SVN Stroke & Vascular Neurology

Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study

Jia You,[1] Yu Guo,[1] Ju-Jiao Kang,[1] Hui-Fu Wang,[1] Ming Yang,[1] Jian-Feng Feng,[1,2,3,4,5] Jin-Tai Yu ,[1] Wei Cheng[1,2,5,6]

To cite: You J, Guo Y, Kang J-J, *et al*. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. *Stroke & Vascular Neurology* 2023;**8**: e002332. doi:10.1136/svn-2023-002332

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/svn-2023-002332).

JY and YG contributed equally.

JY and YG are joint first authors.

Received 19 January 2023
Accepted 3 April 2023
Published Online First 27 April 2023

Check for updates

© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Professor Jin-Tai Yu; jintai_yu@fudan.edu.cn

Dr Wei Cheng; wcheng.fdu@gmail.com

Professor Jian-Feng Feng; jianfeng64@gmail.com

ABSTRACT

Background Previous prediction algorithms for cardiovascular diseases (CVD) were established using risk factors retrieved largely based on empirical clinical knowledge. This study sought to identify predictors among a comprehensive variable space, and then employ machine learning (ML) algorithms to develop a novel CVD risk prediction model.

Methods From a longitudinal population-based cohort of UK Biobank, this study included 473 611 CVD-free participants aged between 37 and 73 years old. We implemented an ML-based data-driven pipeline to identify predictors from 645 candidate variables covering a comprehensive range of health-related factors and assessed multiple ML classifiers to establish a risk prediction model on 10-year incident CVD. The model was validated through a leave-one-center-out cross-validation.

Results During a median follow-up of 12.2 years, 31 466 participants developed CVD within 10 years after baseline visits. A novel UK Biobank CVD risk prediction (UKCRP) model was established that comprised 10 predictors including age, sex, medication of cholesterol and blood pressure, cholesterol ratio (total/high-density lipoprotein), systolic blood pressure, previous angina or heart disease, number of medications taken, cystatin C, chest pain and pack-years of smoking. Our model obtained satisfied discriminative performance with an area under the receiver operating characteristic curve (AUC) of 0.762±0.010 that outperformed multiple existing clinical models, and it was well-calibrated with a Brier Score of 0.057±0.006. Further, the UKCRP can obtain comparable performance for myocardial infarction (AUC 0.774±0.011) and ischaemic stroke (AUC 0.730±0.020), but inferior performance for haemorrhagic stroke (AUC 0.644±0.026).

Conclusion ML-based classification models can learn expressive representations from potential high-risked CVD participants who may benefit from earlier clinical decisions.

INTRODUCTION

As the leading cause of death globally, cardiovascular diseases (CVD) are responsible for >17.3 million deaths per year and the incidence is estimated to increase to 23.6 million by 2030.[1] Due to the long preclinical stage before disease diagnosis, there is an urgent

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Several cardiovascular diseases (CVD) prediction models have been developed to deliver individual-level risk prediction and stratification. These scales were established using predictors largely retrieved based on empirical clinical knowledge, such as age, sex, cholesterol level, systolic blood pressure and smoking status, while many other potential factors can be ignored; in addition, they were conducted through traditional statistical methods of Cox regression restricting their prediction power and solidness. Recently, applications of machine learning (ML) were commonly restricted to limited follow-up time, the inclusion of too many covariates, or utilisation of prespecified domain of variables or participants, narrowing their applications to research or expertise settings.

WHAT THIS STUDY ADDS

⇒ In this study, predictor selections were optimised through a deliberately designed ML-based data-driven pipeline among a rich phenotypic feature space. Predictors of cystatin C, number of current medications, self-reported chest pain and pack-years of smoking (derived variable) have been reported to be associated with CVD risk. Still, they are not commonly used in prediction models. Our predictor selection strategy indicates their significance in the CVD prediction task. Comparable performance can be observed when deploying to subdiagnostic diseases of myocardial infarction and ischaemic stroke, but much inferior to haemorrhagic stroke. The UK Biobank CVD risk prediction model demonstrated a superiority over existing risk prediction models.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Predictors leveraged by the proposed model can be easily accessible, indicating the potential utility in practice and identifying suspected individuals to aid clinical decisions. Further, several identified predictors are intervenable at an early stage, which might be beneficial to reduce or delay the development of CVD.

need for effective approaches to identify high-risk individuals early. Personalised risk assessment is consistently a long-term goal in cardiovascular medicine, and the European Society of Cardiology specifically emphasises the role of risk prediction models in promoting healthcare and population-wide prevention.[2 3]

Numerous multivariable prediction models have been derived and published, such as the Framingham Cardiovascular Risk Score (FGCRS),[4 5] Systematic Coronary Risk Evaluation (SCORE),[6 7] QRISK[8–10] and American Heart Association/Atherosclerotic Cardiovascular Disease (AHA/ASCVD) algorithms.[11] These risk models were mostly developed by integrating information on a variety of traditional CVD risk factors such as age, sex, smoking status, blood pressure, diabetes and cholesterol measurements. However, they may not fully capture other potential factors associated with elevated cardiovascular risk, limiting their model accuracy in discriminating high-risk populations. In addition, these models were primarily implemented by classical statistical modelling techniques (Cox or logistic regressions), which were restricted by assumptions of normality of distributions, non-informative or random censoring, and linearity of risk prediction.

Compared with conventional prediction methods, the machine learning (ML)-based approach has outstanding advantages, as it can effectively handle massive amounts of time-to-event data featuring multidimensional space.[12–21] However, they were commonly restricted to limited follow-up time,[12 13] inclusion of too many covariates,[14] or utilisation of prespecified domain of variables[15–18] or participants,[19 20] narrowing their applications to research or expertise settings. All these limitations underscore the need to construct novel cardiovascular prediction to boost better risk prediction and stratification in real-world clinical practice.

As illustrated in figure 1, in the present study, we used the ML approach to train and validate a novel CVD risk prediction algorithm called UK Biobank CVD risk prediction (UKCRP) in the large prospective UK Biobank cohort with thorough phenotypic and follow-up information. We sought to develop a generalisable model to predict the 10-year risk of incident myocardial infarction (MI), ischaemic stroke (IS) and haemorrhagic stroke (HS) at the individual level. Then we compared the performance of UKCRP with established CVD risk scores to explore the superiority of our proposed model. We hypothesise that the UKCRP Score will be of reference value for clinical CVD risk prediction.

METHODS

Study population and target outcomes

Our study adopted data from the UK Biobank, a longitudinal cohort study of 273 383 women and 229 122 men aged between 37 and 73 years at the time of their baseline assessment starting from March 2006 to October 2010.[22] The cohort enrolled the general population from 22 recruitment centres across the UK to undergo multiple assessments. As shown in figure 2, participants without follow-up records (n=1298), with pre-existing MI (n=20 037) or stroke (n=8984) were excluded from the analysis. We finally included 473 611 CVD-free participants who had at least 10 years of follow-up until March 2021.

