BMJ Open Diabetes Research & Care

Addressing practical issues of predictive models translation into everyday practice and public health management: a combined model to predict the risk of type 2 diabetes improves incidence prediction and reduces the prevalence of missing risk predictions

Martina Vettoretti,[1] Enrico Longato,[1] Alessandro Zandonà,[1] Yan Li,[2] José Antonio Pagán,[3,4] David Siscovick,[5] Mercedes R Carnethon,[6] Alain G Bertoni,[7] Andrea Facchinetti,[1] Barbara Di Camillo [1]

To cite: Vettoretti M, Longato E, Zandonà A, *et al*. Addressing practical issues of predictive models translation into everyday practice and public health management: a combined model to predict the risk of type 2 diabetes improves incidence prediction and reduces the prevalence of missing risk predictions. *BMJ Open Diab Res Care* 2020;**8**:e001223. doi:10.1136/bmjdrc-2020-001223

- Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2020-001223).

For 'Presented at statement' see end of article.

Received 24 January 2020
Revised 3 June 2020
Accepted 10 June 2020

Check for updates

For numbered affiliations see end of article.

Correspondence to
Dr Barbara Di Camillo;
barbara.dicamillo@unipd.it

ABSTRACT

Introduction Many predictive models for incident type 2 diabetes (T2D) exist, but these models are not used frequently for public health management. Barriers to their application include (1) the problem of model choice (some models are applicable only to certain ethnic groups), (2) missing input variables, and (3) the lack of calibration. While (1) and (2) drives to missing predictions, (3) causes inaccurate incidence predictions. In this paper, a combined T2D risk model for public health management that addresses these three issues is developed.

Research design and methods The combined T2D risk model combines eight existing predictive models by weighted average to overcome the problem of missing incidence predictions. Moreover, the combined model implements a simple recalibration strategy in which the risk scores are rescaled based on the T2D incidence in the target population. The performance of the combined model was compared with that of the eight existing models using data from two test datasets extracted from the Multi-Ethnic Study of Atherosclerosis (MESA; n=1031) and the English Longitudinal Study of Ageing (ELSA; n=4820). Metrics of discrimination, calibration, and missing incidence predictions were used for the assessment.

Results The combined T2D model performed well in terms of both discrimination (concordance index: 0.83 on MESA; 0.77 on ELSA) and calibration (expected to observed event ratio: 1.00 on MESA; 1.17 on ELSA), similarly to the best-performing existing models. However, while the existing models yielded a large percentage of missing predictions (17%–45% on MESA; 63%–64% on ELSA), this was negligible with the combined model (0% on MESA, 4% on ELSA).

Conclusions Leveraging on existing literature T2D predictive models, a simple approach based on risk score rescaling and averaging was shown to provide accurate and robust incidence predictions, overcoming the problem of recalibration and missing predictions in practical application of predictive models.

Significance of this study

What is already known about this subject?

- Several models to predict the risk of type 2 diabetes (T2D) onset exist.However, their application in public health management is quite limited due to: (1) problem of model choice, that is, not all the models areapplicable to all subjects; (2) the problem of missing values, that is, theinput variables required by the model are not always available for allsubjects, and (3) the need to recalibrate the models, since parameterswould need to be re-learned on the population of interest to reachgood prediction performance.

What are the new findings?

- In this study, a new approach is presented based on the combination of eight literature models grouped into three different scenarios based on the variables they require, namely: information that can be collected by questionnaires, non-invasive measurements collected by medical instruments (eg, heart rate and blood pressure), and biomarkers measured by blood test.
- The combined model:
  - rescales original models based only on the incidence of T2D in the target population without the need of re-training them;
  - selects the applicable models for each subject and calculates the risk scores of the models for which data are available, thus overcoming the problem of model choice and missing values and reaching a high coverage of subjects for which one or more risk scores can be computed;
  - calculates a weighted average of risk scores choosing weights so as to give more importance to the models based on clinical measurements.

INTRODUCTION

Identifying individuals at risk of type 2 diabetes (T2D) onset is an important goal,

Significance of this study

How might these results change the focus of research or clinical practice?

- The combined model proposed in this study can facilitate the adoption of T2D predictive models in public health management and clinical practice, in particular allowing accurate and robust risk predictions while addressing the issues of model choice, missing predictions, and the need for recalibration

because targeting prevention initiative to at-risk individuals can potentially reduce the incidence and prevalence of T2D (and related complications) with great healthcare cost reduction. Indeed, several studies demonstrated that T2D can be prevented or delayed by early interventions on modifiable risk factors.[1–5] The standard screening for pre-diabetes and T2D is based on diagnostic testing of plasma glucose or glycated hemoglobin[6] and in the United States, it is recommended only for obese individuals.[7] Nevertheless, many other risk factors of T2D onset exist and, over the past two decades, methods to identify subjects at risk of developing T2D based on multiple factors were developed. In particular, predictive models of T2D onset, developed based on physiological and lifestyle risk factors, provide a risk score representing the individual's probability of developing T2D in the future (eg, in the next 5–10 years).[8] Some predictive models only require easily accessible information that can be periodically collected by population health surveys.[9 10] Others require clinical data (eg, fasting plasma glucose),[11 12] which generally result in better predictive performance. However, the collection of such clinical data, which require a laboratory test, may not be feasible for large population screenings in real world (not in clinical trials or research studies).

Although the use of predictive models is recommended by international guidelines,[6 13] their translation into everyday practice and public health management is scarce.[8 14 15] Barriers to the adoption of predictive models include lack of external validation, limited availability of interpretation guidelines, and few recommendations on model selection.[14] In particular, when applied to a new population, predictive models often present suboptimal performance because of different population characteristics that impact the incidence of T2D. In these cases, the model needs a recalibration, that is, its parameters need to be updated to obtain accurate predictions of T2D risk in the new population.[16–18] However, in most cases, a full recalibration of a published model, that is, the re-estimation of all its parameters in the new population of interest is not possible, because this would require the availability of a rich dataset containing knowledge of the model predictors at a baseline time and longitudinal information on T2D incidence, data which often are not available. In addition, when a specific model is applied to a new population, some other issues need to be addressed. First, if some of the input variables required by the model are missing, the risk score cannot be calculated (unless a data imputation technique is used to infer the missing values). Second, the model may be inapplicable to some groups of individuals. For example, some of the literature models that require the variable "ethnicity" cannot be applied to specific ethnic groups because these groups were not present in the original cohorts used for model development and, thus, they were not included among the "ethnicity" variable categories. For all the above reasons, the practical use of T2D risk models is difficult, and the so-called one-score-fits-all approach based on linear combinations of risk indicators seems inappropriate for public health management and clinical reality.[14]

The aim of this paper is to address some of the practical issues currently limiting the use of T2D predictive models, that is, the problems of model choice, missing variables, and recalibration, in an effort to facilitate the use of existing models in real-world applications. For this purpose, a new approach is proposed which consists in combining multiple existing T2D predictive models and using an easy-to-adopt recalibration strategy, in which the risk scores are rescaled based on the T2D incidence in the target population. The performance of the combined T2D model was assessed on data from two large longitudinal studies with different ethnic composition, so to mimic the model applicability to different and previously unseen populations: the Multi-Ethnic Study of Atherosclerosis (MESA) and the English Longitudinal Study of Ageing (ELSA).

RESEARCH DESIGN AND METHODS

Datasets

MESA dataset

The MESA is a longitudinal study funded by the National Heart, Lung, and Blood Institute starting in July 2000 and still ongoing.[19] MESA investigates subclinical cardiovascular disease in participants without prevalent cardiovascular disease at baseline (n=6814): the study population consisted of non-Hispanic Caucasians (38.5%), African-Americans (27.8%), Hispanic-Americans (21.9%), and Asian-Americans, predominantly of Chinese descent (11.8%). Participants were enrolled from six US communities and included males and females aged 45–84 years. In total, six examinations were conducted in the period 2000–2018. At each examination, subjects were interviewed about their health and lifestyles and underwent clinical assessments. Institutional review boards at all sites approved the study and each participant provided written inform consent.

