PULSE

Participatory Urban Living for Sustainable Environments

D5.3 Incorporation of New Variables into the Models

PULSE project

H2020 - 727816

University of Padova

May 2018

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

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
Alessandro Zandonà	University of Padova	alessandro.zandona@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 HISTORY

Date	Version	Editor	Change	Status
04/04/2018	1.0	UNIPD	First template	Draft
22/04/2018	1.1	UNIPD	Draft with partial contents	Draft
27/04/2018	2.0	UNIPD	First complete draft	Draft
01/06/2018	2.1	UNIPD	First revised version	Draft
07/06/2018	2.2	UNIPD	Final version	Final

0.3 DOCUMENT KEYDATA

Key words	H2020 – 727816 – PULSE Deliverable 5.3	
Editor info	Name	Barbara Di Camillo
	Organization	UNIPD
	e-mail	barbara.dicamillo@dei.unipd.it

0.4 DISTRIBUTION LIST

Date	Issue	Distribution list
27/04/2018	Circulate first complete draft for internal revision	Maria Fernanda Cabrera, Riccardo Bellazzi, Myrto Valari, Vladimir Urosevic
07/06/2018	Final revised version	Maria Fernanda Cabrera, Riccardo Bellazzi, Myrto Valari, Vladimir Urosevic
	Final version	All Consortium and the European Commission

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

INDEX OF FIGURES

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

INDEX OF TABLES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

EXECUTIVE SUMMARY

The purpose of this deliverable, entitled “Incorporation of new variables into the models”, is to report the results of task 5.3, i.e., the identification of new variables, known or suspected to be associated to the risk of developing type 2 diabetes (T2D) and asthma, which were not considered by state-of-the-art literature models. In addition, in this deliverable we report the activities of task 5.2 related to the implementation and recalibration of state-of-the-art predictive models on the data of the Multi-Ethnic Study of Atherosclerosis, which were not reported in D5.2, due to delays in the availability of the MESA dataset. Therefore, the previous deliverable D5.2 included only the recalibration of state-of-the-art models on the Health and Retirement Study (HRS) dataset. Note that only two T2D predictive models were implemented on the HRS data, because the variables required for the implementation of the other models were not collected in HRS.

This deliverable is structured as follows. After a brief introduction (Section 1), in Section 2 we describe the MESA dataset, which was used for the activities reported in this deliverable. In the following Section 3, Section 4 Section 5, we document the implementation and recalibration of state-of-the-art predictive models of diabetes onset and asthma adult-onset. Section 6 is dedicated to the identification of new potentially predictive variable of diabetes and asthma onset. The probabilistic relationships between the candidate predictive variables and the onset of diabetes and asthma were studied by using static Bayesian networks, as we describe in Section 7. Finally, in Section 8 we discuss the integration of the predictive models in the PULSE architecture.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

1 INTRODUCTION

In order to test the existing predictive models of T2D and asthma onset and develop new models, longitudinal datasets in which a group of healthy subjects is followed up over time for several years (e.g. more than 5 years) are required, because of the need to observe a sufficient number of new cases of T2D/asthma. Since PULSE cannot collect longitudinal data with such a long time horizon (due to the limited duration of this project), we looked for longitudinal datasets already available in the literature that could be suitable for the implementation and development of predictive models of T2D and asthma. Two datasets were selected for this purpose: the Health and Retirement Study (HRS) dataset and the Multi-Ethnic Study of Atherosclerosis (MESA) dataset. Since the HRS dataset includes information on diabetes onset, but not on asthma adult-onset, it was possible to implement only T2D onset predictive models on this dataset. In particular, the implementation and recalibration of state-of-the-art models on HRS dataset was reported in deliverable D5.2. Conversely, both T2D and asthma onset predictive models can be implemented on the MESA dataset, as information on both T2D and asthma onset was collected in the MESA study. The MESA dataset also contains several variables potentially predictive of T2D/asthma onset, not considered by the state-of-the-art models, such as clinical variables, psychological factors, economic status indicators and neighbourhood characteristics.

In this deliverable, we will focus on two main tasks: i) the implementation and recalibration of state-of-the-art models of T2D and asthma onset on the MESA dataset; ii) the identification of new variables potentially predictive of T2D and asthma onset, which were not used in the state-of-the-art models, and the study of the probabilistic relationships between these variables and the onset of T2D/asthma by static Bayesian networks.

The first task about implementation and recalibration of the state-of-the-art models has a two objectives i.e., to compare the performance of the state-of-the-art models on the same population and to assess the generalizability of the state-of-the-art models, i.e. how the models perform when they are applied to a different population from that in which the models were developed. Assessing the generalizability of the models is important to determine if the models can be applied to new populations, and thus to the population surveyed in the PULSE pilots.

The main objective of the second task is the identification of new variables related to the onset of T2D/asthma that may be incorporated in new predictive models to improve their performance. The incorporation of new variables in the models will be performed in alignment with WP2 to guarantee that the new variables are collected in the PULSE pilots and the models can be implemented in the PULSE architecture.

2 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%) [1]. Participants enrolled were both males and females aged 45-84 years, free of cardiovascular diseases. 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.

Besides the main study, MESA investigators conducted also an ancillary study on respiratory diseases, called MESA Lung, in a subset of MESA original cohort. MESA Lung participants underwent a baseline examination either at exam 3 or 4 and then a follow-up examination at exam 5. The MESA Lung examinations included a spirometry test and a questionnaire on respiratory health.

The number of subjects surveyed at each exam of MESA and MESA Lung is reported in Table 1.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 1. Number of subjects surveyed at each MESA and MESA Lung exam.

Study	Exam 1 (Jul 2000 – Aug 2002)	Exam 2 (Sep 2002 – Feb 2004)	Exam 3 (Mar 2004 – Sep 2005)	Exam 4 (Sep 2005 – May 2007)	Exam 5 (Apr 2010 – Dec 2011)
MESA	6814	6233	5947	5818	4716
MESA Lung	-	-	1381	2574	3228

2.1 DATA FOR DIABETES MODEL DEVELOPMENT

Diabetes was assessed at each MESA exam. In particular, treated diabetes was defined as use of insulin of oral hypoglycemic medications, while untreated diabetes was defined as fasting glucose concentration ≥126 mg/dL, according to the 2003 American Diabetes Association fasting criteria. For the implementation and development of diabetes prediction models, we selected the subsample of subjects who satisfied the following three conditions:

- no diabetes (either treated or untreated) at exam 1
- no history of cancer at exam 1
- information on diabetes available at least at one of the follow-up exams

The selected subsample included 5155 subjects of whom 640 subjects developed diabetes during the study. In particular, 184 subjects developed diabetes at exam 2, 106 at exam 3, 147 at exam 4 and 203 at exam 5. Note that we excluded the subjects having a history of cancer because cancer may have significantly compromised the health of these subjects.

The selected data were split into a training and a test set, including the 80% and 20% of selected subjects, respectively, stratified for incidence of diabetes. In particular, the training set contains 4124 subjects of whom 512 subjects develop diabetes during the follow-up period, while the test set includes the remaining 1031 subjects of whom 128 develop diabetes during the follow-up period.

Dataset for implementation of logistic regression models

The great majority of state-of-the-art diabetes models described in the literature and, thus, in the present deliverable is based on logistic regression. As logistic regression is a static model, i.e., it does not explicitly account for the possible time-variability of the outcome, complete dynamic information cannot be directly used. This is in contrast to survival analysis techniques, which can naturally incorporate information on survival times. To overcome this limitation, a common strategy entails reframing the outcome of the models in terms of a simple yes/no question, such as “Will the subject develop diabetes in a fixed amount of years following the first observation?”, where the temporal meaning is implied but not explicitly modelled. Clearly, the most important parameter in this setting is the choice of a cut-off time: cut-offs too close to the baseline observation are unadvisable because diabetes is a slow-developing illness, while extremely large prediction horizons require observing the training population for longer periods of time, which is often impossible. A cut-off time of 8 years, roughly corresponding to 4 visits, was selected, in line with the typical follow-up times reported in the majority of literature models. Labels were then assigned according to the diagram in Figure 1. Diabetes onset was considered to have happened (value 1) if it had been observed on or before the cut-off, a subject was deemed to be healthy (value 0) at the cut-off if he or she left the study at or after the cut-off and had not developed diabetes before leaving the study. The graph also shows that “NaN” values were assigned to subjects for whom there was no certain status at the cut-off: this happened for those who left the study without diabetes before the cut-off and for those who developed diabetes after the cut-off. Indeed, the former could have fallen ill at an unknown date, after leaving the study and there is insufficient information to say whether the latter had already developed diabetes at the cut-off. The subjects whose status was “NaN” were excluded from the logistic regression analyses related to diabetes. Consequently, the maximum available sample size was 3736 (2997 training and 739 test cases).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

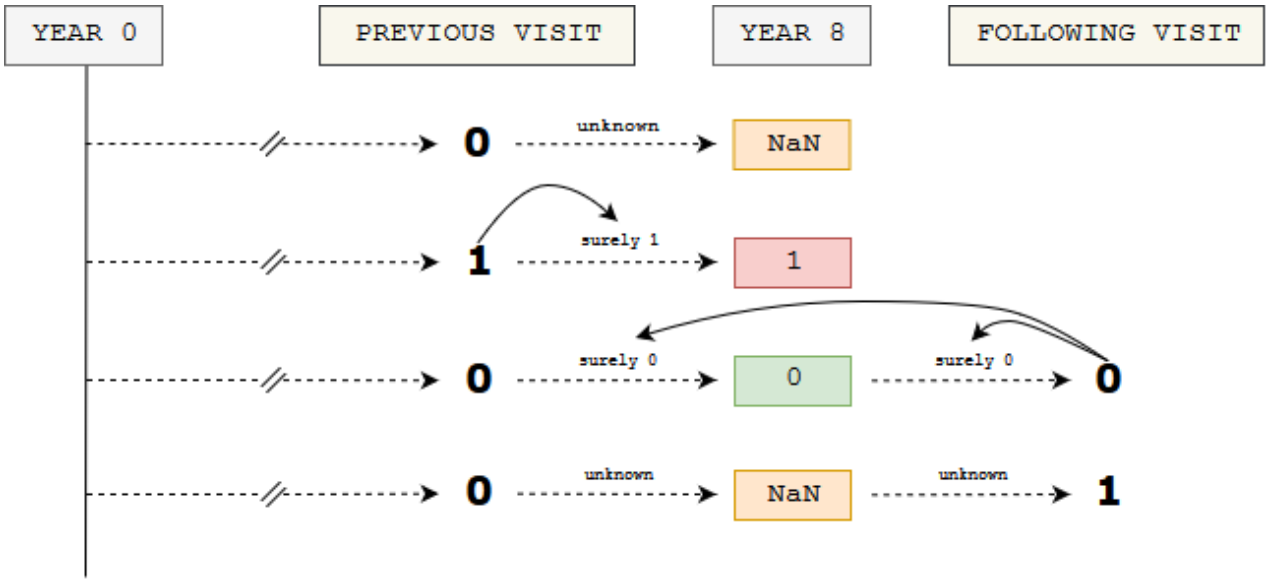

Figure 1. Label assignment strategy. The label value 1 was assigned to cases of diabetes onset by the cut-off date and 0 to subjects who did not report having diabetes after the cut-off date. “NaN” in the diagram means the subjects were excluded from the analysis. For example, the last line is to be interpreted as: "An outcome of 1 is reported in the data, but only after year 8; hence, the subject's health status at year 8 is unknown (NaN).

2.2 DATA FOR ASTHMA MODEL DEVELOPMENT

History of asthma was assessed at the baseline MESA exam by the question “Has a doctor ever told you that you had asthma?”. At the follow-up visits, i.e., exams 2-5, new development of asthma was assessed by the question “Has a doctor ever told you that you have developed asthma since your last MESA visit?”. For the implementation and development of asthma prediction models, we selected the subsample of subjects who satisfied the following three conditions:

- no history of asthma at exam 1
- no history of cancer at exam 1
- information on asthma development available at least at one of the follow-up exams

The selected subsample included 5341 subjects of whom 136 subjects developed asthma during the study. In particular, 29 subjects developed asthma at exam 2, 33 at exam 3, 32 at exam 4 and 42 at exam 5. As for the diabetes data selection, we decided to exclude the subjects having a history of cancer because cancer may have significantly compromise the health of these subjects.

The selected data were split into a training and a test set, including the 80% and 20% of selected subjects, respectively, stratified for incidence of asthma. In particular, the training set contains 4273 subjects of whom 109 subjects develop asthma during the follow-up period, while the test set includes the remaining 1068 subjects of whom 27 develop asthma during the follow-up period.

Dataset for implementation of logistic regression models

As the state-of-the-art models of asthma onset are based on logistic regression, which is a static model used to predict a binary outcome at a fixed amount of time from baseline, we could not fit the logistic regression models on the entire dataset but we had to select a subset of subjects having the outcome defined at a certain cut-off time. A different cut-off time for each state-of-the-art model is chosen in order to maximize the number of subjects with new asthma onset at time lower of equal to the cut-off time. These subjects are assigned label 1. Then, as for the diabetes models, we assumed a subject was healthy at the cut-off (label 0) if he or she left the study at or after the cut-off and had not developed asthma before leaving the study. Subjects who left the study without asthma before the cut-off and subjects who developed asthma after the cut-off, but do not have information on asthma at the cut-off, were excluded because the outcome value at the cut-off time is not known. Since the incidence of adult-

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

onset asthma on the MESA population is low, we decided to consider a different cut-off time for each state-of-the-art model in order to maximize the number of subjects with incident asthma in the selected sample (see Section 5.1.2 and Section 5.2.2).

3 IMPLEMENTATION, RECALIBRATION AND ASSESSMENT OF STATE-OF-THE-ART MODELS

State-of-the-art predictive models of diabetes and asthma onset were implemented and recalibrated on the selected data by performing the following 5 steps:

A. Data selection and pre-processing: suitable subsamples of training and test sets are extracted by selecting the subjects without missing values on the model independent variables at baseline and with sufficient follow-up duration.
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. Model recalibration: the model parameters were estimated on the training set and on 100 sets extracted from the training set by bootstrap resampling, in order to assess the effect of different training/test splits on the model performance.
D. Performance assessment: the performance of the model recalibrated on the entire training set was assessed on the test set, while the performance of the models recalibrated on each of the 100 sets extracted by bootstrap resampling were assessed on the 100 sets of out of bag samples not used for the bootstrap training. This second assessment is in the following called validation phase.
E. Comparison with the original model: the performance of the recalibrated model was compared to the performance of the original state-of-the-art model both in the test set and the validation phase.

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 or asthma onset, and their calibration, i.e., the extent of agreement between observed incidence of diabetes or asthma and that predicted by the model. Two metrics were considered for discriminatory ability: 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 in which higher scores are attributed to subjects at risk for a certain clinical outcome (in this case, diabetes or asthma), 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. See Figure 2 for an example of interpretation of the ROC curves. 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 or asthma onset) is ranked higher than a subject chosen at random from the negative outcome group [2].

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

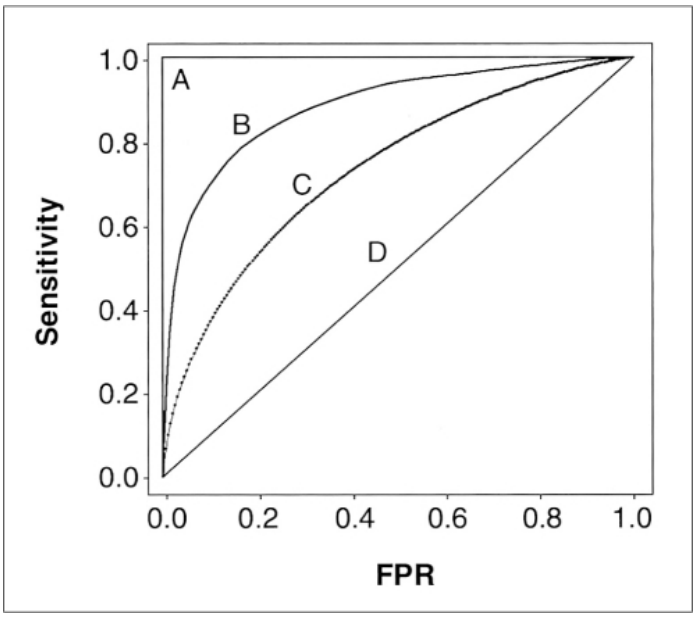

Figure 2. Representative ROC curves for models with different discriminatory ability. D represents the random predictor (AU-ROC = 0.5), A represents the perfect model (AU-ROC = 1), B and C represents profiles typically observed for reasonably good predictive models, where B performs better than C (greater AU-ROC for B than C). Figure taken from [3].

The C-index, proposed by Harrell et al. [4], is an extension of AU-ROC to be used when information on model outcome is available over time. In the case of MESA data, information on the outcome, i.e. diabetes or asthma onset, was collected at each exam. 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 or asthma, 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.

Calibration was assessed by the expected-to-observed event ratio (E/O), i.e., the ratio between the expected number of events at a certain time *t*, obtained as the sum of the probabilities of having diabetes or asthma at time *t* predicted by the model, and the number of observed events at time *t* [5]. Values of E/O close to 1 indicate that the model has good calibration, whereas values significantly higher/lower than 1 indicate that the model tends to over/underestimate the event probability.

In addition, model calibration was also graphically assessed by visualizing the calibration plot at a certain time *t*. The calibration plot represents the number of observed events vs. the number of expected events for increasing deciles of predicted event probability. The more the calibration plot is close to the line with 0 intercept and 45° slope, the better the model calibration. See Figure 3 for an example of interpretation of the calibration plot.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

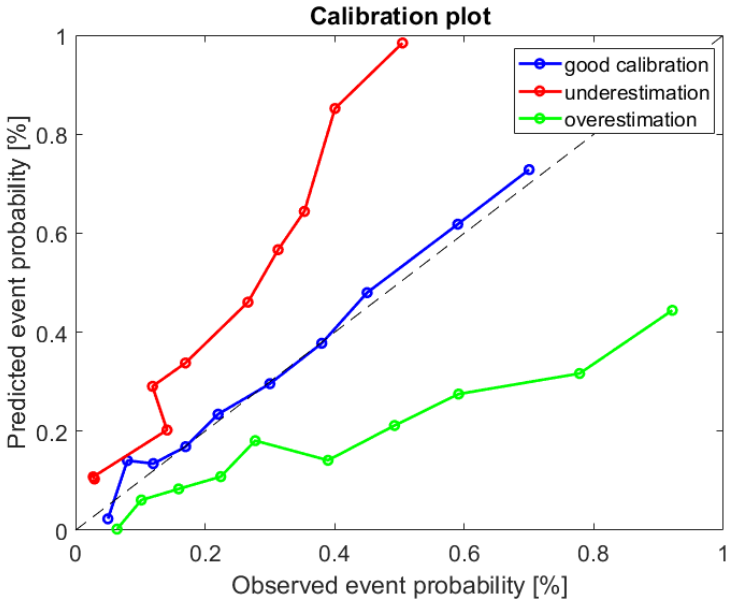

Figure 3. Representative calibration plot for a model with good calibration (blue line), a model that underestimates actual event probability (red line) and a model that overestimates the actual event probability (green line).

