JDI Journal of Diabetes Investigation Open access
Official Journal of the Asian Association for the Study of Diabetes
ORIGINAL ARTICLE

Simple risk score to screen for prediabetes: A cross-sectional study from the Qatar Biobank cohort

Mostafa Abbas[1,2], Raghvendra Mall[1], Khaoula Errafii[3,4], Abdelkader Lattab[1], Ehsan Ullah[1], Halima Bensmail[1*], Abdelilah Arredouani[3,4*]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar, [2]Department of Imaging Science and Innovation, Geisinger, Danville, Pennsylvania, USA, [3]Qatar Biomedical Research Institute, Hamad Bin Khalifa University, Doha, Qatar, and [4]College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar

Keywords
Prediabetes, Qatar Biobank, Risk score

***Correspondence**
Abdelilah Arredouani
Tel.: +974-445-42947
Fax: +974-445-41770
E-mail address:
aarredouani@hbku.edu.qa

Halima Bensmail
Tel.: +974-445-40195
Fax: +974-445-41770
E-mail address:
hbensmail@hbku.edu.qa

J Diabetes Investig 2021; 12: 988–997

doi: 10.1111/jdi.13445

ABSTRACT

Aims/Introduction: The progression from prediabetes to type 2 diabetes is preventable by lifestyle intervention and/or pharmacotherapy in a large fraction of individuals with prediabetes. Our objective was to develop a risk score to screen for prediabetes in the Middle East, where diabetes prevalence is one of the highest in the world.

Materials and Methods: In this cross-sectional, case–control study, we used data of 4,895 controls and 2,373 prediabetic adults obtained from the Qatar Biobank cohort. Significant risk factors were identified by logistic regression and other machine learning methods. The receiver operating characteristic was used to calculate the area under curve, cut-off point, sensitivity, specificity, positive and negative predictive values. The prediabetes risk score was developed from data of Qatari citizens, as well as long-term (≥15 years) residents.

Results: The significant risk factors for the Prediabetes Risk Score in Qatar were age, sex, body mass index, waist circumference and blood pressure. The risk score ranges from 0 to 45. The area under the curve of the score was 80% (95% confidence interval 78–83%), and the cut-off point of 16 yielded sensitivity and specificity of 86.2% (95% confidence interval 82.7–89.2%) and 57.9% (95% confidence interval 65.5–71.4%), respectively. Prediabetes Risk Score in Qatar performed equally in Qatari nationals and long-term residents.

Conclusions: Prediabetes Risk Score in Qatar is the first prediabetes screening score developed in a Middle Eastern population. It only uses risk factors measured non-invasively, is simple, cost-effective, and can be easily understood by the general public and health providers. Prediabetes Risk Score in Qatar is an important tool for early detection of prediabetes, and can help tremendously in curbing the diabetes epidemic in the region.

INTRODUCTION

Diabetes has become a global epidemic, and type 2 diabetes represents 90% of all diabetes cases[1]. Of great concern is that half of all people living with diabetes are undiagnosed[1], often leading to late diagnosis of the diabetes-associated macro- and microvascular complications. Furthermore, it is estimated that 7.5% of the world population has prediabetes[2,3]. Prediabetes is a condition where blood glucose concentrations are higher than normal, but not meeting the absolute definition of diabetes. Specifically, according to the American diabetes Association, prediabetes is characterized by a level of hemoglobin A1C (HbA_{1c}) between 39 and 47 mmol/mol (5.7 and 6.4%), a fasting plasma glucose between 5.6 and 6.9 mmol/L or a 2-h plasma glucose between 7.8 and 11 mmol/L. Prediabetes is a major independent risk factor for developing type 2 diabetes. In fact, prospective studies have shown that 5–10% of individuals with prediabetes progress to type 2 diabetes annually[4]. Additionally, whereas just 5% of individuals with normal glycemia will go on to develop type 2 diabetes, 33–65% of those with prediabetes will develop type 2 diabetes within 6 years[5]. The prevalence of prediabetes is projected to increase to 8.6% by 2045[2,3]. Consequently, the incidence of type 2 diabetes is set

Received 20 April 2020; revised 1 October 2020; accepted 6 October 2020

to worsen without preventive actions. A major challenge to addressing prediabetes is that it can be asymptomatic for a long time, as is type 2 diabetes for that matter. As such, it can silently and gradually damage vital organs, and thus pave the way for diabetes-associated micro- and macro-vascular complications.

Epidemiological studies from different Middle Eastern countries, especially the Gulf Cooperation Council nations, have reported that, like type 2 diabetes, prediabetes is highly prevalent, with reported rates ranging between 20 and 40%[6,7]. These alarming data raise great concern in the region given the above-mentioned annual conversion rates. Fortunately, many lines of recent scientific evidence have shown that the progression from prediabetes to type 2 diabetes can be prevented, or at least delayed, in a large fraction of individuals with prediabetes in response to intensive lifestyle intervention (ILI) or medication, such as with metformin[8–13]. The effect of ILI on reversing type 2 diabetes or preventing the conversion of prediabetes to type 2 diabetes in the Middle East has rarely been investigated. Recently, Alfawaz *et al.*[14] investigated the effects of different dietary and lifestyle modification therapies on metabolic syndrome in Saudi adults with prediabetes in a 12-month longitudinal study. They found that full metabolic syndrome in the ILI program group decreased by 26% compared with 8.2% in the general advice group. That study, despite the short follow-up period, and given the cultural, behavioral and ethnical similarities between the Middle Eastern populations, promisingly shows that ILI can potentially be effective in preventing type 2 diabetes in different countries in the region, and should, thus, be given more attention. Taken together, the available scientific evidence suggests that identifying individuals with prediabetes is crucial, and can be a highly cost-effective step in curbing the type 2 diabetes epidemic sweeping the Middle East region.

Currently, prediabetes is diagnosed with the 2-h oral glucose tolerance test, fasting plasma glucose or HbA_{1c} blood tests. However, these assays are invasive, inconvenient, might take long time and are relatively expensive, especially in poor countries. Furthermore, they are not suitable for routine screening programs in populations with a high prevalence of prediabetes or type 2 diabetes. Therefore, the fight against the relentless type 2 diabetes epidemic in the Middle East might take advantage of a tool that could facilitate screening by identifying a subset of the population for whom laboratory screening is required. To date, we are not aware of any prediabetes screening tool risk score that was developed for Middle Eastern populations. Given the burden of type 2 diabetes and prediabetes in the region, a simple, reliable and cost-effective screening method, such as a prediabetes risk score, which can be easily carried out in clinical or community settings by primary healthcare physicians and by the general public, is of significant clinical importance for the efforts undertaken to curb the diabetes epidemic. The aim of the present study was to build a prediabetes risk score and evaluate its performance for screening prediabetes using a cohort of adult Qatari nationals and long-term residents who lived in Qatar ≥15 years.

