Development and validation of risk prediction model for recurrent cardiovascular events among Chinese: the Personalized CARdiovascular DIsease risk Assessment for Chinese model

Yekai Zhou[1,†], Celia Jiaxi Lin[2,†], Qiuyan Yu[3], Joseph Edgar Blais [3],
Eric Yuk Fai Wan [3,4], Marco Lee[3], Emmanuel Wong[5], David Chung-Wah Siu[5],
Vincent Wong[6], Esther Wai Yin Chan[3,7], Tak-Wah Lam[1], William Chui[6],
Ian Chi Kei Wong[3,7,8], Ruibang Luo [1,*], and Celine Sze Ling Chui [2,7,9,*]

[1]Department of Computer Science, The University of Hong Kong, Rm 301 Chow Yei Ching Building, Pokfulam Road, Pokfulam, Hong Kong Special Administrative Region, 999077, China ; [2]School of Nursing, The University of Hong Kong, 5/F Academic Building, 3 Sassoon Road, Pokfulam, Hong Kong Special Administrative Region, 999077, China ; [3]Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, The University of Hong Kong, Hong Kong Special Administrative Region, 999077, China; [4]Department of Family Medicine and Primary Care, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong Special Administrative Region, 999077, China; [5]Department of Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong Special Administrative Region, 999077, China; [6]Department of Pharmacy, Queen Mary Hospital, Hospital Authority, Hong Kong Special Administrative Region, 999077, China; [7]Laboratory of Data Discovery for Health (D²4H), Hong Kong Science Park, Hong Kong Science and Technology Park, Hong Kong Special Administrative Region, 999077, China; [8]Aston Pharmacy School, Aston University, Birmingham, B4 7ET, United Kingdom; and [9]School of Public Health, The University of Hong Kong, Hong Kong Special Administrative Region, China

Aims	Cardiovascular disease (CVD) is a leading cause of mortality, especially in developing countries. This study aimed to develop and validate a CVD risk prediction model, Personalized CARdiovascular DIsease risk Assessment for Chinese (P-CARDIAC), for recurrent cardiovascular events using machine learning technique.
Methods and results	Three cohorts of Chinese patients with established CVD were included if they had used any of the public healthcare services provided by the Hong Kong Hospital Authority (HA) since 2004 and categorized by their geographical locations. The 10-year CVD outcome was a composite of diagnostic or procedure codes with specific International Classification of Diseases, Ninth Revision, Clinical Modification. Multivariate imputation with chained equations and XGBoost were applied for the model development. The comparison with Thrombolysis in Myocardial Infarction Risk Score for Secondary Prevention (TRS-2°P) and Secondary Manifestations of ARTerial disease (SMART2) used the validation cohorts with 1000 bootstrap replicates. A total of 48 799, 119 672 and 140 533 patients were included in the derivation and validation cohorts, respectively. A list of 125 risk variables were used to make predictions on CVD risk, of which 8 classes of CVD-related drugs were considered interactive covariates. Model performance in the derivation cohort showed satisfying discrimination and calibration with a C statistic of 0.69. Internal validation showed good discrimination and calibration performance with C statistic over 0.6. The P-CARDIAC also showed better performance than TRS-2°P and SMART2.

* Corresponding author. Tel: +852 28592186, Email: rbluo@cs.hku.hk (R.L.); Tel: +852 39176629, Email: cslchui@hku.hk (C.S.L.C.)
†The first two authors are co-first authors.

Conclusion	Compared with other risk scores, the P-CARDIAC enables to identify unique patterns of Chinese patients with established CVD. We anticipate that the P-CARDIAC can be applied in various settings to prevent recurrent CVD events, thus reducing the related healthcare burden.

Structured Graphical Abstract
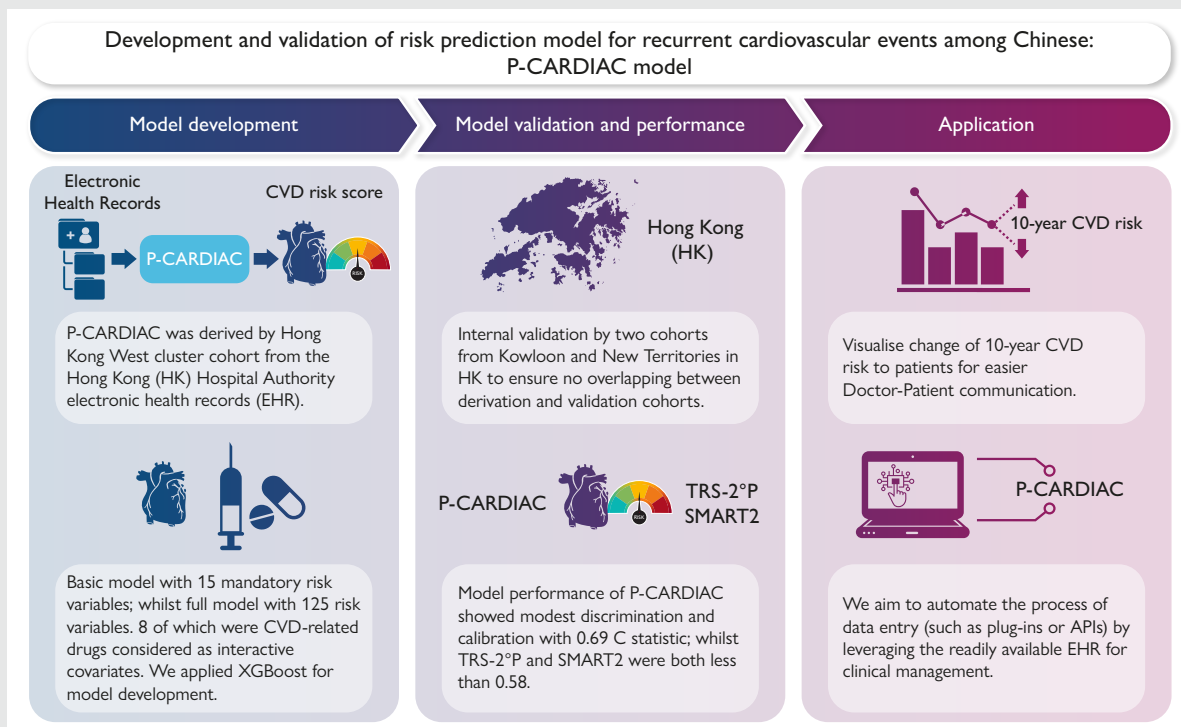
Key Question

Does the Personalized CARdiovascular DIsease risk Assessment for Chinese (P-CARDIAC) perform better than common risk scores among Chinese with established cardiovascular disease (CVD) risk?

