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

External validation of the Hong Kong Chinese non-laboratory risk models and scoring algorithm for case finding of prediabetes and diabetes mellitus in primary care

Will HG Cheng[1], Weinan Dong[1], Emily TY Tse[1,2]*, Carlos KH Wong[1,3,4], Weng Y Chin[1], Laura E Bedford[1], Daniel YT Fong[5], Welchie WK Ko[6], David VK Chao[7,8], Kathryn CB Tan[9], Cindy LK Lam[1,2]

[1]Department of Family Medicine and Primary Care, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong, [2]Department of Family Medicine, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China, [3]Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong, [4]Laboratory of Data Discovery for Health (D24H), Hong Kong Science and Technology Park, Sha Tin, Hong Kong, [5]School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong, [6]Family Medicine and Primary Healthcare Department, Queen Mary Hospital, Hong Kong West Cluster, Hospital Authority, Hong Kong, Hong Kong, [7]Department of Family Medicine & Primary Health Care, United Christian Hospital, Kowloon East Cluster, Hospital Authority, Hong Kong, Hong Kong, [8]Department of Family Medicine & Primary Health Care, Tseung Kwan O Hospital, Kowloon East Cluster, Hospital Authority, Hong Kong, Hong Kong, and [9]Department of Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong

Keywords
Opportunistic case-finding, Prediabetes, Risk prediction models

***Correspondence**
Emily TY Tse
Tel.: +852-2814-7484
Fax: +852-2814-7475
E-mail address:
emilyht@hku.hk

J Diabetes Investig 2024; 15: 1317–1325

doi: 10.1111/jdi.14256

Clinical Trial Registry.
US ClinicalTrial.gov and HKU Clinical Trials Registry.
NCT04881383 and HKUCTR-2808.

ABSTRACT

Aims/Introduction: Two Hong Kong Chinese non-laboratory-based prediabetes/diabetes mellitus (pre-DM/DM) risk models were developed using logistic regression (LR) and machine learning, respectively. We aimed to evaluate the models' validity in case finding of pre-DM/DM in a Chinese primary care (PC) population. We also evaluated the validity of a risk-scoring algorithm derived from the LR model.

Materials and Methods: This was a cross-sectional external validation study on Chinese adults, without a prior DM diagnosis, who were recruited from public/private PC clinics in Hong Kong. A total of 1,237 participants completed a questionnaire on the models' predictors. Of that, 919 underwent blood glucose testing. The primary outcome was the models' and the algorithm's sensitivity in finding pre-DM/DM cases. The secondary outcomes were the models' and the algorithm's specificity, positive/negative predictive values, discrimination and calibration.

Results: The models' sensitivity were 0.70 (machine learning) and 0.72 (LR). Both showed good external discrimination (area under the receiver operating characteristic curve: machine learning 0.744, LR 0.739). The risks estimated by the models were lower than the observed incidence, indicating poor calibration. Both models were more effective among participants with lower pretest probabilities; that is, age 18–44 years. The algorithm's sensitivity was 0.77 at the cut-off score of ≥16 out of 41.

Conclusion: This study showed the validity of the models and the algorithm for finding pre-DM/DM cases in a Chinese PC population in Hong Kong. They can facilitate more cost-effective identification of high-risk individuals for blood testing to diagnose pre-DM/DM in PC. Further studies should recalibrate the models for more precise risk estimation in PC populations.

INTRODUCTION

Prediabetes elevates the risks of long-term health complications even before the onset of diabetes mellitus[1]. Approximately 70%

Received 28 March 2024; revised 20 May 2024; accepted 6 June 2024

of individuals with prediabetes eventually develop diabetes, with an annual conversion rate of 5–10%[2]. Conversely, timely interventions are effective in delaying and/or reversing diabetes progression[3–5]. However, as the current guidelines primarily focus on diabetes screening among high-risk populations, such as adults aged ≥45 years[6], over half of Hong Kong's prediabetes cases are unaware of their conditions[7].

Globally, various risk prediction tools have been developed for identifying prediabetes cases[8,9]. Those solely using non-laboratory-based predictors are often preferred in clinical practice, as they can reduce the cost and inconvenience of unnecessary blood tests. However, to the best of our knowledge, existing non-laboratory-based tools for Chinese adults; that is, the New Chinese Diabetic Risk Score (NCDRS)[10] and the Non-invasive Diabetes Score (NDS)[11], estimate diabetes risks only.

Using the available data from the Hong Kong Population Health Survey 2014–15[7], Dong *et al.* developed two new Hong Kong (HK) Chinese non-laboratory-based prediabetes/diabetes risk models. They incorporated lifestyle variables that are readily available in routine clinical practice by logistic regression (LR) and machine learning (ML) methods (URL: https://www.hk-dm-cx-risk-engine.hku.hk/predm)[12]. They can potentially be applied as an initial screening tool in the general and primary care (PC) populations. Based on the predicted outcomes, they could triage those that require follow-up blood tests and optimize the resources needed for diagnosing prediabetes/diabetes[13]. The models' internal validation – the model's performance using a subsample split from the same dataset that is used for model development[14] – shows good discrimination in detecting individuals with prediabetes/diabetes[12]. However, their external validity – the model's performance in an independent sample that is different from that used in model development[14] – has yet to be confirmed.

