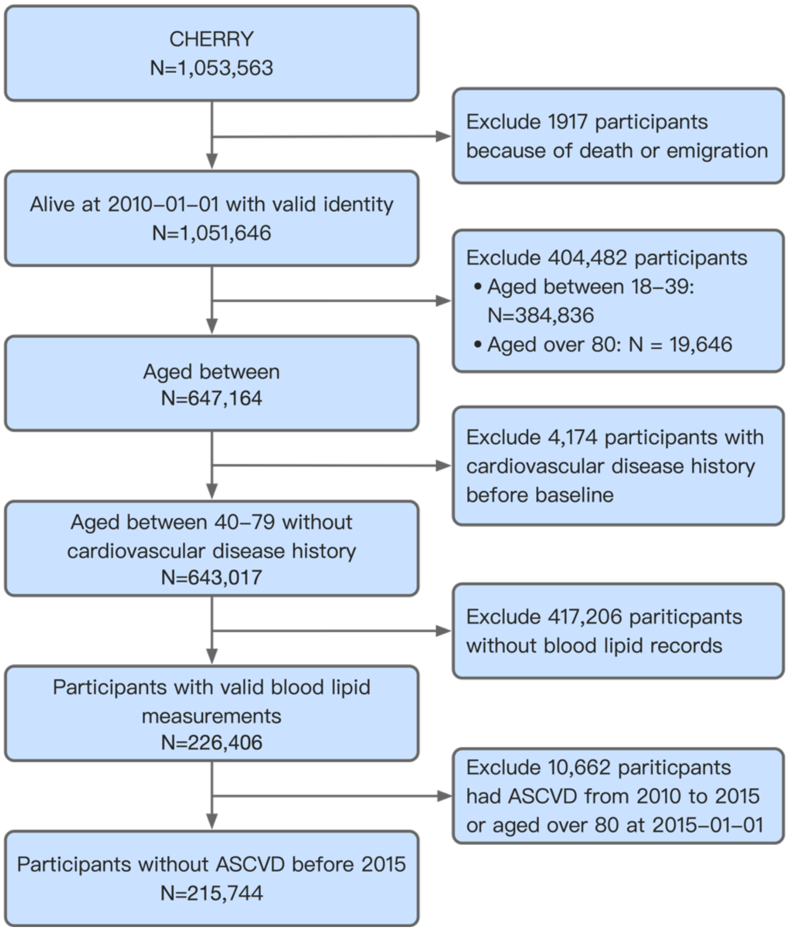
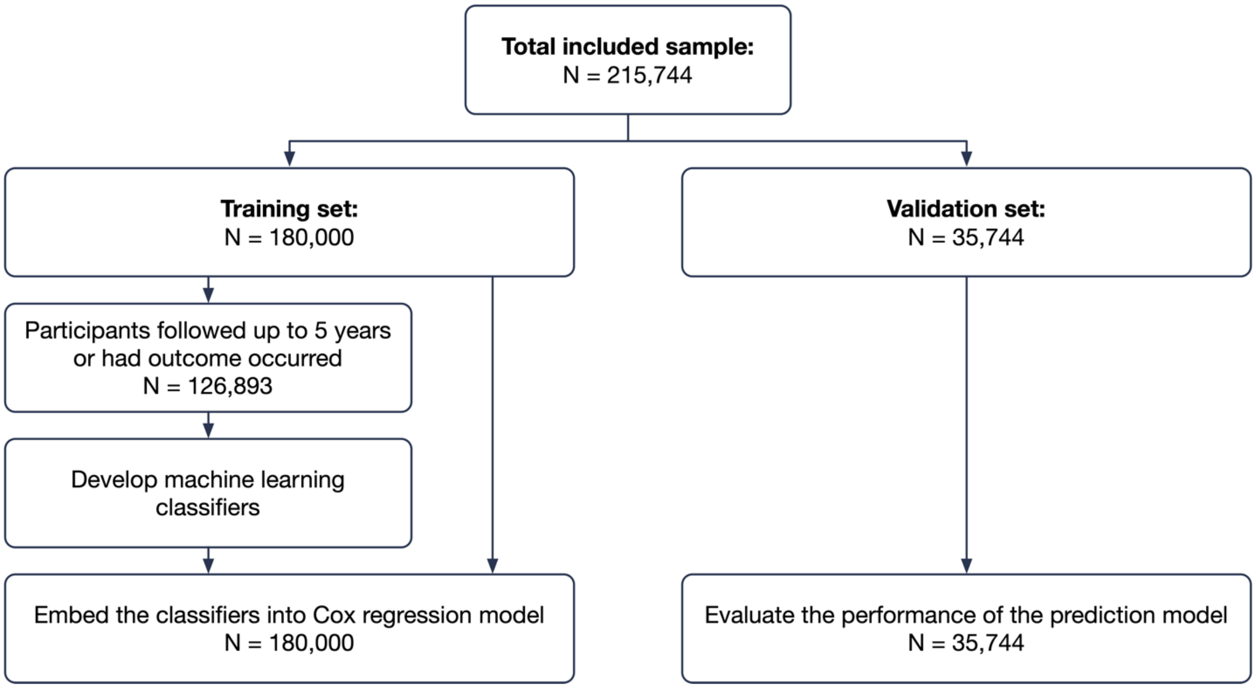
Supplementary Materials

