HAPT2D: high accuracy of prediction of T2D with a model combining basic and advanced data depending on availability

Barbara Di Camillo¹, Liisa Hakaste^{2,3}, Francesco Sambo¹, Rafael Gabriel^{4,5}, Jasmina Kravic⁶, Bo Isomaa³, Jaakko Tuomilehto^{5,7,8,9}, Margarita Alonso^{4,5}, Enrico Longato¹, Andrea Facchinetti¹, Leif C Groop^{6,10}, Claudio Cobelli¹ and Tiinamaija Tuomi^{2,3,10}

¹Department of Information Engineering, University of Padova, Padova, Italy, ²Endocrinology, Abdominal Centre, University of Helsinki and Helsinki University Hospital, Research Program for Diabetes and Obesity, University of Helsinki, Helsinki, Finland, ³Folkhälsan Research Center, Helsinki, Finland, ⁴Department of International Health, National School of Public Health, Instituto de Salud Carlos III, Madrid, Spain, ⁵Asociación Española Para el Desarrollo de la Epidemiología Clínica (AEDEC), Madrid, Spain, ⁶Lund University Diabetes Centre, Department of Clinical Sciences Malmö, Lund University, Skåne University Hospital, Malmö, Sweden, ⁷Dasman Diabetes Institute, Dasman, Kuwait City, Kuwait, ⁸Department of Neuroscience and Preventive Medicine, Danube-University Krems, Krems, Austria, ⁹Saudi Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia, and ¹⁰Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Correspondence should be addressed to T Tuomi
Email
tiinamaija.tuomi@hus.fi

Abstract

Objective: Type 2 diabetes arises from the interaction of physiological and lifestyle risk factors. Our objective was to develop a model for predicting the risk of T2D, which could use various amounts of background information.

Research design and methods: We trained a survival analysis model on 8483 people from three large Finnish and Spanish data sets, to predict the time until incident T2D. All studies included anthropometric data, fasting laboratory values, an oral glucose tolerance test (OGTT) and information on co-morbidities and lifestyle habits. The variables were grouped into three sets reflecting different degrees of information availability. Scenario 1 included background and anthropometric information; Scenario 2 added routine laboratory tests; Scenario 3 also added results from an OGTT. Predictive performance of these models was compared with FINDRISC and Framingham risk scores.

Results: The three models predicted T2D risk with an average integrated area under the ROC curve equal to 0.83, 0.87 and 0.90, respectively, compared with 0.80 and 0.75 obtained using the FINDRISC and Framingham risk scores. The results were validated on two independent cohorts. Glucose values and particularly 2-h glucose during OGTT (2h-PG) had highest predictive value. Smoking, marital and professional status, waist circumference, blood pressure, age and gender were also predictive.

Conclusions: Our models provide an estimation of patient's risk over time and outweigh FINDRISC and Framingham traditional scores for prediction of T2D risk. Of note, the models developed in Scenarios 1 and 2, only exploited variables easily available at general patient visits.

European Journal of
Endocrinology
(2018) **178**, 331–341

Introduction

About 700 million people worldwide are estimated to have T2D by year 2025 (1). To be able to halt this epidemic, we need to improve tools to prevent diabetes and also to predict it accurately. It is well known that the risk increases with increasing age, obesity, sedentary

lifestyle, hypertension, dyslipidemia, depression and unfavorable socioeconomic factors, as well as with history of gestational diabetes and family history of diabetes (2, 3, 4, 5, 6) suggesting a strong genetic predisposition (7).

Clinical Study	B Di Camillo and others	High accuracy of prediction of T2D	178:4	332
----------------	-------------------------	------------------------------------	-------	-----

As hyperglycemia is often asymptomatic, T2D can remain undetected for years and complications can appear already at the time of diagnosis (8, 9). Systematic screening with an OGTT has revealed that even 25–50% of adult patients with T2D had been undiagnosed in various populations (10, 11, 12, 13, 14, 15). As T2D can be prevented by lifestyle management (16, 17, 18, 19) and complications can be avoided with early intervention, it is of paramount importance to detect individuals at risk. However, large-scale screening with an OGTT is impractical and costly, which has led to the development of risk scores as first step in the identification of individuals with highest probability of either having undiagnosed diabetes or developing diabetes (20, 21).

Our aim was to develop a model for prediction of T2D, HAPT2D in the following, which could utilize all possible information depending on data availability, similarly to the approach adopted within the Framingham study (21). We used the ‘least absolute shrinkage and selection operator’ (LASSO) method (22) coupled with survival analysis in order to exploit the entire datasets and the entire time-frame of the study. In particular, we assessed whether adding a larger set of socioeconomic or environmental risk factors or factors derived from advanced modeling techniques, would improve the performance of the widely used the Finnish Diabetes Risk Score FINDRISC (20) and Framingham score (21). For these purposes, we used a joint set of population data from the VIVA, Botnia BPS and PPP-Botnia Studies for a total of 8483 people.

Among the many predictive models available in the literature to predict T2D onset (23, 24, 25, 26, 27), we selected the FINDRISC and the Framingham scores. The FINDRISC is one of the most widely used noninvasive risk scores (especially in Finland and Spain, from where our cohorts originated) for which a questionnaire was designed for self-screening at the population level. The Framingham score was developed, as HAPT2D, for different scenarios reflecting different degrees of available information and is among the best performers when tested on external validation cohorts (26, 27).

Subjects and methods

Within the FP7-MOSAIC Project (MOdels and Simulation techniques for discovering diAbetes Influence faCtors) funded by the European Commission, we studied three large prospective data sets, the Botnia Prospective Study

($N=3331$) and the PPP-Botnia Study ($N=3596$) from Finland and the VIVA Study from Spain ($N=1556$), including a total of 8483 people with no diabetes at the first visit, of whom 533 developed T2D during the follow-up (see below for details). All studies comprised a screening visit and a follow-up visit 1–22 years later (median follow-up time 8 years) with data for fasting blood samples, 75 g OGTT, anthropometric data, as well as information on co-morbidities, socioeconomic status and lifestyle. Variables common to all three datasets were identified, their coding was harmonized and the ones with overall proportion of missing data below 30% were considered for follow-up analysis (Table 1). Missing values were imputed with the k -Nearest Neighbor algorithm, using the Heterogeneous Euclidean Overlap Metric as a measure of distance between records (28). The number of neighbors k was set to the optimal value 25 after a tuning procedure on the available data.