4 IMPLEMENTATION, RECALIBRATION AND ASSESSMENT OF STATE-OF-THE-ART MODELS OF DIABETES ONSET

Eight state-of-the-art predictive models of diabetes were implemented, recalibrated and assessed on the MESA dataset, i.e., the model by Stern et al., the FINDRISC, the ARIC models (3 models), the Framingham model, the basic risk score by Kahn et al. and the DPoRT. It was not possible to implement the German diabetes risk score because the data on diet collected in MESA were not accessible (a special permission is required).

A new model was also recently developed by Di Camillo et al. [7], i.e. the HAPT2D, which was not published at the time of D5.1 and D5.2 writing. The HAPT2D is based on a Cox proportional hazard model combining basic and advanced clinical variables, as well as information on lifestyle habits and socio-economic indicators. However, the HAPT2D model cannot be implemented on the MESA data because some of the model variables cannot be defined by the information available in MESA. In particular, the “country” variable is not available in MESA and the categories of the “professional status” variable (i.e., clerical, manual worker, student, unknown/unemployed/housewife/retired) are not defined in MESA.

4.1 Clinical model by Stern et al.

4.1.1 The original model

In their original work [8], Stern et al. proposed a clinical model for the identification of subjects at high risk of developing diabetes based on a set of readily available variables and clinical variables. Specifically, they used information on age, gender, ethnicity, BMI, family history of diabetes, together with the data from a routine check-up visit (fasting glucose, HDL cholesterol, and diastolic blood pressure), to predict the 7.5-year incidence of type 2 diabetes. They estimated and reported the logistic regression coefficients and the formula they used to calculate the probability of developing diabetes over the 7.5-year follow-up period. Of note, the authors explicitly advise caution in extending the model

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

to different populations or applying it in a clinical setting, as their original cohort comprised only Mexican and white Americans, the former of whom were also overrepresented (1791 vs. 1112) in the study sample.

4.1.2 Data selection and preprocessing

The initial dataset presented in Section 2.1 was further reduced to accommodate the specific characteristics of the clinical model by Stern et al., as follows.

- The authors explicitly state that one of the predictive variables to be used in their model is the distinction between the Mexican American and white non-Hispanic American ethnicities. As such, black and Asian subjects from the MESA dataset were excluded prior to the analyses;
- Subjects for whom one or more model variables were not recorded (i.e., had one or more missing values) were also discarded.

The remaining sample comprised 2280 subjects, divided between a training and a test sets of 1839 and 441 subjects, respectively. Of the 1839 members of the training set, 186 developed diabetes within 8 years vs. 51 in the test set, thus preserving a similar cases to controls ratio.

4.1.3 Model implementation and recalibration

Variable preprocessing

The variables required for the implementation of the clinical model by Stern et al. were treated as in the appendix of the original work [8], with only the following minor deviations.

- The “Mexican American ethnicity” variable was set to 1 when the subject was recorded in the MESA dataset as being of “Hispanic ethnicity”.
- “Family history of diabetes” in the original study only referred to parents and siblings; here, it was extended to also include children.

Recalibration on the training set

The training set was used in the recalibration phase to fit a logistic regression model where the dependent variable was the onset of type 2 diabetes in an 8-year follow-up window and the independent variables were age, gender, Hispanic or white ethnicity, BMI, family history of diabetes, fasting glucose, HDL cholesterol, and diastolic blood pressure. See Table 2 of Section 4.1.4 for further details on variables distribution across the training and test sets.

4.1.4 Results

The data were divided between a training and a test sets, comprising 1839 and 441 subjects, whose characteristics are reported in Table 2. As shown in the table, the predictive variables used in the model present a similar distribution between the two sets and, in particular, the outcome is observed in a very similar percentage of subjects (10.1% vs. 11.6%).

Table 2. Distribution of Stern et al.’s model variables in the training and test sets reported as percentage of subjects in different variable categories for 1/0 variables and as their mean for continuous variables.

Variable	Category	% Subjects or mean in the training set (N=1839)	% Subjects or mean in the test set (N=441)
Age [years]	-	59.7	59.8
Female Gender [Boolean]	Yes	51.5%	57.1%
Hispanic ethnicity [Boolean]	Yes	33.2%	21.4%

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Category	% Subjects or mean in the training set (N=1839)	% Subjects or mean in the test set (N=441)
Fasting glucose [mg/dL]	-	88.5	88.2
Systolic blood pressure [mmHg]	-	121.7	121.2
HDL cholesterol [mg/dL]	-	51.8	50.8
BMI [kg/m^2]	-	28.0	28.3
Family history of diabetes [Boolean]	Yes	34.1%	30.6%
8-year incidence of type 2 diabetes [Boolean]	Yes	10.1%	11.6%

The logistic regression coefficients of the original and recalibrated models are summarized in Table 3. The signs and orders of magnitude for all coefficients are consistent between the models. The most substantial variations are in the importance given to the “age” and “systolic blood pressure” variables, both much closer to zero after recalibration. On the contrary, not surprisingly, the coefficients were re-balanced so that the importance of fasting glucose in the final probability evaluation was greatly increased, from 0.079 to 0.133.

Table 3. Coefficients of the clinical model by Stern et al. (fourth column) compared with those of the recalibrated version on the MESA dataset (third column).

Variable	Value	Recalibrated model coefficient	Original model coefficient
Intercept	-	-17.483	-13.415
Age [years]	-	0.003	0.028
Female Gender [Boolean]	Yes	0.252	0.661
Hispanic ethnicity [Boolean]	Yes	0.129	0.412
Fasting glucose [mg/dL]	-	0.133	0.079
Systolic blood pressure [mmHg]	-	0.001	0.018
HDL cholesterol [mg/dL]	-	-0.014	-0.039
BMI [kg/m^2]	-	0.091	0.070
Family history of diabetes [Boolean]	Yes	0.534	0.481

A summary of the performances in terms of discrimination ability of the original and recalibrated models is presented in Table 4. As it can be seen, the results are quite satisfactory, before and after recalibration: regardless of the way they were assessed (in the validation phase or on the test set, on the original or on the recalibrated models) the AU-ROC was 0.87 and the C-index 0.86, denoting a very high discrimination power. A visual inspection of the ROCs represented in Figure 4 confirms this assessment even though it could be argued that recalibrating the model slightly favoured a steeper initial increase of the curve.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 4. Performance of Stern et al.'s clinical model: AU-ROC and C-index for the recalibrated and original models assessed during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	Metric	Bootstrap validation	Test set
Recalibrated logistic regression	AU-ROC at 8 years	0.871 (± 0.022)	0.873
	C-index	0.858 (± 0.021)	0.857
Original logistic regression	AU-ROC at 8 years	0.872 (± 0.020)	0.871
	C-index	0.858 (± 0.019)	0.856

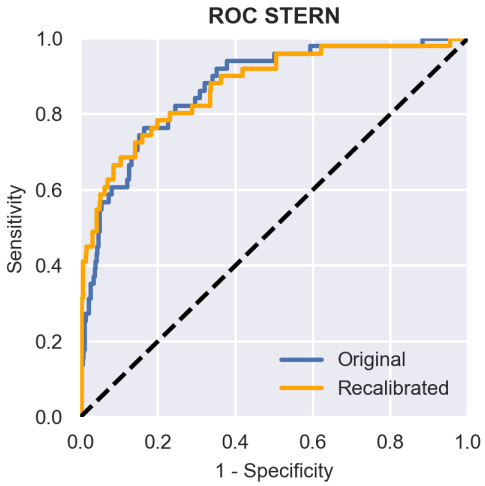

Figure 4. ROC curve for the original (blue) and recalibrated (orange) versions of Stern et al.'s clinical model. The dashed line indicates random chance.

Table 5 and Figure 5– which serves as its graphical counterpart— show that, although discrimination performance was very similar, the recalibrated model greatly outperformed the original one in terms of calibration. Indeed, the latter had a distinct tendency to overestimate the actual probability of developing diabetes (E/O ratio of 1.77, greatly exceeding 1), while the former was very well-calibrated.

Table 5. Calibration of Stern et al.'s clinical model: E/O ratios and their 95% confidence intervals for the recalibrated and original models assessed on the test set.

Model	Metric	Test set
Recalibrated logistic regression	E/O ratio [95% CI]	0.92 [0.70 – 1.21]
Original logistic regression	E/O ratio [95% CI]	1.77 [1.34 – 2.32]

PULSE Participatory Urban Living for Sustainable Environments
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

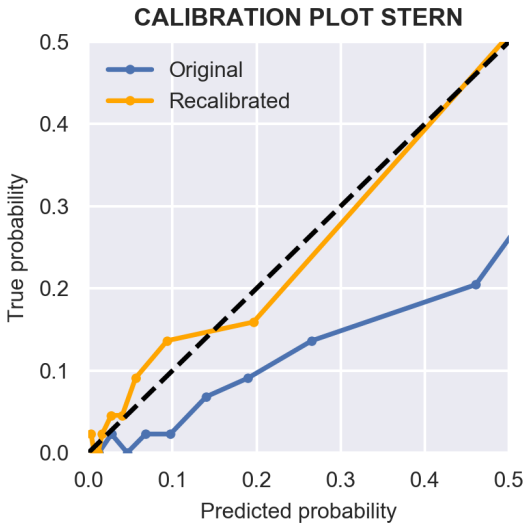

Figure 5. Calibration plot for Stern et al.'s clinical model in its original (blue) and recalibrated (orange) versions. The dashed line represents perfect calibration.

4.2 FINDRISC

4.2.1 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 [9] 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 (10 years) 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 [9] 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 [9] in 4615 not drug-treated subjects that responded to a baseline survey in 1992 and were observed over a follow-up of 5 years (vs. 10 in the development cohort) for incidence of drug-treated diabetes.

4.2.2 Data selection and preprocessing

The initial dataset presented in Section 2.1 was further reduced to accommodate the specific characteristics of the FINDRISC concise model. The full version of the model was impossible to implement on the MESA dataset because of the irrecoverable mismatch between the way physical activity and diet were recorded therein, and the format needed to incorporate them into the full FINDRISC model. Additionally

- subjects for whom one or more model variables were not recorded were discarded.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

The remaining sample comprised 3641 subjects, divided between a training and a test sets of 2990 and 651 subjects, respectively. Of the 2990 members of the training set, 347 developed diabetes within 8 years vs. 87 in the test set, thus preserving a similar cases to controls ratio.

4.2.3 Model implementation and recalibration

Variable preprocessing

The variables required for the implementation of the concise FINDRISC model were treated as in the original work [9], with only the following minor deviation.

- The age category of people aged between 55 and 65 was extended to also include anyone older than 65.

Recalibration on the training set

The training set was used in the recalibration phase to fit a logistic regression model where the dependent variable was the onset of type 2 diabetes in an 8-year follow-up window and the independent variables were age, gender, BMI, waist circumference, use of anti-hypertensive medications, and family history of diabetes. See Table 7 of Section 4.2.4 for further details on variables distribution across the training and test sets. Based on the logistic regression β-coefficients, partial scores were assigned to each variable, applying the same criteria adopted by Lindström and Tuomilehto [9]. These coefficient-to-score conversion rules are reported in Table 6. The recalibrated concise FINDRISC score was calculated by adding together all the partial scores.

Table 6. 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

4.2.4 Results

The characteristics of the training and test subpopulations are reported in Table 7. As shown in the table, the predictive variables used in all the models present a similar distribution between the two sets and, in particular, the outcome is observed in a very similar percentage of subjects (11.6% vs. 11.8%). Of note, the two age categories span the entire dataset: no subjects considered in the analysis are younger than 45. This unintended effect is due to the unavoidable exclusions of subjects detailed in Section 4.2.2.

Table 7. Distribution of FINDRISC variables in the training and test sets reported as percentage of subjects in different variable categories.

Variable	Category	% Subjects training set (N=2990)	% Subjects test set (N=651)
Age [years]	45-54	35.5%	37.3%
	≥55	64.5%	62.7%
BMI [kg/m^2]	25 to <30	39.6%	38.6%

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Category	% Subjects training set (N=2990)	% Subjects test set (N=651)
	≥30	29.1%	31.0%
Waist circumference [cm]	Men: 94 to <102 Women: 80 to <88	25.0%	23.3%
	Men: ≥102 Women: ≥88	50.4%	51.2%
Use of blood pressure medication [Boolean]	Yes	31.0%	26.3%
History of high blood glucose [Boolean]	Yes	1.1%	1.4%
8-year incidence of drug-treated diabetes [Boolean]	Yes	11.6%	11.8%

Table 8 presents a detailed overview of the scores and coefficients that were obtained after recalibration and a comparison between them and those in the published model by Lindström and Tuomilehto. As it is apparent, refitting the model on the MESA dataset had a noticeable impact on many coefficients and partial scores. Indeed, with the exception of "use of blood pressure medication," "history of high blood glucose," and "BMI ≥30", all the other scores were modified: "BMI 25 to <30" changed from the original 1 to a 2-points risk factor, whereas both waist circumference categories were rescaled to the same partial score of 2 vs. 3 and 4 in the original model. The logistic regression coefficients were similarly affected, although less noticeably: their signs and orders of magnitude were consistent with the literature version of the FINDRISC. The difference in coefficients (0 vs. 0.628, 0.064 vs. 0.892) and scores (0 vs. 2, 1 vs. 3) related to the two age categories deserves a special comment. Recall that in the first paragraph of Section 3.2.4 the unavailability of subjects younger than 45 was highlighted as a peculiarity of the reduced training set used for this analysis. As a direct consequence of this, the coefficient associated with one of the two age categories and the intercept are highly collinear. Thus, the problem of estimating their values is ill posed, i.e. it is only possible to give a stable estimate for their sum total. To address this issue, the "Age 45-54" category was considered as the baseline and its coefficient and score were fixed to 0. In this way, it was possible to estimate reasonable values for the related parameters, namely "Age 55-64" and the intercept.

Table 8. 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). The asterisk (*) denotes values that were fixed because of dataset characteristics.

Variable	Value	Recalibrated coefficient	Recalibrated points	Original coefficient	Original points
Intercept	-	-3.440	-	-5.514	-
Age [years]	45-54	0(*)	0(*)	0.628	2
	55-64	0.064	1	0.892	3
BMI [kg/m^2]	25 to <30	0.354	2	0.165	1
	≥30	0.949	3	1.096	3
Waist circumference [cm]	Men: 94 to <102 Women: 80 to <88	0.663	2	0.857	3
	Men: ≥102	0.731	2	1.350	4

PULSE Participatory Urban Living for Sustainable Environments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Value	Recalibrated coefficient	Recalibrated points	Original coefficient	Original points
	Women: ≥88				
Use of blood pressure medication [Boolean]	Yes	0.574	2	0.711	2
History of high blood glucose [Boolean]	Yes	2.716	5	2.139	5

A summary of the performances in terms of discrimination ability of the original and recalibrated models is presented in Table 9. At a close inspection, the table reveals a dichotomous pattern in the effects of recalibration: the logistic regression model, used to estimate a probability value for T2D, slightly benefitted from recalibration, whereas the recalibrated score performance was diminished. A likely explanation of this behaviour lies in the lack of flexibility of the coefficient-to-score conversion rules presented in [9]. Indeed, the original authors only provide hard thresholds (see Table 6) for the construction of the final score, but no hint as to their rationale for coming up with them. As these thresholds are most likely data-dependent, they may be negatively affected by a change in datasets. Figure 6 helps in visualising the extent of the performance deterioration introduced when passing from the continuous-valued logistic regression prediction to the discrete score. It could be argued, then, that if one were interested in defining a score of comparable performance to the logistic regression model, the first logical step would be building a new coefficient-to-score table similar to Table 6. This procedure, however, deviates from the scope of the present work, whose main goal is recalibrating and validating literature models as they are, introducing as little modifications as possible, to achieve a fair comparison.

Table 9. 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 re-samplings) and on the test set.

Model	Metric	Bootstrap validation	Test set
Recalibrated score	AU-ROC at 8 years	0.679 (± 0.020)	0.715
	C-index	0.670 (± 0.019)	0.706
Recalibrated logistic regression	AU-ROC at 8 years	0.683 (± 0.019)	0.738
	C-index	0.674 (± 0.018)	0.729
Original score	AU-ROC at 8 years	0.690 (± 0.018)	0.732
	C-index	0.680 (± 0.017)	0.722
Original logistic regression	AU-ROC at 8 years	0.692 (± 0.018)	0.735
	C-index	0.682 (± 0.017)	0.726

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

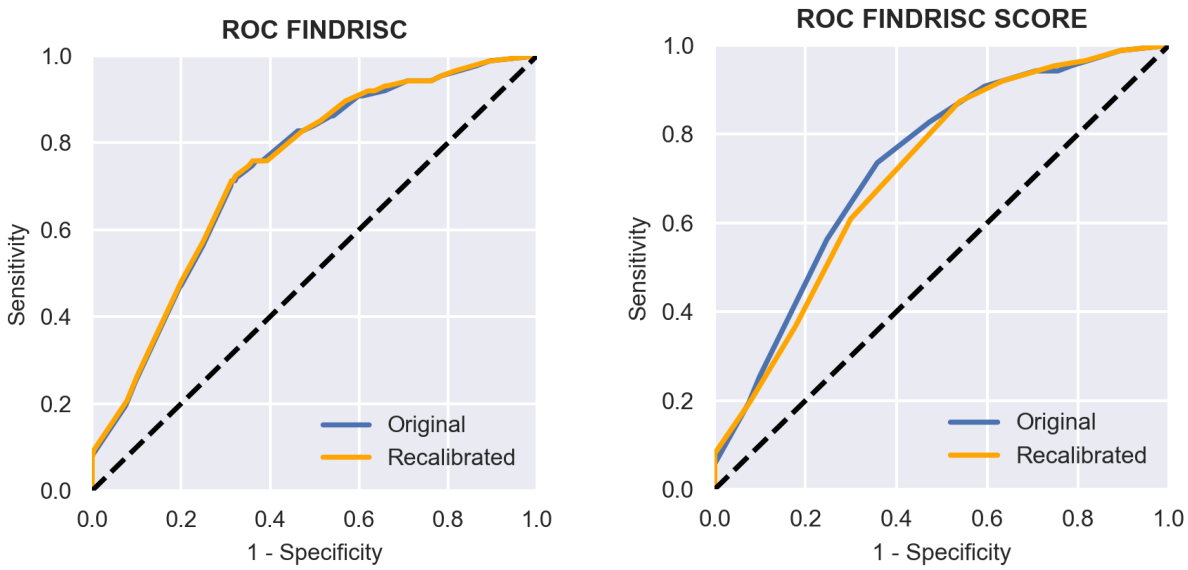

Figure 6.ROC curves for the original (blue) and recalibrated (orange) versions of the concise FINDRISC model (logistic regression on the left, scores on the right). The dashed line indicates random chance.

Table 10 and Figure 7 –which serves as its graphical counterpart— show that, although discrimination performance was very similar, the recalibrated model greatly outperformed the original one in terms of calibration. Indeed, the latter had a distinct tendency to underestimate the actual probability of developing diabetes (E/O ratio of just 0.52), while the former was very well-calibrated. These results are only applicable to the logistic regression version of the FINDRISC model, because calibration curves and E/O ratios are only meaningful if the output is a probability and, as such, are impossible to compute on a discrete score.

