PULSE

Participatory Urban Living for Sustainable Environments

D5.2 Selected models implementation and recalibration against each specific population

PULSE project

H2020 - 727816

UNIPD

January 2018

DOCUMENT INFO

0.1 AUTHORS

Author	Organization	e-mail
Martina Vettoretti	University of Padova	martina.vettoretti@dei.unipd.it
Enrico Longato	University of Padova	enrico.longato@dei.unipd.it
Andrea Facchinetti	University of Padova	andrea.facchinetti@dei.unipd.it
Barbara Di Camillo	University of Padova	barbara.dicamillo@dei.unipd.it

0.2 DOCUMENT KEYDATA

Key words	H2020 – 727816 – PULSE Deliverable 5.2	
Editor info	Name	Martina Vettoretti
	Organization	UNIPD
	e-mail	martina.vettoretti@dei.unipd.it

0.3 DOCUMENT HISTORY

Date	Version	Contributor	Change	Status
10/12/2017	1.0	UNIPD	First template	Draft
31/12/2017	1.1	UNIPD	Draft with partial contents	Draft
05/01/2018	1.2	UNIPD	First complete draft	Draft
12/01/2018	1.3	UNIPD	Revised draft	Draft
26/01/2018	1.4	UNIPD	Final version after internal revision	Final

0.4 DISTRIBUTION LIST

Date	Issue	Distribution list
12/01/2018	Circulate first complete draft for internal revision	Cecilia Vera Muñoz, Riccardo Bellazzi, Vladimir Urosevic
26/01/2018	Final version	All Consortium and the European Commission

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

The purpose of this deliverable, entitled “Selected models implementation and recalibration against each specific population”, is to report the activities of task 5.2 about the implementation and recalibration of state-of-the-art risk models for the prediction of type 2 diabetes and asthma onset, which were previously identified in task 5.1. In particular, in task 5.1 two datasets available to the Consortium were identified, i.e. those of the Health and Retirement Study (HRS) and the Multi-Ethnic Study of Atherosclerosis (MESA). Due to delays in MESA data availability, only the activities related to model implementation and recalibration on the HRS dataset could be completed by the end of December 2017 and presented in this deliverable. After a brief introduction (Section 1), in this document we will describe in detail the steps conducted in the implementation and recalibration of two literature prediction models of diabetes onset (i.e. the FINDRISC and the DPoRT) on HRS data and the obtained results (Section 2). Finally, we will briefly introduce the activities related to model implementation and recalibration on the MESA dataset (Section 3), postponing their complete description to deliverable D5.3.

1. INTRODUCTION

In task 5.1, a literature review of the prediction models of type 2 diabetes (T2D) onset and asthma adult-onset was presented, and some of them were selected to be implemented and recalibrated on the datasets available in the Consortium in task 5.2. The objective of the state-of-the-art model implementation task is to compare the performance of current state-of-the-art T2D and asthma prediction models on the same datasets, which were identified in task 5.1. As later discussed in this deliverable, model recalibration is required in order to maximize the model performance on the available datasets. These models are not intended to be implemented in the PulseAIR app or the Pulse Analytics; rather, they will represent the reference for tasks 5.3 and 5.4, which are focused on the development of new models. Specifically, in task 5.3, the variables known or suspected to be associated with the risk of developing T2D and asthma will be selected from the available datasets and the probabilistic relationship between them will be mathematically studied by means of Bayesian networks. The selected pool of variables will also include variables not previously considered by state-of-the-art models. Then, in task 5.4, new prediction models based on survival analysis and dynamic Bayesian networks will be developed and their performance will be compared with that of the state-of-the-art models assessed in task 5.2. Of note, the critical aspects related to model implementation and variable accessibility in the Pulse architecture will be taken into account in tasks 5.3 and 5.4. Finally, the integration of the models developed in task 5.4 into the Pulse system will be carried out in task 5.5.

In task 5.1, two datasets for prediction models implementation were identified, i.e. the Health and Retirement Study (HRS) and the Multi-Ethnic Study of Atherosclerosis (MESA) datasets. While the HRS dataset became readily available after the distribution agreement was signed, obtaining access to the MESA dataset required a longer process, including the submission of a research proposal and its approval by the Ethics Board. Access to the MESA dataset was obtained on November 30th, 2017, thus it was not possible to complete the activities of model implementation and recalibration on MESA data by the due date of this deliverable, i.e. end of December 2017. For this reason, in this deliverable we will only present our activities about model implementation and recalibration on HRS data, remanding the description of model implementation and recalibration on MESA data to the next deliverable (D5.3). In particular, based on the variables available in HRS, two models to predict the onset of diabetes were selected to be implemented and recalibrated in this dataset, i.e. the FINDRISC score by Lindström and Tuomilehto [1] and the DPoRT model by Rosella et al. [2].

Before presenting the implementation and recalibration of these models on HRS data (Section 2), in the following sections we will introduce the problem of prediction model recalibration (Section 1.1) and summarize the main features of the HRS dataset (Section 1.2).

1.1. RECALIBRATION OF PREDICTION MODELS

Recalibration is a process required when transporting a prediction model from one setting to another, e.g. to a different population. Indeed, as several studies demonstrated [3]-[5], the application of prediction models to a different dataset from that on which the model was trained often results in unsatisfactory model performance, mainly caused by differences in the variables and outcome definition and different population characteristics.

For this reason, several recalibration strategies were proposed in the literature that aim to update existing prediction models for their application to new populations [6][7]. In work by Kangne et al. [3], for example, a simple model recalibration strategy was adopted, consisting in adjusting the model intercept according to observed and predicted outcome's incidence in the studied population. Approaches like this, in which few (e.g. only one) parameters of the model are updated, while the others remain as in the original model, are the only possible recalibration strategies when testing populations present low numerosity, thus the number of subjects available is not sufficient to reliably re-identify the full model and updates all its parameters.

In our work, as the available datasets are sufficiently large, we decided to adopt a full recalibration strategy, in which all the model parameters are re-estimated on the new populations (on appropriate training set).

1.2. THE HRS DATASET

The HRS is a longitudinal study of health, retirement and aging conducted in the United States. The HRS is supported by the National Institute of Aging (NIA U01AG009740) and the Social Security Administration. The HRS dataset includes nationally representative public survey data collected every 2 years since 1992 (wave 1 of the study) to 2014 (wave 12 of the study) in males and females of age 51 or older. The HRS sample is composed by 6 cohorts which entered the study at different years (or waves of data). The study waves of each HRS entry cohort are summarized in Table 1. In total, 37,495 respondents were interviewed during the study.

Public survey data includes information on demographics, family structure, physical health (e.g. current health status and medical history), cognition, functional limitations, disability, physical measures (e.g. weight, height, blood pressure), current and previous employment, health services and insurance, assets and income, internet use and social security. In addition to public survey data, some blood-based biomarkers were collected in 2006 (wave 8) and 2010 (wave 10) for a subgroup of participants, in 2008 (wave 9) and 2012 (wave 11) for another subgroup of participants. Measured biomarkers were total cholesterol, HDL cholesterol, HbA1c, C-reactive protein, cystatin C.

As already discussed in deliverable D5.1, only diabetes prediction models can be implemented on HRS data. Indeed, information on diabetes diagnosis and use of diabetes medication was collected in the physical health section of the survey, but, unfortunately, no information on adult-onset asthma was gathered in the study. As a consequence, models of asthma onset cannot be implemented on HRS data. Based on the variables available in HRS, only two of the state-of-the-art diabetes prediction models can be implemented on this dataset, i.e. the FINDRISC score (both the “concise” version and the “concise” version with addition of physical activity) and the DPoRT model. Other prediction models, e.g. the Framingham score, cannot be implemented because some of the variables they require, e.g. family history of diabetes and fasting plasma glucose, were not collected in HRS.

Table 1. Study waves of the 6 HRS entry cohorts.

Wave	Year	Entry cohort					
		HRS	AHEAD	CODA	WB	EBB	MBB
1	1992	X	X	-	-	-	-
2	1994	X	X	-	-	-	-
3	1996	X	X	-	-	-	-
4	1998	X	X	X	X	-	-
5	2000	X	X	X	X	-	-
6	2002	X	X	X	X	-	-
7	2004	X	X	X	X	X	-
8	2006	X	X	X	X	X	-
9	2008	X	X	X	X	X	-
10	2010	X	X	X	X	X	X
11	2012	X	X	X	X	X	X
12	2014	X	X	X	X	X	X

2. MODEL IMPLEMENTATION AND RECALIBRATION IN HRS

2.1. GENERAL APPROACH AND PERFORMANCE METRICS

The FINDRISC and DPoRT models were implemented and recalibrated on HRS data by performing the following 7 steps:

A. Data selection and pre-processing: a suitable subsample of HRS data was selected for the analysis. Specifically, the selected data for each model comprised the subjects who met the following three inclusion criteria: i) complete information on model independent variables was available; ii) they were free from diabetes at the baseline; iii) follow-up information on the outcome (i.e., the onset of diabetes) was available.
B. Variable pre-processing: model variables were appropriately homogenised to fit their definition in the state-of-the-art models. In addition, independent variables were discretized according to the same criteria adopted in the state-of-the art models.
C. Training and test set definition: selected data were split into a training set and a test set, stratifying for diabetes incidence.
D. Model recalibration: the model parameters were estimated on the training set.
E. Performance assessment: the model discriminatory ability was assessed on the test set.

