Prediction of Type 2 Diabetes Mellitus Onset Using Logistic Regression-Based Scorecards

Yochai Edlitz\textsuperscript{1,2}, Eran Segal\textsuperscript{1,2,\textdagger}

\textsuperscript{1}Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.
\textsuperscript{2}Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.
\textsuperscript{\textdagger}Corresponding author

1. Abstract

1.1 Background

Type 2 diabetes (T2D) accounts for \textasciitilde90\% of all cases of diabetes, resulting in an estimated 6.7 million deaths in 2021, according to the International Diabetes Federation (IDF). Early detection of patients with high risk of developing T2D 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 or access to sophisticated computational resources, there is a need for accurate yet accessible prediction models.

1.2 Methods

In this study, we analyzed data from 44,709 non-diabetic U.K. Biobank participants aged 40-69, predicting the risk of T2D onset within a selected timeframe (mean of 7.3 years with a standard deviation of 2.3 years). We started with 798 features that we identified as potential predictors for T2D onset. We first analyzed the data using gradient boosting decision trees, survival analysis, and logistic regression methods. We devised one non-laboratory model accessible to the general population and one more precise yet simple model that utilizes laboratory tests. We simplified both models to an accessible scorecard form, tested the models on normoglycemic and prediabetes sub cohorts, and compared the results to the results of the general cohort. We established the non-laboratory model using the following covariates: sex, age, weight, height, waist size, hip circumference, waist-to-hip Ratio (WHR), and Body-Mass Index (BMI). For the laboratory model, we used age and sex together with four common blood tests: HDL (high-density lipoprotein), gamma-glutamyl transferase, glycated hemoglobin, and triglycerides. As an external validation dataset, we used the electronic medical record database of Clalit Health Services.

1.3 Results

The non-laboratory scorecard model achieved an Area Under the Receiver Operating Curve (auROC) of 0.81 (0.77-0.84 95\% Confidence Interval (CI)) and an odds ratio (OR) between the upper and fifth prevalence deciles of 17.2 (5-66 95\% CI). Using this model, we classified three risk groups, a group with 1\% (0.8-1\%), 5\% (3-6\%), and the third group with a 9\% (7-12\%) risk of developing T2D. We further analyzed the contribution of the laboratory-based model and devised a blood-test model based on age, sex and the four common blood tests noted above. In this scorecard model, we included age, sex, glycated hemoglobin (HbA1c\%), gamma glutamyl-transferase, triglycerides, and HDL cholesterol. Using
this model, we achieved an auROC of 0.87 (0.85-0.90 95% CI) and a deciles' OR of x48 (12-109 95% CI). Using this model, we classified the cohort into four risk groups with the following risks: 0.5% (0.4%-7%); 3% (2-4%); 10% (8-12%) and a high-risk group of 23% (10-37%) of developing T2D. When applying the blood tests model using the external validation cohort (Clalit), we achieved an auROC of 0.75 (0.74-0.75 95% CI). We analyzed several additional comprehensive models, which included genotyping data and other environmental factors. We found that these models did not provide cost-efficient benefits over the four blood test model. The commonly used German Diabetes Risk Score (GDRS) and Finnish Diabetes Risk Score (FINDRISC) models, trained using our data, achieved an auROC of 0.73 (0.69-0.76) and 0.66 (0.62-0.70), respectively, inferior to the results achieved by the four blood test model and by the Anthropometry models.

1.4 Conclusions

The four blood tests and anthropometric models outperformed the commonly used non-laboratory models, the FINDRISC and the GDRS. We suggest that our models be used as tools for decision-makers to assess populations at elevated T2D risk and thus improve medical strategies. These models might also provide a personal catalyst for changing lifestyle, diet, or medication modifications to lower the risk of T2D onset.

1.5 Funding

No Funders

The funders had no role in study design, data collection, interpretation, or the decision to submit the work for publication.

2. Introduction

Diabetes mellitus is a group of diseases characterized by symptoms of chronic hyperglycemia and is becoming one of the world's most challenging epidemics. The prevalence of Type Two Diabetes (T2D) has increased from 4.7% in 1980 to 10% in 2021, and is considered the cause of an estimated 6.7 million deaths in 2021 (IDF Diabetes Atlas). T2D is characterized by insulin resistance, resulting in hyperglycemia, and accounts for ~90% of all diabetes cases (Zimmet et al., 2016).

In recent years, the prevalence of Diabetes has been rising more rapidly in low and middle-income countries (LMICs) than in high-income countries (“WHO | Diabetes programme,”). In 2019 Eberhard et al. estimated that every other person with diabetes in the world is undiagnosed (Standl et al., 2019). 83.8% of all cases of undiagnosed diabetes are in low-mid income countries (Beagley et al., 2014), and according to the IDF Diabetes Atlas, over 75% of adults with diabetes live in low to middle-income countries (IDF Diabetes Atlas), where laboratory diagnostic testing is limited (Wilson et al., 2018).

According to several studies, a healthy diet, regular physical activity, maintaining normal body weight, and avoiding tobacco use can prevent or delay T2D onset (“WHO | Diabetes programme”; “Home | ADA,”; Knowler et al., 2002; Lindström et al., 2006; Diabetes Prevention Program Research Group, 2015). 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, making it accessible for populations having difficulties using the computer.

Several such tools are in use today (Noble et al., 2011; Collins et al., 2011; Kengne et al., 2014). The Finnish Diabetes Risk Score (FINDRISC), a commonly used, non-invasive T2D risk-score model, estimates the risk of patients between the ages of 35 and 64 of developing T2D within ten years. The FINDRISC was created based on a prospective cohort of 4,746 and 4,615 individuals in Finland in 1987 and 1992, respectively. The FINDRISC model employs 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 can be used as a scorecard model or a logistic regression (LR) model (Bernabe-Ortiz et al., 2018; Lindström and Tuomilehto, 2003; Meijnikman et al., 2018).

Another commonly used scorecard prediction model is the German Diabetes Risk Score (GDRS), which estimates the five-year risk of developing T2D. The GDRS is based on 9,729 men and 15,438 women between the ages of 35-65 from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study (“EPIC Centres - GERMANY,” n.d.). The GDRS is a Cox regression model using age, height, waist circumference, the prevalence of hypertension (yes/no), smoking behavior, physical activity, moderate alcohol consumption, coffee consumption, intake of whole-grain bread, intake of red meat, and parent and sibling history of T2D (Schulze et al., 2007; Mühlenbruch et al., 2014).

Barbara Di Camillo et al. reported in 2019 the development of three survival analysis models using the following features: background and anthropometric information, routine laboratory tests, and results from an Oral Glucose Challenge Test (OGTT). The cohorts consisted of 8483 people from three large Finnish and Spanish data sets. They report achieving auROC scores equal to 0.83, 0.87, and 0.90, outperforming the FINDRISC and Framingham scores (Di Camillo et al., 2018). In 2021 Lara Lama et al. reported using a random forest classifier on 7949 participants from the greater Stockholm area to investigate the key features for predicting prediabetes and T2D onset. They found that body mass index, waist-hip ratio, age, systolic and diastolic blood pressure, and a family history of diabetes were the most significant predictive features for T2D and prediabetes (Lama et al., 2021).

