PULSE

Participatory Urban Living for Sustainable Environments

D5.5 Model validation on the available (retrospective) datasets

PULSE project

H2020 - 727816

University of Padova

February 2019

DOCUMENT INFO

0.1 AUTHORS

Author	Organization	e-mail
Martina Vettoretti	University of Padova	martina.vettoretti@dei.unipd.it
Alessandro Zandonà	University of Padova	alessandro.zandona@dei.unipd.it
Barbara Di Camillo	University of Padova	barbara.dicamillo@dei.unipd.it

0.2 DOCUMENT HISTORY

Date	Version	Editor	Change	Status
06/02/2019	0	UNIPD	First template	Draft
12/02/2019	1	UNIPD	Draft with partial contents	Draft
18/02/2019	2.0	UNIPD	First complete draft	Draft
19/02/2019	2.1	UNIPD	Final version	Final

0.3 DOCUMENT KEYDATA

Key words	H2020 – 727816 – PULSE Deliverable 5.5	
Editor info	Name	Barbara Di Camillo
	Organization	UNIPD
	e-mail	barbara.dicamillo@dei.unipd.it

0.4 DISTRIBUTION LIST

Date	Issue	Distribution list
18/02/2019	Circulate first complete draft for internal revision	Maria Fernanda Cabrera, Riccardo Bellazzi,
19/02/2019	Final revised version	Maria Fernanda Cabrera, Riccardo Bellazzi,
19/02/2019	Final version	All Consortium and the European Commission

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The main purpose of this deliverable, entitled “Model validation on the available (retrospective) datasets”, is to validate the new predictive models of type 2 diabetes (T2D) and asthma onset developed in work package 5. The new models include: a consensus model of T2D onset; COX-LASSO models and survival SVM models for the prediction of both T2D and asthma; a Dynamic Bayesian Network (DBN) describing the trajectories of T2D risk factors over time. The validation of the models on external test sets derived from MESA data has been already reported in deliverable D5.4, where the new models were all validated on previously unseen examples neither used from training or parameter settings. Here, thanks to the availability of another dataset, i.e. the English Longitudinal Study of Ageing (ELSA), some models are validated in a second cohort. In addition, the ELSA dataset is used to develop and validate new predictive models of asthma onset that exploit new variables available in the ELSA dataset, which were not available in the MESA dataset.

1 INTRODUCTION

In the previous task 5.4, we developed new predictive models of type 2 diabetes (T2D) and asthma onset using a variety of statistical learning and machine learning techniques. The models were trained with the data collected in the Multi-Ethnic Study of Atherosclerosis (MESA), a longitudinal study of cardiovascular disease conducted in United States. In particular, we developed: a consensus model for the prediction of T2D onset; different survival models based on Cox-LASSO and survival SVM for the prediction of both T2D and asthma; a Dynamic Bayesian Network (DBN) that models the probabilistic relationships between T2D risk factors over time.

The diabetes consensus model was defined as the weighted average of the risk scores provided by several literature predictive models of T2D onset after partial recalibration is applied. Interestingly, the consensus model can overcome some of the barriers that currently limit the adoption of T2D predictive models in clinical practice, such as the recalibration issue, the problem of missing values, and the limitations imposed by each model's domain of validity (certain models can be applied only to certain ethnic groups) [1].

The new survival models of disease onset included environmental variables (as the ones collected within PULSE) and assessed the ability of different variables to improve predictions of developing T2D or asthma. In particular, Support Vector Machines for survival data (survSVM) [11] and Cox survival analysis model with LASSO (CSA) [12] were embedded on a recursive feature elimination schema [14][15] to select the optimal number of predictive variables and rank them based on their ability to improve prediction performance on previously unseen data.

Finally, the DBN model [13] was trained to detect the biomarkers related to diabetes, and to identify how they influence each other over time in terms of conditional dependencies. Interestingly, the DBN model, besides providing a quantification of the risk of developing T2D, also provides an insight on the most probable health condition trajectories over time driving to T2D onset.