F. Validation via bootstrap resampling: the effect of different training-test splits on model performance was estimated by repeating steps C-E 100 times within the training test, using bootstrap resampling to train the model on 100 different subsets of the training set.
G. Comparison with the original model: the performance of the recalibrated model was compared to the performance of the original state-of-the-art model.

The performance of the prediction models was determined by assessing their discriminatory ability, i.e. their ability to correctly rank the subjects according to their risk of diabetes onset. Two metrics were considered: the area under the receiver-operating characteristic curve (AU-ROC) and the concordance index (C-index). AU-ROC is a metric commonly used to assess classifiers or rankers, like prediction models and risk scores. In particular, in the case of a ranker (like FINDRISC) in which higher scores are attributed to subjects at risk for a certain clinical outcome (in this case, diabetes outcome), a threshold can be defined such that only subjects with scores higher than the threshold are classified as “at risk”. In this setting, the ROC curve represents the plot of the true positive rate (sensitivity) vs. the false positive rate (1-specificity) of the assignment to the “at risk class” for different values of the threshold. The AU-ROC is the area under the ROC curve and, as such, it varies between 0 and 1, with 0.5 corresponding to a random assignment of the scores. The greater the area under the ROC curve, the more accurately discriminatory the score. It can be demonstrated that the AU-ROC is equal to the probability that a subject chosen at random from the positive outcome group (in this case, the positive outcome is diabetes onset) is ranked higher than a subject chosen at random from the negative outcome group [8].

The C-index, proposed by Harrell et al. [9], is an extension of AU-ROC to be used when information on model outcome is available over time. In the case of HRS data, information on the outcome, i.e. diabetes onset, was collected every 2 years. In this setting, the time to event is defined as the time at which the subject first reported the outcome, for the subjects who developed diabetes, and as the time of their last follow-up interview for those who did not. Then, the C-index is defined as the probability that subjects with lower risk score have higher observed time to event, given that the order of two observed times to event can be validly inferred. Values of C-index near 0.5 indicate that the predictive model is no better than tossing a coin in determining which subject will experience the event first, while values of C-index near 0 or 1 indicate the predictive model has good discriminatory ability.

2.2. IMPLEMENTATION AND RECALIBRATION OF FINDRISC

2.2.1. THE ORIGINAL MODEL

The FINDRISC is a risk assessment tool for onset of drug-treated T2D, which is based on easily available individual information that can be collected by questionnaires on medical history and health behaviour and a simple clinical examination without any laboratory tests. The FINDRISC was developed by Lindström and Tuomilehto [1] on the data of 4746 Finnish subjects (aged 34-64, not on antidiabetic drug therapy) who responded to a baseline survey in 1987 and a follow-up survey in 1997. These data were used to fit a logistic regression model with drug-treated diabetes at follow-up as the dependent variable and 7 known risk factors for diabetes as independent variables, i.e. age, BMI, waist circumference, use of blood pressure medication, history of high blood glucose/diabetes, insufficient physical activity and less than daily consumption of fruits, vegetables, and berries. Based on the estimated β coefficients of the logistic regression, a risk score was assigned to each of the risk factors. The FINDRISC score was defined as the sum of the risk scores of each variable. In addition to a "full" model, comprising the entire list of variables, Lindström and Tuomilehto [1] also proposed a "concise" model from which physical activity and fruit and vegetables consumption were omitted as they had not demonstrated a statistically significant association with drug-treated diabetes after the assessment of the "full" model. External validation of the FINDRISC was performed by Lindström and Tuomilehto [1] in 4615 not drug-treated subjects that responded to a baseline survey in 1992 and were observed over a follow-up of 5 years for incidence of drug-treated diabetes.

2.2.2. DATA SELECTION AND PRE-PROCESSING

The FINDRISC was implemented and recalibrated using a subsample of HRS data including subjects who i) did not have drug-treated diabetes at the baseline interview, ii) had complete information on the FINDRISC risk factors at the baseline interview and iii) had information on drug-treated diabetes over a sufficient follow-up time.

Definition of the outcome

The FINDRISC was proposed as a tool to predict drug-treated diabetes. In HRS, drug-treated diabetes was assessed at each interview by the following questions:

Q1. Has a doctor ever told you that you had diabetes or high blood sugar? (for subjects who had never reported having had diabetes at previous interviews) / Our records from your last interview show that you have had diabetes or high blood sugar (for subjects who had reported having had diabetes at their previous interview)
Q2. In order to treat or control your diabetes, are you now taking medication that you swallow?
Q3. Are you now using insulin shots or a pump?

where Q2 and Q3 were asked only to subjects who reported having ever had diabetes or high blood sugar in Q1 (i.e. subjects who answered "yes" to Q1 or disputed the answer they had given to Q1 at their previous interview by stating they had in fact had diabetes). In other words, the answer to Q2 and Q3 is undefined for subjects who did not report having ever had diabetes or high blood sugar. For our purpose, the reasonable assumption was made that subjects who did not report having ever had diabetes or high blood sugar did not take any medications (e.g. insulin) for treating diabetes either. Therefore, drug-treated diabetes was defined as a binary variable equal to 1 for subjects who answered "yes" to either Q2 or Q3 or both, 0 for subjects who did not report diabetes or high blood sugar at Q1, or subjects who reported diabetes or high blood sugar at Q1 but answered "no" to both Q2 and Q3, and NA (i.e. missing) for subjects that did not answered to Q1 (e.g. subjects who answered "don't know" or who refused to answer).

Definition of the model independent variables

All the independent variables included in FINDRISC were collected in HRS, except for daily consumption of fruits, vegetables and berries. Therefore, only the FINDRISC concise model and the FINDRISC concise model with addition of physical activity can be implemented on HRS data. This is not a big issue since in work by Lindström and Tuomilehto [1] daily consumption of fruits, vegetables, and berries showed only a weak, non-statistically significant, association with the outcome.

Independent variables were extracted from the RAND data files, a set of user-friendly data files constructed by the RAND Center for the Study of Aging which contains most of the variables originally collected in HRS [10]. The RAND data files were used instead of the HRS raw data files because in the RAND data files most of variables collected in different HRS waves were homogenized and renamed with user-friendly labels, missing values of static variables (e.g. ethnicity, gender, birth date etc.) were imputed from other waves and between-wave inconsistencies, e.g. in medical history variables, were appropriately solved.

In particular, age, BMI and history of high blood glucose or diabetes and physical activity were extracted from the RAND HRS Data File (version P) [11], while waist circumference and use of antihypertensive medication were extracted from RAND Enhanced Fat Files [12]. Concerning physical activity, three variables were extracted representing the frequency of vigorous, medium and light physical activity, respectively. The use of antihypertensive medication was derived following a similar procedure as that described in the "Definition of outcome" section. Since subjects were asked whether they were taking any antihypertensive medication only if they had ever been diagnosed with high blood pressure, antihypertensive medication was defined as a binary variable equal to 1 for subjects who reported use of such medication, 0 for subjects reporting they had never had high blood pressure, or subjects that reported having had high blood pressure but stated they were not taking any medication for it, and NA (i.e. missing) for subjects that did not answer the first question.

Definition of the baseline

The HRS study involved 6 cohorts with different entry years. If any of the independent variables required for FINDRISC implementation or the information on diabetes medication were missing at the entry year, the baseline was shifted ahead in time till a wave with complete information on FINDRISC variables was found.

Data selection

The HRS dataset includes data regarding a total of 37,495 subjects. Since waist circumference was only measured from wave 8 (in a subset of subjects), 10,959 subjects were excluded because they exited the study before wave 8. An additional 178 subjects were excluded because they entered the study at wave 12 (last wave), and thus no follow-up data was available for them. Another 5,963 subjects were excluded because they had at least one missing value on FINDRISC variables at each wave, or the first wave at which they presented complete information on FINDRISC variables was wave 12.

