Prediction of type 2 diabetes mellitus onset using simple logistic regression models

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

Diabetes mellitus has a world death rate of 1.6 million (2016) of which Type 2 diabetes mellitus (T2DM) accounts for ~90% of all cases. Early detection of T2D high-risk patients can reduce the incidence of the disease through a change in lifestyle, diet, or medication. Since lower socio-demographics layers are more susceptible to T2D and might have limited resources for laboratory testing, there is a need for accurate prediction models based on non-laboratory parameters. Here, we analysed data of 44,879 non-diabetic, UK-Biobank participants at the ages 40-65 within a time frame of 7.3±2.3 years. We devise a non-laboratory prediction model for T2DM onset probability using sex, age, weight, height, waist, hips-circumferences, Waist-Hips Ratio (WHR) and Body-Mass Index (BMI). This model achieved an Area Under the Receiver Operating Curve (auROC) of 0.82 (0.79-0.84 95% CI) and an odds ratio (OR) between the top and lowest prevalence deciles of x42 (33-49). The logistic regression top predictive parameters are WHR with OR of 0.67 (0.49-0.88 95%CI) followed by BMI with OR of 0.53 (0.26-0.79). We further analyse the contribution of laboratory-based parameters and devise a blood-test model based on only five blood tests. In this model, we are using age, sex, Glycated Hemoglobin (HbA1c%), reticulocyte count, Gamma Glutamyl-Transferase, Triglycerides, and HDL cholesterol to predict T2D onset more accurately. This model achieves an auROC of 0.89 (0.87-0.92) and a deciles' OR of x59 (27-75). We also analysed a model that included genotyping data and other environmental factors and found that it did not provide further benefit over the five-blood-tests model. Our models outperform the current state of the art, non-laboratory, Finnish Diabetes Risk Score and the German Diabetes Risk Score, trained on our data, achieving auROC of 0.74 (0.7-0.77) and 0.63 (0.59-0.67), respectively.

Introduction

Diabetes mellitus is classified as a group of diseases characterised by symptoms of chronic hyperglycemia and 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 during 2016. T2D is generally characterised by insulin resistance, which can eventually exhaust the pancreas, resulting in hyperglycemia, and accounting for ~90% of all diabetes cases [1,2].

In recent years, the prevalence of Diabetes has been rising more rapidly in the low and middle-income countries than in high-income countries[3], while the access to laboratory medical tests is of the essence for some of the populations in these low-middle income countries.

According to several studies, a healthy diet, regular physical activity, maintaining normal body weight and avoiding tobacco use can prevent or delay T2D onset [3,4,5,6,7] . As such, having an accessible and accurate, simple and low cost, preferably without the need for laboratory testing screening tool, is of a great need for the identification of high-risk

patients. Such models can delay or even prevent T2D onset through early detection and a conventional change in lifestyle, diet or medications.

Several such models are in use today [8,9,10] . The Finnish Diabetes Risk Score (FINDRISC) which is a commonly used, non-invasive T2D risk-score model, estimates the probability of a person to develop T2D within the next ten years. This model was created and validated using a prospective cohort of 4,746 and 4,615 individuals in Finland in 1987 and 1992, respectively, aged between 35 and 64 years. The FINDRISC model uses gender; age; Body Mass Index (BMI); use of 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 model assigns a score to each answer and uses the sum of scores as the input to a logistic regression model. It provides a predicted probability to develop T2D during the next ten years [11,12,13] .

Another commonly used prediction model is the German Diabetes Risk Score (GDRS) which estimates a 5-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. 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; parents and sibling history of T2D [14,15].

Our objective in this research was to develop clinically usable models which are easy to use and are highly predictable. We developed two simple models that outperform the FINDRISC and GDRS models, tested on a held-out test set from the U.K. Biobank (UKB). We based one model on accessible anthropometric measures, and additional more accurate, though invasive which is based on five blood tests. Our models are most relevant for the U.K. population at the ages of 40-65, but can also be used for people similar to our research cohort [Table 1]

Results

We analysed the data of 20,346 participants from the U.K. Biobank's (UKB) observational study which revisited U.K. biobank's assessment centre during 2012-2013 and 48,705 participants that revisited for a second or third visit from 2014 onwards (Figure 1, see methods). During the screening process of our cohort, we kept only the data of the participants who returned for a second or a third visit and were not treated nor had T2D. We thus continued with research data of 44,879 participants, of which 2.15% developed T2D during a follow-up period of 7.3±2.3 years (Table 1, Figure 1A, See Methods).