The purpose of this new deliverable is to validate the developed predictive models presented in D5.4. Model validation is a mandatory step before applying the models and consists in testing the ability of a model to predict the outcome on previously unseen data, i.e. data not used to train the model. For this purpose, we used a first test sets extracted from the MESA dataset, already presented in the previous deliverables, and a second test set from the English Longitudinal Study of Ageing (ELSA).

Validation on the MESA test set

Model validation is usually performed in the literature by testing the model on an independent sample of previously unseen subjects, extracted from the same dataset used for training the model, so to have the same input variables of the training. This approach has been implemented for validating the models developed in task 5.4; we split the MESA dataset in two independent training and test sets at the purpose of training the models and validating them, respectively. In deliverable D5.4 we had anticipated the results of the validation steps for all the developed models. We report them here briefly for completeness (Section 2).

Validation on the ELSA test set

We used a second dataset, i.e. the ELSA dataset (described in Section 3), in order to:

1) Perform a second, more challenging assessment of the consensus model generalizability (Section 4), testing the model performance on a completely different dataset, which is different from the MESA dataset, in terms population enrolled, variable definition and data collection procedures. Since the consensus model is thought to be used with different population as input, we think this is the proper framework to validate it.
2) Train and test new survSVM and CSA models on the ELSA dataset (Section 5) for two main purposes: i) to validate the variable ranking methodology, ii) to derive new models that can be jointly used with those derived on the MESA dataset. Note that a validation of the survival models (survSVM and CSA) developed using MESA data is not possible with the ELSA dataset, because not all the variables required by these models were collected within ELSA.

2 INTERNAL VALIDATION OF MODELS DEVELOPED IN TASK 5.4

In deliverable D5.4, the new predictive models of T2D and asthma onset were validated using an independent test set extracted from the MESA dataset. Indeed, after suitable pre-processing, the MESA dataset was divided into two independent sets: a training set (4.124 subjects for diabetes, 4.273 subjects for asthma) that was used for model development and a test set (1.031 subjects for diabetes, 1.068 subjects for asthma) that was used to test the models on previously unseen data.

Results of diabetes consensus model on the test set showed that, in terms of discriminatory ability (C-index and AUC), the diabetes consensus model was able to achieve performance comparable to those of the models of scenario 3 (which are the literature models with best performance) and much better than those of scenarios 1 and 2 (i.e. the models not using variables invasively collected). The diabetes consensus model resulted also well calibrated in the MESA population, with E/O equal to 0.82 on the validation set. Interestingly, the diabetes consensus model outperformed the models of scenarios 2 and 3 in terms of number of subjects for which a prediction is possible. In fact, real datasets are characterized by many subjects with missing values. For this reason, applying the literature models is not always possible due to the lack of input data. While the diabetes consensus model presented no missing model predictions, the models of scenarios 2 and 3 had a significant percentage of missing model predictions, which ranged between 17% (Framingham) and 46% (model by Kahn et al.) on the test set.

Regarding the new survival models, the internal validation of the survSVM and CSA models on the MESA test set showed that the new models presented better discriminatory ability than the state-of-the-art models tested in deliverable D5.3. In particular, predictive models of T2D onset showed very good performance in the test set, with C-index for the survSVM and the CSA models equal to 0.82 and 0.88 in scenario 3, 0.74 and 0.84 in scenario 2, 0.73 and 0.81 in scenario 3. The performance of asthma predictive models, even though better that the literature models, were less satisfactory (C-index lower than 0.70 for both survSVM and CSA model).

The DBN model was also validated on the MESA test set. Specifically, for each subject in the test set, the temporal evolution of its baseline variables was simulated by sampling the conditional probability distributions, learnt from the training set, at each time point in accordance with his/her state in the previous time point. For each subject, 100 different simulations were run. The probability of T2D onset over time was compared between real data (test set) and simulated data (predicted by DBN), showing that the DBN model provides a precise estimation of T2D onset probability over time. We also assessed the ability of the DBN to rank subjects according their risk of developing T2D at different time points obtaining very good results (AUC = 0.87, 0.79, 0.77, 0.75 at 24, 36, 48 and 60 months, respectively).

Further details on the validation of the new model performed with the MESA data are reported in deliverable D5.4.

