PULSE

Participatory Urban Living for Sustainable Environments

D5.4 Characterization of the Dynamic Risk of the Disease Onset based on Survival Analysis and Dynamic Bayesian Networks

PULSE project

H2020 - 727816

University of Padova

October 2018

DOCUMENT INFO

0.1 AUTHORS

Author	Organization	e-mail
Martina Vettoretti	University of Padova	martina.vettoretti@dei.unipd.it
Enrico Longato	University of Padova	enrico.longato@dei.unipd.it
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
03/10/2018	0	UNIPD	First template	Draft
17/10/2018	1	UNIPD	Draft with partial contents	Draft
31/10/2018	2	UNIPD	First complete draft	Draft
27/11/2018	2.1	UNIPD	First revised version	Draft
30/11/2018	2.1	UNIPD	Final version	Final

0.3 DOCUMENT KEYDATA

Key words	H2020 – 727816 – PULSE Deliverable 5.4	
Editor info	Name	Barbara Di Camillo
	Organization	UNIPD
	e-mail	barbara.dicamillo@dei.unipd.it

0.4 DISTRIBUTION LIST

Date	Issue	Distribution list
17/10/2018	Circulate first complete draft for internal revision	Maria Fernanda Cabrera, Riccardo Bellazzi,
30/10/2018	Final revised version	Maria Fernanda Cabrera, Riccardo Bellazzi,
30/11/2018	Final version	All Consortium and the European Commission

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

EXECUTIVE SUMMARY

The purpose of this deliverable, entitled “Characterization of the dynamic risk of the disease onset based on survival analysis and dynamic Bayesian networks”, is to provide the consortium with models of disease onset for adult asthma and diabetes.

The major advantage of using Survival Analysis is that it allows accounting for censored observations as well as time to event. This allowed us to use the whole set of available subjects, and to obtain as output more complete information, i.e. the curve of patient risk across years.

Since the data available to the Pulse project contain many dynamic variables, we have also developed new predictive models based on dynamic Bayesian networks. Differently from Survival Analysis models, which can be viewed as deterministic, dynamic Bayesian networks allow modelling the stochastic evolution of variables showing how variables regulate each other over time in terms of probabilistic distribution of each variable at each time point. Once defined, the dynamic Bayesian network model can be exploited to perform simulated trials. It is also possible to evaluate the impact of the single biomarkers for the probability of disease onset and, eventually, efficacy in prevention treatment. It is also possible to identify clusters of patients with similar clinical histories and re-assess their risk profiles accordingly. Once a patient is assigned to a risk group, a predictive model, based on subject specific features, is used to derive the patient-specific probability of the event of interest.

1 INTRODUCTION

The main objective of this work is to adopt different methodological approaches to predict the risk of Type 2 Diabetes(T2D) and Asthma onset on adult population, based on different regression, classification and data mining algorithms and on more traditional risk scores such as FINDRISC (Lindström and Tuomilehto, 2003).

Stemming from the PULSE data assembly activities, we run our analysis on MESA dataset (see deliverable D5.1, Section 3.1.2) that could be useful for the aforementioned objective given that it includes many different variables either directly measured or measured as a proxy (e.g., the energy expenditure due to physical activity, which is derived by a questionnaire in MESA, can be calculated from Fitbit data in PULSE) in PULSE.

The purpose here is many-fold:

- To develop a consensus of existing predicting models when available (as in case of T2D, whereas models of asthma onset are not available in the literature for adult population). This strategy was designed to better generalize on previously unseen data and cope with different nature and availability of variables collected in PULSE and in real-life scenarios.
- To develop new models of disease onset including environmental variables in the models and assessing the ability of different variables to improve predictions. Support Vector Machines (SVMs) with linear and radial kernel and Cox logistic regression coupled with the LASSO (Least Absolute Shrinkage and Selection Operator) were trained on data suitably split in training and validation set. The training set was used to learn the classifier and to rank the variables by using an embedded recursive feature elimination schema coupled with bootstrap. The validation set was used to independently assess method performance.
- To develop a Dynamic Bayesian Network model of disease onset, including the possibility of detecting probabilistic relationships among clinical, behavioural and environmental. Furthermore, the DBN was used to simulate the temporal evolution of variables in time and to stratify subjects by risk factors. The simulation of time-to-onset of the different subgroups was also implemented.