The goal of the present research is to develop easy-to-use, clinically usable models that are highly predictive of T2D onset. We developed two simple scorecard models and compared their predictive power to the established FINDRISC and GDRS models. We trained both models using a subset of data from the U.K. Biobank (UKB) observational study cohort and reported the results using holdout data from the same study. We based one of the models on easily accessible anthropometric measures and the other on four common blood tests. Since we trained and evaluated our models using the UKB database, the models are therefore most relevant for the U.K. population aged 40-65 or for populations with similar characteristics (as presented in Table 1). As an external test case for the four blood tests model, we used the Israeli electronic medical record database of Clalit Health Services (Feldman et al. 2014).
3. Methods

3.1. Data

We analyzed UKB’s observational data of 502,536 participants aged 40-69 recruited in the U.K. voluntarily from 2006 to 2010. During a baseline assessment visit to the U.K. Biobank, the participants self-completed questionnaires, which included lifestyle and other potentially health-related questions. The participants also underwent physical and biological measurements. Out of this cohort, we used the data of 20,346 participants who revisited the U.K. Biobank assessment center from 2012 to 2013 for an additional medical assessment. We also used the data of 48,705 participants who revisited for a second or third visit from 2014 onwards for an imaging visit and underwent an additional, similar medical check. We screened the participants to keep only those not being treated for nor having in the past T2D. We also screened out participants whose average blood sugar level for the past two to three months (hemoglobin A1c - HbA1c%) was below 4% or above 6.5%. We started with 798 features for each participant and removed all the features with more than 50% missing data points in our cohort. We later screened out all the participants who still had more than 25% missing data points from the cohort and imputed the remaining missing data. We further removed those study participants who self-reported as being healthy but had HbA1c% levels higher than the accepted healthy level. We also screened participants who had a record of a prior T2D diagnosis (Data field 2976 at the UKB). As not all participants had HbA1c% measurements, we estimated the bias of participants reporting as healthy while having an HbA1c% level indicating diabetes. For this estimate, we used the data from a subpopulation of our patients and found that 0.5% of participants reported being healthy with a median HbA1c% value of 6.7%, while the cutoff for having T2D is 6.5%. (Table 1).

Of the remaining 44,709 participants in our study cohort, 1.79% developed T2D during a follow-up period of 7.3±2.3 years (Table 1, Figure 1A). As a predicted outcome, we used the data for whether a participant develops T2D between the first and last visit from a self-report using a touchscreen questionnaire. The participants were asked to mark either - “Yes”/“No”/“Do not know”/“Prefer not to answer” for the validity of the sentence “Diabetes diagnosed by a doctor,” which was presented to them on a touch screen questionnaire (Data-field 2443 at the UKB).

Table 1 | Cohort statistical data

<table>
<thead>
<tr>
<th></th>
<th>UKB population</th>
<th>,Train validation, and test sets</th>
<th>Test set</th>
<th>Train set</th>
<th>Validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>502536</td>
<td>44709</td>
<td>8960</td>
<td>25025</td>
<td>10724</td>
</tr>
<tr>
<td>Age at first visit [Years]</td>
<td>56.5±8.1</td>
<td>55.6±7.6</td>
<td>55.5±7.5</td>
<td>55.6±7.6</td>
<td>55.6±7.6</td>
</tr>
<tr>
<td>Age at last visit [Years]</td>
<td>-</td>
<td>62.9±7.5</td>
<td>62.9±7.4</td>
<td>62.9±7.5</td>
<td>62.9±7.5</td>
</tr>
<tr>
<td>The time between visits [Years]</td>
<td>-</td>
<td>7.3±2.3</td>
<td>7.3±2.3</td>
<td>7.3±2.3</td>
<td>7.3±2.3</td>
</tr>
<tr>
<td>Males in the population [%]</td>
<td>45.5</td>
<td>47.8</td>
<td>47.9</td>
<td>47.9</td>
<td>47.5</td>
</tr>
<tr>
<td>% Diabetic at first visit [%]</td>
<td>4.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% Diabetic at last visit [%]</td>
<td>-</td>
<td>1.79</td>
<td>1.76</td>
<td>1.75</td>
<td>1.91</td>
</tr>
<tr>
<td>Hba1c at first visit [%]</td>
<td>5.5±0.6</td>
<td>5.3±0.3</td>
<td>5.3±0.3</td>
<td>5.3±0.3</td>
<td>5.3±0.3</td>
</tr>
<tr>
<td>Hba1c at last return [%]</td>
<td>-</td>
<td>5.4±0.4</td>
<td>5.4±0.3</td>
<td>5.4±0.4</td>
<td>5.4±0.4</td>
</tr>
<tr>
<td>Weight at first visit [Kg]</td>
<td>78.1±15.9</td>
<td>76.6±14.7</td>
<td>76.4±14.6</td>
<td>76.7±14.7</td>
<td>76.8±14.9</td>
</tr>
<tr>
<td>Weight at last visit [Kg]</td>
<td>-</td>
<td>76.2±15.2</td>
<td>76.0±14.9</td>
<td>76.2±15.2</td>
<td>76.5±15.3</td>
</tr>
<tr>
<td>BMI at first visit [Kg/m^2]</td>
<td>27.4±4.8</td>
<td>26.6±4.2</td>
<td>26.5±4.1</td>
<td>26.6±4.2</td>
<td>26.7±4.3</td>
</tr>
<tr>
<td>BMI at last visit [Kg/m^2]</td>
<td>-</td>
<td>26.6±4.4</td>
<td>26.5±4.3</td>
<td>26.6±4.4</td>
<td>26.7±4.5</td>
</tr>
</tbody>
</table>
Table 1 | Characteristics of this study’s cohort population and the UKB population. A “±” sign denotes the standard deviation. While T2D prevalence in the UKB participants is 4.8%, it is 1.79% in our cohort as we screened the cohort at baseline for HbA1c% levels < 6.5%. The age range of the participants at the first visit was 40-69; thus, our models are not suitable for people who develop T2D at younger ages. The models predict the risk of developing T2D between the first visit to the UKB assessment center and the last visit. We refer to this feature 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 of the selection process of participants in this study. We selected participants who came for a repeated second or third visit 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% for the training and validation sets and 20% for the hold-out test set. We excluded an additional 2,285 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, 4) found to be diagnosed with T2D before the first U.K. biobank visit. The final training, validation, and test sets included 25,025 participants (56% of the cohort), 10,724 participants (24%), and 8,960 participants (20%), respectively.

B. The process flow during the training and testing of the models. We first split the data and kept a hold-out test set. We then explored several models using the training and validation data sets. We then compared the selected models using the hold-out test set and reported the results. We calibrated the output of the models to predict the probability of a participant developing T2D.

3.2. Feature selection process

We started with 798 features that we hypothesized as potential predictors for T2D onset. We removed all the features with 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 a genetic input for several models, we used both Polygenic Risk Scores (PRS) and Single-Nucleotide-Polymorphisms (SNPs) from the UKB SNP array (See methods).
We used forty-one PRSs (Polygenic Risk Scores) 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 removing duplicate SNPs, we remained with 2267 SNPs (See “Genetic Data” in methods).

We aggregated the features into thirteen separate groups: age and sex; genetics; early life factors; socio-demographic; mental health; blood pressure and heart rate; family history and ethnicity; medication; diet; lifestyle and physical activity; physical health; anthropometry; and blood test results. We trained models for each group of features separately (Appendix 1 Figure 1, Appendix 1 Table 1). We then added the features groups according to their marginal predictability (Appendix 1 Table 2).

After selecting the leading models from the training and validation data sets, we tested and reported the results of the selected models using the hold-out test set samples (Appendix 1 Table 1). To encourage clinical use of our models, we optimized the number of features the models require. To simplify our models, we iteratively removed the least contributing features of our models using the training data set (See “Missing data” in methods, Appendix 1 Figure1). We examined the normalized coefficient of each model feature to assess its importance in the model. For the four blood tests model, we initially also had “reticulocytes” as one of the model’s features. As we want to use common blood tests only, we dropped this feature from the list after confirming that the impact of removal of this feature on the model accuracy was negligible. Once the models were finalized, we developed corresponding scorecards that were both simple and interpretable (See “scorecards creation” in methods).