The 'Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis' (TRIPOD) statement recommends that all risk prediction models be externally validated in the intended population before clinical implementation[15]. In the present study, we evaluated the external validity of the models in detecting prediabetes/diabetes cases in a Chinese adult population recruited from public and private PC clinics in Hong Kong. We also derived a user-friendly risk-scoring algorithm from the LR model and evaluated its performance in the same population.

MATERIALS AND METHODS

Case definition

We applied the same case definitions as the models' development study[12]. Prediabetes and diabetes cases were determined by oral glucose tolerance testing (OGTT) and/or hemoglobin A1c (HbA1c). Prediabetes was defined as fasting glucose of 6.1–6.9 mmol/L[16], 2-h post-75 g-glucose-solution-OGTT of 7.8–11 mmol/L[16] or HbA1c of 5.7–6.4%[17]. Diabetes was defined as a fasting glucose of ≥7.0 mmol/L[16], 2-h post-75 g-glucose-solution-OGTT of ≥11.1 mmol/L[16] or HbA1c of ≥6.5%[17].

Study design and population

This was a cross-sectional external validation study carried out in Hong Kong between April 2021 and January 2022. Details of the study design, population and data collection procedures for the present study can be found in the published study protocol[18]. In short, Chinese adults aged between 18 and 84 years who did not have any prior diagnoses of diabetes, coronary heart disease, stroke, chronic kidney disease, cancer or anemia were eligible, non-Chinese individuals, individuals who were unable to communicate in Chinese or English and those who were pregnant or too ill to complete a questionnaire were excluded. Convenience sampling was used, and participation was voluntary. Participants were recruited by research assistants in participating clinics, referred by their consulting doctors, self-referred or referred through snowball sampling. Participants younger than 45 years were purposefully recruited to ensure a more representative distribution of age groups.

Outcome measures and sample size calculation

The study's primary outcome was the sensitivity of the models and the algorithm in finding prediabetes/diabetes cases in the PC study population. Secondary outcomes included the models' and the algorithm's specificity, positive predictive value (PPV), negative predictive value (NPV), discrimination and the models' calibration.

Based on the models' hypothesized sensitivity of ≥75% and a 15.08% prevalence of undiagnosed prediabetes/diabetes in Hong Kong[7], 710 adults were needed to achieve the minimal acceptable lower 95% confidence limit of >0.6[19]. We planned to recruit 1,014 participants to account for 30% attrition.

Data collection

Each participant completed a written consent to participate in the study, and completed a questionnaire on sociodemographic information, personal and family history of chronic medical conditions, and lifestyle habits. They then attended designated laboratories for blood tests (OGTT, complete blood count, HbA1c) and anthropometric measurements. The results of the anthropometric and blood test measurements were screened by a clinician in the study team to identify participants who required referral back to their recruiting clinic for further management of abnormalities.

Derivation of a risk-scoring algorithm from the LR model

We derived a scoring algorithm from the LR model based on the methods reported by Sullivan *et al*[20]. The total prediabetes/diabetes risk score, which ranged from 0 (lowest risk) to 41 (highest risk), was obtained by summating each predictor's score.

Statistical analysis

Descriptive statistics were used to present the incidence of newly diagnosed prediabetes and diabetes, and the characteristics of study participants. We applied the study participants' data to each risk model to estimate the prediabetes/diabetes risk

level. We calculated the models' sensitivity, specificity, PPV, and NPV at their optimal and different risk thresholds. The highest value of Youden's index (sensitivity + specificity − 1) determined the models' optimal thresholds. We obtained both models' areas under the receiver operating characteristic curves (AUC-ROC) to determine their discrimination while generating the 95% confidence intervals (CIs) through bootstrapping methods. The models' calibration was assessed using the Hosmer–Lemeshow test and calibration plots.

Subgroup analyses (by age, sex and the presence or absence of a family history of diabetes) on sensitivity, specificity, PPV, NPV and AUC-ROC were carried out to identify the models' performance in specific PC populations. As there is no other Chinese non-laboratory-based prediabetes risk-scoring algorithm, we applied two existing diabetes risk scores, the NCDRS[10] and the NDS[11], to our study population data to compare the performances of the new HK Chinese models to existing models. The sensitivity, specificity, PPV, NPV and discrimination of the risk-scoring algorithm were also determined in the study population.

Statistical analyses were carried out using R 3.5.1 (The R Foundation for Statistical Computing, Vienna, Austria), with statistical significance taken as $P < 0.05$.

RESULTS

A total of 1,237 Chinese adults completed the questionnaire, of which 919 (74.3%) completed the blood tests and anthropometric measurements. The characteristics of those included in the final analyses of this study ($n = 919$) and those lost to follow-up who did not attend the blood test ($n = 318$) are shown and compared in Table S1. The characteristics of participants between the PC validation study population and the models' original development study sample are shown in Table 1. In the PC population, the incidence of newly diagnosed prediabetes/diabetes was 53.43% ($n = 491$), including 49.18% ($n = 452$) pre- diabetes and 4.24% ($n = 39$) diabetes.

