BMJ Open

EFFICACY OF CARDIOVASCULAR DISEASE RISK PREDICTION USING MACHINE LEARNING COMPARED TO WORLD HEALTH ORGANIZATION RISK CHARTS FOR SOUTH-EAST ASIANS

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-081434
Article Type:	Original research
Date Submitted by the Author:	28-Oct-2023
Complete List of Authors:	Mettananda, Chamila; University of Kelaniya, Pharmacology Solangaarachchige, Maheeka; University of Kelaniya, Exam Unit; Sri Lanka Institute of Information Technology, Haddela, Prasanna; Sri Lanka Institute of Information Technology, Department of IT Dassanayake, Anuradha; University of Kelaniya Faculty of Medicine, Pharmacology Kasturiratne, Anuradhani; University of Kelaniya Faculty of Medicine, Public Health Wickremasinghe, Rajitha; University of Kelaniya Faculty of Medicine, Public Health Kato, Norihiro; National Center for Global Health and Medicine Research Institute, Gene Diagnostics and Therapeutics de Silva, Hithanadura; University of Kelaniya Faculty of Medicine, Medicine
Keywords:	Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PREVENTIVE MEDICINE, Primary Prevention, Cardiac Epidemiology < CARDIOLOGY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

EFFICACY OF CARDIOVASCULAR DISEASE RISK PREDICTION USING MACHINE LEARNING COMPARED TO WORLD HEALTH ORGANIZATION RISK CHARTS FOR SOUTH-EAST ASIANS

Mettananda C[1†], Solangaarachchige MB[1,2†], Haddela PS[2], Dassanayake AS[1], Kasturiratne A[1], Wickramasinghe AR[1], Kato N[3], de Silva HJ[1]

[1] Faculty of Medicine, University of Kelaniya, Sri Lanka

[2] Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

[3] National Centre for Global Health and Medicine, Toyama, Shinjuku-ku, Tokyo, Japan

† - Contributed equally to this work

Correspondence to

Chamila Mettananda

Chamila@kln.ac.lk, chamilametta@hotmail.com

Main text – 2900 words. Abstract – 284words

Abbreviated title - efficacy of cardiovascular disease risk prediction using machine learning

Abstract

Introduction and objectives

Models derived from non-Sri Lankan cohorts are currently being used for cardiovascular (CV) risk stratification of Sri Lankans. We aimed to develop a CV risk prediction model using machine learning (ML) based on data from a Sri Lankan cohort followed up for 10 years, and to compare the predictions with World Health Organization (WHO) risk charts.

Methods

Using 10-year follow-up data for 2596 Sri Lankans without CV diseases at baseline, we developed two ML models for predicting 10-year CV risk using 6 conventional CV risk variables (age, gender, smoking status, systolic blood pressure, history of diabetes, and total cholesterol level) and all available variables (n=75). The ML models were derived using classification algorithms of the supervised learning technique. We compared the predictive performance of our ML models with WHO risk charts (2019, Southeast Asia) using "area under the receiver operating characteristic curves" (AUC-ROC). The 6-variable model was further validated in an external cohort.

Results

The baseline cohort consisted of individuals aged 40-64 years, selected by stratified random sampling of a semi-urban health administrative area in Sri Lanka in 2007. During a 10-year follow-up period, 179 incident CV events (CVEs) were recorded. CV risk predictions improved with 6-variable (accurate prediction of 125 CVEs; AUC-ROC: 0.72, CI-0.66-0.78) and 75-variable ML models (124 CVEs; AUC-ROC: 0.74, CI-0.68-0.80), compared to the WHO risk charts (10 CVEs; AUC-ROC: 0.51, CI-0.42-0.60). In the external validation cohort, sensitivity, specificity, positive-predictive-value and negative-predictive-value of the 6-variable model were 70.3%, 94.9%, 87.3%, 86.6% and the WHO risk charts were 23.7%, 79%, 35.8%, 67.7%.

Conclusions

ML-based models derived from a cohort of Sri Lankans improved the overall accuracy of CV-risk prediction compared to the WHO risk charts for South-East Asians.

Keywords – Cardiovascular risk, prediction, World Health Organization risk charts, Machine learning, validation, Sri Lanka

Key messages

What is already known on this topic – World Health Organization (WHO) risk charts are currently the best available for cardiovascular risk prediction of Sri Lankans. However, they were designed to include the whole of South-East Asia and seem less sensitive among high-risk Sri Lankans.

What this study adds – We developed a risk prediction model using machine learning based on data from a Sri Lankan cohort followed up for 10 years and compared the predictions with WHO risk charts. We showed that ML-based models improved the accuracy of cardiovascular risk prediction of Sri Lankans compared to the WHO risk charts in an external cohort.

How this study might affect research, practice, or policy – The new model is more sensitive in predicting Sri Lankans at high cardiovascular risk than the WHO risk charts for South-East Asia. More accurate risk prediction will facilitate the implementation of cost-effective primary prevention strategies. Further, machine learning of data of long-term follow-up cohorts can be used to develop cardiovascular risk prediction models for countries that do not have reliable risk prediction models.

Introduction

There are no cardiovascular (CV) risk prediction models specific to or derived from Sri Lankans. Therefore, different risk prediction models derived from white Caucasians, or models developed for the South-East Asia region (SEAR) are being used for CV risk stratification of Sri Lankans.

Asians behave differently from white Caucasians in terms of CV risk. Asians have a distinct genetic make-up, and a different CV risk factor profile with a higher prevalence of hypertension, diabetes mellitus, central obesity, insulin resistance, and metabolic syndrome than white Caucasians (1). They are also at increased risk of developing CV diseases (CVDs) compared to white Caucasians at a given risk factor level (1). In Sri Lankans, there is low agreement between the CV risk predictions based on the World Health Organization / International Society of Hypertension (WHO/ISH) risk charts and the Framingham General CV risk charts (2). Moreover, the CV risk predictions in a Sri Lankan cohort using three different risk models, the National Cholesterol Education Program - Adult Treatment Panel III (NCEP-ATP III), WHO/ISH charts and Systematic Coronary Risk Evaluation (SCORE) charts, were found discordant (3).

The WHO/ISH CV risk charts for the South-East Asia region-B (SEAR-B) were developed in 2007 together with for another 14 epidemiological sub-regions to risk predict people of those regions that did not have specific risk prediction models derived from their own cohorts (4). These 2007 WHO/ISH risk charts have been validated in Sri Lankans (4) and they showed 81% agreement between predictions and observed events but were less predictive in females and those at high CV risk (5). Later on, the WHO risk charts were revised and re-calibrated in 2019 to improve predictive capacity as well as to include 21 epidemiological sub-regions that did not have specific risk prediction models. These 2019 WHO risk charts are currently the best available for Sri Lankans (6). However, in this also, Sri Lanka is grouped under the South-East Asia epidemiological sub-region together with Indonesia, Cambodia, Laos, Sri Lanka, Maldives, Myanmar, Malaysia, Philippines, Thailand, Timor-Leste, Viet Nam, Mauritius, and Seychelles. This is a heterogeneous population, with different socio-

economic and cultural backgrounds and therefore, the risk predictions may not accurately represent the CV risk of Sri Lankans.

Therefore, we aimed to develop a CV risk prediction model using machine learning (ML) based on data from a Sri Lankan cohort followed up for 10 years, and to compare the predictions with 2019 WHO (South-East Asia) risk charts. Moreover, we aimed to validate the new model in an external cohort of Sri Lankans.

Materials and methods

We developed two CV risk prediction models using ML, based on data from a large community-based study on non-communicable diseases, the "Ragama Health Study (RHS)" (3, 7), where individuals have been followed up from 2007 to date.

The baseline study population (n=2923) in the RHS comprised 35–64 years old, adult residents in the "Ragama Medical Officer of Health (MOH) area" in 2007. Participants were selected by stratified random sampling in the Ragama MOH area, which is a semi-urban health administrative area among 25 districts in Sri Lanka. Participants were followed up for 10 years from 2007 to 2017; during which all CV deaths, non-fatal strokes, and non-fatal myocardial infarctions (including those undergoing percutaneous coronary interventions and coronary artery bypass grafts) were recorded as hard CV events (CVE) by either interviewing patients and their families or perusing clinical notes/death certificates.

Data for participants above 40 years of age, who had no history of CVDs at enrolment in 2007 and completed 10-year follow-up (n=2596), were extracted to develop ML-based risk prediction models, as usually risk predictions are calculated in people over the age of 40 years.

Using the 10-year prospective follow-up data for the cohort, using baseline data of those who developed CVEs and those who did not, we developed two ML-based models to predict the 10-year risk of developing a hard CVE using different risk factor combinations. Individuals who could not be traced in 2017 or those whose cause of death could not be verified were excluded. The ML-based models were developed using classification algorithms of the supervised learning technique. The models were

developed in a recursive process (8) in four steps: project design, data preparation, model fitting and inference & deployment (Figure 1). Using the database, models were built with the publicly available Google Colab ML platform and Scikit-learn library in Python programming language (9) and Train-Test Split method (10). Participant data were split into two groups; the training sample and the testing sample. The training sample was used to build the ML-based models and the testing sample was used to assess the efficacy of the algorithms built using the training sample. Since the ratio of CVE to non-CVE was highly skewed at 7:93, we performed stratified 10-fold cross-validation, using 2336 individuals for the training sample and the remaining 260 for the test sample to prevent over-fitting. Predictive performances of the models were compared using "area under the receiver operating characteristic curve (AUC-ROC)". The mean of AUC-ROCs for the 10 cross-validation samples was taken as the AUC-ROC of the ML-based model in question.

Figure 1

We trialled six standard ML classification algorithms with different modelling approaches, namely, Decision tree, Random forest, k-nearest neighbour, 2D neural networks, AdaBoost and gradient boosting. The best-fitting model in terms of AUC-ROC was selected to develop the final model. Grid search was used to optimize the hyper-parameters of the models (11). Data imputation for all models was done using the statistical imputation of missing values using Python.

We developed two risk prediction models; one using the 6 conventional CV risk variables that are used in the WHO CV risk charts (age, gender, smoking status, systolic blood pressure, history of diabetes, and total cholesterol level) and the other using 75 variables. The total database consisted of 770 variables, including data on demographics, medical history, family history, social history, physical examination, laboratory investigations and non-laboratory investigations like ECG and an ultrasound scan of the abdomen. Following data wrangling and cleaning, we arbitrarily chose 75 (out of 770) variables with a missing value rate of <50% for the ML model development.

We calculated the predicted CVEs over 10 years by 2017, using baseline data (2007 data) and the two ML models separately. Additionally, we calculated the same using the latest 2019 WHO CV risk charts. We compared the predictions of the 6-variable and 75-variable ML models and the WHO model against the observed events using AUC-ROC.

Further, we externally validated the 6-variable ML model in a separate hospital-based database of 357 consecutive patients, 40–74 years of age admitted to Colombo North Teaching Hospital (a tertiary care hospital of Sri Lanka) from 1st of January 2019 to 1st of August 2020 who did not have a history of CVEs and presented with an acute incident CVE (acute myocardial infarction or acute stroke) or a disease other than an acute CVE who had complete data for CVD risk calculation. Their predicted risks of developing a CVE were calculated using the most recent pre-morbid risk factor data available up to one year before developing the incident CVE or the admission to the ward in non-CVE cases. We compared the predictions of the 6-variable model with that of the 2019 WHO risk chart using confusion matrix.

This study was approved by the Ethics Review Committee of the Faculty of Medicine, University of Kelaniya, Sri Lanka and written informed consent was obtained from all the participants.

Patient and Public Involvement statement: It was not appropriate or possible to involve patients or the public in the design or reporting plans of our research but was involved in the conduct and the dissemination of the study. All patients are routinely followed up in a non-communicable disease clinic at the Faculty of Medicine, in collaboration with North Colombo Teaching Hospital (NCTH) Ragama, Sri Lanka as a service component since 2007 to date. Information about their risk factors was available to participants and when necessary they were referred for specialist care at the NCTH. The results of the study will be disseminated to study participants, other patients and the public following the publication of the study.

Results

A total of 2596 participants followed up for 10 years were eligible for the study with a mean age of 53.5 (SD: 6.9) years and 1162 (44.8%) males. The baseline characteristics of the study cohort are shown in Table 1.

Table 1 Baseline characteristics of the cohort

	Male n = 1162	Female n = 1434	Total n = 2596
Ethnicity n (%)			
Sinhalese	1118 (96.2)	1375 (95.9)	2493 (96.0)
Tamil	15 (1.3)	27 (1.9)	42 (1.6)
Muslim	2 (0.2)	2 (0.1)	4 (0.2)
Burgher	15 (1.3)	19 (1.3)	34 (1.3)
Other	12 (1.0)	11 (0.8)	23 (0.9)
Age groups (years), n (%)			
40-49.9	360 (30.9)	456 (31.8)	816 (31.4)
50-59.9	526 (45.3)	669(46.7)	1195 (46.0)
≥ 60.0	276 (23.8)	309(21.5)	585 (22.6)
Smoking, n (%)	416 (35.8)	0 (0.0)	416 (16.0)
Diabetes, n (%)	165 (14.2)	249 (17.4)	414 (15.9)
Hyperlipidaemia, n (%)	98 (8.4)	209 (14.6)	307 (11.8)
SBP (mmHg), n (%)			
<139.9	766 (65.9)	869 (60.6)	1635 (62.9)
140-159.9	260 (22.4)	365 (25.5)	625 (24.1)
160-179.9	88 (7.6)	132 (9.2)	220 (8.5)
≥180.0	48 (4.1)	68 (4.7)	116 (4.5)
Total cholesterol (mmol/L), n (%)			
<4.0	211 (18.2)	207 (14.4)	418 (16.1)

4-4.9	297 (25.6)	269 (18.8)	566 (21.8)
5-5.9	391 (33.6)	476 (33.2)	867 (33.4)
6-6.9	192 (16.5)	322 (22.5)	514 (19.8)
7-7.9	66 (5.7)	123 (8.6)	189 (7.3)
≥8.0	5 (0.4)	37 (2.5)	42 (1.6)
BMI ≥ 23Kg/m², n (%)	590 (50.8)	945(65.9)	1535 (59.1)
BMI ≥ 30Kg/m², n (%)	47 (4.0)	166 (11.6)	213 (8.2)

SBP- systolic blood pressure, BMI- body mass index

Over the 10-year follow-up period, 179 hard CVEs were recorded: 66 (36.9%) in females and 113 (63.1%) in males.

We tested six ML algorithms to find the best predictive CV risk prediction model using 6 variables and 75 variables separately. The Random Forest models showed the highest accuracy with AUC-ROC for both 6-variable and 75-variable ML models and were selected as the final ML-based models (Supplemental Table 1).

The 20 most important variables in terms of predictive performance in the descending order of the 75-variable model developed on the Random Forest algorithm are shown in Table 2.