METHODS

Study population

We initially obtained clinical, anthropometric and demographic data of 7,386 individuals aged between 18 and to 86 years (6,000 Qatari and 1,383 long-term residents) from the Qatar Biobank, which started collecting data from the general population in Qatar since 2012[15]. After exclusion of individuals with body mass index (BMI) <18.5 kg/m^2, we had 7,268 participants (5,903 Qatari and 1,365 long-term resident). Out of these 7,268 samples, there were 4,895 controls (67.4%) and 2,373 with prediabetes (32.6%). According to the American Diabetes guidelines[16], prediabetes cases were defined as those individuals with HbA_{1c} between 39 mmol/mol (5.7%) and 47 mmol/mol (6.4%), whereas controls were those with HbA_{1c} <39 mmol/mol (5.7%). Written consent to use collected data for research, and was obtained by the Qatar Biobank for all the participants. The present project was approved by the institutional review board of the Qatar Biobank (protocol Ex-2018-Res-ACC-0123-0067).

Training and validation populations

For the training stage, we used a case–control design that included 1,902 cases and 3,912 healthy controls. To validate the model developed in the training stage, we used data from 983 healthy controls and 471 cases. The chart of the flow of the analysis starting from acquiring the Qatar Biobank cohort to the Qatar prediabetes risk score development is shown in Figure S1.

Variable categorization

Our aim was to build a prediabetes risk score that uses variables that can be measured routinely, objectively, cheaply, easily and non-invasively in any primary clinical setting or even by the general public. Therefore, 13 variables were requested, including age, sex, BMI, waist circumference (WC) and hip circumference, systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, cigarette smoking, shisha smoking, other household smoker, snoring and occupation. All the variables had <10% missing values, which were imputed using the MICE package in R version 3.5.2 (R Foundation for Statistical Computing, Vienna, Austria). For variable categorization, we used conventional cut-offs or well-accepted clinical guidelines when available, as shown in Table S1. Age (in years) was divided into three groups (18–35, 36–54 and 55 years). For BMI (in kg/m^2), we used the Caucasian cut-offs, and we categorized BMI into three groups: normal (BMI 18.5–24.9 kg/m^2), overweight (BMI 25–29.9 kg/m^2) and obese (BMI ≥30 kg/m^2). The WC (in cm) was categorized into three levels, but with different cut-offs for men and women (level 1: <94 cm for men and <80 cm for women; level 2: 94–102 cm for men and 80–88 cm for women;

level 3: >102 for men and >88 for women). The blood pressure categorization was based on the American Heart Association guidelines[17], but participants were categorized into only the normal state if SBP <130 mmHg and DBP < 80mmHg, or the hypertensive state if SBP ≥130 mmHg or DBP >80 mmHg (Table S1).

Statistical analysis

All statistical analysis was carried out using R version 3.5.2, and R package "h2o" (version 3.17.0.4195) for building the other machine learning (ML) models. Descriptive statistics were used to describe the baseline characteristics of participants. Independent Student's *t*-test was used to compare the means, where the χ^2-test was used to compare proportions and the dependence between the prevalence of prediabetes and the different categories of risk factors. For the optimal cut point to dichotomize the prediabetes score, we used Youden's J statistic[18]. Given the high prevalence of prediabetes in Qatar, and in the Middle East in general, we wanted to prioritize the sensitivity to identify as many cases as possible. Therefore, we used a modified and generalized formula for Youden's J statistic, which is a weighted version for Youden's J statistic[19]. To this end, we chose a weight for specificity that is two-thirds less that the weight of specificity. For the clustering analysis, we used the *K*-means algorithm[20] and the R package "cluster" version 2.1.0[21] to cluster and group our cohort into three clusters based on age, sex, waist size, BMI, SBP and DBP. Statistical significance for all tests was set at $P < 0.05$.

Development of the prediabetes diagnostic model

Two strategies were used for building the prediabetes risk score. The first one relied on multivariate logistic regression (LR), whereas the second relied on more complex ML techniques. Prediabetes was used as an outcome variable; if HbA_{1c} % was in the prediabetes range, the outcome was set as 1; if not, it was 0. For the development of the diagnostic LR model, we apportioned the data into training and validation sets, with an 80/20 split. The most relevant variables were identified using forward then backward binary logistic regression. Initially, we separated the Qatari nationals from the long-term residents. However, the model we developed performed equally in both populations. Therefore, we decided to merge the two populations and treat them as one. We developed two LR models. In the first model, we used the variables sex, age, BMI, blood pressure and waist-to-hip ratio. In the second model, we replaced the waist-to-hip ratio with WC. We compared the two models and did not find any significant difference (comparable areas under the curve [AUCs]; data not shown). Therefore, for reasons of clinical practicality (i.e., no need for hip measurement and calculation of waist-to-hip ratio), we decided to use the second model for building our score. The optimized LR diagnostic model was then tested using the validation sample population. Different metrics can be used to determine the cut-off of a diagnostic model, including Matthew's correlation coefficient, accuracy and balanced accuracy (BACC). We tested all of them, and, because our training and testing data were unbalanced, we decided to use the BACC metric. Similarly, we used the same training and validation sample populations to build four diagnostic models based on four different complex ML techniques, including random forest[22], gradient boosting machine[23], XgBoost[24] (a more scalable and accurate version of gradient boosting machine) and, finally, deep learning[25]. For these models, we used 10-fold cross-validation. Details about these techniques can be found in Appendix S1. The AUC values of the four complex ML models were thereafter compared with the AUC value of the LR diagnostic model.

RESULTS

Basal characteristics of participants

The basal characteristics of the participants in the training and validation datasets are presented in Table 1. The percentage of men in the healthy group and prediabetes participants in the training population was 47.37 and 49.58%, respectively, whereas in the validation population, men represented 46.08% of controls and 48.2% of the cases. People with prediabetes were significantly older than healthy controls in both datasets, but the prevalence of prediabetes was 32% in the two sets. As compared with controls, the cases had significantly higher BMI, WC, hip circumference, SBP and DBP.

Relationship between the prevalence of prediabetes and the categories of different variables

Table S2 shows the relationships between the prevalence of prediabetes and the categories of the five risk factors in the training data. The prevalence of prediabetes increases with increasing age, BMI, waist circumference and high blood pressure. Being male does also increase the prediabetes risk slightly. There was a significant dependence between the prevalence of prediabetes and the categories of the different risk factors ($P < 0.001$), except for sex, where the significance was suggestive ($P = 0.079$).