The primary outcomes for this study were incident (hospitalisation or death) CVD events due to MI (international classification of diseases (ICD)-9 codes 410–412 and 429.79; ICD-10 codes I21–I23, I24.1 and I25.2) and stroke (ICD-9 codes 430, 431, 434 and 436; ICD-10 codes I60, I61, I63 and I64). Dates and causes of hospital admission were obtained through record linkage to Hospital Episode Statistics Admitted Patient Care (England), Patient Episode Database for Wales and Scottish Morbidity Records—General/Acute Inpatient and Day Case Admissions (Scotland). Follow-up visits began from the date attending the assessment centre to the earliest date of CVD diagnosis, death or the last available date from the hospital or general practitioner, whichever occurred first.

Candidate variables and predictor selection

This study included all clinically relevant variables collected at the participants' baseline visits. Data screening was processed to exclude non-informative variables whose missing values were over 70% of all participants. Overall, we adopted 645 measurements (online supplemental eTable 1 and 2) categorised into 10 groups in the analysis: biofluid assays (n=70), cognitive function (n=71), early life factors (n=10), family history (n=28), health and medical history (n=46), lifestyle and environment (n=143), medications (n=9), physical measures (n=197), psychosocial factors (n=34) and sociodemographics (n=37). In addition, we calculated genetic polygenic risk scores (PRSs) (n=45) (online supplemental eMethods).

Predictors for model development were identified in two steps: variable importance ranking and sequential forward selection. The importance of each variable was determined using a preliminary trained light gradient boosting machine (LGBM) classifier,[23] based on which, the top 50 variables were chosen. Next, they were fed into a hierarchical clustering algorithm to further remove redundant ones with multicollinearity. Afterward, we reranked the preselected variables and then employed multiple ML classifiers to consecutively add predictors one at each time. Finally, the classifier was determined on achieving the best performance of area under the receiver operating characteristic curve (AUC), and we arbitrarily chose the top 10 predictors for further model development, as no incremental performance was observed after the 10th iteration (figure 3A).

ML classifiers

We implemented a range of ML techniques, including LGBM, eXtreme Gradient Boosting Machine, random forest, logistic regression, K-nearest neighbours, support vector machine and artificial neural networks. All classifiers performed the classification task of determining

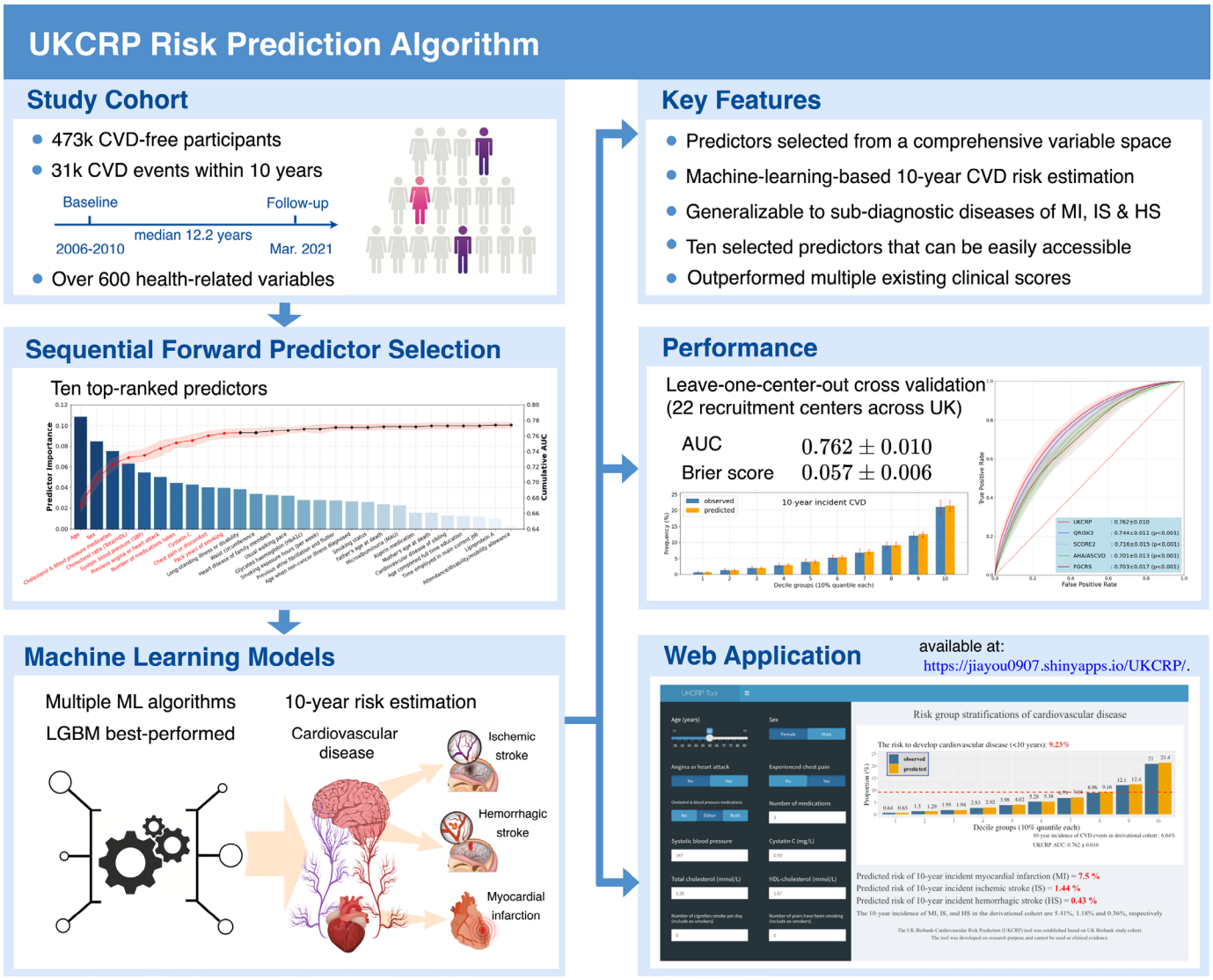

Figure 1 Graphical abstract. The left panels exhibit the development pipeline of the UK Biobank cardiovascular disease risk prediction (UKCRP) model. The right panels demonstrate the key features, performance and an illustrative example of a webpage application. AHA/ASCVD, American Heart Association/Atherosclerotic Cardiovascular Disease; AUC, area under the receiver operating characteristic curve; CVD, cardiovascular diseases; FGCRS, Framingham Cardiovascular Risk Score; HDL, high-density lipoprotein; HS, haemorrhagic stroke; IS, ischaemic stroke; LGBM, light gradient boosting machine; MI, myocardial infarction; ML, machine learning; QRISK V.3, a cardiovascular disease risk score derived through QResearch database V.3; SCORE V.2, Systematic Coronary Risk Evaluation V.2.

whether a participant falls into class 0 (predicted to remain CVD free) or class 1 (predicted to incident CVD within 10 years). Incorporating 10 preidentified predictors, ML models were established based on dichotomised participants of healthy control (n=442 153) and 10-year incident CVD (n=31 466). Postprocessing isotonic regressions[24 25] were adopted to calibrate the raw predicted probabilities into actual incident CVD risks. We used the best-performed method, LGBM, as the final algorithm to develop a cardiovascular risk prediction model, named UKCRP.