Table 10. Calibration of the concise FINDRISC model: E/O ratios and their 95% confidence intervals for the recalibrated and original models assessed on the test set.

Model	Metric	Test set
Recalibrated logistic regression	E/O ratio [95% CI]	0.98 [0.79 – 1.21]
Original logistic regression	E/O ratio [95% CI]	0.52 [0.42 – 0.64]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

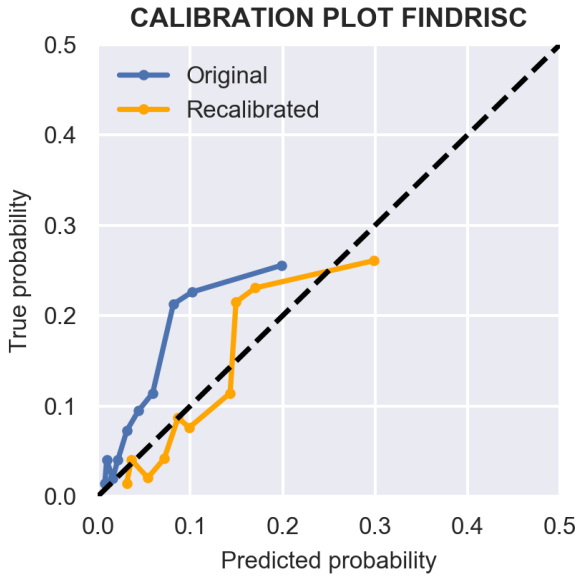

Figure 7. Calibration plot for the concise FINDRISC in its original (blue) and recalibrated (orange) versions. The dashed line represents perfect calibration.

4.3 ARIC models

4.3.1 The original model

The ARIC models are a set of 3 models proposed by Schmidt et al. in 2005 [6] to predict the risk for diabetes. Each of the models builds upon its previous iteration by increasing the total amount of considered variables: the base model includes age, ethnicity, family history of diabetes, systolic blood pressure, waist circumference, and height; the second model adds fasting glucose to the equation; and, finally, the third, most comprehensive model, also incorporates information on HDL cholesterol and triglycerides concentrations. All of these were derived by applying logistic regression to the appropriate set of variables to predict the risk of developing diabetes in middle-aged, white, and African-American adults over the course of 9 years. The original dataset was collected within the Atherosclerosis Risk in Communities (ARIC) study and comprised 7915 participants, 1292 of whom developed diabetes.

4.3.2 Data selection and preprocessing

The initial dataset presented in Section 2.1 was further reduced to accommodate the specific characteristics of the ARIC models.

- The authors explicitly state that one of the predictive variables to be used in their model is the distinction between the African-American and white American ethnicities. As such, Hispanic and Asian subjects from the MESA dataset were excluded prior to the analyses.
- Subjects for whom one or more model variables were not recorded were also discarded.

The remaining sample comprised 2401 subjects, divided between a training and a test sets of 1934 and 467 subjects, respectively. Of the 1934 members of the training set, 196 developed diabetes within 8 years vs. 48 in the test set, thus preserving a similar cases to controls ratio. All the three models were validated and recalibrated on the same subset of MESA data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

4.3.3 Model implementation and recalibration

Variable preprocessing

The variables required for the implementation of the three ARIC models were treated as in the description of the first table in the original work [10], with only the following minor deviation.

- "Family history of diabetes" in the original study only referred to parents; here, it was extended to also include siblings and children.

Recalibration on the training set

The training set was used in the recalibration phase to fit three logistic regression models, where the dependent variable was the onset of type 2 diabetes in an 8-year follow-up window and the independent variables were:

- age, gender, black or white ethnicity, family history of diabetes, systolic blood pressure, waist circumference, height for the base model;
- all the variables in the base model plus fasting glucose for the second model;
- all the variables in the base model plus fasting glucose, HDL cholesterol concentration and triglycerides concentration for the third model.

See Table 11 of Section 4.3.4 for further details on variables distribution across the training and test sets.

4.3.4 Results

The characteristics of the training and test subpopulations are reported in Table 11. As shown in the Table, the predictive variables used in all the models present a similar distribution between the two sets and, in particular, the outcome is observed in a very similar percentage of subjects (10.1% vs. 10.3%).

Table 11. Distribution of ARIC variables in the training and test sets reported as percentage of subjects in different variable categories for 1/0 variables and as their mean for continuous variables.

Variable	Category	% Subjects or mean in the training set (N=1934)	% Subjects or mean in the test set (N=467)
Age [years]	-	60.0	59.8
Black ethnicity [Boolean]	Yes	36.6%	38.8%
Family history of diabetes [Boolean]	Yes	34.8%	31.7%
Systolic blood pressure [mmHg]	-	123.9	123.3
Waist circumference [cm]	-	97.7	97.7
Height [cm]	-	169.3	168.7
Fasting glucose [mg/dL]	-	88.2	87.5
HDL cholesterol [mg/dL]	-	53.6	52.2
Triglycerides [mg/dL]	-	117.6	118.5
8-year incidence of type 2 diabetes [Boolean]	Yes	10.1%	10.3%

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

The logistic regression coefficients of the original and recalibrated models are summarized in Table 12. The models are presented from left to right in order of increasing complexity. Across all versions of the model, coefficients were, for the most part, adjusted in a consistent way as a result of recalibration. For instance, in all recalibrated models the recalibrated waist circumference coefficients are very close to their respective original values (0.0476 vs. 0.412, 0.0276 vs. 0.0328, 0.0257 vs. 0.0273) and all "Family history of diabetes" coefficients are approximately doubled after recalibration (0.9602 vs. 0.5463, 0.8660 vs. 0.5088, 0.8523 vs. 0.4981). Compared to the original model, the coefficients related to subjects' ages and heights had the opposite sign after recalibration 2/3 and 1/3 times, respectively. This suggests that in the MESA dataset, the influences of age and height on T2D risk are not as clear as in the original study cohort. In the second and third model, the coefficients related to fasting glucose were greatly increased after recalibration. This effect suggests the great predictive ability of fasting glucose in the MESA dataset, already seen for Stern's model in Section 4.1. On the contrary, the other biomarkers added in the third ARIC model saw their coefficients shrink to half their original value, possibly signaling their diminished importance relative to fasting glucose.

Table 12. Coefficients of the three ARIC models in their original and recalibrated versions.

Variable	Base model		Base + glucose model		Base + glucose + lipids model	
	Recal.	Original	Recal.	Original	Recal.	Original
Intercept	-10.3354	-7.3359	-15.7518	-12.2555	-14.5921	-9.9808
Age [years]	0.0123	0.0271	-0.0176	0.0168	-0.0155	0.0173
Black ethnicity [Boolean]	0.3715	0.2295	0.2210	0.2361	0.2900	0.4433
Family history of diabetes [Boolean]	0.9602	0.5463	0.8660	0.5088	0.8523	0.4981
Systolic blood pressure [mmHg]	0.0128	0.0161	0.0095	0.0120	0.0092	0.0111
Waist circumference [cm]	0.0476	0.0412	0.0276	0.0328	0.0257	0.0273
Height [cm]	0.0020	-0.0115	-0.0124	-0.0261	-0.0162	-0.0326
Fasting glucose [mg/dL]	-	-	0.1305	0.0913	0.1284	0.0880
HDL cholesterol [mg/dL]	-	-	-	-	-0.0079	-0.0122
Triglycerides [mg/dL]	-	-	-	-	0.0012	0.00271

A summary of the performances in terms of discrimination ability of the original and recalibrated models is presented in Table 13. In all cases, recalibration improved the performance on the test set. However, its advantages were clearly overshadowed by the inclusion of fasting glucose as a variable, as evidenced by the substantial improvement in AU-ROC and C-index between the base model and its more comprehensive variants. Indeed, the original model C-index leapt from a base value of 0.763 to a satisfactory 0.837 when fasting glucose was considered. The same phenomenon can be observed on the recalibrated models: the baseline 0.794 C-index increased to 0.853 in the second ARIC model. The benefit of also including lipids information was not clearly observable in either the original or recalibrated version of the third model, compared to the second one. By examining, it is apparent that recalibration had a greater relative effect on the base model and was only marginally useful for the second and third versions of the ARIC model.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 13. Performance of the ARIC models: AU-ROC and C-index for the recalibrated and original models assessed during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	Metric	Bootstrap validation	Test set
Recalibrated base model	AU-ROC at 8 years	0.750 (± 0.024)	0.812
	C-index	0.737 (± 0.023)	0.794
Original base model	AU-ROC at 8 years	0.750 (± 0.021)	0.776
	C-index	0.736 (± 0.020)	0.763
Recalibrated base + glucose model	AU-ROC at 8 years	0.861 (± 0.023)	0.870
	C-index	0.849 (± 0.022)	0.853
Original base + glucose model	AU-ROC at 8 years	0.861 (± 0.023)	0.852
	C-index	0.850 (± 0.022)	0.837
Recalibrated base + glucose + lipids model	AU-ROC at 8 years	0.861 (± 0.023)	0.873
	C-index	0.848 (± 0.022)	0.849
Original base + glucose + lipids model	AU-ROC at 8 years	0.864 (± 0.023)	0.857
	C-index	0.851 (± 0.022)	0.842

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

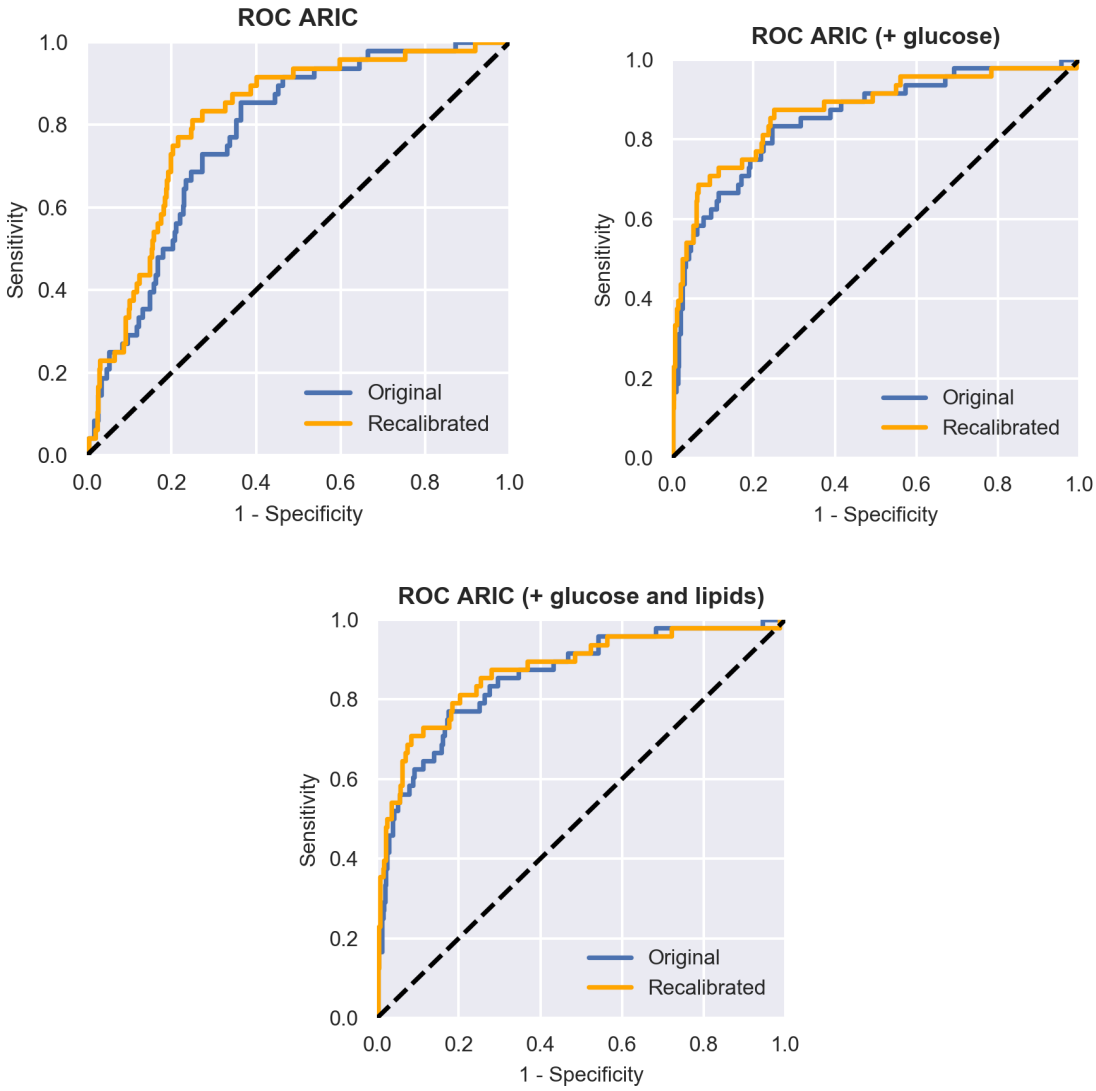

Figure 8. ROC curves for the original (blue) and recalibrated (orange) versions of the three ARIC models (base model on the top left, base + glucose on the top right, base + glucose + lipids on the bottom). The dashed line indicates random chance.

Table 14 and Figure 9 –which serves as its graphical counterpart— show how only the base model was badly calibrated (E/O ratio = 2.29, denoting frequent overestimation) and the clear superiority of its recalibrated version in this sense (E/O = 0.99). They also highlight the overall good quality of the ARIC models that include glucose information, which exhibit E/O ratios close to 1 both before and after recalibration.

Table 14. Calibration of the three ARIC models: E/O ratios and their 95% confidence intervals for the recalibrated and original models assessed on the test set.

Model	Metric	Test set
Recalibrated base model	E/O ratio [95% CI]	0.99 [0.75 – 1.31]
Original base model	E/O ratio [95% CI]	2.29 [1.72 – 3.03]
Recalibrated base + glucose model	E/O ratio [95% CI]	0.94 [0.71 – 1.25]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Model	Metric	Test set
Original base + glucose model	E/O ratio [95% CI]	1.13 [0.85 – 1.50]
Recalibrated base + glucose + lipids model	E/O ratio [95% CI]	0.95 [0.72 – 1.26]
Original base + glucose + lipids model	E/O ratio [95% CI]	1.17 [0.89 – 1.56]

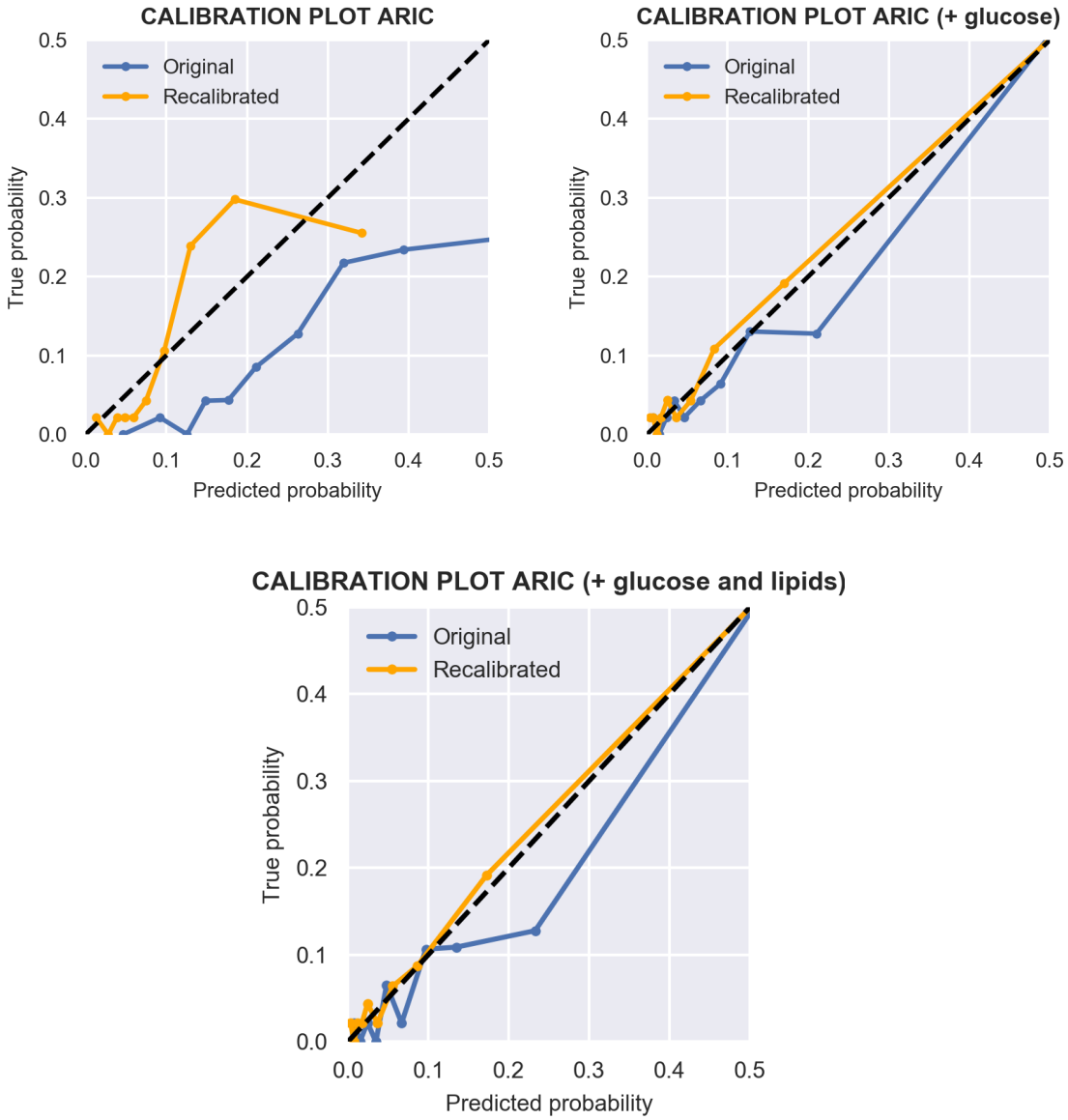

Figure 9. Calibration for the original (blue) and recalibrated (orange) versions of the three ARIC models (base model on the top left, base + glucose on the top right, base + glucose + lipids on the bottom). The dashed line represents perfect calibration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

4.4 Framingham model

4.4.1 The original model

In their original work [11], Wilson et al. proposed a number of clinical models for the estimation of the risk of T2D 7 years after the baseline visit. Among those, we selected the one for "multivariate prediction of T2D according to simple clinical variables", as it reportedly over-performed its simpler variants while not needing excessively specific information such as the result of an oral glucose tolerance test. For that model, they selected age, gender, BMI, family history of diabetes, hypertension medications, fasting glucose, HDL cholesterol, and blood pressure as covariates of interest. They fitted a logistic regression model on 3140 subjects, 160 of whom developed T2D, and reported the odds ratios linking each variable to the 7-year onset of diabetes. They also converted their "simple clinical model" into a scoring system, but did not explicitly state the criteria or the thresholds they used to translate odd ratios into partial scores.

4.4.2 Data selection and preprocessing

The initial dataset presented in Section 2.1 was further reduced to accommodate the specific characteristics of the Framigham model.

- Subjects for whom one or more model variables were not recorded were discarded.