Prediabetes risk score model and risk score scale

The score points contributed by each risk factor to the final prediabetes score were obtained by multiplying the β coefficient in the LR equation by 10 and rounding it to the nearest integer. Therefore, for each participant, the total score was estimated by summing the scores derived for each risk factor category (Table 2). In the final LR model, we used age, BMI, BP, sex and WC as predictors. In the present cohort, age was the most important predictor that influenced the risk of prediabetes, with individuals aged ≥55 years scoring 22 points. Being obese increased the score by 5 points, as did hypertension; whereas level 3 of WC raises the score by 8 points. Being male on the other hand added 4 points to the score. The scale of the score ranged from 0 to 45 points. The minimum score was 0, and was obtained when a participant scored 0 in all risk factors. The maximum score was 45, and is obtained when the

Table 1 | Baseline characteristics of participants in training and validation datasets

	Training dataset			Validation dataset		
	Controls (3,912)	Cases (n = 1,902)	P	Controls (n = 983)	Cases (n = 471)	P
HbA_{1c}%	5.3 ± 0.3	6.2 ± 0.2	<0.001	5.4 ± 0.3	6.4 ± 0.2	<0.001
HbA_{1c} (mmol/mol)	30.5 ± 5.2	40.4 ± 6.3	<0.001	31.6 ± 4.2	42.6 ± 7.3	<0.001
Age (years)	37 ± 11	48 ± 12	<0.001	37 ± 10	49 ± 12	<0.001
Waist (cm)	87 ± 13	97 ± 13	<0.001	87 ± 13	96 ± 12	<0.001
Hip (cm)	106 ± 11	111 ± 11	<0.001	107 ± 11	110 ± 11	<0.001
SBP (mmHg)	112 ± 13	122 ± 16	<0.001	112 ± 13	122 ± 16	<0.001
DBP (mmHg)	68 ± 10	73 ± 11	<0.001	68 ± 10	73 ± 11	<0.001
Pulse (b.p.m.)	70 ± 10	70 ± 10	0.18	70 ± 10	71 ± 10	0.06
BMI (kg/m^2)	29 ± 6	32 ± 6	<0.001	29 ± 6	32 ± 6	<0.001
Sex (% men)	47.37	49.58	0.12	46.08	48.2	0.48
Prediabetes (%)	32			32		

Data shown as the mean ± standard deviation or proportion. Student's t-test was used to compare continuous variables, and the χ^2-test to compare proportions. BMI, body mass index; HBA_{1c}, hemoglobin A1C; DBP, diastolic blood pressure; SBP, systolic blood pressure.

Table 2 | Multivariate logistic regression model and assigned score for each variable category

	β	P-value	OR	Assigned score
Age (years)				
18–36	Reference			0
36–54	1.22	<0.001	3.39	12
≥55	2.2	<0.001	9	22
BMI				
Normal	Reference			0
Overweight	0.18	0.12	1.19	2
Obese	0.47	<0.001	1.6	5
Blood pressure				
Normal	Reference			0
Hypertension	0.61	<0.001	1.85	6
Sex				
Female	Reference			0
Male	0.3	<0.001	1.36	3
Waist circumference				
Level 1	Reference			0
Level 2	0.27	0.01	1.31	3
Level 3	0.86	<0.001	2.36	9
Cutoff point[†]				16

A score is attributed to each variable category by multiplying the β coefficient by 10 and rounding to the nearest integer. The odds ratio (OR) was obtained by exponentiating the b coefficients (OR exp[b]). BMI, body mass index. [†]The cut-off point is the score that gives the best balanced accuracy, sensitivity and specificity.

participant had the maximum risk in all the risk factors (Table 2).

For the sake of clarity, we divided the risk score of having prediabetes into three categories: low risk (0–16 points), moderate risk (17–27 points) and high risk (>27 points; Table S3). We tested these risk categories in the training data, and found that among the 3,098 low-risk participants, there were 2,648 controls and just 450 with prediabetes (14.53%); among the 1,742 moderate-risk participants, 943 are healthy, whereas 781 had prediabetes (45.3%); and finally, among the 992 high-risk participants, there were 321 controls and 671 with prediabetes (67.64%; Table 3).

We compared the baseline characteristics of the low-, medium- and high-risk participants in the training and validation sets, and found that age, SBP, DBP, BMI, waist size and the percentage of men were all significantly higher in the high-risk group in both datasets (Table 4). Additionally, we defined three clusters based on age, SBP, DBP, BMI, waist size and sex (Table 4). The baseline characteristics of the three clusters are shown in Table 5. Interestingly, from the clustering results, we can map each cluster to one of the risk groups. Cluster 1 had 59.97, 9.81 and <1% from the low-, medium- and high-risk groups, respectively. Whereas cluster 2 had 39.34, 60.37 and 10.94 from the low-, medium- and high-risk groups, respectively. Finally, cluster 3 had <1, 29.82 and 88.98% from the low-, medium- and high-risk groups, respectively. Therefore, based on our variables, we could map cluster 1, cluster 2 and cluster 3 to the low-, medium- and high-risk groups, respectively.

The performance of the LR model in the training sample was tested by constructing a receiver operating characteristic curve. The AUC of the LR model in the training data was found to be 79% (95% confidence interval [CI] 77–80%). The optimal cut-off, based on the weighted version for Youden's J statistic, of the total score was 16, which gave a BACC of 70.5% (95% CI 68.8–72.5%) and an accuracy of 65.2% (95% CI 64–66.4%). The specificity, sensitivity, positive predictive value and negative predictive value of the training model at the cut-off of 16 were 55.3% (95% CI 53.7–56.9%), 85.6% (95% CI 84–87.1), 48.2% (95% CI 46.5–50%) and 88.8% (95% CI 87.4–90%), respectively. The AUC, sensitivity, specificity, ACC,

Table 3 | Prediabetes prevalence in training and testing populations based on risk levels defined by the risk score

Risk level	Score range	Control	Case	Total	Percentage of prediabetes
Training dataset					
Low	0–16	2,648	450	3,098	14.53
Moderate	17–27	943	781	1,724	45.3
High	>27	321	671	992	67.64
Testing dataset					
Low	0–16	677	107	784	13.65
Moderate	17–27	226	194	420	46.19
High	>27	80	170	250	68

BACC, positive predictive value and negative predictive value at different cut-off scores of the Prediabetes Risk Score in Qatar (PRISQ) on the training and testing datasets are shown in Table 6.