Key Finding

In this study, the P-CARDIAC showed better discrimination and a C statistic of 0.69, higher than other common risk scores.

Take Home Message

The P-CARDIAC is an innovative model to provide early intervention for Chinese with established CVD

Development and validation of risk prediction model for recurrent cardiovascular events among Chinese: P-CARDIAC model

Model development

P-CARDIAC was derived by Hong Kong West cluster cohort from the Hong Kong (HK) Hospital Authority electronic health records (EHR).

Basic model with 15 mandatory risk variables; whilst full model with 125 risk variables. 8 of which were CVD-related drugs considered as interactive covariates. We applied XGBoost for model development.

Model validation and performance

Hong Kong (HK)

Internal validation by two cohorts from Kowloon and New Territories in HK to ensure no overlapping between derivation and validation cohorts.

P-CARDIAC — TRS-2°P SMART2

Model performance of P-CARDIAC showed modest discrimination and calibration with 0.69 C statistic; whilst TRS-2°P and SMART2 were both less than 0.58.

Application

10-year CVD risk

Visualise change of 10-year CVD risk to patients for easier Doctor-Patient communication.

P-CARDIAC

We aim to automate the process of data entry (such as plug-ins or APIs) by leveraging the readily available EHR for clinical management.
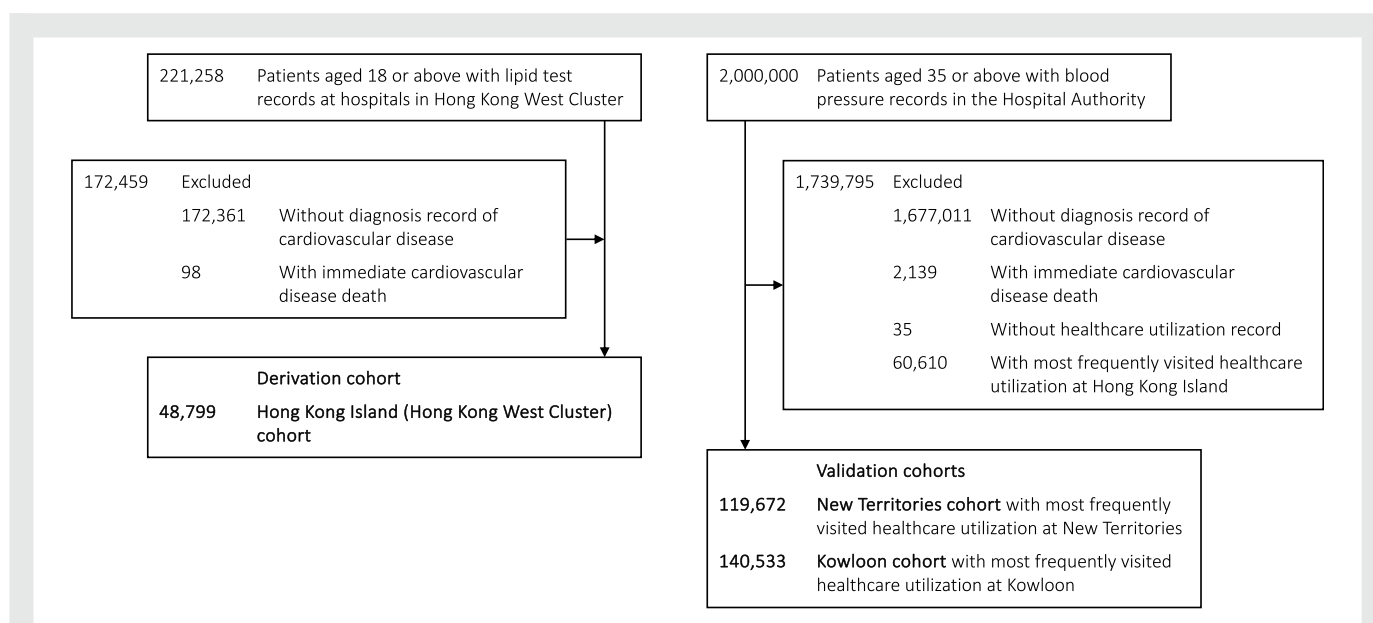
Keywords Cardiovascular diseases • Machine learning • Risk prediction score • Recurrent cardiovascular events

Introduction

Cardiovascular diseases (CVD), including coronary heart disease (CHD) and stroke, are the leading cause of non-communicable deaths globally, with an estimated 18.6 million fatalities recorded in 2019.[1,2] Cardiovascular diseases are also the leading cause of death and disease burden in China, contributing to 3.72 million deaths in 2013 and total hospitalization costs of approximately US $14.5 billion in 2016.[3–5] In Hong Kong, heart disease and cerebrovascular diseases are the third and fourth leading causes of deaths in 2021, respectively.[6] However, according to a World Health Organization report, 80% of premature heart attacks and strokes are preventable.[7]

