.13	2.22	2.17
%Hba1c at first visit [%]	5.5±0.6	5.3±0.3	5.3±0.3	5.3±0.3	5.3±0.3
Hba1c at last visit [%]	-	5.4±0.4	5.4±0.4	5.4±0.4	5.4±0.4
Weight at first visit [Kg]	78.1±15.9	76.7±14.8	76.7±14.7	76.9±15.0	76.5±14.7
Weight at last visit [Kg]	-	76.5±15.2	76.5±15.2	76.8±15.4	76.3±15.0
Hips circumference at first visit [cm]	103.4±9.2	102.1±8.2	102.2±8.2	102.3±8.3	101.9±8.1
Hips circumference at last visit [cm]	-	102.06±8.9	101.6±8.9	102.3±9.0	101.9±8.7
Waist circumference at first visit [cm]	90.3±13.5	88.0±12.5	88.0±12.5	88.2±12.7	87.8±12.5
Waist circumference at last visit [cm]	-	89.4±12.7	89.3±12.7	89.6±12.9	89.2±12.5
Height at first visit [cm]	168.4±9.3	169.5±9.2	169.5±9.2	169.4±9.1	169.5±9.1
Height at last visit [cm]	-	169.0±9.2	169.0±9.3	168.9±9.2	169.0±9.2
Age at first visit [Years]	56.5±8.1	55.6±7.5	55.6±7.6	55.6±7.6	55.6±7.5
Age at last visit [Years]	-	62.9±7.5	62.9±7.5	62.9±7.5	62.9±7.4
BMI at first visit [Kg/m^2]	27.4±4.8	26.6±4.2	26.6±4.2	26.7±4.3	26.5±4.1
BMI at last visit [Kg/m^2]	-	26.7±4.4	26.7±4.5	26.8±4.5	26.6±4.3

Characteristics of the cohort's population and the UKB population. A "±" sign denotes the standard deviation. While T2D prevalence in the UKB participants is 4.8%, in our cohort it was 2.16% because we screened the cohort at baseline for HbA1c% levels < 6.5%, resulting in a lower rate of T2D. The age range of the participants at the first visit was 40-73; and thus, our models are not customised for people who might develop T2D at younger ages. The models predict the

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

risk of developing T2D during the time between the first visit to the UKB assessment center and the last visit. This is referred to as " the time between visits."

Figure 1 A flow chart of the cohort selection process and an illustrative figure of the model's extraction



A. A flowchart which demonstrates the selection process of participants in this study. Participants who came for a repeated second or third visit were selected from the 502,536 participants of the UKB. Next, we excluded 1,652 participants who self-reported having T2D. We then split the data into 80% of the training and validation set, and 20% holdout test set. We excluded an additional 2,115 participants due to (1) having 25% or more missing values from the full feature list, (2) having HbA1c levels above or equal to 6.5%, or (3) being treated with Metformin or Insulin. Finally, the training set included 25,122 participants (56% of the cohort), the validation set included 10,757 (24% of the cohort), and the test set included a total of 8,994 participants (20% of the cohort). **B.** Process flow during training and testing of the models. We first split the data and kept a holdout test set. We later explored several models using the training and validation data sets. In the final stage, we compared the selected models using the holdout test set and reported the results. The output of the models is calibrated to predict the probability of a participant to develop T2D.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2|Anthropometrics and Blood tests scoreboards

A

1)Sex

Sex : Female	0
Sex : Male	3

2)Age [Years]

Age at last visit <= 57	0
Age at last visit > 57 and <= 69	1
Age at last visit > 69	2

3) Years for prediction [Years]

Years for prediction<= 5	0
Years for prediction> 5 and <= 6	6
Years for prediction> 6 and <= 8	8
Years for prediction> 8 and <= 9	9
Years for prediction> 9 and <= 10	12
Years for prediction> 10	16

4)BMI [kg/m^2]

Body mass index (BMI) <= 22.5	0
Body mass index (BMI) > 22.5 and <= 24.1	7
Body mass index (BMI) > 24.1 and <= 25.4	8
Body mass index (BMI) > 25.4 and <= 26.8	15
Body mass index (BMI) > 26.8 and <= 28.4	19
Body mass index (BMI) > 28.4 and <= 30.7	21
Body mass index (BMI) > 30.7	29

5)Weight [Kg]

Weight <= 75.4	0
Weight > 75.4	1

6)Height [cm]

Standing height <= 165	7
Standing height > 165 and <= 174	2
Standing height > 174	0

7)Waist-hips ratio

Waist-hips ratio <= 0.76	0
Waist-hips ratio > 0.76 and <= 0.8	3
Waist-hips ratio > 0.8 and <= 0.84	8
Waist-hips ratio > 0.84 and <= 0.88	12
Waist-hips ratio > 0.88 and <= 0.92	14
Waist-hips ratio > 0.92 and <= 0.95	17
Waist-hips ratio > 0.95	24

8)Waist circumference [cm]

Waist circumference <= 75	0
Waist circumference > 75 and <= 82	2
Waist circumference > 82 and <= 88	6
Waist circumference > 88 and <= 93	7
Waist circumference > 93 and <= 100	10
Waist circumference > 100	14

9)Hip circumference [cm]

Hip circumference <= 97	2
Hip circumference > 97 and <= 101	1
Hip circumference >	0

B

1)Sex

Sex : Female	0
Sex : Male	4

2)Years for prediction

Years for prediction<= 5	0
Years for prediction> 5 and <= 6	8
Years for prediction> 6 and <= 8	11
Years for prediction> 8 and <= 9	12
Years for prediction> 9 and <= 10	16
Years for prediction> 10	21