Statistical analysis

Our model development and validation were conducted using a leave-one-center-out cross-validation. Specifically, we split the dataset based on 22 assessment centres at recruitment. Notably, 3 centres recruited participants less than 1% of the whole study population were merged into 1; thus, we partitioned the data into 20-fold subsets. Within each of the 20 cross-validation loops, 19-fold of data were used for model development and the remaining fold of data was used as a test set. Hyperparameter optimisation and calibrations were performed within an inner-looped cross-validation of the 19-fold of training dataset.

The model's performance was assessed using discrimination and calibration. Discrimination was evaluated by the AUC. Calibration was visually depicted using calibration plots of decile subgroups and Brier scores[26] were calculated. Metrics of accuracy, sensitivity, specificity,

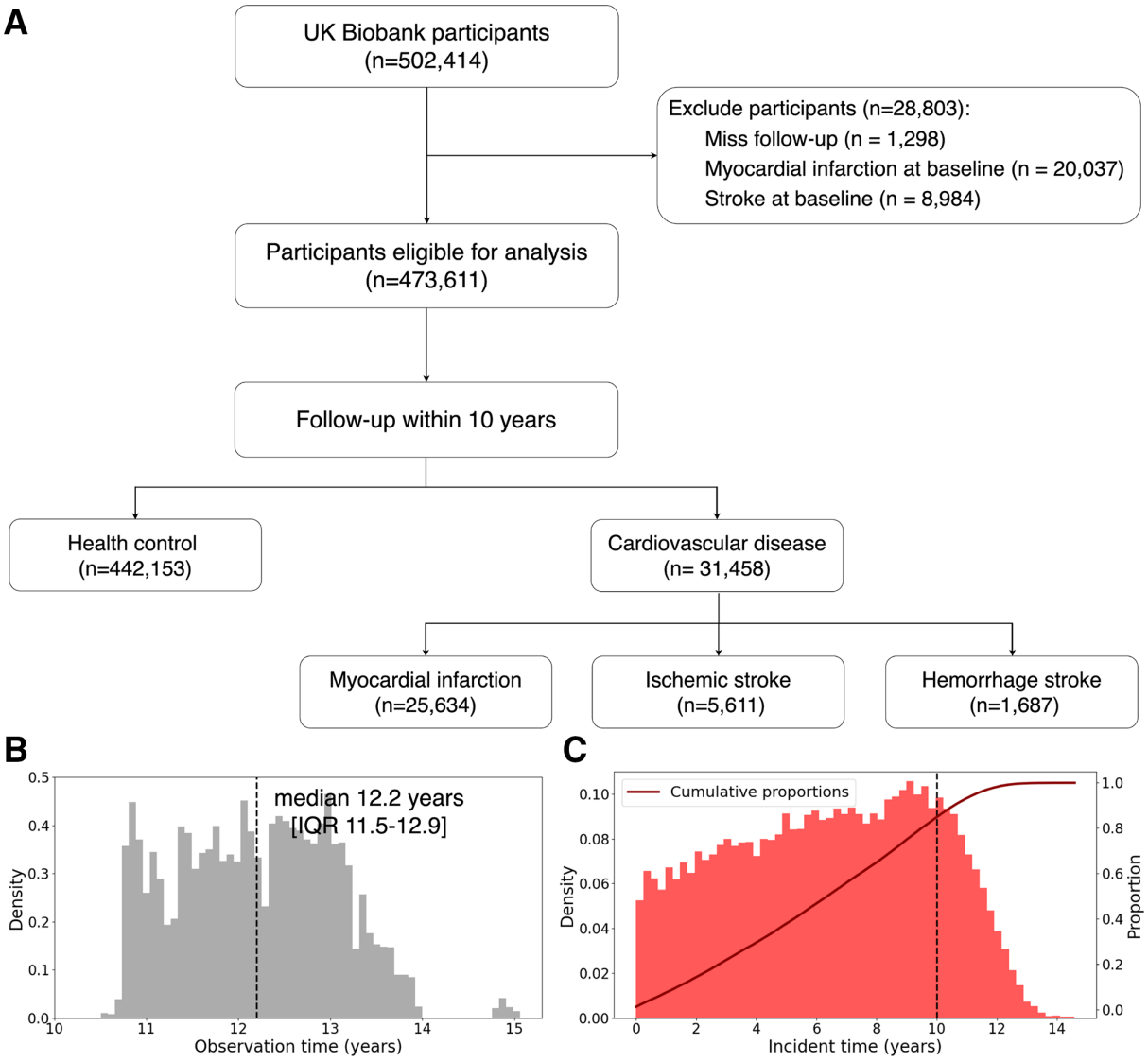

Figure 2 Flowchart of the participants' selection. (A) Individuals in the UK Biobank cohort were excluded if they had reported myocardial infarction at baseline, stroke at baseline or without follow-up records. The remaining participants were classified based on their first reported years of cardiovascular disease after baseline. (B) Distribution of observation times up to 31 March 2021. (C) Distribution of first reported years of cardiovascular disease after baseline.

precision and F1-score were reported as well. In addition, we leveraged SHapley Additive exPlanations (SHAP) plots[27] to graphically interpret the relationship between predictor values and model output.

Hyperparameter optimisation for each classifier was conducted with a grid search on different parameter spaces (online supplemental eTable 3). Missing values were not imputed for the LGBM algorithm as it can automatedly handle missingness in both training and prediction. As for the other ML algorithms, standard scaling and imputation of missing values were performed for the classifiers that typically benefit from these procedures.

We compared the UKCRP with existing CVD risk prediction models, namely, AHA/ASCVD, FGCRS, QRISK V.3 and SCORE V.2 (online supplemental eTable 4). Besides, we investigated the contribution of genetic variables by adding additional PRSs on the top 10 selected predictors. Delong's tests[28] were adopted to assess the significance of AUCs between the UKCRP and previous models. In addition, we deployed the UKCRP to each subdiagnostic group of MI, IS and HS, respectively. For comparison purposes, we repeated the predictors selection and model development procedures under each target population and compared those specific-developed ML models with the UKCRP. Further, we evaluated the UKCRP in subgroups of participants across different age groups and different CVD incident timelines. Lastly, to demonstrate the superiority of our predictor selection strategy, we employed the LGBM classifier, the algorithm to develop UKCRP, to retrain existing prediction scales based on the adopted predictors themselves.