The remaining sample comprised 3595 subjects, divided between a training and a test sets of 2879 and 716 subjects, respectively. Of the 2879 members of the training set, 324 developed diabetes within 8 years vs. 82 in the test set, thus preserving a similar cases to controls ratio.

4.4.3 Model implementation and recalibration

Variable preprocessing

The variables required for the implementation of the Framingham model were treated as in the appendix of the original work [11], with only the following minor deviation.

- "Hypertension or anti-hypertensive therapy" was set to 1 if the subject reported taking anti-hypertensive medications, had systolic blood pressure values ≥130, or diastolic blood pressure values≥85. Compare this with the original "Participants with a blood pressure level of 130/85mmHg or higher or receiving treatment for hypertension" [11].
- "Family history of diabetes" in the original study only referred to parents; here, it was extended to also include siblings and children.

Recalibration on the training set

The training set was used in the recalibration phase to fit a logistic regression model where the dependent variable was the onset of type 2 diabetes in an 8-year follow-up window and the independent variables were age, gender, BMI, fasting glucose, HDL cholesterol, family history of diabetes, triglycerides, and hypertension (medicated or otherwise). See Table 15 of Section 4.4.4 for further details on variables distribution across the training and test sets.

The original score system was not recalibrated on the MESA dataset due to the lack of information concerning the coefficient-to-score conversion system (see Section 4.2.3 for an example of a similar system). Furthermore, the actual rule used by the authors was impossible to reverse-engineer because of the major gaps between scores (e.g., highest score, Fasting glucose = 10 points vs. second highest score, BMI = 5 points).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

4.4.4 Results

The characteristics of the training and test subpopulations are reported in Table 15. As shown in the table, the predictive variables used in all the models present a similar distribution between the two sets and, in particular, the outcome is observed in a very similar percentage of subjects (11.3% vs. 11.5%).

Table 15. Distribution of Framingham model variables in the training and test sets reported as percentage of subjects in different variable categories for 1/0 variables and as their mean for continuous variables.

Variable	Category	% Subjects training set (N=2879)	% Subjects test set (N=716)
Age [years]	50-64	50.3%	47.1%
	≥65	33.8%	34.5%
Gender [Boolean]	Male	47.3%	43.0%
BMI [kg/m^2]	25 to <30	39.8%	39.0%
	≥30	29.6%	30.6%
Fasting glucose [mg/dL]	>100 to 126	13.4%	12.3%
HDL cholesterol [mg/dL]	Men: <40 Women: <50	32.9%	36.6%
Family history of diabetes [Boolean]	Yes	35.9%	31.8%
Triglycerides [mg/dL]	≥150	26.8%	29.1%
Hypertension or anti-hypertensive therapy [Boolean]	Yes	49.5%	46.1%
8-year incidence of drug-treated diabetes [Boolean]	Yes	11.3%	11.5%

The coefficients of the original and recalibrated logistic regression models are presented in

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 16. Note that, while in [11] the coefficients are reported in terms of odds ratios, here they have been calculated using the following simple formula.

$$\text{coefficient} = \ln(\text{odds ratio})$$

The recalibrated coefficients are consistent with the original ones, for the most part. There was a sign inversion between the original (negative) and recalibrated (positive) coefficients associated with age. However, as the former were comparatively very close to 0, this may just have been an effect of the different composition of the MESA dataset (e.g., few subjects younger than 45). Possibly for similar reasons, the absolute value of the male gender coefficient increased twenty-fold, reaching -0.2. A shift in the weight of the biomarkers on the final prediction was also apparent: a parallel could be established between the increase of the fasting glucose coefficient (2.552 vs. the original 1.981) and the decrease of the HDL cholesterol and triglycerides ones (respectively, 0.497 vs. the original 0.944 and 0.092 vs. 0.577).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 16. Coefficients of the Framingham model (fourth column) compared with those of the recalibrated version on the MESA dataset (third column).

Variable	Value	Recalibrated model coefficient	Original model coefficient
Intercept	-	-3.901	-5.517
Age [years]	50-64	0.177	-0.020
	≥65	0.138	-0.083
Gender [Boolean]	Male	-0.202	-0.010
BMI [kg/m^2]	25 to <30	0.216	0.300
	≥30	0.782	0.916
Fasting glucose [mg/dL]	>100 to 126	2.552	1.981
HDL cholesterol [mg/dL]	Men: <40 Women: <50	0.479	0.944
Family history of diabetes [Boolean]	Yes	0.638	0.565
Triglycerides [mg/dL]	≥150	0.092	0.577
Hypertension or anti-hypertensive therapy [Boolean]	Yes	0.392	0.500

The discrimination performance of the Framingham model is summarised in Table 17. Based on that information, it is quite hard to determine whether recalibration had any tangible effects on the AU-ROC and C-index metrics: on the one hand, test results were slightly better for the recalibrated model; on the other, the average bootstrap performance was slightly worse after recalibration, although the difference was not statistically significant. Figure 10 also highlights this behaviour: the recalibrated model slightly outperforms the original ones for low (1 – Specificity) values, but is overtaken by the latter for (1 – Specificity > 0.4).

Table 17. Performance of the Framingham model: AU-ROC and C-index for the recalibrated and original models assessed during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	Metric	Bootstrap validation	Test set
Recalibrated logistic regression	AU-ROC at 8 years	0.813 (± 0.019)	0.855
	C-index	0.799 (± 0.018)	0.839
Original logistic regression	AU-ROC at 8 years	0.815 (± 0.019)	0.853
	C-index	0.802 (± 0.018)	0.837

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

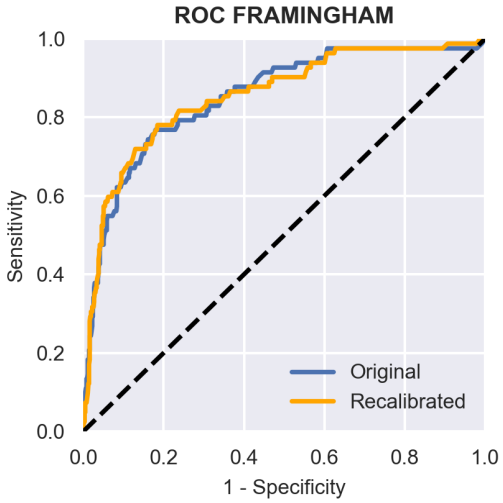

Figure 10. ROC curve for the original (blue) and recalibrated (orange) versions of Framingham model. The dashed line indicates random chance.

Table 18 and Figure 11 – which serves as its graphical counterpart— show that, although discrimination performance was very similar, the recalibrated model greatly outperformed the original one in terms of calibration. Indeed, the latter had a distinct tendency to underestimate the actual probability of developing diabetes (E/O ratio of 0.36), while the former was very well-calibrated.

Table 18. Calibration of the Framingham model: E/O ratios and their 95% confidence intervals for the recalibrated and original models assessed on the test set.

Model	Metric	Test set
Recalibrated logistic regression	E/O ratio [95% CI]	0.98 [0.79 – 1.21]
Original logistic regression	E/O ratio [95% CI]	0.36 [0.29 – 0.44]

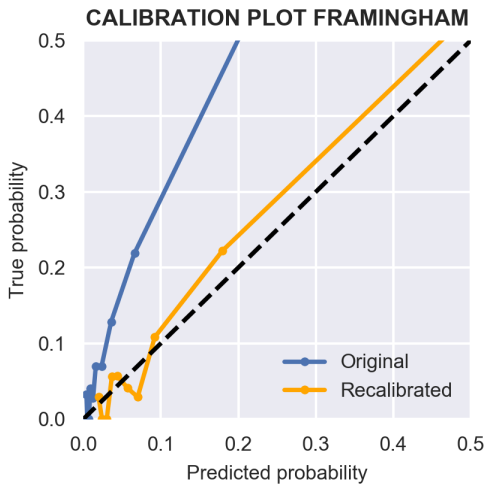

Figure 11. Calibration plot for the Framingham model in its original (blue) and recalibrated (orange) versions. The dashed line represents perfect calibration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

4.5 Basic risk score by Kahn et al.

4.5.1 The original model

Kahn et al. developed a basic risk score system [12] to identify adults at high risk of T2D by using longitudinal data from the ARIC Study, which included 15,792 white and black adults aged 45 to 64 at baseline which were followed-up for 14.9 years. This risk score system includes only simple variables that do not require any clinical examination, i.e., age, parental history of diabetes, ethnicity, hypertension, waist circumference, height, weight, resting pulse and smoking. The risk score system was derived from the coefficients of a Weibull proportional hazard regression model, which was used to model the effect of risk factors on time to diabetes onset. In particular, each variable was assigned a point score proportional to the respective β coefficient in the Weibull proportional hazard regression model. Then, the point score values were normalized so that the maximum total score was equal to 100. The risk score was internally validated to predict the 10-year diabetes incidence.

Note that Kahn et al. also developed an enhanced risk score that additionally includes variables collected in a blood specimen. However, the enhanced risk score could not be assessed because not all the variables in the enhanced score were collected in MESA.

4.5.2 Data selection and preprocessing

The initial dataset presented in Section 2.1 was further reduced to accommodate the specific characteristics of the Kahn risk score system.

- As in the ARIC models of Section 4.3, one of the predictive variables to be used in Kahn's score is the distinction between the African-American and white American ethnicities. As such, Hispanic and Asian subjects from the MESA dataset were excluded prior to the analyses.
- Subjects for whom one or more model variables were not recorded were also discarded.

The remaining sample comprised 2799 subjects, divided between a training and a test sets of 2247 and 552 subjects, respectively. In particular, diabetes was developed during the follow-up by 231 subjects of the training and 63 subjects of the test set, thus a similar diabetes incidence was maintained between training and test sets.

4.5.3 Model implementation and recalibration

Variable preprocessing

The variables required for the implementation of the Kahn basic risk score were selected from the MESA variable list and treated as in the original work [12]. In particular, the "hypertension" variable was defined as a binary variable equal to 1 if either systolic blood pressure is ≥140 mmHg or diastolic blood pressure is ≥90 mmHg.

Recalibration on the training set

In the training set, the Weibull proportional hazard regression model was fitted using the diabetes events and the respective times as outcome, and the following variables as independent variables: diabetic mother, diabetic father, hypertension, age, black race, ever smoking, waist circumference, height, weight and resting heart rate. From the model β coefficient, a recalibrated score was derived by adopting the same criteria of Kahn et al. [12].

4.5.4 Results

The characteristics of the training and test subpopulations are reported in Table 19. As shown in the table, the predictive variables used by the model present a similar distribution between the two sets and, in particular, the outcome is observed in a similar percentage of subjects (10.3% vs. 11.4%).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 19. Distribution of Kahn risk score variables in the training and test sets reported as percentage of subjects in different variable categories.

Variable	Category	% Subjects training set (N=2247)	% Subjects test set (N=552)
Age [years]	≥55	66.9%	65.9%
Race [Boolean]	Black	34.5%	36.6%
Diabetic mother [Boolean]	Yes	15.3%	13.8%
Diabetic father [Boolean]	Yes	11.5%	13.4%
Hypertension [Boolean]	Yes	45.6%	43.7%
Ever smoker [Boolean]	Yes	58.2%	60.0%
Waist circumference [cm]	Men: 90 to <95 Women: 81 to <88	15.5%	15.8%
	Men: 95 to <100 Women: 88 to <96	19.4%	19.8%
	Men: 100 to <106 Women: 96 to <105	20.2%	18.5%
	Men: ≥ 106 Women: ≥105	27.0%	28.8%
Height [cm]	Men: <171 Women <157	18.8%	19.4%
	Men: 171 to <175 Women: 157 to <161	22.1%	21.9%
	Men: 175 to <178 Women: 161 to <164	17.9%	18.1%
Weight [kg]	Men: ≥ 86.4 Women: ≥ 72.7	49.9%	49.1%
Heart rate [beats/min]	Men: ≥ 68 Women: ≥ 70	22.3%	23.2%
Incidence of diabetes [Boolean]	Yes	10.3%	11.4%

The coefficients of the model recalibrated in the whole training set are reported in Table 20, as well as the recalibrated score and the original score. Unfortunately, the coefficients of the recalibrated model cannot be compared to those of the original model because they were not reported in the paper by Kahn et al. [12]. For some of the variables, the recalibrated score differs significantly from the original one. For example, the “age ≥5” class was assigned 5 points in the original score vs. 0 in the recalibrated one. The higher categories of weight and heart rate were also assigned less points in the recalibrated model compared to the original one. However, other risk factors like black race and elevated waist circumference got more points in the recalibrated score vs. the original one. The points assigned to the height categories do not agree with the trend observed in the original score. In particular, the intermediate class is assigned a negative partial score in the recalibrated score, while all the height

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

classes present positive partial score in the original score. This difference may indicate the recalibrated model does not capture correctly the effect of height on the MESA data because of a suboptimal categorization of this variable. Alternatively, in the MESA population height may not have the same predictive power it had in the ARIC population studied by Kahn et al.

Table 20. Coefficients of the recalibrated Weibull proportional hazard model (third column), score points of the recalibrated Kahn's score (fourth column) and score points of the original Kahn's score (fifth column).

Variable	Category	Coefficients of the recalibrated model	Recalibrated score	Original score
Age [years]	≥55	-0.017	0	5
Race [Boolean]	Black	0.516	13	6
Diabetic mother [Boolean]	Yes	0.565	14	13
Diabetic father [Boolean]	Yes	0.319	8	8
Hypertension [Boolean]	Yes	0.046	11	11
Ever smoker [Boolean]	Yes	0.015	4	4
Waist circumference [cm]	Men: 90 to <95 Women: 81 to <88	0.036	9	10
	Men: 95 to <100 Women: 88 to <96	0.060	15	20
	Men: 100 to <106 Women: 96 to <105	1.198	30	26
	Men: ≥ 106 Women: ≥105	1.660	41	35
Height [cm]	Men: <171 Women <157	0.156	4	8
	Men: 171 to <175 Women: 157 to <161	-0.205	-5	6
	Men: 175 to <178 Women: 161 to <164	0.054	1	3
Weight [kg]	Men: ≥ 86.4 Women: ≥ 72.7	0.114	3	5
Heart rate [beats/min]	Men: ≥ 68 Women: ≥ 70	0.072	2	5
Scale parameter	-	$3.945 \cdot 10^{-7}$	-	-
Shape parameter	-	1.336	-	-

The performance of the recalibrated model and the recalibrated score in terms of discriminatory ability are summarized in Table 21 and compared to those of the original score. Results show that the recalibrated model and the recalibrated score achieved similar performance according to both the c-index and the AU-ROC at 8 years. However, the original score outperformed the recalibrated model

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

and the recalibrated score both in the validation phase and on the test set. This may be caused by the counterintuitive effect of the height variable already observed in Table 20. The same result is visible in Figure 12, where the ROC curve at 8 years is displayed for the recalibrated model (red), the recalibrated score (green) and the original score (blue).

Table 21. Performance of the Kahn's score: AU-ROC and C-index for the recalibrated model, the recalibrated score and the original score assessed during the validation phase (mean ± SD over the 100 bootstrap re-samplings) and on the test set.

Model	Metric	Bootstrap validation	Test set
Recalibrated Weibull model	AU-ROC at 8 years	0.729 (± 0.028)	0.799
	C-index	0.706 (± 0.023)	0.747
Recalibrated score	AU-ROC at 8 years	0.729 (± 0.028)	0.800
	C-index	0.705 (± 0.023)	0.747
Original score	AU-ROC at 8 years	0.755 (± 0.026)	0.815
	C-index	0.726 (± 0.022)	0.767

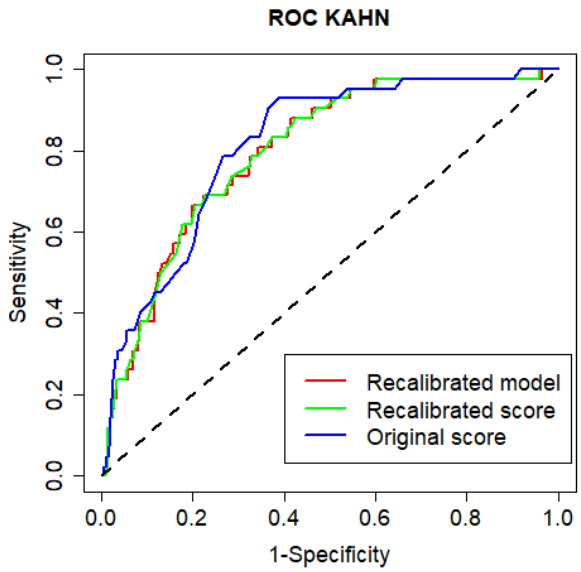

Figure 12. ROC curve for the original Kahn's score (blue), the recalibrated score (green) and the recalibrated model (red). The dashed line indicates random chance.

As showed in Table 22 and Figure 13, the recalibrated model presented good calibration on the test set. Note that it was not possible to assess the calibration of the original model from which the Kahn's score was derived because Kahn et al. did not report the coefficients of the Weibull model in their publication [12].

Table 22. Calibration of the Weibull proportional hazard model by Kahn et al. recalibrated on MESA data: E/O ratio and its 95% confidence interval.

Model	Metric	Test set
Recalibrated Weibull model	E/O ratio [95% CI]	0.93 [0.69 – 1.26]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

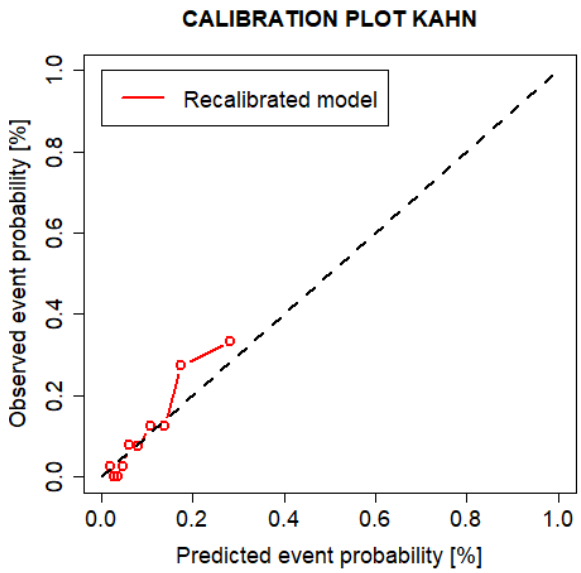

Figure 13. Calibration plot for the Kahn model recalibrated on MESA data. The dashed line represents perfect calibration.

4.6 DPoRT

4.6.1 The original model

The Diabetes Population Risk Tool (DPoRT) is a population-based risk prediction tool developed by Rosella et al. [13] 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. Rosella et al. validated the DPoRT in a cohort of 9899 subjects with 9-year follow-up and a cohort of 26465 subjects with 5-year follow-up.

4.6.2 Data selection and preprocessing

Subjects for whom none of the DPoRT model variables was missing were selected from the initial dataset presented in Section 2.1. The selected sample comprised 5121 subjects, divided between a training and a test sets of 4096 and 1025 subjects, respectively. In particular, diabetes was developed during the follow-up by 508 subjects of the training and 127 subjects of the test set, thus a similar diabetes incidence was maintained between training and test sets.