3 ELSA DATASET

The ELSA dataset was collected in the English Longitudinal Study of Ageing, an ongoing study of health, social, wellbeing and economic circumstances in the English population aged 50 and older. Participants (drawn from the Health Survey for England) have a face-to-face interview every two years and a clinical examination, including blood test, every four years. Currently, the study includes eight waves of data collection covering a period of 15 years (Table 1). At waves 3, 4, 5, 6, and 7, new participants entered the study to maintain the size of the sample.

Table 1. ELSA waves

	Wave 1 2002-2003	Wave 2 2004-2005	Wave 3 2006-2007	Wave 4 2008-2009	Wave 5 2010-2011	Wave 6 2012-2013	Wave 7 2014-2015	Wave 8 2016-2017
Interview	X	X	X	X	X	X	X	X
Visit		X		X		X		X
Sample refreshment	X		X	X	X	X	X	

For PULSE purpose, since the clinical examinations were performed only in waves 2, 4, 6 and 8, we assigned to each subject a baseline wave among waves 2, 4 and 6 (not wave 8 because no follow-up would be available). Specifically, subjects that entered the study in wave 1 were assigned baseline wave 2 (N=9.432), subjects that entered in waves 3 and 4 were assigned to baseline wave 4 (N=4.357), and subjects recruited in waves 5 and 6 were assigned to baseline wave 6 (N=1.557). Then, for each subject, we selected the variables collected at the baseline wave, and the variables related to diabetes and asthma diagnosis at the following waves. Specifically, diabetes and asthma diagnoses were assessed during the interview of each wave by asking whether or not a doctor has told the respondent he/she has diabetes/asthma. The wave of diabetes/asthma diagnosis was defined as the first wave at which the subject reported being diagnosed by a doctor with diabetes/asthma.

Note that while the MESA dataset included subjects of four different ethnicities (white/Caucasian, black/African American, Chinese American and Hispanic), the ELSA dataset mainly includes Caucasian subjects (about 98% of subjects).

3.1 DATA SELECTED FOR T2D MODELS DEVELOPMENT AND VALIDATION

In order to develop and validate predictive models of T2D onset, we extracted from the ELSA dataset the subjects that at baseline wave (either 2, 4 or 6) were free of diabetes and had information on diabetes diagnosis in the subsequent waves. From this sample, we excluded the subjects that did not have the clinical examination. The remaining sample includes 9.641 subjects (6.304 with baseline wave 2, 2.615 with baseline wave 4 and 722 with baseline wave 6) of whom 747 developed diabetes during the observation period after the baseline.

Data were split into a training set containing 80% of subjects (N=7.688) and a test set with the remaining 20% of subjects (N=1.953). Diabetes incidence was similar in the training and the test set.

Variables selected for diabetes consensus model validation

From the baseline variables of each subject, we selected the variables required by the literature models incorporated in the diabetes consensus model. These were: age, gender, ethnicity, education level, immigrant status, family history of diabetes, father history diabetes, mother history of diabetes, smoking, height, weight, body mass index (BMI), waist circumference, heart rate, history of hypertension, use of hypertension medication, systolic blood pressure, diastolic blood pressure, history of diabetes, fasting glucose concentration, HDL cholesterol level, triglycerides level, history of heart disease.

Variables selected for T2D survival model development and validation

From the baseline visit of each subject in the selected sample, we selected the variables already considered for developing diabetes survival models with the MESA dataset. However, unfortunately, some of the variables that were available in the MESA dataset were not collected in ELSA (for this reason we could not directly apply the models developed with the MESA data to the ELSA data). In Table 2, we report the list of variables considered for diabetes survival model development in MESA and ELSA datasets. In particular, the variables available only in MESA are marked in red (in total 10 variables). The variables available in ELSA were pre-processed in order to obtain variable categories/levels as much as possible equal to those of the MESA variables. The levels/categories assumed by the variables in the two datasets are reported in the third and fourth column. Unfortunately, for some variables, like nnoise, ntraffic, mod_vig_pa1, and anxiety_scale1, it was not possible to obtain the same categories/levels of the MESA variables.