### 3.3. Outcome

Our models provide a prediction score for the participant’s risk of developing T2D during a specific timeframe. The mean prediction timeframe in our cohort is 7.3±2.3 years. The results that we report correspond to a hold-out test-set comprising 20% of our cohort that we kept aside until the final report of the results. We also report the results of the four blood tests model using an external electronic medical record database of the Israeli Clalit Health Services. We trained all the models using the same training set, and we then reported the test results of the hold-out test set. We use the area under the receiver operating curve (auROC) and the average precision score (APS) as the main metrics of our models. Using these models, a physician can inform patients about their risk of developing T2D and their predicted risk of developing T2D within a selected timeframe.

We calibrated the models to report the probability of developing T2D during a given timeframe (See calibration in methods). To quantify the risk groups in the scorecards model, we performed a bootstrapping process on our validation data set like the one performed for the calibration. We selected boundaries that showed good separation between risk groups, and we reported the results using the hold-out test set.

### 3.4. Missing data

After removing all features with more than 50% missing data and removing all the participants with more than 25% missing features, we imputed the remaining data. We analyzed the correlations between predictors with missing data and found correlations within anthropometry group features to other features in the same domain — analogous correlations were found in the blood test data. We used SKlearn’s iterative imputer with a maximum of ten iterations for the imputation and tolerance of 0.1 (Alex et al. 2011) We imputed the training and validation sets apart from the imputation of the hold-out test-set. We did not perform imputation on the categorical features but transformed them into “one hot encoding” vectors with a bin for missing data using Pandas categorical tools.
3.5. Genetic data

We use polygenic risk scores (PRS) and single-nucleotide-polymorphisms (SNPs) as genetic input for some models. We calculated the PRS by summing the top correlated risk allele effect sizes derived from Genome-Wide Association Studies (GWAS) summary statistics. We first extracted from each summary statistics the top 1000 SNPs according to their p-value. We then used only the SNPs presented 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 models. After the removal of duplicated SNPs, we kept 2267 SNPs as features. The full PRS summary statistics list can be found in Appendix 1, “References for PRS summary statistics articles.” We calculated the PRS scores according to summary statistics publicly available from studies not derived from the U.K. Biobank to prevent data leakage.

3.6. Baseline models

As the reference models for our results, we used the well-established FINDRISC and GDRS models (Lindström and Tuomilehto, 2003; Schulze et al., 2007; Mühlenbruch et al., 2014), which we retrained and tested on the same data used for our models. These two models are based on Finnish and German populations with similar age ranges as our cohort. We derived a FINDRISC score for each participant using the data for age, sex, body mass index (BMI), waist circumference, and blood pressure medication provided by the UKB. To calculate the score for duration of physical activity, 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 of 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 to 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 of participants’ grandparents, aunts, uncles, first cousins, and children. We also lacked the data about past blood pressure medication, although we do 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 examined 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 when this additional parameter was used. We report here the results for the FINDRISC model without the time of the second visit as a feature.

To derive the GDRS based model, we built a Cox regression model using Python’s lifelines package (Davidson-Pilon et al., 2020). As for the features of the GDRS model, we incorporated the following features: years between visits; height; prevalence of 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 hyperparameter search the same way we used for our models. The hyperparameters we used here are the penalize parameter in the range of 0-10 using a 0.1 resolution and a variance threshold of 0-1 with 0.01 resolution. This last hyperparameter is used to drop columns where the variance of the column was lower than the variance threshold.
3.7. Model building procedures

To test overfitting and biased models, we split the data into three groups: 20% for the holdout test set, used only for the final reporting of results. For the remaining data we used 30% for the validation set and 70% for the training set. We use a two-stage process to evaluate the models’ performance: exploration and test phases (Figure 1, Appendix 1 Figure 1). We selected the optimal features during the exploration stage using the training and validation data sets. We then performed two-hundred iterations of a random hyperparameters selection process for each group of features. We set the selection criterion as the auROC metric using five-fold cross-validation.

We used the validation data set to rank the various models by their auROC scores. We trained each of the models, using the full training set with the top-ranked hyper-parameters determined from the hyperparameters tuning stage and we ranked the models by their score using the validation data set.

During the test phase, we reported the results of our selected models. For this, we evaluated the selected models using the holdout test set. To do so, we reran the hyperparameters selection process using the integrated training and validation data sets. We evaluated the trained model based on the data from the holdout test set. The same data sets were used for all the models.

For the development of the Cox regression models, we used the lifelines survival analysis package (Davidson-Pilon et al., 2020), using the “age diabetes diagnosed” category (data field 2976) as a label. We used SKlearn’s LogisticRegressionCV model for the logistic regression model’s computation (Alex et al. 2011). For the Gradient Boosting Decision Trees (GBDT) models, we used Microsoft’s LightGBM package (Ke et al., 2017). We developed our data pipeline to compute the scorecards. These last three models used the “diabetes diagnosed by a doctor” category of the U.K. biobank as a label (data field 2443).

As part of the models’ calculation process, we used two-hundred iterative random hyperparameters searches for the training of the models. For the GBDT models we used the following parameter values for the searches: 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] (Ke et al., 2017). We used a penalize in the range 0 to 2 with 0.02 resolution for the l2 penalty during the hyperparameters searches for the logistic regression models.

We composed an anthropometric-based scorecard model to provide an accessible, simple, non-laboratory, and non-invasive T2D prediction model. In this model, a patient can easily mark the result in each of the scorecard 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).

In addition, we developed a more accurate tool for predicting T2D onset for those cases where laboratory testing will be available. We started with 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 thirteen categories such as lifestyle, diet, and anthropometrics (Appendix Table 1, Appendix Table 2). We concluded that the blood tests have higher predictability than the other features aggregations. We thus trained a full blood test model using 59 blood tests available in the training dataset. Applying a recursive features’ elimination process to the top 10 predictive features, we established the features of our final model based on four blood tests.

We used the validation data set to rank the various models by their auROC scores. We trained each of the models, using the full training set with the top-ranked hyper-parameters determined from the hyperparameters tuning stage and we ranked the models by their score using the validation data set.

During the test phase, we reported the results of our selected models. For this, we evaluated the selected models using the holdout test set. To do so, we reran the hyperparameters selection process using the integrated training and validation data sets. We evaluated the trained model based on the data from the holdout test set. The same data sets were used for all the models.

For the development of the Cox regression models, we used the lifelines survival analysis package (Davidson-Pilon et al., 2020), using the “age diabetes diagnosed” category (data field 2976) as a label. We used SKlearn’s LogisticRegressionCV model for the logistic regression model’s computation (Alex et al. 2011). For the Gradient Boosting Decision Trees (GBDT) models, we used Microsoft’s LightGBM package (Ke et al., 2017). We developed our data pipeline to compute the scorecards. These last three models used the “diabetes diagnosed by a doctor” category of the U.K. biobank as a label (data field 2443).

As part of the models’ calculation process, we used two-hundred iterative random hyperparameters searches for the training of the models. For the GBDT models we used the following parameter values for the searches: 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] (Ke et al., 2017). We used a penalize in the range 0 to 2 with 0.02 resolution for the l2 penalty during the hyperparameters searches for the logistic regression models.

We composed an anthropometric-based scorecard model to provide an accessible, simple, non-laboratory, and non-invasive T2D prediction model. In this model, a patient can easily mark the result in each of the scorecard 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).

In addition, we developed a more accurate tool for predicting T2D onset for those cases where laboratory testing will be available. We started with 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 thirteen categories such as lifestyle, diet, and anthropometrics (Appendix Table 1, Appendix Table 2). We concluded that the blood tests have higher predictability than the other features aggregations. We thus trained a full blood test model using 59 blood tests available in the training dataset. Applying a recursive features’ elimination process to the top 10 predictive features, we established the features of our final model based on four blood tests.
The four blood tests that we used are the glycated hemoglobin 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; Gamma-Glutamyl Transferase Test (GGT); 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 related to the population’s prevalence. We computed the values of these features’ associated coefficients with their 95%CI to reconstruct the models (Figure 3E).