Of the remaining 20,573 subjects, 5,781 subjects were excluded due to them having a follow-up period of <4 years (i.e. 0 or 2 years), that is not sufficient to observe a significant number of cases of incident diabetes. Finally, 2,338 subjects were excluded because they reported drug-treated diabetes at baseline, and 9 subjects were excluded because no follow-up information on drug-treated diabetes was available for them.

The remaining sample included 12,445 subjects without drug-treated diabetes at baseline and follow-up information on drug-treated diabetes at 4, 6 or 8 years after that. In particular, 12,205 subjects had a follow-up at 4 years (692 cases of incident diabetes), 9,029 subjects had a follow-up at 6 years (686 cases of incident diabetes), 4,425 subjects had a follow-up at 8 years (407 cases of incident diabetes). In the original work by Lindström and Tuomilehto [1], the FINDRISC score was developed by considering a 10-year follow-up period, i.e. it was tuned to predict the 10-year risk of drug-treated diabetes. However, it was also successfully validated on datasets where the follow-up period was only 5 years. Provided that in HRS no subsample of subjects with a 10-year follow-up was suitable for FINDRISC implementation, a shorter follow-up time had to be considered: either 4, 6 or 8 years. As the percentage of newly detected cases of drug-treated diabetes was unsatisfactorily low (5.5%) after 4 years from the baseline observation, and the number of subjects for whom a 8-year follow-up was available was too exiguous, a 6-year follow-up was selected as a good trade-off between the necessity of having both the largest possible sample size and a significant number of cases of incident drug-treated diabetes.

In conclusion, the selected subsample for FINDRISC implementation and recalibration included 9,029 subjects free of drug-treated diabetes at baseline, with complete information on FINDRISC independent variables at baseline, and information on drug-treated diabetes after 6 years.

2.2.3. MODEL IMPLEMENTATION AND RECALIBRATION

Variable discretization

The variables required for FINDRISC implementation were discretized as described in work by Lindström and Tuomilehto [1] with some modifications required because of the different characteristics of the HRS data. Two classes were defined for age, i.e. 45-54 and >=55. In the original FINDRISC score, the second age class was limited to subjects aged 55-64, because the study in which the score was originally developed did not include participants older than 64. In our model implementation, this class was extended to >=55 in order to also include the subjects over 64 participating in HRS.

Concerning physical activity, in the original score this variable was measured in hours per week and categorized into two classes, i.e. <4 hours/week and >=4 hours/week. In the HRS questionnaire, the frequency of physical activity, with different levels of intensity, was measured in times/week. In particular, subjects were asked how often they performed light, moderate and vigorous physical activity, the possible answers being every day, more than once a week, once a week, one to three times a month, hardly ever or never. For our purpose, subjects who reported performing any kind of physical activity more than once a week were assumed to perform physical activity at least 4 hours/week, while subjects who answered they performed physical activity once a week or less were assumed to perform less than 4 hours/week of physical activity.

Definition of the training and test sets

The data were split into a training and a test set, including the 80% and 20% of selected subjects, respectively. The percentage of subjects who developed drug-treated diabetes after 6 years was kept the same in both sets (i.e. 7.6%). In particular, the test set data were extracted by randomly sampling 20% of the subjects who did not develop drug-treated diabetes in the observed follow-up period and 20% of the subjects who did. The remaining subjects were included in the training set. See Table 3 in the Results section for details about training and test set subjects.

Recalibration in the training set

The training set data were used for model recalibration. A multivariable logistic regression model was fitted in the training set using the incidence of drug-treated diabetes at 6 years as the dependent variable, and age, BMI, waist circumference, use of antihypertensive medication, history of diabetes or high blood sugar and physical activity, all discretized as described above, as independent variables. Based on logistic regression β coefficients, partial scores were assigned to each variable, applying the same criteria adopted by Lindström and Tuomilehto [1] which are reported in Table 2. The recalibrated FINDRISC score in its concise version with the addition of physical activity was calculated by adding together all the partial scores. The logistic regression model was also fitted excluding physical activity from the independent variables and, according to the estimated β coefficients, a recalibrated version of the FINDRISC concise score was derived.

Table 2. Point assignment criterion in the FINDRISC

β coefficient	Points
0.01 – 0.2	1
0.21 – 0.8	2
0.81 – 1.2	3
1.21 – 2.2	4
>2.2	5

Performance assessment

The discriminatory ability of the recalibrated score was assessed by calculating AU-ROC and C-index on test set data. In order to determine how much the discriminatory ability of the recalibrated score is affected by the particular training/test set split performed, the mean and standard deviation of AU-ROC and C-index were calculated for the recalibrated score in 100 validation sets obtained by applying bootstrap resampling [13] to the training set. More in detail, a number of subjects equal to the training set size was sampled with replacement from the training set itself. This sample was used to recalibrate the logistic regression model and derive the recalibrated score. The remaining subjects were used as a validation set to assess AU-ROC and C-index. This procedure was repeated for 100 iterations. Finally, the mean and standard deviation of AU-ROC and C-index across the 100 iterations were computed. These statistics are indicative of the performance of the model across different training/test set splits.

The performance of the recalibrated score was also compared to those of the recalibrated logistic model and the original FINDRISC score. When the recalibrated logistic model was used, the probability of having 6-year incidence of drug-treated diabetes was predicted for each subject using the logistic regression model equation:

$$\hat{p}(\boldsymbol{X_j}) = \frac{e^{\hat{\beta}_0 + \boldsymbol{X_J^T} \cdot \hat{\boldsymbol{\beta}}}}{1 + e^{\hat{\beta}_0 + \boldsymbol{X_J^T} \cdot \hat{\boldsymbol{\beta}}}} \qquad (1)$$

where $\mathbf{X}_j$ is the column vector of independent variables, $\hat{\boldsymbol{\beta}}$ is the column vector of related estimated coefficients and $\hat{\beta}_0$ is estimated model intercept.

2.2.4. RESULTS

Data were divided into a training test, including 7,223 subjects, and a test set, including 1,806 subjects. As shown in Table 3, the FINDRISC variables at baseline present similar distributions in the training and test set. Both in the training and the test sets, the percentage of subjects that develop new drug-treated diabetes during the follow-up period is about 7.6%.

Table 3. Distribution of FINDRISC variables in training and test set reported as percentage of subjects in different variable categories

Variable	Category	% Subjects training set (N=7,223)	% Subjects test set (N=1,806)
Age [years]	45-54	11.2%	9.8%
	≥55	87.8%	88.8%
BMI [kg/m^2]	25 to <30	39.8%	39.5%
	≥30	28.1%	27.4%
Waist circumference [cm]	Men: 94 to <102 Women: 80 to <88	21.8%	21.4%
	Men: ≥102 Women: ≥88	61.3%	62.4%
Use of blood pressure medication [Boolean]	Yes	45.5%	46.0%
History of high blood glucose [Boolean]	Yes	3.2%	3.5%
Physical activity [hours/week]	<4	80.0%	80.5%
6-year incidence of drug-treated diabetes [Boolean]	Yes	7.6%	7.6%

The logistic regression coefficients estimated in the training set and related score points are reported in the third column of Table 4 for the concise model with physical activity, and in Table 5 for the concise model. Table 4 and Table 5 also report the coefficients and score points of original full and concise models derived in the original work by Lindström and Tuomilehto [1]. Considering the recalibrated version of the concise model with physical activity, we can observe that the effect of all the considered independent variables is in the same direction (same sign of the coefficient) as in the original model. However, some differences in the model coefficients result in some differences in the score points. In particular, the "BMI 25-30" class is assigned 2 points in the recalibrated score vs 1 point in the original score. Both the waist circumference classes are assigned lower points in the recalibrated model than in the original model, i.e. 1 and 2 points vs 3 and 4 points, respectively. Finally, less than 4 hours/week of physical activity are assigned 1 point in the recalibrated model vs 2 points in the original model. The points of the other variables are the same in the recalibrated and original model. Similar coefficients were estimated in the recalibrated concise model compared to the recalibrated concise model with physical activity, thus, the same observations made for the concise model with physical activity hold when we compare the recalibrated concise score with the original concise score.

Table 4. Coefficients and points of the recalibrated FINDRISC concise model with physical activity (third and fourth column) compared to the coefficients and points of the original FINDRISC full model (fifth and sixth column).