Table 2 Variable ranking by their contribution to CV risk predictions

Ranking	Variable	Importance
1	Age	0.08666
2	Smoking status	0.062
3	Height	0.05601
4	Average systolic blood pressure	0.05274
5	Smoking duration	0.05246
6	Sex	0.05149
7	Sugar control for 3 months	0.03583
8	Hip circumference	0.03004

9	Average diastolic blood pressure	0.02795
10	Serum triglyceride level	0.02524
11	Number of packed smoked a day	0.02387
12	History of hypertension	0.02246
13	Baseline insulin level	0.0222
14	LDL Cholesterol	0.022
15	Fasting blood sugar	0.02166
16	Total cholesterol level	0.0191
17	Weight in 2007	0.01904
18	Alcohol used at least once a week	0.01901
19	Waist in 2007	0.01798
20	Body mass index in 2007	0.01788

The predicted number of CVEs by the newly developed ML-based models (6-variable and 75-variable) and the WHO risk charts (2019) for the next 10 years using baseline data of 2007 were compared with observed CVEs by 2017 using AUC-ROC curves and confusion matrix (Figure 2).

Figure 2

The AUC-ROC of the three models were; 75-variable model: 0.74, (CI-0.68-0.80), 6-variable model: 0.72, (CI-0.66-0.78) and WHO risk charts: 0.51, (CI-0.42-0.60). Accuracy in terms of the rate of prediction of actual CV risk of the population (predicting both true positive and true negative CVEs) was; 75-variable model: 93.1% (2417/2596), 6-variable model: 93.1% (2418/2596) and WHO risk charts: 91.8% (2382/2596) (Figure 2).

The predictive accuracy of the three models was studied. The 75-variable model predicted 125 of 179 CVEs and 2293 of 2417 non-CVE cases correctly; sensitivity -

69.8%, positive predictive value (PPV) - 50.2%, specificity - 94.8%, negative predictive value (NPV) -97.6%. The 6-variable model predicted 124 of 179 CVEs and 2293 of 2417 non-CVE cases correctly; sensitivity - 69.2%, PPV - 50.0%, specificity - 94.8%, NPV - 97.6%. The WHO risk charts predicted only 10 of 179 cases and 2372 of 2417 non-CVE cases correctly; sensitivity - 5.58%, PPV - 18.1%, specificity - 98.1%, NPV - 93.3%. The 75- and 6-variable models correctly predicted 115 and 114 more CVE cases than the 10 CVE cases predicted by the latest WHO risk charts.

The 6-variable ML based model was validated in an external cohort of 357 hospital-based patients. The external validation cohort consisted of 118 incident CVE cases and 239 non-CVE cases, 117 (32.7%) males with a mean age of 63.4 (SD: 7.2) years. Their CVE risk predictions were calculated using the 6-variable model and WHO risk charts separately. The predicted and observed number of CVEs were compared using confusion matrix (Figure 3). The predictive accuracy of the 6-variable model was 83/118 cases (sensitivity 70.3%, PPV 87.3%) and 227/239 non-CVE cases (specificity 95.0%, NPV 86.6%) while that of WHO risk charts was 28/118 cases (sensitivity 23.7%, PPV 35.8%) and 189/239 non-cases (specificity 79.0%, NPV 67.7%). The 6-variable model correctly predicted 55 more cases of CVEs than the 28 cases predicted by the currently used 2019 WHO risk charts.

Figure 3

Discussion

We developed two ML based CV-risk prediction models using longitudinal data of a Sri Lankan cohort that was prospectively followed up for 10 years. This is the first CV risk prediction model developed using individual data from Sri Lankans and the only risk prediction model specific to Sri Lankans. The newly developed 6-variable ML based model was able to predict CVE with a 70% sensitivity and 95% specificity in an external cohort. The overall predictive performances of the ML based models in Sri Lankans were better than that of the reference, WHO CV risk charts developed for the whole of South East Asia Region (2019). The newly developed ML based models appear to be more effective in the risk prediction of people at high CV risk compared to the WHO risk charts and are equally effective as the WHO score in risk predicting

people at low CV risk. Validation of the 6-variable ML based model in an external cohort of Sri Lankans re-confirmed the findings.

Improved CV risk prediction allows for the identification of an increased number of patients who could benefit from preventive treatment while avoiding unnecessary treatment of low-risk people (8). The WHO risk charts developed for the South East Asia region are good in detecting Sri Lankans at low risk of CVDs but are less sensitive in predicting patients who are at high risk of CVDs. The same was observed during the validation of the 2007 WHO/International Society of Hypertension risk charts among Sri Lankans (5). This could be explained by several reasons. The WHO risk charts were developed using available epidemiological data of the member countries to be used in predicting the CV risk of the people of the whole of South East Asia region. However, our ML based models were developed using individual patient data of a Sri Lankan cohort followed up for 10 years and therefore are more specific for Sri Lankans. Further, we developed the prediction models using machine learning of the data of a prospectively followed-up Sri Lankan cohort. ML allows the models to appreciate subtle complex interactions between variables in predicting outcomes than using conventional logistic regression making our ML based models more specific for Sri Lankans.

CV risk prediction using ML is now being used globally and reported to be better than traditional risk prediction models (8-13). Several studies from the UK have shown superiority of ML based models over traditional models in predicting CV risk. Alaa et al. showed that the ML based risk predictions improved the accuracy of CV risk prediction in 423,604 participants of UK Biobank compared to the Framingham risk score (10). Another study of 378,256 patients from UK family practices showed that a new ML model using 8 conventional variables significantly improved the accuracy of CV risk prediction (10). Another recent study using a novel prediction model comprising 10 predictors in a cohort of UK Biobank showed better performance over multiple existing clinical models (13). A study involving 143,043 Chinese patients with hypertension also showed that ML outperforms traditional logistic regression for CV risk prediction (12). Our results for the two ML models in Sri Lankans corroborate these previous findings in other populations.

The study by Alaa et al. using the UK biobank data showed that the predictive capacity of the ML model when using all available 476 variables was better than that when using only the traditional variables (10). However, we did not find a significant difference in predictive performance when using all available variables (n=75) compared to using 6 traditional variables in the ML models in our cohort. Several explanations are possible for the lack of difference between the two ML models in this cohort; e.g., the cohort sample size is too small to identify risk factors with minor contributions and the 75 variables available in this study do not contain enough to provide additional information to the 6 traditional variables.

A recent meta-analysis of ML algorithms utilised for CVD prediction has highlighted the importance of using the optimal algorithm for the datasets being used due to the heterogeneity among ML algorithms (14). A recent review on artificial intelligence (AI) and CV risk prediction has shown that AI-based predictive models may overcome some of the limitations of classic regression models, but successful application of AI requires knowledge of the potential pitfalls in AI techniques to guarantee their safe and effective use in daily clinical practice (15). We trialled six standard ML classification algorithms with different modelling approaches and our models confirmed the importance of the already known conventional CV risk factors in predisposition to CVD. This adds to the validity of our results. In a resource-limited country such as Sri Lanka, our 6-variable model would be more practical than using the 75-variable model to screen individuals at higher CV risk, as it is as predictive as the 75-variable model. The 6-variable ML model is more predictive than WHO risk charts, especially in high-risk people, who should be the main target for primary prevention of CVDs.

There are several strengths in our study. Our cohort is a community-based random sample. The study area consisted of 75,591 multi-ethnic residents in 2007. Participants were prospectively followed up for 10 years. The dropout rate was very low, and only the data of participants who completed 10-year follow-ups were used in the development of the ML models. Patients were initially recruited and followed up by medical officers using face-to-face interviews and medical records and/or death certificates, and therefore self-reporting bias was minimized. The endpoints used (hard CVE) were clear and objective.

There are some limitations to our study. For example, even though our cohort is community-based, it is from a semi-urban area and therefore may not represent the whole of Sri Lanka. According to the 2012 census, however, the overall national distribution of the population in the urban: rural sectors is 1: 4.5, which is comparable to 1: 5.4, in the Gampaha district.

In conclusion, we have shown that the new 6-variable ML model and the 75-variable model were more predictive of CV risk especially of high-risk patients than the WHO CV risk charts for South-East Asians (2019) in a Sri Lankan cohort. We plan to develop a web/mobile interphase of the new 6-variable model to increase its clinical utility.

Declarations

Funding: This study was supported by the Strengthening Research Outputs Grant of the University of Kelaniya, Sri Lanka (RC/SROG/2021/01). The funding bodies played no role in the design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Author Contribution: CM, MBS and PSH conceptualized and designed the study. AK, ASD, ARW, NK and HJdeS were involved in establishing the Ragama Health Study cohort. MBS and PSH analysed the data assisted by CM. CM, MBS and HJdeS prepared and revised the manuscript. All authors read and agreed to the final version of the manuscript.

Availability of data and materials: The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Competing interests: The authors declare that they have no competing interests and no conflicts of interest.

Acknowledgements: We thank all those who have continuously supported the Ragama Health Study, and especially the study participants for their continued cooperation.

Patient and Public Involvement statement: It was not appropriate or possible to involve patients or the public in the design, conduct, reporting, or dissemination plans of our research

References

1. Volgman AS, Palaniappan LS, Aggarwal NT, Gupta M, Khandelwal A, Krishnan AV, et al. Atherosclerotic Cardiovascular Disease in South Asians in the United States: Epidemiology, Risk Factors, and Treatments: A Scientific Statement From the American Heart Association. Circulation. 2018;138(1):e1-e34.
2. Mettananda KCD, Gunasekara N, Thampoe R, Madurangi S, Pathmeswaran A. Place of cardiovascular risk prediction models in South Asians; agreement between Framingham risk score and WHO/ISH risk charts. Int J Clin Pract. 2021;75(7):e14190.
3. Ranawaka U, Wijekoon N, Pathmeswaran P, Kasturiratne A, Gunasekara D, Chackrewarthy S, et al. Risk estimates of cardiovascular diseases in a Sri Lankan community. Ceylon Med J. 2016;61:11.
4. WHO. World Health organization/International Society of Hypertension risk prediction charts for 14 WHO epidemiological sub-regions: WHO; 2007.
5. Thulani UB, Mettananda KCD, Warnakulasuriya DTD, Peiris TSG, Kasturiratne K, Ranawaka UK, et al. Validation of the World Health Organization/ International Society of Hypertension (WHO/ISH) cardiovascular risk predictions in Sri Lankans based on findings from a prospective cohort study. PLoS One. 2021;16(6):e0252267.
6. WHO. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob Health. 2019;7(10):e1332-e45.
7. Dassanayake AS, Kasturiratne A, Rajindrajith S, Kalubowila U, Chakrawarthi S, De Silva AP, et al. Prevalence and risk factors for non-alcoholic fatty liver disease among adults in an urban Sri Lankan population. J Gastroenterol Hepatol. 2009;24(7):1284-8.
8. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE. 2017;12(4):e0174944.
9. Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022;17(1):1100-13.

10. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 2019;14(5):e0213653.

11. Dalal S, Goel P, Onyema EM, Alharbi A, Mahmoud A, Algarni MA, et al. Application of Machine Learning for Cardiovascular Disease Risk Prediction. Computational Intelligence and Neuroscience. 2023;2023:9418666.

12. Xi Y, Wang H, Sun N. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. Front Cardiovasc Med. 2022;9:1025705.

13. Jia Y, Yu G, Ju-Jiao K, Hui-Fu W, Ming Y, Jian-Feng F, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke and Vascular Neurology. 2023:svn-2023-002332.

14. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific Reports. 2020;10(1):16057.

15. Chiarito M, Luceri L, Oliva A, Stefanini GG, Condorelli G. Artificial Intelligence and Cardiovascular Risk Prediction: All That Glitters is not Gold. European Cardiology Review 2022;17:e29. 2022.

Figures

Figure 1 Machine learning model development process.

Figure 2 Comparison of the predictive performance of machine learning based models and the World Health Organization cardiovascular risk charts (South East Asia Region -2019) in a Sri Lankan cohort

ML – machine learning, WHO – World Health Organization, CV – cardiovascular

Figure 3 External validation of the 6-variable machine learning model in cardiovascular risk predicting

Supplemental Table 1 Comparison of predictive performances of 6-variable and 75-variable ML-models

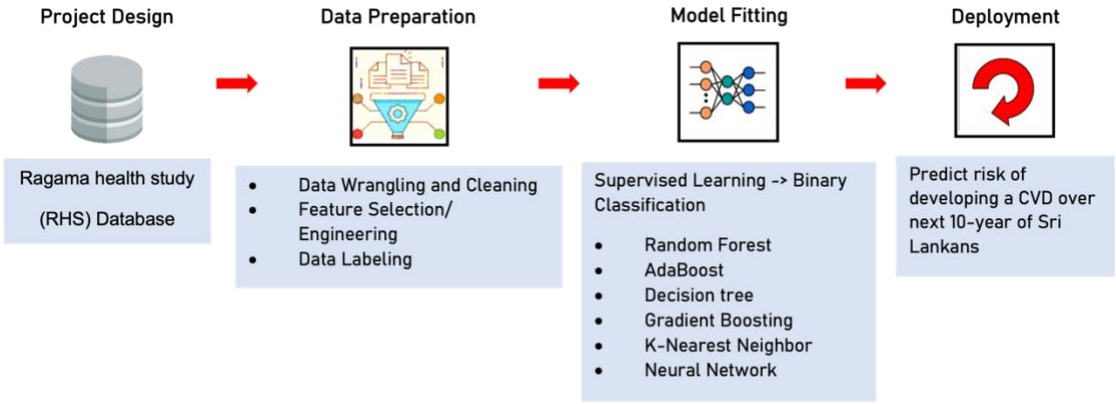
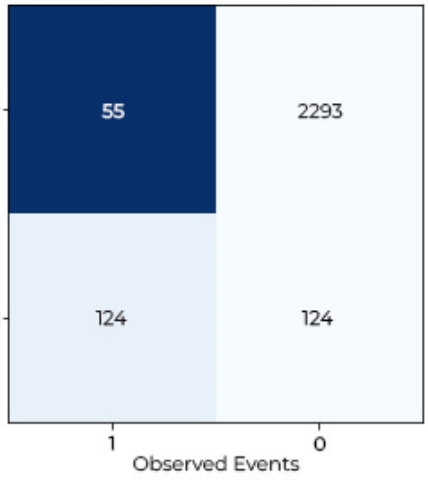
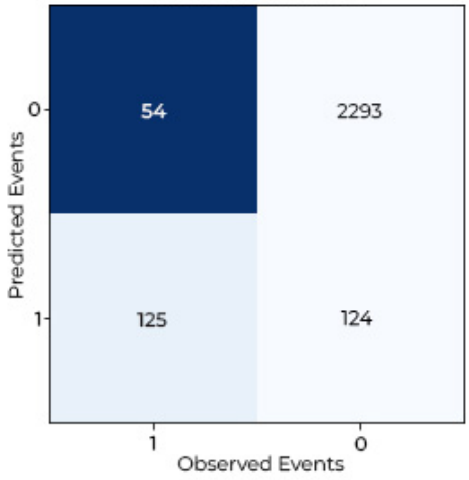
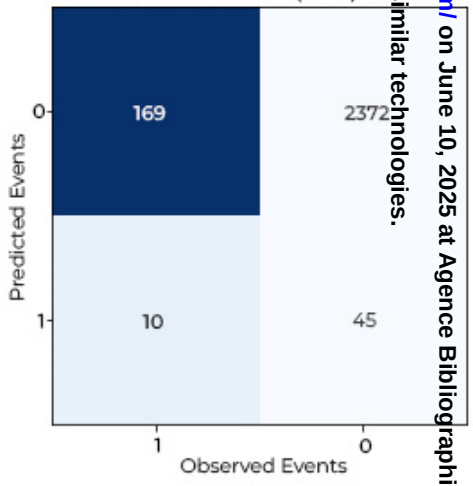
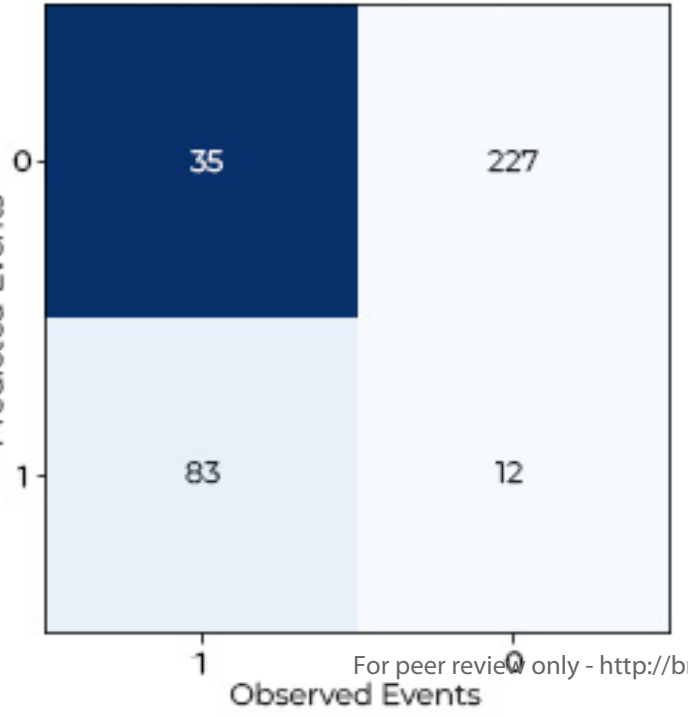