The ML algorithms were implemented with libraries of LightGBM (V.3.3.2)[23] and scikit-learn (V.1.0.2) in Python (V.3.9). Model visualisations were performed using Shap

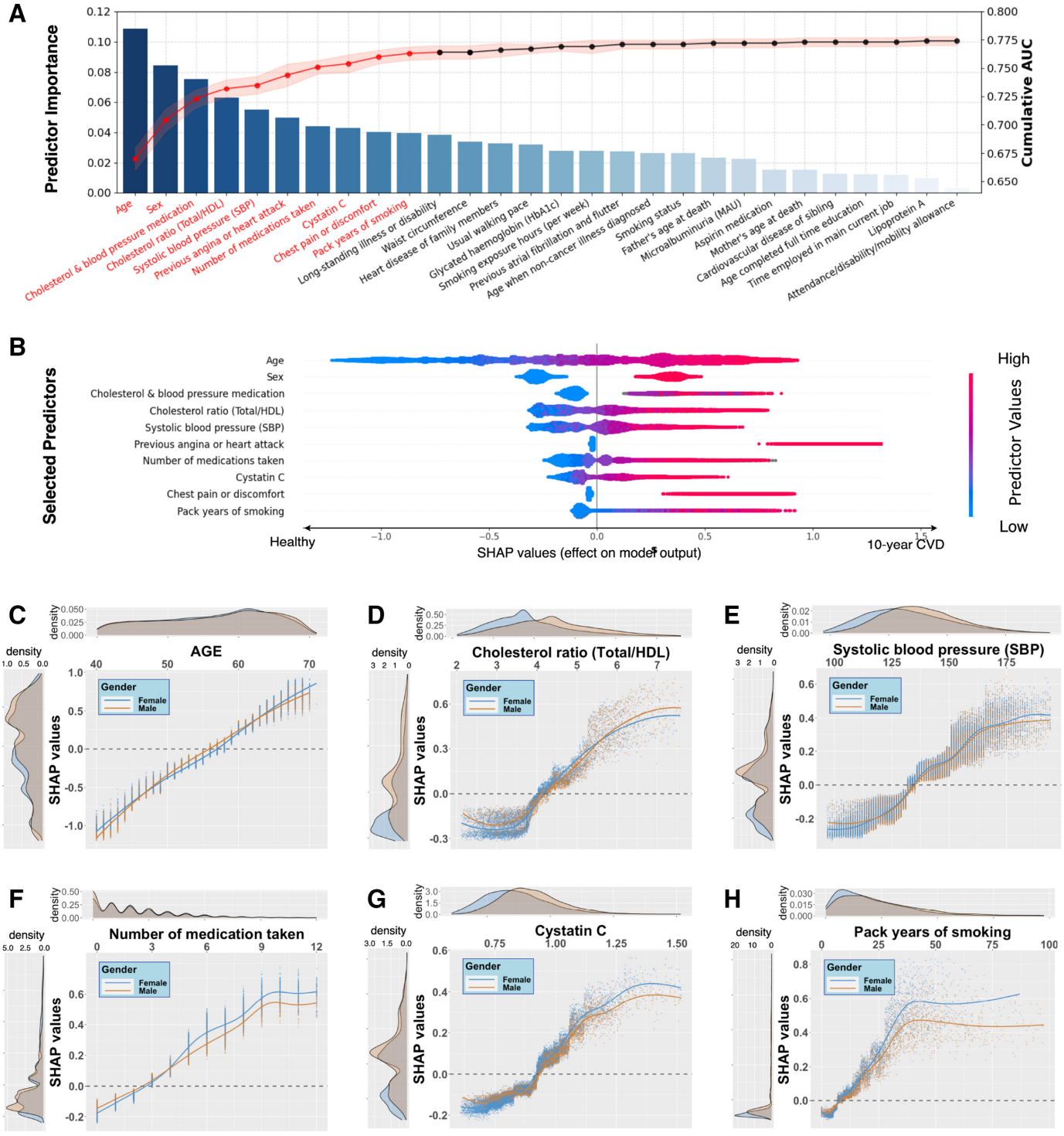

Figure 3 Predictor selection and interpretation of UK Biobank Cardiovascular disease Risk Prediction. (A) Sequential forward selection from preselected candidate variables. The bar chart represents the sorted variables based on their importance to the classification task. The line chart delineated cumulative area under the receiver operating characteristic curves (AUCs) (right axis) on the inclusion of variables one by each iteration. Ten predictors (coloured in red) were finally chosen for model development. (B) SHapley Additive exPlanations (SHAP) visualisation of selected predictors. Each participant was exhibited as a data point and was coded with gradient colours representing the magnitude of the predictor. The predictive power of each predictor can be visually measured by its horizontal range. The effect of each predictor can be depicted by its value magnitude and tendency direction on the x-axis, which represents the extent of the likelihood of developing cardiovascular diseases (CVD). (C–H) Six examples of scatterplots showing the relationship between continuous predictors and model outputs (represented using SHAP values). Horizontal dash lines indicate no effect on model output, whereas above or below the line represents positive or negative effects, respectively. Density plots on the upper and right panels of each graph demonstrate the distributions of each predictor and the corresponding SHAP values. HDL, high-density lipoprotein.

Table 1 The baseline characteristics of UK Biobank participants included in the study

Participants characteristics	Overall (n=473 611)	Healthy control (n=442 145)	10-year incident CVD (n=31 466)
Age, year	57.0 (50–63)	57.0 (49.0–63.0)	62.0 (57.0–66)
Sex (female)	264 308 (55.8%)	252 586 (57.1%)	11 722 (37.3%)
Ethnicity (white)	445 075 (94.0%)	415 611 (94.0%)	29 464 (93.6%)
Systolic blood pressure (mm Hg)	134(123–147)	134.0 (122–147)	141 (129–155)
Total cholesterol (mmol/L)	5.7 (4.99–6.46)	5.71 (5.00–6.46)	5.63 (4.80–6.48)
HDL cholesterol (mmol/L)	1.41 (1.18–1.68)	1.42 (1.19–1.69)	1.28 (1.09–1.54)
Cholesterol ratio (total/HDL)	3.97 (3.37–4.73)	3.95 (3.35–4.70)	4.3 (3.59–5.09)
Cystatin C (mg/L)	0.88 (0.80–0.98)	0.88 (0.8–0.97)	0.95 (0.86–1.06)
Chest pain	24 126 (5.1%)	20 107 (4.5%)	4019 (10.4%)
Current smoker	49 135 (10.4%)	44 372 (10.0%)	4763 (15.1%)
Pack years of smoking	18.4 (9.50–31.0)	17.8 (9.0–30.0)	25.0 (13.8–39.5)
Cholesterol and blood pressure medication			
Either	76 018 (16.1%)	67 842 (15.3%)	8176 (26.0%)
Both	37 376 (7.9%)	31 462 (7.1%)	5914 (18.8%)
Number of medications	2 (0–3)	2.0 (0.0–3.0)	3.0 (1.0–5.0)
Angina or heart attack	7482 (1.6%)	4890 (1.1%)	2592 (8.2%)
Diabetes	21 601 (4.6%)	18 078 (4.1%)	3523 (11.2%)
Hypertension	38 988 (8.2 %)	33 279 (7.5%)	5709 (18.1%)

Data is presented as median (IQR) for continuous variables and number (%) for discrete variables. Statistics of 10-year incident myocardial infarction, ischaemic stroke and haemorrhagic stroke are shown in online supplemental eTable 3. HDL, high-density lipoprotein.
CVD, cardiovascular diseases.

library (V.0.40.0).[27] In addition, we established a webpage application tool based on the Shiny package (V.1.7.1) under R (V.4.1.2).