For the present study, the data collected in the first five examinations were used (at the time of analysis for the present work examination 6 data were not available). The number of participants completing examination 5 was 4655. From the total MESA sample, the subjects who satisfied the following three conditions were selected: (1) no diabetes (either treated or untreated) at enrollment,

(2) information on diabetes available at least at one of the follow-up examinations, and (3) no history of cancer. Diabetes was defined according to the American Diabetes Association (ADA) 2003 fasting criteria algorithm (ie, fasting blood glucose concentration ≥126 mg/dL).[20] The selected subsample included 5155 subjects of whom 640 subjects developed T2D during the follow-up period (184 at examination 2, 106 at examination 3, 147 at examination 4, and 203 at examination 5). Of the remaining 4515 subjects who did not present diabetes during the study, 3299 had follow-up till examination 5 (196, 219, and 801 subjects exited the study after examination 2, 3, and 4, respectively, without diabetes). For each subject, the time of diabetes diagnosis was defined as the time of the first examination at which a diabetes diagnosis was registered. In the analyses of this paper, the variables collected at the first examination (online supplementary table S1) are used to predict the T2D incidence during the follow-up period.

ELSA dataset

The ELSA is an ongoing study of health, social, well-being, and economic circumstances in the English population aged 50 years and older, funded by the US National Institute of Ageing and a consortium of UK government departments.[21] The sample mostly included non-Hispanic Caucasians (about 98% of the sample). Participants have a face-to-face interview every 2 years and a clinical examination every 4 years. Currently, the study includes eight waves of data collection covering a period of 15 years (2002–2017). At waves 3–4 and 6–7, new participants entered the study to maintain the size of the sample.

Since the clinical examinations were performed only in waves 2, 4, 6, and 8, each subject was assigned 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=9432), subjects that entered in waves 3/4 were assigned baseline wave 4 (n=4357), and subjects recruited in waves 5/6 were assigned baseline wave 6 (n=1557). Then the subjects who (1) were free of diabetes at the baseline wave, (2) had the clinical examination at the baseline wave, and (3) had information on diabetes diagnosis at following waves were selected. Diabetes diagnosis was defined as positive answer to the question: "Has a doctor ever told you that you have diabetes or high blood sugar?". The selected sample included 9641 subjects (6304 with baseline wave 2, 2615 with baseline wave 4, and 722 with baseline wave 6) of whom 737 developed diabetes during the observation period after the baseline (114 subjects first reported diabetes at wave 3, 101 at wave 4, 159 at wave 5, 116 at wave 6, 142 at wave 7, and 105 at wave 8). Of the remaining 8904 subjects, 5575 had follow-up till wave 8 (649, 472, 598, 727, and 883 subjects exited the study at wave 3, 4, 5, 6, and 7, respectively, without reporting diabetes). For each subject, the time of diabetes diagnosis was defined as the time of the first wave at which a diabetes diagnosis was registered. In this paper, the variables collected at the baseline wave (online supplementary table S1) are used to predict the T2D incidence observed at waves after the baseline.

Selected literature models

Eight literature predictive models of T2D onset that could be applied with the variables collected both in MESA and ELSA in their original form (ie, without the need of a model revision or re-estimation of model parameters) were selected. Six models are based on logistic regression: the concise version of the Finnish Diabetes Risk Score (FINDRISC),[22] the Atherosclerosis Risk in Communities simple model (ARIC 1),[23] the ARIC clinical model without lipids (ARIC 2),[23] the ARIC clinical model with lipids (ARIC 3),[23] the model published in Ref. 24 (STERN), and the Framingham model (FRAMINGHAM).[11] The other two models are based on Weibull survival model, that is, the Diabetes Population Risk Tool (DPoRT),[9] and the basic model described in Ref. 25 (KAHN). DPoRT, in particular, implements two different Weibull models for the men and the women, respectively.

These models were grouped into three different scenarios based on the variables they require, similarly to what performed in Ref. 12. Scenario 1 (Sc1) includes DPoRT and FINDRISC that use only easily accessible information that can be collected by questionnaires. Scenario 2 (Sc2) includes ARIC 1 and KAHN, which, in addition to the information of Sc1, require some non-invasive measurements collected by medical instruments (eg, heart rate and blood pressure). Finally, scenario 3 (Sc3) includes STERN, ARIC 2, ARIC 3, and FRAMINGHAM, which use, in addition to the variables of Sc2, some biomarkers measured by blood test (eg, fasting plasma glucose and cholesterol concentration). Details on the variables required for the application of the eight selected models (harmonized between the two datasets) are reported in online supplementary table S1.

Definition of the combined T2D model

The problem of missing values is more frequent with models that require clinical variables (eg, models of Sc2/Sc3) that, even if commonly collected in routine clinical encounters, may not be readily available for all the individuals of the general population (not limited to the patients regularly visiting the primary care). Models developed on specific subpopulations recurrently suffer from limited applicability. For example, ARIC and KAHN models were originally developed in a population where only non-Hispanic Caucasians and African-Americans were represented. Therefore, these models include a variable "race/ethnicity" that can only take two values, making the models not applicable to other racial/ethnic groups, such as Asian-Americans and Hispanics, which are present in the MESA dataset. Similarly, the STERN model can be applied only to non-Hispanic white and Hispanic participants because these were the only two racial/ethnic groups represented in its development cohort. One strategy to minimize the issue of missing predictions, caused by missing values and model inapplicability to specific racial/ethnic groups, is to combine the output of multiple models, considering for each subject only the models that can be applied to him/her with the available data. This is the idea behind the development of the combined T2D model evaluated in this report.

The combined T2D model was defined as the weighted average of the risk scores of eight existing models, after rescaling them based on the T2D incidence of the target population. The calculation of the combined T2D model's risk score can be divided into four steps. The first step is to select, for each participant, the applicable models, by excluding the models that are inapplicable to the specific subject because of ethnicity or variable missingness. The second step is to calculate the risk scores for each of the selected models using the original model parameters. The third step is to rescale the risk scores according to the T2D incidence observed in the target population.[26 27] The rescaling is equivalent to adjusting the intercept of the model which represents how much the factors not included in the model contributes to the T2D incidence, thus addressing the issue of possible lack of calibration caused by factors not explicitly taken into account by the model regression. The advantage of using the rescaling method to recalibrate the model is that it only requires information on T2D incidence in the target population (details reported in the online supplementary material). The final step is to combine the rescaled risk scores by weighted average to produce a global T2D risk score. Weights are set according to the scenario number, specifically the weight is 1 for Sc1 models (DPoRT and FINDRISC), 2 for the Sc2 models (ARIC 1 and KAHN), and 3 for the Sc3 models (ARIC 2, ARIC 3, STERN, and FRAMINGHAM).

In figure 1A, an example of application of the combined model to a specific individual's data is shown. Let us consider the imaginary subject John, an African-American, 55-year-old individual of New York with missing values on cholesterol level and heart rate (this is not a real subject, but an invented case used as illustrative example). For this individual, STERN is not applicable because this model can be applied only to Caucasian or Hispanic subjects. Moreover, KAHN is not applicable because of the missing value on heart rate, while ARIC 3 and FRAMINGHAM cannot be applied because of the missing value on cholesterol level. Then, only four models can be applied with John's data and the respective risk scores are rescaled according to the 8-year T2D incidence in the New York population. Finally, the weighted average of the four rescaled risk scores is computed, weighting each risk score for the Sc number. The final result is the combined model risk score which represents John's probability of developing T2D in the next 8 years.

Performance assessment

The diagram in figure 1B summarizes all the analyses performed on the MESA and ELSA datasets to assess the performance of the combined model and the existing models, in their original form, after full recalibration and simple risk score rescaling. A detailed description of the analyses is provided below.

Assessment of recalibrated models

Before evaluating the performance of the combined T2D model, the MESA dataset was used to compare the performance of the models obtained with rescaling to those of the original non-recalibrated models and those of fully recalibrated models, that is, models in which all the parameters (including beta coefficients) were re-estimated in the new population (in this case, the MESA dataset).