4.6.3 Model implementation and recalibration

Variable preprocessing

The variables required for the implementation of the DPoRT model were selected from the MESA variable list and treated as in the original work [13], with only the following deviations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

- Variable “heart disease” was not considered because having heart disease at exam 1 was an exclusion of MESA.
- The BMI-age categories related to age <45 years were not considered because the age of all the MESA participants was ≥45 years at exam 1.

Model recalibration on the training set

Sex-specific Weibull accelerated failure time models were fitted on training set data using the diabetes events and the respective times as outcome, and the following variables as independent variables:

- hypertension, non-white ethnicity, smoking, education, BMI-age, for the men;
- hypertension, non-white ethnicity, immigrant status, education, BMI-age, for the women.

4.6.4 Results

As shown in Table 23 and in Table 24, the predictive variables used by the DPoRT models for men and women presented a similar distribution between the training and the test set. In particular, the outcome is observed in a similar percentage of training and test set subjects.

Table 23. Distribution of DPoRT variables for the men in the training and test sets reported as percentage of subjects in different variable categories.

Variable	Value	% Subjects training set (N=1,948)	% Subjects test set (N=451)
Hypertension [Boolean]	Yes	32.8%	33.2%
Non-white ethnicity [Boolean]	Yes	58.8%	59.0%
Current smoker [Boolean]	Yes	19.5%	16.4%
Education	Post-secondary or higher	55.6%	54.3%
BMI [kg/m^2]	BMI 23-24	14.2%	14.6%
	BMI 25-29	46.1%	49.0%
	BMI 30-34	23.6%	21.5%
	BMI ≥35	5.7%	7.31%
Incidence of diabetes [Boolean]	Yes	12.5%	13.3%

Table 24. Distribution of DPoRT variables for the women in the training and test sets reported as percentage of subjects in different variable categories.

Variable	Value	% Subjects training set (N=2,148)	% Subject test set (N=574)
Hypertension [Boolean]	Yes	37.8%	35.5%
Non-white ethnicity [Boolean]	Yes	60.1%	63.6%
Immigrant status [Boolean]	Yes	31.4%	34.1%
Education	Post-secondary or higher	45.0%	44.6%

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Value	% Subjects training set (N=2,148)	% Subject test set (N=574)
BMI [kg/m^2], age [years]	BMI 23-24, age <65	7.0%	6.3%
	BMI 25-29, age <65	21.0%	20.4%
	BMI 30-34, age <65	13.0%	12.9%
	BMI ≥35, age <65	10.6%	10.3%
	BMI <23, age ≥65	6.4%	4.2%
	BMI 23-24, age ≥65	5.4%	6.1%
	BMI 25-29, age ≥65	15.1%	14.5%
	BMI 30-34, age ≥65	7.4%	10.1%
	BMI ≥35, age ≥65	4.8%	4.4%
Incidence of diabetes [Boolean]	Yes	12.3%	11.7%

The coefficients of the Weibull accelerated failure time models recalibrated on the whole training set are compared with those of the original DPoRT model in Table 25 for the men, in Table 26 for the women. Both for the men and the women, the effect of all the considered independent variables in the recalibrated model is in the same direction (same sign of the coefficient) as in the original model. However, for some variables, especially the BMI-age categories, the coefficients of the recalibrated model are significantly different from those of the original model, probably because of the different characteristics of the MESA population compared to the original DPoRT cohort (e.g., different age range).

Table 25. 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	-	10.5136	10.5971
Hypertension [Boolean]	Yes	-0.3168	-0.2624
Non-white ethnicity [Boolean]	Yes	-0.2471	-0.6316
Heart disease [Boolean]	Yes	-	-0.5355
Current smoker [Boolean]	Yes	-0.2093	-0.1765
Education	Post-secondary or higher	0.1345	0.2344
BMI [kg/m^2], age [years]	BMI 23-24, age <45	-	-1.2378
	BMI 25-29, age <45	-	-1.5490
	BMI 30-34, age <45	-	-2.5437
	BMI ≥35, age <45	-	-3.4717
	BMI <23, age ≥45	0	-1.9749

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Value	Recalibrated model coefficient	Original model coefficient
	BMI 23-24, age ≥45	0.0180	-2.4426
	BMI 25-29, age ≥45	-0.6986	-2.8588
	BMI 30-34, age ≥45	-1.0240	-3.3179
	BMI ≥35, age ≥45	-1.4360	-3.5857
Scale	-	0.7539	0.8049

Table 26. 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	-	10.3016	10.5474
Hypertension [Boolean]	Yes	-0.3761	-0.2865
Non-white ethnicity [Boolean]	Yes	-0.3362	-0.4309
Immigrant status [Boolean]	Yes	-0.3228	-0.2930
Education	Post-secondary or higher	0.1884	0.2042
BMI [kg/m^2], age [years]	BMI 23-24, age <45	-	-0.5432
	BMI 25-29, age <45	-	-0.8453
	BMI 30-34, age <45	-	-1.4104
	BMI ≥35, age <45	-	-2.0483
	BMI <23, age 45-64	0	0.0711
	BMI 23-24, age 45-64	-0.0553	-0.7011
	BMI 25-29, age 45-64	-0.1041	-1.4167
	BMI 30-34, age 45-64	-0.5701	-2.2150
	BMI ≥35, age 45-64	-1.1534	-2.2695
	BMI <23, age ≥65	0.7372	-1.0823
	BMI 23-24, age ≥65	-0.2427	-1.1419
	BMI 25-29, age ≥65	-0.3415	-1.5999
	BMI 30-34, age ≥65	-0.4224	-1.9254
	BMI ≥35, age ≥65	-0.6676	-2.1959
Scale	-	0.7051	0.7814

The performance of the recalibrated model and the original model in terms of discriminatory ability are quantitatively reported in Table 27 (C-index and AU-ROC at 8 years) and graphically represented in Figure 14 (ROC curve at 8 years). In particular, the recalibrated model outperforms the original model in the test set, as showed by the steeper ROC curve of the recalibrated model (red line in Figure 14)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

compared to the original model (blue line in Figure 14), although in the bootstrap validation the two models present comparable average C-index and AU-ROC values.

Table 27. Performance of discriminatory ability of the DPoRT model: AU-ROC and C-index for the recalibrated model and the original score assessed during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Model	Metric	Men		Women	
		Bootstrap validation	Test set	Bootstrap validation	Test set
Recalibrated model	C-index	0.665 (±0.023)	0.707	0.690 (±0.023)	0.706
	AU-ROC at 8 years	0.679 (±0.028)	0.735	0.718 (±0.027)	0.756
Original model	C-index	0.665 (±0.023)	0.697	0.689 (±0.024)	0.687
	AU-ROC at 8 years	0.680 (±0.027)	0.728	0.716 (±0.027)	0.728

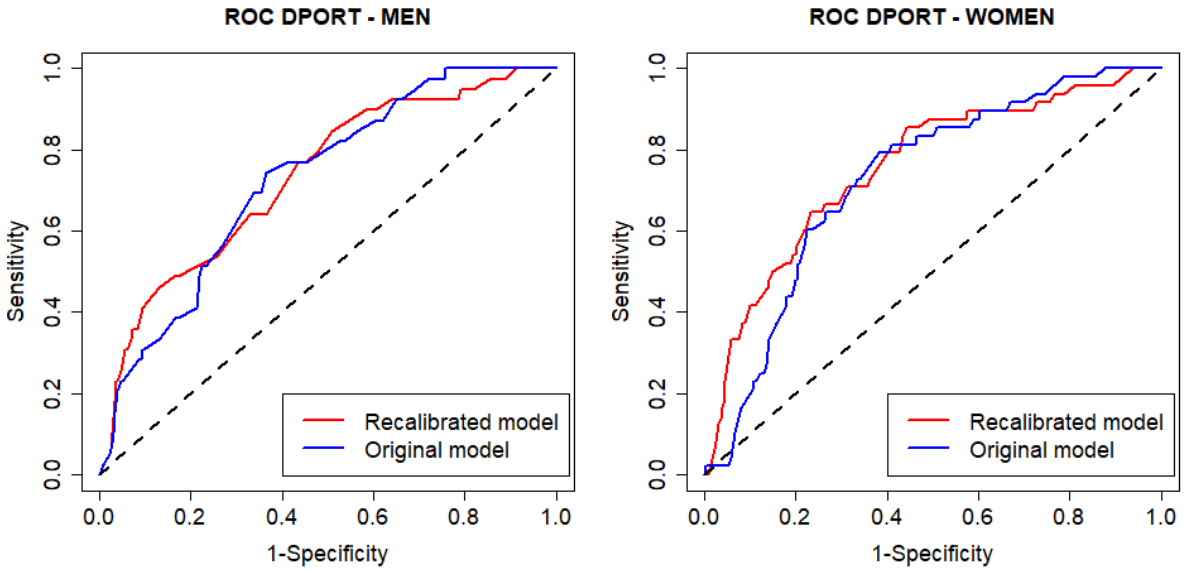

Figure 14. ROC curve at 8 years for the original DPoRT model (blue) and the recalibrated DPoRT model (red). The dashed line indicates random chance.

As far as calibration is concerned, the E/O ratio values reported in Table 28 and the calibration plots of Figure 15 demonstrate that the original DPoRT model significantly overestimates the diabetes outcome 8 year after the baseline. Conversely, the recalibrated model present better calibration, although it slightly underestimates the diabetes incidence for patient with high predicted event probability (see how the red curves in Figure 15 deviates from the dashed line as predicted event probability increases).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 28. Calibration of the DPoRT model: E/O ratio and its 95% confidence interval for the recalibrated and original models assessed on the test set.

Model	Metric	Men	Women
Recalibrated model	E/O ratio	0.67 [0.49 – 0.92]	0.72 [0.54 – 0.96]
Original model	E/O ratio	5.93 [4.33 – 8.12]	2.87 [2.16 – 3.81]

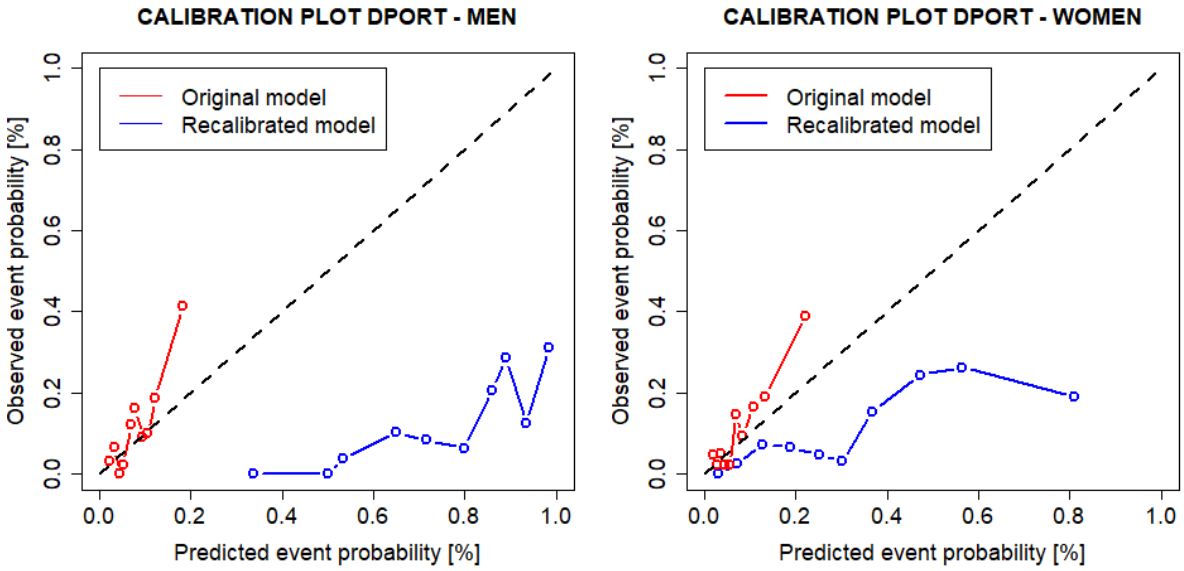

Figure 15. Calibration plot at 8 years for the original DPoRT model and its recalibrated version (men in the left panel, women in the right panel). The dashed line represents perfect calibration.

4.7 Discussion

Eight literature models were assessed on the MESA dataset both in their original version and after recalibration on a suitable training set extracted from MESA. Such models can be divided into the following 2 categories.

- Non-invasive models, i.e., models which do not require the collection of any invasive biomarker. The FINDRISC, the DPoRT, the ARIC simple model and the Kahn's basic score belong to this category.
- Invasive models, i.e., models which require variables deriving from laboratory tests, e.g., blood tests. The Stern's model, the ARIC clinical models and the Framingham model belong to this category.

Non-invasive models present a wider applicability than invasive models as they only use easily accessible information. Nevertheless, invasive models generally perform better than non-invasive models because they include some crucial predictors of diabetes, e.g. fasting plasma glucose. This was confirmed in our assessment: invasive models showed better performance than non-invasive models in terms of their ability to correctly rank subjects according to diabetes risk. The best performance was achieved by the Stern's model (C-index equal to 0.86 on the test set) followed by the ARIC clinical model, which achieved very similar performance with and without using lipids concentration, and the Framingham model. Remarkably, all these models use as independent variable the fasting glucose concentration. Our analysis on the ARIC model evidenced that fasting glucose concentration has, not surprisingly, a remarkable predictive power; indeed, the addition of fasting

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

glucose concentration to the independent variables of the ARIC model increased the C-index value of the test set from 0.76 to 0.84 for the original model, from 0.79 to 0.85 for the recalibrated model. Regarding the non-invasive models, best performance was achieved by the Kahn's score (C-index equal to 0.77 on the test set) followed by the ARIC simple model, the FINDRISC model and the DPoRT model. Note that while the Kahn's score and the ARIC simple model required the measurement of heart rate or blood pressure, the FINDRISC and the DPoRT model use only self-reported information on medical history, biometric parameters and other general variables.

Concerning the comparison between recalibrated vs. original models, the coefficients of the recalibrated models were consistent with those of the original models (same sign). However, in the recalibrated models the effect of age was smaller than in the original models (age coefficients of the recalibrated models close to 0), suggesting that in the MESA population age may have low predictive ability. This is probably due to the limited age range observed in MESA at exam 1, i.e. 45-84, thus young people who, in general, are at lower risk of developing T2D are not present in the MESA dataset. In terms of model performance, our analysis evidenced that recalibration significantly improves the accuracy of the models in predicting the observed diabetes incidence (E/O ratio closer to 1 for the recalibrated models compared to the original models). In particular, the models for which recalibration was most beneficial in terms of E/O ratio are the DPoRT model and the ARIC simple model. However, recalibration does not affect much the ability of the models to correctly rank the subjects according to diabetes risk, as C-index and AU-ROC values were comparable between the recalibrated and the original models for almost all the models. According to the c-index, the model for which recalibration was most beneficial is the ARIC simple model (C-index from 0.76 to 0.79), while for the FINDRISC, the Stern's model and the Framingham model the recalibration did not change at all the C-index of the test set.

5 IMPLEMENTATION, RECALIBRATION AND ASSESSMENT OF STATE-OF-THE-ART MODELS OF ASTHMA ONSET

5.1 Model by Thomsen et al.

5.1.1 The original model

In work by Thomsen et al. [14], a study to establish the risk factors for the development of asthma in young adults was performed using the longitudinal data collected in The Danish Twin Registry for birth cohorts over the period 1953-1982. In particular, the data of 19,349 subjects with no history of asthma in 1994 who answered to the follow-up questionnaire in 2002 were selected for the analysis. The age at baseline of selected subjects was 12-41 years. A logistic regression model was applied to investigate the association of possible risk factors at baseline with asthma onset at follow-up. The analysis was performed separately for subjects of age 12-19 years and 20-41 years. For subjects in the latter group, the model independent variables were gender, age, BMI, smoking and physical activity. Note that the aim of work by Thomsen et al. was to identify risk factors for asthma onset in young adults, not to develop a tool to predict asthma onset. Indeed, Thomsen et al. did not test the ability of the proposed logistic regression model to correctly predict future onset of asthma.

5.1.2 Data selection and preprocessing

A suitable data sample for the recalibration of the Thomsen's model was selected from the initial dataset presented in Section 2.2. Since the Thomsen's model is based on logistic regression, we selected the subjects having the outcome defined at a fixed cut-off time. After analysing the MESA data, we chose the cut-off of 10 years after the exam 1, which was considered as the baseline. This cut-off allowed us to have a good trade-off between the number of subjects discarded because developing asthma after the cut-off and the number of subjects discarded because censored before the cut-off. In addition, the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

subjects for whom at least one of the Thomsen model variables was missing at the baseline were also discarded.

The selected sample comprised 624 subjects, divided between a training and a test sets of 506 and 127 subjects, respectively. Ten years after the baseline asthma was present in 104 subjects of the training set and 26 subjects of the test set, thus a similar asthma incidence was maintained between training and test sets.

5.1.3 Model recalibration

Variable preprocessing

The variables required for the implementation of the Thomsen's model were selected from the MESA variable list and treated as in the original work [14], with only the following deviations.

- Variable "smoking" had 4 levels in the Thomsen's original model, i.e. current daily, current occasional, former, never. Since the MESA data do not allow to distinguish between current daily and occasional smokers, for model recalibration only three levels were considered, i.e., current, former and never.
- Physical activity was defined in hours/week in the Thomsen's original model. In MESA, information on physical activity was collected by the MESA Typical Week Physical Activity Survey, which assesses the time spent in and frequency of various physical activities over the past month [15]. Metabolic equivalents (METs) were assigned to each physical activity, and the total MET-minutes per week of physical activity was determined for each participant for three intensity levels, i.e., light, moderate and vigorous. In our analysis, a variable representing the total MET-min of moderate and vigorous physical activity was used to describe physical activity. The tertiles of the distribution of this variable at the baseline exam were used to define the following three categories: less than 2,698 MET-min/week (low), between 2,698 and 6,165 MET-min/week (medium), and more than 6,165 MET-min/week (high).

Model recalibration on the training set

The training set was used in the recalibration phase to fit a logistic regression model where the dependent variable was the onset of asthma in a 10-year follow-up window and the independent variables were age, gender, BMI, smoking and physical activity.

The original model was impossible to implement on the MESA dataset because the intercept parameter was not reported in the paper by Thomsen et al. [14] and the "smoking" and "physical activity" variables presented a different definition or different categories in MESA compared to the dataset of Thomsen et al. [14].

5.1.4 Results

As shown in Table 29, the predictive variables used by the model present a similar distribution between the training and the test set and, in particular, the outcome is observed in a similar percentage of subjects (20.6% vs. 20.5%).

Table 29. Distribution of Thomsen model variables in the training and test sets reported as percentage of subjects in different variable categories for 0/1 variables, mean (SD) for continuous variables.

Variable	Value	% Subjects training set (N=506)	% Subjects test set (N=127)
Gender	Male	41.5%	38.6%
Age [years]	-	57.9 (9.1)	58.0 (9.1)
BMI [kg/m^2]	-	27.87 (5.5)	27.9 (5.5)
Smoking	Former	32.4%	31.5%

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Value	% Subjects training set (N=506)	% Subjects test set (N=127)
	Current	18.6%	14.2%
Physical activity	Medium	31.0%	38.6%
	High	31.6%	35.4%
Incidence of asthma [Boolean]	Yes	20.6%	20.5%

The coefficients of the logistic regression model recalibrate on the training set are reported in Table 30. We can observe that male gender and physical activity have a negative impact on the outcome, while age, BMI and smoking have a positive effect on the risk of asthma.