The three following sections address the 3 above mentioned goals and corresponding modelling techniques.

2 DIABETES CONSENSUS MODEL AND RANKING OF PATIENTS BASED ON THEIR RISK SCORES

Several predictive models of T2D onset were proposed in the literature to identify subjects at risk of developing T2D. Although T2D predictive models usually perform very well in the populations in which they were developed, they often present suboptimal performance when applied to new populations, mainly because of differences in the variables' definition and different population characteristics. In this case, predictive models need to be recalibrated, i.e., their parameters need to be updated to describe the new population.

In our previous deliverable, D5.3, we implemented eight literature models for prediction of T2D onset and assessed them on the MESA population both in their original version and after recalibrating them by re-estimating all the model parameters on the MESA dataset. By comparing the original vs. the recalibrated models, we could observe that recalibration significantly improved the model accuracy in predicting the T2D onset probability (*model calibration*), but did not significantly improve the model ability to rank subjects according to T2D risk (*model discriminatory ability*). This means that, the models can correctly estimate the relative T2D risk of a subject compared to the population risk without any recalibration. However, the models require a recalibration to estimate correctly the probability of a subject to develop T2D within a certain time.

Model recalibration can be performed by several strategies. In deliverable D5.3, we adopted a "full recalibration" strategy in which all the model parameters are re-estimated in the new population. This strategy is expected to maximise the model performance in the new population because all the model parameters are updated. However, in order to perform a full recalibration, a rich dataset must be available, which must contain measurement of all the model predictors at a baseline time and longitudinal information on T2D for several years after the baseline (e.g. 8-10 years). The problem is that, in practice, such rich information on the target population often is not available, or it is available only for a limited number of subjects, not enough for a robust estimation of the model parameters. For this reason, other simpler recalibration strategies, that we call "partial recalibration", were proposed in which only 1-2 parameters of the model (typically the intercept and/or the scale parameters) are updated according to diabetes incidence in the target population [1][2]. These strategies exploit only information on T2D incidence in the target population.

Besides the problem of recalibration, another issue of applying literature models to new populations is that the model output often cannot be calculated for certain subjects (*missing model prediction*) because some of the model input variables are missing or not defined for certain groups of subjects, e.g., different racial/age groups. For example, among the literature models that we assessed in deliverable D5.3, the ARIC models [3] and the model by Kahn et al. [4] could not be applied to Chinese and Hispanic subjects because they were originally developed in a population in which only Caucasian and Black races were represented and, thus, in these models variable race can only take two values "White/Caucasian" or "Black". Similarly, the model by Stern et al. [5] cannot be applied to Chinese and Black subjects, because in the development cohort of this model only White and Hispanic subjects were represented.

In order to overcome the problem of missing model predictions, we devised a diabetes consensus model that combines multiple existing models of T2D onset risk and manages the issues related to lack of calibration by implementing a suitable recalibration technique. We also propose a diabetes consensus ranker that returns a global ranking of subjects by combining multiple existing models of T2D onset risk.

2.1 DEFINITION OF DIABETES CONSENSUS MODEL

The diabetes consensus model returns a global score calculated as the weighted average of the risk scores of multiple models recalibrated on a target population. More formally, imagine we want to assess the risk of T2D onset in a population of N subjects, described by the covariate vectors $\boldsymbol{X_i}$ i=1,…,N, and there are M different models to perform this task, which return the risk scores $y_j(\boldsymbol{X_i})$ j=1,…,M for subjects i=1,…,N. Imagine also that each model has been recalibrated on the target population (either by full or partial recalibration) and let us define $y'_j(\boldsymbol{X_i})$ the score of the jth recalibrated model for the ith subject. Then, the global score returned by the consensus model is:

$$\bar{y}(\boldsymbol{X_i}) = \frac{\sum_{j=1}^{M} w_j \cdot \delta_{j,i} \cdot y'_j(\mathbf{X_i})}{\sum_{j=1}^{M} w_j \cdot \delta_{j,i}} \quad (1)$$

where w_j is the weight for the j[th] model and $\delta_{j,i}$ is an indicator function that is equal to 1 if the j[th] recalibrated model can be applied to subject i[th] and 0 otherwise (missing model prediction). The weights w_j j=1,..,M should reflect the model performance, in order to assign larger weights to the models with better performance.