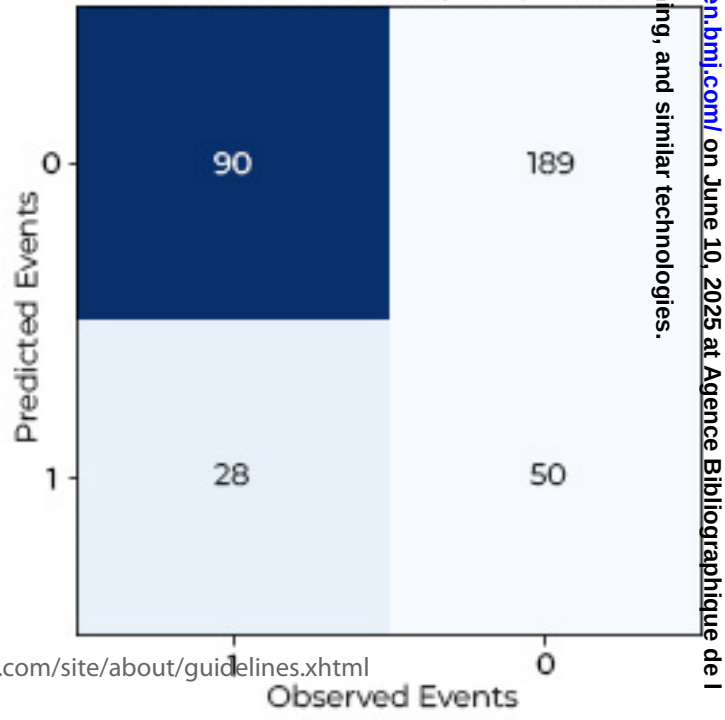
Figure 1











Supplemental Table 1 Comparison of predictive performances of 6-variable and 75-variable ML-models

Algorithm	6-variable ML models		75-variable ML models	
	Accuracy	ROC-AUC	Accuracy	ROC-AUC
Random Forest	0.9314	0.72 ± 0.07	0.9311	0.74 ± 0.06
AdaBoost	0.9291	0.68 ± 0.07	0.9199	0.64 ± 0.08
Decision tree	0.8733	0.55 ± 0.05	0.8663	0.51 ± 0.03
Gradient Boosting	0.9272	0.72 ± 0.06	0.9245	0.72 ± 0.06
k-Nearest Neighbour	0.9310	0.62 ± 0.04	0.9311	0.58 ± 0.06
2D Neural Network	0.8829	0.55 ± 0.02	0.9145	0.60 ± 0.02

ML - machine learning, AUC-ROC - area under the receiver operating characteristic curve

BMJ Open

Efficacy of 10-year cardiovascular risk prediction using machine learning compared to the World Health Organization risk charts; a cohort study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-081434.R1
Article Type:	Original research
Date Submitted by the Author:	30-Nov-2024
Complete List of Authors:	Mettananda, Chamila; University of Kelaniya, Pharmacology Solangaarachchige, Maheeka; University of Kelaniya, Exam Unit; Sri Lanka Institute of Information Technology, Haddela, Prasanna; Sri Lanka Institute of Information Technology, Department of IT Dassanayake, Anuradha; University of Kelaniya Faculty of Medicine, Pharmacology Kasturiratne, Anuradhani; University of Kelaniya Faculty of Medicine, Public Health Wickremasinghe, Rajitha; University of Kelaniya Faculty of Medicine, Public Health Kato, Norihiro; National Center for Global Health and Medicine Research Institute, Gene Diagnostics and Therapeutics de Silva, Hithanadura; University of Kelaniya Faculty of Medicine, Medicine
Primary Subject Heading:	Cardiovascular medicine
Secondary Subject Heading:	Public health
Keywords:	Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PREVENTIVE MEDICINE, Primary Prevention, Cardiac Epidemiology < CARDIOLOGY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Efficacy of 10-year cardiovascular risk prediction using machine learning compared to the World Health Organization risk charts; a cohort study

Mettananda C[1†], Solangaarachchige MB[1,2†], Haddela PS[2], Dassanayake AS[1], Kasturiratne A[1], Wickramasinghe AR[1], Kato N[3], de Silva HJ[1]

[1] Faculty of Medicine, University of Kelaniya, Sri Lanka
[2] Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
[3] National Centre for Global Health and Medicine, Toyama, Shinjuku-ku, Tokyo, Japan

† - Contributed equally to this work

Correspondence to
Chamila Mettananda
Chamila@kln.ac.lk, chamilametta@hotmail.com

Main text – 2900 words. Abstract – 284words

Abbreviated title - efficacy of cardiovascular disease risk prediction using machine learning

Abstract

Introduction

Models derived from non-Sri Lankan cohorts are currently being used for cardiovascular (CV) risk stratification of Sri Lankans. We aimed to develop a CV risk prediction model using machine learning (ML) based on data from a Sri Lankan cohort followed up for 10 years, and to compare the predictions with World Health Organization (WHO) risk charts.

Design: Cohort study

Setting: Ragama health study(RHS), which is a prospective, ongoing, population-based cohort study of patients, randomly selected from the Ragama Medical office of Heath area, Sri Lanka, focusing on the epidemiology of non-communicable diseases was used to develop the model. The external validation cohort included patients admitted to Colombo North Teaching Hospital(CNTH), a tertiary care hospital in Sri Lanka from January 2019 through August 2020.

Participants: All RHS participants, 40-64 years in 2007, without cardiovascular disease (CVD) at baseline, who had complete data for 10-year outcome by 2017 were used for model development. Patients aged 40–74 years admitted to CNTH during the study period with incident CV events or a disease other than an acute CVE who had complete data for CVD risk calculation were used for external validation of the model.

Interventions: No intervention

Using the follow-up data of the cohort, we developed two ML models for predicting 10-year CV risk using 6 conventional CV risk variables(age, gender, smoking status, systolic blood pressure, history of diabetes, and total cholesterol level) and all available variables(n=75). The ML models were derived using classification algorithms of the supervised learning technique. We compared the predictive performance of our ML models with WHO risk charts (2019, Southeast Asia) using area under the receiver operating characteristic curves(AUC-ROC) and calibration plots. We validated the 6-variable model in an external hospital-based cohort.

Results

Of the 2596 participants in the baseline cohort, 179 incident CV events(CVEs) were observed over 10 years. WHO risk charts predicted only 10 CVEs(AUC-ROC: 0.51, CI-0.42-0.60) while the new 6-variable ML-model predicted 125 CVEs(AUC-ROC: 0.72, CI-0.66-0.78) and 75-variable ML-model predicted 124 CVEs(AUC-ROC: 0.74, CI-0.68-0.80). Calibration for the 6-variable ML-model and the WHO risk charts were; the Hosmer-Lemeshow test; $\chi2=12.85$, $p= 0.12$ and $\chi2=15.58$, $p= 0.05$ respectively. In the external validation cohort, the sensitivity, specificity, positive-predictive-value, negative-predictive-value and calibration of the 6-variable ML-model and the WHO risk charts were, 70.3%, 94.9%, 87.3%, 86.6%, $\chi2=8.22$, $p= 0.41$ and 23.7%, 79.0%, 35.8%, 67.7%, $\chi2=81.94$, $p<0.0001$ respectively.

Conclusions

ML-based models derived from a cohort of Sri Lankans improved the overall accuracy of CV-risk prediction compared to the WHO risk charts for this cohort of Southeast Asians.

Keywords – Cardiovascular risk, prediction, World Health Organization risk charts, Machine learning, validation, Sri Lanka

Strengths and limitations of this study

- This is the first CV risk prediction model specific to Sri Lankans.
- We developed the risk prediction models using machine learning of 10-year follow-up data of individual patients.
- 10-year follow-up data of a large, population-based, randomly selected sample was used to develop the model
- Even though the cohort we used to train the ML model was a community-based, multi-ethnic random cohort, representation of the estate sector was less in our cohort compared to the national distribution.
- Imputation of missing data and imbalance of data due to having very few female smokers might have some influence on the model's performance but this was minimized with stratified 10-fold cross-validation

Introduction

There are no cardiovascular (CV) risk prediction models specific to or derived from Sri Lankans. Therefore, different risk prediction models derived from white Caucasians, or models developed for the South-East Asia region (SEAR) are being used for CV risk stratification of Sri Lankans.

Asians behave differently from white Caucasians in terms of CV risk. Asians have a distinct genetic make-up, and a different CV risk factor profile with a higher prevalence of hypertension, diabetes mellitus, central obesity, insulin resistance, and metabolic syndrome than white Caucasians [1]. They are also at increased risk of developing CV diseases (CVDs) compared to white Caucasians at a given risk factor level [1]. In Sri Lankans, there is low agreement between the CV risk predictions based on the World Health Organization / International Society of Hypertension (WHO/ISH) risk charts and the Framingham General CV risk charts [2]. Moreover, the CV risk predictions in a Sri Lankan cohort using three different risk models, the National Cholesterol Education Program - Adult Treatment Panel III (NCEP-ATP III), WHO/ISH charts and Systematic Coronary Risk Evaluation (SCORE) charts, were found discordant [3].

The WHO/ISH CV risk charts for the South-East Asia region-B (SEAR-B) were developed in 2007 together with for another 14 epidemiological sub-regions to risk predict people of those regions that did not have specific risk prediction models derived from their own cohorts [4]. These 2007 WHO/ISH risk charts have been validated in Sri Lankans (4) and they showed 81% agreement between predictions and observed events but were less predictive in females and those at high CV risk [5]. Later on, the WHO risk charts were revised and re-calibrated in 2019 to improve predictive capacity as well as to include 21 epidemiological sub-regions that did not have specific risk prediction models. These 2019 WHO risk charts are currently the best available for Sri Lankans [6]. However, in this also, Sri Lanka is grouped under the Southeast Asia epidemiological sub-region together with Indonesia, Cambodia, Laos, Sri Lanka, Maldives, Myanmar, Malaysia, Philippines, Thailand, Timor-Leste, Viet Nam, Mauritius, and Seychelles. This is a heterogeneous population, with different socio-economic and cultural backgrounds and therefore, the risk predictions may not accurately represent the CV risk of Sri Lankans.

Therefore, we aimed to develop a CV risk prediction model using machine learning (ML) based on data from a Sri Lankan cohort followed up for 10 years, and to compare the predictions with 2019 WHO (South-East Asia) risk charts. Moreover, we aimed to validate the new model in an external cohort of Sri Lankans.

Methods

Machine Learning based model development

We developed two CV risk prediction models using ML, based on data from a large community-based study on non-communicable diseases, the "Ragama Health Study (RHS)" [3 7], where individuals have been followed up from 2007 to date.

The baseline study population (n=2923) in the RHS comprised 35–64 years old, adult residents in the "Ragama Medical Officer of Health (MOH) area" in 2007. Participants were selected by stratified random sampling in the Ragama MOH area, which is a semi-urban health administrative area among 25 districts in Sri Lanka. Participants were followed up for 10 years from 2007 to 2017; during which all CV deaths, non-fatal strokes, and non-fatal myocardial infarctions (including those undergoing percutaneous coronary interventions and coronary artery bypass grafts) were recorded as hard CV events (CVE) by either interviewing patients and their families or perusing clinical notes/death certificates.

Data for participants above 40 years of age, who had no history of CVDs at enrolment in 2007 and completed 10-year follow-up (n=2596), were extracted to develop ML-based risk prediction models, as usually risk predictions are calculated in people over the age of 40 years.

Using the 10-year prospective follow-up data for the cohort, using baseline data of those who developed CVEs and those who did not, we developed two ML-based models to predict the 10-year risk of developing a hard CVE using different risk factor combinations. Individuals who could not be traced in 2017 or those whose cause of death could not be verified were excluded. The ML-based models were developed using classification algorithms of the supervised learning technique. The models were

developed in a recursive process (8) in four steps: project design, data preparation, model fitting and inference & deployment (Figure 1). Using the database, models were built with the publicly available Google Colab ML platform and Scikit-learn library in Python programming language (9) and Train-Test Split method (10). Participant data were split into two groups; the training sample and the testing sample. The training sample was used to build the ML-based models and the testing sample was used to assess the efficacy of the algorithms built using the training sample. Since the ratio of CVE to non-CVE was highly skewed at 7:93, we performed stratified 10-fold cross-validation, using 2336 individuals for the training sample and the remaining 260 for the test sample to prevent over-fitting.