Improving Cardiovascular Risk Prediction through Machine Learning Modelling of Irregular Repeated Electronic Health Records

Supplementary Figure S1 The inclusion and exclusion of eligible participants


CHERRY
N=1,053,563
Exclude 1917 participants because of death or emigration
Alive at 2010–01–01 with valid identity
N=1,051,646
Exclude 404,482 participants
• Aged between 18–39: N=384,836
• Aged over 80: N = 19,646
Aged between
N=647,164
Exclude 4,174 participants with cardiovascular disease history before baseline
Aged between 40–79 without cardiovascular disease history
N=643,017
Exclude 417,206 pariticpants without blood lipid records
Participants with valid blood lipid measurements
N=226,406
Exclude 10,662 pariticpants had ASCVD from 2010 to 2015 or aged over 80 at 2015–01–01
Participants without ASCVD before 2015
N=215,744

Supplementary Figure S2 Flowchart of the study population in training and validation sets


Total included sample:
N = 215,744
Training set:
N = 180,000
Validation set:
N = 35,744
Participants followed up to 5 years or had outcome occurred
N = 126,893
Develop machine learning classifiers
Embed the classifiers into Cox regression model
N = 180,000
Evaluate the performance of the prediction model
N = 35,744

Supplementary Table S1 The descriptions and sources of the predictors included in the study

Predictor	Description	Sources
Settings	Dichotomous, rural or urban	Census data and health insurance databases
Smoke status	Dichotomous, current or not current	Self-reported from GP's survey
Family history of ASCVD	Dichotomous, with or without	Self-reported from GP's survey
Education level	Multiple levels: primary school of lower, middle school, college or above	Census data and health insurance databases
Baseline diabetes	Dichotomous, with or without, defined as E10-14 of ICD-10 code	EMR, health check, disease surveillance, and chronic disease management system databases
Anthropometrics (blood pressure, BMI, and waist circumference)	Continuous, units: mmHg (blood pressure); kg/m^2 (BMI); waist circumference (cm)	Chronic disease management system and health check databases
Laboratory tests (serum lipids, glucose, urinary albumin, and blood creatinine)	Continuous, units: mmol/L (TC, TG, HDL-C, LDL-C, FBG), mg/dL (Apo-a, Apo-b, Lp-(a), urinary albumin), % (HbA1c), μmol/L (blood creatinine); eGFR was calculated according to the CKD-EPI equation	In-patients' EMR, health check, and chronic disease management system databases
Medication history	Dichotomous, with or without	In-patients' EMR, health check, and chronic disease management system databases

Supplementary Table S2 The definition of medication history

Treatments	Definitions
Anti-hypertension	Ever used the following medications before baseline: angiotensin converting enzyme (ACE) inhibitors, beta-blockers, thiazide, angiotensin II receptor blockers (ARB), calcium channel blockers, and alpha-blockers.
Lipid-lowering	Ever used the following medications before baseline: statins, nicotinic acid, cholesterol absorption inhibitors, probucol, cholic acid chelating agent, and fibrates.
Anti-hyperglycemia	Ever used the following medications before baseline: biguanides, sulfonylureas, non-sulfonylurea derivatives of anisic acid, alpha-glucosidase inhibitors, thiazolidinediones, glucagon-like peptide 1 (GLP-1) receptor agonist, dipeptidyl peptidase IV (DPP-4) inhibitors, sodium-glucose transporter 2 (SGLT2) inhibitors, and insulin.
Aspirin	Ever used Aspirin before baseline.