Predictive performance of the HK Chinese risk models in the PC population

The ROC curves showed that both models offered satisfactory external discrimination in detecting prediabetes/diabetes cases, with no significant difference between them (Figure 1; AUC-ROC ML 0.744 [95% CI 0.7127, 0.7734], AUC-ROC LR 0.739 [95% CI 0.7063, 0.7701]; Delong's test $P > 0.05$). The optimal risk thresholds of the models were 17.7% (LR) and 16.7% (ML), which offered similar sensitivities of 0.72 (95% CI 0.68, 0.76) and 0.69 (95% CI 0.65, 0.74), respectively. Detailed results of both models at their optimal, 10, 15, 20 and 25% estimated risk thresholds, as well as the models' performance at fixed sensitivity levels (i.e., 0.9, 0.8 and 0.75), are listed in Table 2. Results from the Hosmer–Lemeshow test ($P < 0.01$) and the calibration plots (Figure 2a,b) showed that the risk levels estimated by both models were significantly lower than the actual incidence in the study population, which indicated poor calibration.

Table 1 | Participants' characteristics in the external validation study population ($n = 919$) and original development dataset ($n = 1857$)

	External validation study participants ($n = 919$)	Participants from the development sample ($n = 1857$)
	Mean (SD)	
Age (years)	51.4 (13.59)	40.7 (15.48)
BMI (kg/m^2)	23.3 (3.65)	23.03 (3.77)
WHR	0.846 (0.072)	0.84 (0.07)
Waist (cm)	82.0 (10.28)	79.68 (10.65)
Hip (cm)	96.8 (7.1)	NA
Mean SBP (mmHg)	120.8 (18.53)	115.8 (17.36)
Mean DBP (mmHg)	71.9 (10.36)	76.6 (10.39)
Vigorous recreational activity (min/week)	41.0 (108.43)	37.2 (111.41)
Fruit consumption (serves/month)	36.3 (24.49)	32.3 (35.53)
Sleep duration (h/day)	6.75 (1.25)	6.9 (1.19)
	%	
Smokers	5.22	12.17

BMI, body mass index; DBP, diastolic blood pressure; NA, not applicable due to unavailability; SBP, systolic blood pressure; WHR, waist-to-hip ratio.

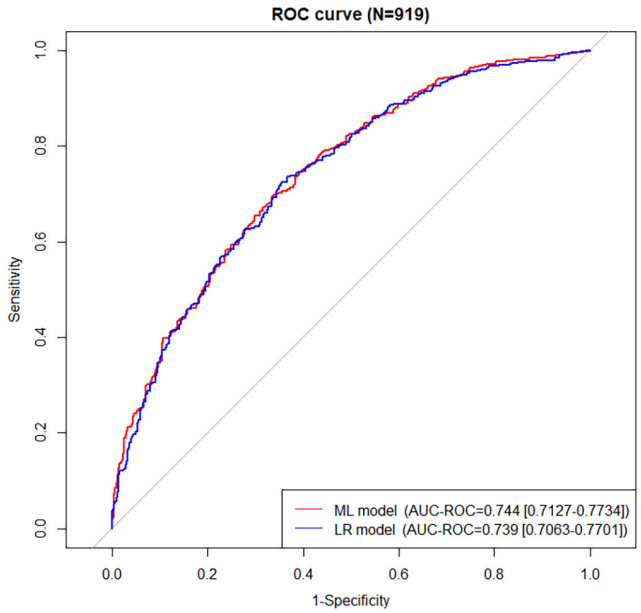

Figure 1 | Receiver operating characteristic (ROC) curves of risk prediction models to detect prediabetes mellitus and diabetes mellitus on the validation study population ($n = 919$). 95% CIs were calculated using bootstrap; AUC, area under curve; LR, logistic regression; ML, machine learning.

Table 2 | Sensitivity, specificity, positive predictive value and negative predictive value of the Hong Kong Chinese non-laboratory-based risk prediction models to detect new prediabetes/diabetes cases in the primary care study population at different risk thresholds and at different sensitivity levels (n = 919)