The predictive performances of the models were compared. We determined the discriminative power using the area under the receiver operating characteristic curve (AUC-ROC, c-index) and the mean F1 scores. The mean of AUC-ROCs for the 10 cross-validation samples was taken as the AUC-ROC of the ML-based model in question. The AUC-ROC and mean F1-score were used to select the best model. A model with a mean F1-score above 0.8, accuracy above 0.85 and AUC c-index closer to 1 was considered good for risk prediction [8 9]. We calibrated the models using calibration plots. A model with a Hosmer-Lemeshow test $\chi 2$ value of greater than 20 or a *p-value* of less than 0.05 was considered to have poor calibration[10].

Figure 1

We trialled six standard ML classification algorithms with different modelling approaches, namely, Decision tree, Random forest, k-nearest neighbour, 2D neural networks, AdaBoost and gradient boosting. The best-fitting model in terms of mean F1-score and AUC-ROC was selected to develop the final model. Grid search was used to optimize the hyper-parameters of the models (11). Data imputation for all models was done using the statistical imputation of missing values using Python.

We developed two risk prediction models; one using the 6 conventional CV risk variables that are used in the WHO CV risk charts (age, gender, smoking status, systolic blood pressure, history of diabetes, and total cholesterol level) and the other

using 75 variables. The total database consisted of 770 variables, including data on demographics, medical history, family history, social history, physical examination, laboratory investigations and non-laboratory investigations like ECG and an ultrasound scan of the abdomen. Following data wrangling and cleaning, we chose 75 (out of 770) variables following the literature review and using domain knowledge for the ML model development. We excluded variables with missing values ≥ 50%. The ML models predicted individuals likely and unlikely to develop a CVE within the next 10 years by machine learning the database.

Internal validation of the machine learning model

We calculated the predicted CVEs over 10 years by 2017, using baseline data (2007 data) and the two ML models separately. Additionally, we calculated the same using the latest 2019 WHO CV risk charts. We compared the predictions of the 6-variable and 75-variable ML models and the WHO model against the observed events using AUC-ROC and mean F1-score.

External Validation of the 6-variable machine learning-based model

We externally validated the 6-variable ML model in a separate hospital-based database of 357 consecutive patients, 40–74 years of age admitted to Colombo North Teaching Hospital (a tertiary care hospital in Sri Lanka) from 1st of January 2019 to 1st of August 2020 who did not have a history of CVEs and presented with an acute incident CVE (acute myocardial infarction or acute stroke) or a disease other than an acute CVE who had complete data for CVD risk calculation. Their predicted risks of developing a CVE were calculated using the most recent pre-morbid risk factor data available up to one year before developing the incident CVE or the admission to the ward in non-CVE cases. We compared the predictions of the 6-variable model with that of the 2019 WHO risk chart using confusion matrix and calibration plots.

Ethical Clearance

This work was approved by the Ethics Review Committee of the Faculty of Medicine, University of Kelaniya, Sri Lanka (original RHS cohort - P38/09/2006, external validation cohort- P61/09/2020) and written informed consent was obtained from all the participants.

Patient and public Involvement statement

It was not appropriate or possible to involve patients or the public in the design or reporting plans of our research but was involved in the conduct and dissemination of the study. All patients are routinely followed up in a non-communicable disease clinic at the Faculty of Medicine, in collaboration with North Colombo Teaching Hospital (NCTH) Ragama, Sri Lanka as a service component since 2007 to date. Information about their risk factors was available to participants and when necessary they were referred for specialist care at the NCTH. The results of the study will be disseminated to study participants, other patients and the public following the publication of the study.

Results

A total of 2596 participants followed up for 10 years were eligible for the study with a mean age of 53.5 (SD: 6.9) years and 1162 (44.8%) males. The baseline characteristics of the study cohort are shown in Table 1.

Table 1 Baseline characteristics of the cohort

	Male n = 1162	Female n = 1434	Total n = 2596
Ethnicity n (%)			
Sinhalese	1118 (96.2)	1375 (95.9)	2493 (96.0)
Tamil	15 (1.3)	27 (1.9)	42 (1.6)
Muslim	2 (0.2)	2 (0.1)	4 (0.2)
Burgher	15 (1.3)	19 (1.3)	34 (1.3)
Other	12 (1.0)	11 (0.8)	23 (0.9)

Age groups (years), n (%)			
40-49.9	360 (30.9)	456 (31.8)	816 (31.4)
50-59.9	526 (45.3)	669(46.7)	1195 (46.0)
≥ 60.0	276 (23.8)	309(21.5)	585 (22.6)
Smoking, n (%)	416 (35.8)	0 (0.0)	416 (16.0)
Diabetes, n (%)	165 (14.2)	249 (17.4)	414 (15.9)
Hyperlipidaemia, n (%)	98 (8.4)	209 (14.6)	307 (11.8)
SBP (mmHg), n (%)			
<139.9	766 (65.9)	869 (60.6)	1635 (62.9)
140-159.9	260 (22.4)	365 (25.5)	625 (24.1)
160-179.9	88 (7.6)	132 (9.2)	220 (8.5)
≥180.0	48 (4.1)	68 (4.7)	116 (4.5)
Total cholesterol (mmol/L), n (%)			
<4.0	211 (18.2)	207 (14.4)	418 (16.1)
4-4.9	297 (25.6)	269 (18.8)	566 (21.8)
5-5.9	391 (33.6)	476 (33.2)	867 (33.4)
6-6.9	192 (16.5)	322 (22.5)	514 (19.8)
7-7.9	66 (5.7)	123 (8.6)	189 (7.3)
≥8.0	5 (0.4)	37 (2.5)	42 (1.6)
BMI ≥ 23Kg/m^2, n (%)	590 (50.8)	945(65.9)	1535 (59.1)
BMI ≥ 30Kg/m^2, n (%)	47 (4.0)	166 (11.6)	213 (8.2)

SBP- systolic blood pressure, BMI- body mass index

Over the 10-year follow-up period, 179 hard CVEs were recorded: 66 (36.9%) in females and 113 (63.1%) in males.

We tested six ML algorithms to find the best predictive CV risk prediction model using 6 variables and 75 variables separately. A comparison of model performances using different ML algorithms is shown in Supp. Table 1. Random Forest models

showed the highest accuracy, mean F1-score and AUC-ROC for both 6-variable and 75-variable ML models and were selected as the final ML-based models.

The 20 most important variables in terms of predictive performance in the descending order of the 75-variable model developed on the Random Forest algorithm are shown in Table 2.

Table 2 Variable ranking by their contribution to CV risk predictions

Ranking	Variable	Importance
1	Age	0.08666
2	Smoking status	0.062
3	Height	0.05601
4	Average systolic blood pressure	0.05274
5	Smoking duration	0.05246
6	Sex	0.05149
7	Sugar control for 3 months	0.03583
8	Hip circumference	0.03004
9	Average diastolic blood pressure	0.02795
10	Serum triglyceride level	0.02524
11	Number of packed smoked a day	0.02387
12	History of hypertension	0.02246
13	Baseline insulin level	0.0222
14	LDL Cholesterol	0.022
15	Fasting blood sugar	0.02166
16	Total cholesterol level	0.0191
17	Weight in 2007	0.01904
18	Alcohol used at least once a week	0.01901
19	Waist in 2007	0.01798
20	Body mass index in 2007	0.01788

The predicted CVEs by the newly developed ML-based models (6-variable and 75-variable) and the WHO risk charts (2019) for the next 10 years using baseline data of 2007 were compared with the observed CVEs by 2017 using AUC-ROC curves and confusion matrices (Figure 2).

Figure 2

Discrimination of the three models using AUC-ROC and c-indexes were; 75-variable model: 0.74, (CI-0.68-0.80), 6-variable model: 0.72, (CI-0.66-0.78) and WHO risk charts: 0.51, (CI-0.42-0.60). Accuracy in terms of the rate of prediction of actual CV risk of the population (predicting both true positive and true negative CVEs) was; 75-variable model: 93.1% (2417/2596), 6-variable model: 93.1% (2418/2596) and WHO risk charts: 91.8% (2382/2596) (Figure 2).

The predictive accuracies of the three models were studied using confusion matrices (Figure 2). The 75-variable model predicted 124 of 179 CVEs and 2293 of 2417 non-CVE cases correctly; sensitivity - 69.2%, positive predictive value (PPV) - 50.0%, specificity - 94.8%, negative predictive value (NPV) -97.6%. The 6-variable model predicted 125 of 179 CVEs and 2293 of 2417 non-CVE cases correctly; sensitivity - 69.8%, PPV - 50.2%, specificity - 94.8%, NPV - 97.6%. The WHO risk charts predicted only 10 of 179 cases but 2372 of 2417 non-CVE cases correctly; sensitivity - 5.58%, PPV - 18.1%, specificity - 98.1%, NPV - 93.3%. The 75- and 6-variable models correctly predicted 114 and 115 more CVEs than the 10 CVEs predicted by the latest WHO risk charts.

The calibration for the 6-variable ML model was good as the Hosmer-Lemeshow test result was $\chi 2=12.85$, $p= 0.12$. the Hosmer-Lemeshow test result for the WHO risk charts was $\chi 2=15.58$, $p= 0.05$. (Supp. Table 2 and Figure 3)

Figure 3

The 6-variable ML-based model was validated in an external cohort of 357 hospital-based patients. The external validation cohort consisted of 118 incident CVE cases and 239 non-CVE cases, 117 (32.7%) males with a mean age of 63.4 (SD: 7.2) years.

Their CVE risk predictions were calculated using the 6-variable model and WHO risk charts separately. The predicted and observed number of CVEs were compared using confusion matrices (Figure 4). The predictive accuracy of the 6-variable model was 83/118 cases (sensitivity 70.3%, PPV 87.3%) and 227/239 non-CVE cases (specificity 95.0%, NPV 86.6%) while that of WHO risk charts was 28/118 cases (sensitivity 23.7%, PPV 35.8%) and 189/239 non-cases (specificity 79.0%, NPV 67.7%). The 6-variable model correctly predicted 55 more cases of CVEs than the 28 cases predicted by the currently used 2019 WHO risk charts. Calibration for the 6-variable ML model in the external validation model was also good with the Hosmer-Lemeshow test result of $\chi2=8.22$, $p= 0.41$ while that of WHO risk charts was $\chi2=81.94$, $p<0.0001$. (Supp Figure 1)

Figure 4

Discussion

We developed two ML-based CV-risk prediction models using longitudinal data of a Sri Lankan cohort that was prospectively followed up for 10 years. This is the first CV risk prediction model developed using individual data from Sri Lankans and the only risk prediction model specific to Sri Lankans. The newly developed 6-variable ML-based model was able to predict CVE with a 70% sensitivity and 95% specificity in an external cohort. The overall predictive performances of the ML-based models in Sri Lankans were better than that of the reference, WHO CV risk charts developed for the whole of South East Asia Region (2019). The newly developed ML-based models appear to be more effective in the risk prediction of people at high CV risk compared to the WHO risk charts and are equally effective as the WHO score in risk predicting people at low CV risk. Validation of the 6-variable ML-based model in an external cohort of Sri Lankans re-confirmed the findings.

Improved CV risk prediction allows for the identification of an increased number of patients who could benefit from preventive treatment while avoiding unnecessary treatment of low-risk people [11]. The WHO risk charts developed for the Southeast Asia region are good in detecting Sri Lankans at low risk of CVDs but are less sensitive in predicting patients who are at high risk of CVDs. The same was observed during the

validation of the 2007 WHO/International Society of Hypertension risk charts among Sri Lankans [5]. This could be explained by several reasons. The WHO risk charts were developed using available epidemiological data of the member countries to be used in predicting the CV risk of the people of the whole of South East Asia region. However, our ML-based models were developed using individual patient data of a Sri Lankan cohort followed up for 10 years and therefore are more specific for Sri Lankans. Further, we developed the prediction models using machine learning of the data of a prospectively followed-up Sri Lankan cohort. ML allows the models to appreciate subtle complex interactions between variables in predicting outcomes rather than using conventional logistic regression making our ML-based models more specific for Sri Lankans.

CV risk prediction using ML is now being used globally and reported to be better than traditional risk prediction models [11-16]. Several studies from the UK have shown the superiority of ML-based models over traditional models in predicting CV risk. Alaa et al. showed that the ML-based risk predictions improved the accuracy of CV risk prediction in 423,604 participants of the UK Biobank compared to the Framingham risk score [13]. Another study of 378,256 patients from UK family practices showed that a new ML model using 8 conventional variables significantly improved the accuracy of CV risk prediction (10). Another recent study using a novel prediction model comprising 10 predictors in a cohort of UK Biobank showed better performance over multiple existing clinical models [16]. A study involving 143,043 Chinese patients with hypertension also showed that ML outperforms traditional logistic regression for CV risk prediction [15]. Our results for the two ML models in Sri Lankans corroborate these previous findings in other populations.

The study by Alaa et al. using the UK biobank data showed that the predictive capacity of the ML model when using all available 476 variables was better than that when using only the traditional variables [13]. However, we did not find a significant difference in predictive performance when using all available variables (n=75) compared to using 6 traditional variables in the ML models in our cohort. Several explanations are possible for the lack of difference between the two ML models in this cohort; e.g., the cohort sample size is too small to identify risk factors with minor contributions and the

75 variables available in this study do not contain enough to provide additional information to the 6 traditional variables.

A recent meta-analysis of ML algorithms utilised for CVD prediction has highlighted the importance of using the optimal algorithm for the datasets being used due to the heterogeneity among ML algorithms [17]. A recent review on artificial intelligence (AI) and CV risk prediction has shown that AI-based predictive models may overcome some of the limitations of classic regression models, but successful application of AI requires knowledge of the potential pitfalls in AI techniques to guarantee their safe and effective use in daily clinical practice [18]. We trialled six standard ML classification algorithms with different modelling approaches and our models confirmed the importance of the already known conventional CV risk factors in predisposition to CVD. This adds to the validity of our results. In a resource-limited country such as Sri Lanka, our 6-variable model would be more practical than using the 75-variable model to screen individuals at higher CV risk, as it is as predictive as the 75-variable model. The 6-variable ML model is more predictive than WHO risk charts, especially in high-risk people, who should be the main target for primary prevention of CVDs.

There are several strengths in our study. Our cohort is a community-based random sample. The study area consisted of 75,591 multi-ethnic residents in 2007. Participants were prospectively followed up for 10 years. The dropout rate was very low, and only the data of participants who completed 10-year follow-ups were used in the development of the ML models. Patients were initially recruited and followed up by medical officers using face-to-face interviews and medical records and/or death certificates, and therefore self-reporting bias was minimized. Individual patient data was used to develop the model. The endpoints used (hard CVE) were clear and objective.

There are some limitations to our study. For example, even though our cohort is community-based, it is from a semi-urban area and therefore may not represent the whole of Sri Lanka. According to the 2012 census, however, the overall national distribution of the population in the urban: rural sectors is 1: 4.5, which is comparable to 1: 5.4, in the Gampaha district. Imputation of missing data and imbalance of data

due to having very few female smokers might have some influence on the model's performance but this was minimized with stratified 10-fold cross-validation.