Supplementary Table S3 Normal ranges of predictors

Predictor	Unit	Normal ranges
HDL-C	mmol/L	(0,10], 0<var≤10
LDL-C	mmol/L	(0,20]
TC	mmol/L	(0,20]
TG	mmol/L	(0,20]
HbA_{1c}	%	[4,15]
Height	cm	[80,250]
Weight	kg	[10,300]
BMI	kg/m2	[10,80]
Systolic pressure	mmHg	[70,270]
Dialystic pressure	mmHg	[30,150]
Waist circumference	cm	[50,130]

Supplementary Table S4 Predictors inputted in the two machine learning models

	China-PAR	Baseline	Repeated measurements-based Variables				
			Number of measurements	Mean	Standard deviation	Range	Difference between first and last measurements
Demography							
Sex	√	√					
Age	√	√					
Smoking status	√	√					
Education level		√					
Settings	√	√					
Family history of ASCVD	√	√					
Blood pressure							
SBP	√	√	√	√	√	√	√
DBP		√	√	√	√	√	√
Obesity							
BMI		√	√	√	√	√	√
Waist circumference	√	√	√	√	√	√	√
Lipid metabolism							
TC	√	√	√	√	√	√	√
TG		√	√	√	√	√	√
HDL-C	√	√	√	√	√	√	√
LDL-C		√	√	√	√	√	√
Apo-a		√	√	√	√	√	√
Apo-b		√	√	√	√	√	√
Lp-a		√	√	√	√	√	√
Glucose metabolism							
FBG		√	√	√	√	√	√
Diabetes	√	√					
HbA1c		√	√	√	√	√	√
Renal function							
eGFR		√	√	√	√	√	√
ACR		√	√	√	√	√	√
Medication							
Anti-hypertension	√	√					
Anti-hyperlipidemia		√					
Anti-hyperglycemia		√					
Aspirin		√					

Supplementary Table S5 The selection ranges of hyperparameters in the two machine learning models

Hyperparameters	**Range**
XGBoost	
Maximum tree depth	6, 7, 8, 9, 10
Learning rate	[0.01, 0.3]
γ	(0.0, 0.2]
Subsample proportion	[0.6, 0.9]
Subspace proportion	[0.5, 0.8]
Minimum children nodes weight	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
LASSO	
Lambda (Grid search for 500 values within the range)	[6.47×10^{-5}, 3.84×10^{-2}]

Supplementary Table S6 Characteristics of the training and validation sets[a]

	Overall (N = 215,744)	Training (n = 180,000)	Validation (n = 35,744)	Statistic (*P*)
Demographic attributes				
Female	115,666 (53.61)	96,568 (53.65)	19,098 (53.43)	0.57 (0.452)
Age, y	56.70 (9.59)	56.71 (9.60)	56.65 (9.55)	-1.02 (0.307)
Rural	65,086 (30.34)	54,417 (30.41)	10,669 (30.01)	2.16 (0.141)
Current smokers	57,961 (26.87)	48,366 (26.87)	9,595 (26.84)	0.01 (0.923)
Finished High school	108,120 (50.11)	90,207 (50.11)	17,913 (50.11)	0.00 (0.999)
Family history of ASCVD	1,318 (0.61)	1,125 (0.62)	193 (0.54)	399.41 (0.065)
Body measurements				
Waist circumference, cm	81.76 (7.94)	81.77 (7.95)	81.69 (7.85)	-1.59 (0.111)
BMI, kg/m^2	23.31 (2.87)	23.31 (2.88)	23.32 (2.85)	0.21 (0.834)
Blood pressure				
SBP, mmHg	134.45 (16.64)	134.45 (16.64)	134.45 (16.62)	-0.03 (0.975)
DBP, mmHg	82.63 (9.87)	82.62 (9.88)	82.65 (9.77)	0.43 (0.670)
Lipid profiles				
Total cholesterol, mmol/L	4.90 (0.98)	4.90 (0.98)	4.91 (0.99)	**2.06 (0.039)**
HDL-C, mmol/L	1.30 (0.34)	1.30 (0.34)	1.30 (0.34)	0.87 (0.385)
TG, mmol/L	1.61 (1.09)	1.61 (1.09)	1.61 (1.09)	-0.01 (0.994)
LDL-C, mmol/L	2.84 (0.82)	2.84 (0.82)	2.84 (0.83)	0.88 (0.378)
Apo-A (mmol/L)	1.22 (0.27)	1.22 (0.27)	1.23 (0.27)	1.09 (0.274)
Apo-B (mmol/L)	0.95 (0.25)	0.95 (0.25)	0.95 (0.25)	1.43 (0.154)
Lp-(a) (mg/dL)	174.17 (184.38)	173.85 (184.23)	175.76 (185.14)	1.07 (0.283)
Glycemia				
FBG, mmol/L	5.67 (1.57)	5.67 (1.56)	5.68 (1.64)	0.82 (0.414)
HbA1c, %	6.86 (1.90)	6.86 (1.91)	6.85 (1.89)	-0.37 (0.712)
Diabetes mellitus	26,090 (12.09)	21,666 (12.04)	4,424 (12.38)	3.22 (0.073)
Renal function				
ACR, mg/g	15.90 (45.36)	15.79 (45.99)	16.42 (42.07)	0.46 (0.642)
eGFR, mL/min/1.73m^2	98.92 (15.30)	98.94 (15.29)	98.86 (15.38)	-0.69 (0.488)