To avoid overfitting of the models to our data, before training the models, we split from our cohort a held-out test set including 8,997 participants. We used this data only for evaluating the final models. We split the remaining data into a training data-set with 25,125 participants, and a validation data-set containing 10,757 participants. We selected the features for our models using these train and validated data sets, and here we report the final results from the held-out test set (Figure S1, See Methods).

Anthropometric based model

To provide an accessible, simple, non-laboratory and non-invasive T2D prediction model, we built the logistic regression anthropometric based model consisting of the following eight parameters: age, sex, weight, height, hips, waist circumference, Body mass index (BMI) and the waist hips ratio (WHR) as the model features.

Testing this model using the held-out test set, we achieve an area under the receiver operating curve (auROC) of 0.82 (0.79-0.84 at 95% CI) and an average precision score (APS) of 0.12 (0.08-0.15 at 95% CI). This model outperforms both the FINDRISC model's results with auROC of 0.74 (0.7-0.77) and an APS of 0.06 (0.05-0.07), and the German diabetes risk score (GDRS) model's results with auROC of 0.63 (0.59-0.67) and APS of 0.04 (0.03-0.06), which we used as the baseline for reference for our models (Figure 2A-B, see methods). With the cohort's baseline prevalence of 2.17%, the anthropometric model achieves deciles' OR of x42 (33-49) compared to x26 (10-41) and x7.3 (2.8-19) of the FINDRISC and GDRS models, respectively (Figure 2C).

Analysing the models' features importance, we conclude that the WHR and the BMI are the most influencing features in the model with the highest logs-odds-ratio (Figure 2D). Both of these features are commonly used in the literature to indicate for body habitus and are predictive of the risk of chronic disease [16,17,18,19]

Five-blood-tests based model

To provide a more predictive tool for T2D prediction, when laboratory data is within reach, we developed a simple logistic regression model based on five blood tests. To derive this model, we first investigated the features importance list of a full-features GBDT model using the train and validation data sets. From this model's features importance, we concluded that the most predictive features group is the blood tests group. We thus trained a full blood test model using 59 blood tests that are available in the train data. We then took the top ten blood-tests predictors and iteratively removed the features which contribute the least impact on the model predictability (see methods). When we reached a model that is based on five-blood-tests, the model achieved an auROC of 0.88 (0.85-0.9) on the training set, compared to a six-blood-tests model's, which resulted with auROC of 0.88 (0.86-0.9). We considered this a tolerable compromise between model simplicity and model predictability. When we compared the five-blood-tests model to a model with only four blood tests, the auROC dropped to 0.86 (0.83-0.89). We performed a similar analysis using a GBDT model, providing an auROC of 0.87 (0.85-0.9) for the five blood tests model. As such, our model of choice is the logistic-regression five-blood-tests model due to its simplicity while keeping the predictability.

Using the five blood-tests model we achieved the following results on the test set: auROC of 0.89 (0.87-0.92), APS of 0.28 (0.21-0.34), and deciles' OR of x59 (27-75) (Figure 2 A-C, Table 2). We then compared these results to the results of a logistic regression model with all 59 blood tests as features, and to the results of a GBDT model of all of the 279 available features. These two models achieved auROC of 0.91 (0.88-0.93) and 0.9 (0.88-0.92) respectively; an APS of 0.33 (0.26-0.4) and 0.28 (0.22-0.35); and deciles' OR of x69 (35-79) and x65(49-73) respectively. These five-blood-tests model results are superior to the previously discussed non-laboratory anthropometrics, FINDRISC and GDRS models as discussed before (Figure 2 A-C, Table 2).

The five blood tests that we are using are glycated haemoglobin (HbA1c%), which measures the average blood sugar for the past 2 to 3 months and is one of the means to diagnose Diabetes; Reticulocyte count; Time to prediction (Time between visits); Gamma-glutamyltransferase (GGT); Triglycerides; Sex (female); age at the repeated visit; HDL cholesterol and a bias term which is related to the prevalence in the population. We compute the values of the associated coefficients with their CI to enable a reconstruction of the models (Figure 2E).