Diabetes was defined as plasma glucose at fasting (FPG) ≥ 7.0 mmol/L or 2-h PG ≥ 11.1 mmol/L. The metabolic syndrome (MS) was defined according to the International Diabetes Federation criteria (29) as central obesity (waist circumference ≥ 94 cm for men and ≥ 80 cm for women), plus any two of the following criteria: (1) serum triglyceride concentration ≥ 1.7 mmol/L, or medication for this; (2) serum HDL cholesterol concentration < 1.03 in men or 1.29 mmol/L in women, or medication for this; (3) systolic blood pressure ≥ 130 or diastolic blood pressure ≥ 85 mmHg, or medication for hypertension and (4) FPG ≥ 5.6 mmol/L, or previously diagnosed T2D. The Homeostatic Model Assessment of insulin resistance index ($HOMA_IR = FPG \cdot FS \cdot \text{insulin} / 22.5$) and beta-cell index ($HOMA_B$; (30)) were calculated. Tertiles of $HOMA_B$ were used because of the high kurtosis of its distribution and negative values were pooled with the third tertile. Categorical variables such as profession and marital status were represented using one-hot variables; for example, the marital status was split in four variables (single, married, widowed, divorced) each taking the value 1 or 2 to indicate yes and no. A variable ‘country’ was added to account for differences in the Spanish and Finnish populations (Fig. 1).

The Botnia Prospective Study and the PPP-Botnia Study

The Botnia Study has been recruiting patients with T2D and their family members in the area of five primary health care centers in western Finland since 1990 aiming at

Clinical Study	B Di Camillo and others	High accuracy of prediction of T2D	178:4	333
-----------------------	--------------------------------	---	--------------	------------

Table 1 Baseline characteristics of the participants in the training and validation data sets. Data before imputation are shown as % or mean $\pm$ S.D. for most variables. Median and interquartile range are reported for the HOMA- β and HOMA-IR since their distribution is skewed.

Description	Training (VIVA+BPS+PPP1) (n=6401)	Validation 1 (VIVA+BPS+PPP1) (n=486)	Validation 2 (PPP2) (n=1596)
Scenario 1			
Sex (male/female)	45/55	44/56	46/54
Country of origin			
Spain	22	32	0
Finland	78	68	100
Age (years)	46.9 $\pm$ 13.7	45.3 $\pm$ 13.5	48.0 $\pm$ 14.5
Years of education	10.5 $\pm$ 4.1	9.8 $\pm$ 4.0	13.2 $\pm$ 3.6
Marital status			
Single	11.2	9.8	12.8
Married	82.2	85.3	79.1
Widowed	3.2	3.4	2.9
Divorced	3.4	1.5	5.2
History of cardio-vascular events	2.8	1.7	1.8
History of stroke	0.7	0.2	0.9
Anti-hypertensive medication	11.7	12.8	16.1
Lipid lowering medication	4.1	2.4	8.2
History of high glucose	1.2	2.0	2.1
Family history of diabetes			
None	28	26	22
Second-degree	30	15	27
First-degree	42	59	51
Occupation			
Clerical	36.6	31.1	52.2
Manual worker	42.8	52.9	22.2
Student	2.3	2.5	3.2
Unknown, unemployed, housewife, retired	18.3	13.5	22.4
Alcohol consumption, servings per week			
0–4	34.3	37.7	66.7
5–10	34.8	26.9	22.3
11–20	11.9	11.2	8.9
>20	19.0	24.2	2.1
Currently smoking (%)	18.1	19.6	13.1
History of smoking (%)	43.6	42.7	35.2
Physical activity at work			
Light	46.4	40.6	66.0
Mid	40.4	43.7	34.0
Heavy	13.2	15.7	0.0
Regular physical activity in leisure time (%)	60.1	65.8	56.4
Body mass index (kg/m ²)	26.3 $\pm$ 4.3	26.4 $\pm$ 4.6	26.3 $\pm$ 4.3
Waist circumference (cm)			
Male	94 $\pm$ 11	94 $\pm$ 11	94 $\pm$ 11
Female	85 $\pm$ 12	86 $\pm$ 14	83 $\pm$ 12
Added for the scenario 2			
Systolic blood pressure (mmHg)	129 $\pm$ 19	127 $\pm$ 19	130 $\pm$ 18
Diastolic blood pressure (mmHg)	79 $\pm$ 11	78 $\pm$ 12	80 $\pm$ 10
Pulse (beats per minute)	66 $\pm$ 15	65 $\pm$ 16	66 $\pm$ 10
Cholesterol (mmol/L)	5.49 $\pm$ 1.10	5.54 $\pm$ 1.17	5.24 $\pm$ 1.01
Triglycerides (mmol/L)	1.27 $\pm$ 0.78	1.24 $\pm$ 0.65	1.27 $\pm$ 0.78
HDL cholesterol (mmol/L)			
Male	1.23 $\pm$ 0.29	1.26 $\pm$ 0.32	1.48 $\pm$ 0.39
Female	1.48 $\pm$ 0.34	1.45 $\pm$ 0.33	1.63 $\pm$ 0.41
Fasting plasma glucose (mmol/L)	5.3 $\pm$ 0.6	5.4 $\pm$ 0.6	5.4 $\pm$ 0.6
Fasting serum insulin (μ U/mL)	8.12 $\pm$ 6.80	7.99 $\pm$ 6.00	6.59 $\pm$ 4.98
HOMA- β index	73.7 (45.7–125.3)	67.0 (41.0–124.6)	59.0 (40.1–88.4)
HOMA-IR index	1.5 (1.0–2.4)	1.4 (0.9–2.3)	1.3 (0.9–1.9)
Metabolic syndrome (%)	30.9	30.1	24.6
Added for the scenario 3			
2h-PG (mmol/L)	5.7 $\pm$ 1.6	6.0 $\pm$ 1.7	5.2 $\pm$ 1.6