The performance in terms of discriminatory ability is shown in Table 31 and Figure 16. Although the C-index and the AU-ROC are close to 0.7 on the test set, lower average performance are obtained in the bootstrap validation. Model calibration on the test set was good, with E/O ratio equal to 1 (Table 32).

Table 30. Coefficients of the Thomsen model recalibrated in the training set.

Variable	Value	Recalibrated model coefficient
Intercept	-	-5.399
Gender	Male	-0.628
Age [years]	per year	0.054
BMI [kg/m^2]	per unit	0.034
Smoking	Former	0.615
	Current	0.717
Physical activity	Medium	-0.336
	High	-0.481

Table 31. Performance of discriminatory ability of the Thomsen model: AU-ROC and C-index for the recalibrated model during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Metric	Validation	Test set
C-index	0.637 (±0.037)	0.688
AU-ROC at 10 years	0.648 (±0.040)	0.690

PULSE Participatory Urban Living for Sustainable Environments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

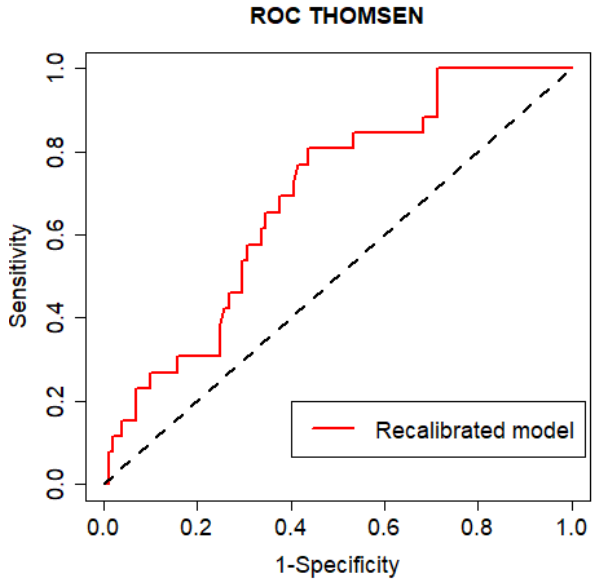

Figure 16. ROC curve at 10 years for the recalibrated Thomsen model (red). The dashed line indicates random chance.

Table 32. Calibration of the Thomsen model: E/O ratio and its 95% confidence interval for the recalibrated model on the test set.

Metric	Test set
E/O ratio	1.00 [0.68 – 1.46]

5.2 Model by Verlato et al.

5.2.1 The original model

Verlato et al. investigated the association of smoking with new onset of asthma in adults, taking into account also other variables as confounding factors [16]. For this purpose, data collected in 3 population cohorts extracted from the Italian Study on Asthma in Young Adults and the Italian Study on the Incidence of Asthma were used. In particular, 5241 subjects without history of asthma at baseline were selected. Subjects participated in a follow-up survey on average 9 years after the baseline survey. On these data, a multivariate logistic regression model was fitted using asthma onset at follow-up as dependent variable, smoking habits, age, gender, occupation, asthma symptoms, chronic bronchitis, allergic rhinitis and cohort as independent variables.

As for previous studies, the aim of work by Verlato et al. was to study risk factors associated with asthma onset, not to develop a tool to predict asthma onset. Therefore, the prediction ability of the multivariable logistic regression model was not tested by Verlato et al.

5.2.2 Data selection and preprocessing

A suitable data sample for the recalibration of the Verlato's model was selected from the initial dataset presented in Section 2.2. Since the Verlato's model requires some variables (e.g., chronic bronchitis, allergic rhinitis and asthma symptoms) that were collected only in MESA Lung, our analysis was restricted to the MESA Lung cohort. For each subject the baseline was defined as either exam 3 or exam 4 depending on when he or she entered MESA Lung. Then, since the Verlato's model is based on logistic regression, we selected the subjects having the outcome defined at a fixed cut-off time. After

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

analysing the MESA Lung data, we chose the cut-off of 6 years after the baseline. This cut-off allowed us to have the better trade-off between the number of subjects discarded because developing asthma after the cut-off and the number of subjects discarded because censored before the cut-off. In addition, the subjects for whom at least one of the Verlato's model variables was missing at the baseline were also discarded.

The selected sample comprised 561 subjects. Because of the short duration of the follow-up, only 34 subjects presented asthma at 6 years after the baseline. Given the small number of cases, we decided to assess the recalibrated model performance by performing the bootstrap validation on the entire dataset, without extracting the test set.

5.2.3 Model recalibration

Variable preprocessing

The variables required for the implementation of the Verlato's model were selected from the MESA variable list and treated as in the original work [16], with the following deviations.

- Since some of the age categories of the Verlato's model are not represented in MESA, we defined different categories for age. In particular, two classes were defined: <65 years and ≥65 years.
- We did not consider the variable "occupation" because we did not find a correspondence between the categories defined in Verlato's model and those available in MESA
- We did not consider the variable "cohort" because a single cohort is available in MESA
- Allergic rhinitis was defined as the answer to question "Do you have any nasal allergies including hay fever?" in Verlato's model, "Have you ever had hay fever (allergies involving the nose and/or eyes)?" in MESA.
- In Verlato's model asthma symptoms included wheezing, tightness in the chest or shortness of breath in the last 12 months. In MESA, participants were not asked for tightness in the chest, thus only wheezing and shortness of breath in the last 12 months were used to define the "asthma symptoms" variable.

Model recalibration

A logistic regression model where the dependent variable was the onset of asthma in a 6-year follow-up window and the independent variables were age, gender, smoking, asthma symptoms, allergic rhinitis and chronic bronchitis was fitted on the entire selected dataset and on 100 sets extracted by bootstrap resampling.

The original model was impossible to implement on the MESA dataset because the intercept parameter was not reported in the paper by Verlato et al. [16] and because of all the discrepancies in variable definition reported in the "Variable preprocessing" section.

5.2.4 Results

The model variable distribution in the sample selected for model recalibration is summarized in Table 33. The coefficients of the logistic regression model recalibrated on the entire dataset are reported in Table 34. We can observe that male gender and allergic rhinitis have a negative impact on the outcome, while age, smoking, asthma symptoms and chronic bronchitis have a positive effect on the risk of asthma. The performance in terms of discriminatory ability are shown in Table 35. The C-index presents average value of 0.65 and a high standard deviation, which suggests the model performance are highly sensitive to the particular set of data used for model training. This was expected given the small number of subjects who develop asthma in this subset of data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 33. Distribution of Verlato model variables in the training and test sets reported as percentage of subjects in different variable categories for 0/1 variables, mean (SD) for continuous variables.

Variable	Value	% Subjects (N=561)
Gender	Male	49.4%
Age [years]	≥65	52.4%
Asthma symptoms [Boolean]	Yes	23.5%
Chronic bronchitis [Boolean]	Yes	6.2%
Allergic rhinitis [Boolean]	Yes	31.0%
Smoking	Former	32.2%
	Current	7.3%
Incidence of asthma [Boolean]	Yes	4.8%

Table 34. Coefficients of Verlato's logistic regression model recalibrated on the data selected from the MESA dataset.

Variable	Value	Model coefficient
Intercept	-	-3.384
Gender	Male	-0.746
Age [years]	≥65	0.058
Asthma symptoms [Boolean]	Yes	1.320
Chronic bronchitis [Boolean]	Yes	1.147
Allergic rhinitis [Boolean]	Yes	-0.594
Smoking	Former	0.127
	Current	0.510

Table 35. Performance of discriminatory ability of the Verlato's model: AU-ROC and C-index for the recalibrated model during the validation phase (mean ± SD over the 100 bootstrap resamplings) and on the test set.

Metric	Validation phase
C-index	0.654 (±0.109)
AU-ROC at 6 years	0.689 (±0.090)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

5.3 Discussion

We used the data collected in MESA to recalibrate two literature models of asthma adult-onset, both based on logistic regression. The Thomsen's model uses only general information and variables related to subject lifestyle, such as smoking and physical activity, while the Verlato's model includes variables more related to the respiratory health, such as asthma symptoms and chronic bronchitis.

The performance of the two models were not satisfactory. In particular, the Thomsen's model showed acceptable discriminatory ability on the test set (C-index close to 0.7), but lower performance in the bootstrap validation (slightly higher than a random predictor). The Verlato's model presented highly variable discriminatory ability in the bootstrap validation.

However, we would like to remark that important limitations of our analysis are the small size of the datasets used for model recalibration and, for the Verlato's model, the short duration of the follow-up (having exam 3 or 4 as baseline it was not possible to observe a follow-up longer than 6 years), which did not allow to observe a sufficient number of cases.

6 SELECTION OF NEW POTENTIALLY PREDICTIVE VARIABLES

Several studies were published in the literature that investigate possible risk factors of T2D and asthma onset. Based on the evidences provided by the literature studies, we identified new variables potentially predictive of T2D and asthma onset, which were not used in the state-of-the-art models. These variables include medication use, psychological factors, habits and indicators of individual and neighbourhood socio-economic status. Then, the full set of candidate predictive variables for T2D and asthma onset was obtained merging the variables already considered by the state-of-the-art models and the new identified variables. As later described in Section 7, we studied the probabilistic relationships between the candidate predictive variables and the onset of T2D and asthma by static Bayesian networks. Provided that many of the candidate predictive variables, especially clinical variables, were not collected in the Health and Retirement Study, and this study was focused on a particular population, i.e. old subjects, we decided to use only the MESA dataset to study the predictive ability of the new variables. Therefore, the new candidate predictive variables of diabetes and asthma that we assessed in the Bayesian network models were selected from the MESA baseline survey (exam 1). Such variables are described in Section 6.1 and Section 6.2, respectively.

6.1 Variables for diabetes model

Several studies have shown that the socio-economic status of an individual can affect his or her health status. In particular, diabetes incidence was found associated with socio-economic factors, like occupation, education level, income and marital status [7][17]-[19]. In MESA the socio-economic status was assessed in the MESA Personal History questionnaire. In particular, the following socio-economic variables were selected as candidate predictors of diabetes: marital status, education level, occupation, family income, number of dependents sustained by the family income and ongoing financial strain.
Recent studies have also indicated that the neighbourhood characteristics can affect health status, and in particular diabetes incidence, independent of the individual socio-economic status [20]. Indeed, neighbourhood environment can influence diet and physical activity through the availability of grocery stores, parks and other recreational facilities [21]. In addition, presence of noise, violence and poverty in the neighbourhood are sources of chronic stress that can affect the health status. In MESA, participants were surveyed about their neighbourhood characteristics by the MESA Neighbourhood questionnaire. In particular, the following neighbourhood variables were selected as candidate predictors of diabetes: lack of parks or playgrounds, lack of sidewalks, lack of access to adequate food shopping, heavy traffic, excessive noise, presence of trash, violence.
Psychological factors may also play a role in the onset of chronic disease, such as diabetes. For example, a positive association between anxiety and depression and the onset of diabetes was demonstrated in the study by Engum [22]. The study by Schmitz et al. highlighted the interaction between depressive symptoms and metabolic dysregulation as a risk factor for T2D onset [23]. In another study performed on the MESA data, trait anger was found positively associated with future

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

development of T2D [24]. In MESA, the psychological status was assessed by the MESA Health and Life questionnaire. Based on the literature evidences, we selected the Spielberg Trait Anger Scale, the Spielberg Trait Anxiety Scale and the Center for Epidemiologic Studies Depression Scale as candidate risk factors for diabetes onset. We also selected the variable labelled as "chronic burden", which assigned a score between 0 and 5 according to the ongoing experience of the following problems: serious personal health problem, serious health problem of a close person, difficulties with the job or the ability to work, difficulties in a relationship with a close person, financial strain. For the characterization of depression, we also extracted from the MESA dataset the variables related to the use of anti-depressants.
In literature studies, the use of certain drugs was also find associated with diabetes. For instance, in study by Mikkelsen et al. the use of antibiotics was identified as a potential risk factor of T2D [25]. Other studies investigated the role of aspirin in stimulating insulin and glucagon secretion [26]. The MESA dataset includes a section on medication use. According to literature evidences, we selected the variables related to the use of antibiotics, aspirin and other anti-inflammatory medications. In addition, we selected the variables related to use of lipid lowering medication, since some state-of-the-art models include high lipids concentration as risk factors, and thyroid medications, because there is literature evidence suggesting thyroid disorders and diabetes affect each other [27].
Another potential risk factor for diabetes onset is alcohol consumption. In particular, according to a recent review article, light-to-moderate alcohol consumption decreases the incidence of diabetes in the majority of the studies, whereas heavy drinkers and binge drinkers are at increased risk for diabetes [27]. In MESA, alcohol consumption was assessed in the MESA Personal History questionnaire; therefore, variables related to the alcohol consumption were selected from this questionnaire as candidate predictive variables.
A summary of the new potentially predictive variables of T2D is reported in Table 36.

Table 36. New potentially predictive variable of T2D extracted from the MESA dataset.

Category	Variables
Socio-economic variables	Marital status, education level, occupation, family income, number of dependents sustained by the family income, ongoing financial strain
Neighbourhood characteristics	Lack of parks or playgrounds, lack or sidewalks, lack of access to adequate food shopping, heavy traffic, excessive noise, presence of trash, violence
Psychological factors	Trait anxiety scale, trait anger scale, depression scale, chronic burden
Medications	Anti-depressants, antibiotics, aspirin, other anti-inflammatory medications, thyroid medications
Habits	Alcohol consumption

6.2 Variables for asthma model

Many literature studies demonstrated an association between personal socio-economic status and incidence of adult asthma [29]-[31]. As for diabetes, we therefore selected marital status, education level, occupation, family income, number of dependents sustained by the family income and ongoing financial strain from the MESA dataset, as socio-economic variables potentially predictive of asthma onset.
Living neighbourhood characteristics were also selected as candidate risk factors for asthma. Indeed, neighbourhood characteristics may affect physical activity, diet and stress, which were all found associated with asthma onset or lung function deterioration in literature studies [32]-[34].
Some literature studies also suggested that psychological conditions, such as depression, anxiety and anger, can have a role in the onset and progression of asthma and other respiratory diseases [34][35]. According to these evidences, the Spielberg Trait Anger Scale, the Spielberg Trait Anxiety Scale, the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Center for Epidemiologic Studies Depression Scale and the chronic burden variable were selected from the MESA variable lists as candidate risk factors for asthma onset.
Regarding the use of medications, a positive association between use of antibiotics and asthma onset in children was reported in the literature [36]. In a study by Thomsen et al. [37], the regular use of non-steroidal anti-inflammatory drugs was demonstrated to increase the risk of asthma adult-onset. Use of aspirin was also shown to have a causal effect on asthma symptoms [38]. Based on these evidences, variables related to use of antibiotics, anti-inflammatory medications and aspirin were extracted from the MESA dataset as candidate predictors of asthma.
Exposure to second hand smoke was identified as a predictor of asthma symptoms in many literature studies (e.g., [39]). In a study by Lajunen et al., exposure to second hand smoke and parental history of asthma were demonstrated to have a synergistic effect on the risk of asthma onset [40]. Then, both exposure to second hand smoke and family history of asthma were selected from the MESA dataset and included in the candidate predictive variable set.
Concerning alcohol consumption, evidences reported for asthma are similar to those observed for diabetes: a moderate alcohol consumption seems associated with reduced risk of asthma onset, while heavy daily drinkers have an increased risk of asthma onset [41].
Finally, two variables related to respiratory problems while sleeping, i.e., use of two pillows to help breathing and waking breathless at night, were selected from the MESA variables collected at the baseline visit, because linked to commonly observed asthma symptoms.
Note that other potential risk factors of asthma were assessed in the MESA Lung study, such as wheezing, chough and exposure to dust and fumes. However, we decided not to select those variables because, as shown when recalibrating the Verlato's model (Section 5.2), only few subjects developed asthma in the MESA Lung cohort, thus any model relying on those data would be affected by high uncertainty.
A summary of the new potentially predictive variables of asthma is reported in Table 37.

Table 37. New potentially predictive variable of asthma extracted from the MESA dataset.

Category	Variables
Socio-economic variables	Marital status, education level, occupation, family income, number of dependents sustained by the family income, ongoing financial strain
Neighbourhood characteristics	Lack of parks or playgrounds, lack or sidewalks, lack of access to adequate food shopping, heavy traffic, excessive noise, presence of trash, violence
Psychological factors	Trait anxiety scale, trait anger scale, depression scale, chronic burden
Medications	Anti-depressants, antibiotics, aspirin, other anti-inflammatory medications
Habits	Alcohol consumption, exposure to second-hand smoke
Symptoms	Sleep with two pillows to help breathing, waking breathless at night

7 STATIC BAYESIAN NETWORK MODELS

Bayesian Networks (BNs) are descriptive models that encode the probabilistic relationships among variables; thus, BNs can be used to assess the most probable values of a specific variable, based on the values of some others. For instance, Bayesian Networks identify the combination of factors maximizing the probability of diabetes or asthma outcome. Thus, BN models provide a further level of information in addition to the models presented in Sections 4 and 5, which rank the subjects according to their risk of diabetes or asthma onset.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

More in detail, a Bayesian Network [42][43] is a mathematical description of a joint probability distribution of a set of random variables based on a set of conditional independence assumptions. The structure of a Bayesian Network is a directed acyclic graph (DAG) such that each random variable corresponds to a node and the influence of one node (parent) on another (child) corresponds to a directed edge. The network structure induces a set of conditional probability distributions (CPDs), since each variable is a stochastic function of its parents. The network structure annotated with its CPDs, define a Bayesian Network (BN).

BNs can thus detect probabilistic relationships among variables and the exploitation of prior knowledge in the learning process allows the BNs to deal even with variables with missing values.

A two-step iterative procedure can be adopted to infer the BN on a training dataset: i) learning the graph topology (i.e., the parents-children dependencies among nodes) and ii) learning the parameters of each CPD (i.e., the probability that a variable assumes a specific value conditional to each possible joint assignment of values to its parents). The structure learning can be performed through three types of algorithms: score-based, constraint-based and hybrid. They all rely on a set of assumptions: the relationships between variables are conditional independencies, the observations are considered as independent and identically distributed samples of a population, and two different nodes cannot be a deterministic function of a single variable. The parameters learning is an estimation problem that is usually solved through techniques like Bayesian estimation or (regularized) maximum likelihood.

In detail, the structure learning was performed using Hill-Climbing (HC) algorithm [44], a greedy search-and-score method that starts with an initial graph (empty graph in our case) and searches the complete space of possible graph structures, by adding, reversing or deleting edges. The HC repeats as long as a specific score (Bayesian Information Criterion scoring was our choice) is maximized or a specific number of iterations has been recorded. Thus, the structure learning phase provided the topology of the BN with the highest probability to have generated the data. Subsequently, a Maximum a Posteriori estimation computed the set of parameters of the conditional probability distribution at each node. This procedure was implemented in R as a combination of a set of in-house script with bnstruct [45], an R package that makes use of state-of-the-art algorithms for network learning.