As expected, as it is one of the criteria for T2D diagnosis, the most predictive feature of the five blood tests model is the HbA1c% value. The next feature in the feature importance list is the high-light-scatter-reticulocytes-count, which reflects the number of new red blood cells in the body[20]. HDL cholesterol, which is known to be beneficial for health, especially in the context of cardiovascular diseases and T2D [13,21,22] also contributes in this model to reduce the predicted probability of T2D. When using the blood-tests results, interestingly, the age and sex values of the patients

merely contribute to the model's prediction result - probably as a result of having the relevant information of these features latent within the blood-tests' data.

	UKB population	Train, validation and test sets	Train set	Validation set	Test set
Number of participants	502536	44879	25125	10757	8997
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 [%]	45.5	47.8	47.9	47.5	48
Diabetic at first visit [%]	4.8	0	0	0	0
Diabetic at last visit [%]	-	2.15	2.13	2.21	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±14.8	76.9±15.0	76.5±14.7
Weight at last visit [Kg]	-	76.3±15.2	76.2±15.2	76.5±15.3	76.0±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]	-	101.6±8.9	101.6±8.8	101.8±9.0	101.4±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]	-	88.8±12.7	88.7±12.7	89.1±12.9	88.6±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±7.6	55.6±7.6	55.6±7.6	55.5±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]	-	26.6±4.4	26.6±4.4	26.7±4.5	26.6±4.3

Table 1|Cohort's Statistics table. Characteristics of the cohort's population and the general UKB population. When a "±" sign appears, it denotes the standard deviation from the mean. While the prevalence in the general UKB participants is 4.8%, In our cohort we screened the population at baseline to consist of people who had HbA1c% levels<6.5%, and thus, a lower rate of participants developed T2D during the research period (2.15%). The age range of the participants at the baseline visit is 40-73 ; as such, our models are not customised for people who might develop T2D at younger ages. The predictions of the models are representing the risk to develop T2D during the time between the first visit to the UKB assessment centre and the last visit, as indicated by the time between visits which is 7.3±2.3 years.