RESULTS

Population characteristics

After quality control, our study included 473 611 participants who were predominantly white (94.0%). The median age of the participants was 57 years (IQR 50–63) and 55.8% were women. During a median follow-up time of 12.2 years (IQR 11.5–12.9), 31 466 (6.6%) participants developed CVD within 10 years after their baseline visits, among whom 25 634 (5.4%), 5611 (1.2%) and 1687 (0.4%) developed MI, IS and HS, respectively. The critical baseline predictors are presented by incident CVD status in table 1 and by incident subdiagnostic disease in online supplemental eTable 5.

Data-driven predictors selection

Among the 645 candidate variables, we initially chose the top 50 ones based on a naïve-trained LGBM classifier and further adopted hierarchical clustering to select variables that were not correlated with each other (online supplemental eFigure 1). As shown in the bar chart of figure 3A, a set of 28 variables was determined and reranked based on their importance to the prediction task. We employed a sequential forward selection strategy, which can be delineated using the line chart to balance the trade-off between the model performance (AUC on the right axis) and the number of variables involved. The line chart witnessed a steep increase when taking in the first several variables and reached a plateau when additional ones joined in. The LGBM was witnessed as the best-performed classifier (online supplemental eTable 6) and we chose the first 10 variables as the final predictors for ML model development. The top-selected variables selected under subdiagnostic diseases were reported in online supplemental eTable 7. Detailed notations of finally included variables are given in online supplemental eTable 8, and their OR and p values calculated based on a multivariate logistic regression are shown in online supplemental eTable 9.

Performance of UKCRP

We employed multiple ML algorithms based on the selected predictors, and all models demonstrated comparable performance of AUCs around 0.75–0.76 as shown in online supplemental eTable 6. LGBM exhibited the best performance (AUC 0.762±0.010) and was adopted as the final risk prediction model of UKCRP.

Figure 4E–H exhibits the calibrated risks of UKCRP on CVD and three subdiagnostic groups. All calibration plots demonstrated closely matched proportions of observed events versus predicted risks throughout all deciles. Brier

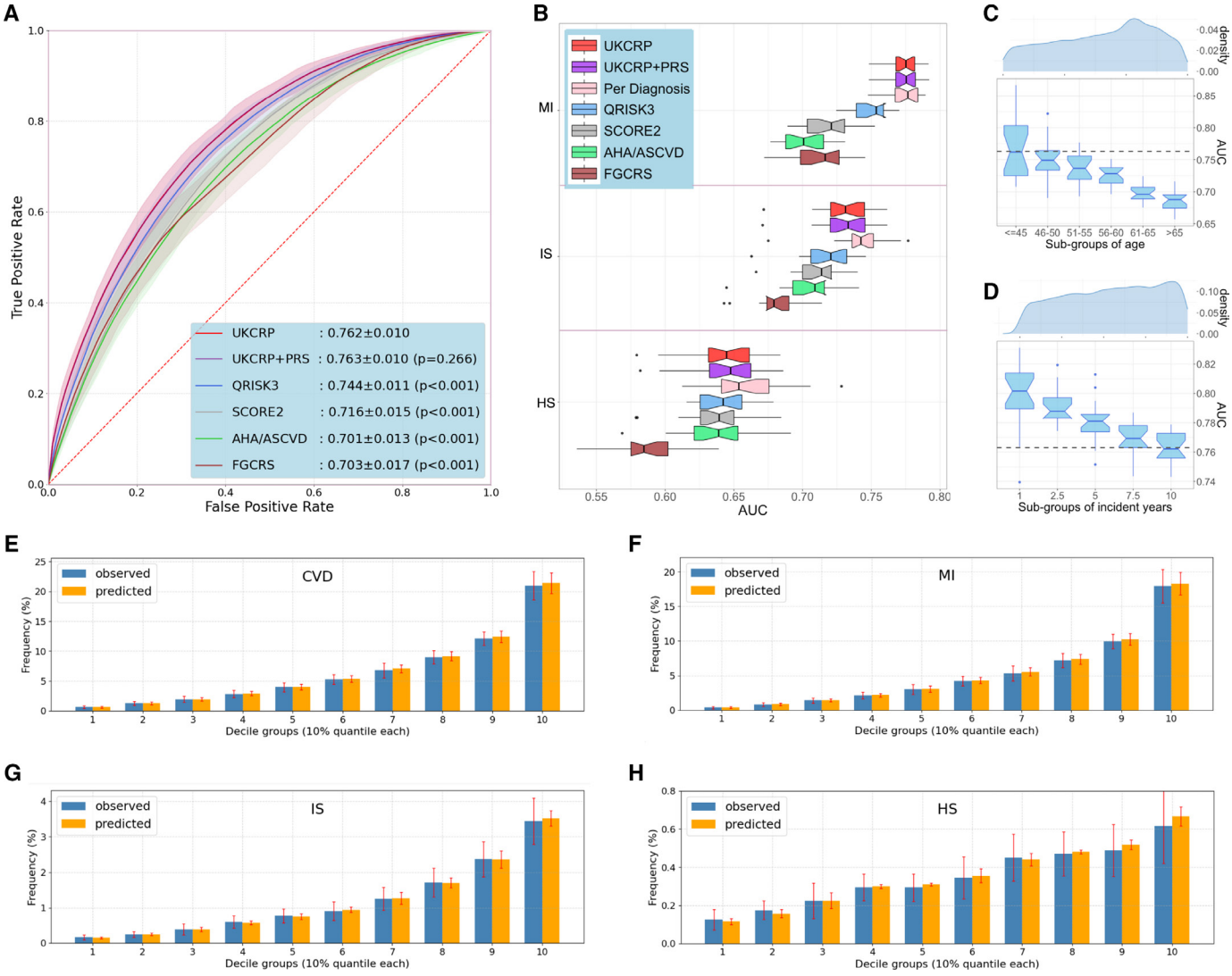

Figure 4 Performance of the UK Biobank Cardiovascular disease Risk Prediction (UKCRP) and subgroup analysis. (A) Area under the receiver operating characteristic curve (AUC) plots of different models of cardiovascular disease. (B) AUC of different models evaluated on different diagnostic subgroups. (C) AUC of the UKCRP that evaluated based on different age subgroups. (D) AUC of the UKCRP that evaluated based on different incident time subgroups. (E–H) Calibration plots of UKCRP on different diagnostic populations. Specific isotonic regressions were trained for each population. AHA/ASCVD, American Heart Association/Atherosclerotic Cardiovascular Disease; FGCRS, Framingham Cardiovascular Risk Score; HS, haemorrhagic stroke; IS, ischaemic stroke; MI, myocardial infarction; PRS, polygenic risk score; QRISK V.3, a cardiovascular disease risk score derived through QResearch database V.3; SCORE V.2, Systematic Coronary Risk Evaluation V.2.

Score for CVD was 0.057±0.006, and that for MI, IS and HS were 0.047±0.006, 0.012±0.002 and 0.004±0.001, respectively. More evaluation metrics were reported in online supplemental eTable 10.