Some research groups advocate the use of risk prediction models on patients to identify those at high risk of CVD who are more likely to benefit from preventive strategies.[8–11] The development and applicability of CVD risk prediction models are highly dependent on the ethnic and socioeconomic factors of the population of interest.[12] Currently, there are several risk scores for recurrent CVD risk prediction among individuals with established CVD, including The Thrombolysis in Myocardial Infarction (TIMI) Risk Score for Secondary Prevention (TRS-2°P) and Secondary Manifestations of ARTerial disease (SMART2) risk score.[13,14] These risk scores provide an estimated risk of recurrent CVD and thus help provide early intervention to patients with less resource implications.[15] However, these models are tailored

Figure 1 Selection of patients into the study cohorts. N.B. Hong Kong West Cluster is a part of Hong Kong Island.

to western populations and validated on similar model derivation data set, whose applicability to other ethnicities is uncertain. There has been limited validation of the influence of ethnicity on the application of CVD risk scores, and these results are poorly calibrated for Asian populations in Southeast Asia.[16] In addition, although treatment options such as lipid-modifying therapies are effective in secondary prevention among those with established CVD, the estimation of treatment effect is often not considered in current risk scores.[17–19] Therefore, a risk prediction model specifically tailored to the Chinese population for secondary prevention, incorporating dynamic medication treatment with drugs proven to reduce CVD risk, is of paramount importance to identify the means to reduce the CVD healthcare burden.

In this study, we developed and validated the Personalized CARdiovascular DIsease risk Assessment for Chinese (P-CARDIAC) among the Chinese population in Hong Kong using machine learning (ML) technique. The ML technique has been used to identify patterns in large data sets to enable delivery of healthcare services by facilitating effective patient–provider decision-making.[20] The P-CARDIAC was developed to provide early intervention for patients at high risk of recurrent CVD by leveraging the rich data source of electronic health records (EHR). It estimates the 10 years of recurrent CVD risk for high-risk individuals with consideration of an array of risk variables captured in the EHR. We also validated the performance of the P-CARDIAC, TRS-2°P, and SMART2 on the representative study cohorts from Hong Kong, a city in Southeast Asia where over 90% of inhabitants are of Chinese ethnicity.[21]

Methods

Study cohorts

Three cohorts of patients with established CVD were identified based on geographical location of residence in Hong Kong (Hong Kong West Cluster, Hong Kong Island; Kowloon; New Territories). The Hong Kong Island (Hong Kong West Cluster) cohort was used for model derivation, while the Kowloon and New Territories cohorts were used for model validation to ensure no overlapping between derivation and validation cohorts. Patients were included if they had used any of the public healthcare services provided by the Hong Kong Hospital Authority (HA) since 2004 (inclusion

and exclusion criteria detailed in *Figure 1* and Supplementary material online, *Information S1*). The HA is a statutory body and the largest public healthcare provider of Hong Kong. It provides government subsidized primary, secondary, and tertiary care to all residents, capturing over 70% of all hospitalizations in Hong Kong.[22] Previous studies demonstrated high validity of the data source with a positive predictive value of 85% for myocardial infarction (MI) and 91% for stroke.[23] The database was also used for over 200 studies published in peer-reviewed journals, including cardiovascular diseases and cardiovascular drug studies, ensuring the creditability of the data source for research purposes.[23–26]

Each patient was categorized as Hong Kong Island (Hong Kong West Cluster), Kowloon, and New Territories based on the region of their most frequently visited healthcare facility within the study period. Cohort entry date was the date of their first diagnosis of CVD in any inpatient and outpatient setting. Patients were censored at the earliest date of the second record of CVD diagnosis, date of registered death, or study end date (31 December 2019). Patients were excluded from the cohort if they had no diagnosis record of CVD or died on the same day as the first CVD event.

Outcomes and risk variables

The outcome was a composite diagnosis of CHD, ischaemic or haemorrhagic stroke, peripheral artery disease, and revascularization (see Supplementary material online, *Table S5*). The diagnosis of CVD was defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), codes. We estimated the incidence of recurrent CVD events for each cohort with reference to the total person-years of each cohort.

The full list of 125 risk variables (see Supplementary material online, *Table S6*) includes the commonly known risk factors such as age, sex, lipid profile, blood pressure, haemoglobin A1c, and blood glucose, of which 15 were mandatory risk variables and were derived based on clinical evidence, statistically strong correlation, and data completeness to predict CVD risk. Eight classes of medications including lipid-modifying (fibrates, niacin, cholesterol absorption inhibitors, PCSK9 inhibitors, and statins), antihypertensive, antidiabetic, and antiplatelet drugs were considered CVD-related drug use options to observe any changes in CVD risk in the model. Diagnoses and procedures were defined by ICD-9-CM codes (see Supplementary material online, *Table S7*), and medication exposure was defined by the British National Formulary (BNF) sections (see Supplementary material online, *Tables S8 and S9*).

Model derivation

The design of the hybrid statistical–ML model is illustrated in Supplementary material online, *Figure S3*. Feature selection procedure was first applied to all available risk variables to identify mandatory risk variables for model interpretability. Multivariate imputation with chained equations (MICE) was used to generate one imputed data set to replace the missing values of clinical laboratory tests.[27] Multivariate imputation with chained equations is a principled method for dealing with missing data and is extremely reliable on high-dimensional data sets with various missing patterns.[28] For better statistical reliability and clinical utility, risk variables with missing rates below 10% (e.g. clinical laboratory tests) and event rate above 5% (e.g. disease and medication history) were passed for feature selection. We employed a Cox proportional hazards model (CPH) with the least absolute shrinkage and selection operator (LASSO) regularization to shortlist statistically significant (P value < 0.05) risk variables.[29,30] Assumptions were tested to ensure risk associated with the covariates are proportional over time. Cox proportional hazards model is the most widely used multivariate statistical model for survival analysis.[31,32] Its regression coefficients can be interpreted as hazard ratios which can be easily understood by clinicians for better decision-making. The LASSO is a robust feature selection method. It selects the most representative yet independent set of risk variables, which is reliable when downstream manual prioritization is required. Mandatory risk variables were also determined based on clinical relevance to ensure the final set of risk variables are comprehensive and relevant to CVD prognosis. Mandatory risk variables were included in the final model as linear covariates.