To perform the full recalibration, the MESA selected data were split into a training and a test set, including the 80% and 20% of selected subjects, respectively, stratified for incidence of T2D. The stratification was performed to balance the percentage of T2D cases in the training and test set. Indeed, an unstratified split could drive to the unlucky situation in which most cases occur in the training set and just few in the test set, or vice versa, which would impair the model training and its assessment. The training set contained 4124 subjects (512 with incident T2D during the study), while the test set included the remaining 1031 subjects (128 with incident T2D during the study).

The eight selected models were fully recalibrated by re-estimating all the model parameters in the MESA training set. The logistic regression models were fitted to the outcome at 8 years after the baseline (ie, the first examination). The 8-year threshold was chosen because it well approximates the average follow-up duration of the cohort studies used to develop the selected models (7 years for FRAMINGHAM, 7.5 years for STERN, 9 years for ARIC and 10 years for FINDRISC). Moreover, this threshold allowed for a large sample of subjects with incident T2D, while limiting the number of censored subjects (ie, exiting the study without developing T2D) prior to the end of the follow-up period.

In the rescaling approach, the intercept of the models was adjusted based on the T2D incidence at 8 years after the baseline calculated on the MESA training set.

Performance of the models after full recalibration and simple risk score rescaling was assessed on the MESA test set and compared with those of the original non-recalibrated models.

Assessment of the combined T2D model

The combined T2D model was assessed both on the MESA and ELSA test sets. In MESA, the risk score rescaling was performed using the 8-year T2D incidence extracted from the training set used for the full recalibration. Regarding ELSA, the selected data were split into two equal parts. From the first part (n=4821), called reference set in figure 1B, the 8-year T2D incidence of the ELSA population was estimated. The second part (n=4820) was used as test set to assess the model performance. For each subject in the MESA/ELSA test set, the applicable models were selected, their risk scores were calculated by using the coefficients provided in the original publications, such risk scores were rescaled using

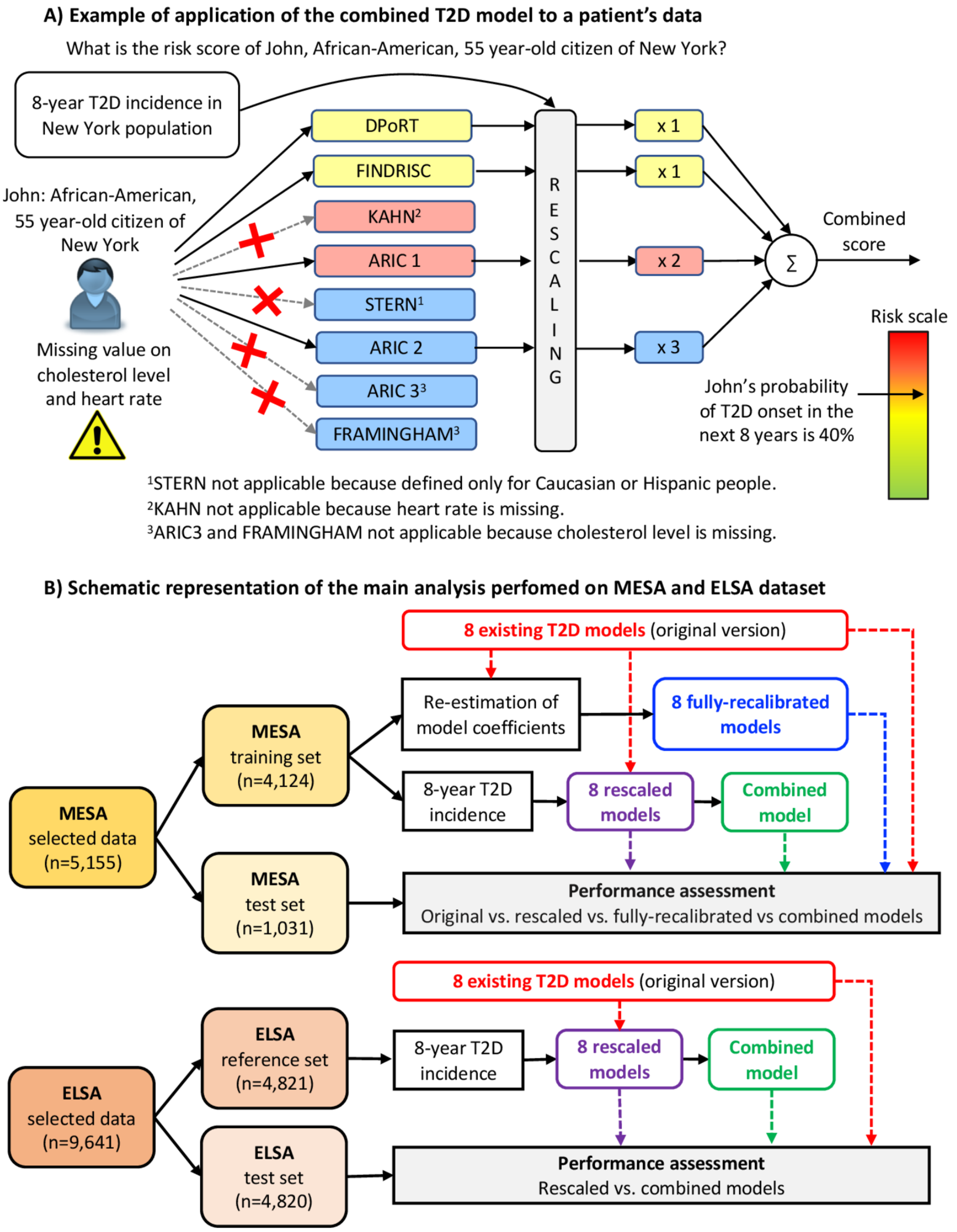
A) Example of application of the combined T2D model to a patient's data
What is the risk score of John, African-American, 55 year-old citizen of New York?
8-year T2D incidence in New York population
John: African-American, 55 year-old citizen of New York
Missing value on cholesterol level and heart rate
DPoRT
FINDRISC
KAHN²
ARIC 1
STERN¹
ARIC 2
ARIC 3³
FRAMINGHAM³
RESCALING
x 1
x 1
x 2
x 3
Σ
Combined score
Risk scale
John's probability of T2D onset in the next 8 years is 40%
¹STERN not applicable because defined only for Caucasian or Hispanic people.
²KAHN not applicable because heart rate is missing.
³ARIC3 and FRAMINGHAM not applicable because cholesterol level is missing.
B) Schematic representation of the main analysis perfomed on MESA and ELSA dataset
8 existing T2D models (original version)
Re-estimation of model coefficients
8 fully-recalibrated models
MESA selected data (n=5,155)
MESA training set (n=4,124)
8-year T2D incidence
8 rescaled models
Combined model
MESA test set (n=1,031)
Performance assessment
Original vs. rescaled vs. fully-recalibrated vs combined models
8 existing T2D models (original version)
ELSA selected data (n=9,641)
ELSA reference set (n=4,821)
8-year T2D incidence
8 rescaled models
Combined model
ELSA test set (n=4,820)
Performance assessment
Rescaled vs. combined models

Figure 1 (A) Illustrates the steps for calculating the combined model risk score, considering the data of an imaginary individual, that is, John, an African-American, 55-year-old citizen of New York with missing values on cholesterol level and heart rate (this is not a real subject, but an invented case used as illustrative example). (B) Schematizes the analyses performed on the MESA and ELSA datasets for the assessment of the combined model and the eight existing models (original version, rescaled models, and fully recalibrated models). ARIC 1, Atherosclerosis Risk in Communities simple model; ARIC 2, ARIC clinical mode lwithout lipids; ARIC 3, ARIC clinical model with lipids; DPoRT, Diabetes Population Risk Tool; ELSA, English Longitudinal Study of Ageing; FINDRISC, Finnish Diabetes Risk Score; FRAMINGHAM, Framingham model; MESA, Multi-Ethnic Study of Atherosclerosis; T2D, type 2 diabetes.

the 8-year T2D incidence of the MESA/ELSA population (derived from a separate group of MESA/ELSA subjects, not present in the test set), and finally the weighted average of rescaled risk scores is computed.