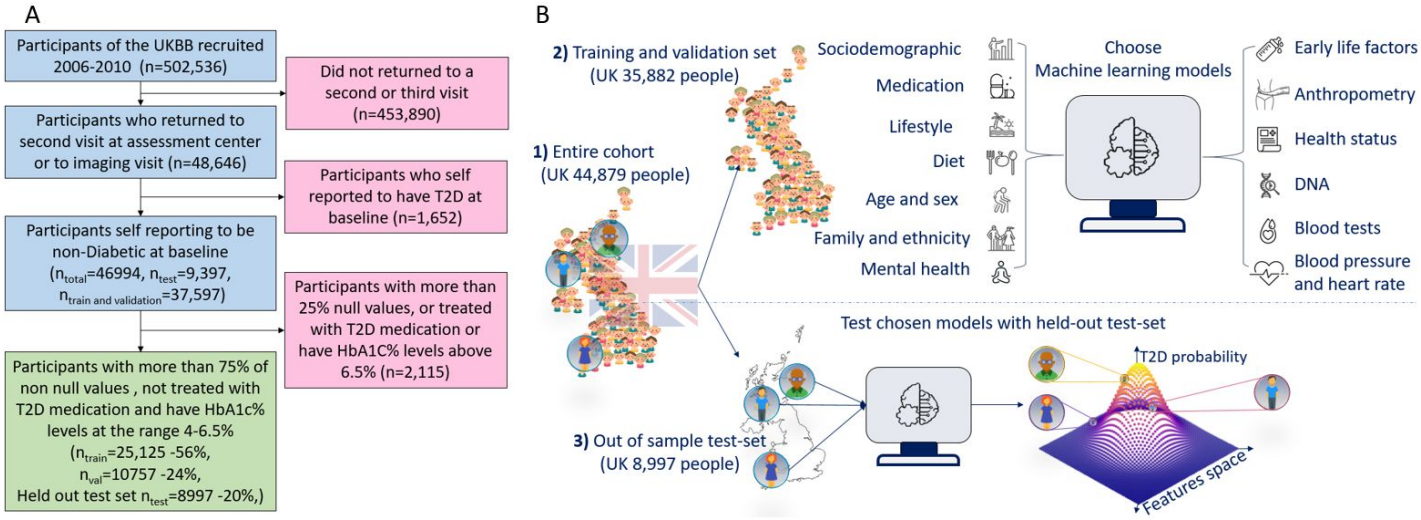

Figure 1 |A flow chart of the cohort selection process and an Illustrative figure of the model's extraction . **A.** A flowchart which demonstrates the selection process of participants in this study. Participants who came for repeated second or a third visit were selected from the 502,536 participants of the UKB. Next, we excluded 1,652 participants who self-reported as having T2D. We then split the data into 80% of the train and validation set, and 20% held-out test set. We excluded Additional 2,115 participants due to either having 25% or more missing values from the full features list or having HbA1c% levels above 6.5%, or participants who were treated with Metformin or Insulin. Finally, the train-set holds 25,125 participants (56% of the total cohort), the validation set holds 10,757 (24% of the whole cohort), and the test set contains a total of 8,997 participants (20% of cohort).**B.** Cohort flow during training, and testing of the models. We first split the data and kept a held-out test-set. We later explored several models using the train and validation data sets. In the final stage, we compared the selected models using the held-out test-set and reported the results. The output of the models is calibrated to provide the probability of a participant to develop T2D.