Variable	Value	Recalibrated coefficient	Recalibrated points	Original coefficient	Original points
Intercept	-	-4.716	-	-5.658	-
Age [years]	45-54	0.789	2	0.650	2
	55-64	0.808	3	0.940	3
BMI [kg/m^2]	25 to <30	0.403	2	0.015	1
	≥30	0.971	3	0.938	3
Waist circumference [cm]	Men: 94 to <102 Women: 80 to <88	0.062	1	1.021	3
	Men: ≥102 Women: ≥88	0.532	2	1.424	4
Use of blood pressure medication [Boolean]	Yes	0.386	2	0.714	2

Variable	Value	Recalibrated coefficient	Recalibrated points	Original coefficient	Original points
History of high blood glucose [Boolean]	Yes	2.955	5	2.263	5
Physical activity [hours/week]	<4	0.061	1	0.268	2
No daily consumption of vegetables, fruits or berries [Boolean]	Yes	-	-	0.165	1

Table 5. Coefficients and points of the recalibrated FINDRISC concise model (third and fourth column) compared to the coefficients and points of the original FINDRISC concise model (fifth and sixth column).

Variable	Value	Recalibrated coefficient	Recalibrated points	Original coefficient	Original points
Intercept	-	-4.669	-	-5.514	-
Age [years]	45-54	0.787	2	0.628	2
	55-64	0.806	3	0.892	3
BMI [kg/m^2]	25 to <30	0.401	2	0.165	1
	≥30	0.969	3	1.096	3
Waist circumference [cm]	Men: 94 to <102 Women: 80 to <88	0.062	1	0.857	3
	Men: ≥102 Women: ≥88	0.540	2	1.350	4
Use of blood pressure medication [Boolean]	Yes	0.387	2	0.711	2
History of high blood glucose [Boolean]	Yes	2.957	5	2.139	5

Model performance in terms of discriminatory ability are summarized in Table 6 and Figure 1 for the model with physical activity, and in Table 7 and Figure 2 for the concise model. The AU-ROC and C-index values obtained by the model with physical activity in the validation phase (mean ± over the 100 bootstrap resamplings) and the test set are reported in Table 6 for the recalibrated score, the recalibrated logistic regression and the original score. Best performance is achieved by the recalibrated logistic regression, with validation AU-ROC equal to 0.7459 (± 0.0158) and test set AU-ROC equal to 0.7492. Slightly lower performance is achieved by the recalibrated score, with validation AU-ROC equal to 0.7328 (± 0.0161) and test set AU-ROC equal to 0.7482. Both the recalibrated logistic regression and the recalibrated score show slightly better performance than the original score whose validation AU-ROC is 0.7306 (± 0.0151) and test set AU-ROC is 0.7369. Similar observations can be made according to C-index. The ROC profiles obtained by the model with physical activity on the test set are reported in Figure 1 for the recalibrated score (green solid line), the recalibrated logistic regression (red solid line) and the original score (blue solid line). After an initial phase in which the three curves exactly overlap, the recalibrated score and recalibrated logistic regression curves are generally steeper (which is indicative of better performance).

When physical activity is not included in the model, the concise model performance in the validation set are very close to those of the model with physical activity (see the AU-ROC and C-index values reported in Table 7). In the test set, however, the concise model presents slightly higher AU-ROC and C-index values compared to the model with physical activity for both the recalibrated score/model and the original score.

Table 6. Performance of the of the FINDRISC concise model with physical activity: AU-ROC and C-index for the recalibrated score, recalibrated logistic regression and original score assessed during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	Metric	Validation	Test set
Recalibrated score	AU-ROC at 6 years	0.7328 (± 0.0161)	0.7482
	C-index	0.7213 (± 0.0148)	0.7344
Recalibrated logistic regression	AU-ROC at 6 years	0.7459 (± 0.0158)	0.7492
	C-index	0.7358 (± 0.0139)	0.7397
Original score	AU-ROC at 6 years	0.7306 (± 0.0151)	0.7369
	C-index	0.7214 (± 0.0133)	0.7249

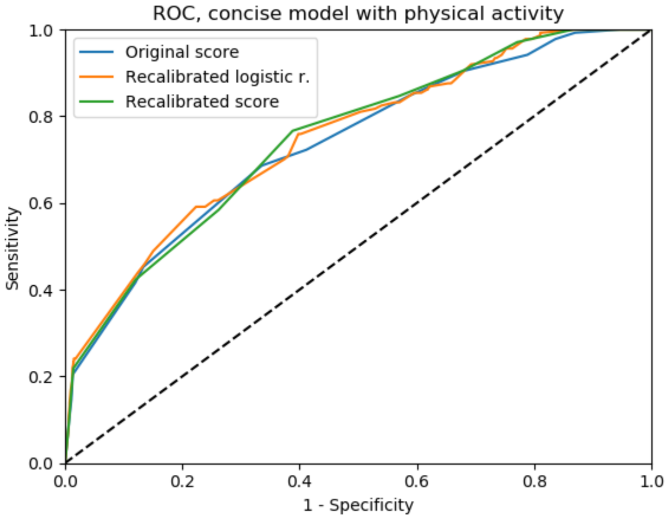

Figure 1. Performance of the FINDRISC model with physical activity: receiver-operating curve for the recalibrated score (green solid line), the recalibrated logistic regression (red solid line), and the original score (blue solid line) assessed on the test set.

Table 7. Performance of the of the FINDRISC concise model: AU-ROC and C-index for the recalibrated score, recalibrated logistic regression and original score assessed during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	Metric	Validation	Test set
Recalibrated score	AU-ROC at 6 years	0.7343 (± 0.0157)	0.7514
	C-index	0.7230 (± 0.0143)	0.7363
Recalibrated logistic regression	AU-ROC at 6 years	0.7464 (± 0.0158)	0.7510
	C-index	0.7364 (± 0.0138)	0.7409
Original score	AU-ROC at 6 years	0.7348 (± 0.0153)	0.7462
	C-index	0.7260 (± 0.0143)	0.7316

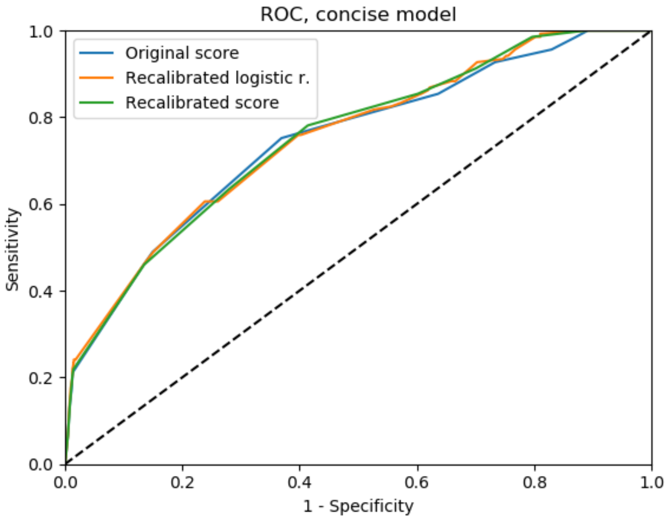

Figure 2. Performance of the FINDRISC concise model: receiver-operating curve for the recalibrated score (green solid line), the recalibrated logistic regression (red solid line), and the original score (blue solid line) assessed on the test set.

2.3. IMPLEMENTATION AND RECALIBRATION OF DPORT

2.3.1. THE ORIGINAL MODEL

The Diabetes Population Risk Tool (DPoRT) is a population-based risk prediction tool developed by Rosella et al. [2] to predict T2D onset using national survey data. The DPoRT was derived using the data of the participants from Ontario of the 1996/7 National Population Health Survey conducted by Statistics Canada. Such data included the records of 9177 male and 10618 female subjects free of diabetes at baseline who could be individually linked to a registry of physician-diagnosed diabetes. The data were used to fit a Weibull accelerated failure time model separately for men and women. In particular, variable selection according to predictive significance was performed separately for men and women. The variables selected for inclusion in the model were: age, ethnicity, education, smoking, BMI, hypertension and heart disease for men; age, ethnicity, education, immigrant status, BMI and hypertension for women. The DPoRT was externally validated in a cohort of 9899 subjects with 9-year follow-up and a cohort of 26465 subjects with 5-year follow-up.

2.3.2. DATA SELECTION AND PRE-PROCESSING

The DPoRT was implemented and recalibrated using a subsample of HRS data including subjects who i) did not have physician-diagnosed diabetes at the baseline interview, ii) had complete information on the DPoRT risk factors at the baseline interview and iii) had information on physician-diagnosed diabetes at least at one of the follow-up interviews.

Definition of the outcome

In HRS, physician-diagnosed diabetes was assessed at each interview by the following question:

Q1. Has a doctor ever told you that you had diabetes or high blood sugar? (for subjects who never reported diabetes at previous interviews) / Our records from your last interview show that you have had diabetes or high blood sugar (for subjects who reported having had diabetes at last interview)

The answer to Q1 was coded in a variable in the RAND HRS Data File (v.P), representing whether a doctor had ever told the respondent he or she had high blood glucose or diabetes. This variable was used as the outcome for the DPoRT model.