Note that the MESA and ELSA test sets were exclusively used to assess the models' performance.

Metrics

The performance of each model was assessed in terms of discriminatory ability, calibration and missing predictions. The discriminatory ability is the ability of the model to correctly rank the subjects according to their risk of T2D onset. Discriminatory ability was assessed by the receiver operating characteristic (ROC) curve (sensitivity vs 1-specifity) at time t, the area under the ROC curve (AU-ROC)[28] at time t and the concordance index (C-index).[29] A CI of 95% was derived for the AU-ROC as detailed in Ref. 28 and for the C-index by using the Noether estimator.[30]

Calibration represents the ability of the model to correctly predict the incidence of T2D over a certain period of duration t after the baseline. Model calibration was graphically assessed by visualizing the calibration plot at a certain time t.[31 32] Calibration was also quantitatively assessed by the expected to observed event ratio (E/O), that is, the ratio between the expected number of events within time t, obtained as the sum of the probabilities of developing T2D within time t predicted by the model, and the number of observed events in the same time period.[33] 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 overstimate/underestimate the event probability. 95% CIs for the E/O were calculated as in Ref. 33.

The C-index calculation, not depending on time, was performed considering as an outcome the time-to-event, which is defined as the time, in years from the baseline, to the first report of diabetes for subjects with incident diabetes, or time to censoring for subjects that exit the study without diabetes. The C-index is thus calculated using all the subjects in the test sets.

Conversely, the ROC curve, AU-ROC, E/O, and calibration plot, which do not consider the time of events and lost to follow-up subjects, were assessed considering as an outcome a binary variable representing the diabetes status at t=8 years after the baseline time (consistently with the rescaling and the full recalibration). Subjects who first reported diabetes within 8 years from the baseline were assigned outcome “1,” while those that exited the study after 8 years from baseline without reporting diabetes were assigned class “0.” The subjects who were either censored within 8 years from the baseline or first reported T2D after 8 years from the baseline were excluded from the calculation of these metrics, because the class of censored subjects is unknown at 8 years, whereas assigning a class label equal to “healthy” to subjects that might develop T2D just after this threshold might lead to unwanted bias. As a result, on MESA test set, time-dependent metrics were calculated on a subset of 739 subjects (87 with incident diabetes, 652 without incident diabetes); on ELSA test set, the computation was performed on a subset of 3295 subjects (291 with incident diabetes, 3004 without incident diabetes).

In addition, the missing model predictions were assessed, that is, the percentage of subjects for whom the model cannot return a valid risk score, because some of the input variables are missing or the model is not applicable to the racial/ethnic group of the subject.

RESULTS

Efficacy of model recalibration

On the MESA test set, performance of the original literature models are compared with those of the models recalibrated with either rescaling or full recalibration. In terms of discriminatory ability, the rescaling does not affect the model performance, as it simply adjusts the model intercept parameter (common to all the subjects). Therefore, as far as discrimination is concerned, the comparison is limited to original versus fully recalibrated models. As visible in figure 2, the ROC curves and the AU-ROC values of original and fully recalibrated models are very similar for all the models, suggesting that the full recalibration only marginally affects the model discriminatory ability. Therefore, in principle, it is not necessary to recalibrate a model on a new population if the investigator is only interested in ranking subjects based on their risk.

Regarding calibration, in figure 3, the observed versus predicted event probability plot, and the corresponding E/O values, are reported for original, rescaled, and fully recalibrated models. The calibration plots and the E/O values show that almost all the original models (except ARIC 2 and ARIC 3) are not well calibrated on the MESA population, with models that significantly overestimate (DPoRT, ARIC 1, KAHN, and STERN) or underestimate (FINDRISC and FRAMINGHAM) the T2D incidence. Calibration performance significantly improves with model recalibration. Indeed, both the rescaled models and the fully recalibrated models present a calibration plot very close to the line with intercept 0 and slope 45°, and E/O values close to 1, indicating a good agreement between the observed event probability and the one predicted by the model.

In summary, these results suggest that: (1) model discriminatory ability is good even without recalibration, (2) the original models suffer from a lack of calibration when applied to different popuations, and (3) the rescaling is sufficient to achieve a good incidence prediction, comparable to that of the full recalibration.

Combined T2D model versus literature models

The comparison of the different literature models on the MESA and ELSA test sets (table 1) shows that the best-performing models in terms of discrimination (C-index) are the models of Sc3, followed by Sc2 and finally Sc1. This is expected because, while in Sc2/Sc3 important risk factors such as hypertension and high blood glucose are quantitatively assessed, in Sc1 such conditions are approximated by self-reported indicators. However, the models of Sc2 and Sc3 present a lot of missing predictions compared with the models of Sc1. For MESA, such high percentage of missing predictions is due to the multiethnicity of the MESA population: the models of Sc2 and Sc3 (except FRAMINGHAM) are not applicable to all the MESA ethnic groups. For ELSA, the high percentage of missing predictions is due to the high level of missingness of variables required by Sc2 and Sc3 models (see online supplementary table S1). Specifically, the percentage of missing predictions of Sc3’s models ranged between 17% and 45% in the MESA test set, and between 63% and 64% in the ELSA test set.

This limitation can be overcome by the combined T2D model for which the percentage of missing predictions is 0% in the MESA test set and 4% in the ELSA test set