Model interpretation of selected predictors

We leveraged SHAP values to assess the effect of each selected predictor visually. Figure 3B aims to interpret the specific impact of each predictor by its value magnitude (coded by a gradient of colours) and tendency direction on the horizontal axis (the likelihood of developing CVD). Take the predictor age as an example; older participants (coloured in red) are more likely to develop CVD (right side) compared with younger ones (coloured in blue) who tend to keep healthy (left side). Similarly, for the rest predictors, participants of males, and those who took cholesterol and blood pressure medications, had higher cholesterol ratios, had higher systolic blood pressure (SBP), previous experienced angina or heart attack, took more medications, had higher cystatin C, suffered chest pains and had more pack-years of smoking, were prone to CVD.

We examine the relationships between continuous predictors versus their risk or protective effect on the CVD in figure 3C–H, where the horizontal dash lines indicate no effect. It can be inferred that age of 55, cholesterol ratio of 4, SBP of 130 and cystatin C of 0.9 can be regarded as cut-offs in positive contributions to developing CVD. The number of current medications and

pack-years of smoking donates relatively no protective mechanism as minimal proportions of data distributed under the dash lines; in another saying, medication taken and smoking only cause increased risks of developing CVD. Specifically, although the participant's sex individually has a significant impact on the CVD risk predictions, it does not witness any strong interactive effects regarding these predictors.

Performance comparison between existing clinical models

According to figure 4A, UKCRP observed an AUC of 0.762±0.010 on 10-year incident CVD, which is significantly greater than those of existing risk scales of QRISK V.3 (0.744±0.011), SCORE V.2 (0.716±0.015), ASCVD (0.701±0.013) and FGCRS (0.703±0.017), and Delong tests indicated statistical significance with p values all less than 0.001. Further, by using LGBM algorithm to retrain predictors adopted in each pre-existing prediction scale, UKCRP exhibited higher AUC in the prediction of CVD, MI; as for IS, AUC of UKCRP surpassed SCORE V.2, AHA/ASCVD and FGCRS and comparable to QRISK V.3. As for HS, no obvious superiority was discovered (online supplemental eTable 10).

Subgroup analysis for different diagnosis

figure 4B demonstrated the discrimination performance in participants with different diagnoses. In general, all models demonstrated a similar pattern that performed best in discriminating MI, followed by IS, and performed worst on HS. Specifically, the UKCRP performed considerably better on MI and IS when compared with existing risk scales, but no obvious superiority in HS prediction. Individualised models for each diagnosis subgroup ('per diagnosis' in figure 4B) witnessed slightly better performance than the UKCRP, and it is conceivable as they were developed on customised predictor selection and hyperparameter optimisation procedures for each target outcome. Their customised predictor sets are reported in online supplemental eTable 7 and eFigure 2–4.

Subgroup analysis of age and incident years

As shown in figure 4C, the predictive power of UKCRP dropped as participants' age increased. This might be due to the drop in sensitivity over the other risk factors for older participants. As for different CVD incident times, the UKCRP seems to perform better when constraining to shorter incident timelines (figure 4D). This can be partially explained by the timing effectiveness of the data, the longer period after the baseline visits, the less effective of data's predictive power.

Predictive contribution of genetic information

We investigated the added values of PRS, which did not show any significance in all predictions of CVD and subdiagnosis (figure 4A,B).

Webpage deployment tool

We implemented the UKCRP model into a web application (figure 5) that provides risk predictions for individuals based on input predictors. It can calculate the risk of CVD and exhibit it on the calibration plots to delineate explicit visualisations. The web application was made accessible online (https://jiayou0907.shinyapps.io/UKCRP/).

DISCUSSION

In this individual-level analysis of a large prospective cohort, we developed a novel ML-based tool to predict the 10-year risk of incident CVD. From a massive number of health-related variables, we employed a series of data-driven selection schemes and identified the 10 most important predictors. The proposed model of UKCRP yielded an AUC of 0.762 for CVD, outperforming multiple existing clinical models. The UKCRP was well-calibrated with excellent agreement between predicted risks and observed proportions of events. Its deployment to the prediction of MI and IS achieved comparable performance, but inferior performance for HS. Added values of genetic information of PRS did not observe significant improvement in model discriminations. Our proposed risk tool is easy to implement in practice and will optimise the identification of suspected individuals to aid clinical decision-making.

As to the deployment of subdiagnostic groups, the proposed UKCRP model demonstrated consistent results with existing models of AHA/ASCVD, FGCRS, QRISK V.3 and SCORE V.2 that exhibited the best predictive ability for MI, followed by IS, and the worst for HS. The actual incidence of MI, IS and HS ranging from high to low may be partly responsible for this result. Given the heterogeneity of disorders, distinct models leveraging disease-specific training were indeed demonstrated to be superior to the UKCRP model. In subgroup analysis, we observed gradually decreased AUCs along with each 5-year increase in age. This indicates that the association between risk factors and incident CVD may be stronger in younger people, as supported by recent publications.[29–31] Although the prediction accuracy reduced with the increment of incident years, the AUC of 10-year CVD risk remained above 0.76. Accordingly, our model is robust enough to predict short-term and long-term CVD risk. In line with former studies,[32 33] we subsequently demonstrated that the effect of PRS addition on risk discrimination improvement was trivial. This further highlights the real-world transportability of our proposed model, which could achieve good predictive performance using only routinely available parameters.

Overcoming the weakness of previous algorithms incorporating only a few traditional predictors,[5 7 10 11 34] our predictor selection pipeline allows identifying significant predictors from 645 variables. All the top 10 predictors for model development can be easily obtained through quick questionnaires or blood sampling, which provides the general population with the opportunity to perform automated and rapid health screening. Advanced age and male sex are the two most critical risk factors, with a