In conclusion, we have shown that the new models developed by machine learning individual participant follow-up data of a Sri Lankan cohort were more predictive of CV risk, especially of high-risk Sri Lankans than the WHO CV risk charts meant for South-East Asia region (2019). We plan to improve predictions of the model by using data from a larger sample and to develop a web/mobile interphase of the new 6-variable model to increase its clinical utility.

Declarations

Funding: This study was supported by the Strengthening Research Outputs Grant of the University of Kelaniya, Sri Lanka (RC/SROG/2021/01). The funding bodies played no role in the design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Author Contribution: CM, MBS and PSH conceptualized and designed the study. AK, ASD, ARW, NK and HJdeS were involved in establishing the Ragama Health Study cohort. MBS and PSH analysed the data assisted by CM. CM, MBS and HJdeS prepared and revised the manuscript. All authors read and agreed to the final version of the manuscript. CM acted as guarantor

Availability of data and materials: The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Competing interests: The authors declare that they have no competing interests and no conflicts of interest.

Acknowledgements: We thank all those who have continuously supported the Ragama Health Study, and especially the study participants for their continued cooperation.

Patient and Public Involvement statement: It was not appropriate or possible to involve patients or the public in the design, conduct, reporting, or dissemination plans of our research

References

1. Volgman AS, Palaniappan LS, Aggarwal NT, et al. Atherosclerotic Cardiovascular Disease in South Asians in the United States: Epidemiology, Risk Factors, and Treatments: A Scientific Statement From the American Heart Association. Circulation 2018;138(1):e1-e34. doi: doi:10.1161/CIR.0000000000000580
2. Mettananda KCD, Gunasekara N, Thampoe R, et al. Place of cardiovascular risk prediction models in South Asians; agreement between Framingham risk score and WHO/ISH risk charts. Int J Clin Pract 2021;75(7):e14190. doi: 10.1111/ijcp.14190 [published Online First: 2021/03/30]
3. Ranawaka U, Wijekoon N, Pathmeswaran P, et al. Risk estimates of cardiovascular diseases in a Sri Lankan community. Ceylon Med J 2016;61:11. doi: 10.4038/cmj.v61i1.8253
4. WHO. World Health organization/International Society of Hypertension risk prediction charts for 14 WHO epidemiological sub-regions: WHO, 2007:40.
5. Thulani UB, Mettananda KCD, Warnakulasuriya DTD, et al. Validation of the World Health Organization/ International Society of Hypertension (WHO/ISH) cardiovascular risk predictions in Sri Lankans based on findings from a prospective cohort study. PLoS One 2021;16(6):e0252267. doi: 10.1371/journal.pone.0252267 [published Online First: 2021/06/08]
6. WHO. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob Health 2019;7(10):e1332-e45. doi: 10.1016/s2214-109x(19)30318-3 [published Online First: 2019/09/07]
7. Dassanayake AS, Kasturiratne A, Rajindrajith S, et al. Prevalence and risk factors for non-alcoholic fatty liver disease among adults in an urban Sri Lankan population. J Gastroenterol Hepatol 2009;24(7):1284-8. doi: 10.1111/j.1440-1746.2009.05831.x [published Online First: 2009/05/30]
8. Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol 2002;20(2):96-107. doi: 10.1053/suro.2002.32521
9. Cichosz P. Assessing the quality of classification models: Performance measures and evaluation procedures. Open Engineering 2011;1(2):132-58. doi: doi:10.2478/s13531-011-0022-9

10. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024;384:e074820. doi: 10.1136/bmj-2023-074820
11. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE 2017;12(4):e0174944. doi: 10.1371/journal.pone.0174944
12. Pal M, Parija S, Panda G, et al. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars) 2022;17(1):1100-13. doi: 10.1515/med-2022-0508 [published Online First: 20220617]
13. Alaa AM, Bolton T, Di Angelantonio E, et al. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One 2019;14(5):e0213653. doi: 10.1371/journal.pone.0213653 [published Online First: 20190515]
14. Dalal S, Goel P, Onyema EM, et al. Application of Machine Learning for Cardiovascular Disease Risk Prediction. Computational Intelligence and Neuroscience 2023;2023:9418666. doi: 10.1155/2023/9418666
15. Xi Y, Wang H, Sun N. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. Front Cardiovasc Med 2022;9:1025705. doi: 10.3389/fcvm.2022.1025705 [published Online First: 20221114]
16. Jia Y, Yu G, Ju-Jiao K, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke and Vascular Neurology 2023:svn-2023-002332. doi: 10.1136/svn-2023-002332
17. Krittanawong C, Virk HUH, Bangalore S, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific Reports 2020;10(1):16057. doi: 10.1038/s41598-020-72685-1
18. Chiarito M, Luceri L, Oliva A, et al. Artificial Intelligence and Cardiovascular Risk Prediction: All That Glitters is not Gold. European Cardiology Review 2022;17:e29 2022 doi: 10.15420/ecr.2022.11

Figures

Figure 1 Machine learning model development process.

Figure 2 Comparison of the predictive performance of machine learning-based models and the World Health Organization cardiovascular risk charts (South East Asia Region -2019) in a Sri Lankan cohort

ML – machine learning, WHO – World Health Organization, CV – cardiovascular

Figure 3 External validation of the 6-variable machine learning model in cardiovascular risk predicting

Figure 4 Calibration for 6-variable machine learning model and World Health Organization risk charts in the original cohort

Supplementary Table 1 Comparison of predictive performances of 6-variable and 75-variable machine learning models

Supplementary Table 2 Comparison of predictions in 2007 of 6-variable machine learning model and the World Health Organization risk charts and observed events in 2017

Supplementary Figure 1 Calibration for 6-variable machine learning model and World Health Organization risk charts in the external validation cohort

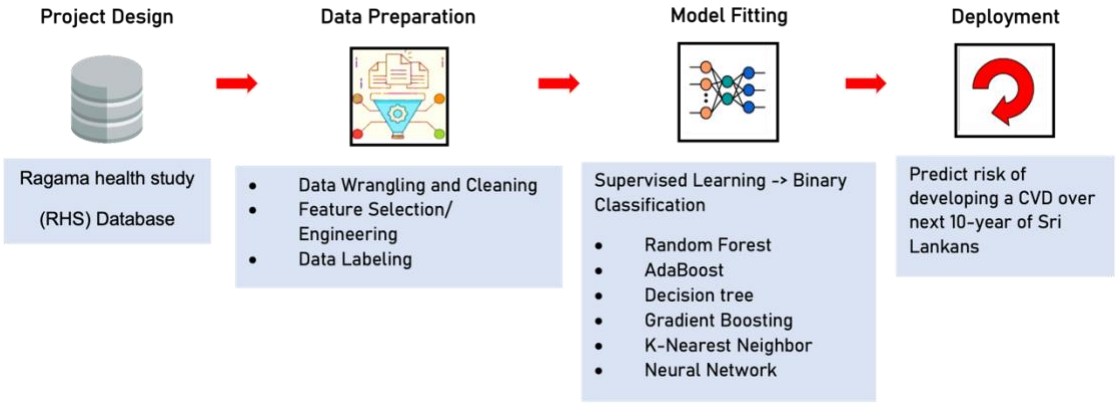
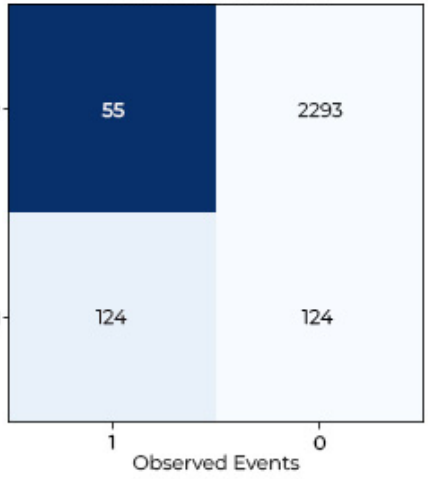
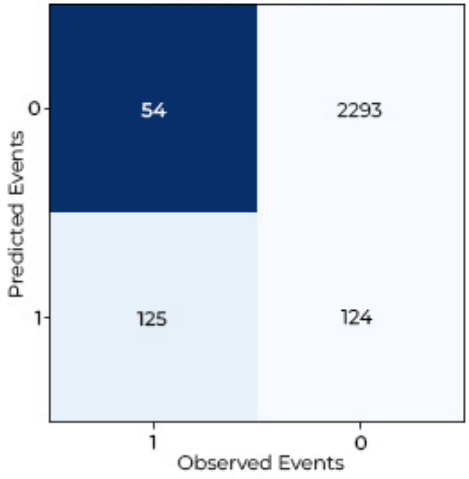
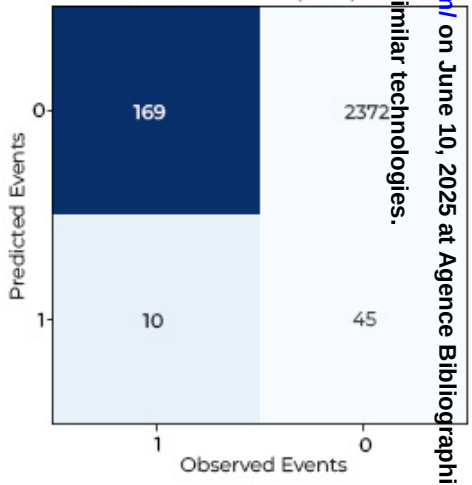
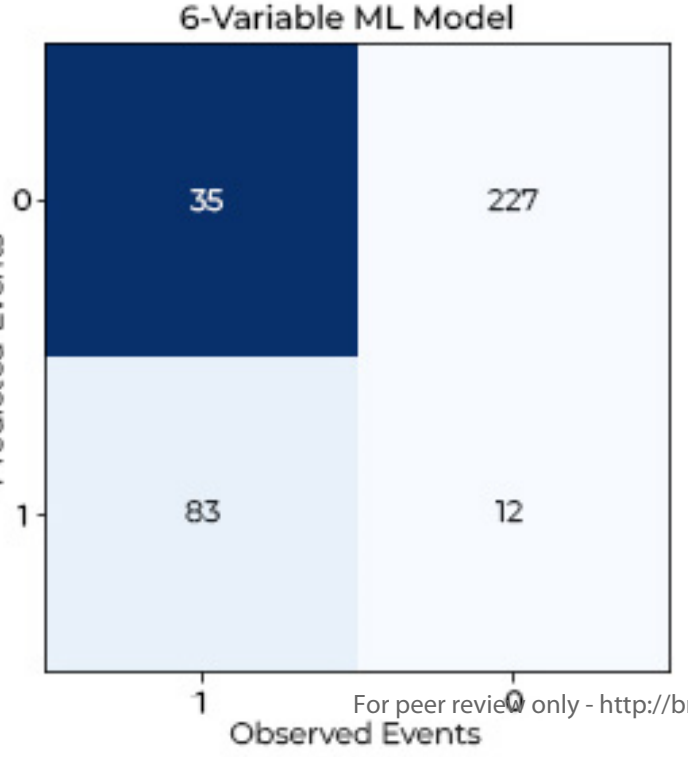


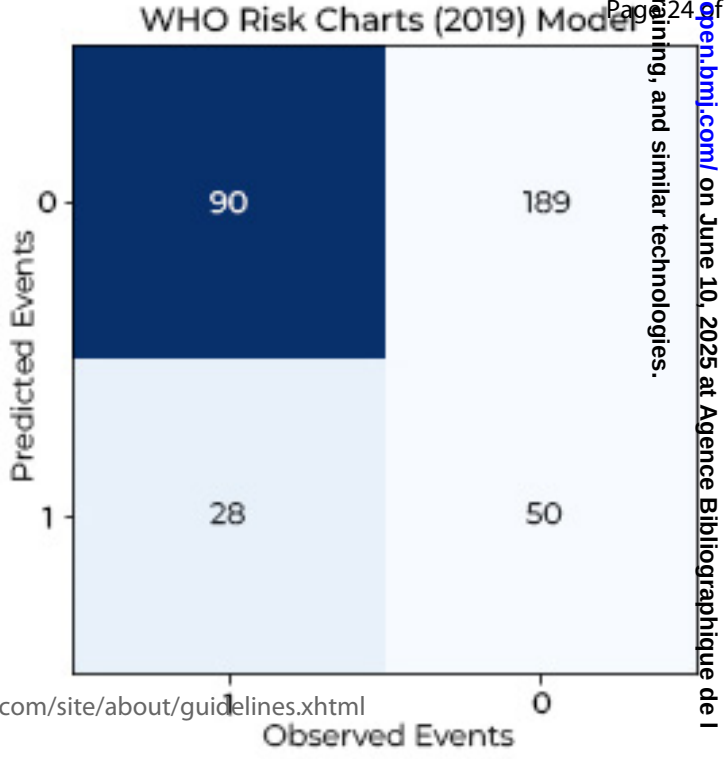
Figure 1











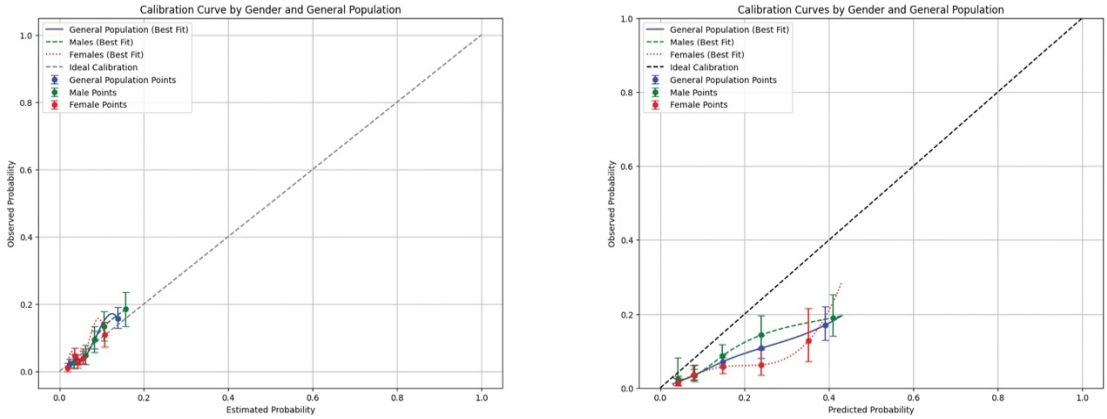
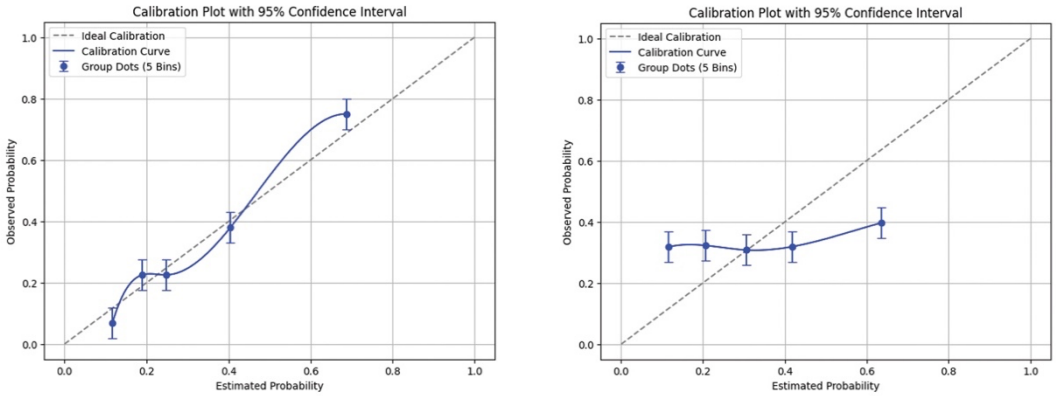