Risk threshold	Model	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Models' performance at different risk thresholds (10%/15%/20%/25%)					
Optimal[†]					
16.7%	ML	0.69 (0.65, 0.74)	0.67 (0.62, 0.71)	0.70 (0.66, 0.75)	0.65 (0.61, 0.70)
17.7%	LR	0.72 (0.68, 0.76)	0.65 (0.60, 0.69)	0.70 (0.66, 0.74)	0.67 (0.63, 0.72)
10%	ML	0.87 (0.84, 0.90)	0.43 (0.38, 0.47)	0.63 (0.60, 0.67)	0.74 (0.69, 0.79)
	LR	0.89 (0.86, 0.91)	0.41 (0.36, 0.46)	0.63 (0.60, 0.67)	0.76 (0.70, 0.81)
15%	ML	0.74 (0.70, 0.77)	0.61[‡] (0.57, 0.66)	0.69 (0.65, 0.73)	0.67 (0.62, 0.72)
	LR	0.78 (0.74, 0.81)	0.55 (0.51, 0.60)	0.67 (0.63, 0.71)	0.68 (0.64, 0.73)
20%	ML	0.60[‡] (0.55, 0.64)	0.74[‡] (0.69, 0.78)	0.72 (0.68, 0.77)	0.61 (0.57, 0.66)
	LR	0.64 (0.60, 0.68)	0.69 (0.65, 0.73)	0.70 (0.66, 0.75)	0.63 (0.58, 0.67)
25%	ML	0.48 (0.44, 0.52)	0.82 (0.78, 0.85)	0.75 (0.70, 0.80)	0.58 (0.54, 0.62)
	LR	0.48 (0.44, 0.53)	0.82 (0.78, 0.85)	0.75 (0.70, 0.80)	0.58 (0.54, 0.62)
Models' performance at different sensitivity levels (0.9/0.8/0.75)					
8.7%	ML	0.90	0.38 (0.33, 0.42)	0.62 (0.59, 0.66)	0.77 (0.71, 0.82)
8.9%	LR		0.37 (0.32, 0.41)	0.62 (0.58, 0.66)	0.76 (0.70, 0.82)
13.2%	ML	0.80	0.53 (0.48, 0.58)	0.66 (0.62, 0.70)	0.70 (0.65, 0.75)
14.1%	LR		0.53 (0.48, 0.57)	0.66 (0.62, 0.70)	0.70 (0.65, 0.75)
14.9%	ML	0.75	0.60 (0.55, 0.65)	0.68 (0.64, 0.72)	0.68 (0.63, 0.73)
16.4%	LR		0.60 (0.55, 0.64)	0.68 (0.64, 0.72)	0.68 (0.63, 0.72)

[†]Determined by Youden's Index. [‡]McNemar's χ^2-test, $P < 0.05$. DM, diabetes mellitus; LR, logistic regression; ML, machine learning; NPV, negative predictive value; PC, primary care; PPV, positive predictive value; Pre-DM, prediabetes.

Models' performance in specific PC subgroups

The two models had similar discrimination performances in the various age subgroups (Delong's test P-value >0.05). Both models had the best discrimination and sensitivity in the 18–44-years-old subgroup, for whom the incidence of prediabetes/diabetes was the lowest among all age subgroups (AUCs-ROC ML 0.738 [95% CI 0.6708, 0.8031], LR 0.728 [95% CI 0.6600, 0.7693]). For the male subgroup, the ML model had a better specificity of 0.80 (95% CI 0.73, 0.87), but a low sensitivity of 0.58 (95% CI 0.51, 0.66). In contrast, the LR model had a better sensitivity of 0.88 (95% CI 0.83, 0.93), but a low specificity of 0.49 (95% CI 0.40, 0.57). Both models had higher discrimination and sensitivity in the subgroup without a family history (AUCs-ROC ML 0.756 [95% CI 0.7181, 0.7949], LR 0.754 [95% CI 0.7151, 0.7937]) than that with a positive family history (AUCs-ROC ML 0.721 [95% CI 0.6691, 0.7764], LR 0.706 [95% CI 0.6500, 0.7591]). AUCs-ROC by subgroups are shown in Figures S1–S3, and the results of the models' performances by subgroups are tabulated in Tables S2–S4.

Models' performance compared with existing diabetes risk-scoring algorithm

The AUC-ROCs of the new models were generally larger than those of the NCDRS and NDS in our PC study population (ML 0.744 [95% CI 0.7127, 0.7734], LR 0.739 [95% CI 0.7063, 0.7701], NCDRS 0.738 [95% CI 0.7066, 0.7697], NDS 0.720 [95% CI 0.6866, 0.7524]), whereas the ML model's AUC-ROC was significantly higher than the NDS (DeLong's test P = 0.018). The NDS was the most sensitive in detecting prediabetes/diabetes among the risk tools (ML 0.69 [95% CI 0.65, 0.74], LR 0.72 [95% CI 0.68, 0.76], NCDRS 0.71 [95% CI 0.67, 0.75], NDS 0.78 [95% CI 0.75, 0.82]), but it had the lowest specificity (ML 0.67 [0.62, 0.71], LR 0.65 [95% CI 0.60, 0.69], NCDRS 0.66 [95% CI 0.61, 0.70], NDS 0.55 [95% CI 0.51, 0.60]). At a fixed sensitivity of 75%, the ML and LR had a higher specificity than the others (ML 0.60 [95% CI 0.55, 0.65], LR 0.60 [95% CI 0.55, 0.64], NCDRS 0.54 [95% CI 0.50, 0.59], NDS 0.55 [95% CI 0.51, 0.60]). The AUC-ROCs of the new models and the existing risk scores are shown in Figure S4, and the performance at their respective optimal risk thresholds/score cut-offs are shown in Table S5.