2.2 DEFINITION OF DIABETES CONSENSUS RANKER

The diabetes consensus ranker returns a global ranking of subjects for their T2D onset risk, which is obtained as the weighted average of the risk scores of multiple models, expressed as relative scores with respect to a reference dataset. The method is described more in details in the following. As in previous section, let us suppose we have N subjects described by the covariate vectors $\boldsymbol{X_i}$ i=1,…,N, and M models that return the risk scores $y_j(\boldsymbol{X_i})$ j=1,…,M for subjects i=1,…,N. Imagine also that the M models have already been applied to a reference dataset, with N_{ref} subjects, and have produced M lists of ordered risk scores L_j j=1,…,M. Now we define the relative risk score of the j[th] model for the i[th] subject, $r_j(\boldsymbol{X_i})$, as the percentage of values in L_j that are lower of equal than $y_j(\boldsymbol{X_i})$. Then, for each subject an average relative risk score is computed as:

$$\bar{r}(\boldsymbol{X_i}) = \frac{\sum_{j=1}^{M} w_j \cdot \delta_{j,i} \cdot r_j(\boldsymbol{X_i})}{\sum_{j=1}^{M} w_j \cdot \delta_{j,i}} \quad (2)$$

where, as in eq. (1), w_j is the weight for the j[th] model and $\delta_{j,i}$ is an indicator function that is equal to 1 if the j[th] model can be applied to subject i[th] and 0 otherwise (missing model prediction). Finally, a global ranking of the N subjects is obtained by ordering them according to their average relative risk score.

2.3 IMPLEMENTATION AND ASSESSMENT ON MESA DATASET

2.3.1 SELECTED DATA

The diabetes consensus model and the diabetes consensus ranker were tested on the same data that we selected from MESA dataset in deliverable D5.3 to perform the recalibration of literature T2D models. In particular, from the total MESA sample, we selected the 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.

Then, 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.

2.3.2 SELECTED MODELS

The diabetes consensus model and the diabetes consensus ranker were tested considering the same models implemented and recalibrated in deliverable D5.3. In particular, we considered M=8 literature models of diabetes: the model by Stern et al. [5], FINDRISC [6], the three ARIC models (ARIC 1, ARIC 2 and ARIC 3) [3],

Framingham model [7], the basic risk score by Kahn et al. [4] and DPoRT [8]. These models can be grouped in three different scenarios based on the variables they require (Figure 1). In particular, scenario 1 includes DPoRT and FINDRISC that use only easily accessible information; scenario 2 includes ARIC 1 and the model by Kahn et al., which, in addition to easily accessible information, require some non-invasive measurements collected by medical instruments (e.g. heart rate and blood pressure); finally, scenario 3 includes the models that use biomarkers measured in blood test, i.e. the model by Stern et al., ARIC 2, ARIC 3 and Framingham.

The assessment performed in deliverable D5.3 showed that models in scenario 3 have the best performance in terms of discriminatory ability, followed by the models in scenario 2 and finally the models in scenario 1. This is an expected result because, while in scenario 2 and 3 important risk factors as hypertension and high blood sugar are quantitatively assessed, in the models of scenario 1 such conditions are approximated by self-reported indicators. Nevertheless, models of scenario 1 generally have less missing predictions, because they rely only on easily accessible information. Conversely, models of scenarios 2 and 3 are more likely to have missing model predictions, as clinical measurements and biomarkers may not always be available.