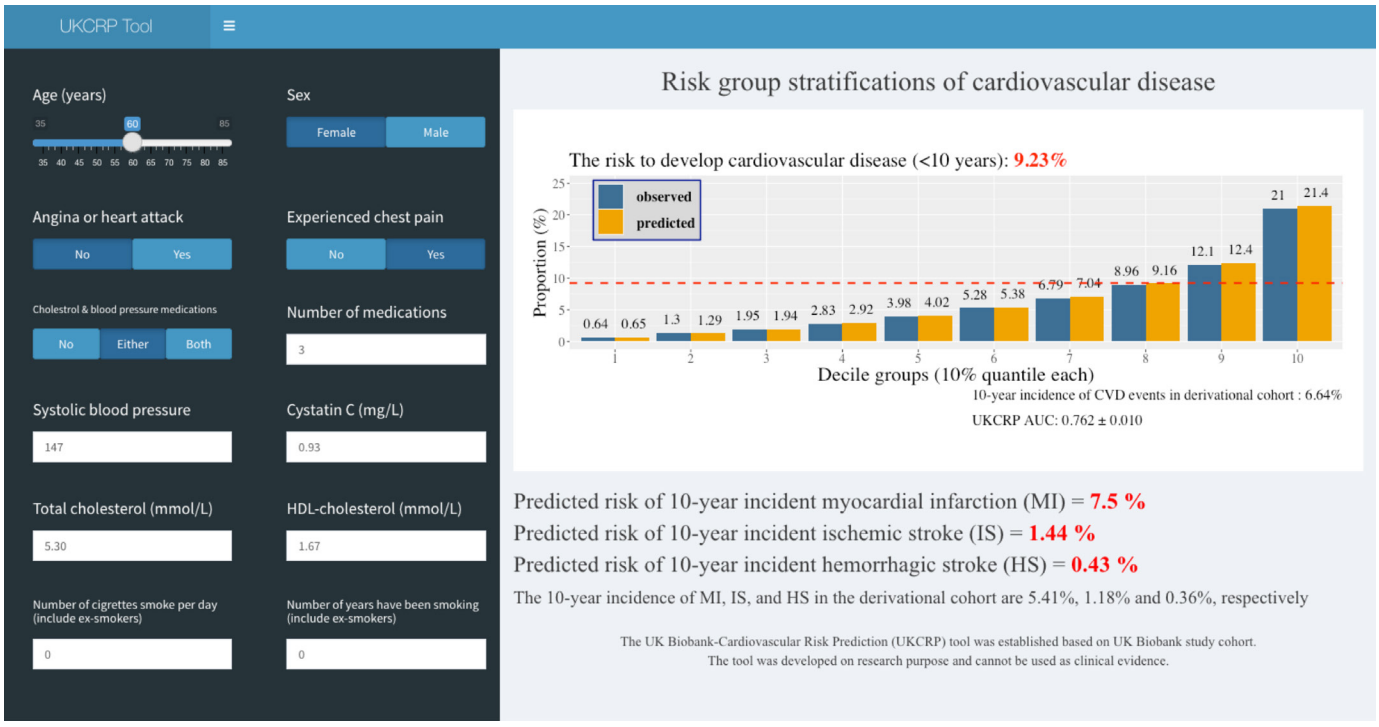

Figure 5 Webpage interface of UKCRP tool. Users can input the baseline characteristics on the left panel, and the calibration plot displayed on the right panel delineates the stratified risk groups of 10-year incident cardiovascular disease based on decile partitions. An example of 60-year-old female participants with no previous angina nor heart attack, experienced chest pain, takes three medications, including a cholesterol treatment, has a systolic blood pressure of 147 mm Hg, cystatin C of 0.93 mg/L, total cholesterol of 5.30 mmol/L, HDL cholesterol of 1.67 mmol/L (cholesterol ratio of 5.30/1.67=3.17) and is a non-smoker, her risk of developing cardiovascular disease in 10 years is predicted as 9.23% and corresponding risks of myocardial infarction, ischaemic stroke, and haemorrhagic stroke are 7.5%, 1.44% and 0.43%, respectively. AUC, area under the receiver operating characteristic curve; CVD, cardiovascular diseases; HDL, high-density lipoprotein.

combined AUC of around 0.7. As previously reported,[7 10] treated hypertension, SBP and ratio of total cholesterol/high-density lipoprotein cholesterol played imperative roles in the prediction of CVD risk. In addition, our model included cholesterol medication. Considering that a subset of the population in the study cohort may have already initiated preventive therapies (eg, statins or antihypertensive medication), the incorporation of drug usage could improve the modelling accuracy.[7] Taking multiple medications, often driven by managing multiple comorbidities, is common in older CVD patients and has been linked to increased risk of CVD outcomes and adverse consequences such as disability, hospitalisation and death.[35–37] Thereby, deprescribing has been an accumulating focus in clinical settings to minimise tangible harm. Ascertaining the predictive value of prior anginal or heart attack and chest pain or discomfort is pivotal, as patients with these symptoms often seek emergency or outpatient assistance and have a greater willingness to engage in proactive risk factor management.[38] Cystatin C is a predictor of cardiovascular risk[39]; however, it has rarely been adopted into previous predictive models. Current smoking status has been frequently reported as a risk factor; in this study, we leveraged pack-years of smoking, which is a derived variable calculated using the number of cigarettes smoked per day and years smoked, and it was found to be more sensitive than simply using a binary variable of smoking status. Overall, the predictors derived in our data-driven pipeline have been validated by numerous studies, proving the reliability of our model; however, it is the first time that the ten predictors were combined to establish a CVD risk prediction model.

The UKCRP model developed in this study can serve as a tool for CVD prediction to evaluate those suspected individuals who may benefit from effective preventive measures. Individuals with a higher CVD risk (eg, a 10-year risk>20%) require more aggressive risk factor interventions. The strategies may include maintaining cholesterol at a reasonable level, intensive blood pressure control, rational and standardised drug use, lowering cystatin C and smoking cessation. Moreover, our study revealed the heightened CVD risk in young adults. Because most young people at high risk tend to ignore potential health hazards, there is a need to raise their self-awareness of the condition and encourage more rigorous interventions or treatments as early as possible to reduce the burden of CVD. Pending external validation, the UKCRP Score is promising not only to help physicians assess CVD risk and make appropriate clinical decisions but also to monitor preventative or therapeutic effectiveness.

One notable strength of our study is that the combination of included predictors was carefully selected from

a comprehensive and massive multidimensional variable space, and the predictors used for modelling were easily accessible and proven to be reliable. The powerful LGBM algorithm we used could perfectly fit the enormous datasets and better deal with the missingness and potential nonlinear interactions compared with traditional Cox regressions. The development of the UKCRP model was underpinned by exceedingly thorough and extensive data of contemporary relevance to European populations, comprising over half a million participants with prolonged follow-up. The above characteristics improve the accuracy, versatility and validity of the model.

Several caveats should be concerned. First, the UK Biobank individuals suffer a lower CVD risk relative to the general primary care population. Prior to widespread implementation, the model needs to be recalibrated using related datasets such as the UK Clinical Practice Research Datalink. Second, despite that the UKCRP model was well calibrated over spatially different recruitment centres, its value in the pragmatic clinical application should be verified in entirely independent prospective cohorts to ensure that such implementation does improve patient outcomes. Third, because the population of the UK Biobank is predominantly white, the generalisability of the model across ancestrally distinct individuals will help to determine whether more appropriate and ethnically relevant decisions are required.

CONCLUSION

In this study, we introduced an ML-based algorithm for CVD risk prediction that outperformed previously established scores. The new model incorporating simple and accessible predictors can be conveniently applied in clinical practice to evaluate CVD risk and guide preventive primary care. Its validity and generalisability need to be tested in future studies.

Author affiliations
[1]Institute of Science and Technology for Brain-Inspired Intelligence, Department of Neurology, Huashan Hospital, State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China
[2]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Ministry of Education, Shanghai, China
[3]Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai, China
[4]School of Data Science, Fudan University, Shanghai, China
[5]Fudan ISTBI—ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Zhejiang, China
[6]Shanghai Medical College and Zhongshan Hospital Immunotherapy Technology Transfer Center, Fudan University, Shanghai, China

Acknowledgements This study used the UK Biobank Resource under application number 19542. We would like to thank all the participants and researchers from the UK Biobank. We thank the Shanghai Center for Brain Science and Brain-Inspired Technology, Zhangjiang Lab, Tianqiao and Chrissy Chen Institute.

Contributors WC, J-TY and J-FF conceived, designed and supervised the project. JY implemented models' development, validation and statistical analysis. YG, J-JK, H-FW and MY supported the analysis and contributed to the discussion of the results. JY and YG drafted the manuscript and verified the underlying data reported. WC, J-TY and J-FF reviewed and edited the manuscript. All authors had full access to all the data in the study and accept the responsibility to submit it for publication.