Validation of the prediabetes risk score model

We assessed the performance and the robustness of our scoring system using the testing dataset (983 control and 471 cases). The AUC of the validation model was 80% (95% CI 78–83%; Figure 1;Table 4), which is comparable to the AUC of the training model 79% (95% CI 77–80%), showing the steadiness of our prediabetes risk score model. At the cut-off point of 16, the specificity, sensitivity and BACC for the model in the validation population were 57.88% (95% CI 54.73–61%), 86.2% (95% CI 82.75–89.19%) and 72.04% (95% CI 68.74–75.09%), respectively. The positive predictive value and negative predictive value values are 49.5% (95% CI 46.04–53%) and 89.75% (95% CI 887.12–92%), respectively. We tested the model in the validation dataset and found that among the low-risk participants, just 13.65% had prediabetes; this prevalence jumped to 68% among the high-risk individuals (Table 3).

Comparison of logistic regression and complex ML prediabetes models

We used the five risk factors included in the LR diagnostic model in four prediabetes diagnostic models developed using more complex ML techniques, including random forest, gradient boosting machine, XGBoost and deep learning (Appendix S1). The receiver operating characteristic curve curves and AUC values for the four ML models are shown in Figure 1. The details about the different metrics for the four models are shown in Table S4. The performance of the ML models was comparable to that of the LR model (Figure 1).

DISCUSSION

To our knowledge, this is the first study to describe a screening risk score tool for prediabetes using a population from the Middle East. PRISQ is a non-invasive tool to screen prediabetes. It is easy to interpret, as the higher the score the greater the risk of having prediabetes. This scoring system can be used either as a simple web application open to any individual who has the measurements of the risk factors, or implemented in primary care centers and used routinely by health personnel. Given the evidence of the prediabetes-associated harms, including development of type 2 diabetes, risk of cardiovascular disease[26,27], kidney[28] and nerve damage[29], as well as the strong data showing that progression from prediabetes to type 2 diabetes can be prevented in individuals with prediabetes by intensive lifestyle intervention and/or pharmacotherapy[9,30–32], this tool has the potential to play an important role in curbing the type 2 diabetes epidemic sweeping the Middle East[33]. Efficient early identification of prediabetes, as is the case for many other conditions, will significantly affect the intervention strategies and improve outcomes. Furthermore, PRISQ can also have a potential utility in research settings to help recruit participants.

Despite the scientific evidence supporting the prediabetes-associated harms, the threat posed by prediabetes is being overlooked, either because of its asymptomatic nature or because of a lack of public awareness or of knowledge in the medical community[1,34,35]. In the present study, we developed a prediabetes risk scoring system that is simple, easy to use and understand by health providers in primary care. It is quick and does not require any blood assays. The required parameters can all be obtained by a nurse in the triage room, before the encounter with the physician who can immediately see the score and make an appropriate decision in the case of individuals with a high risk of having prediabetes. We do not know of any prediabetes risk score that has been developed for Middle Eastern populations as yet.

Although our score was built using data from the Qatari population, we anticipate that it can be used in the neighboring countries, such as the other nations of the Gulf Cooperation Council, given the shared genetic background, climate, ethnicity and lifestyle habits. Furthermore, we separately analyzed Qatari nationals and long-term residents, including people from other neighboring Arab countries, such as Egypt, Jordan, Lebanon and Iraq, and did not see any difference in terms of performance of our diagnostic model. This suggests that PRISQ might be useful in other Middle Eastern countries. Other prediabetes scores have been developed for other populations, but

Table 4 | Baseline characteristics of the low, medium and high-risk groups in the training and validation sets, and clusters obtained with K-means clustering analysis based on age, body mass index, systolic blood pressure, diastolic blood pressure, waist size and sex

	Training data set				Validation dataset			
	Low (n = 2,438)	Moderate (n = 2,069)	High (n = 1,307)	P	Low (n = 634)	Moderate (n = 488)	High (n = 332)	P
Basic characteristics								
Age (years)	30.71 ± 6.91	43.08 ± 8.01	54.07 ± 9.55	<0.0001	31.61 ± 7.43	42.87 ± 7.74	54.61 ± 9.86	<0.0001
SBP (mmHg)	106.90 ± 9.93	115.78 ± 2.06	129.87 ± 15.10	<0.0001	106.42 ± 9.20	115.81 ± 12.35	130.65 ± 14.99	<0.0001
DBP (mmHg)	64.79 ± 8.63	70.63 ± 10.28	75.97 ± 11.45	<0.0001	64.36 ± 8.39	71.40 ± 10.48	75.64 ± 11.50	<0.0001
BMI (kg/m^2)	26.53 ± 4.80	31.04 ± 5.16	33.77 ± 5.24	<0.0001	26.40 ± 4.49	31.37 ± 5.18	33.53 ± 5.71	<0.0001
Waist (cm)	81.01 ± 10.44	93.07 ± 10.79	102.28 ± 11	<0.0001	80.66 ± 9.69	93.61 ± 10.79	101.77 ± 11.43	<0.0001
Sex								
Women	1,399 (57.38%)	1,041 (50.31%)	578 (44.22%)	<0.0001	377 (59.46%)	253 (51.84%)	144 (43.37%)	<0.0001
Sex (% men)	1,039 (42.62%)	1,028 (49.69%)	729 (55.78%)	<0.0001	257 (40.54%)	235 (48.16%)	188 (56.63%)	<0.0001
Clustering								
Cluster 1	1,462 (59.97%)	203 (9.81%)	1 (0.08%)	<0.0001	382 (60.25%)	48 (9.84%)	1 (0.30%)	<0.0001
Cluster 2	959 (39.34%)	1,249 (60.37%)	143 (10.94%)	<0.0001	250 (39.43%)	291 (59.63%)	30 (9.04%)	<0.0001
Cluster 3	17 (0.70%)	617 (29.82%)	1,163 (88.98%)	<0.0001	2 (0.32%)	149 (30.53%)	301 (90.66%)	<0.0001

The P-value shows significance among the three groups based on ANOVA test. Data are mean ± standard deviation or proportions. BMI, body mass index; DBP, diastolic blood pressure; SBP, systolic blood pressure.