As in deliverable D5.4, the variables were grouped according to 3 different scenarios reflecting different degrees of information availability, in which to assess the performance of the developed survival models. Scenario 1 includes all the easily accessible variables that do not require particular measurements; Scenario 2 adds to scenario 1 non-invasive measurements of health parameters (e.g. blood pressure); Scenario 3 adds to scenario 2 invasive measurements of biomarkers (e.g. fasting glucose concentration). The scenario of each variable is specified in the fifth column of Table 2.

Table 2. Variables selected as candidate predictive variables for the survival models of diabetes onset.

Variable	Description	Levels/categories in MESA	Levels/categories in ELSA	Scenario
ethnicity	ethnicity	White, Caucasian Chinese American Black, African-American Hispanic	White Other	1
gender	gender	female male	Same in MESA	1
marital_status	marital status	married/living as married widowed/divorced/separated never married	Same in MESA	1
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	Lower than high school High-school graduate or some college College and above	1

Variable	Description	Levels/categories in MESA	Levels/categories in ELSA	Scenario
		graduate or professional school		
nparks	lack of parks in neighbourhood	very serious/somewhat serious problem minor problem not really a problem	N.a.	1
nsidewalks	lack of sidewalks in neighbourhood	very serious/somewhat serious problem minor problem not really a problem	N.a.	1
nfshop	lack of adequate food shopping in neighbourhood	very serious/somewhat serious problem minor problem not really a problem	N.a.	1
ntraffic	heavy traffic or speeding cars in neighbourhood	very serious/somewhat serious problem minor problem not really a problem	no yes	1
nnoise	excessive noise in neighbourhood	very serious/somewhat serious problem minor problem not really a problem	no yes	1
nviolence	violence problem in neighborhood	very serious/somewhat serious problem minor problem not really a problem	N.a.	1
ntrash	Trash and litter problem in neighborhood	very serious/somewhat serious problem minor problem not really a problem	N.a.	1
fam_hx_diab	family history of diabetes	no yes	Same in MESA	1
hx_htn1	History of hypertension	no yes	Same in MESA	1
hx_high_chol1	History of high cholesterol	no yes	Same in MESA	1
hx_diab1	history of high blood sugar or diabetes	no yes	Same in MESA	1
ever_aspirin_regularuse1	ever used aspirin regularly	no yes	N.a.	1
age1	age	Continuous values in years	Same in MESA	1
bmi1	body mass index	Continuous values in kg/m^2	Same in MESA	1

Variable	Description	Levels/categories in MESA	Levels/categories in ELSA	Scenario
waist1	waist circumference	Continuous values in cm	Same in MESA	2
smoking1	smoking status	never former current	Same in MESA	1
alcohol_drinking1	alcohol drinking status	never moderate frequent	Same in MESA	1
heart_rate1	heart rate	Continuous values in beats/min	Same in MESA	2
systolic_bp1	systolic blood pressure	Continuous values in mmHg	Same in MESA	2
diastolic_bp1	diastolic blood pressure	Continuous values in mmHg	Same in MESA	2
htn_med1	use of anti-hypertensive medication	no yes	Same in MESA	1
ldl1	LDL cholesterol	Continuous values in mg/dl	Same in MESA	3
hdl1	HDL cholesterol	Continuous values in mg/dl	Same in MESA	3
tot_chol1	Total cholesterol	Continuous values in mg/dl	Same in MESA	3
trig1	Triglycerides	Continuous values in mg/dl	Same in MESA	3
lipid_med1	Use of lipid-lowering medication	no yes	Same in MESA	1
metabolic_syndrome1	Diagnosis of metabolic syndrome	no yes	N.a.	1
thyroid_med1	Use of thyroid medication	no yes	N.a.	1
depression1	Depression symptoms according to depression scale	no yes	no yes	1
antidepr_med1	Use of antidepressants	no yes	no yes	1
curr_job1	Current occupation	homemaker employed unemployed or retired	Same in MESA	1