Figure 1

Kaplan–Meier plot for people remaining without T2D in Finnish (BPS, PPP1 and PPP2 studies) and Spanish (VIVA study) populations.

identifying clinical and genetic risk factors predisposing to T2D. During 1994–1998, the study was extended to other parts of Finland and southern Sweden. Those without diabetes at baseline (relatives or spouses of patients with T2D) have been invited for follow-up examinations every 3–5 years (the Botnia Prospective Study, BPS) (31, 32, 33). By the end of 2013, follow-up data were available for 3331 participants with an average follow-up time of 9.1 (s.d. 4.5) years and with 215 incident T2D.

The Prevalence, Prediction and Prevention of diabetes (PPP) – Botnia Study is a population-based study in the same region designed to obtain accurate estimates of prevalence and risk factors for diabetes, impaired glucose tolerance and the MS in the population aged 18–75 years (34). A random sample of individuals was invited from the population registry, of which 5208 (54.7%) participated in the baseline study in 2004–2008 representing 6–7% of the population in the target age group (comprising 96 000 individuals). At baseline, 327 (6.3%) had diabetes, 354 (6.8%) impaired glucose tolerance and 416 (8.0%) impaired fasting glucose (33). Of the 3850 (76.5% of those alive) people who participated in the follow-up examination in 2011–2015, data for 3596 were used in the present study (254 people were excluded due to insufficient follow-up information). Of them, 152 (4%) had developed T2D. Data for the first 2000 individuals

participating in the follow-up, 97 of whom had developed T2D, were retrospectively used for building the predictive model (PPP1). The latter part of the follow-up cohort (1596 people, 55 with incident T2D), examined after the initiation of the present study, was used for the validation of the model (PPP2, validation set 2). The participants in the BPS and PPP-Botnia studies gave their written informed consent and the study protocol was approved by the Ethics Committee of Helsinki University Hospital, Finland.

All people participated in a 75 g OGTT after a 10- to 12-h fast. Samples for plasma glucose and serum insulin were drawn at –10, 0, 30, 60 and 120 min (BPS) or 0, 30 and 120 min (PPP-BOTNIA), and for total cholesterol, HDL cholesterol and triglycerides at 0 min. Body weight, height, waist and hip circumference as well as heart rate were measured. The mean value was calculated from two blood pressure recordings obtained from the right arm of a sitting person after 30 min of rest at 5-min intervals. The study participants filled in a structured questionnaire to provide information on marital status, occupation, length of education, exercise, alcohol consumption, smoking, cardiovascular and other diseases as well as family history of diabetes. Medication was recorded by a trained study nurse. Diagnosis of diabetes was based on the OGTT or a history of previously known diabetes applying WHO criteria. In uncertain cases, the diagnosis was confirmed from patient records.

Plasma glucose was measured with a hexokinase (Boehringer Mannheim, Mannheim, Germany; BPS) or a glucose dehydrogenase method (Hemocue, Angelholm, Sweden; PPP-BOTNIA). Serum insulin was measured by radioimmunoassay (RIA, Linco; Pharmacia, Uppsala, Sweden) or enzyme immunoassay (EIA; DAKO) in the first part of BPS, and then by fluoroimmunoassay (FIA, AutoDelfia; Perkin Elmer Finland) in the latter part of BPS and the whole PPP-BOTNIA. For the analysis, the insulin concentrations obtained using the other assays were transformed to cohere with those obtained using the EIA. The correlation coefficient between RIA and EIA as well as FIA and EIA was 0.98 ($P < 0.0001$). Serum total cholesterol, HDL and triglyceride concentrations were measured first on a Cobas Mira analyzer (Hoffman LaRoche, Basel, Switzerland) and since January 2006 with an enzymatic method (Konelab 60i analyser; Thermo Electron Oy, Vantaa, Finland). LDL cholesterol concentrations were calculated using the Friedewald formula.

The Variability of Insulin with Visceral Adiposity study (VIVA)

VIVA is an observational multi-center study aiming at identifying risk factors related to MS and cardiovascular diseases and developing a specific scale for Spanish population for detecting people at risk of developing T2D. The baseline study included 2959 non-diabetic and non-pregnant people aged 30–64 years, randomly derived from nine population registries in Spain in 1998 (35). Ten years later, (68.3%) people participated in the follow-up examinations, and data for 1556 who participated in an OGTT (14, 15) were used for the present study; 166 people had incident diabetes. Information on clinical history, metabolic and cardiovascular diseases as well as cardiovascular disease risk factors was obtained through interview, and body weight, height, waist circumference and blood pressure (after 5 min of rest with the person sitting) were measured. FPG and 2-h PG and fasting serum insulin, total cholesterol, HDL cholesterol and triglycerides were analyzed in a central laboratory with standard assays using quality assurance by the Spanish Society of Clinical Chemistry (Fundación Jiménez Díaz of Madrid).

The prediction model

For the modeling purposes, data from the BPS, the PPP1 and VIVA studies were combined to include 6887 people, of whom 478 had developed incident T2D during the follow-up. A subset ($n=486$; 7%) of these data, stratified for sex, age, BMI and incident T2D, was set aside for validation purposes (validation set 1) leaving a training set of 6401 people.

The models were created for three different scenarios reflecting different degrees of available information (Table 1). Scenario 1 includes questionnaire data and basic measurements (weight, height, waist circumference, BMI), resembling information found in citizens clinical history records. Scenario 2 complements Scenario 1 with the usual information available for a clinician after a general visit including blood pressure and routine laboratory tests. Scenario 3 includes all the above-mentioned variables and 2h-PG.