We tested the models in mixed and stratified populations of 1,006 prediabetes participants with a T2D prevalence of 9.4% and a separate 7,948-normoglycemic participants with a T2D onset prevalence of 0.8% (See Table 4).

3.8. Shapley Additive Explanations (SHAP)
We used the SHAP method, which approximates Shapley values, for the feature importance analysis of the GBDT model. This method originated in game theory to explain the output of any machine-learning model. SHAP approximates the average marginal contributions of each model feature across all permutations of the other features in the same model (Lundberg and Lee, 2017).

3.9. Predictors
To estimate the contribution of each feature’s domain and for the initial screening of features, we built 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) designed to find correlations between known genetic variants and a phenotype of interest. We used only GWASs from outside the UKB population to avoid data leakage (See supplementary material Appendix 1, “References for PRS summary statistics articles”).

We trained and tested the full-features model using the training and validation cohort to select the most predictive features for the anthropometry and the blood-tests models. We then used this model’s feature importance to extract the most predictive features. We omitted data concerning family relatives with T2D from the model, as we did not see any major improvement over the anthropometrics model. For the last step, we tested and reported the model predictions using data in the holdout section of the cohort.

For the extraction of the four blood tests model, we performed a features selection process using the training and validation datasets. We executed models starting with twenty and down to four features of blood tests together with age and sex as features, each time removing the feature with the smallest importance score. We then selected the model with four blood tests (HbA1c%, Gamma glutamyl transferase (GGT), triglycerides, HDL cholesterol) plus age and sex as the optimal balance between model simplicity (a small number of features) and model accuracy. We reported model results against data from the holdout test set.

We normalized all the continuous predictors using the standard “z-score”. We normalized the train-validation sets apart from the holdout test set to avoid data leakage.

3.10. Model calibration
Calibration refers to the concurrence between the real T2D onset occurrence in a subpopulation and the predicted T2D onset probability in this population. Since our data are highly imbalanced between healthy and T2D ill patients, with a prevalence of 1.79% T2D, we used one thousand bootstrapping iterations of each model to improve the calibration. To do this, we first split each model’s prediction into ten deciles bins from zero to
one to calculate the calibration curves. Using Sklearn's isotonic regression calibration, we scale the results with a five-fold cross-validation (Alex et al. 2011). We do so for each of the bootstrapping iterations. Lastly, we concatenate all the calibrated results and calculate each probability decile's overall mean predicted probability. We split the probabilities range (0-1) into deciles (Figure 3F) and assigned each prediction sample to a decile bin according to the calibrated predicted probability of T2D onset.

3.11. Scorecards creation

We used the training and validation data sets for our scorecards building process. We reported the results on the holdout data set. We calculated our data's weight of evidence (WoE) by splitting each feature into bins. We binned greater importance features in a higher resolution while maintaining a monotonically increasing WoE (Kinden 2019). For quantizing the risk groups of the scorecards model, we performed one thousand iterations of a bootstrapping process on our validation data set. We considered several potential risk score limits that separate T2D onset probability in each score group using the validation data set. Once we set the final boundaries of the score, we reported the prevalence in each risk group on the test set. For the Cox regression based scorecards, we used the parameters coefficient the same way we used the coefficient in the logistic regression model for binning the model. When using a Cox-regression based scorecard, we compute the probability to develop based on a fix timeframe for all participants (5 and 10 years’ timeframes models) (Table 3). To enable an easy way for choosing the desired timeframe as part of scorecard usage, we chose to use the logistic regression based scorecards as our model of choice for an additional development and validation.

3.12. External validation cohort - EHR database of Clalit Health Services

As an external validation cohort for the four blood tests scorecard model, we used the Clalit retrospective cohort's electronic health records. Clalit is the largest Israeli healthcare organization, serving more than 4.4 million patients (about half the population of Israel). The Clalit database holds electronic health records of over 11 million patients, dating back to 2002. It is considered one of the world’s largest EHR databases (Artzi et al., 2020). We extracted data from patients who visited Clalit clinics from 2006 to 2011 and had a minimum of three HbA1C% tests, with the following inclusion criteria: the first sample below 6.5%, and two consecutive tests consistent with either HbA1c%>=6.5 for each of the tests or both tests with HbA1C%<6.5%. These criteria were some of the criteria we used to indicate if the patient had developed T2D. We started with 179,000 patients that met the HbA1c% criteria noted above. We then included data from the following tests: GGT (80,000 patients), HDL (151,969 patients), and triglycerides (157,321 patients). In addition to the HbA1c% exclusion criteria, we added the following: patients who did not have all four blood tests; patients older than 70 or younger than 40; patients who were diagnosed with T2D before the first visit; patients who had a diabetes diagnosis without a clear indication that it was T2D; patients who had taken diabetes-related drugs (ATC code A10) before the first visit or before being diagnosed with T2D either based on HbA1c% levels or by a physician.

As a criterion for T2D, we considered two consecutive tests with HbA1c>=6.5% or a physician diagnosis of T2D. After excluding patients according to the above criteria, the remaining cohort included 17,132 patients with anthropometric characteristics similar to the U.K. Biobank cohort (Table 2). The remaining cohort’s T2D onset prevalence is 4.1%, considerably greater than the 1.79% in the U.K. Biobank cohort. We further tested the model on a stratified normoglycemic sub cohort with 10,064 patients and a prevalence of 2% T2D and a prediabetes sub cohort with 7,059 patients with 7.1% T2D prevalence.

We tested the four blood tests model on the data from the above cohorts by calculating a raw score for each
participant based on all relevant scorecard features apart from the time “years for prediction” feature. We then randomly sampled out of a normal distribution resembling the UKB cohort (mean=7.3 years, std=2.3 years) one thousand time periods for a returning visit for each participant. We limited the patients’ time of returning between 2 and 17 years to emulate the U.K. Biobank data. The cut-off data for last visit was December 31, 2019, the last date reported in the Clalit database. We then estimated the mean and 95%CI of these cohort’s auROC and APS results.

We did not evaluate the FINRISC, GDRS, and Anthropometrics models using the Clalit database as these models required some features that do not appear in the Clalit database. The FINRISC model requires data regarding physical activity, waist circumference, and consumption of vegetables, fruit, or berries. The GDRS requires the following missing data fields: Physical activity, waist circumference, consumption of whole-grain bread/rolls and muesli, consumption of meat, and coffee consumption. The Anthropometrics model requires data regarding waist and hips circumference.

Table 2 | External validation cohort (“Clalit”) statistical data

<table>
<thead>
<tr>
<th>males</th>
<th>HbA1c%</th>
<th>GGT</th>
<th>Reticulocyte Count</th>
<th>HDL</th>
<th>Triglycerides</th>
<th>Age</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean value</td>
<td>45%</td>
<td>5.56</td>
<td>32.31</td>
<td>56.33</td>
<td>49.77</td>
<td>141.33</td>
<td>56.40</td>
<td>79.00</td>
<td>1.66</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.41</td>
<td>49.29</td>
<td>36.97</td>
<td>13.33</td>
<td>82.09</td>
<td>8.06</td>
<td>49.90</td>
<td>0.09</td>
<td>19.58</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>5.30</td>
<td>17.00</td>
<td>38.35</td>
<td>40.00</td>
<td>90.00</td>
<td>50.14</td>
<td>67.00</td>
</tr>
<tr>
<td>5.60</td>
<td>23.00</td>
<td>58.00</td>
<td>48.00</td>
<td>123.00</td>
<td>57.02</td>
<td>77.00</td>
<td>1.65</td>
<td>27.68</td>
<td></td>
</tr>
<tr>
<td>5.90</td>
<td>33.00</td>
<td>78.60</td>
<td>57.00</td>
<td>170.00</td>
<td>87.82</td>
<td>1.72</td>
<td>31.25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Results

4.1. Anthropometric based model

Using the Anthropometrics scorecard model the patient’s final score relates to three risk groups (see Model building procedures in Methods). Participants within the score range between 1 and 69 has a 1% [0.8-1% 95% CI] probability of developing T2D. The second group, with a score range between 70 and 78, predicts a 5% [3-6% 95%CI] probability of developing T2D. The third group, with a score range of between 79 and 96, predicts a 9% [7-12% 95%CI] probability of developing T2D (Figure 2C).