Variable	Description	Levels/categories in MESA	Levels/categories in ELSA	Scenario
anger_scale1	Spielberg trait anger scale	Integer values in between 10 and 40	N.a.	1
anxiety_scale1	Spielberg trait anxiety scale	Integer values in between 0 and 40	no yes	1
chronic_burden1	Chronic burden scale (indicator of chronic stress)	Integer values in between 0 and 5	N.a.	1
mod_vig_pa1	Moderate and vigorous physical activity	Continuous values in MET-min/week	Hardly ever or never 1-3 times per month Once per week More than once per week	1
gluc1	Fasting glucose [mg/dl]	Continuous values in mg/dl	Same in MESA	3

3.2 DATA SELECTED FOR ASTHMA MODELS DEVELOPMENT AND VALIDATION

We selected the subjects that were free of asthma at baseline wave (either 2, 4 or 6) and had information on asthma diagnosis in the subsequent waves. From this sample, we excluded the subjects that did not have the clinical examination. The remaining sample includes 9.132 subjects (6.001 with baseline wave 2, 2.460 with baseline wave 4 and 671 with baseline wave 6) of whom 276 developed asthma during the observation period after the baseline. 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. The resulting training set contains 7.297 subjects, while the test set includes the remaining 1.835 subjects.

Variables selected for asthma survival model development and validation

Unfortunately, there was not a good overlap of variables between the MESA and the ELSA datasets. Indeed, on the one hand, some of the most predictive variables in MESA, e.g. family history of asthma and sleep with two pillows, were not available in ELSA. On the other hand, some possible predictive variables, not available in MESA, were collected in ELSA. Therefore, we selected from the ELSA dataset a new set of candidate variables that only partly overlap with the variables used for model development in MESA, and includes some new variables. The list of selected variables, their levels/categories (in ELSA) and the scenario number (1 or 2) are reported in Table 3. New variables include variables related to respiratory symptoms, e.g. chest pain, phlegm, wheezing, problems of the accommodation where the subject lives, e.g. damp and condensation, and parents' death for respiratory disease.

Table 3. Variables selected from the ELSA dataset as candidate predictive variables for the development of survival models of asthma onset.

Variable	Description	Levels/categories	Scenario
ethnicity	ethnicity	White Other	1
gender	gender	female male	1
marital_status	marital status	married/living as married widowed/divorced/separated never married	1
education	education	Lower than high school High-school graduate or some college College and above	1
ntraffic	Accommodation have a problem of pollution grime or other environmental problems caused by traffic	No Yes	1
nnoise	Accommodation have a problem of noise from neighbours or other street noise	No Yes	1
age1	age	Continuous values in years	1
bmi1	body mass index	Continuous values in kg/m^2	1
waist1	waist circumference	Continuous values in cm	2
smoking1	smoking status	never former current	1
alcohol_drinking1	alcohol drinking status	never moderate frequent	1
heart_rate1	heart rate	Continuous values in beats/min	2
depression1	Depression symptoms according to depression scale	no yes	1
curr_job1	Current occupation	homemaker employed unemployed or retired	1
anxiety1	Have anxiety symptoms	Yes No	1
mod_vig_pa1	Moderate and vigorous physical activity	Hardly ever or never 1-3 times per month	1

Variable	Description	Levels/categories	Scenario
		Once per week More than once per week	
wake_breath1	Awakened at night for trouble breathing	no yes	1
chest_pain	Ever had severe pain across the front of your chest lasting for half an hour or more	no yes	1
Phlegm	Usually have phlegm in winter	no yes	1
short_breath_walking	Shortness of breath when walking	no yes	1
wheezing	Had attacks of wheezing or whistling in the last 12 months	no yes	1
damp	Accommodation problem - rising damp in floors and walls	no yes	1
water_in	Accommodation problem - Water getting in from roof, gutters or windows	no yes	1
condensation	Accommodation problem - bad condensation problem	no yes	1
cold	Accommodation problem - too cold in winter	no yes	1
parents_death_respd	Either mother or father died from respiratory disease	no yes	1

4 SECOND VALIDATION OF THE DIABETES CONSENSUS MODEL

In this deliverable, a second validation of the diabetes consensus model is performed using a sample extracted from the ELSA dataset (subsection 3.1). Validation is performed adopting the same procedure and the same metrics used for the validation on the MESA dataset (subsection 4.1). Results of the validation on the ELSA dataset (subsection 4.2) are compared with those of the validation on the MESA dataset.