Table 5 | Baseline characteristics for the three clusters in training and validation datasets

	Training dataset				Validation dataset			
	Cluster 1 (n = 1,666)	Cluster 2 (n = 2,351)	Cluster 3 (n = 1,797)	P	Cluster 1 (n = 431)	Cluster 2 (n = 571)	Cluster 3 (n = 452)	P
Age (years)	31.58 ± 8	38.14 ± 9.16	51.42 ± 10.34	<0.0001	32.10 ± 8.07	38.27 ± 8.94	51.78 ± 10.64	<0.0001
SBP (mmHg)	101.16 ± 6.71	113.19 ± 7.49	130.93 ± 12.92	<0.0001	101.23 ± 6.51	112.80 ± 7.32	131.24 ± 3.15	<0.0001
DBP (mmHg)	60.28 ± 6.84	69.59 ± 7.65	77.55 ± 10.79	<0.0001	60.45 ± 7.21	69.80 ± 7.57	77.10 ± 11.18	<0.0001
BMI (kg/m^2)	25.49 ± 3.95	30.16 ± 4.89	33.20 ± 5.91	<0.0001	25.48 ± 3.79	30.23 ± 5.06	33.06 ± 5.97	<0.0001
Waist (cm)	76.68 ± 7.64	91.51 ± 9.45	100.65 ± 12.20	<0.0001	76.66 ± 7.27	91.31 ± 9.32	100.51 ± 2.20	<0.0001
Women	1,230 (73.83%)	1,013 (43.09%)	775 (43.13%)	<0.0001	322 (74.71%)	255 (44.66%)	197 (43.58%)	<0.0001
Men	436 (26.17%)	1,338 (56.91%)	1,022 (56.87%)	<0.0001	109 (25.29%)	316 (55.34%)	255 (56.42%)	<0.0001

Data are the mean ± standard deviation or proportion. P for difference across clusters for the different variables. BMI, body mass index; DBP, diastolic blood pressure; SBP, systolic blood pressure.

we believe that the genetic and environmental determinants of prediabetes might render these scores unsuitable for Middle Eastern populations. Indeed, previous studies have shown that scores to screen for prediabetes developed for certain populations did not perform as well in other populations, and sometimes even between two populations from the same country, but different regions[36]. Also, Glümer *et al.*[37] showed that the risk score they developed for undiagnosed type 2 diabetes performed well in different Caucasian populations, but poorly in non-Caucasian populations.

Compared with other LR-based prediabetes risk scores developed in other countries for adult populations, the discrimination capability of the PRISQ risk score, as measured by the AUC of the receiver operating characteristic curve, was 80%, higher than many other scores developed in other countries, such as Indonesia (AUC 62.3%)[38], China (two scores: AUC 74% and AUC 70%)[36,39] and the USA (AUC 74%)[40]. According to our risk score, a person with a score between 17 and 27 would be considered as being at moderate risk of prediabetes, whereas a score >27 would be considered high risk. Our score showed a good sensitivity (86.02%) at the cut-off point of 16.

Given the high prevalence of prediabetes in the Middle Eat in general, as well as the high risk of converting from prediabetes to type 2 diabetes, one would like to have a very sensitive test to detect as many cases as possible. This will obviously come at a price of increasing the number of false positives. However, given the low cost of the confirmatory blood tests and the fact that prediabetes is not a life-threatening condition, at least in the short term, the benefits of being falsely identified with the condition outweighs the risk of not being diagnosed.

Table 6 | Specificity, sensitivity, balanced accuracy, accuracy, positive predictive value, negative predictive value and area under the curve for selected risk point scores of Prediabetes Risk Score in Qatar in the training and testing datasets using the logistic regression model

Cut-off point	Specificity % (95% CI)	Sensitivity % (95% CI)	BACC % (95% CI)	ACC% (95% CI)	PPV% (95% CI)	NPV % (95% CI)	AUC % (95% CI)
Training data							
1	9.71 (8.8–10.68)	98.79 (98.19–99.23)	54.25 (53.5–54.96)	38.85 (37.6–40.12)	34.73 (33.46–36.01)	94.29 (91.56–96.35)	0.79
5	33.03 (31.55–34.53)	95.43 (94.39–96.32)	64.2 (62.97–65.42)	53.4 (52.15–54.73)	40.9 (39.47–42.39)	93.6 (92.28–94.92)	(0.77–0.80)
10	39.19 (37.65–40.74)	93.9 (92.73–94.93)	66.5 (65.19–67.84)	57.0 (55.8–58.36)	42.8 (41.37–44.4)	92.9 (91.62–94.15)	
15	55.34 (53.77–56.91)	85.65 (83.99–87.19)	70.4 (68.88–72.05)	65.2 (64.02–66.48)	48.2 (46.55–49.95)	88. (87.48–90.03)	
16	55.34 (53.77–56.91)	85.65 (83.99–87.19)	70.4 (68.88–72.05)	65.2 (64.02–66.48)	48.2 (46.55–49.95)	88. (87.48–90.03)	
17	67.38 (65.89–68.85)	76.66 (74.69–78.54)	72.0 (70.29–73.7)	70.4 (69.22–71.59)	53.3 (51.44–55.21)	85.5 (84.29–86.81)	
18	67.69 (66.2–69.15)	76.34 (74.36–78.24)	72.0 (70.28–73.69)	70.5 (69.33–71.69)	53.4 (51.56–55.35)	85.4 (84.18–86.7)	
19	67.69 (66.2–69.15)	76.34 (74.36–78.24)	72.0 (70.28–73.69)	70.5 (69.33–71.69)	53.4 (51.56–55.35)	85.4 (84.18–86.7)	
20	73.29 (71.87–74.67)	70.4 (68.29–72.44)	71.8 (70.08–73.56)	72.3 (71.17–73.49)	56.1 (54.15–58.17)	83.5 (82.3–84.81)	
25	79.65 (78.36–80.9)	61.72 (59.5–63.92)	70.6 (68.93–72.41)	73.7 (72.64–74.91)	59.5 (57.39–61.77)	81.0 (79.79–82.29)	
30	91.79 (90.89–92.64)	35.28 (33.13–37.47)	63.5 (62.01–65.05)	73.3 (72.15–74.44)	67.6 (64.63–70.55)	74.4 (73.22–75.7)	
35	96.09 (95.43–96.67)	19.98 (18.2–21.85)	58.0 (56.82–59.26)	71.1 (70.01–72.35)	71.2 (67.25–75.1)	71.1 (69.94–72.4)	
40	98.34 (97.89–98.72)	9.1 (7.84–10.48)	53.7 (52.86–54.6)	69.1 (67.94–70.33)	72.6 (66.56–78.25)	68.9 (67.76–70.2)	
45	100 (99.91–100)	0 (0–0.19)	5 (49.95–50.1)	67.2 (66.06–68.49)	NaN(0–100)	67.2 (66.06–68.49)	
Testing data							
1	9.46 (7.7–11.46)	99.15 (97.84–99.77)	54.3 (52.77–55.62)	38.5 (36–41.07)	34.4 (31.89–37.01)	95.8 (89.78–98.87)	0.80
5	33.06 (30.12–36.1)	97.03 (95.06–98.37)	65.0 (62.59–67.23)	53.7 (51.18–56.37)	40.9 (38.08–43.94)	95.8 (93.17–97.72)	(0.78–0.83)
10	38.56 (35.5–41.68)	94.69 (92.26–96.54)	66.6 (63.88–69.11)	56.7 (54.15–59.31)	42.4 (39.46–45.53)	93.8 (91–95.96)	
15	57.88 (54.73–60.99)	86.2 (82.75–89.19)	72.0 (68.74–75.09)	67.0 (64.57–69.47)	49.5 (46.04–52.99)	89.7 (87.12–92)	
16	57.88 (54.73–60.99)	86.2 (82.75–89.19)	72.0 (68.74–75.09)	67.0 (64.57–69.47)	49.5 (46.04–52.99)	89.7 (87.12–92)	
17	68.57 (65.56–71.46)	77.28 (73.23–80.99)	72.9 (69.39–76.23)	71.3 (68.99–73.7)	54.0 (50.24–57.9)	86. (83.69–88.63)	
18	68.87 (65.87–71.76)	77.28 (73.23–80.99)	73.0 (69.55–76.37)	71. (69.2–73.9)	54.3 (50.47–58.15)	86.3 (83.75–88.68)	
19	68.87 (65.87–71.76)	77.28 (73.23–80.99)	73.0 (69.55–76.37)	71. (69.2–73.9)	54.3 (50.47–58.15)	86.3 (83.75–88.68)	
20	75.28 (72.46–77.95)	71.34 (67.02–75.38)	73.3 (69.74–76.67)	7 (71.67–76.24)	58.0 (53.89–62.09)	84.5 (82.01–86.9)	
25	80.16 (77.53–82.61)	62.42 (57.87–66.81)	71.2 (67.7–74.71)	74.4 (72.09–76.64)	60.1 (55.63–64.49)	81.6 (79.07–84.05)	
30	91.86 (89.97–93.49)	36.09 (31.75–40.61)	63.9 (60.86–67.05)	73. (71.46–76.04)	6 (61.83–73.74)	7 (72.45–77.42)	
35	97.05 (95.79–98.02)	23.14 (19.41–27.22)	60. (57.6–62.62)	73.1 (70.75–75.37)	78.9 (71.23–85.45)	72.4 (69.99–74.89)	
40	99.39 (98.68–99.78)	10.83 (8.17–13.99)	55.1 (53.42–56.88)	70. (68.29–73.03)	89.4 (78.48–96.04)	69.9 (67.46–72.33)	
45	100 (99.63–100)	0 (0–0.78)	5 (49.81–50.39)	67.6 (65.13–70.01)	NaN (0–100)	67.6 (65.13–70.01)	