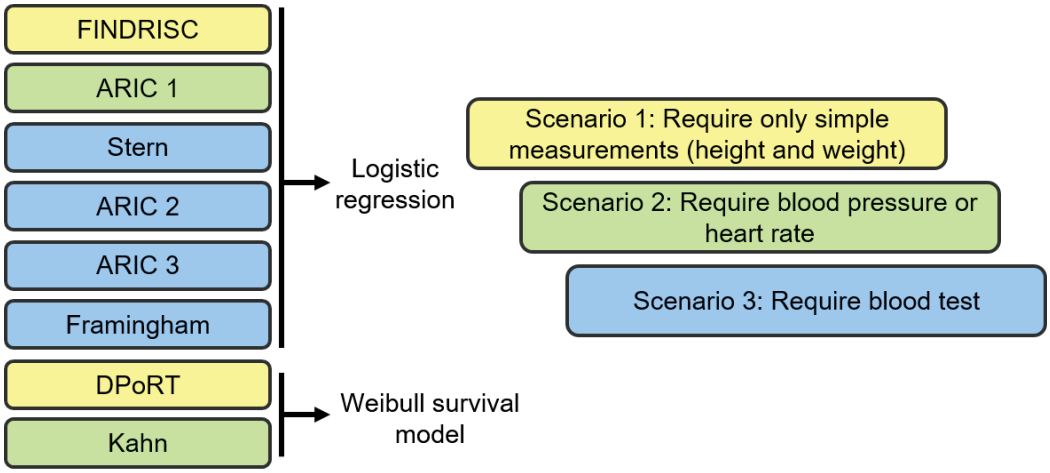

Figure 1. Summary of the selected models. Models of scenario 1, 2 and 3 are evidenced in yellow, green and blue respectively.

2.3.3 IMPLEMENTATION AND ASSESSMENT

The diabetes consensus model was assessed on the test set extracted from the MESA dataset (N=1031). Recalibration of the eight selected models was performed by the partial recalibration strategy adopted in work by Kangne et al. [9]. In particular, for logistic regression models, the recalibrated risk score was calculated as [2]:

$$y'(\boldsymbol{X_i}) = \frac{e^{\boldsymbol{X_i}+\varphi}}{1+e^{\boldsymbol{X_i}+\varphi}} \tag{3}$$

where ϕ is 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:

$$\varphi = \frac{\log(\rho_O/(1-\rho_O))}{\rho_P/(1-\rho_P)} \tag{4}$$

For survival models, the recalibrated incident risk score for follow-up time T was calculated as [1]:

$$y'(\boldsymbol{X_i}) = 1 - \exp(-\exp(\gamma + \log(-\log(1-y(\boldsymbol{X_i}))) \tag{5}$$

where $y(\boldsymbol{X_i})$ is the risk of T2D onset predicted by the model for subject i at follow-up time T, while γ is a correction coefficient calculated based on ρ_O and ρ_P:

$$\gamma = \log(-\log(1-\rho_O)) - \log(-\log(1-\rho_P)) \quad (6)$$

Specifically, in our implementation the recalibration was performed for a follow-up time of 8 years, to be consistent with the recalibration performed in deliverable D5.3. Thus, ρ_O was calculated using the training set as the fraction of subjects in the training set that developed diabetes within 8 years after the baseline visit, while ρ_P was calculated using test set data as the mean predicted diabetes risk of subjects in the test set. In order to assess the efficacy of the partial recalibration, the performance of the models with partial recalibration was compared to those of the original models and the models recalibrated with a full recalibration strategy (derived in deliverable D5.3).

Once the scores of different models had been recalibrated, the diabetes consensus model's global scores were calculated by eq. (1). The weights w_j j=1,…,M were defined as the scenario number, thus models in scenario 1 had weight 1, models in scenarios 2 had weight 2 and models in scenario 3 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).

A description of metrics of discriminatory ability and calibration was provided in deliverable D5.3, Section 3. 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.

The diabetes consensus ranker was assessed on the MESA dataset using a similar scheme to that adopted for the diabetes consensus model: the global ranking was assessed on the test set (N=1031), using the training set as reference set to calculate the relative risk scores (N_{ref}=4124). As for the diabetes consensus model, different models in the average relative risk scores of eq. (2) were weighted according to the model scenario.

The discriminatory ability of the diabetes consensus ranker was assessed on the test using the C-index and 8-year AUC. Confidence intervals for these metrics were derived by performing a bootstrap validation with 100 repetitions on the training set (as done for the diabetes consensus model), in which the ranker performance was assessed on the out-bag-samples considering the respective sample extracted by bootstrap sampling as reference set.