For better model performance, the measurement and integration of complex effects from all risk variables in the EHR is important for our model. However, real-world EHR data like our cohorts are highly heterogeneous in form, distribution, and especially completeness. Therefore, we used XGBoost in the P-CARDIAC to fit a tree-ensembled hazard ratio based on all risk variables (see Supplementary material online, *Information S2*). Most ML methods require complete data sets which will cause huge imputation bias in high-dimensional data sets. Compared with other state-of-the-art ML methods, e.g. deep learning (neural network), XGBoost is a gradient boosting decision tree method, for better dealing with heterogeneous tabular data.[33] More importantly, it can work with missing values without imputation. To cancel out the non-linear distribution bias in the raw output of XGBoost, the raw output hazard ratio was first mapped to discrete percentiles, which was tested to largely benefit model calibration performance. To balance the significance between the XGBoost risk score and other risk variables in the final model, the percentiles are then mapped onto a hinge loss-like function (see Supplementary material online, *Information S3*). The P-CARDIAC full model with all 125 risk variables is a CPH model with ridge regularization regressed on the mandatory risk variables and the XGBoost risk score.[34] Ridge regularization is widely used as a stabilizer of regression coefficients, which provides reliable estimates of the hazard ratios of the risk variables. For comparison, a CPH model with only the mandatory risk variables was built as a P-CARDIAC basic model.

Model validation

Internal consistency of model performance was evaluated on the derivation cohort by 100 repeats of 10-fold cross-validation. Model performance of the P-CARDIAC, TRS-2°P, and SMART2 was compared using the validation cohorts with 1000 bootstrap replicates. A high number of repeats were employed to ensure accurate estimation (mean and confidence interval) of model performance statistics.

Calibration performance was assessed graphically by categorizing patients into deciles of predicted 10-year CVD risk and plotting mean 10-year predicted risk against observed 10-year risk. The observed 10-year risk was obtained by the Kaplan–Meier method.[35] Means and confidence intervals of Harrell's C statistic, calibration-in-the-large, and calibration slope were calculated.[36,37] The calibration slope was the slope of linear regression of the observed risk against the predicted risk of each decile. Recalibration was performed if there was overall overestimation or underestimation observed in the calibration curves.[38]

Decision curve analysis was used to estimate the effect of different treatment options across different threshold risks.[39–41] This can identify the range of threshold risks where the model has clinical value (with positive net benefit) and the magnitude of the clinical value. First, a threshold probability (pt) would be chosen to define when a patient is positive. Second, we defined $x = 1$ if the

patient had a predicted probability from the model ≥ pt (the threshold probability) and $x = 0$ otherwise; s(t) was the Kaplan–Meier survival probability at our chosen landmark time t, and N was the number of subjects in the data set. The number of true positives (TP) = $[1 - (s(t) \mid x = 1)] \times P(x = 1) \times N$ and the false positives (FP) = $(s(t) \mid x = 1) \times P(x = 1) \times N$. We calculated the net benefit = TP/N − FP/N × [pt/(1 − pt)] and repeated the above calculation for a reasonable range of threshold probabilities. Finally, we repeated all steps for each model in the study, as well as the default strategies of treat-all and treat-none as if the result is positive. The model with higher net benefits across a larger range of threshold risks is the preferred model. We use decision curve analysis to describe and compare the 10-year clinical value of the P-CARDIAC, TRS-2°P, and SMART2 on the two validation cohorts. TRS-2°P has proposed the specific 3-year risk regarding different risk scores, and we extrapolated the predicted 3-year risk to 10-year risk by multiplying the ratio of the corresponding Kaplan–Meier estimated risks for each of the two cohorts.

All analysis were conducted using Python (version 3.9.1) with add-on package lifelines.[42] This study report is in accordance with the TRIPOD statement.[43] Ethical approval for this study was granted by the Institutional Review Board of the University of Hong Kong/HA Hong Kong West Cluster (UW20-073).

Results

Study cohorts

For the derivation cohort, we identified 221 258 patients aged 18 or above with lipid test records between 1 January 2004 and 31 December 2019. We excluded 172 459 patients from the cohort who had no diagnosis record of CVD or died of the first CVD event on the same date. Overall, 48 799 patients were included in the derivation cohort.

For the validation cohorts, we initially identified a cohort of 2 million patients aged 35 or above with blood pressure records in the HA between 1 January 2005 and 31 December 2019. We excluded 1 679 150 patients who had no diagnosis record of CVD or died of the first CVD event on the same date. We excluded 60 645 patients without healthcare utilization records or with the most frequently visited healthcare facility at Hong Kong Island. Overall, 119 672 patients were included in the New Territories cohort, and 140 533 patients were included in the Kowloon cohort. A flowchart of patient selection is illustrated in *Figure 1*.

Incidence rates of cardiovascular disease and baseline characteristics

Supplementary material online, *Table S1*, shows the event rates of CVD across three cohorts. The event rate per 1000 person-years was 219 to 241, while the median estimated 10-year event rate was 71.7–76.1%, respectively. During a median follow-up of 0.3–1.0 year, 55–64% of patients had cardiovascular disease recurrences. Regarding the composition of incident CVD events, CHD was the most common, with composition around 61–65%, of which MI had a ratio of ~9–10%. Stroke was the second most common outcome with a ratio of ~33–39%. The ratio of peripheral arterial disease (PAD) was around 3–4%.

All subtypes of incidence events in the derivation cohort had significantly different distribution from the validation cohorts. The proportion of total CVD events was higher. The proportion of CHD, MI, PAD, and revascularization was higher, while the proportion of stroke and fatal events was lower. Supplementary material online, *Tables S2 and S10*, showed the baseline characteristics of the risk variables across three cohorts.