Supplementary Table S6 Characteristics of the training and validation sets (continued)[a]

	Overall (N = 215,744)	Training (n = 180,000)	Validation (n = 35,744)	Statistic (*P*)
Medications				
Anti-hypertension treatment	75,857 (35.16)	63,299 (35.17)	12,558 (35.13)	0.01 (0.910)
Anti-hyperlipidemia treatment	35,561 (16.48)	18,960 (10.53)	3,887 (10.87)	3.63 (0.057)
Anti-hyperglycemia treatment	22,847 (10.59)	29,514 (16.40)	6,047 (16.92)	**5.84 (0.016)**
Aspirin treatment	19,064 (8.84)	15,896 (8.83)	3,168 (8.86)	0.03 (0.854)
Outcomes				
ASCVD events	6,081 (2.82)	5,112 (2.84)	969 (2.71)	1.77 (0.184)
Average follow-up time, years	5.41 (1.36)	5.41 (1.36)	5.41 (1.36)	-0.18 (0.861)
Incidence rate of ASCVD, per 10^6 person-years	6,178 (6177-6179)	6,225 (6224-6226)	5,944 (5943-5945)	1.36 (0.174)
Kaplan-Meier survival	0.969 (0.968-0.970)	0.969 (0.968-0.970)	0.970 (0.968-0.972)	1.74 (0.187)

[a] Categorical variables were presented by counts and percentages, using the Chi-square test to compare differences; Continuous variables were presented by means and standard deviations, using Student’s t-test to compare the difference. The difference in survival was given by the log-rank test.

Supplementary Table S7 The missing proportions of predictors[a]

	Baseline	Repeated measurements-based Variables				
		Number of measurements	Mean	Standard deviation	Range	Difference between first and last measurements
Demography						
Sex	0					
Age	0					
Smoking status	0					
Education level	6.5%					
Settings	0.6%					
Family history of ASCVD	0					
Blood pressure						
SBP	31.0%	42.4%	42.4%	62.3%	42.4%	42.4%
DBP	31.0%	42.4%	42.4%	62.3%	42.4%	42.4%
Obesity						
BMI	3.8%	35.9%	35.9%	70.0%	35.9%	35.9%
Waist circumference	14.6%	39.3%	39.3%	58.9%	39.3%	39.3%
Lipid metabolism						
TC	4.9%	4.9%	4.9%	31.3%	4.9%	4.9%
TG	4.5%	4.5%	4.5%	31.3%	4.5%	4.5%
HDL-C	4.6%	4.6%	4.6%	32.1%	4.6%	4.6%
LDL-C	5.1%	5.3%	5.3%	33.3%	5.3%	5.3%
Apo-A	50.1%	50.1%	50.1%	74.8%	50.1%	50.1%
Apo-B	50.1%	50.1%	50.1%	74.8%	50.1%	50.1%
Lp-(a)	63.7%	63.7%	63.7%	83.4%	63.7%	63.7%
Glucose metabolism						
FBG	13.8%	14.0%	14.0%	36.2%	14.0%	14.0%
HbA1c	89.3%	89.4%	89.4%	94.8%	89.4%	89.4%
Diabetes	0					
Kidney function related						
eGFR	33.7%	33.8%	33.8%	60.9%	33.8%	33.8%
ACR	96.6%	96.2%	96.6%	98.9%	96.6%	96.6%
Medication						
Anti-hypertension	0					
Anti-hyperlipidemia	0					
Anti-hyperglycemia	0					
Aspirin	0					