3) Glycated hemoglobin-HbA1c[%]

HbA1c<= 30.7	0
HbA1c> 30.7 and <= 32.4	11
HbA1c> 32.4 and <= 33.6	16
HbA1c> 33.6 and <= 34.6	18
HbA1c> 34.6 and <= 35.7	27
HbA1c> 35.7 and <= 36.9	32
HbA1c> 36.9 and <= 38.5	39
HbA1c> 38.5	58

4)HDL [mmol/l]

HDL cholesterol <= 1.1	27
HDL cholesterol > 1.1 and <= 1.3	19
HDL cholesterol > 1.3 and <= 1.4	18
HDL cholesterol > 1.4 and <= 1.5	14
HDL cholesterol > 1.5 and <= 1.6	12
HDL cholesterol > 1.6 and <= 1.9	9
HDL cholesterol > 1.9	0

5) Gamma glutamyltransferase [U/L]

GGT <= 14.8	0
GGT> 14.8 and <= 18	7
GGT> 18 and <= 21.4	11
GGT> 21.4 and <= 25.3	12
GGT> 25.3 and <= 30.4	16
GGT> 30.4 and <= 37.7	20
GGT> 37.7 and <= 53.3	22
GGT> 53.3	24

6)Reticulocyte count [10^{12} cells/Liter]

Reticulocyte count <= 0.035	0
Reticulocyte count > 0.035 and <= 0.043	5
Reticulocyte count > 0.043 and <= 0.051	7
Reticulocyte count > 0.051 and <= 0.058	11
Reticulocyte count > 0.058 and <= 0.067	14
Reticulocyte count > 0.067 and <= 0.081	17
Reticulocyte count > 0.081	22

7)Triglycerides [mmol/L]

Triglycerides <= 0.89	0
Triglycerides > 0.89 and <= 1.2	3
Triglycerides > 1.2 and <= 1.4	4
Triglycerides > 1.4 and <= 1.8	6
Triglycerides > 1.8 and <= 2.4	7
Triglycerides > 2.4	9

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

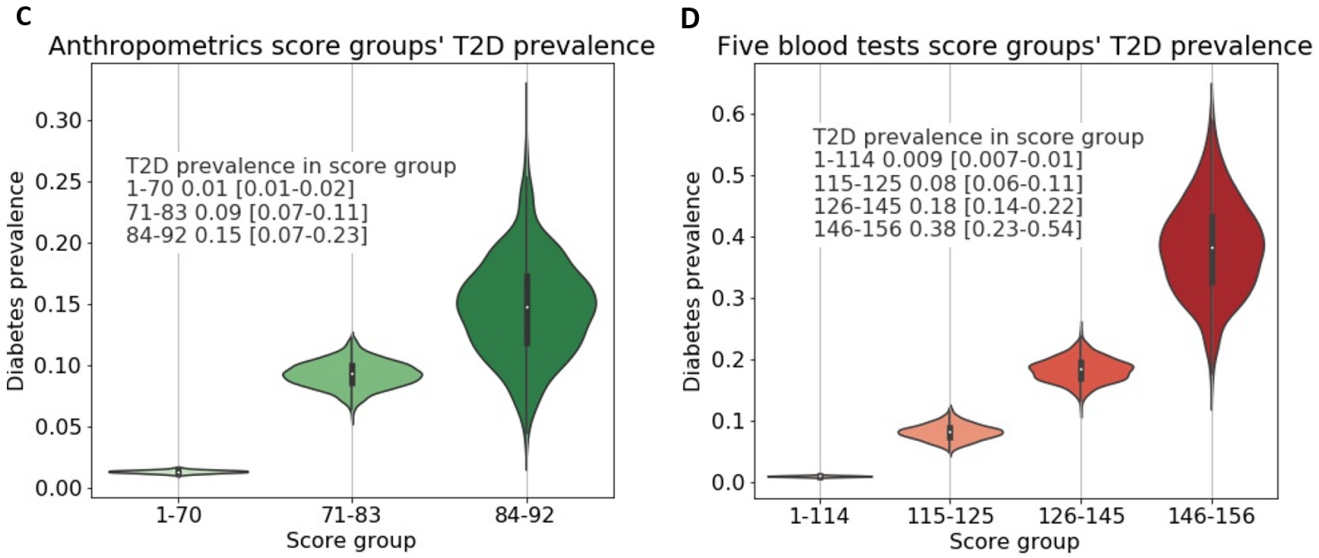


Figure 2 **Anthropometrics and Blood tests scoreboards** A)Anthropometrics based scoreboard. Scoreboard, summing the scores of the various features provides a final score that is quantified into one of three risk groups. B)Five blood tests scoreboard. Summing the scores of the various features provides a final score that is quantified into one of four risk groups (Figure 2D). C) Anthropometrics scoreboards risk groups - first group score range [1-70] 1% [1-2%]95%CI of developing T2D; Second group, score range 71-83 predicts a 9% [7-11%] 95%CI of developing T2D.; Third group 84-92 15% [7-23%]95%CI of developing T2D. D) Five blood tests scoreboards risk groups - first group score range [1-114] <1% [0.7-1%]95%CI of developing T2D; Second group, score range 115-125 predicts an 8% [6-11%]95%CI of developing T2D.; Third group 126-145 18% [14-22%] 95%CI of developing T2D. Fourth group 146-156 predicts 38% [23-54%] 95%CI of developing T2D.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3 |Main result