Model derivation

We identified 15 mandatory risk variables and 8 CVD-related drug use options (*Table 1*) that were statistically significant and medically coherent for CVD pathogenesis. Multivariate imputation with chained equations was conducted once with <2% missing rate among the 15 mandatory risk variables. For both the basic and full models, all risk

Table 1 Adjusted hazard ratios in the Personalized CARdiovascular DIsease risk Assessment for Chinese models

	Basic model		Full model	
	(Mandatory risk variables)		(Mandatory + supplementary risk variables)	
	HR (95% CI)	P value	HR (95% CI)	P value
General				
Age per year	1.02 (1.01–1.02)	<0.0001	1.01 (1.01–1.01)	<0.0001
Female	0.84 (0.82–0.86)	<0.0001	0.86 (0.84–0.88)	<0.0001
Accident and emergency visits per year (prior to incident cardiovascular events)	1.07 (1.06–1.08)	<0.0001	1.06 (1.05–1.07)	<0.0001
Clinical laboratory tests				
Low-density lipoprotein cholesterol (mmol/L)	1.06 (1.05–1.08)	<0.0001	1.05 (1.04–1.06)	<0.0001
Neutrophil (10^9/L)	1.02 (1.02–1.03)	<0.0001	1.02 (1.02–1.02)	<0.0001
Aspartate transaminase: alanine aminotransferase ratio	1.02 (1.02–1.03)	<0.0001	1.02 (1.01–1.02)	<0.0001
Disease and medication history				
Statins	0.84 (0.82–0.87)	<0.0001	0.88 (0.85–0.90)	<0.0001
Hypertension	1.16 (1.13–1.19)	<0.0001	1.13 (1.10–1.16)	<0.0001
Diabetes	1.38 (1.34–1.43)	<0.0001	1.30 (1.25–1.35)	<0.0001
Atrial fibrillation	1.09 (1.05–1.13)	<0.0001	1.08 (1.04–1.12)	0.0001
Myocardial infarction	2.13 (2.06–2.21)	<0.0001	1.71 (1.65–1.78)	<0.0001
Angina	0.92 (0.88–0.96)	0.0003	0.93 (0.89–0.97)	0.0022
Revascularization	0.91 (0.88–0.95)	<0.0001	0.93 (0.90–0.96)	<0.0001
Family history of diabetes	1.37 (1.32–1.43)	<0.0001	1.28 (1.23–1.33)	<0.0001
Drug use (interactive covariates)				
Antihypertensive drugs	0.67 (0.65–0.69)	<0.0001	0.77 (0.74–0.79)	<0.0001
Antidiabetic drugs	0.71 (0.69–0.74)	<0.0001	0.77 (0.74–0.80)	<0.0001
Antiplatelet drugs	0.78 (0.75–0.80)	<0.0001	0.85 (0.83–0.87)	<0.0001
Fibrates	0.78 (0.73–0.84)	<0.0001	0.78 (0.73–0.84)	<0.0001
Niacin	0.53 (0.38–0.75)	0.0003	0.56 (0.40–0.78)	0.0007
Cholesterol absorption inhibitors	0.55 (0.49–0.63)	<0.0001	0.56 (0.49–0.63)	<0.0001
PCSK9 inhibitors	0.24 (0.09–0.68)	0.0066	0.25 (0.09–0.69)	0.0078
Statins	0.87 (0.85–0.90)	<0.0001	0.89 (0.86–0.91)	<0.0001
XGBoost risk score			1.03 (1.02–1.03)	<0.0001

HR, hazard ratio; CI, confidence interval; PCSK9, proprotein convertase subtilisin/kexin type 9.

variables were statistically significant (P value < 0.05) when compared with those without recurrent CVD. Both models had similar estimates on the linear effects of the risk variables, while the basic model's hazard ratios deviated more from 1 than the full model with wider 95% CIs, indicating more precise estimates in the full model. The similar hazard ratios between the models reassure the consistent risk estimation across the two models.

Model validation

Validation results on the derivation cohort of the P-CARDIAC full model showed satisfying discrimination and calibration performance. The C statistic was 0.69, the calibration slope was 1.00, and the calibration-in-the-large was 0.03. There was slight overestimation across risk deciles. The P-CARDIAC basic model showed modest discrimination and calibration performance but was inferior to the full model. The C statistic was 0.66, the calibration slope was 0.86, and the calibration-in-the-large was 0.01. There was slight overestimation in high-risk patients and underestimation in low-risk patients. The internal validation results are shown in *Figure 2* and *Table 2*.