ACC, accuracy; AUC, area under the curve; BACC, balanced accuracy; CI, confidence interval; NPV, negative; predictive value; PPV, positive predictive value.

PRISQ includes five risk factors, of which two – BMI and WC – are known to be associated with insulin resistance[41,42], a hallmark of prediabetes and type 2 diabetes. Fortunately, these two factors are modifiable and provide the opportunity for preventive intervention, such as intensive lifestyle modification, which, as above-mentioned, was shown to reduce the incidence of type 2 diabetes in individuals with prediabetes.

The present study showed that age is a major risk factor for the development of prediabetes in the Qatari population. Previous studies have reported that some populations from central America (Mexico) and the Caribbean (Jamaica)[43], as well as non-white UK minorities, including black people, people from the Caribbean and south Asian people[44], show an earlier diabetes onset than white people. Furthermore, a study from Israel reported that Arabic people have an earlier diabetes onset than Jewish people[45]; by the age of 57 years, 25% of Arabic people had diagnosed diabetes; the corresponding age among Jewish people was 68 years, a difference of 11 years. Similarly, Arabic men developed diabetes earlier than UK men living in Canada, likely due to unhealthy lifestyle[46]. Furthermore, a study comparing diabetes onset in Iraqi immigrants living in Sweden versus Swedes showed that independent of a family history of diabetes and obesity – two major risk factors for type 2 diabetes – Iraqi immigrants had a significantly earlier age of diabetes onset (47.6 years vs 53.4 years) and a higher risk of diabetes onset[47]. These observations might, at least partly, explain why age makes a major contribution to PRISQ. Given that age is a non-modifiable factor, special attention should be paid to screening for prediabetes at an early age, mainly in individuals with other risk factors, such as obesity or hypertension, for example. The American Diabetes Association already recommends that diabetes screening for most adults begin at age 45 years, and advises diabetes screening before age 45 years if the person has additional risk factors. No data are available about the onset of prediabetes in Arabic people. However, the fact that Arabic people from Middle East develop diabetes earlier that white people suggests they are also prone to early prediabetes onset. It is, therefore, recommended that the age for screening for prediabetes, and