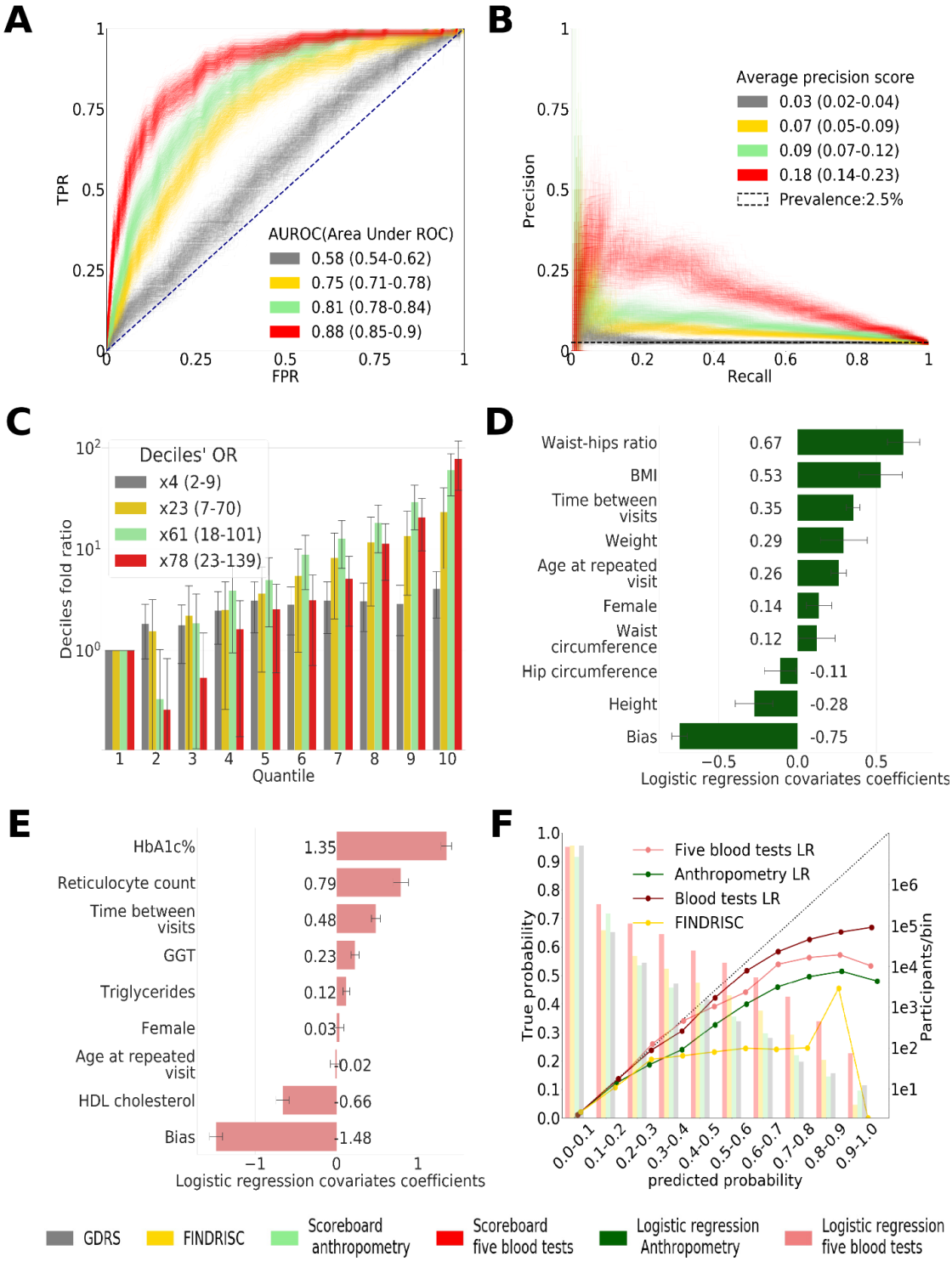


***Figure 3 Main results calculated using 1000 bootstraps of the cohort population**. Each point in the graphs represents a bootstrap iteration result. The color legend is shown at the bottom of the figure. **A.** ROC curves comparing the models developed in this research: a GBDT model of all features; logistic-regression models of five blood tests and the*

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

anthropometry based model compared to the well established GDRS and FINDRISC. ***B.*** *Precision-Recall (P-R) curves, showing the precision versus the recall for each model, with the prevalence of the population marked with the dashed line.* ***C.*** *Deciles' odds-ratio graph, the ratio of prevalence in each decile to the prevalence in the first decile. We bounded the prevalence in the first decile to be at least a tenth of the T2D prevalence in the full cohort.* ***D.*** *A feature importance graph of the logistic regression anthropometry model for a model with normalised features values. The bars indicate the standard deviation (SD) of the feature importance values. The top predictive features of this model are the body mass index (BMI) and waist to hip ratio (WHR).* ***E.*** *Feature importance graph of logistic regression Blood-tests model with SD bars. While the HbA1c% and Reticulocyte positively contribute to the T2D prediction, and HDL cholesterol lowers the T2D prediction probability, the information provided age and sex which is relevant for the prediction of T2D onset is overpowered by other features.* ***F.*** *A calibration plot of the anthropometry; five blood tests; full blood-test and the FINDRISC models. Calibration of the models' predictions allow reporting the probability of developing T2D (see Methods). The calibration was performed using an isotonic regression method.*