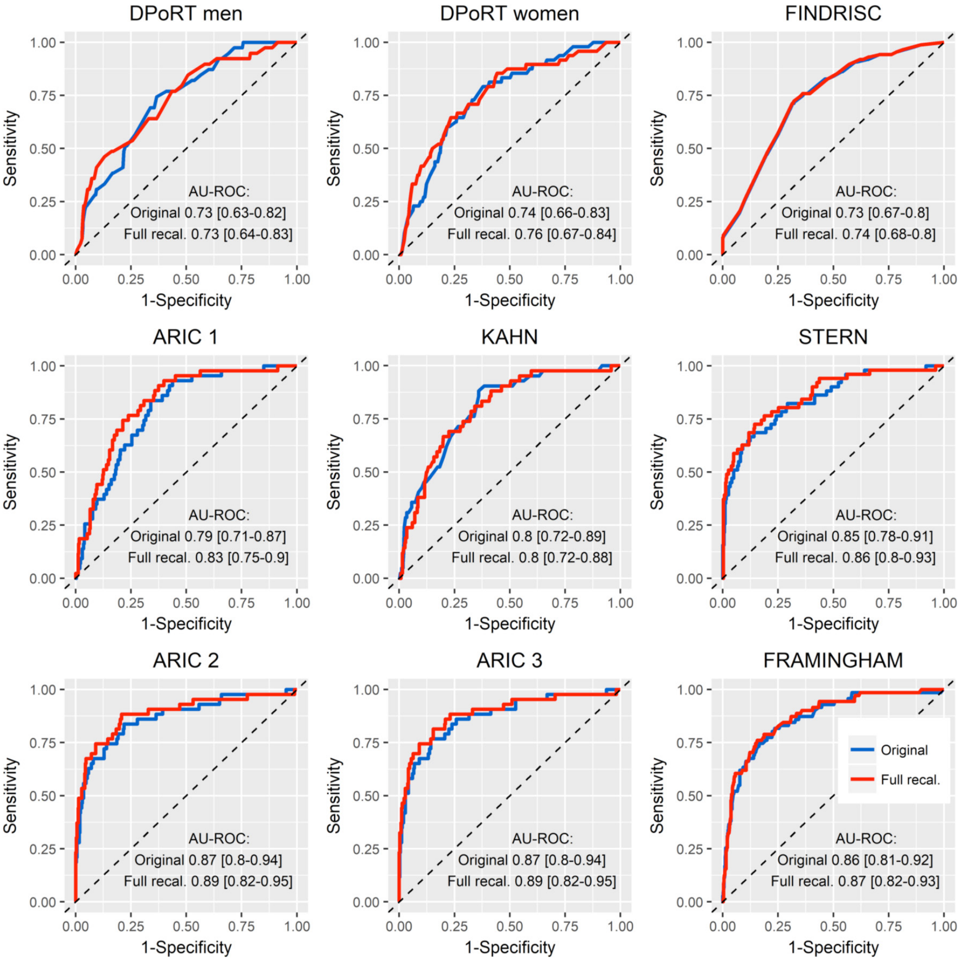

Figure 2 ROC curve at 8 years on the MESA test set for the original literature models (blue) and the models after full recalibration is performed on the MESA training set (red). In a setting in which the subjects with risk scores above a certain threshold T are predicted to develop T2D within a certain time t from the baseline, the ROC curve represents the plot of the true positive rate (sensitivity) versus the false positive rate (1-specificity) for different values of the threshold T. The greater the AU-ROC, the more accurately the score discriminates between subjects at high versus low risk. ARIC 1, Atherosclerosis Risk in Communities simple model; ARIC 2, ARIC clinical model without lipids; ARIC 3, ARIC clinical model with lipids; AU-ROC, area under the receiver operating characteristic curve; DPoRT, Diabetes Population Risk Tool; FINDRISC, Finnish Diabetes Risk Score; FRAMINGHAM, Framingham model; MESA, Multi-Ethnic Study of Atherosclerosis; recal, recalibration; T2D, type 2 diabetes.

(table 1). Indeed, for most participants there is at least one applicable model with the available data, thus the combined T2D model can provide a valid T2D risk score for these subjects.

In terms of C-index, the performance of the combined model is much better than that of the models of Sc1 and Sc2 and comparable with that of the models of Sc3. In particular, in the MESA test set, where the level of missing predictions for Sc3 models is modest, the combined T2D model achieves equivalent discrimination performance to the best-performing literature model, which in the case of the MESA test set is ARIC 3, but in general on a new previously unseen population is unknown a priori. This is still valid if the analysis is restricted to the subset of subjects without missing predictions (n=347; table 2). Regarding the ELSA test set, the combined model cannot reach the performance of the best model (FRAMINGHAM), as the level of missing predictions for this model is very high (64%). However, as visible in table 2, if the C-index computation is restricted to the subset of test set subjects to whom all the literature models can be applied (n=2340), the combined model reaches the discrimination performance of FRAMINGHAM, confirming that the averaging of multiple scores does not deteriorate the performance of the combined model. The 8-year ROC curves obtained on the MESA test set and the ELSA test set confirm that the combined model presents similar discrimination performance to the models of Sc3 (figure 4, left panels).

Thanks to the rescaling step, the combined model is also well calibrated, with an E/O very close to 1 both on the MESA and ELSA test set (table 1). In particular, both on the MESA and ELSA test set, the combined T2D model achieves E/O values comparable to those of the rescaled model with best calibration performance, that is ARIC 3 on the MESA test set and FINDRISC on the ELSA

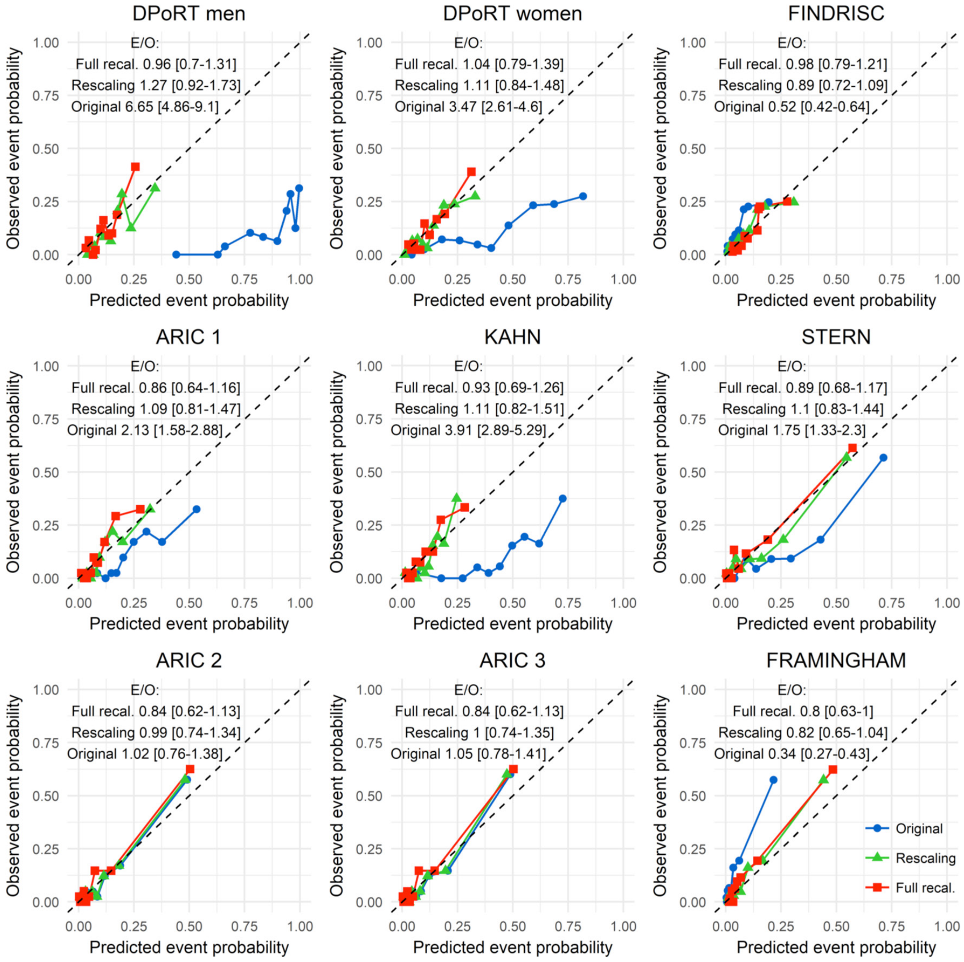

Figure 3 Calibration plot at 8 years on the MESA test set for the original literature models (blue), the rescaled models (green), and the fully recalibrated models (red). The calibration plot represents the number of observed events versus the number of expected events, at time t, for increasing deciles of predicted event probability. The more the calibration plot is close to the line with 0 intercept and 45° slope, the better the model is calibrated. ARIC 1, Atherosclerosis Risk in Communities simple model; ARIC 2, ARIC clinical model without lipids; ARIC 3, ARIC clinical model with lipids; DPoRT, Diabetes Population Risk Tool; E/O, expected to observed event ratio; FINDRISC, Finnish Diabetes Risk Score; FRAMINGHAM, Framingham model; MESA, Multi-Ethnic Study of Atherosclerosis; recal, recalibration.

test set, but in general is not known a priori. The good calibration performance of the combined model is also visible from the 8-year calibration plot, which is very close to the diagonal line both on the MESA and the ELSA test sets (figure 4, right panels).

Some additional analyses to test the robustness of the combined model approach were performed and reported in the online supplementary material for sake of space. In particular, a comparison of different weighting strategies performed on the MESA test set showed that the performance of the combined model is not sensitive to variations of the weights, as long as higher weights are used for higher scenario's models (online supplementary table S2). Empirical confidence intervals for all the performance metrics (online supplementary table S3) confirmed the results obtained on the MESA/ELSA test sets are not sensitive to the particular test set choice. Finally, we tested the performance of the combined T2D model considering only the models of Sc1 and Sc2, to mimic the case in which blood test results are not available for any individual. In this configuration, the combined model performs similarly to Sc2's models on both MESA (C-index=0.74 (0.70–0.78), E/O=1.05 (0.85–1.29)), and ELSA (C-index=0.74 (0.71–0.76), E/O=1.18 (1.05–1.33)) test sets.

