Online-Only Supplemental Material

Table S1. Characteristics of the subjects of MESA and ELSA datasets at baseline. Relative frequencies [%] are reported for categorical variables, mean (standard deviation) for continuous variables. The percentage of missing values in each dataset is also reported. The last column shows the models that use the variables.

	Variable	MESA dataset		ELSA dataset		Used by
		Relative freq. or mean (SD)	Percentage of missing values	Relative freq. or mean (SD)	Percentage of missing values	
Categorical variables	Male sex	47%	0%	45%	0%	DPoRT, FINDRISC, KAHN, STERN, FRAMINGHAM
	White ethnicity	40%	0%	98%	0%	DPoRT, ARIC1, KAHN, STERN, ARIC2, ARIC3
	African-American ethnicity	26%	0%	-	100%	ARIC1, KAHN, ARIC2, ARIC3
	Asian/Chinese ethnicity	12%	0%	-	100%	None
	Hispanic ethnicity	21%	0%	-	100%	STERN
	Non-white ethnicity	60%	0%	2%	0%	DPoRT
	Less than post-secondary education	50%	0.3%	83%	20%	DPoRT
	Born outside living country	32%	0.3%	8%	0.08%	DPoRT
	Have family history of diabetes	35%	4%	17%	22%	STERN
	Mother with diabetes	16%	9%	7%	21%	ARIC1, KAHN, ARIC2, ARIC3, FRAMINGHAM
	Father with diabetes	12%	15%	10%	22%	ARIC1, KAHN, ARIC2, ARIC3, FRAMINGHAM
	Currently smoking	15%	0.4%	15%	0.2%	DPoRT, KAHN
	Ever had hypertension	35%	0.3%	34%	0.01%	DPoRT
	Use of hypertension medication	32%	0.04%	15%	15%	FINDRISC, KAHN, FRAMINGHAM
	Ever had heart disease	0%	0%	14%	0.01%	DPoRT
Continuous variables	Age [years]	61.10 (10.17)	0%	63.41 (9.45)	0%	All
	BMI [Kg/m^2]	28.00 (5.31)	0%	27.86 (4.93)	4%	DPoRT, FINDRISC, STERN, FRAMINGHAM
	Waist circumference [cm]	96.94 (13.99)	0.02%	95.25 (16.10)	2%	FINDRISC, ARIC1, KAHN, ARIC2, ARIC3

	Height [cm]	166.56 (9.97)	0%	166.02 (9.45)	3%	ARIC1, KAHN, ARIC2, ARIC3
	Weight [Kg]	77.90 (16.92)	0%	76.98 (15.77)	3%	KAHN
	Resting heart rate [beats/min]	62.44 (9.27)	0.8%	57.40 (14.65)	11%	KAHN
	Systolic blood pressure [mmHg]	124.93 (20.88)	0.02%	133.18 (18.28)	11%	ARIC1, KAHN, STERN, ARIC2, ARIC3, FRAMINGHAM
	Diastolic blood pressure [mmHg]	71.94 (10.23)	0.02%	75.78 (10.83)	11%	KAHN, FRAMINGHAM
	Fasting glucose [mg/dL]	89.46 (10.47)	0%	88.82 (13.23)	50%	STERN, ARIC2, ARIC3, FRAMINGHAM
	HDL cholesterol [mg/dL]	51.63 (14.92)	0.08%	60.35 (15.58)	21%	STERN, ARIC3, FRAMINGHAM
	Triglycerides concentration [mg/dL]	127.24 (77.50)	0.04%	155.24 (97.64)	21%	ARIC3, FRAMINGHAM

Comparison of different scenarios' weights

We assessed the impact of the weights' choice on the combined model performance, comparing different weighting schemas on the MESA test set. In order to have each model contributing to the weighted average for each subject, the analysis was focused on the subset of subjects without missing predictions. Moreover, to have the three scenarios contributing with an equal number of models to the weighted average, only one model for each scenario was considered, i.e DPoRT (Sc1), KAHN (Sc2) and FRAMINGHAM (Sc3). In this setting, the configurations of model weights reported in Table S2 were tested.

The results show that there is no a drastic difference between the performance achieved with different weights configurations in terms of C-index. Using equal weights for all the models (configuration 1-1-1) drives to the worst performance, i.e. C-index=0.8, the use of only two weights (configurations 0.25-0.25-0.5 and 0.2-0.4-0.4) results in C-index=0.82, while best results are achieved when a different weight is used for each scenario and in this case C-index was 0.83 with almost all the configurations tested.

From this analysis, we can conclude that the performance of the combined model are not that sensitive to the specific model weights, provided that different weights are used for each Sc and higher weights are used for higher-Sc models.