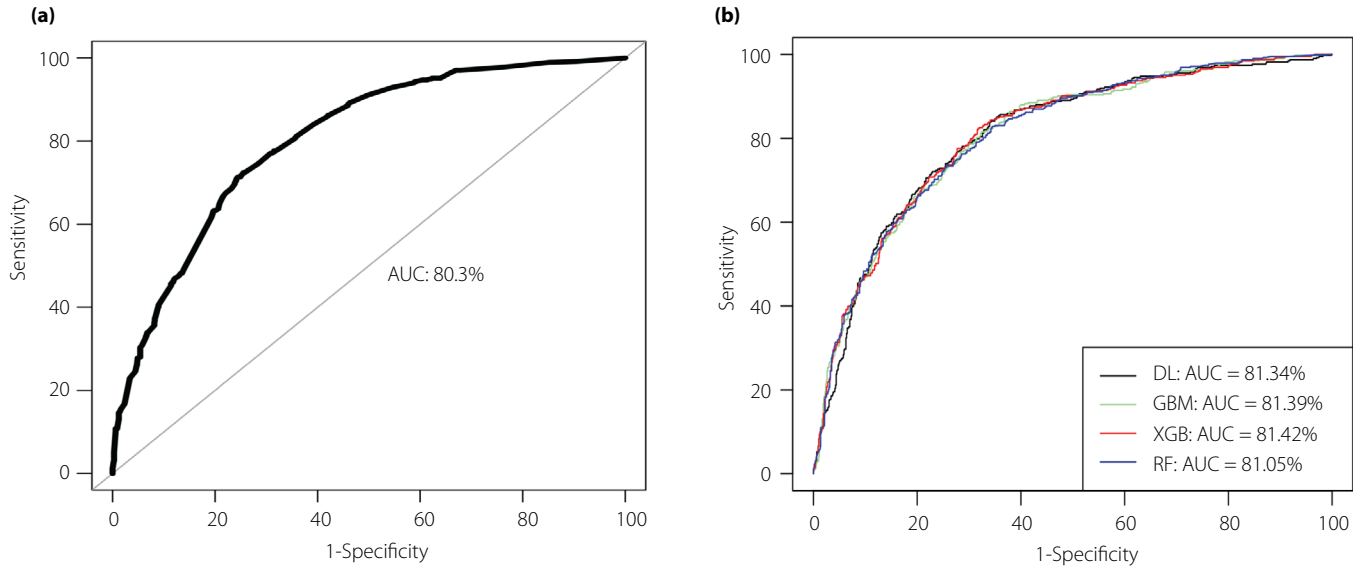

Figure 1 | Receiver operating characteristic curves for diagnosing prediabetes using either the (a) logistic regression model or (b) more complex machine learning models (Qatar Biobank cohort). AUC, area under the curve; DL, deep learning; GBM, gradient boosting machine; RF, random forest; XGB, XgBoost.

diabetes for that matter, is reduced to 40 years, and even less in individuals with additional risk factors.

PRISQ does not use family history of diabetes, which is a major independent risk factor for the development of prediabetes and type 2 diabetes[1,48,49], as a predictor, since many individuals might not know about the medical history of their relatives[50], give unreliable information[51] or because their relatives have type 2 diabetes or prediabetes, but are undiagnosed. Indeed, according to the International Diabetes Federation, one in two people with diabetes was undiagnosed globally[3]. Other risk factors, such as eating habits, smoking, physical activity, income and occupation, might improve the performance of our model, but they are all subjective and it is not easy to obtain accurate data, mainly when, for cultural reasons, people do not like to talk about their habits, such as smoking (mainly women) or drinking alcohol. Although the present study focused on prediabetes identification, if high-risk individuals undergo confirmatory blood tests, we might expect to find undiagnosed type 2 diabetes too. Consequently, the present tool might also help identify many early type 2 diabetes cases, which have a better chance of being treated.

To improve the screening of prediabetes in the present sample population, we also built four prediabetes risk scores based on complex ML techniques and compared their performance with the LR-based model. In our hands, the ML models did not outperform the LR model. The AUCs of the four ML models are comparable to the AUC of the LR model (Figure 1). We opted for logistic regression, as it is a direct competitor of many other ML methods, and outperforms many of them when predictors act mainly additively[52].

The main strength of the present study was that it is the first study to use a population from the Middle East to develop a prediabetes risk score. We also used a large sample size (7,386) compared with other prediabetes risk scores, if we consider the small size of the population of Qatar (2.5 million). Furthermore, we had a balanced population in terms of sex, which means that our model can be used for both sexes. Compared with other prediabetes risk scores, PRISQ uses only objective risk factors that can accurately and easily measured. The main limitation of the present study was the lack of validation in external populations. We believe, however, that given the shared environmental factors and lifestyle habits, as well as the genetic background and ethnicity between many Middle Eastern countries, the PRISQ might perform as well in many of these nations.

The PRISQ is a prediabetes risk score that is the first of its kind in the Middle East. The risk model showed a good validation performance. The categorization of the total risk level makes it easy for a health provider to decide about a prediabetes intervention or prevention program if required. As a web application, PRISQ can be used by any individual anywhere if he/she has the measures of the five risk factors. The score can be easily applied in clinical settings using electronic health records, and can improve the efficiency of population-based screening. It also increases the chance to identify undiagnosed type 2 diabetes. Finally, it can be useful for research in that it might aid in the identification of potential research participants with prediabetes.

ACKNOWLEDGMENTS

We thank Qatar Biobank staff, particularly Dr Nahla Afifi and Dr Asma AL Thani, for facilitating access to the data and providing us with expert advice. We are also grateful to all the participants of the study. The project was funded by internal

grants from Qatar Biomedical Research Institute and Qatar Computing Research Institute to AA and HB, respectively.

DISCLOSURE

The authors declare no conflict of interest. Open Access funding provided by the Qatar National Library.

REFERENCES

1. Cho NH, Shaw JE, Karuranga S, *et al*. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* 2017; 138: 271–281.
2. Saeedi P, Petersohn I, Salpea P, *et al*. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9(th) edition. *Diabetes Res Clin Pract* 2019; 157: 107843.
3. International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Brussels, Belgium: International Diabetes Federation, 2020.
4. Nathan DM, Davidson MB, DeFronzo RA, *et al*. Impaired fasting glucose and impaired glucose tolerance: implications for care. *Diabetes Care* 2007; 30: 753–759.
5. de Vegt F, Dekker JM, Jager A, *et al*. Relation of impaired fasting and postload glucose with incident type 2 diabetes in a Dutch population: the Hoorn Study. *JAMA* 2001; 285: 2109–2113.
6. Aldossari KK, Aldiab A, Al-Zahrani JM, *et al*. Prevalence of prediabetes, diabetes, and its associated risk factors among males in Saudi Arabia: a population-based survey. *J Diabetes Res* 2018; 2018: 2194604.
7. Alkandari A, Longenecker JC, Barengo NC, *et al*. The prevalence of pre-diabetes and diabetes in the Kuwaiti adult population in 2014. *Diabetes Res Clin Pract* 2018; 144: 213–223.
8. Jiang L, Johnson A, Pratte K, *et al*. Long-term outcomes of lifestyle intervention to prevent diabetes in American Indian and Alaska Native communities: the Special Diabetes Program for Indians Diabetes Prevention Program. *Diabetes Care* 2018; 41: 1462–1470.
9. 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.
10. Knowler WC, Fowler SE, Hamman RF, *et al*. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* 2009; 374: 1677–1686.
11. Pan XR, Li GW, Hu YH, *et al*. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. *Diabetes Care* 1997; 20: 537–544.
12. Ramachandran A, Snehalatha C, Mary S, *et al*. The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia* 2006; 49: 289–297.
13. Tuomilehto J, Lindstrom J, Eriksson JG, *et al*. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001; 344: 1343–1350.
14. Alfawaz HA, Wani K, Alnaami AM, *et al*. Effects of different dietary and lifestyle modification therapies on metabolic syndrome in prediabetic arab patients: a 12-month longitudinal study. *Nutrients* 2018; 10: 383.
15. Al Kuwari H, Al Thani A, Al Marri A, *et al*. The Qatar Biobank: background and methods. *BMC Public Health* 2015; 15: 1208.
16. Association AD. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2019. *Diabetes Care* 2019; 42(Suppl 1): S13–S28.
17. Bakris G, Ali W, Parati G. ACC/AHA versus ESC/ESH on hypertension guidelines: JACC guideline comparison. *J Am Coll Cardiol* 2019; 73: 3018–3026.
18. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3: 32–35.
19. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* 2006; 163: 670–675.
20. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. *J Roy Stat Soc: Ser C (Appl Stat)* 1979; 28: 100–108.
21. cluster: Cluster Analysis Basics and Extensions [computer program]. Version 2.1.0. R2019.
22. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
23. Mall R, Cerulo L, Garofano L, *et al*. RGBM: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. *Nucl Acids Res* 2018; 46: e39.
24. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2016.
25. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Massashusettes, USA: The MIT Press, 2016.
26. Huang Y, Cai X, Mai W, *et al*. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis. *BMJ* 2016; 355: i5953.
27. Vistisen D, Witte DR, Brunner EJ, *et al*. Risk of cardiovascular disease and death in individuals with prediabetes defined by different criteria: the Whitehall II study. *Diabetes Care* 2018; 41: 899–906.
28. Kim GS, Oh HH, Kim SH, *et al*. Association between prediabetes (defined by HbA1C, fasting plasma glucose, and impaired glucose tolerance) and the development of chronic kidney disease: a 9-year prospective cohort study. *BMC Nephrol* 2019; 20: 130.
29. Wilson ML. Prediabetes: beyond the borderline. *Nurs Clin North Am* 2017; 52: 665–677.