The risk-scoring algorithm derived from the LR model

The prediabetes/diabetes risk-scoring algorithm is shown in Table 3. The AUC-ROC in the PC study population was 0.731 (95% CI 0.6969, 0.7615). There was no significant difference in discrimination between the LR model and the risk-scoring algorithm (Delong's test $P > 0.05$). The Youden's index suggested that ≥18 out of 41 was the optimal cut-off score, with a sensitivity, specificity, PPV and NPV of 0.63 (95% CI 0.58, 0.67), 0.72 (95% CI 0.68, 0.76), 0.72 (95% CI 0.68, 0.76) and 0.63 (95% CI 0.58, 0.67), respectively. At the study's proposed sensitivity of 75%, the cut-off score was ≥16 out of 41 with a specificity, PPV and NPV of 0.56 (95% CI 0.51, 0.61), 0.67 (95% CI 0.63, 0.70) and 0.68 (95% CI 0.63, 0.73), respectively. The performance of the risk-scoring algorithm in finding prediabetes/

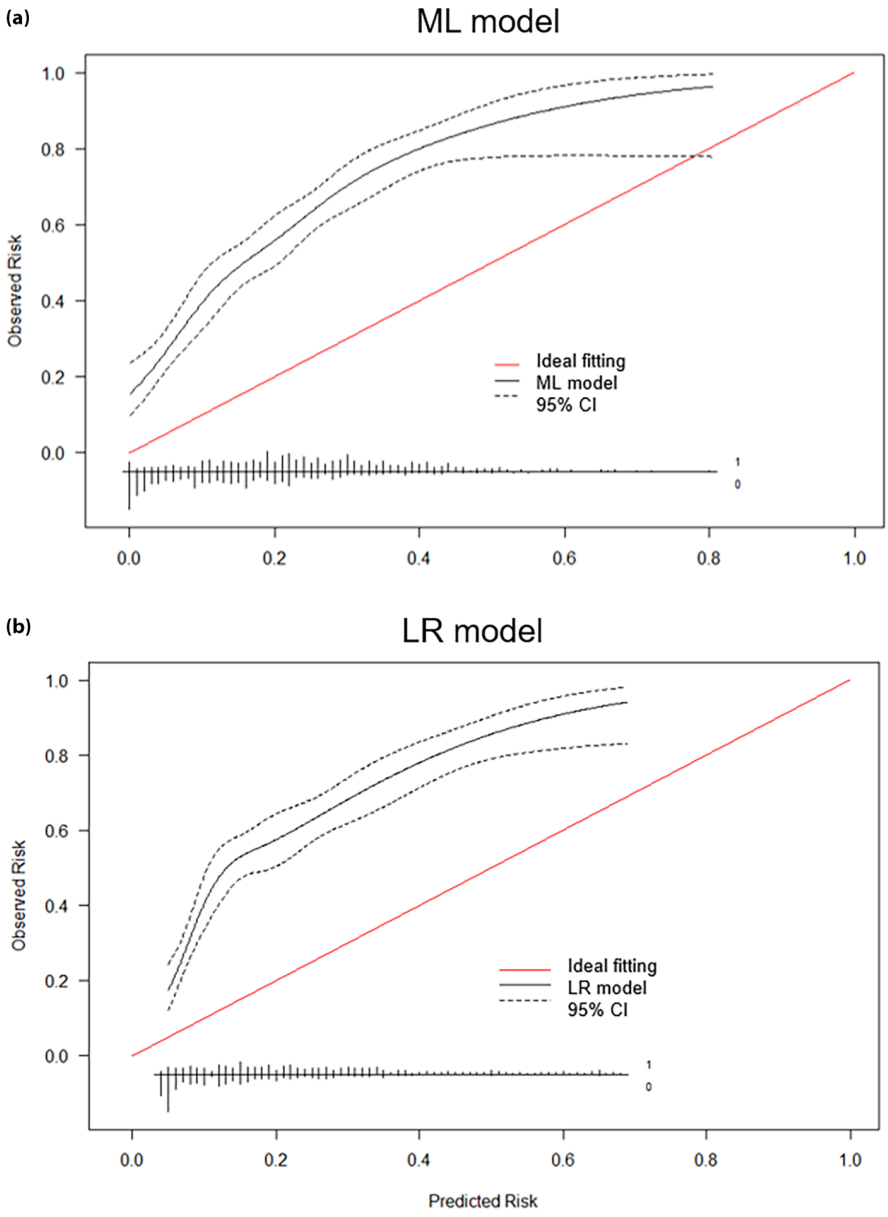

Figure 2 | Calibration plots of the (a) machine learning (ML) and (b) logistic regression (LR) model to detect prediabetes mellitus and diabetes mellitus on the validation study population (n = 919). The x-axis is the predicted risk of prediabetes mellitus and diabetes mellitus, and the y-axis is the observed risk of prediabetes mellitus and diabetes mellitus. The curves were fitted based on restricted cubic splines. At the bottom of the graphs, histograms of the predicted risks are shown for the participants with (1) and without (0) prediabetes and diabetes mellitus.

diabetes cases in the study population at different cut-off scores is shown in Table 4.