[a] All the missing proportions were calculated in the whole study population of 215,744.

Supplementary Table S8 The median time intervals between measurements of key predictors

Predictor	Median number (IQR) of measurements from each individual[a]	Median (IQR) of the mean time interval between each measurement of each individual[b]	Maximum (IQR) of the mean time interval between each measurement of each individual[c]
Total cholesterol	3 (4)	269 days (337)	559 days (418)
Systolic blood pressure	2 (2)	136 days (384)	320 days (694)
Body mass index	1 (1)	267 days (525)	671 days (785)
Fasting blood glucose	3 (5)	251 days (371)	536 days (458)

[a] The median of the number of each individual's measurement.

[b] Mean time intervals of each individual's measurements were calculated first and the medians of each individual's mean were given above.

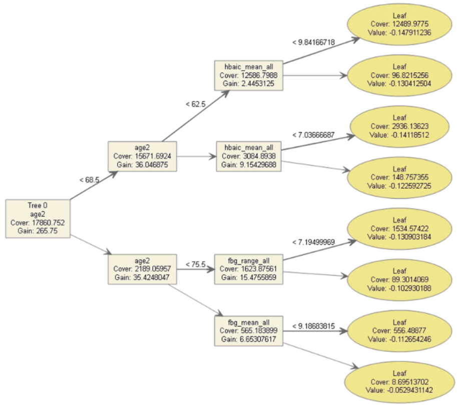
[c] Mean time intervals of each individual's measurements were calculated first and the maximums of each individual's mean were given above.

Supplementary Table S9 Hyperparameters of XGBoost models and LASSO regression model[a]

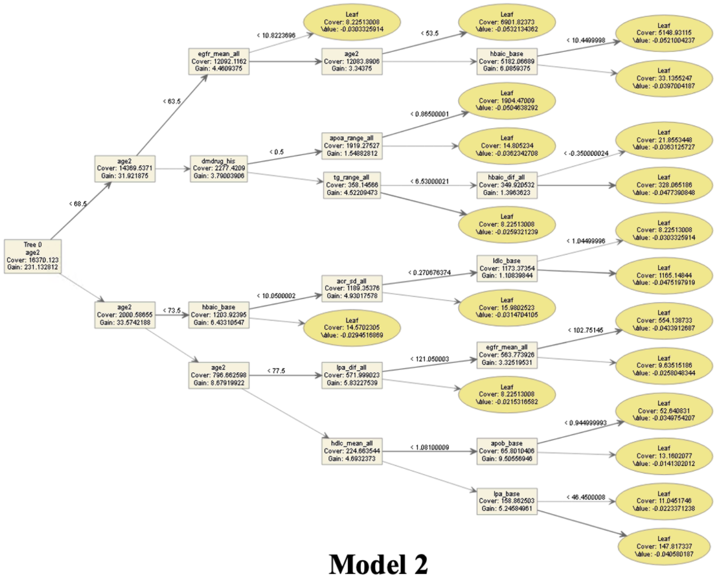
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
XGBoost					
Maximum tree depth	3	5	5	3	3
Learning rate	0.0568	0.0203	0.0215	0.0971	0.0689
γ	0.191	0.138	0.0308	0.139	0.0336
Subsample proportion	76.2%	69.6%	66.1%	73.4%	63.7%
Subspace proportion	74.5%	58.4%	59.1%	63.0%	56.8%
Minimum children nodes weight	8	8	6	8	3
LASSO					
Lambda	0.000638	0.000650	0.000650	0.000712	0.000662

[a] The five models were generated according to the five imputation subsets for predictors in the China-PAR model and repeated measurements derived predictors without being imputed.