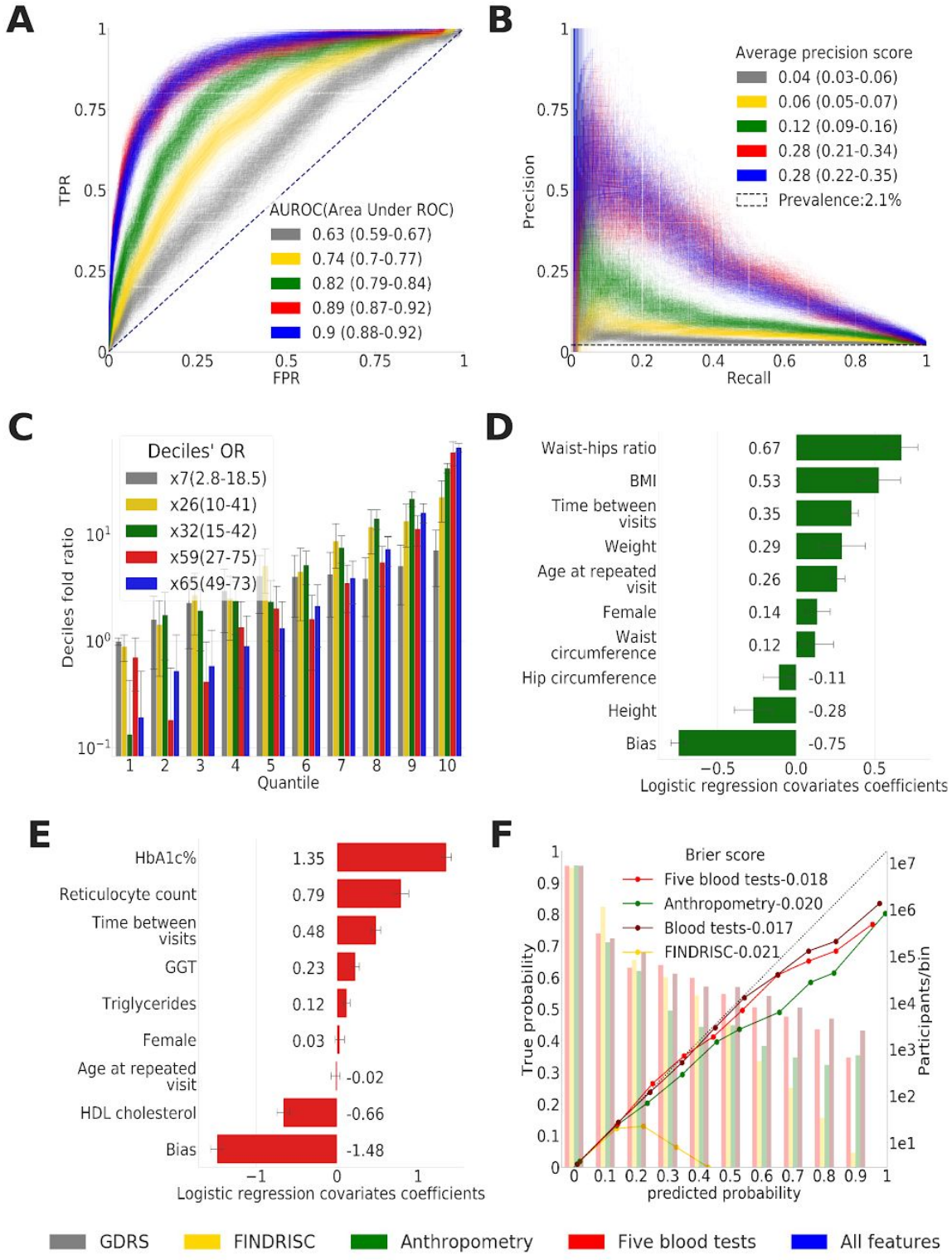

Figure 2 |Main results calculated using 1000 bootstraps of the cohort population. **Main results calculated using 1000 bootstraps of the cohort population**. Each point in the graphs represents a bootstrap iteration result. *The color legend is shown at the bottom of the figure. **A.** ROC curves comparing the models developed in this research: a GBDT model of all features; logistic-regression models of five-blood-tests and the 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 (S.D.) of the features importance values. The top predictive features of this model are the body mass index (BMI) and waist to hips ratio (WHR).**E.** Feature importance graph of logistic regression Blood-tests model with S.D. bars. While the HbA1c% and Reticulocyte are positively contributing to the T2D prediction, and HDL cholesterol lowers the T2D prediction probability, the age and sex of the participants are screened out by the other features. **F.** A calibration plot of the anthropometry; five blood tests; full blood-test and the FINDRISC models. Providing the ability to report to a patient the calculated probability to develop T2D (see methods) Calibrated using isotonic regression.

Label	Model type	Average precision score	Area under ROC	Deciles prevalence fold ratio
GDRS SA	Cox regression	0.04 (0.03-0.06)	0.63 (0.59-0.67)	7.3(2.8- 19)
FINDRISC LR	Logistic regression	0.06 (0.05-0.07)	0.74 (0.70-0.77)	26 (10-41)
Anthropometry	Logistic regression	0.12 (0.09-0.16)	0.82 (0.79-0.84)	42 (33-49)
Anthropometry and in family T2D history	Logistic regression	0.12 (0.08-0.15)	0.82 (0.79-0.85)	42 (22-51)
Five blood tests	Logistic regression	0.28 (0.21-0.34)	0.89 (0.87-0.92)	59 (27-75)
All features	Boosted decision trees	0.28 (0.22-0.35)	0.90 (0.88-0.92)	65 (49-73)
Blood Tests	Logistic regression	0.33 (0.26-0.40)	0.91 (0.88-0.93)	69 (35-79)

Table 2|Models comparison*: Comparing the "Five blood tests," and "anthropometrics" models to the "all features," "FINDRISC," and "GDRS" models. The bracketed values indicate 95% CI. The deciles' OR is a measure of the ratio of prevalence in the top risk score decile bin to the prevalence in the lowest decile bin. In the lowest decile, in case the actual prevalence in that bin was zero, we used a threshold of a tenth of the general prevalence, i.e 0.215% (see methods). Using a logistic regression model of blood tests achieves auROC and APS that are close to the full GBDT model results.*

Prediction within an HbA1c% stratified population

To verify that our models are capable of discriminating within a group of participants with normoglycemic or a group of participants with pre-diabetic participants, we tested the models separately on each group that we extracted from our data . We separated the groups based on their HbA1c% levels during the first visit to the UKB assessment centres. We allocated participants with 4%<HbA1c%<=5.7% to be in the normoglycemic group, and participants with 5.7%<HbA1c%<6.5% levels to be in the pre-diabetic group[23]. As HbA1c% is one of the Diabetes identifiers, this measure in itself is a strong predictor of T2D. The prevalence within the normoglycemic group of participants is only 1% versus a prevalence of 12% T2D onset in the pre-diabetic group. We examined what are the driving factors of T2D in each of these stratified groups (Table 1S). Within the normoglycemic group, the anthropometry model provides 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 achieves 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. The Anthropometry model's results in this normoglycemic HbA1c% range are similar to those of the five blood tests model's results with auROC of 0.82 (0.77-0.87), APS of 0.06 (0.04-0.1) and deciles' OR of x29 (7.5-56). (FigureS2 A-C)

Calibration of the models

To test our models' goodness of fit to the actual probability of developing T2D, we performed a calibration of the anthropometry; five-blood-tests; full-blood-test and the FINDRISC models (see methods). In each of the models, we calculated the deviation of the mean predicted probability from the actual T2D prevalence of each bin. We then report, for each model, the highest bin's index which achieves less than 10% and 20% deviation from the actual bin's T2D prevalence (See Table 3). In all of the three models (Anthropometry, Five-blood-tests and Full-blood-tests), we find monotonically increasing values of the actual T2D prevalence versus the bins' mean predicted probabilities throughout the whole prediction range, while the FINDRISC model shows a decline after the second bin (Figure 2F).