First we used Bayesian Network analyses to study several indices for profiling glucose variability in type 2 diabetes (T2D) and to provide a probabilistic model of the relations between different T2D risk factors in the three datasets. A Bootstrap aggregation strategy on top of

the Bayesian Network structure learning algorithm was exploited to obtain a ranked list of variables useful for the model, and an optimal number of variables.

For the model predicting T2D, we used the LASSO ‘least absolute shrinkage and selection operator’ regression analysis (LASSO, (22)) coupled with Cox survival model (36) to be able to use the whole set of available data in the entire time-frame of the study (2–20 years, Fig. 1), not just the data at a certain time point. This produces a curve of patient risk across follow-up years, which facilitates the interpretation of prediction.

The training set ($n=6401$) was further split in 100 training and 100 test sets and the model was trained in a Monte Carlo bootstrap resampling scheme (37, 38, 39, 40) with $B=100$ external training/test splits separately for each of the three scenarios – one model was learned from each of these training sets and validated on the corresponding test set resulting in 100 different scores of the method performance. In addition, as LASSO selects a number of variables significant for prediction of T2D for each model, we obtained 100 lists of variables that can be ranked based on the number of times a variable is selected by different models allowing us to rank the variables based on their ability to predict the onset of T2D being robust to overfitting (38, 39, 40). The training data were scaled with respect to their maximum and used to train three models, one for each scenario, using the package ‘survival’ in R (36).

The final survival model was trained on the entire training set of 6401 people using the top ranking variables. The optimal number of variables was chosen as the one optimizing the average performance across the 100 test sets.

The final model performance on the three scenarios described earlier was assessed on validation sets 1 and 2. The outcome produced by each of the three predictive algorithms is twofold: (i) a predictive model of the risk of developing T2D in the future assessed in terms of Integrated Areas Under the ROC Curve (iAUCs) (41, 42) and (ii) a selection and ranking of the variables based on their ability to predict the risk of developing T2D. We also considered AUC at 6 years from the baseline visit, which corresponds to a reasonable number of cases developing diabetes in our dataset (Fig. 1) and is in the time range of FINDRISC and Framingham predictions, thus allowing a comparison with these two traditional scores. We assessed FINDRISC and Framingham performance on the 100 internal test splits in addition to the external validation data sets 1 and 2 and got 100 performance scores for each model. Finally, the performance of the method was

compared with the FINDRISC (20) and Framingham (21) risk scoring methods.

FINDRISC and Framingham score implementation

The FINDRISC score was implemented according to the paper by Lindström and Tuomilehto (20) and subsequent updates, using the following variables: sex, age, BMI, waist circumference, antihypertensive medication, information on physical activity habits and physical work and history of high blood glucose. The variable high blood glucose includes any occasional measurement of abnormal glucose in the past, could be fasting or post-prandial, pre-diabetic or diabetic range. In practice, the patient has been told that he/she has an abnormal value, but no diagnosis of permanent diabetes has been made (for example, this could have been e.g. in conjunction with an infection). The Framingham score was implemented using the variables sex, age, BMI, family history of diabetes, systolic and diastolic blood pressure, HDL cholesterol, triglycerides and FPG, as summarized in the following equation, according to Wilson *et al.* (21) and subsequent updates:

$$p = \frac{1}{1 + e^{-z}}; z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_k \cdot x_k$$

where z indicates regression equation with beta coefficients taken from (21), and x 's indicate values of corresponding risk factor.

Results

The three models developed for Scenarios 1, 2 and 3 predicted T2D risk during the whole follow-up time with iAUC (SD) on the test set by 0.83 (0.03), 0.87 (0.02) and 0.90 (0.02), respectively (boxplots in Fig. 2, left panel). This performance was corroborated in the validation data sets 1 and 2, in which all scenarios showed performance higher or equal to an iAUC of 0.85 (Scenarios 1, 2, 3; validation set 1: 0.85, 0.85, 0.95; validation set 2: 0.87, 0.91, 0.91, respectively). The average performance obtained including 2-h-PG (Scenario 3) outperformed those obtained in the other scenarios ($P < 10^{-15}$ for Scenario 3 vs either Scenario 2 or Scenario 1, t -test).

AUCs at year 6 obtained with Scenarios 1, 2 and 3 (Fig. 2, right panel) were consistent with iAUCs for the whole follow-up time (Fig. 2, right panel) on both validation sets 1 (0.87, 0.89 and 0.95, respectively)

and 2 (0.90, 0.98, 0.96, respectively). When testing the FINDRISC score, AUC at year 6 was lower than using HAPT2D for both validation sets 1 and 2 (AUC equal to 0.80 and 0.88, respectively). Moreover, on the test sets, HAPT2D performed better than FINDRISC on all the three scenarios (Fig. 2, right panel, $P < 10^{-10}$). Also, compared with the Framingham score, HAPT2D performed better on validation set 1 (AUC=0.71) and on the test sets ($P < 10^{-15}$ for all the three scenarios), but the performance of Framingham score was similar to those obtained in Scenario 1 using validation set 2 (AUC=0.9).

The variables selected and ranked for their ability to predict T2D are shown in Table 2. The lower the rank, the more frequently the variable was selected as a good

Figure 2

Boxplots of the iAUC (top panel; median follow-up time 8 years) or AUC at year 6 (bottom panel) for the 100 internal train/test splits obtained using different models (Scenarios 1, 2, and 3) and FINDRISC and Framingham scores. The iAUCs or AUCs for the external validation data sets 1 and 2 (Val1, Val2) are shown in the middle of the panels. iAUC, integrated Area Under the ROC Curve for the 100 internal train/test splits.

Clinical Study	B Di Camillo and others	High accuracy of prediction of T2D	178:4	337
-----------------------	--------------------------------	---	--------------	------------

Table 2 Variables selected (recursive feature elimination) in different scenarios and their average ranking based on ability to predict the risk of developing T2D. The average rank of the variable across the 100 bootstrap samples from the training set is shown. The lower the rank, the more frequently the variable was selected as a good predictor using different data subsamples.