Figure 4 *Calibration for 6-variable machine learning model and World Health Organization risk charts in the original cohort*

Supplementary Table 1 Comparison of predictive performances of 6-variable and 75-

Algorithm	6-variable ML models			75-variable ML models		
	Accuracy	F1-Score	ROC-AUC	Accuracy	F1-Score	ROC-AUC
Random Forest	0.9314	0.8123	0.72 ± 0.07	0.9311	0.8102	0.74 ± 0.06
AdaBoost	0.9291	0.7632	0.68 ± 0.07	0.9199	0.7601	0.64 ± 0.08
Decision tree	0.8733	0.5812	0.55 ± 0.05	0.8663	0.5808	0.51 ± 0.03
Gradient Boosting	0.9272	0.5410	0.72 ± 0.06	0.9245	0.5401	0.72 ± 0.06
k-Nearest Neighbour	0.9310	0.6100	0.62 ± 0.04	0.9311	0.6023	0.58 ± 0.06
2D Neural Network	0.8829	0.5645	0.55 ± 0.02	0.9145	0.5623	0.60 ± 0.02

ML - machine learning, AUC-ROC - area under the receiver operating characteristic curve



Supplementary Figure 1 *Calibration for 6-variable machine learning model and World Health Organization risk charts in the external validation cohort*

Supplementary Table 2 ***Comparison of predictions in 2007 of 6-variable machine learning model and the World Health Organization risk charts and observed events in 2017***

10-year risk predictions of developing a CVD using 6-variable ML model in 2007 (n)	10-year risk predictions of developing a CVD using the WHO risk charts in 2007 (n)			Number of observed CVDs over 10-years from 2007-2017 (n)	Total Cohort (n)
	<10%	10-19.9%	≥20%		
Low risk	1957	415	45	54	2347
High risk	102	67	10	125	249
Total	2059	482	55	179	2596

BMJ Open

Comparison of cardiovascular risk prediction models developed using machine learning based on data from a Sri Lankan cohort with World Health Organization risk charts for predicting cardiovascular risk among Sri Lankans: a cohort study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-081434.R2
Article Type:	Original research
Date Submitted by the Author:	25-Dec-2024
Complete List of Authors:	Mettananda, Chamila; University of Kelaniya, Pharmacology Solangaarachchige, Maheeka; University of Kelaniya, Exam Unit; Sri Lanka Institute of Information Technology, Haddela, Prasanna; Sri Lanka Institute of Information Technology, Department of IT Dassanayake, Anuradha; University of Kelaniya Faculty of Medicine, Pharmacology Kasturiratne, Anuradhani; University of Kelaniya Faculty of Medicine, Public Health Wickremasinghe, Rajitha; University of Kelaniya Faculty of Medicine, Public Health Kato, Norihiro; National Center for Global Health and Medicine Research Institute, Gene Diagnostics and Therapeutics de Silva, Hithanadura Janaka; University of Kelaniya Faculty of Medicine, Medicine
Primary Subject Heading:	Cardiovascular medicine
Secondary Subject Heading:	Public health
Keywords:	Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PREVENTIVE MEDICINE, Primary Prevention, Cardiac Epidemiology < CARDIOLOGY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Comparison of cardiovascular risk prediction models developed using machine learning based on data from a Sri Lankan cohort with World Health Organization risk charts for predicting cardiovascular risk among Sri Lankans: a cohort study

Mettananda C[1†], Solangaarachchige MB[1,2†], Haddela PS[2], Dassanayake AS[1], Kasturiratne A[1], Wickramasinghe AR[1], Kato N[3], de Silva HJ[1]

[1] Faculty of Medicine, University of Kelaniya, Sri Lanka
[2] Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
[3] National Centre for Global Health and Medicine, Toyama, Shinjuku-ku, Tokyo, Japan

† Contributed equally to this work.

Correspondence to:
Chamila Mettananda
Chamila@kln.ac.lk, chamilametta@hotmail.com

Main text – 3180 words. Abstract – 284words

Abstract

Introduction: Models derived from non-Sri Lankan cohorts are used for cardiovascular (CV) risk stratification of Sri Lankans. We aimed to develop a CV risk prediction model using machine learning (ML) based on data from a Sri Lankan cohort followed up for 10 years and to compare the predictions with World Health Organization (WHO) risk charts.

Design: Cohort study.

Setting: Ragama health study (RHS), an ongoing, prospective, population-based cohort study of patients randomly selected from the Ragama Medical office of Heath area, Sri Lanka, focusing on the epidemiology of non-communicable diseases, was used to develop the model. The external validation cohort included patients admitted to Colombo North Teaching Hospital(CNTH), a tertiary care hospital in Sri Lanka, from January 2019 through August 2020.

Participants: All RHS participants, aged 40-64 years in 2007, without cardiovascular disease (CVD) at baseline, who had complete data of 10-year outcome by 2017, were used for model development. Patients aged 40–74 years admitted to CNTH during the study period with incident CV events or a disease other than an acute CVE with complete data for CVD risk calculation were used for external validation of the model.

Methods: Using the follow-up data of the cohort, we developed two ML models for predicting 10-year CV risk using six conventional CV risk variables(age, gender, smoking status, systolic blood pressure, history of diabetes, and total cholesterol level) and all available variables(n=75). The ML models were derived using classification algorithms of the supervised learning technique. We compared the predictive performance of our ML models with WHO risk charts (2019, Southeast Asia) using area under the receiver operating characteristic curves(AUC-ROC) and calibration plots. We validated the 6-variable model in an external hospital-based cohort.

Results: Of the 2596 participants in the baseline cohort, 179 incident CV events (CVEs) were observed over 10 years. WHO risk charts predicted only 10 CVEs (AUC-ROC: 0.51, CI-0.42-0.60), while the new 6-variable ML-model predicted 125 CVEs (AUC-ROC: 0.72, CI-0.66-0.78) and 75-variable ML-model predicted 124 CVEs (AUC-ROC: 0.74, CI-0.68-0.80). Calibration results (Hosmer-Lemeshow test) for the 6-variable ML-model and the WHO risk charts were $\chi2=12.85$ ($p= 0.12$) and $\chi2=15.58$ ($p= 0.05$), respectively. In the external validation cohort, the sensitivity, specificity, positive predictive value, negative predictive value and calibration of the 6-variable ML-model and the WHO risk charts, respectively, were: 70.3%, 94.9%, 87.3%, 86.6%, $\chi2=8.22$, $p= 0.41$ and 23.7%, 79.0%, 35.8%, 67.7%, $\chi2=81.94$, $p<0.0001$.

Conclusions: ML-based models derived from a cohort of Sri Lankans improved the overall accuracy of CV-risk prediction compared with the WHO risk charts for this cohort of Southeast Asians.

Keywords: Cardiovascular risk, prediction, World Health Organization risk charts, Machine learning, validation, Sri Lanka

Strengths and limitations of this study

- We developed the risk prediction models using machine learning of 10-year follow-up data of individual patients.
- We used 10-year follow-up data from a large, population-based, randomly selected sample to develop the model
- Even though the cohort we used to train the ML model was a community-based, multi-ethnic random cohort, representation of the state sector was less in our cohort compared to the national distribution.
- The data imbalance due to having very few female smokers might have influenced the model's performance, but this was minimised with stratified 10-fold cross-validation.

INTRODUCTION

No cardiovascular (CV) risk prediction models are specific to or derived from Sri Lankans. Therefore, different risk prediction models derived from white Caucasians or models developed for the Southeast Asia region (SEAR) are used for the CV risk stratification of Sri Lankans.

Asians behave differently from white Caucasians in terms of CV risk. Asians have a distinct genetic make-up and a different CV risk factor profile with a higher prevalence of hypertension, diabetes mellitus, central obesity, insulin resistance, and metabolic syndrome than white Caucasians [1]. They are also at increased risk of developing CV diseases (CVDs) compared to white Caucasians at a given risk factor level [1]. There is little agreement between the CV risk predictions of Sri Lankans based on the World Health Organization / International Society of Hypertension (WHO/ISH) risk charts and the Framingham General CV risk charts [2]. Moreover, the CV risk predictions in a Sri Lankan cohort using three different risk models, the National Cholesterol Education Program - Adult Treatment Panel III (NCEP-ATP III), WHO/ISH charts and Systematic Coronary Risk Evaluation (SCORE) charts, were found discordant [3].

The WHO/ISH CV risk charts for the Southeast Asia region-B (SEAR-B) were developed in 2007 with another 14 for different epidemiological sub-regions to predict CV risk of people of those regions that did not have specific risk prediction models derived from their cohorts [4]. Thulani et al. validated 2007 WHO/ISH risk charts among Sri Lankans and observed 81% agreement between predictions and observed events, but were less predictive in females and those at high CV risk [5]. Later, the WHO risk charts were revised and re-calibrated in 2019 to improve predictive capacity and expanded to 21 epidemiological sub-regions that did not have specific risk prediction models. These 2019 WHO risk charts are currently the best available for Sri Lankans [6]. However, in this also, Sri Lanka is grouped under the Southeast Asia epidemiological sub-region together with Indonesia, Cambodia, Laos, Sri Lanka, Maldives, Myanmar, Malaysia, Philippines, Thailand, Timor-Leste, Viet Nam, Mauritius, and Seychelles. Southeast Asians are a heterogeneous population with different socio-economic and cultural backgrounds, and therefore, the risk predictions may not accurately represent Sri Lankans' CV risk.

Therefore, we aimed to develop a CV risk prediction model using machine learning (ML) based on data from a Sri Lankan cohort that was followed up for 10 years and compare the predictions with 2019 WHO (Southeast Asia) risk charts. Moreover, we aimed to validate the new model in an external cohort of Sri Lankans.

METHODS

Machine learning model development

We developed two CV risk prediction models using ML, based on data from a large community-based study on non-communicable diseases, the "Ragama Health Study (RHS)" [3 7], where individuals have been followed up from 2007 to date.

The baseline study population (n=2923) in the RHS was comprised of 35–64-year-old adult residents in the "Ragama Medical Officer of Health (MOH) area" in 2007[7]. Participants were selected by stratified random sampling in the Ragama MOH area, which is a semi-urban health administrative area among 25 districts in Sri Lanka. Participants were followed up for 10 years from 2007 to 2017, during which all CV deaths, non-fatal strokes, and non-fatal myocardial infarctions (including those undergoing percutaneous coronary interventions and coronary artery bypass grafts) were recorded as hard CV events (CVE) by either interviewing patients and their families or perusing clinical notes/death certificates[8].

Data for participants above 40 years of age, who had no history of CVDs at enrolment in 2007 and completed 10-year follow-up (n=2596), were extracted to develop ML-based risk prediction models, as usually risk predictions are calculated in people over the age of 40 years.

Using the 10-year prospective follow-up data for the cohort, using baseline data of those who developed CVEs and those who did not, we developed two ML-based models to predict the 10-year risk of developing a hard CVE using different risk factor combinations. Individuals who could not be traced in 2017 or those whose cause of death could not be verified were excluded. The ML-based models were developed using classification algorithms of the supervised learning technique. The models were

developed in a recursive process (8) in four steps: project design, data preparation, model fitting and inference & deployment (Figure 1). Models were built using the publicly available Google Colab ML platform, the Scikit-learn library in Python (9), and the Train-Test Split method (10). Participant data were split into two groups: the training and testing samples. The training sample was used to build the ML-based models, and the testing sample was used to assess the efficacy of the algorithms built using the training sample. Since the ratio of CVE to non-CVE was highly skewed at 7:93, we performed stratified 10-fold cross-validation, using 2336 individuals for the training sample and the remaining 260 for the test sample to prevent over-fitting.

The predictive performances of the models were compared. We determined the discriminative power using the area under the receiver operating characteristic curve (AUC-ROC, c-index) and the mean F1 scores. The mean of AUC-ROCs for the 10 cross-validation samples was taken as the AUC-ROC of the ML-based model in question. The AUC-ROC and mean F1-score were used to select the best model. A model with a mean F1-score above 0.8, accuracy above 0.85 and AUC c-index closer to 1 was considered suitable for risk prediction [9 10]. We calibrated the models using calibration plots. A model with a Hosmer-Lemeshow test $\chi 2$ value of greater than 20 or a *p-value* of less than 0.05 was considered poor calibration[11].

We trialled six standard ML classification algorithms with different modelling approaches: Decision tree, Random forest, k-nearest neighbour, 2D neural networks, AdaBoost and gradient boosting. We selected the best-fitting model in terms of mean F1-score and AUC-ROC to develop the final model. Grid search was used to optimise the hyper-parameters of the models (11). Data imputation for all models was done using Python's statistical imputation of missing values.

We developed two risk prediction models; one using the six conventional CV risk variables used in the WHO CV risk charts (age, gender, smoking status, systolic blood pressure, history of diabetes, and total cholesterol level) and the other using 75 variables. The total database consisted of 770 variables, including data on demographics, medical history, family history, social history, physical examination, laboratory investigations and non-laboratory investigations like ECG and an

ultrasound scan of the abdomen. Following data wrangling and cleaning, we chose 75 (out of 770) variables following the literature review and using domain knowledge for the ML model development. We excluded variables with missing values ≥ 50%. By machine learning the database, the models predicted individuals likely and unlikely to develop a CVE within the next 10 years.

Internal validation of the machine learning model

We calculated the predicted CVEs over 10 years by 2017, using baseline data (2007 data) and the two ML models separately. Additionally, we calculated the same using the latest 2019 WHO CV risk charts. We compared the predictions of the 6-variable and 75-variable ML models and the WHO model against the observed events using AUC-ROC and mean F1-score.

External validation of the 6-variable machine learning-based model

We externally validated the 6-variable ML model in a separate hospital-based database of 357 consecutive patients, 40–74 years of age, admitted to Colombo North Teaching Hospital (a tertiary care hospital in Sri Lanka) from 1st of January 2019 to 1st of August 2020 who did not have a history of CVEs and presented with an acute incident CVE (acute myocardial infarction or acute stroke) or a disease other than an acute CVE who had complete data for CVD risk calculation. Their predicted risks of developing a CVE were calculated using the most recent pre-morbid risk factor data available up to one year before the incident CVE or the admission to the ward in non-CVE cases. We compared the predictions of the 6-variable model with that of the 2019 WHO risk chart using confusion matrices and calibration plots.