Internal validation of the P-CARDIAC full model across validation cohorts showed good discrimination and calibration performance. The C statistic for the Kowloon and New Territories cohorts were 0.62 and 0.64, the calibration slope was 0.75 and 0.93, and the calibration-in-the-large was 0.04 and 0.01, respectively. There was overestimation for high-risk patients (predicted 10-year risk >80%) for the Kowloon cohort. There was overestimation on all patients for the New Territories cohort which was largely mitigated by recalibration (see Supplementary material online, *Figure S4*). The P-CARDIAC basic model showed good discrimination and calibration performance but was inferior to the full model. The C statistic for Kowloon and New Territories cohorts were 0.60 and 0.62, the calibration slope was 0.66 and 0.75, and the calibration-in-the-large was 0.01 and 0.03, respectively. There was overestimation in high-risk patients and underestimation in low-risk patients for both cohorts which could not be mitigated by recalibration. Validation of both TRS-2°P and SMART2 risk scores underperformed regarding discrimination and risk stratification performance. The C statistic was lower than 0.55 for both validation cohorts. The validation results are summarized in Supplementary material online, *Figure S1.1–1.2*, *Table 3*, and Supplementary material online, *Tables S3 and S4*.
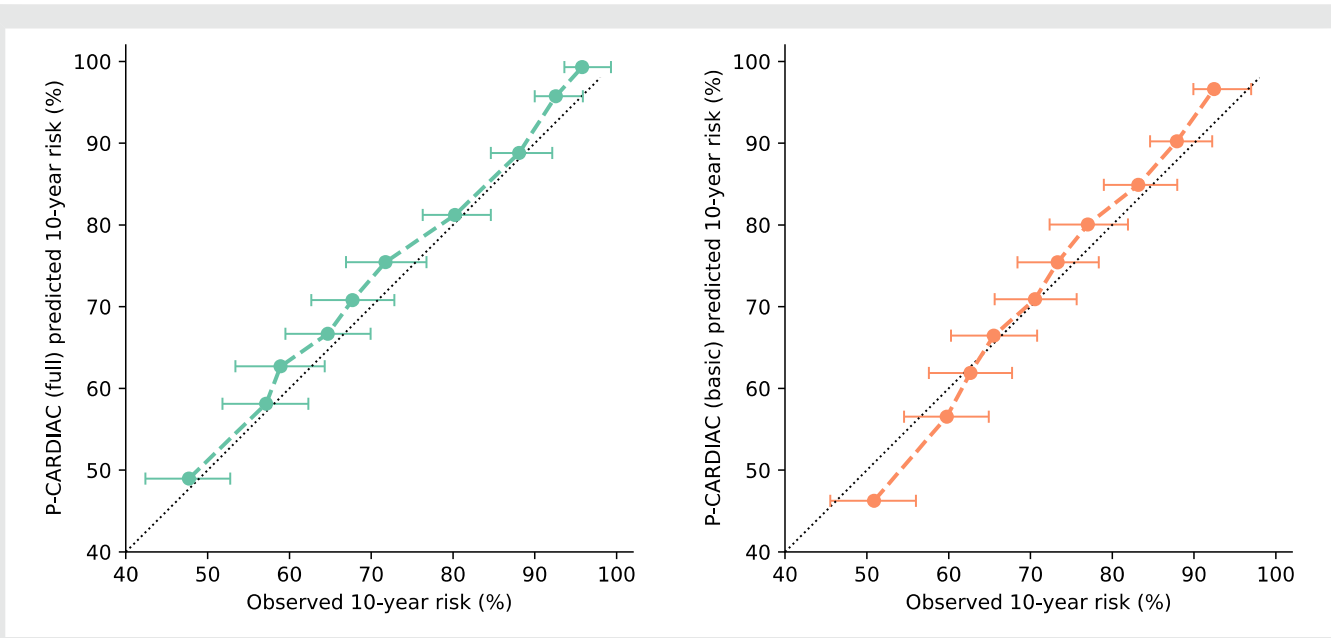
Figure 2 Calibration plots for the P-CARDIAC (full) model in the Hong Kong Island (Hong Kong West Cluster) derivation cohort with 95% confidence interval. Results were measured from 10-fold cross-validation.

Table 2 Discrimination and calibration performance of the Personalized CARdiovascular DIsease risk Assessment for Chinese on derivation cohort

	Harrell's C statistic	Calibration slope	Calibration-in-the-large
Basic model	0.66 (0.66, 0.66)	0.86 (0.86, 0.86)	0.01 (0.01, 0.01)
Full model	0.69 (0.69, 0.69)	1.00 (1.00, 1.00)	0.03 (0.03, 0.03)

Harrell's C statistic is a measure of model discrimination with values ranging from 0.5 to 1, i.e. probability of correct ordering for a randomly selected pair of subjects. Calibration slope is a measure of model calibration with target value of 1. Values smaller than 1 indicate overfitting, i.e. too low for low-risk patients and/or too high for high-risk patients. Values >1 indicate underfitting, i.e. too high for low-risk patients and/or too low for high-risk patients. Calibration-in-the-large is a measure of model calibration with target value of 0. Values >0 means the model overestimates risk in general. Values smaller than 0 means the model underestimates risk in general. Results were measured from 100 repeats of 10-fold cross-validation.

In summary, the P-CARDIAC showed great performance on the three derivation and validation cohorts. The full model had better performance than the basic model as it accurately accounted for the nonlinear effects and the effects from supplementary risk variables. On the other hand, TRS-2°P and SMART2 underperformed when adapted to the two cohorts for Chinese populations.

Clinical utility

Decision curve analysis of the two validation cohorts was similar (see Supplementary material online, *Figure S2*). The P-CARDIAC full model performed better than the P-CARDIAC basic model. Both P-CARDIAC models had similar and greater net benefits across a larger range of threshold risks compared with the treat-all strategy, TRS-2°P, and SMART2. The P-CARDIAC had clinical values for decision-making when the threshold risk was under 90%.

Website design

The website interface at p-cardiac.com was designed to be flexible and interactive (see Supplementary material online, *Information S4,* for example screenshots). Users can input up to 15 risk variables in the mandatory field for a quick evaluation of CVD risk. More than 100 risk variables can be further inputted in the supplementary field for a more comprehensive evaluation. The more risk variables submitted in the supplementary field, the more accurate the prediction. Furthermore, the drug use risk variables were designed as interactive selection options, where up to 8 types of drug classes could be selected for evaluation of potential synergetic treatment effects to guide possible treatment plans.

Discussion

To the best of our knowledge, this is the first model to predict recurrent CVD events in a Chinese population from a large contemporary Chinese cohort using ML technique. The P-CARDIAC demonstrated reliable performance of recurrent CVD risk prediction in 10 years on three derivation and validation cohorts. We demonstrated that P-CARDIAC models have better performance in risk prediction than existing CVD risk scores such as TRS-2°P and SMART2 that were developed on western populations. Our results also demonstrate that the P-CARDIAC full model has superior performance to the basic model.