Scenario 1		Scenario 2		Scenario 3	
Variable	Rank	Variable	Rank	Variable	Rank
Country	2.9	FPG	2.7	2h-PG	1.8
Waist circumference	3.7	Heart rate	11.5	FPG	8.2
Family history of diabetes	6.7	DBP	11.7	Country	9.7
Antihypertensive medication	9.0	Country	12.0	Heart rate	12.5
Age	10.2	Physical activity at work	13.0	Physical activity at work	13.7
Regular physical activity at work	10.8	Waist circumference	13.3	Age	13.9
Sex	10.9	HOMA_B	14.3	Marital status	14.2
Marital status	11.2	Sex	14.3	Waist circumference	14.4
History of high glucose	11.2	Age	14.3	HOMA_B	14.6
Professional status	11.9	Marital status	14.4	Smoking	14.6
Current smoking	12.3	Triglycerides	15.4	Professional status	14.9
		Metabolic syndrome	16.4	DBP	16.1
				Sex	16.8
				HOMA_IR	17.5

predictor using different data subsamples. 2-h PG and FPG were among the best predictors in the scenarios together with other known predictors of T2D, i.e. waist circumference, age, sex, family history of diabetes and blood pressure (Table 2). In addition, smoking habits, marital and occupational status got selected in the prediction models, as well as the variable country (Table 2), which accounts for a higher risk of Finnish population with respect to Spanish (Fig. 1 and Supplementary data, see section on Supplementary data given at the end of this article).

In the most informative scenarios, we included some variables that are highly related to each other such as history of high glucose (positive answer from a patient to a questionnaire) and FPG levels or waist circumference, triglycerides and MS. However, the feature selection step automatically selects the most informative variables, ranking them and discarding those that do not add any further advantage to the model performance. For example, as shown in Table 2, FBG is selected in Scenarios 2 and 3, whereas 'history of high glucose' is selected only in Scenario 1, where FBG is not available. On the opposite, in Scenario 2, waist circumference, triglycerides and MS are all selected, although this latter with a low rank, indicating that MS helps improving the prediction possibly because it introduces further information on HDL and blood pressure status.

The model equations and parameters are available in Supplementary data and its implementation is available at <http://sysbiobig.dei.unipd.it/?q=Software> \l "HAPT2D" \t "_blank").

Discussion

The three models predicted T2D risk with an average integrated area under the ROC curve equal to 0.83, 0.87 and 0.90, respectively. The average performance obtained including 2-h PG (Scenario 3) outperformed those obtained in the other scenarios, and the Scenario 2 outperformed Scenario 1, the difference being more clear between the first and second scenarios compared to the difference between the second and the third scenarios. The first scenario included variables available without blood samples. For the Scenario 2, fasting blood sample with glucose, insulin and lipids is needed, as well as measurement of pulse and blood pressure. For Scenario 3, OGTT is needed. The performance at six years (AUC 0.87–0.95) overcame that of the state-of-the-art FINDRISC and Framingham risk scores (AUC 0.80 and 0.75, respectively). Even the simplest model using only noninvasive data was a good predictor, and the model with the richest data predicted incident diabetes better than 1-h plasma glucose previously shown to be highly predictive in the BPS (43).

Several risk scores for predicting T2D have been developed utilizing various amounts of background and laboratory data (27). Many factors affect their performance, including the richness of data, handling of missing data, generalizability to other populations and whether individual or population risk is targeted. To construct HAPT2D, we adopted the approach outlined in the Framingham study (20) and assessed the ability of the model to predict the onset of T2D with different degrees of available information. As hypothesized, but differing

from the Framingham score, feeding the model with more data progressively improved the performance in terms of iAUC, and this was even more apparent as regards AUC at 6 years.

The variables that predicted best the risk of T2D included: country, waist circumference, family history of diabetes, fasting glucose, heart rate, diastolic blood pressure and 2-h glucose at OGTT. In addition, although with lower rank, variables related to smoking habits, as well as marital and professional status, together with HOMA_B, waist circumference, blood pressure, triglycerides, age and sex, were selected as predictive (Table 2). These are mainly known risk factors for T2D and were expected. Country of origin covers differences in the Spanish and Finnish populations, including lifestyle habits, study design and genetic background, a known risk factor for T2D (32). This highlights the importance of recalibrating the predictive models when using them on different populations (44).

Also heart rate, somewhat surprisingly, got included as a predictive variable – although contrary to the expected direction, low heart rate was not protective in the model. The use of beta blockers might explain the finding, but as all antihypertensive medications were pooled in the same variable, this could not be assessed. It is also possible, that the finding is a true association, as reduced heart rate variability has been shown to associate with insulin resistance and lower insulin sensitivity, and decreased insulin sensitivity index has been linked with parasympathetic dysfunction, primarily in non-overweight individuals (45).

The variable physical work was included in all three scenarios as a detrimental factor increasing the risk of developing T2D. If we think of physical work as a proxy of physical activity, this is somehow surprising. We think that in this model, such as in other models including environmental variables, some variables might be a proxy of a number of different variables which are maybe not directly measured, but that are predictive. For example, in our case physical work might be a proxy of a low socioeconomic status.

Unlike the previous modeling approaches, we adapted 'least absolute shrinkage and selection operator' (LASSO) method for creating the predictive models. This method enables us to tackle the low incidence of diabetes (7%) and varying follow-up time (Fig. 1) in our data. The low incidence of T2D in our data may have led to results by chance only, but on the other hand, the results were consistent. One benefit of LASSO is its output, a curve of patient risk over the years, which makes the interpretation of the result easier and usable also at daily clinical practice.

For an individual, a visualized risk of T2D might be an incentive for lifestyle changes and a clinician might be more alert as regards T2D risk.