Definition of model independent variables

All the independent variables needed for DPoRT were collected in HRS and included in the RAND HRS Data File (v.P). In particular, sex, age, BMI, ethnicity, education level and smoking habits were recorded directly. To represent hypertension, we used a variable indicating whether or not a doctor had ever told the respondent he or she had high blood pressure or hypertension. Analogously, heart disease was defined using a variable which indicated whether a doctor had ever told the respondent he or she had had heart attack, coronary heart disease, angina, congestive heart failure, or other heart-related problems. Finally, immigrant status was derived from a variable representing the respondent's birth place. In particular, immigrant status was set to 1 for all the subjects who were born outside the U.S., and to 0 for the others.

Definition of the baseline

The HRS study involved 6 cohorts with different entry years. If any of the independent variables required for DPoRT implementation or the information on physician-diagnosed diabetes was missing at the entry year, the baseline was shifted ahead in time till a wave with complete information was found.

Data selection

The complete HRS dataset included a total of 37,495 subjects. 178 subjects were excluded because no follow-up was available for them, as they entered the study at wave 12 (last wave). Another 352 subjects were excluded because they had at least one missing value on DPoRT variables at each wave, or the first wave at which they presented complete information on DPoRT variables was wave 12. Of the remaining 36,965 subjects, 4,493 were excluded because they presented physician-diagnosed diabetes at the baseline, and 2,198 because they did not have follow-up information on physician-diagnosed diabetes.

In conclusion, the subsample selected for DPoRT implementation and recalibration comprised 30,274 subjects free of physician-diagnosed diabetes at the baseline, with complete information on DPoRT independent variables at the baseline and information on physician-diagnosed diabetes at least at one of the follow-up waves. Table 8 and Table 9 show how the baseline wave and the duration of follow-up are distributed in the selected subsample. In this subsample, 4,725 subjects had new onset of physician-diagnosed diabetes during the follow-up period.

Table 8. Distribution of baseline waves in the HRS subsample selected for DPoRT implementation and recalibration.

Baseline wave	# subjects
1	10921
2	6249
3	218
4	4211
5	243
6	175
7	2845
8	231
9	128
10	4794
11	259

Table 9. Distribution of follow-up durations in the HRS subsample selected for DPoRT implementation and recalibration.

Follow-up duration	# subjects
2	2440
4	6140
6	1774
8	1800
10	3580
12	1364
14	1532
16	3432
18	929
20	1532
22	5751

2.3.3. MODEL IMPLEMENTATION AND RECALIBRATION

Variable discretization

The variables required for DPoRT implementation were discretized as described in the original work by Rosella et al. [2]. Regarding age and BMI, 10 classes were defined for men, using 45 as the cut-off value for age and 23, 25, 30, and 35 as the cut-off values for BMI, 15 classes were defined for women, using 45 and 65 as the cut-off values for age and 23, 25, 30, and 35 as the cut-off value for BMI. Ethnicity was discretized into two classes, i.e. “white/Caucasian” and “other”. Education was discretized in “less than post-secondary” and “post-secondary or higher”. Hypertension, heart disease and immigrant status were considered as binary variables, equal to 1 if the subject had the condition and 0 otherwise.

Definition of time to event

Survival analysis aims to model the time after which an event of interest happens. In our case the event was the onset of physician-diagnosed diabetes. We defined a variable representing the event that was equal to 1 for subject who developed physician-diagnosed diabetes at some point during follow-up period and 0 for subjects who were free of physician-diagnosed diabetes at the end of their observation time. For the latter category of subjects, the time of diabetes onset was not known: in survival analysis such data are commonly referred to as “right-censored”. Then we defined T as the variable representing the time to event or “survival time”. In particular, T was equal to the time of diabetes onset, i.e. the time at which the subject first reported being diagnosed with diabetes, for the subjects that experienced the event before the end of their follow-up period, and to the time of their last follow-up interview for those who did not. Time was expressed in days from the baseline, as in [2].

Definition of the training and test sets

The data were split into a training and a test set, including the 80% and 20% of selected subjects, respectively, stratified for incidence of physician-diagnosed diabetes and sex. In particular, the test set data were extracted by randomly sampling 20% of men and women who did not develop physician-diagnosed in the observed follow-up period and 20% of men and women who did. The remaining subjects were included in the training set.

Recalibration in the training set

The training set data were used for model recalibration. Two Weibull accelerated failure time models were fitted, one on the men and one on the women, using defined events and respective times, T, as outcome. The age-BMI categories, ethnicity, education level, smoking habits, hypertension, and heart disease were used as the independent variables for the men, while age-BMI, ethnicity, education level, immigrant status and hypertension were used as the independent variables for the women.

Performance assessment

The discriminatory ability of the recalibrated model was assessed on the test set by using the C-index. In particular, the recalibrated model was applied to calculate the survival probability of the j^{th} subject at time to event T_j, i.e. the probability of being free of physician-diagnosed diabetes at time T_j, by using the following equation:

$$\hat{s}\left(T_j, \boldsymbol{X_j}\right) = e^{-e^{\frac{log\left(T_j\right) - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \cdot \boldsymbol{X_j}}{\hat{\sigma}}}} \qquad (2)$$

where $\mathbf{X_j}$ was the vector of the subject-related covariates, $\hat{\boldsymbol{\beta}}^{T}$ was the vector of the corresponding model coefficients, $\hat{\beta}_0$ was the intercept parameter, and $\hat{\sigma}$ was the scale parameter. Then, the level of agreement between the estimated survival probability $\hat{s}$ and the observed time to event T was measured by computing the C-index [9].

Besides the C-index, the AU-ROC at fixed time points T* was also computed, with T* equal to 6, 10 and 14 years. The AU-ROC compares the model-predicted probability of having developed physician-diagnosed diabetes before/at time T*, which can be calculated as $1 - \hat{s}\left(T^*, \mathbf{X_j}\right)$ for the j^{th} subject, and the actual presence of physician-diagnosed diabetes at time T*. Note that, for this metric calculation, only the subjects that develop diabetes before/at T* were considered to have incident diabetes, while the subjects developing diabetes after T* were considered free from diabetes and their data treated as right-censored. Moreover, subjects with follow-up periods shorter than T* who did not develop diabetes during their observation time were excluded from the calculation of AU-ROC at time T* since their status at T* was unknown.

As we did for the FINDRISC, to determine how the discriminatory ability of the recalibrated model was affected by the particular training/test set split we performed, we calculated the mean and standard deviation over 100 bootstrap resamplings of the training set [6] of both the C-index and the AU-ROCs at fixed time points.

The performance of the recalibrated model was also compared to that of the original DPoRT model. In this case, the survival probability was computed using the model coefficients reported in the paper by Rosella et al. [2]. Then, the same procedure for the calculation of the C-index and the AU-ROCs at fixed times described above was applied to the predictions obtained with the original model.

2.3.4. RESULTS

Data were divided into a training test, including 10,396 men and 13,823 women, and a test set, including 2,599 men and 3,456 women. As shown in Table 10, for the men, and in Table 11, for the women, the DPoRT variables at the baseline present similar distributions in the training and test sets. Both in the training and the test sets, the percentage of subjects who developed physician-diagnosed diabetes during the follow-up period was about 16.7% for the men and 14.8% for the women.

Table 10. Distribution of DPoRT variables for the men in the training and test sets reported as percentage of men in each variable category.

Variable	Value	% Subjects training set (N=10,396)	% Subjects test set (N=2,599)
Hypertension [Boolean]	Yes	34.7%	33.7%
Non-white ethnicity [Boolean]	Yes	21.2%	22.0%
Heart disease [Boolean]	Yes	14.9%	14.7%
Current smoker [Boolean]	Yes	24.1%	26.3%
Education	Post-secondary or higher	44.1%	42.9%
BMI [kg/m^2], age [years]	BMI 23-24, age <45	0.3%	0.1%
	BMI 25-29, age <45	0.6%	0.4%
	BMI 30-34, age <45	0.4%	0.4%
	BMI ≥35, age <45	0.2%	0.2%
	BMI <23, age ≥45	11.1%	10.5%
	BMI 23-24, age ≥45	17.0%	17.1%
	BMI 25-29, age ≥45	46.1%	47.1%
	BMI 30-34, age ≥45	18.5%	18.8%
	BMI ≥35, age ≥45	5.9%	5.3%
Incidence of physician-diagnosed diabetes during follow-up [Boolean]	Yes	16.7%	16.7%

Table 11. Distribution of DPoRT variables for the women in the training and test sets reported as percentage of women in each variable category.