DISCUSSION

Potential applications of T2D predictive models

T2D predictive models have a variety of potential users, apart from researchers. General practitioners can use T2D predictive models to identify the patients who are at risk of developing T2D and provide them recommendation on healthy behavioral changes or prescribing specific tests if undiagnosed diabetes is suspected. Local health departments can take advantage of T2D predictive models to identify local communities with increased risk of diabetes and develop policies to mitigate this risk. Health departments may use information about diabetes risk in different regions to make health cost analysis

Table 1 Performance of the literature models (original and rescaled) and the combined T2D model on the MESA and ELSA test sets

Test set	Scenario	Model	C-index*	E/O† original model	E/O† rescaled model	Missing predications‡
MESA	Sc1	DPoRT men	0.70 (0.64–0.77)	6.65 (4.86–9.10)	1.27 (0.92–1.73)	1%
		DPoRT women	0.70 (0.63–0.76)	3.47 (2.61–4.60)	1.11 (0.84–1.48)	
		FINDRISC	0.72 (0.67–0.76)	0.52 (0.42–0.64)	0.89 (0.72–1.09)	0%
	Sc2	ARIC 1	0.73 (0.68–0.78)	2.13 (1.58–2.88)	1.09 (0.81–1.47)	45%
		KAHN§	0.75 (0.69–0.81)	3.91 (2.89–5.29)	1.11 (0.82–1.51)	46%
	Sc3	STERN	0.81 (0.75–0.86)	1.75 (1.33–2.30)	1.10 (0.83–1.44)	42%
		ARIC 2	0.82 (0.76–0.87)	1.02 (0.76–1.38)	0.99 (0.74–1.34)	45%
		ARIC 3	0.83 (0.77–0.88)	1.05 (0.78–1.41)	1.00 (0.74–1.35)	45%
		FRAMINGHAM	0.83 (0.79–0.87)	0.34 (0.27–0.43)	0.82 (0.65–1.04)	17%
		Combined model	0.83 (0.79–0.87)		1.00 (0.81–1.24)	0%
ELSA	Sc1	DPoRT men	0.72 (0.67–0.76)	9.29 (7.60–11.36)	1.40 (1.14–1.71)	25%
		DPoRT women	0.71 (0.67–0.75)	3.63 (3.02–4.37)	1.23 (1.02–1.47)	
		FINDRISC	0.73 (0.70–0.77)	0.65 (0.57–0.74)	1.14 (0.99–1.30)	19%
	Sc2	ARIC 1	0.74 (0.72–0.77)	3.00 (2.61–3.46)	1.40 (1.22–1.62)	34%
		KAHN§	0.76 (0.73–0.79)	5.30 (4.57–6.15)	1.41 (1.22–1.64)	38%
	Sc3	STERN	0.79 (0.74–0.83)	2.46 (1.99–3.04)	1.81 (1.46–2.23)	64%
		ARIC 2	0.77 (0.73–0.82)	1.67 (1.35–2.05)	1.58 (1.29–1.95)	63%
		ARIC 3	0.80 (0.76–0.84)	1.60 (1.30–1.97)	1.56 (1.27–1.92)	64%
		FRAMINGHAM	0.82 (0.79–0.86)	0.39 (0.32–0.44)	1.22 (1.00–1.49)	64%
		Combined model	0.77 (0.74–0.79)		1.17 (1.04–1.31)	4%

*C-index varies between 0 and 1, with 0.5 corresponding to a random assignment of the scores and 1 representing the perfect score.

†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 overestimate/underestimate the event probability.

‡Percentage of missing predictions, that is, percentage of subjects for whom the model cannot return a valid risk score.

§Note that in Ref. 25, only the risk scoring system derived from the Weibull model was reported, whereas the parameters of the original Weibull model were not published. Therefore, to obtain the probability scores for this model, we divided the KAHN's risk scores (range 0–100) by 100.

ARIC 1, Atherosclerosis Risk in Communities simple model; ARIC 2, ARIC clinical model without lipids; ARIC 3, ARIC clinical model with lipids; C-index, concordance index; DPoRT, Diabetes Population Risk Tool; ELSA, English Longitudinal Study of Ageing ; E/O, expected to observed event ratio; FINDRISC, Finnish Diabetes Risk Score; FRAMINGHAM, Framingham model; MESA, Multi-Ethnic Study of Atherosclerosis; Sc, scenario; T2D, type 2 diabetes.

and make informed decisions about the distribution of health resources. Finally, T2D predictive models can be integrated in mobile health (mHealth) apps as tools to provide the users with a feedback about their health status (eg, a diabetes risk indicator) and recommendations on how to improve it, also resorting to gamification strategies to promote healthy behaviors.[34] Although the many potential users, T2D predictive models are scarsely adopted in real-world applications.

Limitations of using a single existing T2D predictive model

When applying T2D predictive models in real-world applications, the investigator (researcher, clinician, public health officer, or app designer) has to struggle with some practical issues. First, the investigator needs to choose the model that is most suitable for the target population. For example, some models can only be applied to specific ethnic groups, because they contain a variable race/ethnicity whose possible values (eg, black/white) can be restrictive for the application of interest. If the population considered is heterogeneous, like the MESA population, the investigator is forced to choose a model that does not take race/ethnicity as an input, although this might imply the choice of a model with lower performance.

Then, the investigator must make sure all the information required by the selected model is available for all the individuals. Missing values can represent a big issue when T2D predictive models are applied for a screening of the general population, for example, by a local health department, or integrated on mHealth apps, where data are self-provided by the user. In such applications, invasively collected biomarkers (required by Sc3 models) are frequently missing. Indeed, healthy individuals of the general population may not have recent blood test

Table 2 Discrimination performance of the literature models and the combined T2D model on the subset of MESA and ELSA test sets without missing predictions (ie, all the models can be applied without missing values in the input variables)

		C-index	
Scenario	Model	MESA test set	ELSA test set
Sc1	DPoRT men	0.79 (0.70–0.88)	0.72 (0.62–0.81)
	DPoRT women	0.68 (0.57–0.79)	0.75 (0.67–0.83)
	FINDRISC	0.76 (0.69–0.83)	0.75 (0.69–0.80)
Sc2	ARIC 1	0.77 (0.69–0.84)	0.73 (0.67–0.83)
	KAHN	0.78 (0.70–0.86)	0.74 (0.69–0.80)
Sc3	STERN	0.80 (0.72–0.89)	0.80 (0.75–0.85)
	ARIC 2	0.82 (0.74–0.90)	0.79 (0.73–0.85)
	ARIC 3	0.84 (0.76–0.91)	0.81 (0.76–0.86)
	FRAMINGHAM	0.83 (0.75–0.90)	0.82 (0.77–0.87)
	Combined model	0.84 (0.76–0.91)	0.83 (0.78–0.87)

ARIC 1, Atherosclerosis Risk in Communities simple model; ARIC 2, ARIC clinical model without lipids; ARIC 3, ARIC clinical model with lipids; C-index, concordance index; DPoRT, Diabetes Population Risk Tool; ELSA, English Longitudinal Study of Ageing; FINDRISC, Finnish Diabetes Risk Score; FRAMINGHAM, Framingham model; MESA, Multi-Ethnic Study of Atherosclerosis; Sc, scenario; T2D, type 2 diabetes.