Performance of the diabetes consensus model and the diabetes consensus ranker were compared to those of the eight recalibrated literature models considered for their construction.

2.4 RESULTS OF THE ASSESSMENT

The first part of our assessment focused on comparing the efficacy of different recalibration strategies, i.e. the partial recalibration and the full recalibration. In Table 1, the values of AUC and E/O at 8 years are compared for the eight literature models in their original version (columns "Original") and their recalibrated versions obtained with partial recalibration (columns "Partial recal.") and full recalibration (columns "Full recal."). As already shown in deliverable D5.3, with a full recalibration the AUC slightly improves (e.g., for Stern and ARIC models) or remains unchanged (e.g., for FINDRISC and Kahn), while the E/O significantly improves compared to the original models for most of the models. With a partial recalibration, AUC does not change at all, because only the intercept parameter, which is the same for all the subjects, is updated with this strategy, and the E/O generally improves, although, as expected, this improvement is smaller than with the full recalibration. These results evidenced that:

- model discriminatory ability is good even without recalibration;
- to solve the lack of calibration issue, the re-estimation of all model parameters (full recalibration) is preferable, provided that a sufficiently rich dataset is available for this purpose;
- if the full recalibration cannot be performed, reasonably good calibration performance can still be obtained by a partial calibration approach.

Table 1. Performance of original models, models with partial recalibration and fully-recalibrated models assessed on the test set. Reported metrics are AUC and E/O at 8 years.

Model	AUC at 8 years			E/O at 8 years		
	Original	Partial recal.	Full recal.	Original	Partial recal.	Full recal.
DPoRT men	0.724	0.724	0.735	6.349	1.764	0.956
DPoRT women	0.744	0.744	0.756	3.474	0.534	1.044
FINDRISC	0.735	0.735	0.738	0.521	0.697	0.980
ARIC 1	0.789	0.789	0.825	2.134	0.839	0.861
Kahn	0.803	0.803	0.798	3.909	0.852	0.930
Stern	0.846	0.846	0.864	1.748	0.890	0.891
ARIC 2	0.868	0.868	0.887	1.025	0.812	0.837
ARIC 3	0.872	0.872	0.890	1.046	0.807	0.837
Framingham	0.864	0.864	0.873	0.864	0.864	0.796

The second part of our assessment focused on the testing of the diabetes consensus model and the diabetes consensus ranker. Performance metrics of the diabetes consensus model, the diabetes consensus ranker and the recalibrated literature models used for their derivation (with partial recalibration) are reported in Table 2 for both the test set and the bootstrap validation.

Results show that, in terms of discriminatory ability (C-index and AUC), the diabetes consensus model is able to achieve performance comparable to those of the models of scenario 3, and much better than those of scenarios 1 and 2. The diabetes consensus model results also well calibrated in the MESA population, with E/O equal to 0.82 on the test set and 0.83 [0.72-1.09] in the bootstrap validation. Interestingly, the diabetes consensus model outperforms the models of scenarios 2 and 3 in terms of missing model predictions. Indeed, while the diabetes consensus model presents no missing model predictions, the models of scenarios 2 and 3 have a significant percentage of missing model predictions, which, e.g. on the test set, ranges between 17% (Framingham) and 46% (model by Kahn et al.).

Similar considerations can be done in terms of discriminatory ability and missing model predictions when comparing the diabetes consensus ranker with the literature models. Note that we cannot assess E/O for the diabetes consensus ranker, because this is a ranking tool, which can only rank the subjects according to T2D onset risk, but it cannot estimate the probability of T2D onset and thus the expected number of events in 8 years needed to calculate E/O. Comparing the diabetes consensus model with the diabetes consensus ranker, performances in terms of discriminatory ability are similar (slightly lower for the ranker).

In Figure 2, the ROC curve at 8 years on the test set is plotted for the diabetes consensus model (green), the diabetes consensus ranker (light blue) and ARIC 3 (red), which is the literature model with best discriminatory ability. The three curves are very similar, confirming that the three models have similar ranking performance. In Figure 3, the calibration plot at 8 years obtained for the diabetes consensus model and the ARIC 3 model on the

test set is shown. The plot shows that the two models are generally well calibrated, although they both underestimate the cases of T2D onset for high-risk subjects.