4.1 METHOD

The diabetes consensus model was assessed on the test set extracted from the ELSA dataset, as described in subsection 3.1, following the same procedure adopted for internal validation in MESA. The eight literature models included in the diabetes consensus model, i.e. the model by Stern et al. [4], FINDRISC [5], the three ARIC models (ARIC 1, ARIC 2 and ARIC 3) [2], Framingham model [6], the basic risk score by Kahn et al. [3] and DPoRT [7], were recalibrated by the partial recalibration strategy adopted in work by Kangne et al. [10] and implemented in our consensus model. Such partial recalibration strategy basically adjust the risk scores of the original models using a correction factor based on observed incident diabetes rate at a certain follow-up, ρ_O, and the respective incident diabetes rate predicted by the original model, ρ_P. As for the MESA dataset, for model recalibration, we used incidence rates at 8 years calculated on the training set data. Then, the diabetes consensus model's risk score for each subject was obtained as the weighted average of the scores of the recalibrated models that could be applied to that subject. Models of scenario 1 (DPoRT and FINDRISC) had weight 1, models of scenario 2 (ARIC 1 and model by Kahn et al.) had weight 2, and models of scenario 3 (model by Stern et al., ARIC 2, ARIC 3 and Framingham) had weight 3.

Performance of the diabetes consensus model was assessed in terms of:

- discriminatory ability, by calculating the concordance index (C-index) and the area under the ROC curve (AUC) at 8 years;
- calibration, by calculating the expected to observed event ratio (E/O) at 8 years;
- missing model predictions, by calculating the percentage of subject for whom the model cannot return a valid risk score (MMP).

Discriminatory ability is the ability to correctly rank the subjects according to their risk of diabetes or asthma onset. Two metrics were considered for discriminatory ability: AUC and C-index. AUC 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 AUC 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.

The C-index, proposed by Harrell et al. [8], is an extension of AUC to be used when information on model outcome is available over time. 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 is the extent of agreement between observed incidence of diabetes or asthma and that predicted by the model. 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* [9]. 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.

Confidence intervals for these metrics were constructed by a bootstrap validation in the training set. Specifically, 100 sets of subjects were extracted from the training set by bootstrap resampling and the 8-year diabetes incidence rate for the kth set, $\rho_{O,k}$, was calculated. Then, performance of the consensus model was assessed on each of the 100 out-of-bag samples, using the rate $\rho_{O,k}$ for the model recalibration in the respective kth out-of-bag sample, for k=1,…,100. Finally, median and 95% confidence intervals were calculated for all the metrics on the 100 out-of-bag samples.

4.2 RESULTS

Performance metrics of the diabetes consensus model and the recalibrated literature models used for its derivation (with partial recalibration) are reported in Table 4 for both the test set and the bootstrap validation. Results of this second validation confirm the results of the validation performed on the MESA dataset (Table 5). Indeed, in terms of discriminatory ability (C-index and AUC), the diabetes consensus model is able to achieve better performance than the models of scenario 1 (DPoRT and FINDRISC) and 2 (ARIC 1 and model by Kahn et al.) and similar performance to the models of scenario 3 (model by Stern et al., ARIC 2, ARIC 3 and Framingham). The diabetes consensus model has also good calibration in the ELSA population, with E/O equal to 0.86 on the test set and 0.82 [0.71-0.94] in the bootstrap validation. The advantage of the diabetes consensus model is that it presents a much lower percentage of missing model predictions compared to the other literature models, i.e. 4% for the diabetes consensus model vs a value ranging between 19% and 64% for the literature models.

Table 4. Validation of the diabetes consensus model on the ELSA dataset. Performance metrics are reported for the diabetes consensus model and the literature models recalibrated with partial recalibration. Metrics for the bootstrap validation are reported as median [2.5 percentile – 97.5 percentile] of the values obtained in the 100 bootstrap repetitions.