For clinical use, it might be of interest to fix a threshold of risk. Such a threshold can be chosen based on a compromise between false positives and false negative cases on the available data. For example, in the model implementation available at <http://147.162.226.104:8000/> (temporary link accessible to the reviewers with Username: HAPT2D and password dicamilloetal) the prediction is shown as the cumulative risk of developing T2D in the course of 12 years (Fig. 3).

There are a few other previously developed scores for prediction of T2D. In some of them, Cambridge, DPoRT, FINDRISC and T2D diagnosis were based on registry data or was self-reported and some used fasting glucose measurement (ARIC, DESIR) instead of OGTT (27). The

Figure 3

Exemplifying curve of patient risk over the years. Each point on the black line corresponds to the probability of developing T2D within the number of years shown on the x axis. As a guideline to determine how high the shown risk level actually is, compared to the observed prognoses in the training population, four color-coded bands are shown. The area in dark gray represents high risk (50% of the subjects with a risk curve in this area eventually developed diabetes within 12 years); the area in mid gray represents medium-high risk (25% of the subjects with a risk curve in this area eventually developed diabetes within 12 years); the light gray represents a medium risk (10% of the subjects with a risk curve in this area eventually developed diabetes within 12 years); the white area represents a low risk (less than 10% of the subjects with a risk curve in this area eventually developed diabetes within 12 years).

strength of our study is the reliable diagnosis of T2D based on OGTT and careful examination of the patient records. The methods in our population studies were also well defined.

When comparing the results with Framingham and FINDRISC models, we found that feeding only non-laboratory data (Scenario 1), HAPT2D performed better in the prediction of T2D than the FINDRISC score utilizing a smaller selection of non-laboratory data and the Framingham risk score, which also includes data on blood pressure measurements, fasting glucose, insulin and lipids.

The improvement in prediction is probably due to several reasons. Firstly, with LASSO coupled with survival analysis, we could take advantage of the entire follow-up regardless of its length and censored data, while FINDRISC (20), Framingham (21) and other approaches (46) used a fixed time horizon between the visits. Full exploitation of the available data results in a more robust model identification procedure. In fact, when we trained different prediction models such as Support Vector Machines and LASSO not coupled with survival analysis, we obtained performance similar to those obtained by Framingham (data not shown). Secondly, the bootstrap approach we used to obtain an average estimate of the models accuracy and to select the predicting variables has previously been shown able to predict the average performance on the internal test splits in a way that resembles closely the actual performance on an external, previously unseen, validation set (37, 39). Third, we were able to include a large number of variables in HAPT2D: traditionally used risk factors, socioeconomic and environmental factors, biochemical data as well as OGTT-derived data like the HOMA_IR and HOMA_B indices. Similar difference in performance between the models and the FINDRISC and Framingham scores was observed on the first but not the second validation set, where the Framingham score performed as well as the Scenario 1 model (AUC=0.9). Also Framingham score utilizes a wide variety of data that may explain this finding. However, Framingham AUC showed high variability on different datasets, ranging from 0.61 to 0.84 on different bootstrap samples. Furthermore, the performance of different models has been shown to vary with country, age, sex and adiposity, and discrimination can vary across BMI and waist circumference strata (27), but there were no obvious differences in these variables between the training and the validation datasets. In the validation study of Kengne *et al.* (27) comparing different prediction tools, FINDRISC acceptably predicted the overall rate of incident diabetes in age subgroups, but the discrimination of Framingham model was better.

As an external validation data, we used the latter part of the Botnia-PPP study (PPP2). It can be argued that this validation was not totally independent of the data, the model construction was based on (PPP1). However, firstly, the original study population was randomly selected from the population registry, and secondly, the diabetes status of the PPP2 individuals was unknown at the time the models were constructed. Thirdly, the order of the follow-up examination was not related to the diabetes status in any way.

It can be debated whether there is a real need for one more T2D risk calculator. However, the world and the clinical practice are changing rapidly to be more computerized and personalized. While previously, counting a risk for T2D was more or less cumbersome even with a help of manual risk calculation aids, in the future, data mining techniques enable an automatic risk calculation in seconds. Also, an easy to visualize output, such as ours, would certainly benefit clinicians and patients at risk. It is also clear that there are differences across populations regarding the risk of T2D. With the current worldwide T2D epidemic, it is only an advantage to have one more risk calculator at hand.

Supplementary data

This is linked to the online version of the paper at <https://doi.org/10.1530/EJE-17-0921>.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this study.

Funding

This study was performed as part of the FP7-MOSAIC project funded by the European Union under the 7th framework program (grant agreement FP7-600914). The Botnia Prospective Study and the PPP-Botnia study have also been financially supported by grants from the Sigrid Juselius Foundation, Folkhälsan Research Foundation, Nordic Center of Excellence in Disease Genetics, EU (EXGENESIS), Ollqvist Foundation, Signe and Ane Gyllenberg Foundation, Swedish Cultural Foundation in Finland, Finnish Diabetes Research Foundation, Foundation for Life and Health in Finland, Finnish Medical Society, Paavo Nurmi Foundation, Helsinki University Central Hospital Research Foundation, Perklén Foundation, Närpes Health Care Foundation and Ahokas Foundation. The study has also been supported by the Ministry of Education in Finland, Municipal Health Care Center and Hospital in Jakobstad and Health Care Centers in Vasa, Närpes and Korsholm. The VIVA Study has also been supported by Fondo de Investigación Sanitaria, Instituto de Salud Carlos III: PI95/0029, PI06/90270 and PI15/00308.

Author contribution statement

All authors contributed to study concept and design, analysis and interpretation of the data, and the drafting and reviewing of the

Clinical Study	B Di Camillo and others	High accuracy of prediction of T2D	178:4	340
----------------	-------------------------	------------------------------------	-------	-----

manuscript. All authors approved the final version of the manuscript. B D C performed the computational data analysis and developed the model with the contribution from Folkhälsan, Lh, TT, A F and C C. Barbara Di Camillo, Liisa Hakaste and Tiinamaija Tuomi wrote the manuscript. Barbara Di Camillo and Tiinamaija Tuomi are the guarantors of this work and, as such, take responsibility for the integrity of the data and the accuracy of the data analysis.