Variable	Value	% Subjects training set (N=13,823)	% Subject test set (N=3,456)
Hypertension [Boolean]	Yes	34.4%	33.9%
Non-white ethnicity [Boolean]	Yes	22.8%	24.7%
Immigrant status [Boolean]	Yes	11.8%	12.8%

Variable	Value	% Subjects training set (N=13,823)	% Subject test set (N=3,456)
Education	Post-secondary or higher	39.8%	40.4%
BMI [kg/m^2], age [years]	BMI 23-24, age <45	0.8%	0.7%
	BMI 25-29, age <45	1.5%	1.2%
	BMI 30-34, age <45	0.9%	0.7%
	BMI ≥35, age <45	0.7%	0.6%
	BMI <23, age 45-64	13.0%	13.6%
	BMI 23-24, age 45-64	9.7%	9.3%
	BMI 25-29, age 45-64	21.9%	22.6%
	BMI 30-34, age 45-64	11.7%	12.0%
	BMI ≥35, age 45-64	7.5%	8.2%
	BMI <23, age ≥65	8.5%	9.0%
	BMI 23-24, age ≥65	5.6%	5.3%
	BMI 25-29, age ≥65	11.2%	10.0%
	BMI 30-34, age ≥65	4.2%	3.8%
	BMI ≥35, age ≥65	1.3%	1.2%
Incidence of physician-diagnosed diabetes during follow-up [Boolean]	Yes	14.8%	14.8%

In Table 12 and Table 13, the coefficients of the original DPoRT model and of the DPoRT model recalibrated on the training set are reported for the men and the women, respectively. For the men, the effect of all the considered independent variables in the recalibrated model was in the same direction (same sign of the coefficient) as in the original model. For the women, almost all the coefficients had the same sign in the recalibrated and the original model, except the coefficient of the “BMI <23, age 45-64” class, which was positive in the original model and negative in the recalibrated model, and that of the “BMI ≥35, age ≥65” class, which was negative in the original model and positive in the recalibrated model. Regarding the absolute values of the coefficients, relevant differences were present between the recalibrated and the original model, especially for men. Such differences suggested that the original DPoRT model would not achieve optimal levels of performance on the HRS population.

Table 12. Coefficients of the DPoRT model recalibrated in the training set for men (third column) compared to coefficients of the original model (fourth column).

Variable	Value	Recalibrated model coefficient	Original model coefficient
Intercept	-	20.3138	10.5971
Hypertension [Boolean]	Yes	-0.3063	-0.2624
Non-white ethnicity [Boolean]	Yes	-0.2145	-0.6316
Heart disease [Boolean]	Yes	-0.1934	-0.5355
Current smoker [Boolean]	Yes	-0.1938	-0.1765
Education	Post-secondary or higher	0.0516	0.2344
BMI [kg/m^2], age [years]	BMI 23-24, age <45	-0.2972	-1.2378
	BMI 25-29, age <45	-10.2006	-1.5490
	BMI 30-34, age <45	-10.3726	-2.5437
	BMI ≥35, age <45	-11.1878	-3.4717
	BMI <23, age ≥45	-9.7221	-1.9749
	BMI 23-24, age ≥45	-10.0920	-2.4426
	BMI 25-29, age ≥45	-10.5263	-2.8588
	BMI 30-34, age ≥45	-10.9611	-3.3179
	BMI ≥35, age ≥45	-11.3147	-3.5857
Scale	-	0.6721	0.8049

Table 13. Coefficients of the DPoRT model recalibrated in the training set for women (third column) compared to coefficients of the original model (fourth column).

Variable	Value	Recalibrated model coefficient	Original model coefficient
Intercept	-	10.7653	10.5474
Hypertension [Boolean]	Yes	-0.3215	-0.2865
Non-white ethnicity [Boolean]	Yes	-0.1938	-0.4309
Immigrant status [Boolean]	Yes	-0.2513	-0.2930

Variable	Value	Recalibrated model coefficient	Original model coefficient
Education	Post-secondary or higher	0.1627	0.2042
BMI [kg/m^2], age [years]	BMI 23-24, age <45	-0.6690	-0.5432
	BMI 25-29, age <45	-0.9694	-0.8453
	BMI 30-34, age <45	-1.1769	-1.4104
	BMI ≥35, age <45	-1.3287	-2.0483
	BMI <23, age 45-64	-0.2169	0.0711
	BMI 23-24, age 45-64	-0.5851	-0.7011
	BMI 25-29, age 45-64	-1.0053	-1.4167
	BMI 30-34, age 45-64	-1.3319	-2.2150
	BMI ≥35, age 45-64	-1.4807	-2.2695
	BMI <23, age ≥65	-0.3274	-1.0823
	BMI 23-24, age ≥65	-0.6183	-1.1419
	BMI 25-29, age ≥65	-0.8724	-1.5999
	BMI 30-34, age ≥65	-1.1367	-1.9254
	BMI ≥35, age ≥65	0.1661	-2.1959
Scale	-	0.6689	0.7814

The performance levels of the original and the recalibrated DPoRT models are summarized in Table 14 for both the men and the women. According to the C-index, the recalibrated model performed significantly better than the original model both in the validation phase and on the test set (the C-index values obtained by the recalibrated model were always higher than those achieved by the original model). The original model performance was particularly unsatisfactory for the women, with a C-index less than 0.6 both in the validation phase and on the test set.

The AU-ROCs at 6, 10 and 14 years were also calculated on the subsamples of subjects for which the follow-up time is sufficient to define the condition of physician-diagnosed diabetes at 6, 10 and 14 years. In the test set, these subsamples comprised 1,864 men and 2,574 women, 1,588 men and 2,201 women, and 1,234 men and 1,741 women, respectively. The ROC curves calculated at 6, 10 and 14 years are reported in Figure 3, Figure 4, and Figure 5, respectively, for the recalibrated model (blue solid line) and the original model (red solid line) tuned on the men (left panel) and the women (right panel). The corresponding AU-ROC values are reported in Table 14. We can observe that the recalibrated model achieved comparable values of AU-ROC to the original model, whereas the C-index highlighted a higher capability of the recalibrated model to rank subjects according to their risk of developing diabetes.

Table 14. Performance of the of the DPoRT model: C-index and AU-ROC at 6, 10, and 14 years of the recalibrated model and the original model assessed in the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	**Metric**	**Men**		**Women**	
		Validation	**Test set**	**Validation**	**Test set**
Recalibrated model	C-index	0.7020 (±0.0105)	0.7109	0.6847 (±0.0091)	0.6567
	AU-ROC at 6 years	0.7206 (±0.0140)	0.7176	0.7411 (±0.0109)	0.7600
	AU-ROC at 10 years	0.7194 (±0.0116)	0.7034	0.7392 (±0.0091)	0.7549
	AU-ROC at 14 years	0.7219 (±0.0118)	0.6969	0.7317 (±0.0094)	0.7610
Original model	C-index	0.6664 (±0.0083)	0.6766	0.5691 (±0.0087)	0.5407
	AU-ROC at 6 years	0.7158 (±0.0139)	0.7051	0.7430 (±0.0101)	0.7619
	AU-ROC at 10 years	0.7141 (±0.0114)	0.6904	0.7406 (±0.0091)	0.7547
	AU-ROC at 14 years	0.7173 (0.0111)	0.6914	0.7309 (±0.0095)	0.7593

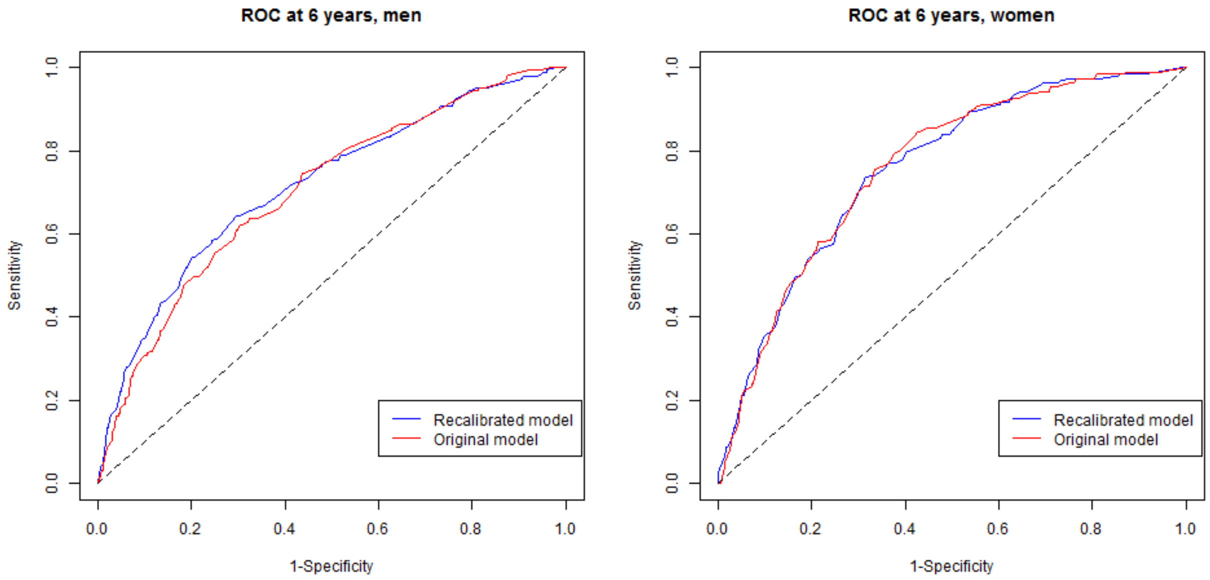

Figure 3. Performance of DPoRT model: ROC curve at 6 years of the recalibrated model (blue solid line) and the original model (red solid line) for the men (left panel) and the women (right panel).