Table 2|Comparing models main results

Label	Model type	APS	AUROC	Deciles prevalence odds ratio
GDRS SA	Cox regression	0.03 (0.02-0.04)	0.58 (0.54-0.62)	4.1 (1.75-9.24)
FINDRISC LR	Scoreboard	0.07 (0.05-0.09)	0.75 (0.71-0.78)	23 (6.80-70.4)
Anthropometry	Scoreboard	0.09 (0.07-0.12)	0.81 (0.78-0.84)	61 (17.7-101)
Anthropometry	Logistic regression	0.12 (0.09-0.16)	0.82 (0.79-0.85)	77 (27.7-98.1)
Five blood tests	Scoreboard	0.18 (0.14-0.23)	0.88 (0.85-0.90)	78 (23.4-139)
Five blood tests	Logistic regression	0.26 (0.20-0.33)	0.89 (0.86-0.91)	87 (26.7-152)
Full blood tests	Logistic regression	0.32 (0.25-0.39)	0.91 (0.89-0.93)	117 (36.5-163)
All features DT	Gradient boosting decision trees	0.34 (0.28-0.42)	0.92 (0.90-0.94)	133 (45.1-167)

The values in parentheses indicate 95% CI. The deciles' OR is a measure of the ratio between T2D prevalence in the top risk score decile bin and the prevalence in the lowest decile bin. For the lowest decile, in case the actual prevalence in that bin was zero, we used a threshold of one diagnosed participant (see Methods).

2.3 Prediction within an HbA1c% stratified population

To verify that our models are capable of discriminating within a group of normoglycemic participants and within a group of pre-diabetic participants, we tested the models separately on each group extracted from our data. We separated the groups based on their HbA1c% levels during the first visit to the UKB assessment centers. We allocated participants with 4%<HbA1c%<=5.6% to the normoglycemic group, and participants with 5.7%<HbA1c%<6.5% levels to the pre-diabetic group[26]. As HbA1c% is one of the identifiers of T2D, this measure is a strong predictor of T2D. The prevalence of T2D onset within the normoglycemic group was only 1% versus a prevalence of 12% in the pre-diabetic group. We examined the driving factors of T2D in each of these stratified groups (Table 1S). Within the normoglycemic group, the anthropometry model yielded an auROC of 0.81 (0.76-0.85) with an APS of 0.05 (0.03-0.08) and deciles' OR of x31 (8.2-51). When testing the models within the pre-diabetic group, the anthropometry model achieved an auROC of 0.75 (0.7-0.79), APS of 0.32 (0.24-0.41) and deciles' OR of x26 (9.6-37). Both of these results outperform the FINDRISC and the GDRS results. For the normoglycemic HbA1c% range the anthropometry model resulted in an auROC of 0.82 (0.77-0.87), APS of 0.06 (0.04-0.1) and deciles' OR of x29 (7.5-56). These results are similar to those of the five blood tests model's results (Figure 3 A-C).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

3. Discussion and conclusions

In this study, we analyzed several models for predicting the onset of T2D, which we trained and tested on a UKB based cohort, aged 40 - 69. Due to their accessibility and high predictability, we suggest two simple logistic regression models: the anthropometric and the five blood tests models. These models are suited for the UKB cohort, or populations with similar characteristics (See Table 1).

To provide an accessible and simple, yet predictive model, we based our first proposed model on eight non-laboratory anthropometric measures. We then developed an additional, straightforward model which is more accurate than the anthropometric model, to be used when laboratory blood tests are within reach. We based our second proposed model on five blood tests, including the age and sex of the participants. Both models are provided in their logistic regression form which is more accurate' yet requires a computer aided analysis, and in an easy to use scoreboards form. For both models, we obtained results that were superior to those of the current clinically validated non-laboratory models, the Finnish Diabetes Risk SCore (FINDRISC) and the German Diabetes Risk Score (GDRS). To have a fair comparison, we trained these reference models and evaluated the predictions on the data sets of our models.

Our models achieved a better auROC, APS, and decile prevalence OR, and better-calibrated predictions than the FINDRISC and GDRS models. The anthropometrics model and the five blood tests logistic regression model delivered prevalence OR of x77 and x87, respectively, while their scoreboard forms achieved OR of x61 and x78 respectively.

Analysing our models' feature importance, we conclude that the most predictive features of the anthropometry model are the waist to hip ratio (WHR) and body mass index (BMI), both of which are body measures that also encapsulate data regarding body type or shape. These features are known in the literature as being related to T2D, for instance in the metabolic syndrome[19]. The most predictive features of the five blood tests model are the HbA1c%, which is a measure of the glycated-haemoglobin carried by the red blood cells often used to diagnose Diabetes, and the Reticulocyte count which is a measure of the number of young red blood cells. Using both these features may provide a better indication of the average blood sugar level during the last two - three months than using just the standard HbA1C% measure. Interestingly, Age and Sex had a very low OR value, meaning that they hardly contributed to the model. One explanation might be that the T2D related information of these features is already latent within the blood test data. For instance, the SHBG feature contains a continuous measure regarding the sex hormone of each participant thus making the Sex feature redundant.

One of the limitations of our study is that our cohort is biased from the actual U.K. population. Our cohort's T2D prevalence was only 2.16% during the time of the research, while the general UK population T2D prevalence is 6.3%, and 8% among adults aged 45-54 in the general UK population (2019)[27]. This bias is commonly reported as a "healthy volunteer" selection bias [28,29], which reduces the T2D prevalence from 6% in the general UK population to 4.8% in the UKB population. An additional screening bias is caused by including only healthy participants at the first visit. This reduced the prevalence of T2D in our cohort to 2.08%. As such, to apply our models to additional populations, further research on their ethnicity and fine tuning of the feature coefficients might be required.