Ethical clearance

This work was approved by the Ethics Review Committee of the Faculty of Medicine, University of Kelaniya, Sri Lanka (P38/09/2006), ML development and external validation cohort (P61/09/2020). Written informed consent was obtained from all the participants.

Patient and public involvement

It was not appropriate or possible to involve patients or the public in the design or reporting plans of our research, but they were involved in the conduct and dissemination of the study. All patients are routinely followed up in a non-communicable disease clinic at the Faculty of Medicine, in collaboration with North Colombo Teaching Hospital (NCTH) Ragama, Sri Lanka, as a service component since 2007. Information about their risk factors was available to participants, and when necessary, they were referred for specialist care at the NCTH. The study results will be disseminated to study participants, other patients, and the public following publication.

RESULTS

A total of 2596 participants followed up for 10 years were eligible for the study with a mean age of 53.5 (SD: 6.9) years and 1162 (44.8%) males. The baseline characteristics of the study cohort are shown in Table 1.

Table 1. Baseline characteristics of the cohort

	Male n = 1162	Female n = 1434	Total n = 2596
Ethnicity n (%)			
Sinhalese	1118 (96.2)	1375 (95.9)	2493 (96.0)
Tamil	15 (1.3)	27 (1.9)	42 (1.6)
Muslim	2 (0.2)	2 (0.1)	4 (0.2)
Burgher	15 (1.3)	19 (1.3)	34 (1.3)
Other	12 (1.0)	11 (0.8)	23 (0.9)
Age groups (years), n (%)			
40-49.9	360 (30.9)	456 (31.8)	816 (31.4)
50-59.9	526 (45.3)	669(46.7)	1195 (46.0)

≥ 60.0	276 (23.8)	309(21.5)	585 (22.6)
Smoking, n (%)	416 (35.8)	0 (0.0)	416 (16.0)
Diabetes, n (%)	165 (14.2)	249 (17.4)	414 (15.9)
Hyperlipidaemia, n (%)	98 (8.4)	209 (14.6)	307 (11.8)
SBP (mmHg), n (%)			
<139.9	766 (65.9)	869 (60.6)	1635 (62.9)
140-159.9	260 (22.4)	365 (25.5)	625 (24.1)
160-179.9	88 (7.6)	132 (9.2)	220 (8.5)
≥180.0	48 (4.1)	68 (4.7)	116 (4.5)
Total cholesterol (mmol/L), n (%)			
<4.0	211 (18.2)	207 (14.4)	418 (16.1)
4-4.9	297 (25.6)	269 (18.8)	566 (21.8)
5-5.9	391 (33.6)	476 (33.2)	867 (33.4)
6-6.9	192 (16.5)	322 (22.5)	514 (19.8)
7-7.9	66 (5.7)	123 (8.6)	189 (7.3)
≥8.0	5 (0.4)	37 (2.5)	42 (1.6)
BMI ≥ 23Kg/m^2, n (%)	590 (50.8)	945(65.9)	1535 (59.1)
BMI ≥ 30Kg/m^2, n (%)	47 (4.0)	166 (11.6)	213 (8.2)

SBP- systolic blood pressure, BMI- body mass index

Over the 10-year follow-up period, 179 hard CVEs were recorded: 66 (36.9%) in females and 113 (63.1%) in males.

We tested six ML algorithms to find the best predictive CV risk prediction model using 6-variables and 75-variables separately. A comparison of model performances using different ML algorithms is shown in Supplementary Table 1. Random Forest models showed the highest accuracy, mean F1-score and AUC-ROC for both 6-variable and 75-variable ML models and were selected as the final ML-based models.

The 20 most important variables in terms of predictive performance in the descending order of the 75-variable model developed on the Random Forest algorithm are shown in Table 2.

Table 2. Variable ranking by their contribution to CV risk predictions

Ranking	Variable	Importance
1	Age	0.08666
2	Smoking status	0.062
3	Height	0.05601
4	Average systolic blood pressure	0.05274
5	Smoking duration	0.05246
6	Sex	0.05149
7	Sugar control for 3 months	0.03583
8	Hip circumference	0.03004
9	Average diastolic blood pressure	0.02795
10	Serum triglyceride level	0.02524
11	Number of packed smoked a day	0.02387
12	History of hypertension	0.02246
13	Baseline insulin level	0.0222
14	LDL Cholesterol	0.022
15	Fasting blood sugar	0.02166
16	Total cholesterol level	0.0191
17	Weight in 2007	0.01904
18	Alcohol used at least once a week	0.01901
19	Waist in 2007	0.01798
20	Body mass index in 2007	0.01788

The predicted CVEs by the newly developed ML-based models (6-variable and 75-variable) and the WHO risk charts (2019) for the next 10 years using baseline data of

2007 were compared with the observed CVEs by 2017 using AUC-ROC curves and confusion matrices (Figure 2).

Discrimination of the three models using AUC-ROC and c-indexes were; 75-variable model: 0.74, (CI-0.68-0.80), 6-variable model: 0.72, (CI-0.66-0.78) and WHO risk charts: 0.51, (CI-0.42-0.60). Accuracy in terms of the rate of prediction of actual CV risk of the population (predicting both true positive and true negative CVEs) was; 75-variable model: 93.1% (2417/2596), 6-variable model: 93.1% (2418/2596) and WHO risk charts: 91.8% (2382/2596) (Figure 2).

The predictive accuracies of the three models were studied using confusion matrices (Figure 2). The 75-variable model predicted 124 of 179 CVEs and 2293 of 2417 non-CVE cases correctly; sensitivity - 69.2%, positive predictive value (PPV) - 50.0%, specificity - 94.8%, negative predictive value (NPV) -97.6%. The 6-variable model correctly predicted 125 of 179 CVEs and 2293 of 2417 non-CVE cases; sensitivity - 69.8%, PPV - 50.2%, specificity - 94.8%, NPV - 97.6%. The WHO risk charts predicted only 10 of 179 cases but 2372 of 2417 non-CVE cases correctly; sensitivity - 5.58%, PPV - 18.1%, specificity - 98.1%, NPV - 93.3%. The 75- and 6-variable models correctly predicted 114 and 115 more CVEs than the 10 CVEs predicted by the latest WHO risk charts.

The calibration for the 6-variable ML model was good as the Hosmer-Lemeshow test result was $\chi 2=12.85$, $p= 0.12$. the Hosmer-Lemeshow test result for the WHO risk charts was $\chi 2=15.58$, $p= 0.05$. (Supplementary Table 2, Figure 3)

The 6-variable ML-based model was validated in an external cohort of 357 hospital-based patients. The external validation cohort consisted of 118 incident CVE cases and 239 non-CVE cases, 117 (32.7%) males with a mean age of 63.4 (SD: 7.2) years. Their CVE risk predictions were calculated using the 6-variable model and WHO risk charts separately. The predicted and observed number of CVEs were compared using confusion matrices (Figure 4). The predictive accuracy of the 6-variable model was 83/118 cases (sensitivity 70.3%, PPV 87.3%) and 227/239 non-CVE cases (specificity

95.0%, NPV 86.6%). In comparison, that of WHO risk charts were 28/118 cases (sensitivity 23.7%, PPV 35.8%) and 189/239 non-cases (specificity 79.0%, NPV 67.7%). The 6-variable model correctly predicted 55 more cases of CVEs than the 28 cases predicted by the currently used 2019 WHO risk charts. Calibration for the 6-variable ML model in the external validation cohort was also good, with the Hosmer-Lemeshow test result of $\chi 2=8.22$, $p= 0.41$, while that of WHO risk charts was $\chi 2=81.94$, $p<0.0001$ (Supplementary Figure 1).

DISCUSSION

We developed two ML-based CV-risk prediction models using longitudinal data of a Sri Lankan cohort prospectively followed up for 10 years. The ML-based models were the first CV risk prediction model developed using individual data from Sri Lankans and the only risk prediction model specific to Sri Lankans. The newly developed 6-variable ML-based model predicted CVE with a 70% sensitivity and 95% specificity in an external cohort. The overall predictive performances of the ML-based models in Sri Lankans were better than that of the reference WHO CV risk charts developed for the whole of South East Asia Region (2019). The newly developed ML-based models appear to be more effective in the risk prediction of people at high CV risk compared to the WHO risk charts and are equally effective as the WHO score in risk predicting people at low CV risk. Validation of the 6-variable ML-based model in an external cohort of Sri Lankans re-confirmed the findings, showing very good calibration for the 6-variable ML model and poor calibration for the WHO risk charts.

Improved CV risk prediction allows for identifying more patients who could benefit from preventive treatment while avoiding unnecessary treatment of low-risk people [12]. The WHO risk charts developed for the Southeast Asia region are good in detecting Sri Lankans at low risk of CVDs but are less sensitive in predicting patients who are at high risk of CVDs. The same was observed while validating the 2007 WHO/International Society of Hypertension risk charts among Sri Lankans [5]. The low accuracy in predicting high-risk individuals using the WHO risk charts could be explained by several reasons. The WHO risk charts were developed using the epidemiological data of the member countries available to predict the CV risk of the

people of the South East Asia region. However, our ML-based models were developed using individual patient data from a Sri Lankan cohort that had been followed up for 10 years and, therefore, are more specific for Sri Lankans. Further, we developed the prediction models using machine learning data from a prospectively followed-up Sri Lankan cohort. ML allows the models to appreciate subtle, complex interactions between variables in predicting outcomes rather than using conventional logistic regression, making our ML-based models more specific for Sri Lankans.

CV risk prediction using ML is now being used globally and reported to be better than traditional risk prediction models [12-17]. Several studies from the UK have shown the superiority of ML-based models over conventional models in predicting CV risk. Alaa et al. showed that the ML-based risk predictions improved the accuracy of CV risk prediction in 423,604 participants of the UK Biobank compared to the Framingham risk score [14]. Another study of 378,256 patients from UK family practices showed that a new ML model using eight conventional variables significantly improved the accuracy of CV risk prediction (10). Another recent study using a novel prediction model comprising 10 predictors in a cohort of UK Biobank showed better performance over multiple existing clinical models [17]. A study involving 143,043 Chinese patients with hypertension also showed that ML outperforms traditional logistic regression for CV risk prediction [16]. Our results for the two ML models in Sri Lankans corroborate these previous findings in other populations.

The study by Alaa et al. using the UK biobank data showed that the predictive capacity of the ML model when using all available 476 variables was better than that when using only the traditional variables [14]. However, we did not find a significant difference in predictive performance when using all available variables (n=75) compared to 6 traditional variables in the ML models in our cohort. Several explanations are possible for the lack of difference between the two ML models in this cohort; e.g., the cohort sample size is too small to identify risk factors with minor contributions, and the 75 variables available in this study do not contain enough to provide additional information to the six traditional variables.

A recent meta-analysis of ML algorithms utilised for CVD prediction has highlighted the importance of using the optimal algorithm for the datasets being used due to the

heterogeneity among ML algorithms [18]. A recent review on artificial intelligence (AI) and CV risk prediction has shown that AI-based predictive models may overcome some of the limitations of classic regression models. Still, the successful application of AI requires knowledge of the potential pitfalls in AI techniques to guarantee their safe and effective use in daily clinical practice [19]. We trialled six standard ML classification algorithms with different modelling approaches, and our models confirmed the importance of the already known conventional CV risk factors in predisposition to CVD. This finding also adds to the validity of our results. In a resource-limited country such as Sri Lanka, our 6-variable model would be more practical than the 75-variable model to screen individuals at higher CV risk, as it is as predictive as the 75-variable model. The 6-variable ML model is more predictive than WHO risk charts, especially in high-risk people, who should be the main target for primary prevention of CVDs.

There are several strengths in our study. Our cohort is a community-based random sample. The study area consisted of 75,591 multi-ethnic residents in 2007. Participants were prospectively followed up for 10 years. The dropout rate was very low, and only the data of participants who completed 10-year follow-ups were used to develop the ML models. Patients were recruited and followed up by medical officers using face-to-face interviews and perusing medical records, including death certificates where applicable, and therefore self-reporting bias was minimised. Individual patient data was used to develop the model. The endpoints used (hard CVE) were clear and objective.

There are some limitations to our study. For example, even though our cohort is community-based, it is from a semi-urban area and may not represent the whole of Sri Lanka. According to the 2012 census, however, the overall national distribution of the population in the urban-rural sectors is 1: 4.5, comparable to 1: 5.4, in the Gampaha district. Imputation of missing data and imbalance of data due to having very few female smokers might have some influence on the model's performance, but this was minimised with stratified 10-fold cross-validation.

In conclusion, we have shown that the new models developed by machine learning individual participant follow-up data of a Sri Lankan cohort were more predictive of CV risk, especially of high-risk Sri Lankans, than the WHO CV risk charts for the Southeast

Asia region (2019). We plan to improve predictions of the model by using data from a larger sample and to develop a web/mobile interphase of the new 6-variable model to increase its clinical utility.

Funding: This study was supported by the Strengthening Research Outputs Grant of the University of Kelaniya, Sri Lanka (RC/SROG/2021/01). The funding bodies played no role in the design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Contributors: CM, MBS, and PSH conceptualised and designed the study. AK, ASD, ARW, NK and HJdeS were involved in establishing the Ragama Health Study cohort. MBS and PSH analysed the data assisted by CM. CM, MBS and HJdeS prepared and revised the manuscript. All authors read and agreed to the final version of the manuscript. CM acted as guarantor

Data availability statement: The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Competing interests: The authors declare that they have no competing interests and no conflicts of interest.

Acknowledgements: We thank all those who have continuously supported the Ragama Health Study, and especially the study participants for their continued cooperation.