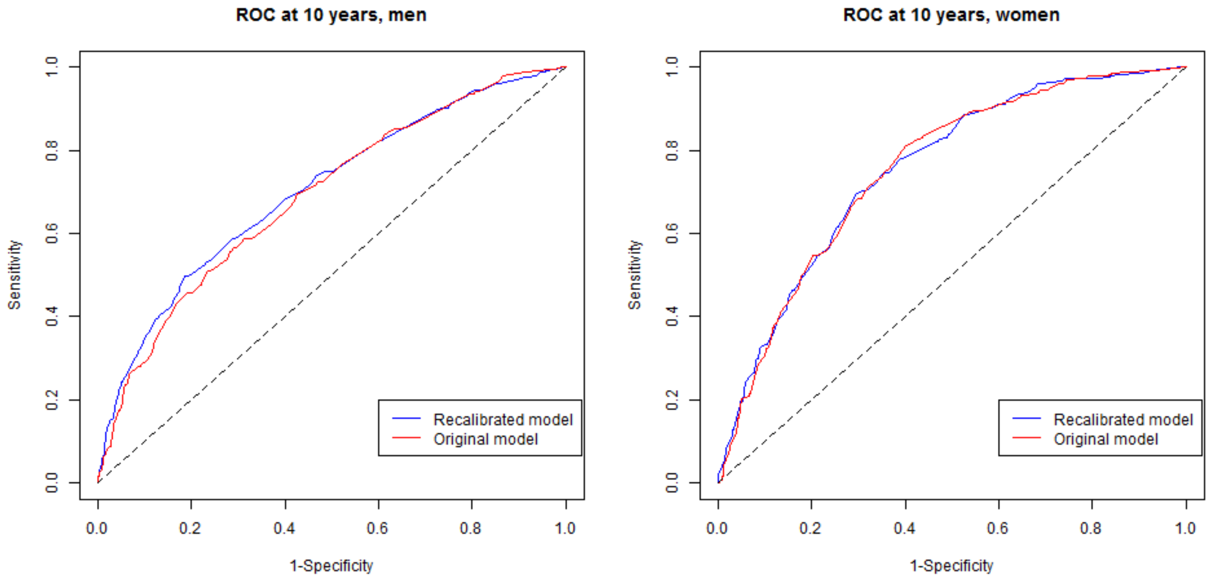

Figure 4. Performance of DPoRT model: ROC curve at 10 years of the recalibrated model (blue solid line) and the original model (red solid line) for the men (left panel) and the women (right panel).

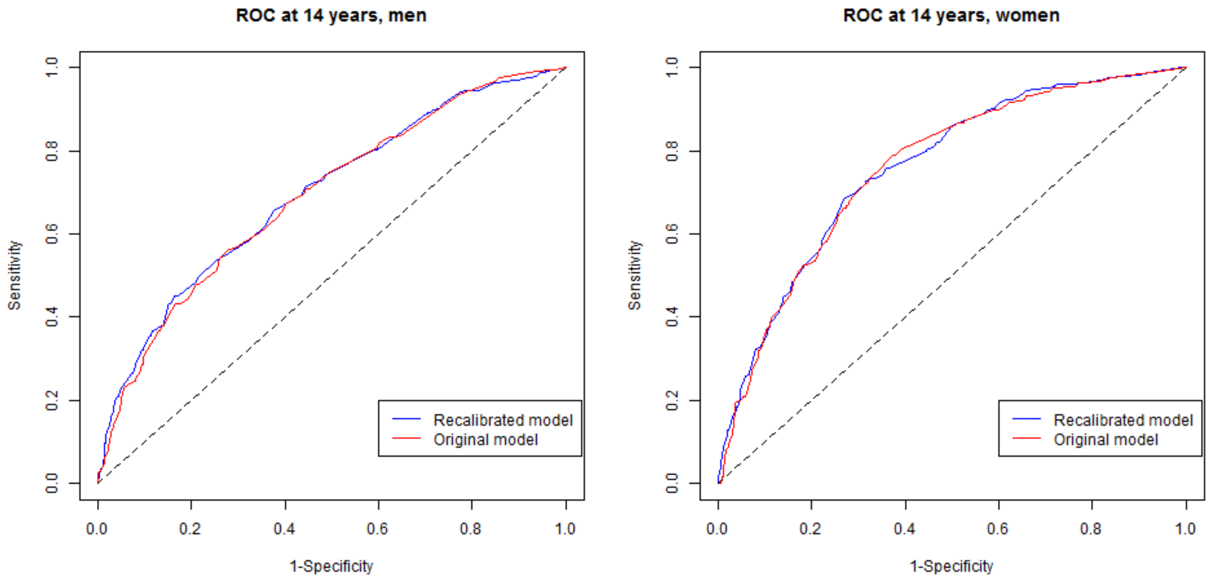

Figure 5 Performance of DPoRT model: ROC curve at 14 years of the recalibrated model (blue solid line) and the original model (red solid line) for the men (left panel) and the women (right panel).

2.4. DISCUSSION

Two non-invasive literature models to predict future diabetes were implemented and recalibrated on HRS data. The two models were designed to predict different diabetes outcomes, i.e. drug-treated diabetes for the FINDRISC and physician-diagnosed diabetes for the DPoRT, and, thus, cannot be directly compared.

The FINDRISC model was originally developed, using logistic regression analysis, to predict the onset of drug-treated diabetes at 10 years. In our work, the FINDRISC logistic regression model and the corresponding score were tuned and assessed using a subsample of the HRS data with 6 years of follow-up, since no suitable data with 10-year follow-up were available in that study. In particular, we recalibrated and assessed the concise version of the FINDRISC with and without considering physical activity as an independent variable (no information on fruit and vegetables consumption was collected in HRS, rendering the implementation of the full FINDRISC model impossible).

After recalibration, all the regression coefficients estimated on the training set presented the same sign of the coefficients of the original model developed by Lindström and Tuomilehto [1], indicating that the FINDRISC risk factors had the same kind of effect—increasing or decreasing— on drug-treated diabetes risk as in the original model. However, due to differences in the estimated coefficient absolute values, the recalibrated score presented different partial scores for BMI, waist circumference, and physical activity. Differences in the physical activity score may be caused by the different definition of physical activity categories performed in our study compared to the original model (see paragraph entitled “Variable discretization” of Section 2.2.3 for detailed description).

In an independent test set, the recalibrated logistic regression and the recalibrated score showed slightly better performance than the original score according to AU-ROC and C-index, especially in the case of the concise model with physical activity. Models without physical activity (both the original and the recalibrated score) presented comparable performance than the models with physical activity. This may be caused by the fact that the population interviewed in HRS only included older adults with reduced mobility who were mostly sedentary (about 80% of subjects performed less than 4 hours per week of physical activity). In this population, physical inactivity was only weakly associated with the risk of new incidence of drug-treated diabetes (estimated regression coefficient equal to 0.061), differently than in the younger population used for deriving the original FINDRISC score in which a stronger association was apparent (estimated regression coefficient equal to 0.268). This suggests that, in the HRS population, physical activity may be a confounding factor for many subjects (e.g. those who were recommended by a doctor to exercise because already at risk of diabetes or other pathologies), thus not improving model performance.

The DPoRT model was originally developed to predict the onset of physician-diagnosed diabetes at different time points in the future using a Weibull accelerated failure time model. In survival analysis, time is considered as an independent variable of the model and, thus, data with different follow-up periods can be modelled at the same time. This allowed us to recalibrate and assess the DPoRT model on the entire HRS data sample.

In the recalibration step, almost all the model coefficients, estimated on the training set, presented the same sign of the coefficients of the original model proposed by Rosella et al. [2], except the coefficient of the “BMI <23, age 45-64” and the “BMI ≥35, age ≥65” classes in the model tuned on the women. Such differences can be due to the different age-BMI composition of the HRS population compared to the development cohort of the original DPoRT model, which comprised a generally younger population (about 52% of the women were under 45) with on average lower BMI (about 40% of the women had BMI <23).