Table S2. Discrimination performance of the combined model for different weight settings. Results are shown in terms of C-index [95% confidence interval] calculated in the subset of MESA test set data without missing predictions.

Model weights			**C-index Subset without missing**
Sc1	**Sc2**	**Sc3**	
1	1	1	0.81 [0.74-0.89]
0.25	0.25	0.5	0.82 [0.75-0.90]
0.2	0.4	0.4	0.82 [0.75-0.90]
1 (0.17)	2 (0.33)	3 (0.5)	0.83 [0.75-0.90]
0.25	0.33	0.42	0.82 [0.75-0.90]
0.1	0.3	0.6	0.83 [0.76-0.91]
0.1	0.2	0.7	0.83 [0.76-0.91]
0.05	0.3	0.65	0.83 [0.76-0.91]
0.05	0.10	0.85	0.83 [0.76-0.91]

Empirical confidence intervals by bootstrap resampling

Empirical confidence intervals for performance metrics were computed by applying bootstrap resampling both on the MESA training set and ELSA reference set. In particular, 100 iterations were performed and, at each iteration, a new set of data was generated from the original one by sampling with replacement a number of elements equal to the original set cardinality. Then, this new set of data was used to estimate the 8-year T2D incidence and the respective out-of-bag sample was used for testing the 8 existing rescaled models and the combined model by computing C-index, E/O and missing

predictions. At the end of the 100 iterations, 100 values for each metric were obtained, from which the median, the 2.5% and the 97.5% percentiles were computed. Such empirical estimates of the metrics' median and 95% confidence interval are reported in Table S3. These results confirm the ones obtained on the MESA and ELSA test sets, i.e. the results we obtained are not sensitive to the specific test set choice.

Table S3. Performance of the rescaled literature models and the combined model on 100 out-of-bag sets generated by bootstrap resampling on the MESA training set and the ELSA reference set. Results reported as median [95% confidence interval].

Test set	Scenario	Model	Results on training set with bootstrap conf. intervals		
			C-index	E/O Rescaled model	Missing pred.
MESA	Sc1	DPoRT men	0.67 [0.62-0.72]	1.30 [1.05-1.75]	1% [0-1]%
		DPoRT women	0.69 [0.65-0.75]	1.07 [0.88-1.44]	
		FINDRISC	0.68 [0.65-0.72]	0.89 [0.75-1.20]	0% [0-0]%
	Sc2	ARIC 1	0.71 [0.66-0.75]	1.15 [0.91-1.77]	43% [41-45]%
		KAHN	0.73 [0.68-0.77]	1.21 [0.93-1.86]	46% [44-47]%
	Sc3	STERN	0.81 [0.76-0.83]	1.23 [1.04-1.68]	41% [39-42]%
		ARIC 2	0.83 [0.80-0.90]	1.09 [0.88-1.61]	43% [41-45]%
		ARIC 3	0.83 [0.80-0.86]	1.09 [0.88-1.61]	43% [41-45]%
		FRAMIN-GHAM	0.78 [0.75-0.81]	0.84 [0.72-1.16]	16% [15-17]%
	Combined model		0.80 [0.77-0.83]	1.01 [0.86-1.31]	0% [0-0]%
ELSA	Sc1	DPoRT men	0.75 [0.71-0.80]	1.00 [0.82-1.24]	23% [22-25]%
		DPoRT women	0.73 [0.68-0.80]	1.49 [1.15-2.05]	
		FINDRISC	0.75 [0.71-0.78]	1.03 [0.86-1.32]	20% [18-21]%
	Sc2	ARIC 1	0.75 [0.71-0.78]	1.37 [1.08-1.68]	35% [32-36]%

		KAHN[d]	0.75 [0.72-0.79]	1.37 [1.09-1.71]	40% [38-42]%
	Sc3	STERN	0.81 [0.75-0.85]	1.55 [1.26-2.03]	63% [61-64]%
		ARIC 2	0.79 [0.72-0.84]	1.42 [1.13-1.85]	62% [60-64]%
		ARIC 3	0.81 [0.75-0.85]	1.40 [1.12-1.82]	63% [61-64]%
		FRAMIN-GHAM	0.83 [0.80-0.87]	1.08 [0.85-1.46]	63% [61-64]%
	Combined model		0.79 [0.76-0.81]	1.10 [0.92-1.33]	4% [3-5]%