Model name	Highest bin's mean predicted probability with less than 10% deviation from actual bin's prevalence	Highest bin's mean predicted probability with less than 20% deviation from actual bin's prevalence	Top bin Mean predicted prevalence	Brier score
FINDRISC	2% (2%)	14% (12%)	43% (0%)	0.021
Anthropometrics	14% (13%)	46% (40%)	99% (80%)	0.020
Five blood-tests	65% (61%)	75% (65%)	95% (-24%)	0.018
Full blood-tests	75% (68%)	98% (83%)	97% (-17%)	0.017

Table 3: Calibration results. A comparison of the calibration results for the Anthropometrics, the Five-blood-tests and the Full-blood-tests models to the baseline FINDRISC model.

Conclusions/Discussion

In this study, we devised several prediction models that we trained and tested on a UKBB based cohort. Out of the models that we analysed, we suggest two simple logistic regression models for predicting the onset of T2D in the British population aged 40-69 or populations with characteristics that resemble our cohort (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 provide an additional simple model which is more accurate than the anthropometric model, though it requires laboratory blood tests. We based our second proposed model on five blood tests, including the age and sex of the participants. For both of the models, we show superior results over the current state of the art, non-laboratory models, such as 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 at the last visit of each participant to the clinical centre - on the same data sets that we used for our models.

Our models achieved better auROC, APS, decile prevalence OR, and better-calibrated predictions than the FINDRISC and GDRS models. The anthropometrics model and the five-blood-tests model deliver a fold enrichment between the highest and lowest deciles of x42 and x59, respectively. For the calibration process, we used a threshold of 10% deviation of the mean predicted value in each bin from the actual prevalence in the bin. Within these thresholds boundary, we achieved calibrated predictions of up to 10-20% and 60-70% probability of developing T2D using the anthropometrics and five blood tests model, respectively. Both models achieve better-calibrated probabilities than the calibrated FINDRISC model, which provides calibrated results up to the 0-10% probability bin on our cohort.

Analysing the features' importance of our models, we conclude that the most predictive features of the anthropometry model are the waist to hips ratio (WHR) and body mass index (BMI) – both are body measures that also encapsulate data regarding body type or shape. These features are known in the literature as being related to T2D, such as in the metabolic syndrome [16] . The top 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 and is often used to diagnose Diabetes, and the Reticulocyte-count which is a measure for the number of young red blood cells in the blood. Although we did not analyse the causality between these two features, it might be that the combination of these two features, as we use them , is a better indication of the average blood sugar level during the last 2-3 months as compared to the standard HbA1C% measure alone.

One of the limitations of our study is that our cohort is not representative of the true U.K. population, but somewhat biased towards a healthier population. Our cohort's T2D prevalence is only 2.15% during the time of the research, which is x3-x4 folds less than 6.3% prevalence in the general UK population and 8% among adults aged 45-54 in 2019 in the general UK population[24]. This bias is considered to be caused by a "healthy volunteer" selection bias [25,26] which reduces the T2D prevalence from 6% to 4.8% in the entire UKB population. An additional screening bias is caused by having only healthy participants at the first visit, which reduces the prevalence of T2D in our cohort to 2.08%. In order to make usage of our models relevant to other ethnic communities than the U.K., further research on additional ethnicities groups is required. We suggest applying the features of the anthropometrics and five-blood-tests models on new cohorts, with a preliminary stage of tuning the feature coefficients of the models on the new cohorts.

As several studies have concluded [5,6,7], promoting a healthy lifestyle and diet modifications before the inception of T2D are expected to reduce the probability of developing T2D. Having the anthropometrics model accessible online, and the five-blood-tests model as an accurate T2D predictive tool could be used for early detection of patients in a high risk of developing T2D. Further, a lifestyle, diet or medication intervention could delay or even prevent the onset of T2D, thus having the potential to improve millions of people's lives and reducing a substantial economic burden from the medical system.

Methods:

Data

We analysed UKB's observational study data of 500,000 participants recruited voluntarily during 2006-2010 from across the UK at the ages of 40-69. During the baseline assessment visit for UK Biobank, the participants self-completed questionnaires, including lifestyle and other potentially health-related information. The participants also went through physical measurements, and biological samples were collected from them.