Acknowledgements

The authors thank their collaborators in the MOSAIC study for fruitful discussions and the Botnia Study Group for clinically studying the BPS and PPP-Botnia participants.

References

- NCD Risk Factor Collaboration. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* 2016 **387** 1513–1530. ([https://doi.org/10.1016/S0140-6736\(16\)00618-8](https://doi.org/10.1016/S0140-6736(16)00618-8))
- American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2010 **33** S62–S69. (<https://doi.org/10.2337/dc10-S062>)
- Mezuk B, Eaton WW, Albrecht S & Golden SH. Depression and T2D over the lifespan. *Diabetes Care* 2008 **31** 2383–2390. (<https://doi.org/10.2337/dc08-0985>)
- Knol MJ, Twisk JW, Beekman AT, Heine RJ, Snoek FJ & Pouwer F. Depression as a risk factor for the onset of T2D mellitus. A meta-analysis. *Diabetologia* 2006 **49** 837. (<https://doi.org/10.1007/s00125-006-0159-x>)
- Zimmet P. Nauru and Mauritius: barometers of a global diabetes epidemic. *Journal of Medical Sciences* 2010 **3** 78–81. (<https://doi.org/10.2174/1996327001003020078>)
- Tuomilehto J. T2D is a preventable disease-lifestyle is the key. *Journal of Medical Sciences* 2010 **3** 82–86.
- Genetics and Diabetes – Report from World Health Organization. (available from: www.who.int/genomics/about/Diabetis-fin.pdf). Last accessed on April 10, 2012.
- Harris MI, Klein R, Welborn TA & Knudman MW. Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes Care* **15** 815–819. (<https://doi.org/10.2337/diacare.15.7.815>)
- Kohner EM, Aldington SJ, Stratton IM, Manley SE, Holman RR, Matthews DR & Turner RC. United Kingdom Prospective Diabetes Study, 30: diabetic retinopathy at diagnosis of non-insulin-dependent diabetes mellitus and associated risk factors. *Archives of Ophthalmology* 1998 **116** 297–303. (<https://doi.org/10.1001/archophth.116.3.297>)
- Gregg EW, Cadwell BL, Cheng YJ, Cowie CC, Williams DE, Geiss L, Engelgau MM & Vinicor F. Trends in the prevalence and ratio of diagnosed to undiagnosed diabetes according to obesity levels in the US. *Diabetes Care* 2004 **27** 2806–2812. (<https://doi.org/10.2337/diacare.27.12.2806>)
- Mayor S. Quarter of people with diabetes in England are undiagnosed. *BMJ* 2005 **331** 656. (<https://doi.org/10.1136/bmj.331.7518.656-a>)
- Koopman RJ, Mainous AG & Jeffcoat AS. Moving from undiagnosed to diagnosed diabetes: the patient's perspective. *Family Medicine* 2004 **36** 727–732.
- Saaristo TE, Barengo NC, Korpi-Hyövälti E, Oksa H, Puolijoki H, Saltevo JT, Vanhala M, Sundvall J, Saarikoski L, Peltonen M *et al*. High prevalence of obesity, central obesity and abnormal glucose tolerance in the middle-aged Finnish population. *BMC Public Health* 2008 **8** 423. (<https://doi.org/10.1186/1471-2458-8-423>)
- DECODE Study Group. Age- and sex-specific prevalences of diabetes and impaired glucose regulation in 13 European cohorts. *Diabetes Care* 2003 **26** 61–69. (<https://doi.org/10.2337/diacare.26.1.61>)
- The DECODE Study Group. Comparison of three different definitions for the metabolic syndrome in non-diabetic Europeans. *British Journal of Diabetes and Vascular Disease* 2005 **5** 161–168. (<https://doi.org/10.1177/14746514050050030901>)
- Diabetes Prevention Program Research Group. Reduction in the incidence of T2D with lifestyle intervention or metformin. *New England Journal of Medicine* 2002 **346** 393–403. (<https://doi.org/10.1056/NEJMoa012512>)
- Pan XR, Li GW, Hu YH, Wang JX, Yang WY, An ZX, Hu ZX, Xiao JZ, Cao HB, Liu PA *et al*. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance: the Da Qing IGT and Diabetes Study. *Diabetes Care* 1997 **20** 537–544. (<https://doi.org/10.2337/diacare.20.4.537>)
- Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, Keinänen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M *et al*. Prevention of T2D mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *New England Journal of Medicine* 2001 **344** 1343–1350. (<https://doi.org/10.1056/NEJM200105033441801>)
- Panel AD. Guidelines for computer modeling of diabetes and its complications. *Diabetes Care* 2004 **27** 2262–2265. (<https://doi.org/10.2337/diacare.27.9.2262>)
- Lindström J & Tuomilehto J. The diabetes risk score. *Diabetes Care* 2003 **6** 725–731. (<https://doi.org/10.2337/diacare.26.3.725>)
- Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM & D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Archives of Internal Medicine* 2007 **167** 1068–1074. (<https://doi.org/10.1001/archinte.167.10.1068>)
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, Methodological* 1996 **58** 267–288.
- Buijsse B, Simmons RK, Griffin SJ and Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. *Epidemiologic Reviews* 2011 **33** 46–62. (<https://doi.org/10.1093/epirev/mxq019>)
- Noble D, Mathur R, Dent T, Meads C & Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011 **343** d7163. (<https://doi.org/10.1136/bmj.d7163>)
- Cichosz SL, Johansen MD & Hejleson O. Toward big data analytics: review of predictive models in management of diabetes and its complications. *Journal of Diabetes Science and Technology* 2015 **10** 27–34. (<https://doi.org/10.1177/1932296815611680>)
- Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AMW, van der ADL, Moons KG, Navis G, Bakker SJ *et al*. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* 2012 **345** e5900. (<https://doi.org/10.1136/bmj.e5900>)
- Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, Spijkerman AM, Griffin SJ, Grobbee DE, Palla L *et al*. Non-invasive risk scores for prediction of T2D (EPIC-InterAct): a validation of existing models. *Lancet Diabetes and Endocrinology* 2014 **2** 19–29. ([https://doi.org/10.1016/S2213-8587\(13\)70103-7](https://doi.org/10.1016/S2213-8587(13)70103-7))
- García-Laencina PJ, Sancho-Gómez JL & Figueiras-Vidal AR. Pattern classification with missing data: a review. *Neural Computing and Applications* 2010 **19** 263–82. (<https://doi.org/10.1007/s00521-009-0295-6>)
- The IDF consensus worldwide definition of the metabolic syndrome. (available at: http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf)
- Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF & Turner RC. Homeostasis model assessment: insulin resistance