In addition, the effects of concurrent drug use are often neglected in existing CVD risk scores. In this study, we included exposures of various drug classes as interactive covariates in the model to evaluate their bias-mitigated, risk-stratified, and Chinese-specific treatment effects. Among the eight drug classes included in the interactive covariates, all classes had hazard ratios lower than 1, while PSCK9 inhibitors had the lowest.

Table 3 Mean (95% confidence interval) of Harrell's C statistic on validation cohorts

	P-CARDIAC (full)	P-CARDIAC (basic)	SMART2	TRS-2°P
Kowloon	0.62 (0.62, 0.62)	0.60 (0.60, 0.60)	0.55 (0.55, 0.55)	0.53 (0.53, 0.53)
New Territories	0.64 (0.64, 0.64)	0.62 (0.62, 0.62)	0.55 (0.55, 0.55)	0.54 (0.54, 0.54)

A measure of model discrimination with values ranging from 0.5 to 1, i.e. probability of correct ordering for a randomly selected pair of subjects. Values were measured from 1000 bootstrap replicates.

This observation indicates that drug treatment with indications for risk variable CVD such as lipid-modifying drugs, antihypertensive, and antidiabetic drugs has a beneficial effect on reducing CVD risk. In addition, our model also considers prior statin use for primary prevention prior to the first CVD event. We found that patients who received statins as primary prevention prior to the first CVD event had a lower risk of recurrent CVD events, independent of whether they continued statin therapy. We believe the P-CARDIAC is the first risk prediction model to include these risk variables and offer CVD-related drug use for recalculation in CVD risk prediction, highlighting the novelty of our approach.

The P-CARDIAC was developed using hybrid statistical–ML algorithms, which is novel in the field of CVD risk prediction. To facilitate efficient clinical management, comprehensive electronic systems were developed, thus providing sizable clinical data for better development of computational models. However, as the pool of covariates becomes increasingly larger, there is a dilemma in the development of medical prediction models, where it may be challenging to balance interpretability and performance. Traditional prediction tools rely on the linear combinations of a selected pool of small number of covariates, which are easily interpreted but do not consider the massive non-linear effects and often lack accuracy. On the other hand, in recent years, many ML and deep learning methods have emerged that take into consideration the complex relationships of all massive covariates to yield high accuracy. However, since these models lack linear representations of the covariates, the effects of the risk variables are uncertain and unclear.[44] Therefore, the ML approach is described as the 'black box approach. Our proposed methodology adopts the traditional approach by selecting a pool of clinically relevant covariates using statistical methods and then considers the large number of covariates and their complex effects using the ML method as another non-interfering component for better model fit. We used XGBoost as the ML method. XGBoost is a tree-based ensemble method that does not require complete values in the large pool of covariates which circumvents the potential imputation bias. This novel hybrid method showed significantly better performance than the traditional statistical method by comprehensively considering a large pool of covariates, including commonly known risk factors, such as blood pressure, haemoglobin A1c, blood glucose, and lipid profile where its interpretability is still evident. The novel hybrid method is customisable and can be used for other studies.

This study has limitations. First, the P-CARDIAC was developed using real-world data and any change in clinical practice in the future may result in changes to the predicted recurrent CVD risk among patients. The advantage of the ML approach is that recalibration and fine-tuning the model can be done as more data are accrued. Therefore, the model can be calibrated periodically to account for any changes in clinical practice. Second, the P-CARDIAC was developed based on a population of predominantly Chinese in Hong Kong. Although the P-CARDIAC demonstrated superior performance than TRS-2°P and SMART2 in our study cohorts, we cannot rule out the possibility that the P-CARDIAC may underperform if applied to the study cohorts used to develop TRS-2°P and SMART2. Hence, recalibration is needed for use in populations of other ethnicities and Chinese from other regions such as Mainland China. Third, manual input of more than 100 risk variables is time-consuming and not practical in fast-paced clinical

settings. Therefore, we aim to automate the process of data entry (such as plug-ins or APIs) by leveraging the readily available EHR for clinical management to provide timely risk estimation. Last, the P-CARDIAC serves as a risk stratification tool to better utilize healthcare resources rather than a diagnostic tool; thus, a composite risk score was given for a spectrum of CVD diseases rather than a score for each specific disease. In the near future, we plan to generate evidence to support the effective implementation of the P-CARDIAC in clinical setting. The advanced technologies currently available enable the harnessing of the power of Big Data. However, we believe that the empathy of healthcare providers and their connection with patients which influences the best decision on care will not be replaced by artificial intelligence (AI) in the near future.

Conclusions

We developed and validated the P-CARDIAC, a new CVD risk prediction model for recurrent CVD events among Chinese adults with established CVD. Compared with TRS-2°P and SMART2, the P-CARDIAC was able to identify unique patterns of Chinese patients with established CVD with good performance. The consideration of treatment effects of various drug use could also guide improved and individualized secondary prevention.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Commission, and the National Health and Medical Research Council in Australia; has received speaker fees from Janssen and Medice in the previous 3 years; and is an independent non-executive director of Jacobson Medical in Hong Kong. C.S.L.C. has received grants from the Food and Health Bureau of the Hong Kong Government, Hong Kong Research Grant Council, Hong Kong Innovation and Technology Commission, Pfizer, IQVIA, MSD, and Amgen and personal fees from PrimeVigilance, outside the submitted work. All other authors declare no competing interests.

Data availability

Data will not be available for others as the data custodians have not given permission.