Table 2. Performance of diabetes consensus model and the diabetes consensus ranker compared to the original 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]%
Diabetes consensus ranker	0.82	0.86	-	0%	0.78 [0.76-0.81]	0.82 [0.79-0.85]	-	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]%

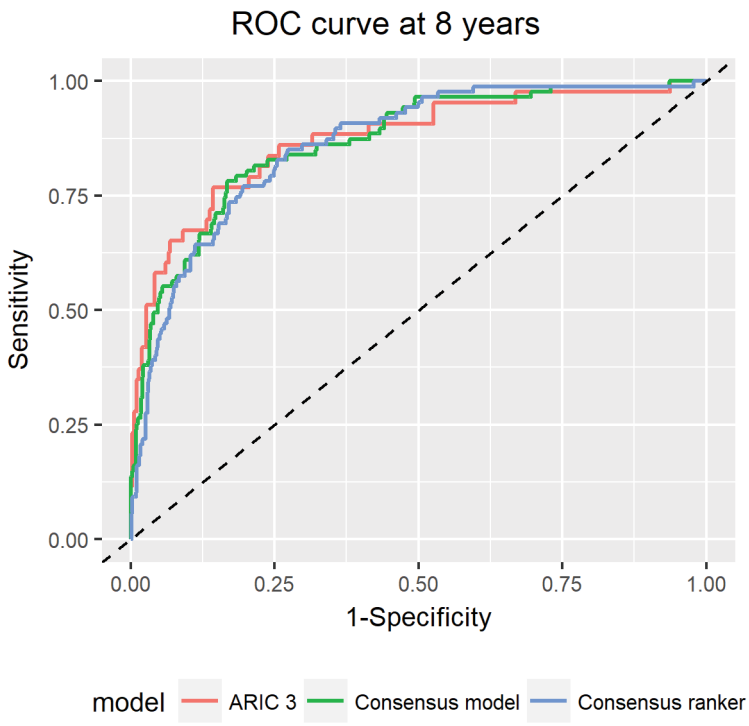

Figure 2. ROC curve at 8 years for the diabetes consensus model (green), the diabetes consensus ranker (blue) and ARIC 3 (red).

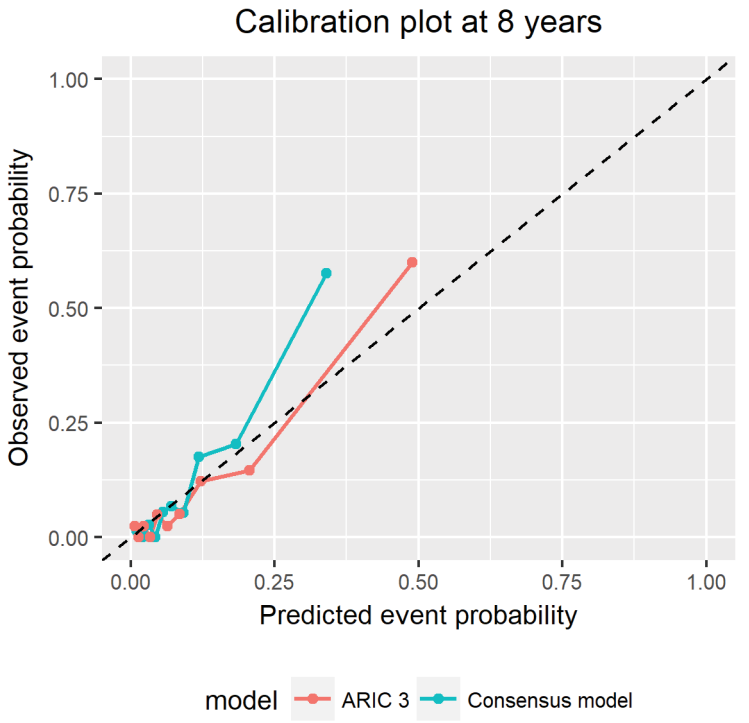

Figure 3. Calibration plot at 8 years for the diabetes consensus model (water green) and ARIC 3 (red).