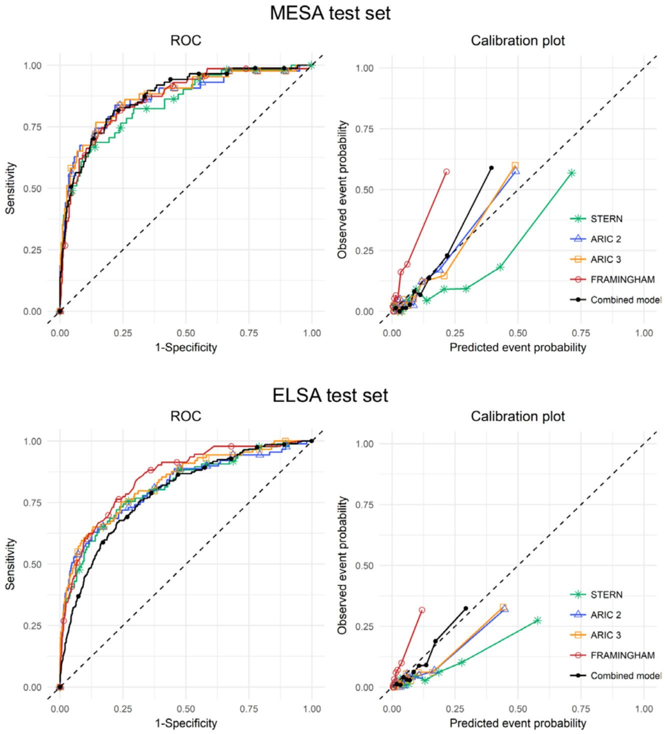

Figure 4 Receiver operating characteristic (ROC) curve and calibration plot at 8 years on the Multi-Ethnic Study of Atherosclerosis (MESA) test set (top panel) and the English Longitudinal Study of Ageing (ELSA) test set (bottom panels) for the combined type 2 diabetes model (black) and the original models of scenario 3 (STERN in green, Atherosclerosis Risk in Communities clinical model without lipids (ARIC 2) in blue, Atherosclerosis Risk in Communities clinical model with lipids (ARIC 3) in orange, Framingham model (FRAMINGHAM) in red).

results. Missing values can occur also with non-invasively collected data, especially with waist circumference (not as common as body mass index for measuring obesity) or family history of diabetes (some people may not know if their parents or siblings had/have diabetes). If any of the required inputs is missing for an individual, then the investigator has two choices: either not calculating the risk score (this generates a missing prediction) or imputing the missing values. The simpler imputation method consists in using a population average, for continuous variables, or the most frequent value in the population, for categorical variables, but this method obviously bring to a deterioration of the model performance, with underestimation of diabetes risk in at-risk individuals (see online supplementary table S4). More sophisticated imputation algorithms exist but they require the availability of a training set, they are computationally demanding and typically they can be applied only with missing values at random (ie, if the variables related to blood test are missing because blood test was not perform, then those missing values are not random and should not be imputed). Thus, missing data imputation is not always possible in practical applications.

Finally, the application of a model to a population different from the one in which the model was developed often drives to poor incidence prediction performance because the model could be not well calibrated for the target population.

Addressing practical issues by the combined T2D model approach

The purpose of this work was not to develop a new T2D predictive model, but to address the practical issues the investigators must face when applying an existing model to a new population, in particular the issues of model choice, missing values, and recalibration. This paper proposes a strategy that combines eight existing T2D models by risk score rescaling and averaging, resulting in a combined model for T2D onset prediction. It is important to remark that the combined model is not a new model trained on a specific population, rather it is a combination of multiple existing models, trained on different populations, designed to extend the applicability of the existing T2D predictive models. This approach is different from other ensemble learning approaches for T2D onset prediction,[35–37] in which a set of new models, typically all using the same variables, are developed and combined to provide the final prediction.

The combined model requires in input the subject's variables (missing values are allowed) and the expected T2D incidence in the target population. Then, the combined model automatically selects for each subject the applicable models based on the subject's ethnicity and excludes the theoretically applicable models for which some variables are missing. Then, the combined model calculates the risk scores of the remaining models, rescales them according to the expected diabetes incidence, and finally computes the weighted average of the rescaled risk scores, using different weights based on the level of information the models require.

In summary, the combined model automatically manages the issues of model choice, missing values, and recalibration, thus facilitating the investigators when applying existing models to new populations, especially when these are very heterogeneous.

Why using the combined model? High prediction performance with high coverage

The combined model was assessed on two external datasets, MESA and ELSA, not used for training any of the models included in the combined model. The two datasets present different characteristics: MESA includes a multiethnic US population, while ELSA includes a mostly white English population.

On the MESA dataset, it was demonstrated that risk score rescaling allows to achieve good calibration performance, comparable to that of the full recalibration. Despite the simple risk score rescaling and the full recalibration perform similarly, the former is much more easy to apply than the latter because it does not require the re-estimation of model coefficients. Indeed, to perform the risk score rescaling, the only information required is prior knowledge on the T2D incidence in the target population (eg, derived from historical data). Conversely, to implement a full recalibration, that is, re-estimating all coefficients, it is mandatory to have a sufficiently large training set containing information on all the predictors at a baseline time and knowledge on the T2D incidence during a sufficiently long follow-up period, information often unavailable in real-world applications. For this reason, the risk score rescaling was embedded in the combined model, as a much more practical, but equally effective, recalibration procedure than the full recalibration.

The benefits of using the combined T2D model over single predictive models were assessed on both MESA and ELSA datasets, confirming in both cases that the combined T2D model has better generalization ability than single predictive models. Indeed, considering the entire MESA and ELSA test sets, the combined T2D model was able to achieve comparable performance to the models of Sc3 (the best-performing existing models) but with almost zero missing predictions. On the MESA test set, the model with best coverage was FINDRISC (0% missing predictions) which presented C-index=0.72, while the models with highest C-index were the Sc3 models (C-index=0.81–0.83) that, however, presented a percentage of missing prediction between 17% and 45%. The combined model was able to achieve both high coverage (0% missing predictions), like FINDRISC, and high discrimination performance (C-index=0.83), like the Sc3 models. The same result was obtained with the ELSA dataset, in which achieving good prediction performance with high coverage was even more challenging because of the high prevalence of missing values. In such a challenging scenario, the existing model with best coverage (FINDRISC) presented 19% missing predictions and C-index=0.72; the combined model was able to achieve an excellent coverage (4% missing predictions) with discrimination performance comparable to the Sc3 models (C-index=0.77 vs C-index=0.77–0.82).

When all the models could be applied, the discrimination performance of the combined model was equivalent to that of the model with best discrimination performance (in this case ARIC 3 for MESA, FRAMINGHAM for ELSA). The combined T2D model also presented calibration performance similar to the best-performing rescaled model (in this case ARIC 3 for MESA, FINDRISC for ELSA), thus averaging the models' score does not negatively affect the discrimination and calibration performance of the combined model. Note that in general the model that will show the best performance on a new population is not known a priori. Thus, another important advantage of the combined model is that it allows to achieve the performance of the best model, without knowing it a priori.

Limitations of current study and future developments

A possible limitation of this study is that MESA and ELSA datasets differ for the definition of the diabetes outcome, defined by the ADA 2003 fasting criteria algorithm in MESA and self-reported in ELSA. This could reflect on the slightly worse performance on ELSA dataset, as the outcome definition in this dataset is suboptimal, but overall it shows the robustness of our approach to different datasets and data collection procedures. Note that while in general the use of self-reported diabetes diagnosis as an outcome might lead to misclassification of undiagnosed T2D subjects, this was not an issue for the ELSA dataset, as only the 0.9% of the subjects that exited the study without reporting diabetes actually met the ADA 2003 fasting criteria algorithm, that is, they might have had undiagnosed diabetes.

The encouraging results achieved in this proof-of-concept study for the ELSA and MESA datasets may be not necessarily true for other cohorts, calling for the need of a broader validation of the methodology. Future work will include the validation of the combined T2D model on other datasets. Moreover, while in this proof-of-concept study the combined model included only eight literature models, which were selected because they could be applied with the MESA and ELSA datasets, in future work the combined model can be easily extended by incorporating other literature models, provided that the information required by such models is available in the target population. Finally, the methodology adopted to develop the combined T2D model is general and can be applied to the prediction of other diseases such as cardiovascular diseases[38] and chronic respiratory diseases.

The R code implementing the combined T2D model, presented in this article, is made available on request for investigators willing to apply the model on their own dataset. As a future development, a web application implementing the combined T2D model will be developed and published.

Final remarks

This work has been performed within the Participatory Urban Living for Sustainable Environment (PULSE) project,[39] a H2020 project funded by the European Commission, focused on big data analytics for monitoring citizens' health and promoting positive behavioral changes. Currently, we are living in the era of the internet of things in which a huge amount of health data are collected by wearable sensors and mobile apps.[40] In such a scenario, methods like the combined T2D model proposed in this paper, that allows to automatically handle the issues related to previously unseen data, missing data, model choice and model generalization, are fundamental to guarantee a successful application of existing models of T2D onset.