As several studies have concluded [7,8,9], a healthy lifestyle and diet modifications before the inception of T2D are expected to reduce the probability of T2D onset. Therefore, identifying people at risk for this disease is crucial. We assert that our models make a significant contribution to such identification in two ways: The laboratory five blood tests model for clinical use is highly predictive of T2D onset, and the anthropometrics mode, mainly in its coreboard

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

form, is an easily accessible and accurate tool. Thus these models carry the potential to improve millions of people's lives and reduce the economic burden on the medical system.

4. Methods

4.1 Data

We analysed UKB's observational data of 500,000 participants that were recruited voluntarily during 2006-2010 from across the UK, aged 40-69. During the baseline assessment visit to the UK Biobank, the participants self-completed questionnaires, including lifestyle and other potentially health-related information. The participants also went through physical and biological measurements. Out of this cohort, we used the data of 20,346 participants who revisited the UK Biobank assessment center during 2012-2013 during the longitudinal research, and we also used the data of 48,705 participants that revisited for a second or third visit from 2014 onwards for an imaging visit and went through a medical check very similar to the one in the first visit to the assessment centre. We performed a screening process on the participants to keep only the ones who were not treated nor have T2D. We thus kept data of 44,873 participants in our study cohort, from which 2.16% developed T2D during a follow-up period of 7.3±2.3 years (Table 1, Figure 1A).

We started with 798 features for each participant and removed all the features which had more than 50% missing data points in our cohort. We later removed from the cohort all the participants who still had more than 25% missing data points. We then imputed the remaining missing data. We further removed from the study the participants who self-reported as being healthy but had HbA1c% levels higher than the healthy level of glycated haemoglobin (HbA1c%) test, which is often used to identify T2D, measuring the average blood sugar for the past 2 to 3 months. As not all of the participants had HbA1c% measurements, we estimate the bias of participants reporting as being healthy while having an HbA1c% level indicating as being diabetic. To do so, we used the data we have from a subpopulation of our patients and found it to be 0.5% of participants who reported as being healthy with a median HbA1c% value of 6.7%, while the cutoff for having T2D is 6.5%. (Table 1)

4.2 Feature selection process

For the feature selection process, we started with 798 features that we estimated as potential predictors for T2D onset. We then removed all the features which had more than 50% missing data values, leaving 279 features for the research. Next, we imputed the missing data of the remaining records (See methods). As the genetic input for some of the models, we used for each participant both Polygenic Risk Scores (PRS) and Single-Nucleotide-Polymorphisms (SNPs) from the UKB SNP array (See methods). We used forty-one PRSs with 129±37.8 SNPs on average for each PRS. We also used the single SNPs of each PRS as some of the models' features; after removal of duplicate SNPs, we remained with 2267 SNPs (See methods).

Out of the screened features and the genetic data, we aggregated the features into thirteen sperate groups: age and sex; genetics; early life factors; sociodemographics; mental health; blood pressure and heart rate; family and ethnicity; medication; diet; lifestyle and physical activity; physical health; anthropometry; blood tests. We then ran models for each group of features separately; later, we trained models where we added the features groups according to their marginal predictability. (Figure 1A, supplementary material).

After selecting our leading models from the training and validation data sets, we tested and reported the results of the selected models from the holdout test set (Figure S1, see methods, supplementary material). To encourage extensive clinical use of our models, we optimised the number of features we use. We chose the logistic regression models as our final models due to their simplicity and interpretability while providing similar results to the GBDT models that we

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

validated (See methods). For the purpose of simplifying our models, we analysed the validation models' features' importance, and iteratively removed the least contributing features (See methods, supplementary material). We used each normalised feature's coefficient as a measure of its importance in the model.

4.3 Outcome

Our models provide a prediction score for the participant risk of developing T2D during a specific timeframe. The mean time between the first visit and the prediction time point in our cohort is 7.3±2.3 years. The results that we report are of a holdout test-set comprising 20% of our cohort that we kept aside up until the final report of the results. We trained all the models using the same training set, and we then reported the test results of the holdout test-set. We present the area under the receiver operating curve (auROC) and also the average precision score (APS) as the main metrics of our models. Using these models, a physician can inform patients regarding the risk fold of developing T2D vs the participants in the lowest risk decile or vs any other risk decile. Using the scoreboards models, a patient can obtain its predicted risk of developing T2D during the following years.

We calibrate the models to enable reporting of the probability to develop a T2D during a given timeframe. Calibration refers to the concurrence between the real T2D onset occurrence in a subpopulation and predicted T2D onset probabilities in this population. Since our data is highly imbalanced, with the prevalence of 2.17% T2D, we used one thousand bootstrapping iterations of each model to better estimate the mean predicted value in each calibration bin. To calculate the calibration curves, we first split the prediction of each model to ten deciles bins in the range of zero to one. We then scale the results using SKlearn's isotonic regression calibration with five-fold cross-validation [30] . We do so for each of the bootstrapping iterations. We then concatenate all the calibrated results and calculate the overall mean predicted probability for each probability decile.

For the quantisation of the risk groups of the scoreboards model, we performed a similar bootstrapping process on our validation data set. We considered several potential risk score limits that separate T2D onset probability, and we chose boundaries that showed good separation between risk groups. We then measured the prevalence in each risk group on the test set and we report these results.

4.4 Missing data

After removing all features with more than 50% missing data, and removing all the data of participants with more than 25% missing features, we imputed the remaining missing data. We analysed the correlations between predictors with missing data and found mainly correlations within anthropometry group features to other features in the same domain and same for blood-tests. We used SKlearn's iterative imputer with a maximum of 10 iterations for the imputation and a tolerance of 0.1[30] . We imputed the training and validation sets apart from the imputation of the holdout test-set. We did not perform imputation on the categorical features but rather transformed them into one hot encoding with a bin for missing data using Pandas categorical tools.