As longitudinal data, we used the data of 20,346 participants who revisited the UK biobank assessment centre during 2012-2013. 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 returned for a second or third visit and were not treated nor have T2D. We thus kept data of 44,879 participants in our study cohort, from which 2.15% 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 had to estimate the bias of participants reporting as being healthy while having an HbA1c% levels indicating as being diabetic. 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)

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 we selected our leading models from the train and validation data sets, we tested and reported the results of the selected models from the held-out 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 validated (See methods). For the screening of the features, we analysed the feature importance from the models that we validated and iteratively chose the top importance features (See methods, supplementary material). When we used the logistic regression model, we normalised our data and used each feature's coefficient as a measure of its importance in the model.

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 held-out 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 train set, and we then reported the test results of the held-out test-set. We present the area under the receiver operating curve (auROC) and also the average precision score (APS) as the 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.

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 [27] . 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.

To verify the accuracy of a probability forecast, we use the Brier-score over a deciles calibration curve [28]. As our data is imbalanced, Brier score merely provides meaningful insights into the results. We thus also report the top deciles which have a deviation of the mean predicted values lower than 10% and 20% from the actual prevalence of the same decile.

Missing data:

After removal of 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[27] . We imputed the train and validation sets apart from the imputation of the held-out 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.

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 alleles 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.

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

Baseline models:

As the reference models for our results, we used the well established FINDRISC and GDRS models[12,14,15], 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 [29]. 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.

Model building procedures

To prevent overfitting and biased models, we split the data to twenty percent of a held-out 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 train 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 train set.

We later trained a model on the full train 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 held-out test-set. To do so, we rerun the hyperparameters selection process using the train and validation data sets. We train the selected models with the selected hyperparameters on the pooled train and validation data sets. Lastly, we calculate the results of the trained model based on the held-out test-set. We use the same datasets for all of the discussed models.

For the logistic regression models we used SKlearn's LogisticRegressionCV model [27]. For the GBDT models we used Microsoft's LightGBM package [30] , and for the survival analysis models, we used the lifelines package [29].

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] [30].
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 .

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 [31].

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 [31] , 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 train and validation cohort, we 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 train 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 held-out test.

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 held-out 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 held-out test set.

Models calibration

We split the probabilities range (0-1) to ten prediction probabilities bins with probabilities resolution of 0.1 (Figure 2F). 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.

According to the calibration results and to our metrics, the FINDRISC model can provide calibrated results to participants with 0-10% probabilities (First bin). The mean predicted probability in this bin is 2% with a similar actual bin's prevalence. For participants in the second bin (10-20% predicted probability) the FINDRISC's mean predicted value is 14%, which is more than 10% deviation from the actual 12% prevalence in this bin. In the next prediction bin, the FINDRISC mean predicted probability is 22% while the true prevalence in that bin is only 13%. From this bin

onwards, inclusive, the calibration curve is not monotonically increasing anymore and thus is a challenge for calibration usage.

The Anthropometry model provides a calibrated result with less than 10% deviation up to the second calibration bin with a mean prediction of 14% versus an actual 13% prevalence in this bin. In the fifth calibrated bin the mean predicted probability is 46% while the true prevalence in that bin is 40% - This bin is the top bin below the 20% deviation threshold.

Analysing the five-blood-tests model, we find that for participants in the seventh bin the mean predicted value is 65% versus the actual prevalence of 61%, which is less than the 10% deviation threshold. This model provides calibrated results with less than 20% deviation threshold with a mean predicted probability of 75% versus the actual 65% T2D prevalence in the seventh bin.

The full blood tests model provides a calibrated probability with less than 10% deviation from actual prevalence, up to the eighth bin, in which the mean predicted probability is 75%, this is versus the actual 68% prevalence in that bin. In the last decile of this model (90-100%), the mean predicted probability is 98% vs an actual prevalence of 83% (reflecting 17% deviation).

References for PRS summary statistics articles.

HbA1c[43,44,45] ; Cigarettes per day, ever smoked, age start smoking[46] ; HOMA-IR, HOMA-B, diabetes BMI unadjusted, diabetes BMI adjusted, fasting glucose [47]; Fasting glucose, 2 hours glucose level,fasting insulin, fasting insulin adjusted BMI'-(MAGIC_Scott)[48]; Fasting glucose, fasting glucose adjusted for BMI,fasting insulin adjusted for BMI[49]; Two hours glucose level[50]; Fasting insulin [51]; Fasting Proinsulin[52]; Leptin adjusted for BMI, Leptin unadjusted for BMI[53]; Triglycerides, Cholesterol, ldl, hdl[54]; BMI[55]; Obesity class1, obesity_class2, overweight [56];Anorexia[57]; Height[58]; Waist circumference, hips circumference[59]; Cardio[60]; Heart_Rate[61]; Alzheimer[62]; Asthma [63]

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. Home | ADA. at <https://www.diabetes.org/>
5. 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).