We also provide models with the same features in their logistic regression form and a Cox regression form for more accurate computer-aided results. Testing these models using the holdout test set achieved an area under the receiver operating curve (auROC) of 0.81 (0.78-0.84) and an average precision score (APS) of 0.09 (0.06-0.13) at 95% CI. Applying a survival-analysis Cox-regression model to the same features resulted in comparable results (Table 3). Using the model in its scorecard form, we achieved an auROC of 0.81 (0.77-0.84) and an APS of 0.07 (0.05-0.10). All these models outperformed the two models that we used as a reference: applying the FINDRISC model resulting in an auROC of 0.73 (0.69-0.76) and an APS of 0.04 (0.03-0.06), and applying the GDRS model resulting in an auROC of 0.66 (0.62-0.70) and an APS of 0.04 (0.03-0.06) (Figure 3A-B, table 3, and Methods). With the cohort’s baseline prevalence of 1.79%, the Cox regression model achieved deciles’ odds ratio (OR) of x10.65 (4.99-23.67), the L.R (Logistic Regression). Anthropometric model achieved deciles’ OR of x16.9 (4.84-65.93), and its scorecard derivative achieved deciles OR of x17.15 (5.65-97) compared to the FINDRISC model’s x4.13 (2.29-7.37) and the x2.53 (1.46-4.45) deciles’ OR achieved by the GDRS model (Figure 3C, Table 3,
methods). The WHR and the BMI have the highest predictability in the anthropometric model (Figure 3D). These two body habitus measures are indicators associated with chronic illness (Eckel et al., 2005; Cheng et al., 2010; Jafari-Koshki et al., 2016; Qiao and Nyamdorj, 2010).

### 4.2. Model based on four blood tests

Using the four blood tests scorecard (Methods, Figure 2B), we binned the resulting scores into four groups. Participants with a score within the score range of between 1 and 104 has a 0.5% [0.4-0.7%, 95% CI] probability of developing T2D. The second group, with a score range of between 105 and 116 predicts a 3% [2-4%, 95% CI] probability of developing T2D. The third group score with a range of between 117 and 146 predicts a 10% [8-12%, 95% CI] of developing T2D. The fourth group score range was between 147 and 162, participants in this score range have a 23% [10-37% 95% CI] probability of developing T2D.

We used four common blood test scores as input to the survival analysis and the logistic regression model. Applying the survival analysis Cox regression model for the test set, we achieved an ROC of 0.88 (0.85-0.90), an APS of 0.25 (0.18-0.32), and a deciles OR of x42.9 (13.7-109.1). Using the four blood tests logistic regression model, we achieved comparable results: an auROC of 0.88 (0.85-0.91), an APS of 0.24 (0.17-0.31), and a deciles’ OR of 32.5 (10.8-110.1). Applying the scorecard model, we achieve an auROC of 0.87 (0.85-0.9), an APS of 0.13 (0.10-0.17), and a deciles’ OR of 47.7(79-115) (Figure 3 A-C, Table 3). The four blood tests model results are superior to those of the non-laboratory anthropometric model and those of the commonly used FINDRISC and GDRS models (Figure 3 A-C, Table 3).

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 HDL cholesterol, which is known to be beneficial for health, especially in the context of cardiovascular diseases, with high levels being negatively correlated to T2D onset. (Meijnikman et al., 2018; Kontush and Chapman, 2008; Bitzur et al., 2009) was found to be inversely correlated to the predicted probability of T2D onset. Interestingly, age and sex had a low OR value, meaning that they hardly contributed to the model, probably because of the T2D relevant information of these features latent within the blood tests data.

We compared these results to those of fifty-nine blood tests input features of the logistic regression model and those of a GBDT model, including thirteen features aggregations composed of two hundred seventy-nine individual features and genetics data available in the dataset. These two models achieved an auROC of 0.91 (0.89-0.93) and 0.91 (0.9-0.93), an APS of 0.26 (0.19-0.33) and 0.27 (0.20-0.34), and a deciles’ OR of x75.4 (17.74-133.45) and x72.6 (15.09-134.9), respectively.
Figure 2 | Anthropometrics and Blood tests scorecards

**A**

1) **Sex**

- Female: 0
- Male: 4

2) **Age at the end of timeframe**

- Age<57: 0
- Age>57 and <=64: 4
- Age>64 and <=69: 5
- Age>69: 7

3) **Timeframe for prediction**

- Timeframe <=5: 0
- Timeframe >5 and <=6: 8
- Timeframe >6 and <=8: 9
- Timeframe >8 and <=9: 11
- Timeframe >9 and <=10: 13
- Timeframe >10: 17

**B**

1) **HDL [mmol/L]**

- HDL<1.1: 29
- HDL>1.1 and <=1.3: 21
- HDL>1.3 and <=1.4: 16
- HDL>1.4 and <=1.5: 13
- HDL>1.5 and <=1.6: 9
- HDL>1.6 and <=1.9: 0

5) **Gamma glutamyltransferase [U/L]**

- GGT<14.8: 0
- GGT>14.8 and <=18: 15
- GGT>18 and <=21.3: 19
- GGT>21.3 and <=25.2: 21
- GGT>25.2 and <=30.3: 25
- GGT>30.3 and <=37.6: 30
- GGT>37.6 and <=53.1: 33
- GGT>53.1: 35

6) **Glycated hemoglobin [mmol/mol]**

- HbA1c<30.7: 0
- HbA1c>30.7 and <=32.4: 8
- HbA1c>32.4 and <=33.6: 11
- HbA1c>33.6 and <=34.6: 17
- HbA1c>34.6 and <=35.6: 22
- HbA1c>35.6 and <=36.8: 29
- HbA1c>36.8 and <=38.4: 35
- HbA1c>38.4: 54

7) **Triglycerides [mmol/L]**

- Triglycerides<0.89: 0
- Triglycerides>0.89 and <=1.2: 5
- Triglycerides>1.2 and <=1.4: 6
- Triglycerides>1.4 and <=1.8: 8
- Triglycerides>1.8 and <=2.4: 11
- Triglycerides>2.4: 14
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 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 four blood tests; an 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 prevalence ratio in each decile to the prevalence in the fifth decile. D. A feature importance graph of the logistic regression anthropometry model for a model with normalized features values. The bars indicate the feature importance values’ standard deviation (S.D.). 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 S.D. bars. While higher levels of HbA1c% positively contribute to T2D prediction, and HDL cholesterol levels are negatively correlated with the predicted probability of T2D, the information provided by age and sex relevant for predicting T2D onset is screened by other features. F. A calibration plot of the anthropometry, four blood tests, full blood test, and the FINDRISC models. Calibration of the models’ predictions allows reporting the probability of developing T2D (see Methods).