DISCUSSION

The present results supported the external validity of the new HK Chinese non-laboratory-based risk prediction models and the derived risk-scoring algorithm in identifying prediabetes/diabetes cases in a PC population. At their optimal risk thresholds (LR 17.7%, ML 16.7%), the LR and ML models had a sensitivity of detecting prediabetes/diabetes cases of 0.72 (95% CI 0.68, 0.76) and 0.69 (95% CI 0.65, 0.74), respectively. The scoring algorithm had a sensitivity of 0.77 (95% CI 0.73, 0.81) at the cut-off score of ≥16 out of 41. The AUCs-ROC of the models and scoring algorithm were 0.74 (95% CI 0.71, 0.77;

Table 3 | The prediabetes/diabetes risk scoring algorithm converted from the Hong Kong Chinese non-laboratory-based logistic regression risk prediction model

Risk prediction variables	Categories	Scores
Age (years)	18–40	0
	41–49	8
	50–59	10
	60–69	11
	70–84	10
	≥85	6
Body mass index (kg/m^2)	<21	0
	21–21.9	2
	22–22.9	3
	23–25.9	5
	≥26	10
Waist-to-hip ratio	<0.85	0
	0.85–0.89	2
	0.9–0.99	3
	≥1	5
Smoker	No	0
	Yes	4
Sleep duration (h/day)	<6, and you are aged	
	<45 years	6
	≥45 years	2
	≥6	
	For all ages	0
Vigorous exercise (min/week)	None	3
	10–119	2
	≥120	0
Fruit consumption (serve/day)	None	2
	<1	1
	≥1	0

ML), 0.74 (95% CI 0.71, 0.77; LR) and 0.73 (95% CI 0.70, 0.76; algorithm). However, the models were poorly calibrated for estimating the risk of prediabetes/diabetes in the PC study population. Subgroup analyses showed that both models were most effective in case finding among participants with a lower pretest probability, specifically those aged 18–44 years and those without any family history of diabetes.

Although the ML model had a better discrimination performance than the LR model in their internal validation[12], there was no significant difference in AUC-ROCs between them in our study. Similar studies reported comparable performances between LR and ML models[21–23]. A systematic review that compared clinical prediction models also reported that ML models did not offer incremental benefits in performance over LR models[24]. The reduced performance of the ML model in external validation could be due to ML models being more prone to overfitting than LR models[25–28]. As ML models operate by best-fitting relationships between predictors and outcomes in the original development sample[29,30], their performance might not be as generalizable when applied to other populations[31]. In the present study, the LR model offered a higher sensitivity, but a lower specificity in the PC population, which might be preferred as an initial screening tool.

The present findings suggested that the HK Chinese models and risk-scoring algorithm would be most useful for case finding among population groups with lower prediabetes/diabetes incidences. These people are considered as "low risk" and are often not included in current guidelines on diabetes screening. The Hong Kong Reference Framework for Diabetes Care for Adults in Primary Care Settings recommends diabetes screening only for adults at or over the age of 45 years[6]. Our tools can supplement the current guidelines to enhance case findings among individuals aged younger than 45 years.

The prediabetes/diabetes risk score is obtained by simple addition, facilitating opportunistic case finding at the point of care in busy PC clinics and by self-assessment. Although the statistically optimal cut-off score determined by Youden's index was ≥18/ out of 41, the sensitivity was low at 0.63. This indicates that 37% of prediabetes/diabetes cases could be missed. Using ≥16 out of 41 as the practical cut-off score, the sensitivity, specificity, PPV and NPV of the algorithm are 0.77, 0.56, 0.67 and 0.68, respectively. The selection of the cut-off score should strike a balance between the potential health risks of undiagnosed prediabetes/diabetes, and the burden of cost and

Table 4 | Sensitivity, specificity, positive predictive value and negative predictive value of the Hong Kong Chinese risk scoring algorithm risk scoring algorithm to detect new prediabetes/diabetes cases and the percentage of individuals identified as high-risk by the algorithm in the primary care study population at different cut-off scores (n = 919)

Cut-off score (out of 41)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
≥14	0.87 (0.84, 0.90)	0.42 (0.37, 0.46)	0.63 (0.59, 0.67)	0.73 (0.68, 0.79)
≥15	0.82 (0.78, 0.85)	0.49 (0.44, 0.54)	0.65 (0.61, 0.68)	0.70 (0.65, 0.75)
≥16	0.77 (0.73, 0.81)	0.56 (0.51, 0.61)	0.67 (0.63, 0.70)	0.68 (0.63, 0.73)
≥17	0.68 (0.64, 0.73)	0.66 (0.61, 0.70)	0.70 (0.65, 0.74)	0.64 (0.60, 0.69)
≥18[†]	0.63 (0.58, 0.67)	0.72 (0.68, 0.76)	0.72 (0.68, 0.76)	0.63 (0.58, 0.67)
≥19	0.52 (0.48, 0.57)	0.79 (0.75, 0.83)	0.74 (0.70, 0.79)	0.59 (0.55, 0.63)

[†]Optimal cut-off score determined by Youden's Index. DM, diabetes mellitus; NPV, negative predictive value; PC, primary care; PPV, positive predictive value; Pre-DM, prediabetes.

time of the diagnostic blood test for the "screened-positive" individuals. As the confirmatory blood test for prediabetes/diabetes is relatively inexpensive and causes little harm, having a more sensitive cut-off score at the expense of more false positive cases and resultant additional blood tests might be justifiable.