The performance of the recalibrated DPoRT model was assessed on an independent test set and compared to the original model by Rosella et al. [2]. According to the C-index, the recalibrated model performed better than the original model, both in men and women. Results were more controversial when considering AU-ROC at 6, 10, and 14 years, not indicating a clear superiority of the recalibrated model compared to the original model. However, we should consider the C-index as the most representative metric of model performance, as it was especially designed as an extension of the AU-ROC for survival analysis, and, thus, all the available data could be used for its calculation (on the contrary, only a subset of the data can be used to calculate AU-ROCs at a fixed time points in the presence of censored data).

As a final comment, both the FINDRISC and the DPoRT model showed lower discriminatory ability when implemented on HRS data than in the studies by Lindström and Tuomilehto [1] and Rosella et al. [2] in which the models were originally developed. This result was expected since the selection and categorization of model independent variables were optimized for the original development cohorts and can be suboptimal for the HRS population. The performances of FINDRISC and DPoRT on HRS data, after simple model recalibration, are reasonably good, yet not excellent. Since the HRS population mainly includes old people, we expect that a better variable selection and discretization scheme would improve the performance of the models. Such investigations will be carried out in the next tasks of work package 5.

3. MODEL IMPLEMENTATION AND RECALIBRATION IN MESA

Another dataset available in the Consortium for diabetes and asthma prediction model implementation is the MESA dataset. MESA is a longitudinal study funded by the National Heart, Lung, and Blood Institute starting in July 2000 and still ongoing. MESA investigates subclinical cardiovascular disease in a sample (n=6,814) of population consisting of African-Americans (27.8%), Hispanics (21.9%), Chinese (11.8%), and Whites (38.5%). Participants enrolled were both males and females aged 45-84 years. Data were collected from 6 U.S. communities (Baltimore City and Baltimore County, Maryland; Chicago, Illinois; Forsyth County, North Carolina; Los Angeles County, California; Northern Manhattan and the Bronx, New York; and St. Paul, Minnesota). In total, five exams were conducted in the period 2000-2012. At each exam, subjects were interviewed about their health and lifestyle and underwent some clinical assessments. Specifically, information on both diabetes and asthma diagnosis was collected in the survey by questions like "Has a doctor ever told you that you have diabetes/asthma?".

For its characteristics, the MESA dataset is suitable to implement and recalibrate both diabetes and asthma prediction models. In particular, the list of the state-of-the-art models that were selected in task 5.1 for implementation and recalibration on MESA data, according to the variables collected in this study, is reported in Table 15.

As mentioned at the beginning of Section 1, obtaining access to the MESA data required a long process, including the submission of a research proposal and its approval by the Ethics Board. Access to the MESA data was obtained on November 30th, 2017, therefore we did not have sufficient time to complete the activities of model implementation and recalibration on MESA data by the end of December 2017 and present them in this deliverable. The models listed in Table 15 will be implemented and recalibrated on MESA data using the same methodologies presented in Section 2 of this deliverable for FINDRISC and DPoRT implementation and recalibration on HRS. The description of model implementation and recalibration on MESA data and the obtained results will be included in the next deliverable (D5.3).

Table 15. State-of-the-art prediction models of T2D and asthma onset selected for implementation on MESA data.

T2D prediction models	Asthma prediction models
• Clinical model by Stern et al. [14] • FINDRISC [1] • ARIC models [15] • Framingham personal and clinical models [16] • GDR score [17] • Simple risk score by Kahn et al. [18] • DPoRT [2]	• Model by Thomsen et al. [19] • Model by Verlato et al. [20]

4. REFERENCES

[1] Lindström J. and Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003 Mar; 26(3):725-31.

[2] Rosella L.C., Manuel D.G., Burchill C., Stukel T.A., for the PHIAT-DM team. A Population-Based Risk Algorithm for the Development of Diabetes: Development and Validation of the Diabetes Population Risk Tool (DPoRT). J Epidemiol Community Health. 2011 Jul; 65(7):613-620.

[3] Kengne A.P., Beulens J.W., Peelen L.M., Moons K.G., van der Schouw Y.T., Schulze M.B., Spijkerman A.M., Griffin S.J., Grobbee D.E., Palla L., Tormo M.J., Arriola L., Barengo N.C., Barricarte A., Boeing H., Bonet C., Clavel-Chapelon F., Dartois L., Fagherazzi G., Franks P.W., Huerta J.M., Kaaks R., Key T.J., Khaw K.T., Li K., Mühlenbruch K., Nilsson P.M., Overvad K., Overvad T.F., Palli D., Panico S., Quirós J.R., Rolandsson O., Roswall N., Sacerdote C., Sánchez M.J., Slimani N., Tagliabue G., Tjønneland A., Tumino R., van der A D.L., Forouhi N.G., Sharp S.J., Langenberg C., Riboli E., Wareham N.J. Non-Invasive Risk Scores for Prediction of Type 2 Diabetes (EPIC-InterAct): A Validation of Existing Models. Lancet Diabetes Endocrinol. 2014 Jan; 2(1):19-29.

[4] D'Agostino R.B., Grundy S., Sullivan L.M., Wilson P.; CHD Risk Prediction Group. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA. 2001 Jul; 286(2):180-7.

[5] van der Leeuw J., van Dieren S., Beulens J.W., Boeing H., Spijkerman A.M., van der Graaf Y., van der A D.L., Nöthlings U., Visseren F.L., Rutten G.E., Moons K.G., van der Schouw Y.T., Peelen L.M. The validation of cardiovascular risk scores for patients with type 2 diabetes mellitus. Heart. 2015 Feb; 101(3):222-9.

[6] Janssen K.J., Moons K.G., Kalkman C.J., Grobbee D.E., Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008 Jan; 61(1):76-86.

[7] Moons K.G., Kengne A.P., Grobbee D.E., Royston P., Vergouwe Y., Altman D.G., Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012 May; 98(9):691-8.

[8] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006 Jun; 27(8):861–874.

[9] Harrell F., Califf R., Pryor D., Lee K. and Rosati R. Evaluating the yield of medical tests. JAMA. 1982 May; 247(18):2543–2546.

[10] Data products of the Center for the Study of Aging. [Online]. Available: https://www.rand.org/labor/aging/dataprod/hrs-data.html (accessed on Jan 4th, 2018).

[11] The RAND HRS Data (version P). [Online]. Available: http://hrsonline.isr.umich.edu/modules/meta/rand/randhrsp/rnd_Pdd.pdf (accessed on Jan 4th, 2018).

[12] RAND Enhanced HRS Fat Files. [Online]. Available: https://www.rand.org/labor/aging/dataprod/enhanced-fat.html (accessed on Jan 4th, 2018).

[13] Efron B. and Tibshirani R.J. An introduction to the bootstrap. CRC press, 1994. Management of Data (ACM SIGMOD '97), 265-276.

[14] Stern M. P., Williams K., Haffner S.M. Identification of Persons at High Risk for Type 2 Diabetes Mellitus: Do We Need the Oral Glucose Tolerance Test? Ann Intern Med. 2002 Apr; 136(8):575-581.

[15]Schmidt M.I., Duncan B.B., Bang H., Pankow J.S., Ballantyne C.M., Golden S.H., Folsom A.R., Chambless L.E., Atherosclerosis Risk in Communities Investigators. Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. Diabetes Care. 2005 Aug; 28(8):2013-8.

[16]Wilson P. W., Meigs J.B., Sullivan L., Fox C.S., Nathan D.M., D'Agostino R.B. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007 May; 167(10):1068-74.

[17]Schulze M.B., Hoffmann K., Boeing H., Linseisen J., Rohrmann S., Möhlig M., Pfeiffer A.F., Spranger J., Thamer C., Häring H.U., Fritsche A., Joost H.G. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care. 2007 Mar; 30(3):510-5.

[18]Kahn H.S., Cheng Y.J., Thompson T.J., Imperatore G., Gregg E.W. Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults aged 45 to 64 years. Ann Intern Med. 2009 Jun; 150(11):741-51.

[19]Thomsen S.F., Ulrik C.S., Kyvik K.O.,Larsen K., Skadhauge L.R., Steffensen I., Backer V. The Incidence of Asthma in Young Adults. Chest. 2005 Jun; 127(6):1928-34.

[20]Verlato G., Nguyen G., Marchetti P., Accordini S., Marcon A., Marconcini R., Bono R., Fois A., Pirina P., de Marco R. Smoking and New-Onset Asthma in a Prospective Study on Italian Adults. Int Arch Allergy Immunol. 2016 Aug; 170(3):149-57.