4.5 Genetic data

We use both Polygenic Risk scores (PRS) and Single-Nucleotide-Polymorphisms (SNPs) as genetic input for some of the models. We calculate the PRS by summing the top correlated risk allele effect-sizes derived from Genome-Wide Association Studies (GWAS) summary statistics. To do so we first extracted from each summary statistics the top 1000 SNPs according to their p-value. We then took only the SNP's that were presented also in the UKB SNP-array. We used 41 PRSs with 129±37.8 SNPs on average for each PRS. We also used the single SNPs of each PRS as features for some of the models, after removal of duplicate SNPs, we kept 2267 SNPs as features. The full list of the PRS summary statistics is given in the supplementary material.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

To prevent data leakage, we calculated the PRS scores according to summary statistics publicly available from studies that were not derived from the UK Biobank. We also provided the models which include genetic data and raw SNPs data as an input.

4.6 Baseline models:

As the reference models for our results, we used the well established FINDRISC and GDRS models[14,17,18], which we retrained and tested on the same data that we used for our models (See methods). These two models are based on the Finnish and German populations which are relatively close to the U.K. population and on similar age ranges.

We trained and tested these models on the same data that we use for our models. We derive a FINDRISC score for each participant using the data for age, sex, Body Mass Index (BMI), waist circumference, and blood pressure medication as provided from the UKB. To calculate the duration of the physical activity score, as required by the FINDRISC model, we summed up the values of "duration of moderate activity" and "duration of vigorous activity" as provided by the UKB. As a measure for the consumption of vegetables and fruits we summed up the categories "cooked vegetable intake", "Salad/raw vegetable intake", and "fresh fruit intake" categories from the UKB. As an answer regarding the question *"Have any members of your patient's immediate family or other relatives been diagnosed with diabetes (type 1 or type 2)? This question applies to blood relatives only"* we used the fields for the illness of the mother, the father and the siblings of each participant.

We lacked the data regarding participant's grandparents, aunts, uncles, first cousins and children. We also lack the data regarding past blood pressure medication, but rather have the data for the current medication usage. Following the calculation of the FINDRISC score for each participant, we trained a logistic regression model using the score for each participant as the model input, and the probability of developing T2D as the output. We also calculated an additional model, in which we added the time of the second visit as an input for the FINDRISC mode, but found no major differences between the two. We report here the results for the FINDRISC model without time of the second visit as a feature.

To derive the GDRS model, we built a Cox regression model using Python's lifelines package [32]. As the features for the GDRS model we calculated the following features: years between visits; height; prevalent hypertension; physical activity (h/week); smoking habits (Former smoker<20 units per day or >=20 units per day, current smoker >=20 units per day or <20 units per day); whole bread intake; coffee intake; red meat consumption; one parent with diabetes; both parents with diabetes and a sibling with diabetes. We performed a random hyperparameters search in the same way that we used for our models. The hyperparameters we used here are: the penaliser parameter in the range of 0-10 using a 0.1 resolution; variance threshold 0-1 with 0.01 resolution to drop columns where the variance of the column was lower than the variance threshold.

4.7 Model building procedures

To prevent overfitting and biased models, we split the data to twenty percent of a holdout test set which we used only for the final reporting of results. From the remaining data, we split again into a thirty percent validation set and a seventy percent for the training set. We then use a two-stage process to evaluate the models' performance: an exploration phase and a test phase (Figure 1, S1). During the exploration stage, we select the optimal features for our models using the training and validated data sets. For each group of features, we optimised the hyperparameters using two-hundred iterations of a random selection process. In each iteration, we measured the performance using the auROC metric with a five-fold cross-validation within the training set.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We later trained a model on the full training set with the top ranked hyper-parameters from the previous step. We test this model using the validation data set. We use this stage to compare various models and for the features selection process for our models.

At the final phase, the test phase, we report the results of our selected models. In this phase, we evaluate the selected models on the holdout test-set. To do so, we rerun the hyperparameters selection process using the training and validation data sets. We train the selected models with the selected hyperparameters on the pooled training and validation data sets. Lastly, we calculate the results of the trained model based on the holdout test-set. We use the same datasets for all of the discussed models.

For the logistic regression models we used SKlearn's LogisticRegressionCV model [30]. For the GBDT models we used Microsoft's LightGBM package [33] , and for the survival analysis models, we used the lifelines package [32].

During the models' calculation process we used two-hundred iterations of random hyperparameters-search for the training of the models. For the GBDT models we used the following parameters values for the search: number of leaves -[2, 4, 8, 16, 32, 64, 128]; Number of boosting iterations - [50, 100, 250, 500, 1000, 2000, 4000]; learning rate - [0.005, 0.01, 0.05]; minimum child samples - [5, 10, 25, 50]; subsample - [0.5, 0.7, 0.9, 1]; features fraction - [0.01, 0.05, 0.1, 0.25, 0.5, 0.7, 1]; lambda l1 - [0, 0.25, 0.5, 0.9, 0.99, 0.999]; lambda l2 - [0, 0.25, 0.5, 0.9, 0.99, 0.999]; bagging frequency - [0, 1, 5]; bagging fraction- [0.5, 0.75, 1] [33].
For the logistic regression models, during the hyperparameters search we used penaliser at the raNGE OF 0-2 with 0.02 resolution for the l2 penalty .

4.8 SHAP

As the feature importance analysis for the GBDT model, we used the SHAP method, which approximates Shapley values. SHAP (SHapley Additive exPlanations) originated in a game theory, intended to explain the output of any machine learning model. SHAP Approximates the average marginal contributions of each feature of a model across all permutations of the other features in the same model [34].