Table 3 | Comparing models’ main results

<table>
<thead>
<tr>
<th>Measure type</th>
<th>Model type</th>
<th>APS</th>
<th>auROC</th>
<th>Decile’s prevalence odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDRS</td>
<td>Score card cox regression for 5 years</td>
<td>0.04</td>
<td>0.66</td>
<td>2.5 (1.46-4.45)</td>
</tr>
<tr>
<td>FINDRISC</td>
<td>Score card logistic regression</td>
<td>0.04</td>
<td>0.73</td>
<td>4.13 (2.29-7.37)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Score card cox regression for 5 years</td>
<td>0.04</td>
<td>0.79</td>
<td>8.8 (3.6-36)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Score card cox regression for 10 years</td>
<td>0.06</td>
<td>0.79</td>
<td>10 (4.6-32.9)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Score card logistic regression</td>
<td>0.07</td>
<td>0.81</td>
<td>17.2 (5-66)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Logistic regression</td>
<td>0.09</td>
<td>0.81</td>
<td>16.9 (4.8-66)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Cox regression</td>
<td>0.10</td>
<td>0.82</td>
<td>10.7 (5-24)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Score card cox regression for 10 years</td>
<td>0.13</td>
<td>0.87</td>
<td>22.4 (9.8-54)</td>
</tr>
<tr>
<td>Four blood tests</td>
<td>Cox regression</td>
<td>0.25</td>
<td>0.88</td>
<td>43 (13.6-109)</td>
</tr>
<tr>
<td>Four blood tests</td>
<td>Score card cox regression for 5 years</td>
<td>0.09</td>
<td>0.89</td>
<td>53.2 (18.9-84.2)</td>
</tr>
<tr>
<td>Four blood tests</td>
<td>Cox regression</td>
<td>0.25</td>
<td>0.88</td>
<td>43 (13.6-109)</td>
</tr>
<tr>
<td>Four blood tests</td>
<td>Logistic regression</td>
<td>0.24</td>
<td>0.88</td>
<td>32.5 (10.89-110)</td>
</tr>
<tr>
<td>Blood tests</td>
<td>Logistic regression</td>
<td>0.26</td>
<td>0.91</td>
<td>75.4 (17.7-133)</td>
</tr>
<tr>
<td>All features</td>
<td>DT Boosting decision trees</td>
<td>0.27</td>
<td>0.91</td>
<td>72.6 (15.1-135)</td>
</tr>
</tbody>
</table>

Table 3 | The values in parentheses indicate a 95% CI. The deciles’ OR measures the ratio between T2D prevalence in the top risk score decile bin and the prevalence in the fifth decile bin (see Methods).
4.3. Prediction within an HbA1c% stratified population

To verify that our scorecard models can discriminate within a group of normoglycemic participants and within a group of pre-diabetic participants, we tested the models separately on each group. We separated the groups based on their HbA1c% levels during the first visit to the UKB assessment centers. We allocated participants with 4%<=HbA1c%<5.7% to the normoglycemic group and participants with 5.7%=<HbA1c%<6.5% levels to the pre-diabetic group (Cohen et al. 2010). 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 is only 0.8% versus a prevalence of 9.4% in the pre-diabetic group. The anthropometry model yielded an auROC of 0.81 (0.76-0.86) within the normoglycemic group with an APS of 0.04 (0.02-0.07). When testing the models within the pre-diabetic group, the anthropometry model achieved an auROC of 0.73 (0.68-0.77), and an APS of 0.2 (0.15-0.26). Both results outperform the FINDRISC and the GDRS results. For the normoglycemic HbA1c% range, the four blood tests model yielded an auROC of 0.81 (0.76-0.85) and an APS of 0.03 (0.02-0.05). These results are similar to those of the anthropometry model's results. To explore the option of developing scorecard models dedicated to these stratified populations, we developed and tested two such models, which achieved results similar to the mixed-cohort model (Table 4).
Table 4 | Comparing model results applied to an HbA1c% stratified population

<table>
<thead>
<tr>
<th>Population</th>
<th>Mixed population-based model. Tested on a stratified population</th>
<th>Models built using a stratified training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre diabetic (N=1006, Prevalence=9.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDRS</td>
<td>0.64 [0.57-0.70]</td>
<td>0.17 [0.12-0.23]</td>
</tr>
<tr>
<td>FINRISC</td>
<td>0.66 [0.61-0.72]</td>
<td>0.20 [0.14-0.27]</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>0.73 [0.68-0.77]</td>
<td>0.20 [0.15-0.26]</td>
</tr>
<tr>
<td>Four blood tests</td>
<td>0.73 [0.68-0.77]</td>
<td>0.20 [0.15-0.26]</td>
</tr>
</tbody>
</table>

| Normoglycemic (N=7948, Prevalence=0.8%) | | |
| GDRS       | 0.67 [0.61-0.74] | 0.02 [0.01-0.03] | - |
| FINRISC    | 0.74 [0.69-0.79] | 0.04 [0.02-0.07] | - |
| Anthropometry | 0.81 [0.76-0.86] | 0.04 [0.02-0.07] | 0.81 [0.76-0.85] | 0.03 [0.02-0.06] |
| Four blood tests | 0.81 [0.76-0.85] | 0.03 [0.02-0.05] | 0.82 [0.77-0.86] | 0.05 [0.03-0.09] |

Table 4 | The values in parentheses indicate 95% CI. Results of the models applied to a stratified population. The mixed population-based model column provides the results of the scorecard models presented in Figure 2 applied to normoglycemic and prediabetes subcohorts.

4.4. Validating the four blood tests model on an external independent cohort.

To validate our four blood tests model, we utilized the Israeli electronic medical record database of Clalit Health Services as an external cohort. Applying our model to non-diabetic participants of the same age range population (See methods), the four blood tests model achieved an auROC of 0.75 [0.74-0.75 95%CI] and an APS of 0.11 [0.10-0.11 95%CI] on a population of 17,132 participants with a 4.1% T2D onset prevalence. We then tested the model on stratified normoglycemic and prediabetes sub-cohorts. In the normoglycemic population (N=10,064 participants) with T2D onset prevalence of 2% the model achieved an auROC of 0.69 [0.66-0.69, 0.95%CI] and an APS of 0.04 [0.04-0.05, 95% CL]. Within the prediabetes population (N=7059) with a T2D onset prevalence of 7.1%, the model achieved an auROC of 0.68 [0.67-0.69, 95% CL] and an APS of 0.12 [0.12-0.13, 95% CL] (Table 5). These results validate the general applicability of our models applied to an external cohort. As this database lacks data required for the anthropometry, GDRS and FINRISC scorecards, we could not apply these models to the Clalit database (See “External validation cohort” in methods).

5. 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 scorecard
models: the anthropometric and the four 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 model on age, sex and six non-laboratory anthropometric measures. We then developed an additional, more accurate, straightforward model that can be used when laboratory blood test data are available. We based our second proposed model on four blood tests, in addition to the age and sex of the participants. We reported results of both models according to their scoring on survival-analysis Cox regression and logistic regression models. As these models require computer-aided analysis, we developed an easy-to-use scorecard form. For all 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). As a fair comparison, we trained these reference models and evaluated their predictions on the same data sets used with all our models.

Our models achieved a better auROC, APS, decile prevalence OR, and better-calibrated predictions than the FINDRISC and GDRS models. The anthropometrics and the four blood tests survival-analysis models achieved deciles prevalence OR of x10.7 and x42.9, respectively, while the scorecard forms achieved OR of x17.15 and x47.7, respectively.

The anthropometry-based model retained its auROC performance of 0.81[0.76-0.86] in the normoglycemic population but its performance worsened to 0.73[0.68-0.77] in the prediabetes population. The four blood tests model’s performance showed a similar trend in these two sub-cohorts [Table 4]. Training a sub-cohort-specific model did not improve these results.

Analyzing 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), body metrics that characterize body type or shape data. These features are known in the literature as being related to T2D, such conditions known as part of the metabolic syndrome (Eckel et al., 2005). The most predictive feature of the four blood tests model is the HbA1c%, which is a measure of the glycated hemoglobin carried by the red blood cells, often used to diagnose diabetes. Interestingly, age and sex had a very low feature importance value, implying that they hardly contributed to the model results. One potential explanation is that the T2D related information of these features is already latent within the blood test data. For instance, the sex hormone binding globulin (SHBG) feature contains a continuous measure regarding the sex hormone of each participant, thus making the sex feature redundant.