6. 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).
7. 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).
8. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* **343,** d7163 (2011).
9. 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).
10. 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).
11. 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).
12. Lindström, J. & Tuomilehto, J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* **26,** 725–731 (2003).
13. 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).
14. 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).
15. 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).
16. Eckel, R. H., Grundy, S. M. & Zimmet, P. Z. The metabolic syndrome. *Lancet* **365,** 1415–1428 (2005).
17. 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).

18. 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).
19. 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).
20. Fekete, T. & Sopon, E. Glycaemic control and reticulocyte count in diabetic patients. *Horm. Metab. Res.* **18,** 141 (1986).
21. Kontush, A. & Chapman, M. J. Why is HDL functionally deficient in type 2 diabetes? *Curr. Diab. Rep.* **8,** 51–59 (2008).
22. 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).
23. Understanding A1C | ADA. at <https://www.diabetes.org/a1c>
24. Diabetes Prevalence 2019 | Diabetes UK. at <https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevalence-2019>
25. 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).
26. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A structural approach to selection bias. *Epidemiology* **15,** 615–625 (2004).
27. Alex, F., ALEX, G., Bertr, R. G. F., BERTR, T. & THIRION. Scikit-learn: Machine Learning in Python.
28. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63,** 938–9; author reply 939 (2010).
29. Davidson-Pilon, C. *et al.* CamDavidsonPilon/lifelines: v0.24.16. *Zenodo* (2020). doi:10.5281/zenodo.3937749
30. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. (2017).
31. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017).
32. Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2,** 56–67 (2020).

33. Ali, O. Genetics of type 2 diabetes. *World J. Diabetes* **4,** 114–123 (2013).

34. Perry, J. R. B. *et al.* Genetic evidence that raised sex hormone binding globulin (SHBG) levels reduce the risk of type 2 diabetes. *Hum. Mol. Genet.* **19,** 535–544 (2010).

35. Kannel, W. B. Lipids, diabetes, and coronary heart disease: insights from the Framingham Study. *Am. Heart J.* **110,** 1100–1107 (1985).

36. Femlak, M., Gluba-Brzózka, A., Ciałkowska-Rysz, A. & Rysz, J. The role and function of HDL in patients with diabetes mellitus and the related cardiovascular risk. *Lipids Health Dis.* **16,** 207 (2017).

37. Haase, C. L., Tybjærg-Hansen, A., Nordestgaard, B. G. & Frikke-Schmidt, R. HDL cholesterol and risk of type 2 diabetes: A mendelian randomization study. *Diabetes* **64,** 3328–3333 (2015).

38. Wannamethee, S. G., Shaper, A. G., Lennon, L. & Morris, R. W. Metabolic syndrome vs Framingham Risk Score for prediction of coronary heart disease, stroke, and type 2 diabetes mellitus. *Arch. Intern. Med.* **165,** 2644–2650 (2005).

39. Micha, R. & Mozaffarian, D. Trans fatty acids: effects on metabolic syndrome, heart disease and diabetes. *Nat. Rev. Endocrinol.* **5,** 335–344 (2009).

40. Aubert, R. *Diabetes in America*. (DIANE Publishing, 1995).

41. Sebastiani, P. *et al.* Biomarker signatures of aging. *Aging Cell* **16,** 329–338 (2017).

42. Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological Age Predictors. *EBioMedicine* **21,** 29–36 (2017).

43. 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).

44. 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).

45. 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).

46. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking

behavior. *Nat. Genet.* **42,** 441–447 (2010).

47. 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).
48. 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).
49. 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).
50. 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).
51. 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).
52. 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).
53. Kilpeläinen, T. O. *et al.* Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nat. Commun.* **7,** 10494 (2016).
54. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).
55. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).
56. 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).
57. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* **19,** 1085–1094 (2014).
58. 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).
59. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518,** 187–196

(2015).

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

61. 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).

62. 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).

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

Acknowledgements:

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