30. Lindstrom J, Peltonen M, Eriksson JG, *et al*. Improved lifestyle and decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish Diabetes Prevention Study (DPS). *Diabetologia* 2013; 56: 284–293.
31. Galaviz KI, Narayan KMV, Lobelo F, *et al*. Lifestyle and the prevention of type 2 diabetes: a status report. *Am J Lifestyle Med* 2018; 12: 4–20.
32. Group DPPR. 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: 866–875.
33. Azizi F, Hadaegh F, Hosseinpanah F, *et al*. Metabolic health in the Middle East and north Africa. *Lancet Diabetes Endocrinol* 2019;7:866–879.
34. Tseng E, Greer RC, O'Rourke P, *et al*. Survey of primary care providers' knowledge of screening for, diagnosing and managing prediabetes. *J Gen Intern Med* 2017; 32: 1172–1178.
35. Khan T, Wozniak GD, Kirley K. An assessment of medical students' knowledge of prediabetes and diabetes prevention. *BMC Med Educ* 2019; 19: 285.
36. Wang H, Liu T, Qiu Q, *et al*. A simple risk score for identifying individuals with impaired fasting glucose in the Southern Chinese population. *Int J Environ Res Public Health* 2015; 12: 1237–1252.
37. Glumer C, Vistisen D, Borch-Johnsen K, *et al*. Risk scores for type 2 diabetes can be applied in some populations but not all. *Diabetes Care* 2006; 29: 410–414.
38. Fujiati II, Damanik HA, Bachtiar A, *et al*. Development and validation of prediabetes risk score for predicting prediabetes among Indonesian adults in primary care: cross-sectional diagnostic study. *Interv Med Appl Sci* 2017; 9: 76–85.
39. Ouyang P, Guo X, Shen Y, *et al*. A simple score model to assess prediabetes risk status based on the medical examination data. *Can J Diabetes* 2016; 40: 419–423.
40. Koopman RJ, Mainous AG 3rd, Everett CJ, *et al*. Tool to assess likelihood of fasting glucose impairment (TAG-IT). *Ann Fam Med* 2008; 6: 555–561.
41. Cheng YH, Tsao YC, Tzeng IS, *et al*. Body mass index and waist circumference are better predictors of insulin resistance than total body fat percentage in middle-aged and elderly Taiwanese. *Medicine* 2017; 96: e8126.
42. Sasaki R, Yano Y, Yasuma T, *et al*. Association of waist circumference and body fat weight with insulin resistance in male subjects with normal body mass index and normal glucose tolerance. *Intern Med* 2016; 55: 1425–1432.
43. Irving R, Tusie-Luna MT, Mills J, *et al*. Early onset type 2 diabetes in Jamaica and in Mexico. Opportunities derived from an interethnic study. *Rev Invest Clin* 2011; 63: 198–209.
44. Winkley K, Thomas SM, Sivaprasad S, *et al*. The clinical characteristics at diagnosis of type 2 diabetes in a multi-ethnic population: the South London Diabetes cohort (SOUL-D). *Diabetologia* 2013; 56: 1272–1281.
45. Kalter-Leibovici O, Chetrit A, Lubin F, *et al*. Adult-onset diabetes among Arabs and Jews in Israel: a population-based study. *Diabet Med* 2012; 29: 748–754.
46. Tenkorang EY. Early onset of type 2 diabetes among visible minority and immigrant populations in Canada. *Ethn Health* 2017; 22: 266–284.
47. Bennet L, Lindblad U, Franks PW. A family history of diabetes determines poorer glycaemic control and younger age of diabetes onset in immigrants from the Middle East compared with native Swedes. *Diabetes Metab* 2015; 41: 45–54.
48. Scott RA, Langenberg C, Sharp SJ, *et al*. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the EPIC-InterAct study. *Diabetologia* 2013; 56: 60–69.
49. Wagner R, Thorand B, Osterhoff MA, *et al*. Family history of diabetes is associated with higher risk for prediabetes: a multicentre analysis from the German Center for Diabetes Research. *Diabetologia* 2013; 56: 2176–2180.
50. Goergen AF, Ashida S, Skapinsky K, *et al*. What you don't know: improving family health history knowledge among multigenerational families of Mexican origin. *Public Health Genomics* 2016; 19: 93–101.
51. Daelemans S, Vandevoorde J, Vansintejan J, *et al*. The use of family history in primary health care: a qualitative study. *Adv Prev Med* 2013; 2013: 695763.
52. McMahan HB, Streeter M. Open problem: better bounds for online logistic regression. Proceedings of the 25th Annual Conference on Learning Theory; 2012; Proceedings of Machine Learning Research.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 | Chart of the flow of data from the Qatar Biobank cohort to building a prediabetes risk score.
Table S1 | Variable categorization.
Table S2 | Relationship between prediabetes and variable categories.
Table S3 | Risk status categorization based on the prediabetes risk score.
Table S4 | Specificity, sensitivity, accuracy and balanced accuracy for selected risk point scores of the machine learning models.
Appendix S1 | Research design and methods: Complex machine learning models.