Funding This study was funded by grants from the National Natural Science Foundation of China (82071997, 82071201), National Key R&D Program of China (2018YFC1312904, 2019YFA0709502), Science and Technology Innovation 2030 Major Projects (2022ZD0211600), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), the 111 Project (B18015), Shanghai Rising-Star Program (21QA1408700), Research Start-up Fund of Huashan Hospital (2022QD002), Excellence 2025 Talent Cultivation Program at Fudan University (3030277001), Shanghai Municipal Health Commission New Interdisciplinary Research Project (2022JC014).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants. All data used in this study were accessed from the publicly available UK Biobank Resource under application number 19542. The UK Biobank, and it has research tissue bank approval from the North West Multicentre Research Ethics Committee and provided oversight for this study. Written informed consent was obtained from all participants. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data used in this study were accessed from the publicly available UK Biobank Resource under application number 19542. These data cannot be shared with other investigators.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD
Jin-Tai Yu http://orcid.org/0000-0002-7079-8041

REFERENCES

1 Writing Group Members, Mozaffarian D, Benjamin EJ, *et al*. Heart disease and stroke statistics-2016 update: a report from the American heart association. *Circulation* 2016;133:e38–360.
2 Mach F, Baigent C, Catapano AL, *et al*. 2019 ESC/EAS guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. *Eur Heart J* 2020;41:111–88.
3 Rossello X, Dorresteijn JA, Janssen A, *et al*. Risk prediction tools in cardiovascular disease prevention: a report from the ESC prevention of CVD programme led by the European association of preventive cardiology (EAPC) in collaboration with the acute cardiovascular care association (ACCA) and the association of cardiovascular nursing and allied professions (ACNAP). *Eur J Prev Cardiol* 2019;26:1534–44.
4 Wilson PW, D'Agostino RB, Levy D, *et al*. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837–47.
5 D'Agostino RB Sr, Vasan RS, Pencina MJ, *et al*. General cardiovascular risk profile for use in primary care. *Circulation* 2008;117:743–53.
6 Conroy RM, Pyörälä K, Fitzgerald AP, *et al*. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the score project. *Eur Heart J* 2003;24:987–1003.
7 SCORE2 working group and ESC Cardiovascular risk collaboration, Hageman S, Pennells L, *et al*. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* 2021;42:2439–54.
8 Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335:136.
9 Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.
10 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.
11 Goff DC, Lloyd-Jones DM, Bennett G, *et al*. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the American College of cardiology/american heart association Task force on practice guidelines. *J Am Coll Cardiol* 2014;63:2935–59.
12 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:e0213653.
13 Pylypchuk R, Wells S, Kerr A, *et al*. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. *The Lancet* 2018;391:1897–907.
14 Dolezalova N, Reed AB, Despotovic A, *et al*. Development of an accessible 10-year digital cardiovascular (dicava) risk assessment: a UK Biobank study. *Eur Heart J Digit Health* 2021;2:528–38.
15 Lees JS, Welsh CE, Celis-Morales CA, *et al*. Glomerular filtration rate by differing measures, albuminuria and prediction of cardiovascular disease, mortality and end-stage kidney disease. *Nat Med* 2019;25:1753–60.
16 Gepner AD, Young R, Delaney JA, *et al*. Comparison of coronary artery calcium presence, carotid plaque presence, and carotid intima-media thickness for cardiovascular disease prediction in the multi-ethnic study of atherosclerosis. *Circ Cardiovasc Imaging* 2015;8:e002262.
17 Welsh CE, Celis-Morales CA, Ho FK, *et al*. Grip strength and walking pace and cardiovascular disease risk prediction in 406,834 UK biobank participants. *Mayo Clinic Proceedings* 2020;95:879–88.
18 Welsh P, Welsh C, Celis-Morales CA, *et al*. Lipoprotein (a) and cardiovascular disease: prediction, attributable risk fraction, and estimating benefits from novel interventions. *Eur J Prev Cardiol* 2022;28:1991–2000.
19 Bundy JD, Rahman M, Matsushita K, *et al*. Risk prediction models for atherosclerotic cardiovascular disease in patients with chronic kidney disease: the CRIC study. *J Am Soc Nephrol* 2022;33:601–11.
20 Cunningham R, Poppe K, Peterson D, *et al*. Prediction of cardiovascular disease risk among people with severe mental illness: a cohort study. *PLoS ONE* 2019;14:e0221521.
21 Quer G, Arnaout R, Henne M, *et al*. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. *J Am Coll Cardiol* 2021;77:300–13.
22 Sudlow C, Gallacher J, Allen N, *et al*. Uk Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779.
23 Ke G*et al*. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017:30.
24 Chakravarti N. Isotonic median regression: a linear programming approach. *Mathematics of OR* 1989;14:303–8.
25 de Leeuw J. Correctness of kruskal's algorithms for monotone regression with ties. *Psychometrika* 1977;42:141–4.
26 Brier GW. Verification of forecasts expressed in terms of probability. *Mon Wea Rev* 1950;78:1–3.
27 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017:30.
28 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
29 Wang C, Yuan Y, Zheng M, *et al*. Association of age of onset of hypertension with cardiovascular diseases and mortality. *Journal of the American College of Cardiology* 2020;75:2921–30.
30 Niiranen TJ, McCabe EL, Larson MG, *et al*. Heritability and risks associated with early onset hypertension: multigenerational, prospective analysis in the Framingham heart study. *BMJ* 2017;357:j1949.
31 Zhao M, Song L, Sun L, *et al*. Associations of type 2 diabetes onset age with cardiovascular disease and mortality: the kailuan study. *Diabetes Care* 2021;44:1426–32.
32 Steinfeldt J, Buergel T, Loock L, *et al*. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. *Lancet Digit Health* 2022;4:e84–94.
33 Khan SS, Page C, Wojdyla DM, *et al*. Predictive utility of a validated polygenic risk score for long-term risk of coronary heart disease in young and middle-aged adults. *Circulation* 2022;146:587–96.
34 Than MP, Pickering JW, Sandoval Y, *et al*. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation* 2019;140:899–909.
35 Malik N, Weintraub WS. The cardiovascular risk of proton pump inhibitors and the dangers of polypharmacy. *Mayo Clin Proc* 2021;96:2511–3.
36 Rea F, Biffi A, Ronco R, *et al*. Cardiovascular outcomes and mortality associated with discontinuing statins in older patients receiving polypharmacy. *JAMA Netw Open* 2021;4:e2113186.
37 Wastesson JW, Morin L, Tan ECK, *et al*. An update on the clinical consequences of polypharmacy in older adults: a narrative review. *Expert Opin Drug Saf* 2018;17:1185–96.
38 Black JA, Sharman JE, Marwick TH. Absolute risk assessment for guiding cardiovascular risk management in a chest pain clinic. *Med J Aust* 2021;215:486:266–71.:.
39 Taglieri N, Koenig W, Kaski JC. Cystatin C and cardiovascular risk. *Clin Chem* 2009;55:1932–43.