4.9 Predictors

To estimate the contribution of each feature's domain and for initial screening of features, we started by building a GBDT model based on 279 features plus genetics data originating from the UKB SNPs array. We used T2D related summary statistics from Genome-Wide-Association-Studies (GWAS). These are genetic studies designed to find correlations between known genetic variants and a phenotype of interest. To avoid data leakage, we used only GWASs that derived from outside the UKB population (See supplementary material for the full list of PRSs). As the feature importance analysis for the GBDT model, we used the SHAP method [34] , which approximates Shapley values (See methods).

To select the most predictive features for the anthropometry and the blood-tests models, we trained and tested the full-features model using the training and validation cohort, and then used this model's feature importance to extract the most predictive features. We also analysed models which included data of family relatives with T2D using only the training and validation sets. As we did not observe any major improvement over the anthropometrics model, for the simplicity of the model, we decided to omit this feature. At the last step, we tested and reported the model on the holdout test.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

For the extraction of the five blood tests model, we performed a features selection process by evaluating logistic regression models using the training and validation datasets. We ran models with twenty, ten, and down to four features of blood tests together with age and sex as features, each time removing the blood test with the least essential feature score. We then selected the model with five blood tests (HbA1c%, Reticulocytes count, Gamma Glutamyl Transferase (GGT), Triglycerides, HDL cholesterol, age and sex) as the optimal balance between model's simplicity (low number of features) and model's accuracy (using more features) and report its results on the holdout test set.

We normalised all the continuous predictors using the standard z-score. In order to avoid data leakage, the train-validation sets were normalised apart from the holdout test set.

4.10 Models calibration

For each of the models, we calculated the deviation of the mean predicted probability from the actual T2D prevalence of each bin.

We split the probabilities range (0-1) to ten prediction probabilities bins with probabilities resolution of 0.1 (Figure 3F). We assign each prediction's sample to a decile bin according to the calibrated predicted probability of T2D onset. Since our data is highly imbalanced, with a prevalence of 2.17% T2D onset, we used one thousand bootstrapping iterations to better calibrate the models. As such, each participant might be present at several bins according to each prediction iteration of the bootstrapping process. We repeated this process also for the scoreboards models.

Figure 4|Models calibration plots

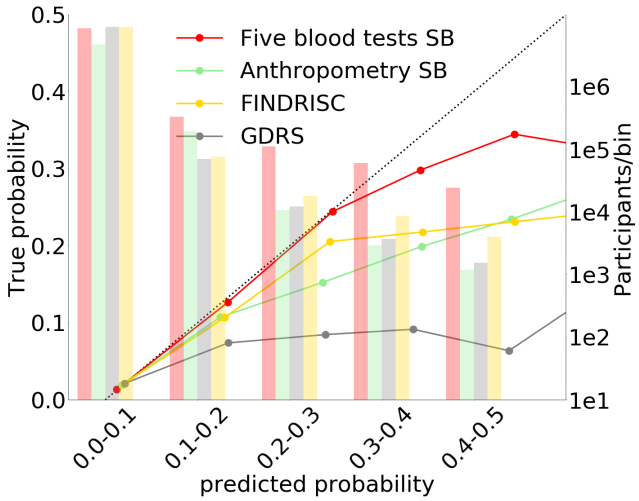


Figure 4 Anthropometric, five blood tests FINDRISC and GDRS scoreboards calibration graphs*. While Five blood tests, Anbthropometrics and GDRS are monotonically rising, the FINDRISC model starts to decline after the third bin. The calibration of the Anthropometric and Five blood tests deteriorates compared to the continuous logistic regression model due to the scores quantisation effect.*

4.11 Extracting scoreboards

To extract our scoreboards, we explored the train and validation data sets, and reported the results on the hold out data set. We calculated the weight of evidence (WoE) of our data by splitting each of our features into bins. We binned

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

in higher resolution features that have greater importance, while maintaining monotonically increasing WoE. (Anon n.d.) For the quantisation of the risk groups of the scoreboards model, we performed one thousand iterations of the bootstrapping process on our validation data set. We considered several potential risk score limits that separate T2D onset probability in each of the scores groups, and we chose boundaries that showed a separation between the risk groups. We then measured the prevalence in each risk group on the test set and we report these results.

4.12 References for PRS summary statistics articles.

HbA1c[353,36,37] ; Cigarettes per day, ever smoked, age start smoking[38] ; HOMA-IR, HOMA-B, diabetes BMI unadjusted, diabetes BMI adjusted, fasting glucose [39]; Fasting glucose, 2 hours glucose level,fasting insulin, fasting insulin adjusted BMI'-(MAGIC_Scott)[40]; Fasting glucose, fasting glucose adjusted for BMI,fasting insulin adjusted for BMI[41]; Two hours glucose level[42]; Fasting insulin [43]; Fasting Proinsulin[44]; Leptin adjusted for BMI, Leptin unadjusted for BMI[45]; Triglycerides, Cholesterol, ldl, hdl[46]; BMI[47]; Obesity class1, obesity_class2, overweight [48];Anorexia[49]; Height[50]; Waist circumference, hips circumference[51]; Cardio[52]; Heart_Rate[53]; Alzheimer[54]; Asthma [55]

5. Acknowledgements:

This research has been conducted using the UK Biobank Resource under Application Number 28784