- and β -cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 1985 **28** 412–419. (<https://doi.org/10.1007/BF00280883>)
- 31 Lyssenko V, Almgren P, Anevski D, Perfekt R, Lahti K, Nissén M, Isomaa B, Forsén B, Homström N, Saloranta C *et al.* Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of T2D. *Diabetes* 2005 **54** 166–174. (<https://doi.org/10.2337/diabetes.54.1.166>)
 - 32 Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P & Groop L. Clinical risk factors, DNA variants, and the development of T2D. *New England Journal of Medicine* 2008 **359** 2220–2232. (<https://doi.org/10.1056/NEJMoa0801869>)
 - 33 Lundgren VM, Isomaa B, Lyssenko V, Laurila E, Korhonen P, Groop LC, Tuomi T & Botnia Study Group. GAD antibody positivity predicts T2D in an adult population. *Diabetes* 2010 **59** 416–422. (<https://doi.org/10.2337/db09-0747>)
 - 34 Isomaa B, Forsén B, Lahti K, Holmström N, Waden J, Matintupa O, Almgren P, Eriksson JG, Lyssenko V, Taskinen MR *et al.* A family history of diabetes is associated with reduced physical fitness in the Prevalence, Prediction and Prevention of Diabetes (PPP)–Botnia study. *Diabetologia* 2010 **53** 1709–1713. (<https://doi.org/10.1007/s00125-010-1776-y>)
 - 35 Gabriel R, Alonso M, Parra J, Fernández-Carreira JM, Rojo-Martínez G, Brotons C, Segura A, Cabello J, Muñoz J, Vega S *et al.* Aggregation pattern and factorial analysis of cardiovascular risk factors included in the metabolic syndrome in a Spanish non-diabetic population: the VIVA study. *Avances en Diabetologia* 2009 **25** 131–138.
 - 36 Therneau TM & Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media, 2013.
 - 37 Ambrose C & McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS* 2002 **99** 6562–6566. (<https://doi.org/10.1073/pnas.102102699>)
 - 38 Di Camillo B, Sanavia T, Martini M, Jurman G, Sambo F, Barla A, Squillario M, Furlanello C, Toffolo G & Cobelli C. Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. *PLoS ONE* 2012 **7** e32200. (<https://doi.org/10.1371/journal.pone.0032200>)
 - 39 Furlanello C, Serafini M, Merler S & Jurman G. Semisupervised learning for molecular profiling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005 **2** 110–118. (<https://doi.org/10.1109/TCBB.2005.28>)
 - 40 Pegolo S, Di Camillo B, Montesissa C, Cannizzo FT, Biolatti B & Bargelloni L. Toxicogenomic markers for corticosteroid treatment in beef cattle: integrated analysis of transcriptomic data. *Food and Chemical Toxicology* 2015 **77** 1–11. (<https://doi.org/10.1016/j.fct.2014.12.001>)
 - 41 Sing T, Sander O, Beerenwinkel N & Lengauer T. *ROCR: visualizing classifier performance in R*. *Bioinformatics* 2005 **21** 3940–3941. (<https://doi.org/10.1093/bioinformatics/bti623>)
 - 42 Blanche P, Dartigues JF & Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine* 2013 **32** 5381–5397. (<https://doi.org/10.1002/sim.5958>)
 - 43 Alyass A, Almgren P, Akerlund M, Dushoff J, Isomaa B, Nilsson P, Tuomi T, Lyssenko V, Groop L & Meyre D. Modelling of OGTT curve identifies 1 h plasma glucose level as a strong predictor of incident T2D: results from two prospective cohorts. *Diabetologia* 2015 **58** 87–97. (<https://doi.org/10.1007/s00125-014-3390-x>)
 - 44 Van Der Leeuw J, Visseren FL, Woodward M, Zoungas S, Kengne AP, Van Der Graaf Y, Glasziou P, Hamet P, MacMahon S, Poulter N *et al.* Predicting the effects of blood pressure–lowering treatment on major cardiovascular events for individual patients with T2D mellitus: novelty and significance. *Hypertension* 2015 **65** 115–121. (<https://doi.org/10.1161/HYPERTENSIONAHA.114.04421>)
 - 45 Saito I, Hitsumoto S, Maruyama K, Nishida W, Eguchi E, Kato T, Kawamura R, Takata Y, Onuma H, Osawa H *et al.* Heart rate variability, insulin resistance, and insulin sensitivity in Japanese adults: the Toon Health Study. *Journal of Epidemiology* 2015 **25** 583–591. (<https://doi.org/10.2188/jea.JE20140254>)
 - 46 Abdul-Ghani MA, Abdul-Ghani T, Stern MP, Karavic J, Tuomi T, Bo I, DeFronzo RA & Groop L. Two-step approach for the prediction of future T2D risk. *Diabetes Care* 2011 **34** 2108–2112. (<https://doi.org/10.2337/dc10-2201>)

Received 3 November 2017

Revised version received 16 January 2018

Accepted 25 January 2018