Considering the low calibration of the models at this stage, they are not suitable for estimating the absolute levels of prediabetes/diabetes risk for individuals in PC. Future studies that recalibrate and/or update the models according to the characteristics of the patient population in PC, such as higher prediabetes/diabetes incidence, are warranted to improve their prediction precision for clinical application.

The present study had several points of strength. First, we offered an easy-to-use, initial, preblood test screening tool to allocate resources better to diagnose prediabetes/diabetes among individuals in PC. Second, we validated the models in the intended population with participants recruited from Hong Kong's public and private PC clinics. This enhanced the local applicability and representativeness of our results. Third, as the development and validation populations were sampled independently, it increased the reliability of our findings while avoiding data leakage between study populations, which could overestimate performances[32]. In contrast, some limitations of our study should be acknowledged. Voluntary convenience sampling tended to attract individuals with higher risks. The pre diabetes/diabetes incidence (53.42%) in the study population was much higher than that of the original development sample (15.08%), which used a population-representative random sampling method[7]. Such differences lowered the models' calibration in the PC population, and might mislead users and clinicians during clinical decision-making[33]. The high prevalence in our study population could also lead to overestimating the PPVs and underestimating the NPVs of the models, as predictive values depend on the pretest probability[34,35]. Finally, the results from the HK Chinese PC population might not be generalizable to Chinese populations in other parts of the world due to potential lifestyle and environmental differences.

The present findings supported the validity and discrimination of two new HK Chinese non-laboratory-based prediabetes/diabetes risk prediction models in a Chinese PC population in HK. A simplified risk-scoring algorithm performed similarly to its original model. We recommend a cut-off score of ≥16 out of 41 for case finding of individuals at risk of prediabetes/diabetes cases in PC. We hope this algorithm can serve as an easy-to-use initial screening tool at the point of care to facilitate more effective identification of individuals for further blood tests to detect prediabetes/diabetes in busy clinical practices.

ACKNOWLEDGMENTS

We thank all the participants who participated in our study. We also thank all the doctors and staff in the participating public/private primary care clinics. We also thank Dr Amy Ng for proofreading the manuscript. This work was supported by the Health and Medical Research Fund, the Health Bureau, Government of the Hong Kong Special Administrative Region (reference number: 17181641). The funding organization did not play any role in the design and conduct of the study, collection, management, analysis, interpretation of the data or manuscript preparation.

DISCLOSURE

KCBT is an Editorial Board member of the *Journal of Diabetes Investigation* and a co-author of this article. To minimize bias, they were excluded from all editorial decision-making related to the acceptance of this article for publication.

Approval of the research protocol: Published in a peer-reviewed journal, BMJ Open, 12 (5), e059430. https://bmjopen.bmj.com/content/12/5/e059430.

Informed consent: Written informed consent was obtained from all participants in the study.

Registry and the registration no. of the study/trial: The study is registered at (i) the US ClinicalTrial.gov (NCT04881383; approval date: 10 May 2021) and (ii) the HKU clinical trials registry (HKUCTR-2808; approval date 27 December 2019).

Animal studies: N/A. Our study did not involve any animal subjects.

REFERENCES

1. 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.
2. Tabák AG, Herder C, Rathmann W, *et al*. Prediabetes: A high-risk state for developing diabetes. *Lancet* 2012; 379: 2279–2290.
3. Diabetes Prevention Program Research Group. The Diabetes Prevention Program (DPP) description of lifestyle intervention. *Diabetes Care* 2002; 25: 2165–2171.
4. Gillies CL, Abrams KR, Lambert PC, *et al*. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: Systematic review and meta-analysis. *BMJ* 2007; 334: 299.
5. Tuomilehto J, Lindström 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.
6. Department of Health HKSAR Government. Hong Kong reference framework for diabetes care for adults in primary care settings. 2017.
7. Department of Health HKSAR Government. Report of population health survey 2014/15. 2017.
8. Cheng WH-G, Mi Y, Dong W, *et al*. Non-laboratory-based risk prediction tools for undiagnosed pre-diabetes: A systematic review. *Diagnostics* 2023; 13: 1294.
9. Barber SR, Davies MJ, Khunti K, *et al*. Risk assessment tools for detecting those with pre-diabetes: A systematic review. *Diabetes Res Clin Pract* 2014; 105: 1–13.