6. Bibliography

1. Zimmet, P., Alberti, K. G., Magliano, D. J. & Bennett, P. H. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. *Nat. Rev. Endocrinol.* **12,** 616–622 (2016).
2. International Diabetes Federation - Type 2 diabetes. at <https://www.idf.org/aboutdiabetes/type-2-diabetes.html>
3. WHO | Diabetes programme. at <https://web.archive.org/web/20140329084830/http://www.who.int/diabetes/en/>
4. Beagley, J., Guariguata, L., Weil, C. & Motala, A. A. Global estimates of undiagnosed diabetes in adults. *Diabetes Res. Clin. Pract.* **103,** 150–160 (2014).
5. Wilson, M. L. *et al.* Access to pathology and laboratory medicine services: a crucial gap. *Lancet* **391,** 1927–1938 (2018).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

6. Home | ADA. at <https://www.diabetes.org/>

7. Knowler, W. C. *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **346,** 393–403 (2002).

8. Lindström, J. *et al.* Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. *Lancet* **368,** 1673–1679 (2006).

9. Diabetes Prevention Program Research Group. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. *Lancet Diabetes Endocrinol.* **3,** 866–875 (2015).

10. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* **343,** d7163 (2011).

11. Collins, G. S., Mallett, S., Omar, O. & Yu, L.-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* **9,** 103 (2011).

12. Kengne, A. P. *et al.* Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. *The Lancet Diabetes & Endocrinology* **2,** 19–29 (2014).

13. Bernabe-Ortiz, A., Perel, P., Miranda, J. J. & Smeeth, L. Diagnostic accuracy of the Finnish Diabetes Risk Score (FINDRISC) for undiagnosed T2DM in Peruvian population. *Prim. Care Diabetes* **12,** 517–525 (2018).

14. Lindström, J. & Tuomilehto, J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* **26,** 725–731 (2003).

15. Meijnikman, A. S., De Block, C. E. M., Verrijken, A., Mertens, I. & Van Gaal, L. F. Predicting type 2 diabetes mellitus: a comparison between the FINDRISC score and the metabolic syndrome. *Diabetol. Metab. Syndr.* **10,** 12 (2018).

16. EPIC Centres - GERMANY. at <https://epic.iarc.fr/centers/germany.php>

17. Schulze, M. B. *et al.* An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care* **30,** 510–515 (2007).

18. Mühlenbruch, K. *et al.* Update of the German Diabetes Risk Score and external validation in the German MONICA/KORA study. *Diabetes Res. Clin. Pract.* **104,** 459–466 (2014).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

19. Eckel, R. H., Grundy, S. M. & Zimmet, P. Z. The metabolic syndrome. *Lancet* **365,** 1415–1428 (2005).
20. Cheng, C.-H. *et al.* Waist-to-hip ratio is a better anthropometric index than body mass index for predicting the risk of type 2 diabetes in Taiwanese population. *Nutr. Res.* **30,** 585–593 (2010).
21. Jafari-Koshki, T., Mansourian, M., Hosseini, S. M. & Amini, M. Association of waist and hip circumference and waist-hip ratio with type 2 diabetes risk in first-degree relatives. *J. Diabetes Complicat.* **30,** 1050–1055 (2016).
22. Qiao, Q. & Nyamdorj, R. Is the association of type II diabetes with waist circumference or waist-to-hip ratio stronger than that with body mass index? *Eur. J. Clin. Nutr.* **64,** 30–34 (2010).
23. Fekete, T. & Sopon, E. Glycaemic control and reticulocyte count in diabetic patients. *Horm. Metab. Res.* **18,** 141 (1986).
24. Kontush, A. & Chapman, M. J. Why is HDL functionally deficient in type 2 diabetes? *Curr. Diab. Rep.* **8,** 51–59 (2008).
25. Bitzur, R., Cohen, H., Kamari, Y., Shaish, A. & Harats, D. Triglycerides and HDL cholesterol: stars or second leads in diabetes? *Diabetes Care* **32 Suppl 2,** S373-7 (2009).
26. Understanding A1C | ADA. at <https://www.diabetes.org/a1c>
27. Diabetes Prevalence 2019 | Diabetes UK. at <https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevalence-2019>
28. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186,** 1026–1034 (2017).
29. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A structural approach to selection bias. *Epidemiology* **15,** 615–625 (2004).
30. Alex, F., ALEX, G., Bertr, R. G. F., BERTR, T. & THIRION. Scikit-learn: Machine Learning in Python.
31. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63,** 938–9; author reply 939 (2010).
32. Davidson-Pilon, C. *et al.* CamDavidsonPilon/lifelines: v0.24.16. *Zenodo* (2020). doi:10.5281/zenodo.3937749
33. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. (2017).
34. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

35. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. *Diabetes* **59,** 3229–3239 (2010).

36. Walford, G. A. *et al.* Genome-Wide Association Study of the Modified Stumvoll Insulin Sensitivity Index Identifies BCL2 and FAM19A2 as Novel Insulin Sensitivity Loci. *Diabetes* **65,** 3200–3211 (2016).

37. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med.* **14,** e1002383 (2017).

38. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42,** 441–447 (2010).

39. Morris, G. P. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci USA* **110,** 453–458 (2013).

40. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44,** 991–1005 (2012).

41. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44,** 659–669 (2012).

42. Saxena, R. *et al.* Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42,** 142–148 (2010).

43. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).

44. Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60,** 2624–2634 (2011).

45. Kilpeläinen, T. O. *et al.* Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nat. Commun.* **7,** 10494 (2016).

46. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20165092; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

47. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).

48. Berndt, S. I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45,** 501–512 (2013).

49. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* **19,** 1085–1094 (2014).

50. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).

51. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518,** 187–196 (2015).

52. CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45,** 25–33 (2013).

53. den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat. Genet.* **45,** 621–631 (2013).

54. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).

55. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363,** 1211–1221 (2010).