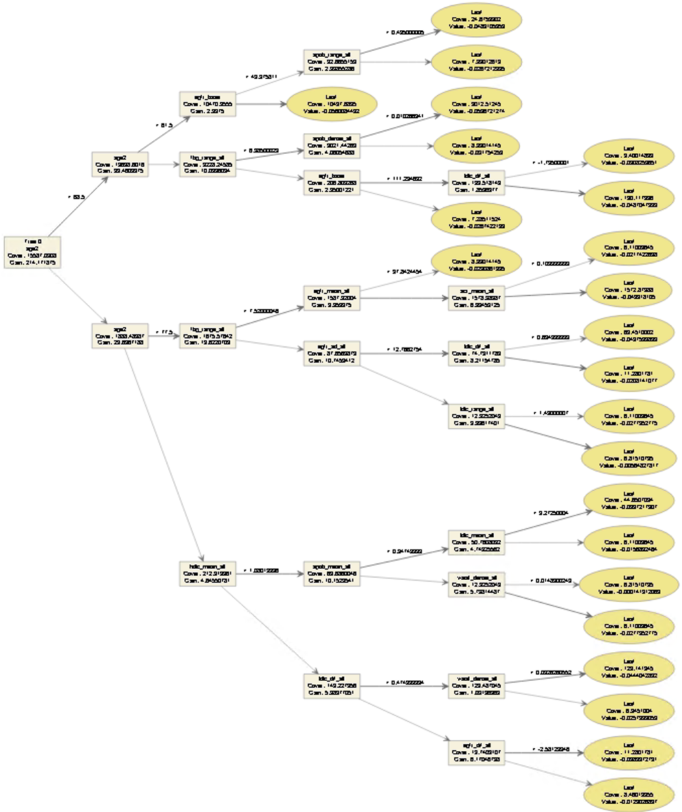
Supplementary Figure S3 The structures of the first tree in each XGBoost model