10. Zhou X, Qiao Q, Ji L, *et al*. Nonlaboratory-based risk assessment algorithm for undiagnosed type 2 diabetes developed on a nation-wide diabetes survey. *Diabetes Care* 2013; 36: 3944–3952.
11. Woo YC, Gao B, Lee CH, *et al*. Three-component non-invasive risk score for undiagnosed diabetes in Chinese people: Development, validation and longitudinal evaluation. *J Diabetes Investig* 2020; 11: 341–348.
12. Dong W, Tse TYE, Mak LI, *et al*. Non-laboratory-based risk assessment model for case detection of diabetes mellitus and pre-diabetes in primary care. *J Diabetes Investig* 2022; 13: 1374–1386.
13. Khunti K, Gillies CL, Taub NA, *et al*. A comparison of cost per case detected of screening strategies for type 2 diabetes and impaired glucose regulation: Modelling study. *Diabetes Res Clin Pract* 2012; 97: 505–513.
14. Steyerberg EW. Applications of Prediction Models. Cham: Springer, 2009.
15. Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med* 2015; 162: 55–63.
16. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: Report of a WHO/IDF consultation. 2006.
17. American Diabetes Association. 2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2019. *Diabetes Care* 2019; 42: S13–S28.
18. Dong W, Cheng WHG, Tse ETY, *et al*. Development and validation of a diabetes mellitus and prediabetes risk prediction function for case finding in primary care in Hong Kong: A cross-sectional study and a prospective study protocol paper. *BMJ Open* 2022; 12: e059430.
19. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005; 58: 859–862.
20. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* 2004; 23: 1631–1660.
21. Abbas M, Mall R, Errafii K, *et al*. Simple risk score to screen for prediabetes: A cross-sectional study from the Qatar Biobank cohort. *J Diabetes Investig* 2021; 12: 988–997.
22. Yu W, Liu T, Valdez R, *et al*. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak* 2010; 10: 16.
23. Sadek K, Abdelhafez I, Al-Hashimi I, *et al*. Screening for diabetes and impaired glucose metabolism in Qatar: Models' development and validation. *Prim Care Diabetes* 2022; 16: 69–77.
24. Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; 110: 12–22.
25. Cabitza F, Campagner A, Soares F, *et al*. The importance of being external. Methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed* 2021; 208: 106288.
26. Navarro CLA, Damen JA, Takada T, *et al*. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ* 2021; 375: n2281.
27. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Technol* 2010; 11: 2079–2107.
28. von Neumann J. Model selection and overfitting. *Nat Methods* 2016; 13: 703–704.
29. Gravesteijn BY, Nieboer D, Ercole A, *et al*. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol* 2020; 122: 95–107.
30. Liew BX, Kovacs FM, Rügamer D, *et al*. Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. *Eur Spine J* 2022; 31: 2082–2091.
31. van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *J Clin Epidemiol* 2016; 78: 83–89.
32. Samala RK, Chan H-P, Hadjiiski L, *et al*. (eds). Hazards of data leakage in machine learning: A study on classification of breast cancer using deep neural networks. In: *Proceedings of SPIE Medical Imaging 2020: Computer-Aided Diagnosis*. International Society for Optics and Photonics.
33. Van Calster B, McLernon DJ, Van Smeden M, *et al*. Calibration: The Achilles heel of predictive analytics. *BMC Med* 2019; 17: 1–7.
34. Akobeng AK. Understanding diagnostic tests 2: Likelihood ratios, pre-and post-test probabilities and their use in clinical practice. *Acta Paediatr* 2007; 96: 487–491.
35. Altman D. Diagnostic test 2: Predictive values. *BMJ* 1994; 309: 102.

SUPPORTING INFORMATION

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

Figure S1. (a–c) ROC curves of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population by age groups.

Figure S2. (a–c) ROC curves of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population by gender.

Figure S3. (a–c) ROC curves of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population by family history of DM.

Figure S4. ROC curves of different risk prediction models to detect new pre-DM/DM cases in the PC study population (N = 919).

Table S1. The characteristics of participants included in the external validation study population who completed the blood test (n = 919) and participants lost to follow-up who did not attend the blood test (n = 318).

Table S2. (a) Sensitivity, specificity, PPV, and NPV of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population at the optimal risk threshold by age subgroups (N = 919). (b) Risk threshold, specificity, PPV, and NPV of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population at different sensitivity levels by age subgroups (N = 919).

Table S3. (a) Sensitivity, specificity, PPV, and NPV of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population at the optimal risk threshold by gender subgroups (N = 919). (b) Risk threshold, specificity, PPV, and NPV of pre-diabetes mellitus risk models at different sensitivity levels by gender subgroups (N = 919).

Table S4. (a) Sensitivity, specificity, PPV, and NPV of the new Hong Kong Chinese models to detect new pre-DM/DM cases in the PC study population at the optimal risk threshold by subgroups with/without family history of diabetes mellitus (N = 919). (b) Risk threshold, specificity, PPV, and NPV of the new Hong Kong Chinese risk prediction models to detect new pre-DM/DM cases in the PC study population at different sensitivity levels by subgroups with/without family history of diabetes mellitus (N = 919).

Table S5. Sensitivity, specificity, PPV, and NPV of different risk prediction models to detect new pre-DM/DM cases in the PC study population at their optimal risk thresholds (N = 919).