Applying the four blood tests model to the Clalit external cohort, we achieved an auROC of 0.75 [0.74-0.75]. While we obtained a sound prediction indication, the results are inferior to the scores when applied to the U.K. Biobank population. This sound prediction indication and degradation in performance are seen both in the general population and in the HbA1C% stratified sub-cohorts. We expected degradation in results when transitioning from the U.K. biobank to the Clalit cohort, as these two cohorts vary in many aspects. While the U.K. Biobank is a U.K. population-based prospective study suffering from “healthy volunteer selection bias” and from “attrition bias” (Fry et al., 2017; Hernán et al., 2004), the Clalit cohort is a retrospective cohort based on an Israeli population and suffers from ascertainment bias and diagnostic suspicion biases, as people with higher risk for T2D are sent to perform the related blood tests. In both studies, there is a need to handle missing data. In the Clalit database, we had to drop patients with inconclusive diagnoses (e.g., diabetes diagnosis, without referring to the type of diabetes) (See External validation cohort section in methods). One of the most apparent
differences is seen when comparing the T2D prevalence of the two cohorts: 1.79% for the U.K. biobank vs. 4.1% for the Clalit data base.

One main limitation of our study is that our cohorts’ T2D prevalence is biased away from the general U.K. populations’ T2D prevalence. Our cohort’s T2D prevalence was only 1.79%, while the general UK population's T2D prevalence is 6.3%, and 8% among adults aged 45-54 in the general UK population (2019) (Diabetes UK). This bias is commonly reported as a "healthy volunteer" selection bias (Fry et al., 2017; Hernán et al., 2004), which reduces the T2D prevalence from 6% in the general UK population to 4.8% in the UK Biobank (UKB) population. An additional screening bias is caused by including only healthy participants on their first visit. This contributed to the reduced prevalence of T2D in our cohort of 1.79% T2D onset. Applying our models to additional populations requires further research, and fine-tuning of the feature coefficients might be required.

As several studies have concluded (Knowler et al., 2002; Lindström et al., 2006; Diabetes Prevention Program Research Group, 2015), a healthy lifestyle and diet modifications are expected to reduce the probability of T2D onset; therefore, identifying people at risk for T2D is crucial. We assert that our models make a significant contribution to such identification in two ways: The laboratory four blood tests model for clinical use is highly predictive of T2D onset, and the anthropometrics model, mainly in its scorecard form, is an easily accessible and accurate tool. Thus, these models have the potential to improve millions of people’s lives and reduce the economic burden on the medical system.

6. Acknowledgments:
This research has been conducted using the U.K. Biobank Resource under Application Number 28784.
We thank I. Kalka, N. Bar, M.D Yotam Reisner, M.D, Ph.D. Smadar Shilo, Ph.D. E. Barkan, and members of the Segal group for discussions.

7. Code availability
Source code for analysis is available at https://github.com/yochaiedlitz/T2DM_UKB_predictions 1e6b22e

8. Data availability
We used two data sets, The U.K. Biobank data set and the Clalit Healthcare dataset. The two databases can be accessed upon specific requests and approval as described below.
UKBiobank - The U.K. Biobank data is available from the U.K. Biobank subject to standard procedures (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). The U.K. Biobank resource is open to all bonafide researchers at bonafide research institutes to conduct health-related research in the public interest. U.K. Biobank welcomes applications from academia and commercial institutes.
Clalit - The data that support the findings of the external Clalit cohort originate from Clalit Health Services (http://clalitresearch.org/about-us/our-data/). Due to restrictions, these data can be accessed only by request to the authors and/or Clalit Health Services. Requests for access to all or parts of the Clalit datasets should be addressed to Clalit Healthcare Services via the Clalit Research Institute (
The Clalit Data Access committee will consider requests given the Clalit data-sharing policy.

Bibliography


Qiao Q, Nyamador J. 2010. 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. doi:10.1038/ejcn.2009.93


Appendix 1

1. Exploring the full features’ space using gradient boosted decision trees.

To select model features, we analyzed the importance of features that we sought to relate to T2D. We analyzed the power of a predictive model with a vast amount of information and compared it to our minimal features models.

We started by sorting out a list of 279 preliminary features from the first visit to the UKB assessment center. In addition to these features, we used the UKB single-nucleotide polymorphisms (SNPs) genotyping data and its calculated PRS.

We inspected the impact of various features groups using the lightGBM (Ke et al., 2017) gradient decision trees model using SHAP (Lundberg et al., 2020; Lundberg and Lee, 2017) (Figure S1, See SHAP at methods). We aggregated the features into thirteen separate groups: age and sex; genetics; early life factors; socio-demographic; mental health; blood pressure and heart rate; family background and ethnicity; medication; diet; lifestyle and physical activity; physical health; anthropometry; and blood tests. All of these groups included age and sex features.

Appendix 1 Figure 1 - Models testing and training process

Appendix 1 Figure 1 | Models’ development | Scheme of the models’ exploration and evaluation process. For the models’ selection process, we used a fivefold CV with two hundred iterations of the random hyper-parameters process for each group of features. We then selected the top-scored hyper-parameters for each feature’s group. We trained a new model based on the training set, and measured the auROC using the validation set. Out of the validated models, we chose the models that had a minimal number of features and provided high-performance. The reported results are of the held-out test set.
We also tested the impact of HbA1c% with age and sex and genetics without age and sex. We list the top five predictive GBDT models, according to their auROC and APS, in descending order. “All features without genetic sequencing” model - auROC of 0.92 (0.89-0.94 95%CI) and APS of 0.28 (0.20-0.36 95% CI). Adding genetics to this model degraded the results to an auROC of 0.91(0.89-0.93) and an APS of 0.27 (0.2-0.34), probably due to overfitting. The “full-blood-tests” model - auROC of 0.90(0.88-0.93) and an APS of 0.28 (0.21-0.36). The “four blood tests” model - auROC of 0.88(0.85-0.90) and an APS of 0.20(0.14-0.17). The HbA1c% based model - auROC of 0.84(0.80-0.87) and an APS of 0.17 (0.12-0.23). The anthropometry model - auROC of 0.79 (0.75-0.82) and an APS of 0.07(0.05-0.11) (Appendix 1 table 1).

The lifestyle and physical activity model includes ninety-eight features related to physical activity; addictions; alcohol, smoking, cannabis use, electronic device use; employment; sexual factors; sleeping; social support; and sun exposure. This model achieved an auROC of 0.73(0.69-0.77), providing better prediction scores than the diet features group. The diet-based model includes thirty-two diet features from the UKB touchscreen questionnaire on the reported frequency of type and intake of common food and drink items. This model achieved an auROC of 0.66 (0.60-0.71).