Author affiliations

[1]Department of Information Engineering, School of Engineering, University of Padova, Padova, Italy
[2]Icahn School of Medicine at Mount Sinai, New York, New York, USA
[3]Department of Public Health Policy and Management, New York University, New York, New York, USA
[4]Center for Health Innovation, New York Academy of Medicine, New York, New York, USA
[5]Research, Evaluation & Policy, New York Academy of Medicine, New York, New York, USA
[6]Department of Preventive Medicine, Northwestern University, Chicago, Illinois, USA
[7]Division of Public Health Sciences, Wake Forest University Health Sciences, Winston-Salem, North Carolina, USA

Presented at

Part of this study was presented in abstract form at the 12th International Conference on Advanced Technologies & Treatments for Diabetes, Berlin, Germany, February 20–23, 2019.

Acknowledgements The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions.

Contributors MV conceived and designed the work, analyzed and interpreted the data, and drafted and critically revised the manuscript. EL and AZ analyzed and interpreted the data, and critically revised the manuscript. BDC conceived the work, interpreted the data, drafted, and critically revised the manuscript. AF, YL, JAP, DS, MRC, and AGB interpreted the data and critically revised the manuscript. BDC is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding The MESA study was supported by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by grants UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 from the National Center for Advancing Translational Sciences (NCATS). The ELSA study is funded by the National Institute on Aging (R01AG017644), and a consortium of UK government departments coordinated by the National Institute for Health Research. This research was funded by: the European Commission within the Horizon 2020 PULSE project, ID 727816, from January 11, 2016 to October 31, 2019; MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethic approval was not required for this study because this study analysed datasets collected by previous studies (MESA and ELSA) which were available to the authors of this paper in anonymized form, without any link to the real subjects.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data are available from the website of the MESA study (https://www.mesa-nhlbi.org/Publications.aspx) and the ELSA study (https://www.elsa-project.ac.uk/accessing-elsa-data) on reasonable request.

Author note The ELSA was developed by a team of researchers based at the University College London, NatCen Social Research, and the Institute for Fiscal Studies. The data were collected by NatCen Social Research.

ORCID iD

Barbara Di Camillo http://orcid.org/0000-0001-8415-4688

REFERENCES

1 Knowler WC, Barrett-Connor E, Fowler SE, *et al*. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393–403.
2 Lindström J, Ilanne-Parikka P, Peltonen M, *et al*. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish diabetes prevention study. *Lancet* 2006;368:1673–9.
3 Li G, Zhang P, Wang J, *et al*. Cardiovascular mortality, all-cause mortality, and diabetes incidence after lifestyle intervention for people with impaired glucose tolerance in the dA Qing diabetes prevention study: a 23-year follow-up study. *Lancet Diabetes Endocrinol* 2014;2:474–80.
4 Diabetes Prevention Program Research Group, Knowler WC, Fowler SE, *et al*. 10-Year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. *Lancet* 2009;374:1677–86.
5 Diabetes Prevention Program Research Group. Long-Term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the diabetes prevention program outcomes study. *Lancet Diabetes Endocrinol* 2015;3:866–75.
6 American Diabetes Association. 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes-2019*. *Diabetes Care* 2019;42:S13–28.
7 Siu AL, U S Preventive Services Task Force. Screening for abnormal blood glucose and type 2 diabetes mellitus: U.S. preventive services Task force recommendation statement. *Ann Intern Med* 2015;163:861–8.
8 Noble D, Mathur R, Dent T, *et al*. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011;343:d7163.
9 Rosella LC, Manuel DG, Burchill C, *et al*. 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;65:613–20.
10 Hippisley-Cox J, Coupland C, Robson J, *et al*. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880.
11 Wilson PWF, Meigs JB, Sullivan L, *et al*. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham offspring study. *Arch Intern Med* 2007;167:1068–74.
12 Di Camillo B, Hakaste L, Sambo F, *et al*. HAPT2D: high accuracy of prediction of T2D with a model combining basic and advanced data depending on availability. *Eur J Endocrinol* 2018;178:331–41.

13 National Institute for Health and Care Excellence. Type 2 diabetes: prevention in people at high risk [article online], 2012. Available: https://www.nice.org.uk/guidance/ph38/resources/type-2-diabetes-prevention-in-people-at-high-risk-pdf-1996304192197 [Accessed 28 May 2019].

14 Nowak C, Ingelsson E, Fall T. Use of type 2 diabetes risk scores in clinical practice: a call for action. *Lancet Diabetes Endocrinol* 2015;3:166–7.

15 Martinez-Millana A, Argente-Pla M, Valdivieso Martinez B, *et al*. Driving type 2 diabetes risk scores into clinical practice: performance analysis in hospital settings. *J Clin Med* 2019;8:107.

16 Moons KGM, Kengne AP, Grobbee DE, *et al*. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart* 2012;98:691–8.

17 Abbasi A, Peelen LM, Corpeleijn E, *et al*. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* 2012;345:e5900.

18 Vettoretti M, Longato E, Camillo BD, *et al*. Importance of recalibrating models for type 2 diabetes onset prediction: application of the Diabetes Population Risk Tool on the health and retirement study. *Conf Proc IEEE Eng Med Biol Soc* 2018;2018:5358–61.

19 Bild DE, Bluemke DA, Burke GL, *et al*. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol* 2002;156:871–81.

20 American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2011;34 Suppl 1:S62–9.

21 Steptoe A, Breeze E, Banks J, *et al*. Cohort profile: the English longitudinal study of ageing. *Int J Epidemiol* 2013;42:1640–8.

22 Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* 2003;26:725–31.

23 Schmidt MI, Duncan BB, Bang H, *et al*. Atherosclerosis Risk in Communities Investigators. identifying individuals at high risk for diabetes: the Atherosclerosis Risk in Communities study. *Diabetes Care* 2005;28:2013–8.

24 Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? *Ann Intern Med* 2002;136:575–81.

25 Kahn HS, Cheng YJ, Thompson TJ, *et al*. Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years. *Ann Intern Med* 2009;150:741–51.

26 Janssen KJM, Vergouwe Y, Kalkman CJ, *et al*. A simple method to adjust clinical prediction models to local circumstances. *Can J Anaesth* 2009;56:194–201.

27 Steyerberg EW. Updating for a New Setting. In: *Clinical prediction models: a practical approach to development, validation, and updating*. Rotterdam: Springer, 2009: 361–90.

28 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.

29 Harrell FE, Califf RM, Pryor DB, *et al*. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–6.

30 Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004;23:2109–23.

31 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.

32 Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Stat Methods Med Res* 2016;25:1692–706.

33 Rockhill B, Spiegelman D, Byrne C, *et al*. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* 2001;93:358–66.

34 Ottaviano M, Beltrán-Jaunsarás ME, Teriús-Padrón JG, *et al*. Empowering citizens through perceptual sensing of urban environmental and health data following a participative citizen science approach. *Sensors* 2019;19:2940.

35 Anderson JP, Parikh JR, Shenfeld DK, *et al*. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *J Diabetes Sci Technol* 2015;10:6–18.

36 Liu Y, Ye S, Xiao X, *et al*. Machine learning for tuning, selection, and ensemble of multiple risk scores for predicting type 2 diabetes. *Risk Manag Healthc Policy* 2019;12:189–98.

37 Nguyen BP, Pham HN, Tran H, *et al*. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Programs Biomed* 2019;182:105055.

38 Damen JAAG, Hooft L, Schuit E, *et al*. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.

39 PULSE. Participatory Urban Living for Sustainable Environments [online], 2019. Available: http://www.project-pulse.eu/ [Accessed 19 Mar 2019].

40 Vettoretti M, Cappon G, Acciaroli G, *et al*. Continuous glucose monitoring: current use in diabetes management and possible future applications. *J Diabetes Sci Technol* 2018;12:1064–71.