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, *et al.* Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol* 2020;**76**:2982–3021.
2. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020;**396**:1204–1222.
3. Zhou M, Wang H, Zhu J, Chen W, Wang L, Liu S, *et al.* Cause-specific mortality for 240 causes in China during 1990–2013: a systematic subnational analysis for the Global Burden of Disease Study 2013. *Lancet* 2016;**387**:251–272.
4. He J, Gu D, Wu X, Reynolds Ki, Duan X, Yao C, *et al.* Major causes of death among men and women in China. *N Engl J Med* 2005;**353**:1124–1134.
5. National Center for Cardiovascular Disease. Report on cardiovascular diseases in China. Encyclopedia of China Publishing House; 2019. https://www.nccd.org.cn/Sites/Uploaded/File/2021/3/中国心血管病报告2018(English).pdf.
6. Death rates by leading causes of death, 2001–2021. https://www.chp.gov.hk/en/statistics/data/10/27/117.html
7. Cardiovascular diseases: avoiding heart attacks and strokes. 2015. https://www.who.int/news-room/questions-and-answers/item/cardiovascular-diseases-avoiding-heart-attacks-and-strokes#:~:text=Healthy%20diet%2C%20regular%20physical%20activity,diabetes%20is%20also%20very%20important.
8. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, *et al.* 2016 European guidelines on cardiovascular disease prevention in clinical practice: the sixth joint task force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J* 2016;**37**:2315–2381.
9. van der Leeuw J, Ridker PM, van der Graaf Y, Visseren FL. Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects. *Eur Heart J* 2014;**35**:837–843.
10. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AMJ, Paynter NP, Steyerberg EW, *et al.* Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ* 2011;**343**:d5888.
11. Cooney MT, Selmer R, Lindman A, Tverdal A, Menotti A, Thomsen T, *et al.* Cardiovascular risk estimation in older persons: SCORE OP. *Eur J Prev Cardiol* 2016;**23**:1093–1103.
12. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;**353**:i2416.
13. Hageman SHJ, McKay AJ, Ueda P, Gunn LH, Jernberg T, Hagström E, *et al.* Estimation of recurrent atherosclerotic cardiovascular event risk in patients with established cardiovascular disease: the updated SMART2 algorithm. *Eur Heart J* 2022;**43**:1715–1727.
14. Bohula EA, Bonaca MP, Braunwald E, Aylward PE, Corbalan R, De Ferrari GM, *et al.* Atherothrombotic risk stratification and the efficacy and safety of vorapaxar in patients with stable ischemic heart disease and previous myocardial infarction. *Circulation* 2016;**134**:304–313.
15. Grant SW, Collins GS, Nashef SA. Statistical primer: developing and validating a risk prediction model. *Eur J Cardiothorac Surg* 2018;**54**:203–208.
16. Huang D, Cheng YY, Wong YT, Yung SY, Chan KW, Lam CC, *et al.* TIMI risk score for secondary prevention of recurrent cardiovascular events in a real-world cohort of post-non-ST-elevation myocardial infarction patients. *Postgrad Med J* 2019;**95**:372–377.
17. Temporelli PL, Arca M, D'Erasmo L, De Caterina R. Lipid-lowering therapy in patients with coronary heart disease and prior stroke: mission impossible? *J Clin Med* 2021;**10**:886.
18. Gutierrez J, Ramirez G, Rundek T, Sacco RL. Statin therapy in the prevention of recurrent cardiovascular events: a sex-based meta-analysis. *Arch Intern Med* 2012;**172**:909–919.
19. Gynnild MN, Hageman SH, Spigset O, Lydersen S, Saltvedt I, Dorresteijn JAN, *et al.* Use of lipid-lowering therapy after ischaemic stroke and expected benefit from intensification of treatment. *Open heart* 2022;**9**:e001972.
20. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC Med Inform Decis Mak* 2021;**21**:54.
21. The demographics in Hong Kong: ethnic groups. In: Race Relations Unit, Home Affairs Department - The Government of the Hong Kong Special Administrative Region. https://www.had.gov.hk/rru/english/info/demographics.htm.
22. Census and Statistics Department. Hong Kong Special Administrative Region. Thematic Household Survey Report No. 68. https://www.censtatd.gov.hk/en/data/stat_report/product/B1130201/att/B11302682019XXXXB0100.pdf.
23. Wong AY, Root A, Douglas IJ, Chui CSL, Chan EW, Ghebremichael-Weldeselassie Yonas, *et al.* Cardiovascular outcomes associated with use of clarithromycin: population based study. *BMJ* 2016;**352**:h6926.
24. Li X, Tong X, Yeung WWY, Kuan P, Yum SHH, Chui CSL, *et al.* Two-dose COVID-19 vaccination and possible arthritis flare among patients with rheumatoid arthritis in Hong Kong. *Ann Rheum Dis* 2022;**81**:564–568.
25. Chui CSL, Fan M, Wan EYF, Leung MTY, Cheung E, Yan VKC, *et al.* Thromboembolic events and hemorrhagic stroke after mRNA (BNT162b2) and inactivated (CoronaVac) COVID-19 vaccination: a self-controlled case series study. *EClinicalMedicine* 2022;**50**:101504.
26. Lai FTT, Li X, Peng K, Huang L, Ip P, Tong X, *et al.* Carditis after COVID-19 vaccination with a messenger RNA vaccine and an inactivated virus vaccine: a case–control study. *Ann Intern Med* 2022;**175**:362–370.
27. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;**45**:1–67.
28. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;**20**:40–49.
29. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodol)* 1996;**58**:267–288.
30. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B (Methodol)* 1972;**34**:187–202.
31. Deo SV, Deo V, Sundaram V. Survival analysis—part 2: cox proportional hazards model. *Indian J Thorac Cardiovasc Surg* 2021;**37**:229–233.
32. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*: New York: Springer; 2001.
33. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016. p. 785–794.
34. Tihonov AN. Solution of incorrectly formulated problems and the regularization method. *Soviet Math* 1963;**4**:1035–1038.
35. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;**53**:457–481.
36. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;**247**:2543–2546.
37. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;**17**:230.
38. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Stat Methods Med Res* 2016;**25**:1692–1706.
39. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–138.
40. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak* 2008;**8**:53.
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565–574.
42. Davidson-Pilon C. Lifelines: survival analysis in Python. *J Open Source Softw* 2019;**4**:1317.
43. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;**13**:1.
44. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018;**320**:2199–2200.