To estimate a level of confidence on the probabilistic relations between variables, the BN analysis was iterated on a set of bootstrap samples of the original data and then the results were combined in an ensemble of BN. In detail, the resulting DAGs were converted into the corresponding partially directed acyclic graph (PDAG) representing the corresponding I-equivalent classes and aggregated in a Weighted Partially Directed Acyclic Graph (WPDAG). WPDAGs encode the confidence on the presence of each edge as the fraction of bootstrap samples with that edge.

7.1 Diabetes BN model

7.1.1 Data selection and preprocessing

The diabetes BN model was developed using the data of subjects selected in Section 2.1. In particular, the variables used by the state-of-the-art models and the new variables selected in Section 6.1 were considered as candidate predictive variables of diabetes onset. First, continuous variables were discretized according either to their distribution percentiles or to thresholds adopted from the state-of-the art diabetes models or clinical practice. For instance, age was discretized in 3 levels based on its distribution: younger than 55, between 55 and 65, and older than 65. On the contrary, fasting glucose was discretized according to the thresholds given by the American Diabetes Association: lower than 100 mg/dl, between 100 and 125 mg/dl, and higher than 125 mg/dl. Discrete variables with a strong imbalance in the number of subjects over its possible values were either filtered out or re-discretized (reducing number of levels), to avoid a strong imbalance in the number of parameters associated to specific variable with respect to other variables. For instance, the number of family members sustained by family income was originally associated to 6 levels and was re-discretized into 3 levels: one, two, more than two dependents.

Furthermore, the dataset was subsampled in order to accommodate the outcome of BN model, which is the probability of developing diabetes within a specific lapse of time since the first exam. Thus, BN

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

required to be trained on both subjects with and without diabetes outcome at a specific time. To maximize the number of diabetic subjects considered and to balance as much as possible the number of diabetic and non-diabetic subjects, a 10-year horizon was chosen as model outcome. In detail, the dataset was reduced to 633 subjects that developed diabetes and 1,415 without diabetes outcome within 10 years since the first exam. Subjects that did not develop diabetes within this lapse of time were excluded because they could have developed diabetes before 10 years but this information was not available. At this point, a check was performed on the reduced dataset, in order to equally distribute the number of samples among the discretization levels of each variable. Some variables were thus removed, such as the one accounting for violence in the neighbourhood, while some other were aggregated, such as steroidal and non-steroidal anti-inflammatory medications considered as one variable.

The dataset used for BN training comprised 2,048 subjects over 40 variables. The entire set of variables and their respective discretization levels are reported in Table 38.

Table 38. Variables included in the dataset used for BN training, with description and discretization levels.

Variable	Description	Levels/categories
ethnicity	ethnicity	White, Caucasian Chinese American Black, African-American Hispanic
gender	gender	female male
Immigrant	immigrant status	no, born in the U.S. yes, born in another country
marital_status	marital status	married/living as married widowed/divorced/separated never married
education	education	grade 11 or less completed high school/ged, or some college but no degree technical school certificate, associate degree or bachelor's degree graduate or professional school
nsidewalks_parks	lack of sidewalks or parks in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
nfshop	lack of adequate food shopping in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
ntraffic	heavy traffic or speeding cars in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
nnoise	excessive noise in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
fam_hx_diab	family history of diabetes	no yes
hx_diab1	history of high blood sugar or diabetes	no yes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Description	Levels/categories
ever_aspirin_regularuse1	ever used aspirin regularly	no yes
age1	age [years]	<55 55-65 >65
bmi1	body mass index [kg/m^2]	<25 25 - 29.99 >=30
waist1	waist circumference [cm]	<80 from 80 to <88 from 88 to <94 from 94 to <102 >=102
smoking1	smoking status	never former current
alcohol_drinking1	alcohol drinking status	never moderate frequent
heart_rate1	heart rate [beats/min]	<60 60 -75 >75
systolic_bp1	systolic blood pressure [mmHg]	<130 130-139 >=139
diastolic_bp1	diastolic blood pressure [mmHg]	<85 85-89 >89
htn_med1	use of anti-hypertensive medication	no yes
ldl1	LDL cholesterol [mg/dl]	<130 130 -159 >159
hdl1	HDL cholesterol [mg/dl]	<40 40-59 >59
tot_chol1	Total cholesterol [mg/dl]	<200 200-239 >239
trig1	Triglycerides [mg/dl]	<150 150-199 >199
lipid_med1	Use of lipid-lowering medication	no yes
thyroid_med1	Use of thyroid medication	no yes
depression1	Use of antidepressants or depression symptoms according to depression scale	no yes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Description	Levels/categories
anti_inflammatory1	Use of anti-inflammatory meds (steroidal or non-steroidal including cox 2 inhibitors)	<150 150-199 >199+D32:D43
antibiotics1	Treated with antibiotics in the past year	no yes
curr_job1	Current occupation	homemaker employed unemployed or retired
income1	Total gross family income in the past 12 months	< $30,000 $30,000-74,999 >= $75,000
num_dependents1	Number of family members sustained by family income (including the respondent)	1 2 >2
fin_strain1	Ongoing financial strain	no yes
anger_scale1	Spielberg trait anger scale	10-14 15-21 22-40
anxiety_scale1	Spielberg trait anxiety scale	0 -13 14 -17 18 - 40
chronic_burden1	Chronic burden scale (indicator of chronic stress)	0 1 2-5
mod_vig_pa1	Moderate and vigorous physical activity [MET-min/week]	0 - 2698 2699 - 6165 >6165
gluc1	Fasting glucose [mg/dl]	<100 101-125 >=126
diab_10y	Diagnosis of diabetes within 10 years since exam 1	no yes

7.1.2 Method for BN training

BN training was performed through Hill-Climbing algorithm (structure learning) and a Maximum a Posteriori estimation (parameters learning); sense constraints was also applied to the network structure to codify the domain knowledge. For example, clinically or biologically non-sense relations among variables were forbidden, such as the dependence of ethnicity from the use of anti-inflammatory medications. To this purpose, variables were divided into four layers (Table 39) where variables in layer j could be dependent only on variables in layers i ≤ j. In particular, outcome variable (diabetes outcome at 10 years since the first exam) could be dependent on all the other variables of the datasets.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Table 39. Layering of variables in Bayesian Network

Layer	Variables
1. Unpredictable variables	Ethnicity, gender, immigrant ,marital_status ,education, nsidewalks_parks, nfshop, ntraffic, nnoise, fam_hx_diab, age1, curr_job1, income1, num_dependents1, fin_strain1
2. Habits	smoking1, alcohol_drinking1, mod_vig_pa1
3. Phenotypic/Metabolic variables	hx_diab1, ever_aspirin_regularuse1, bmi1, waist1, heart_rate1, systolic_bp1, diastolic_bp1, htn_med1, ldl1, hdl1, tot_chol1, trig1, lipid_med1, thyroid_med1, depression1, anti_inflammatory1, antibiotics1, anger_scale1, anxiety_scale1, chronic_burden1,gluc1
4. Outcome	diab_10y

BN training was run first on the entire dataset, then 500 Bayesian Networks were inferred from 500 bootstrap samples of the original dataset, in order to estimate a level of confidence on the edges among variables. The 500 BNs were merged into a WPDAG registering, for each pair of variables, the number of BNs with an edge between them; the presence of that edge among all the BNs encoded its confidence.

7.1.3 Results

The DAG resulting from the BN learning on the entire dataset comprised 40 nodes and 107 edges (Figure 17). The BN model identified some expected dependencies between variables, such as the influence of fasting glucose level and family history of diabetes on the probability of developing diabetes within 10 years. Besides, the model confirmed the relationship between diabetes and waist circumference, since it is known that being overweight or obese is a risk factor for T2D. Notably, the antibiotics were found to influence the probability of diabetes outcome as already reported in [25].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

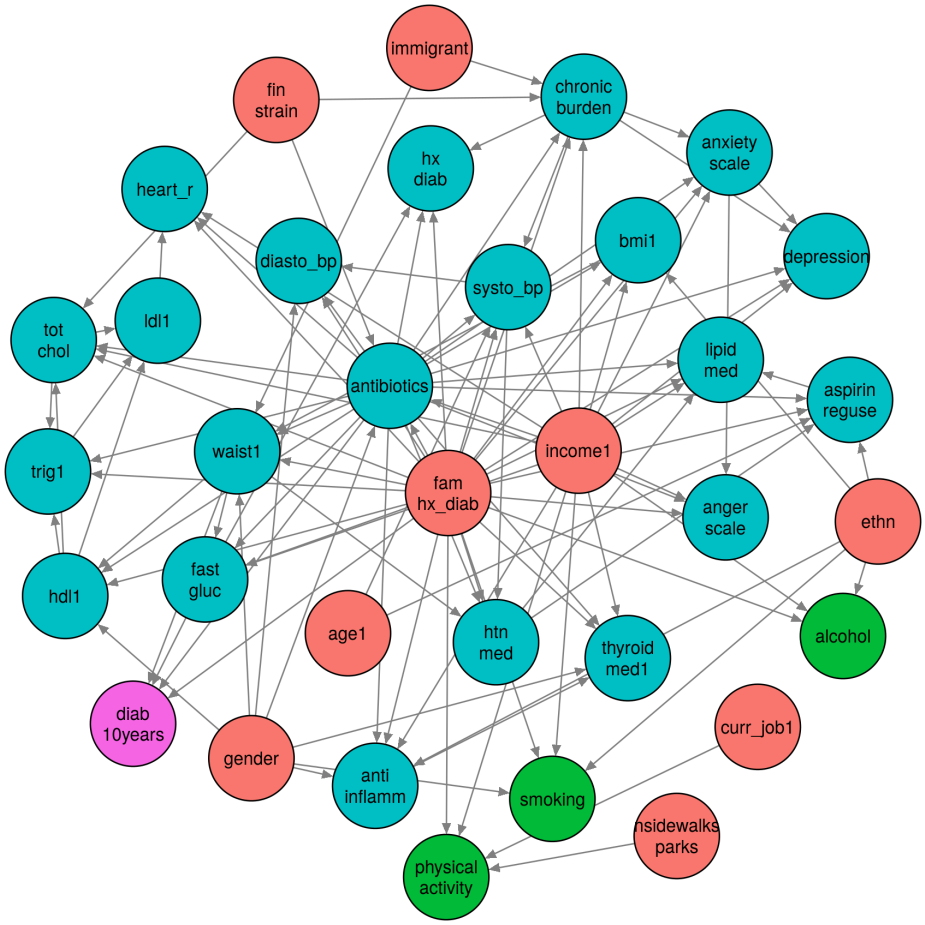

Figure 17. Subset of the DAG obtained on the entire training dataset. Only nodes (34) with at least one direct edge are shown. Red: unpredictable variables (layer 1); green: habits (layer 2); cyan: phenotypic/metabolic variable (layer 3); purple: outcome (layer 4).

Besides, Figure 18 reports the WPDAG resulting from the 500 BNs learnt on 500 bootstrap samples of the training dataset; the edge thickness is proportional to the confidence of that specific relation between variables. The model confirmed that diabetes is highly dependent on fasting glucose level and family history of diabetes, with these relations found in 100% and 83% of all the 500 DAGs, respectively. Furthermore, diabetes outcome was influenced by antibiotics, family income and BMI in 99%, 48% and 25% of the DAGs, respectively. Education was the only node with edges in no more than 10% of DAGs, thus this variable should not be considered in further BN models.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

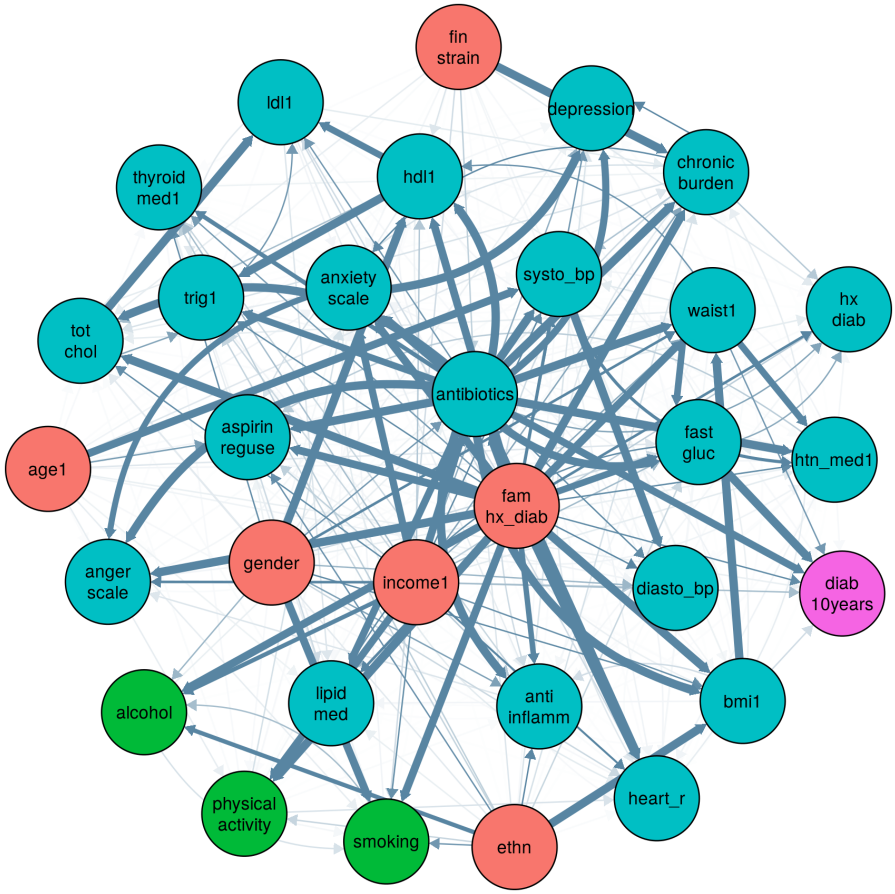

Figure 18. Subset of the WPDAG obtained on the 500 bootstrap samples of the entire training dataset. Edge thickness is proportional to the number of times that edge is observed in the 500 DAGs. Red: unpredictable variables (layer 1); green: habits (layer 2); cyan: phenotypic/metabolic variable (layer 3); purple: outcome (layer 4).

7.2 Asthma BN model

7.2.1 Data selection and preprocessing

The BN model for asthma was developed using the data of subjects selected in Section 2.2. In particular, the variables used by the state-of-the-art variables and the new variables selected in Section 6.2 were considered as candidate predictive variables of diabetes onset. First, continuous variables underwent a discretization process, based on thresholds either computed on their distribution percentiles or taken from the most recent literature or clinical practice. For instance, physical activity was discretized in 3 levels according to its distribution: less than 2,698 MET-min/week, between 2,698 and 6,165 MET-min/week, and more than 6,165 MET-min/week. The number of levels of some discrete variables were reduced to homogenize the number of subjects over variable levels.

Subsequently, the dataset was subsampled in order to accommodate the outcome of BN model, which is the probability of developing asthma within a specific lapse of time. Since BN training required subjects both with and without asthma at a specific time, a 11-year horizon was adopted as model outcome; consequently, the number of asthmatic subjects considered was maximized and the ratio asthmatic/non-asthmatic subjects was balanced. In detail, the dataset was reduced to 136 subjects that developed asthma and 163 without asthma outcome within 11 years since the first exam. Subjects that did not develop asthma within this lapse of time were excluded because they could have become asthmatic before 11 years but this information was not available. Consequently, for each variable, the number of subjects within the discretization levels was homogenized as much as possible. Some

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

variables were thus filtered out, such as the one accounting for trash problem in the neighbourhood, while some other were aggregated, such as steroidal and non-steroidal anti-inflammatory medications considered as one variable.

The dataset used for BN training included 299 subjects over 32 variables. The entire set of variables and their respective discretization levels are reported in Table 40.

Table 40. Variables included in the dataset used for BN training, with description and discretization levels.

Variable	Description	Levels/categories
ethnicity	ethnicity	White, Caucasian Chinese American Black, African-American Hispanic
gender	gender	female male
Immigrant	immigrant status	no, born in the U.S. yes, born in another country
marital_status	marital status	married/living as married widowed/divorced/separated never married
education	education	grade 11 or less completed high school/ged, or some college but no degree technical school certificate, associate degree or bachelor's degree graduate or professional school
nsidewalks_parks	lack of sidewalks or parks in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
nfshop	lack of adequate food shopping in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
ntraffic	heavy traffic or speeding cars in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
nnoise	excessive noise in neighbourhood	very serious/somewhat serious problem minor problem not really a problem
fam_hx_asthma	family history of asthma	no yes
ever_aspirin_regularuse1	ever used aspirin regularly	no yes
age1	age [years]	<55 55-65 >65
bmi1	body mass index [kg/m^2]	<25 25 - 29.99 >=30
waist1	waist circumference [cm]	<80 from 80 to <88

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Description	Levels/categories
		from 88 to <94 from 94 to <102 >=102
smoking1	smoking status	never yes (former or current)
alcohol_drinking1	alcohol drinking status	never moderate frequent
heart_rate1	heart rate [beats/min]	<60 60 -75 >75
depression1	Use of antidepressants or depression symptoms according to depression scale	no yes
anti_inflammatory1	Use of anti-inflammatory meds (steroidal or non-steroidal including cox 2 inhibitors)	<150 150-199 >199+D32:D43
antibiotics1	Treated with antibiotics in the past year	no yes
curr_job1	Current occupation	homemaker employed unemployed or retired
income1	Total gross family income in the past 12 months	< $30,000 $30,000-74,999 >= $75,000
num_dependents1	Number of family members sustained by family income (including the respondent)	1 2 >2
fin_strain1	Ongoing financial strain	no yes
anger_scale1	Spielberg trait anger scale	10-14 15-21 22-40
anxiety_scale1	Spielberg trait anxiety scale	0 -13 14 -17 18 - 40
chronic_burden1	Chronic burden scale (indicator of chronic stress)	0 1 2-5
mod_vig_pa1	Moderate and vigorous physical activity [MET-min/week]	0 - 2698 2699 - 6165 >6165
sndh_smoke1	Second hand smoke [hours/week]	no yes
two_pillow1	Sleep with two or more pillows to help breathe	no yes
wake_breath1	Awakened at night for trouble breathing	no yes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

Variable	Description	Levels/categories
asthma_11y	Diagnosis of diabetes within 11 years since exam 1	no yes

7.2.2 Method for BN training

The Bayesian Network on asthma dataset was learnt through Hill-Climbing algorithm (structure learning) and a Maximum a Posteriori estimation (parameters learning); constraints was also applied to the network structure in order to forbid clinically or biologically non-sense relations among variables. For instance, the influence of antibiotics medication on gender was forbidden. Hence, variables were grouped into five layers (Table 41) where variables in layer j could be dependent only on variables in layers i ≤ j. In particular, outcome variable (asthma outcome at 11 years since the first exam) could be dependent on all the other variables of the datasets.