Model	Test set				Bootstrap validation			
	C-index	AUC at 8 years	E/O at 8 years	MMP [%]	C-index	AUC at 8 years	E/O at 8 years	MMP [%]
Diabetes consensus model	0.77	0.80	0.86	4%	0.77 [0.74-0.80]	0.79 [0.76-0.82]	0.82 [0.71-0.94]	4% [4-5]%
DPoRT men	0.74	0.80	1.98	24%	0.71 [0.68-0.77]	0.73 [0.70-0.79]	1.53 [1.26-1.79]	24% [23-25]%
DPoRT women	0.72	0.73	0.43		0.71 [0.66-0.75]	0.72 [0.67-0.76]	0.46 [0.36-0.56]	

FINDRISC	0.73	0.76	0.77	19%	0.72 [0.69-0.74]	0.72 [0.70-0.75]	0.74 [0.63-0.84]	20% [19-21]%
ARIC 1	0.75	0.77	0.89	33%	0.74 [0.71-0.77]	0.74 [0.71-0.78]	0.97 [0.84-1.13]	34% [33-35]%
Kahn	0.74	0.76	0.88	38%	0.74 [0.71-0.78]	0.75 [0.72-0.78]	0.99 [0.80-1.17]	39% [38-40]%
Stern	0.77	0.80	1.20	64%	0.79 [0.76-0.84]	0.81 [0.77-0.87]	1.26 [1.06-1.60]	63% [61-64]%
ARIC 2	0.75	0.78	1.06	63%	0.78 [0.74-0.82]	0.80 [0.75-0.86]	1.13 [0.95-1.46]	62% [60-64]%
ARIC 3	0.78	0.81	1.06	64%	0.80 [0.77-0.84]	0.82 [0.78-0.87]	1.12 [0.95-1.45]	63% [61-64]%
Framingham	0.85	0.90	0.85	64%	0.81 [0.77-0.85]	0.83 [0.78-0.88]	0.85 [0.70-1.08]	63% [62-65]%

Table 5. Validation of the diabetes consensus model on the MESA dataset. Performance metrics are reported for the diabetes consensus model and the literature models recalibrated with partial recalibration. Metrics for the bootstrap validation are reported as median [2.5 percentile – 97.5 percentile] of the values obtained in the 100 bootstrap repetitions.

Model	**Test set**				**Bootstrap validation**			
	C-index	**AUC at 8 years**	**E/O at 8 years**	**MMP [%]**	**C-index**	**AUC at 8 years**	**E/O at 8 years**	**MMP [%]**
Diabetes consensus model	0.83	0.87	0.82	0%	0.79 [0.76-0.82]	0.83 [0.80-0.86]	0.83 [0.72-1.09]	0% [0-0]%
DPoRT men	0.70	0.72	1.76	1%	0.67 [0.62-0.71]	0.68 [0.63-0.74]	1.80 [1.47-2.40]	1% [0-1]%
DPoRT women	0.70	0.74	0.53		0.69 [0.65-0.74]	0.71 [0.67-0.77]	0.51 [0.42-0.71]	
FINDRISC	0.70	0.74	0.70	0%	0.67 [0.64-0.71]	0.70 [0.66-0.74]	0.70 [0.59-0.94]	0% [0-0]%
ARIC 1	0.73	0.79	0.84	45%	0.71 [0.66-0.75]	0.74 [0.68-0.78]	0.90 [0.72-1.39]	43% [41-45]%
Kahn	0.75	0.80	0.85	46%	0.73 [0.68-0.77]	0.74 [0.68-0.78]	0.90 [0.72-1.42]	46% [44-47]%
Stern	0.81	0.85	0.89	42%	0.81 [0.76-0.84]	0.85 [0.81-0.88]	1.02 [0.86-1.38]	41% [39-42]%
ARIC 2	0.82	0.87	0.81	45%	0.83 [0.80-0.86]	0.87 [0.80-0.86]	0.83 [0.80-0.86]	43% [42-45]%
ARIC 3	0.83	0.87	0.81	45%	0.83 [0.80-0.86]	0.87 [0.83-0.90]	0.88 [0.72-1.29]	43% [41-45]%

Framingham	0.83	0.86	0.86	17%	0.78 [0.74-0.81]	0.82 [0.78-0.85]	0.68 [0.58-0.92]	16% [15-17]%