**Appendix 1 Table 1. Predicting using feature domain groups. Results of GBDT models for various feature domains.**

<table>
<thead>
<tr>
<th>Label</th>
<th>APS</th>
<th>auROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features without genetic sequencing</td>
<td>0.28 (0.20-0.36)</td>
<td>0.92 (0.89-0.94)</td>
</tr>
<tr>
<td>All features</td>
<td>0.27 (0.20-0.34)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>All blood tests</td>
<td>0.28 (0.21-0.36)</td>
<td>0.90 (0.88-0.93)</td>
</tr>
<tr>
<td>Four blood tests</td>
<td>0.20 (0.14-0.27)</td>
<td>0.88 (0.85-0.90)</td>
</tr>
<tr>
<td>Blood tests without HbA1c%</td>
<td>0.13 (0.09-0.18)</td>
<td>0.84 (0.81-0.87)</td>
</tr>
<tr>
<td>HbA1c%</td>
<td>0.17 (0.12-0.23)</td>
<td>0.84 (0.80-0.87)</td>
</tr>
<tr>
<td>Blood tests without HbA1c% nor Glucose</td>
<td>0.10 (0.07-0.13)</td>
<td>0.82 (0.79-0.86)</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>0.07 (0.05-0.11)</td>
<td>0.79 (0.75-0.82)</td>
</tr>
<tr>
<td>Lifestyle and physical activity</td>
<td>0.05 (0.04-0.07)</td>
<td>0.73 (0.69-0.77)</td>
</tr>
<tr>
<td>Blood pressure and heart rate</td>
<td>0.05 (0.03-0.07)</td>
<td>0.69 (0.64-0.73)</td>
</tr>
<tr>
<td>Non diabetes related medication</td>
<td>0.04 (0.03-0.06)</td>
<td>0.67 (0.62-0.73)</td>
</tr>
<tr>
<td>Mental health</td>
<td>0.04 (0.03-0.06)</td>
<td>0.67 (0.62-0.71)</td>
</tr>
<tr>
<td>Family and Ethnicity</td>
<td>0.04 (0.03-0.05)</td>
<td>0.66 (0.60-0.71)</td>
</tr>
<tr>
<td>Diet</td>
<td>0.04 (0.03-0.06)</td>
<td>0.66 (0.60-0.71)</td>
</tr>
<tr>
<td>Socio demographics</td>
<td>0.03 (0.02-0.05)</td>
<td>0.65 (0.60-0.70)</td>
</tr>
<tr>
<td>Early Life Factors</td>
<td>0.03 (0.02-0.05)</td>
<td>0.64 (0.59-0.69)</td>
</tr>
<tr>
<td>Age and Sex</td>
<td>0.03 (0.02-0.04)</td>
<td>0.61 (0.56-0.67)</td>
</tr>
<tr>
<td>Only genetics</td>
<td>0.03 (0.02-0.04)</td>
<td>0.57 (0.51-0.63)</td>
</tr>
</tbody>
</table>

We then examined the additive contribution of each predictive group up to the total predictive power of the "all features" model (Appendix 1 Table 2). We started with the baseline model of "age and sex" and added feature groups, sorted by their predictive power as separate GBDT models. We concluded that using the four-blood-test model substantially increases the accuracy of the prediction results compared to a model based only on HbA1c% and age and sex. The auROC and APS increases from 0.84 (0.80-0.87), 0.17 (0.12-0.23) to 0.88(0.85-0.90), 0.20 (0.14-0.27), respectively.
The full blood tests model increased the auROC and APS to 0.90 (0.88-0.93) and 0.28 (0.21-0.36), respectively. We did not identify any major increase in accuracy of the predictions by adding any other specific group to this list, suggesting that most of the predictive power of our models are either captured by the blood test features or possess collinear information. Using all features together provided an increase in performance to auROC, APS, and the deciles' OR of 0.9(0.88-0.92), 0.28(0.22-0.35), and x65(49-73), respectively (Appendix 1 table 2).

Appendix 1 Table 2 | Summary of Incremental feature’s model: Comparison table of APS and auROC for the GBDT models, where each model includes the preceding model's features plus an additional feature domain. The largest increase in prediction accuracy was the result of adding the HbA1C% feature, which is also a biomarker for T2D diagnosis. Adding the DNA sequencing data did not significantly contribute to the prediction power of the model.

<table>
<thead>
<tr>
<th>Label</th>
<th>APS</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age and Sex</td>
<td>0.03 (0.02-0.04)</td>
<td>0.61 (0.56-0.67)</td>
</tr>
<tr>
<td>HbA1c%</td>
<td>0.17 (0.12-0.23)</td>
<td>0.84 (0.80-0.87)</td>
</tr>
<tr>
<td>Four Blood Tests</td>
<td>0.20 (0.14-0.27)</td>
<td>0.88 (0.85-0.90)</td>
</tr>
<tr>
<td>All Blood Tests</td>
<td>0.28 (0.21-0.36)</td>
<td>0.90 (0.88-0.93)</td>
</tr>
<tr>
<td>Adding Anthropometrics</td>
<td>0.23 (0.17-0.30)</td>
<td>0.90 (0.87-0.92)</td>
</tr>
<tr>
<td>Adding Physical Health DT</td>
<td>0.28 (0.21-0.36)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>Adding Lifestyle DT</td>
<td>0.24 (0.18-0.32)</td>
<td>0.91 (0.88-0.93)</td>
</tr>
<tr>
<td>Adding Blood pressure and heart rate</td>
<td>0.25 (0.19-0.33)</td>
<td>0.91 (0.88-0.93)</td>
</tr>
<tr>
<td>Adding non T2D related medical diagnosis</td>
<td>0.24 (0.18-0.32)</td>
<td>0.91 (0.88-0.93)</td>
</tr>
<tr>
<td>Adding Mental health</td>
<td>0.28 (0.20-0.36)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>Adding Medication</td>
<td>0.28 (0.20-0.35)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>Adding Diet</td>
<td>0.24 (0.18-0.31)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>Adding Family related information</td>
<td>0.28 (0.21-0.35)</td>
<td>0.91 (0.89-0.94)</td>
</tr>
<tr>
<td>Adding early life factors</td>
<td>0.24 (0.17-0.31)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>Adding sociodemographic</td>
<td>0.27 (0.20-0.36)</td>
<td>0.92 (0.89-0.94)</td>
</tr>
<tr>
<td>Adding Genetics</td>
<td>0.27 (0.20-0.34)</td>
<td>0.91 (0.89-0.93)</td>
</tr>
</tbody>
</table>

1. Deprivation Index differences between sick and healthy populations in our UKB cohort

Here we analyzed the Townsend deprivation index differences of participants diagnosed with T2D in one of their returning visits to the assessment center vs. the healthy population. The Townsend deprivation index measures deprivation based on employment status, ownership of car and home, and overcrowded household. Higher index values represent lower socioeconomic status. According to our analysis, a higher deprivation index is correlated with a higher risk of developing T2D (Appendix 1 Figure 2A). We analyzed the data using a Mann-Whitney U test with a sample of 1000 participants from each group; we achieved a p-value of 2.37e-137. When we analyzed the full cohort, the p-value dropped below the computational threshold, indicating a significant correlation between a Townsend deprivation index and our cohort’s tendency to develop T2D.

We performed a SHAP analysis on the Socio-demographic features of the patients, and features such as higher Townsend deprivation Index or being in the lower-income groups (Less than 40,000 GBP) push towards a prediction of developing T2D while being in the top two groups (52,000 GBP or more) is predictive of having less risk of developing T2D (Figure S3B). The full meaning of the codes is available at the U.K. Biobank data showcase.
Appendix 1 Figure 2 | Socioeconomic impact on prediction of risk of developing T2D | A. Deprivation Index differences between T2D sick and healthy populations in our data: A density histogram showing the differences in deprivation index of participants who were diagnosed with T2D in one of their returning visits to the assessment center and for healthy participants. Executing a Mann-Whitney test on this data yields a P-value lower than $2.37 \times 10^{-13}$, indicating a correlation between lower socio-demographic state with higher T2D prevalence. B. SHAP analysis of the socio-demographic features for a GBDT predictor of T2D: Each dot represents a participant’s value for each feature along the Y-axis. The colors indicate the values of the features: red indicates higher feature values, blue indicates lower feature values. The X-axis is the SHAP value, where higher SHAP values indicate a stronger positive impact on the positive prediction of the GBDT predictor, i.e., higher risk for T2D onset. The analysis indicates that higher values of deprivation index and lower household income push the probability of T2D onset to higher values. The full meaning of the codes is provided at the U.K. Biobank data showcase.

2. References for PRS summary statistics articles

- HbA1c (Soranzo et al., 2010; Walford et al., 2016; Wheeler et al., 2017)
- Cigarettes per day, ever smoked, age start smoking (Tobacco and Genetics Consortium, 2010)
- HOMA-IR, HOMA-B, diabetes BMI unadjusted, Diabetes BMI adjusted, fasting glucose (Morris et al., 2013)
- Fasting glucose, 2 hours glucose level, fasting insulin, fasting insulin adjusted BMI' - (MAGIC Scott) (Scott et al., 2012)
- Fasting glucose, fasting glucose adjusted for BMI, fasti