References

1. Volgman AS, Palaniappan LS, Aggarwal NT, et al. Atherosclerotic Cardiovascular Disease in South Asians in the United States: Epidemiology, Risk Factors, and Treatments: A Scientific Statement From the American Heart Association. Circulation 2018;138(1):e1-e34. doi: doi:10.1161/CIR.0000000000000580
2. Mettananda KCD, Gunasekara N, Thampoe R, et al. Place of cardiovascular risk prediction models in South Asians; agreement between Framingham risk score and WHO/ISH risk charts. Int J Clin Pract 2021;75(7):e14190. doi: 10.1111/ijcp.14190 [published Online First: 2021/03/30]
3. Ranawaka U, Wijekoon N, Pathmeswaran P, et al. Risk estimates of cardiovascular diseases in a Sri Lankan community. Ceylon Med J 2016;61:11. doi: 10.4038/cmj.v61i1.8253
4. WHO. World Health Organisation/International Society of Hypertension risk prediction charts for 14 WHO epidemiological sub-regions: WHO, 2007:40.
5. Thulani UB, Mettananda KCD, Warnakulasuriya DTD, et al. Validation of the World Health Organization/ International Society of Hypertension (WHO/ISH) cardiovascular risk predictions in Sri Lankans based on findings from a prospective cohort study. PLoS One 2021;16(6):e0252267. doi: 10.1371/journal.pone.0252267 [published Online First: 2021/06/08]
6. WHO. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob Health 2019;7(10):e1332-e45. doi: 10.1016/s2214-109x(19)30318-3 [published Online First: 2019/09/07]
7. Dassanayake AS, Kasturiratne A, Rajindrajith S, et al. Prevalence and risk factors for non-alcoholic fatty liver disease among adults in an urban Sri Lankan population. J Gastroenterol Hepatol 2009;24(7):1284-8. doi: 10.1111/j.1440-1746.2009.05831.x [published Online First: 2009/05/30]
8. Niriella MA, Kasturiratne A, Beddage TU, et al. Metabolic syndrome, but not non-alcoholic fatty liver disease, increases 10-year mortality: A prospective, community-cohort study. *Liver International* 2020;40(1):101-06. doi: https://doi.org/10.1111/liv.14237

9. Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol 2002;20(2):96-107. doi: 10.1053/suro.2002.32521
10. Cichosz P. Assessing the quality of classification models: Performance measures and evaluation procedures. Open Engineering 2011;1(2):132-58. doi: doi:10.2478/s13531-011-0022-9
11. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024;384:e074820. doi: 10.1136/bmj-2023-074820
12. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE 2017;12(4):e0174944. doi: 10.1371/journal.pone.0174944
13. Pal M, Parija S, Panda G, et al. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars) 2022;17(1):1100-13. doi: 10.1515/med-2022-0508 [published Online First: 20220617]
14. Alaa AM, Bolton T, Di Angelantonio E, et al. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One 2019;14(5):e0213653. doi: 10.1371/journal.pone.0213653 [published Online First: 20190515]
15. Dalal S, Goel P, Onyema EM, et al. Application of Machine Learning for Cardiovascular Disease Risk Prediction. Computational Intelligence and Neuroscience 2023;2023:9418666. doi: 10.1155/2023/9418666
16. Xi Y, Wang H, Sun N. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. Front Cardiovasc Med 2022;9:1025705. Doi: 10.3389/fcvm.2022.1025705 [published Online First: 20221114]
17. Jia Y, Yu G, Ju-Jiao K, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke and Vascular Neurology 2023:svn-2023-002332. doi: 10.1136/svn-2023-002332
18. Krittanawong C, Virk HUH, Bangalore S, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific Reports 2020;10(1):16057. doi: 10.1038/s41598-020-72685-1

19. Chiarito M, Luceri L, Oliva A, et al. Artificial Intelligence and Cardiovascular Risk Prediction: All That Glitters is not Gold. European Cardiology Review 2022;17:e29 2022 doi: 10.15420/ecr.2022.11

FIGURES

Figure 1. Machine learning model development process

Figure 2. Comparison of the predictive performance of machine learning-based models and the World Health Organization cardiovascular risk charts (South East Asia Region, 2019) in a Sri Lankan cohort

ML – machine learning, WHO – World Health Organization, CV – cardiovascular

Figure 3. External validation of the 6-variable machine learning model in cardiovascular risk predicting

Figure 4. Calibration for 6-variable machine learning model and World Health Organization risk charts in the original cohort

SUPPLEMENTARY MATERIAL

Supplementary Table 1. Comparison of predictive performances of 6-variable and 75-variable machine learning models

Supplementary Table 2. Comparison of predictions in 2007 of 6-variable machine learning model and the World Health Organization risk charts and observed events in 2017

Supplementary Figure 1. Calibration for 6-variable machine learning model and World Health Organization risk charts in the external validation cohort

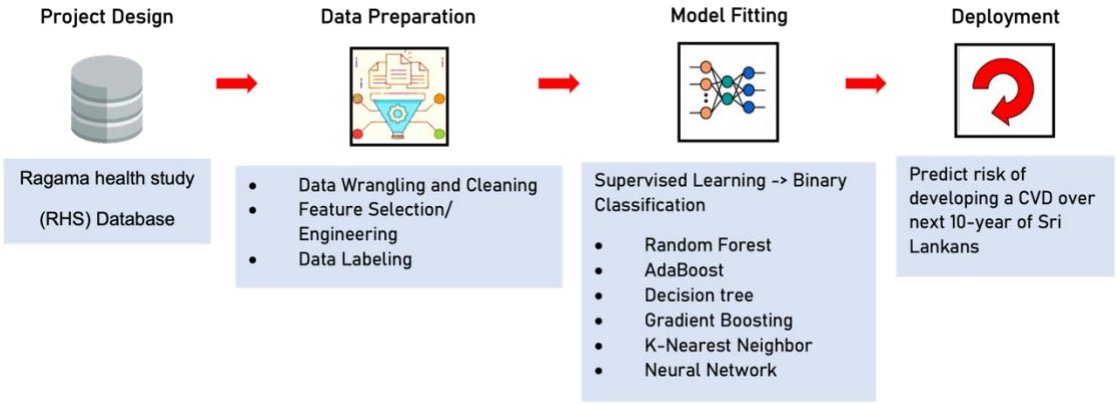
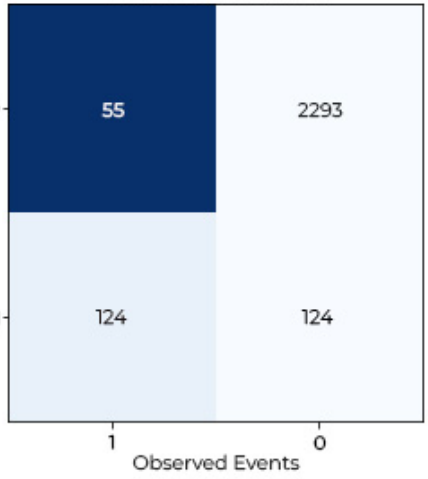
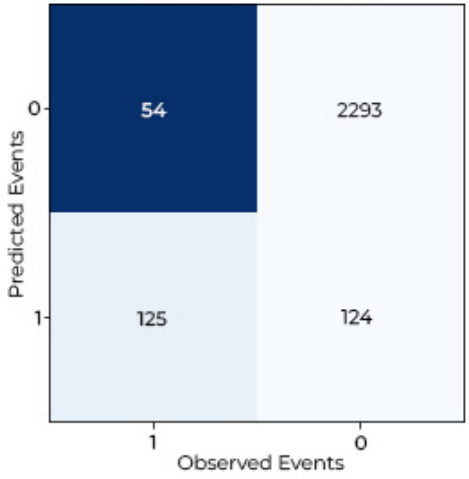
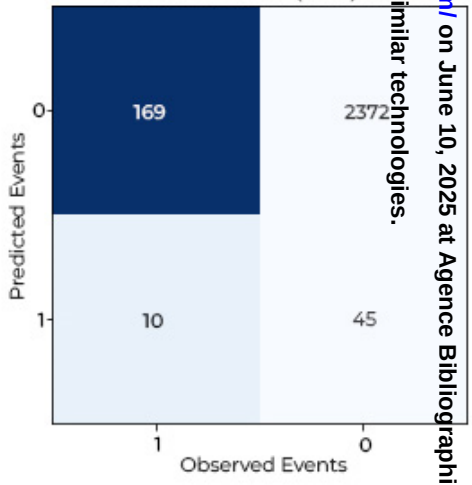
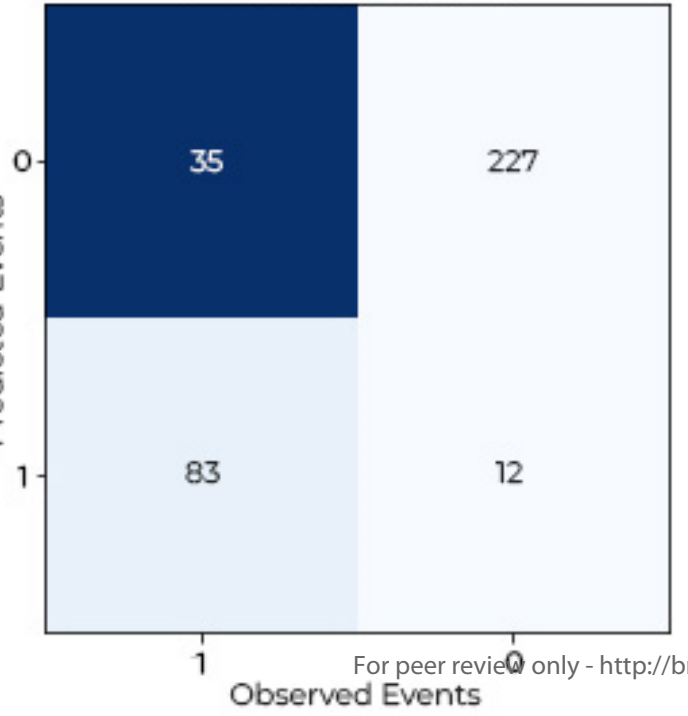


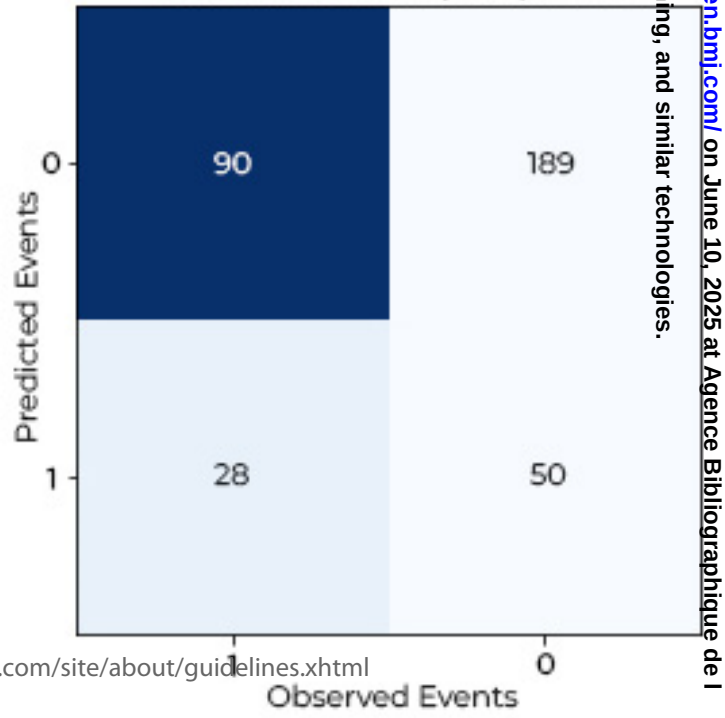
Figure 1











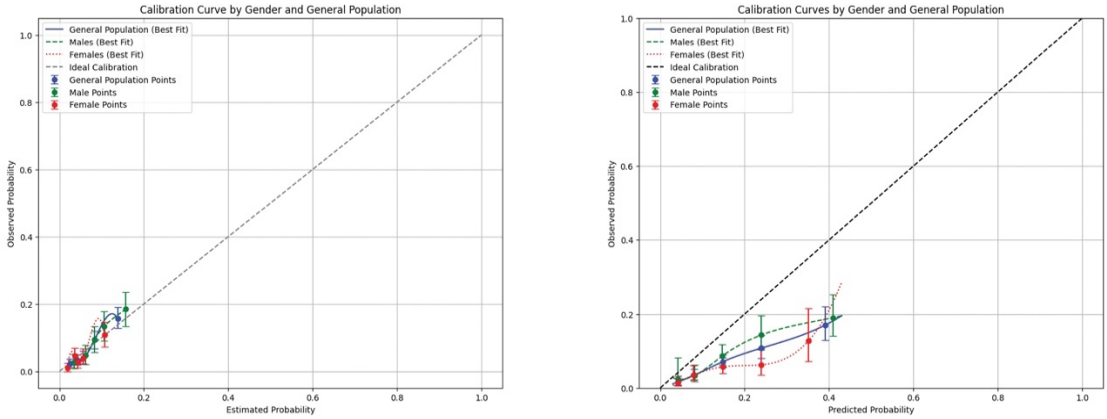
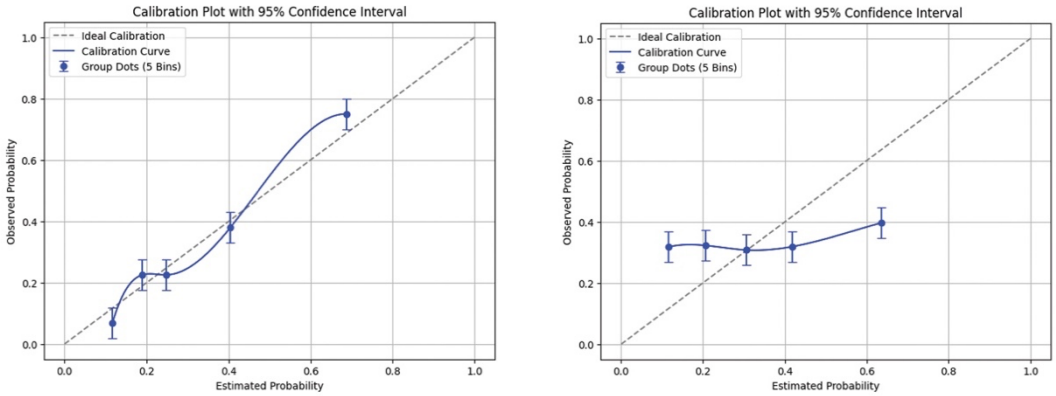


Figure 4 *Calibration for 6-variable machine learning model and World Health Organization risk charts in the original cohort*

Supplementary Table 1 Comparison of predictive performances of 6-variable and 75-

Algorithm	6-variable ML models			75-variable ML models		
	Accuracy	F1-Score	ROC-AUC	Accuracy	F1-Score	ROC-AUC
Random Forest	0.9314	0.8123	0.72 ± 0.07	0.9311	0.8102	0.74 ± 0.06
AdaBoost	0.9291	0.7632	0.68 ± 0.07	0.9199	0.7601	0.64 ± 0.08
Decision tree	0.8733	0.5812	0.55 ± 0.05	0.8663	0.5808	0.51 ± 0.03
Gradient Boosting	0.9272	0.5410	0.72 ± 0.06	0.9245	0.5401	0.72 ± 0.06
k-Nearest Neighbour	0.9310	0.6100	0.62 ± 0.04	0.9311	0.6023	0.58 ± 0.06
2D Neural Network	0.8829	0.5645	0.55 ± 0.02	0.9145	0.5623	0.60 ± 0.02

ML - machine learning, AUC-ROC - area under the receiver operating characteristic curve



Supplementary Figure 1 *Calibration for 6-variable machine learning model and World Health Organization risk charts in the external validation cohort*

Supplementary Table 2 ***Comparison of predictions in 2007 of 6-variable machine learning model and the World Health Organization risk charts and observed events in 2017***

10-year risk predictions of developing a CVD using 6-variable ML model in 2007 (n)	10-year risk predictions of developing a CVD using the WHO risk charts in 2007 (n)			Number of observed CVDs over 10-years from 2007-2017 (n)	Total Cohort (n)
	<10%	10-19.9%	≥20%		
Low risk	1957	415	45	54	2347
High risk	102	67	10	125	249
Total	2059	482	55	179	2596