Performance of the original models with imputation of 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. This method, however, brings to a deterioration of the model performance, with underestimation of diabetes risk in at risk individuals. This is visible in Supplementary Table S4, that reports the performance of the 8 literature models on the MESA and ELSA test set by replacing the missing values with the mean/mode values observed for each variable on the training/reference set. In particular, on the MESA test set, where the percentage of missing values is low (see Table S1), the deterioration of discrimination performance is small. For example, the C-index for Sc3 models drops from 0.81-0.83 without missing values imputation to 0.80-0.82 with missing values imputation. However, even with missing value imputation, most of the models of Sc2 and Sc3 still present a large percentage of missing predictions (34-40%), due to the fact that these models can be applied only to specific ethnic groups. On the ELSA test, where the level of missing values is higher (see Table S1), a larger deterioration of discrimination performance is obtained with missing data imputation. For example, the C-index for Sc3 models drops from 0.77-0.82 without missing values imputation to 0.73-0.77 with missing values imputation.

Table S4. Performance of the rescaled literature models on the MESA and ELSA test sets after substituting the missing variables with the mean values, for continuous variables, or the most frequent value, for categorical variables, obtained from the MESA training set and the ELSA reference set.

Test set	Scenario	Model	Results with missing values imputation		
			C-index	E/O Rescaled model	Missing pred.
MESA	Sc1	DPoRT men	0.71 [0.64-0.77]	1.27 [0.93-1.74]	0%
		DPoRT women	0.69 [0.63-0.76]	1.13 [0.85-1.50]	
		FINDRISC	0.72 [0.67-0.76]	0.89 [0.72-1.09]	0%
	Sc2	ARIC 1	0.70 [0.65-0.76]	1.11 [0.84-1.46]	34%
		KAHN	0.73 [0.68-0.79]	1.13 [0.86-1.50]	34%
	Sc3	STERN	0.81 [0.76-0.86]	1.06 [0.81-1.39]	40%
		ARIC 2	0.80 [0.75-0.85]	1.03 [0.78-1.36]	34%
		ARIC 3	0.81 [0.76-0.86]	1.05 [0.80-1.39]	34%
		FRAMIN-GHAM	0.82 [0.78-0.86]	0.80 [0.65-0.99]	0%
ELSA	Sc1	DPoRT men	0.72 [0.68-0.76]	1.29 [1.09-1.54]	0%
		DPoRT women	0.70 [0.67-0.74]	1.14 [0.98-1.33]	
		FINDRISC	0.71 [0.69-0.74]	1.03 [0.92-1.16]	0%
	Sc2	ARIC 1	0.73 [0.71-0.75]	1.21 [1.07-1.36]	2%
		KAHN	0.73 [0.71-0.75]	1.21 [1.07-1.36]	2%
	Sc3	STERN	0.73 [0.71-0.76]	1.13 [1.01-1.27]	2%
		ARIC 2	0.75 [0.72-0.78]	1.02 [0.90-1.14]	2%
		ARIC 3	0.77 [0.74-0.79]	1.06 [0.95-1.20]	2%
		FRAMIN-	0.74	0.83	0%

		GHAM	[0.71-0.77]	[0.74-0.93]	

The rescaling method for model recalibration

In the combined T2D model, the 8 selected models were recalibrated by the rescaling method adopted in work by Kengne et al. (1). In particular, for a logistic regression model of equation:

$$y(\boldsymbol{X_i}) = \frac{e^{\boldsymbol{X_i}}}{1 + e^{\boldsymbol{X_i}}}$$

where $\boldsymbol{X_i}$ is the linear regression of the variables of subject i, the rescaled model equation is obtained as:

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

where φ is a correction factor (2) calculated 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:

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

For survival models, the rescaled incident risk score, $y'(\boldsymbol{X_i})$, for follow-up time *t*, is calculated as (3):

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

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

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

Specifically, in our implementation the rescaling was performed for a follow-up time of 8 years.

References:

(1) Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, Spijkerman AM, Griffin SJ, Grobbee DE, Palla L, Tormo MJ, Arriola L, Barengo NC, Barricarte A, Boeing H, Bonet C, Clavel-Chapelon F, Dartois L, Fagherazzi G, Franks PW, Huerta JM, Kaaks R, Key TJ, Khaw KT, Li K, Mühlenbruch K, Nilsson PM, Overvad K, Overvad TF, Palli D, Panico S, Quirós JR, Rolandsson O, Roswall N, Sacerdote C, Sánchez MJ, Slimani N, Tagliabue G, Tjønneland A, Tumino R, van der A DL, Forouhi NG, Sharp SJ, Langenberg C, Riboli E, Wareham NJ. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol 2014; 2(1):19-29.

(2) Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth 2009; 56: 194–201.

(3) Steyerberg EW. Updating for a New Setting. In *Clinical prediction models: a practical approach to development, validation, and updating*. Rotterdam, Springer, 2009, p. 361-390.