Table 41. Layering of variables in Bayesian Network

Layer	Variables
1. Unpredictable variables	ethnicity, gender, immigrant ,marital_status ,education, nsidewalks_parks, nfshop, ntraffic, nnoise, fam_hx_asthma, age1, curr_job1, income1, num_dependents1, fin_strain1
2. Habits	smoking1, alcohol_drinking1, mod_vig_pa1, sndh_smoke
3. Phenotypic/Metabolic variables	ever_aspirin_regularuse1, bmi1, waist1, heart_rate1, depression1, anti_inflammatory1, antibiotics1, anger_scale1, anxiety_scale1, chronic_burden1
4. Derived variables	two_pillow1, wake_breath1
5. Outcome	asthma_11y

First, BN was inferred from the entire dataset, secondly 500 Bayesian Networks were trained on 500 bootstrap samples of the original dataset, in order to estimate a level of confidence on the edges among variables. The 500 BNs were aggregated into a WPDAG registering, for each pair of variables, the number of BNs with an edge between them; the presence of that edge among all the BNs encoded its confidence.

7.2.3 Results

The BN learning on the entire dataset resulted into a DAG with 32 nodes and 55 edges (Figure 19).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

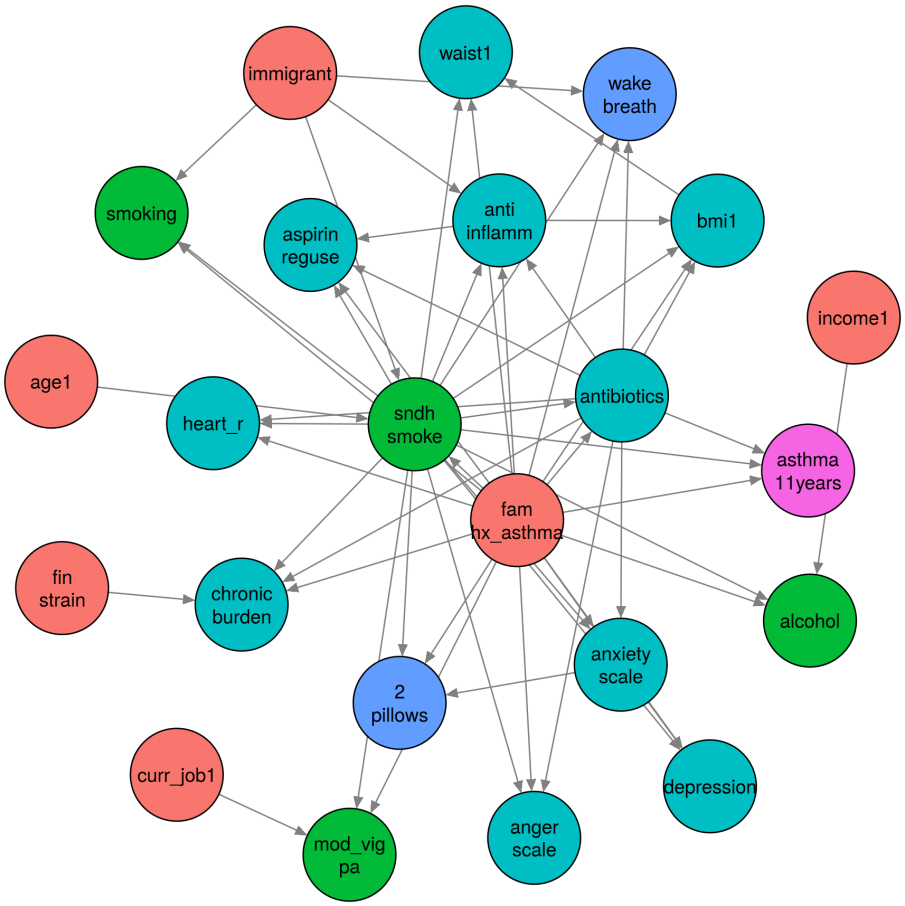

Figure 19. Subset of the DAG obtained on the entire training dataset. Only nodes (23) with at least one direct edge are shown. Red: unpredictable variables (layer 1); green: habits (layer 2); cyan: phenotypic/metabolic variable (layer 3); blue: derived variables (layer 4); purple: outcome (layer 5).

The BN model revealed some new dependencies between variables that were not considered by the state-of-the-art models for asthma: the influence of second hand smoke, antibiotics and family history of asthma on the probability of developing asthma within 11 years.

Besides, the WPDAG resulting from the 500 BNs learnt on 500 bootstrap samples of the training dataset is reported in Figure 20; the edge thickness is proportional to the number of times that edge is observed in the 500 DAGs. Our model confirmed that asthma is highly dependent on family history of asthma, second-hand smoke and antibiotics with these relations found in 100%, 99% and 91% of all the 500 DAGs, respectively. Four variables showed edges in no more than 10% of DAGs: education, marital status, lack of food shopping and excessive noise in neighbourhood. These variables will probably not considered in further BN models, given their low influence over other variables.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

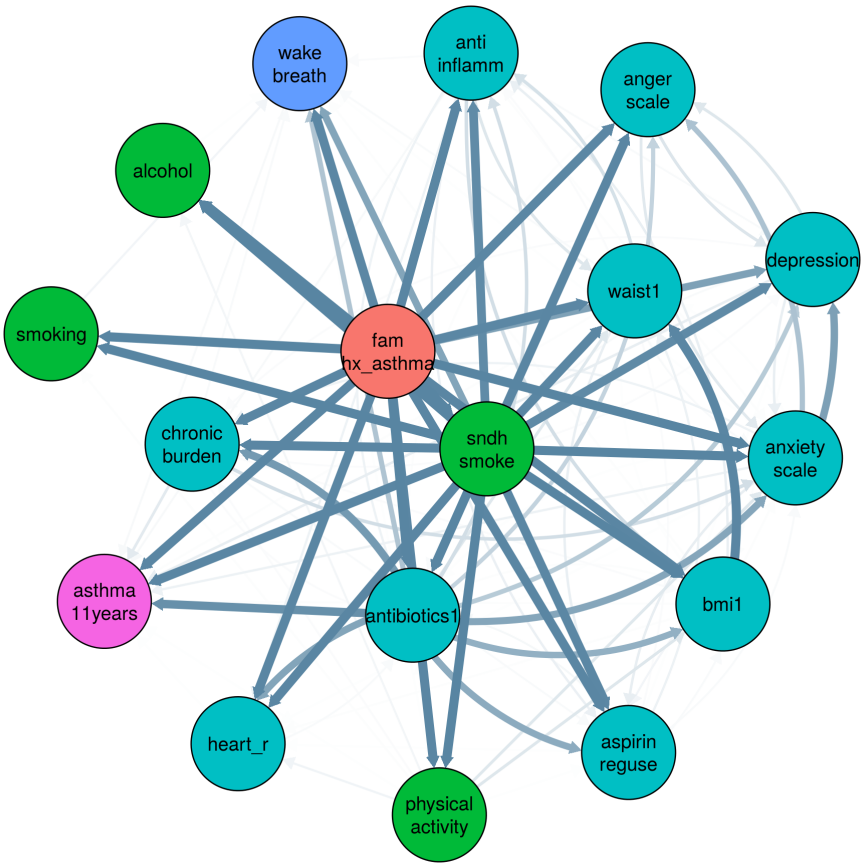

Figure 20. Subset of the WPDAG obtained on the 500 bootstrap samples of the entire training dataset. Edge thickness is proportional to the number of times that edge is observed in the 500 DAGs. Red: unpredictable variables (layer 1); green: habits (layer 2); cyan: phenotypic/metabolic variable (layer 3); blue: derived variables (layer 4); purple: outcome (layer 5).

8 INTEGRATION OF THE MODELS IN THE PULSE APP

Many predictive models of diabetes onset were proposed in the literature based on simple logistic regression analysis, proportional hazard models or accelerated failure time models. The variables required by the state-of-the-art models were all included in the first version of PulsAir questionnaires, thus the state-of-the-art models of diabetes could be integrated in the PulsAir app as risk calculators to provide feedback to all the users. In addition, for the users having the Fitbit, the variables related to physical activity and heart rate can be derived from the Fitbit API with suitable measurement unit conversion. Specifically, as in MESA physical activity was measured in MET-min/week, which is an indicator of amount of energy expenditure per week, an equivalent information can be extracted from the Fitbit tracking of calories and minutes of activity per week.

The assessment of state-of-the-art models for the prediction of diabetes onset (Section 4) showed that the recalibration on a different population does not impact significantly the discriminatory ability of the models. In other words, when the state-of-the-art models are applied to a new population their performance in correctly ranking the subjects according to diabetes risk is as good as the one of a model with equal structure, but whose parameters are estimated in the new population. As a consequence, the state-of-the-art models of diabetes can be used to assess the diabetes risk in the population enrolled in the PULSE pilots without a significant deterioration of their ranking ability.

Nevertheless, our analysis showed that when the models are applied to a different population (e.g. in the MESA population in this deliverable) their performance in terms of calibration, i.e., the ability to correctly predict the observed probability of diabetes incidence, may not be satisfactory, unless a recalibration of the model is performed. Unfortunately, state-of-the-art models cannot be recalibrated

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

on the data collected in the PULSE pilots, because developing new models or recalibrating existing ones require longitudinal datasets in which a healthy population is followed up for several years (e.g. 5-10 years) in order to observe a sufficient number of new cases of diabetes. To overcome this limitation, we are now validating a consensus model for application in PULSE that calculates the scores using different literature models, it interprets them as a relative risk rather than an absolute risk and, finally, it combines them in an aggregated risk score. In other words, the risk value returned by each model on a certain individual must be compared to the risk value of the other subjects in the pilot in order to really understand if the individual can be classified as “at risk” or not.

Concerning the prediction of asthma onset, only few models were proposed in the literature to predict the adult-onset of asthma and none of those was validated. In this deliverable, we recalibrated two literature models and tested them on the MESA dataset (Section 5). However, the achieved performance was not satisfactory, although our results might be affected by the low incidence of asthma in the datasets used for model recalibration. For this reason, we are now developing new models including new variables (see below) to understand if new informative data might improve the model performances. In the meantime, we recommend 1) collecting data related to new variables and 2) designing a feedback strategy to provide the PULSE users with simple recommendations to address potential risk factors of asthma, which are known from literature studies.

In this deliverable, we also identified new variables potentially predictive for diabetes and asthma onset based on the evidences published in the literature (Section 6). Some of them have been already included in the PulsAir questionnaires, e.g., depression symptoms, socio-economic indicators, neighbourhood characteristics and alcohol consumption. The probabilistic relationships between the candidate predictive variables and the onset of diabetes or asthma were assessed using static BNs (Section 7). Static BNs are good descriptive models that provide useful insight into the relationships between variables. Regarding asthma onset, in particular, the BN analysis evidenced that some of the new candidate variables have a direct effect on asthma adult-onset, e.g. exposure to second hand smoke and family history of asthma. Therefore, questions related to these variables were added in the last version of the PulsAir questionnaires.

Next steps will include the development of new predictive models of diabetes and asthma onset using dynamic Bayesian networks and survival analysis (task 5.4). The new models will be integrated in the PulsAir app and/or the PULSE dashboard. The new models will incorporate both variables used by the state-of-the-art models and new variables identified in this deliverable. If needed, we will recommend further updates of the PULSE questionnaires based on our future analysis with the new models.

Finally, as possible strategies to take advantage of the data collected by the PULSE system for the development of new health risk models, we consider that data collected within the PULSE project, especially the environmental data, would have a great value for modelling the risk of asthma attacks. Even if not originally present in WP5 tasks, the prediction of asthma attacks might be an outcome of high interest to study within the PULSE project. Therefore, the inclusion of a questionnaire to collect information related to asthma attacks in the PulsAir app would be of great value.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

9 REFERENCES

[1] Bild D.E., Bluemke D.A., Burke G.L., Detrano R., Diez Roux A.V., Folsom A.R., Greenland P., Jacob D.R. Jr, Kronmal R., Liu K., Nelson J.C., O'Leary D., Saad M.F., Shea S., Szklo M., Tracy R.P. Multi-Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol. 2002 Nov 1; 156(9):871-81.

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

[3] Park S.H., Goo, J.M., Jo C.-H. Receiver Operating Characteristic (ROC) curve: Practical review for radiologists. Korean J Radiol. 2004 Jan-Mar; 5(1):11–18.

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

[5] Rockhill B.,Spiegelman D.,Byrne C., Hunter D.J., Colditz G.A. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst. 2001 Mar 7; 93(5):358-66.

[6] 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.

[7] Di Camillo B., Hakaste L., Sambo F., Gabriel R., Kravic J., Isomaa B., Tuomilehto J., Alonso M., Longato E., Facchinetti A., Groop L.C., Cobelli C., Tuomi T. “HAPT2D: High accuracy of prediction of T2D with a model combining basic and advanced data depending on availability.” Eur J Endocrinol, 2018 Jan; 178(4):331-341.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Cohen R., Gasca N.C., McClelland R.L., Alcántara C., Jacobs D.R., Roux A.D., Rozanski A. and Shea S. Effect of Physical Activity on the Relation between Psychosocial Factors and Cardiovascular Events (From the Multi-Ethnic Study of Atherosclerosis [MESA]). Am J Cardiol. 2016 May 15; 117(10): 1545–1551.

[16] 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.

[17] Robbins J.M., Vaccarino V., Zhang H., et al. Socioeconomic status and diagnosed diabetes incidence. Diabetes Res Clin Pract. 2005; 68(3): 230-236.

[18] Rabi DM, Edwards AL, Southern DA, et al. Association of socio-economic status with diabetes prevalence and utilization of diabetes care services. BMC Health Services Research. 2006;6:124. doi:10.1186/1472-6963-6-124.

[19] Cornelis MC, Chiuve SE, Glymour MM, et al. Bachelors, Divorcees, and Widowers: Does Marriage Protect Men from Type 2 Diabetes? Sen U, ed. PLoS ONE. 2014;9(9):e106720.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA727816.

[20] Krishnan S., Cozier Y.C., Rosenberg L., Palmer J.R. Socioeconomic Status and Incidence of Type 2 Diabetes: Results From the Black Women's Health Study. Am J Epidemiol. 2010 Mar 1;171(5):564-70.
[21] Christine P.J., Auchincloss A.H., Bertoni A.G., Carnethon M.R., Sánchez B.N., Moore K., Adar S.D., Horwich T.B., Watson K.E., Diez Roux A.V. Longitudinal Associations Between Neighborhood Physical and Social Environments and Incident Type 2 Diabetes Mellitus: The Multi-Ethnic Study of Atherosclerosis (MESA). JAMA Intern Med. 2015 Aug; 175(8):1311-20.
[22] Engum A. The role of depression and anxiety in onset of diabetes in a large population-based study. J Psychosom Res. 2007; 62(1):31–38.
[23] Schmitz N., Deschênes S.S., Burns R.J., Smith K.J., Lesage A., Strychar I., Rabasa-Lhoret R., Freitas C., Graham E., Awadalla P., Wang J.L. Depression and risk of type 2 diabetes: the potential role of metabolic factors. Mol Psychiatry. 2016 Dec; 21(12): 1726-1732.
[24] Abraham S., Shah N.G., Diez Roux A., Hill-Briggs F., Seeman T., Szklo M., Schreiner P.J., Golden S.H. Trait anger but not anxiety predicts incident type 2 diabetes: The Multi-Ethnic Study of Atherosclerosis (MESA). Psychoneuroendocrinology. 2015 Oct; 60:105-13.
[25] Mikkelsen K.H., Knop F.K., Frost M., Hallas J., Pottegård A. Use of Antibiotics and Risk of Type 2 Diabetes: A Population-Based Case-Control Study. J Clin Endocrinol Metab. 2015; 100(10):3633-3640.
[26] Micossi P., Pontiroli A.E., Baron S.H., Tamayo R.C., Lengel F., Bevilacqua M., Raggi U., Norbiato G., Foà P.P. Aspirin stimulates insulin and glucagon secretion and increases glucose tolerance in normal and diabetic subjects. Diabetes. 1978 Dec; 27(12):1196-1204.
[27] Hage M, Zantout MS, Azar ST. Thyroid Disorders and Diabetes Mellitus. J Thyroid Res. 2011; 2011: 439463.
[28] Polsky S., Akturk H.K. Alcohol Consumption, Diabetes Risk, and Cardiovascular Disease Within Diabetes. Curr Diab Rep. 2017 Nov 4;17(12):136.
[29] Ekerljung L., Sundblad B.M., Rönmark E., Larsson K., Lundbäck B. Incidence and prevalence of adult asthma is associated with low socio-economic status. Clin Respir J. 2010; 4(3):147-156.
[30] Ellison-Loschmann L., Sunyer J., Plana E., et al. European Community Respiratory Health Survey. Socioeconomic status, asthma and chronic bronchitis in a large community-based study. Eur Respir J. 2007;29(5):897-905.
[31] Hedlund U., Eriksson K., Rönmark E. Socio-economic status is related to incidence of asthma and respiratory symptoms in adults. Eur Respir J. 2006; 28(2):303-310.
[32] Simons E., Dell S.D., Moineddin R., To T. Associations between Neighborhood Walkability and Incident and Ongoing Asthma in Children. Ann Am Thorac Soc. 2018 Apr 17. doi: 10.1513/AnnalsATS.201708-693OC. [Epub ahead of print].
[33] Wickens K., Barry D., Friezema A., Rhodius R., Bone N., Purdie G., Crane J. Fast foods - are they a risk factor for asthma? Allergy. 2005 Dec; 60(12):1537-41.
[34] Lehrer P. Anger, stress, dysregulation produces wear and tear on the lung. *Thorax*. 2006;61(10):833-834.
[35] Brumpton B.M., Leivseth L., Romundstad P.R., Langhammer A., Chen Y., Camargo C.A., Mai X.M. The joint association of anxiety, depression and obesity with incident asthma in adults: the HUNT study. Int J Epidemiol. 2013 Oct; 42(5):1455-63.
[36] Subbarao P., Mandhane P.J., Sears M.R. Asthma: epidemiology, etiology and risk factors. CMAJ : Canadian Medical Association Journal. 2009; 181(9):E181-E190.
[37] Thomsen S.F., Kyvik K.O., Skadhauge L.R., Steffensen I., Backer V. Regular use of non-steroidal anti-inflammatory drugs increases the risk of adult-onset asthma: a population-based follow-up study. Clin Respir J. 2009 Apr;3(2):82-4.
[38] Hamad A.M., Sutcliffe A.M., Knox A.J. Aspirin-induced asthma: clinical aspects, pathogenesis and management. Drugs. 2004; 64(21):2417-32.
[39] Merianos A.L., Jandarov R.A., Mahabee-Gittens E.M. Association of Secondhand Smoke Exposure with Asthma Symptoms, Medication Use, and Healthcare Utilization among Asthmatic Adolescents. J Asthma. 2018 Apr 11:1-30.
[40] Lajunen T.K., Jaakkola J.J., Jaakkola M.S. The synergistic effect of heredity and exposure to second-hand smoke on adult-onset asthma. Am J Respir Crit Care Med. 2013 Oct 1;188(7):776-82.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.

[41] Lieberoth S., Backer V., Kyvik K.O., Skadhauge L.R., Tolstrup J.S., Grønbæk M., Linneberg A., Thomsen S.F. Intake of alcohol and risk of adult-onset asthma. Respir Med. 2012 Feb; 106(2):184-8.
[42] Jensen, Finn V. An introduction to Bayesian networks. Vol. 210. London: UCL press, 1996.
[43] Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.
[44] Tsamardinos I, Brown L, Aliferis C. The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning. 2006;65(1):31-78.
[45] Franzin A, Sambo F, Di Camillo B. 2016. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. Bioinformatics. 33.8: 1250-1252.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA727816.