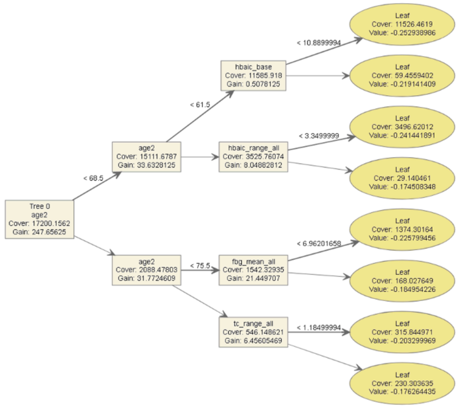
Model 1



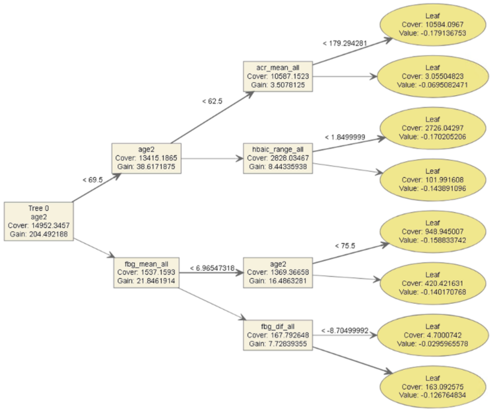
Model 2



Model 3



Model 4



Model 5

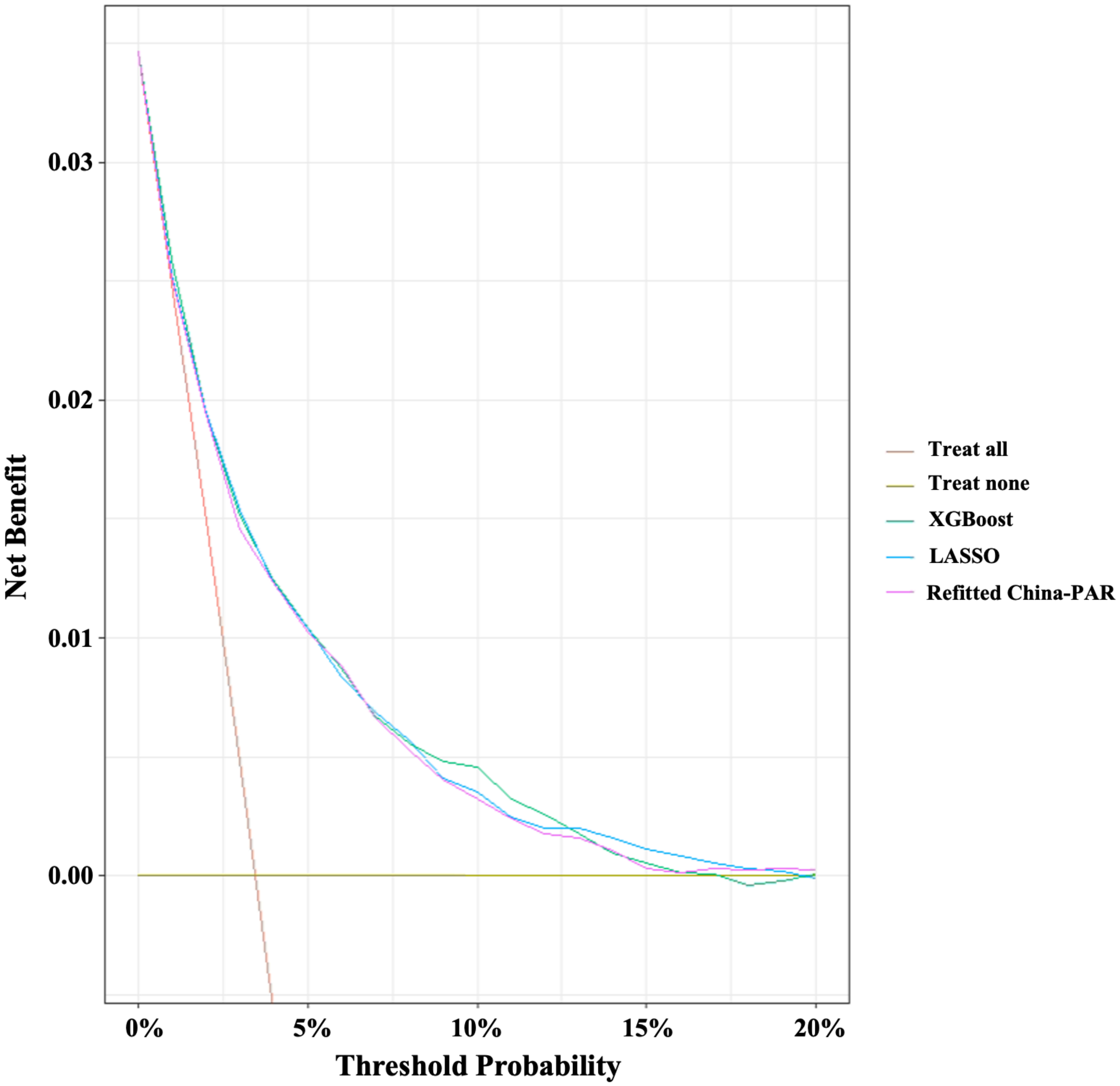
Supplementary Table S10 Fifteen β coefficients with largest absolute value in each LASSO regressions

Predictor	β coefficients	Odds ratio
Model 1		
Age	0.9861	2.681
Anti-hypertension treatment	0.1713	1.187
Female	-0.0830	0.920
Family history of ASCVD	0.0708	1.073
Baseline diabetes	0.0708	1.073
Fifth quintile of mean fasting blood glucose	0.0617	1.064
Aspirin treatment history	0.0593	1.061
Missing of education records	-0.0526	0.949
Baseline systolic blood pressure	0.0457	1.047
Fifth quintile of the standard deviation of fasting blood glucose	0.0398	1.041
Baseline HDLC	-0.0330	0.968
Second quintile of baseline LDLC	-0.0322	0.968
Fifth quintile of mean triglycerides	0.0287	1.029
Fourth quintile of mean fasting blood glucose	0.0278	1.028
Third quintile of baseline eGFR	-0.0276	0.973
Model 2		
Age	0.9854	2.679
Anti-hypertension treatment	0.1708	1.186
Female	-0.0827	0.921
Family history of ASCVD	0.0707	1.073
Baseline diabetes	0.0704	1.073
Fifth quintile of mean fasting blood glucose	0.0617	1.064
Aspirin treatment history	0.0592	1.061
Missing of education records	-0.0518	0.950
Baseline systolic blood pressure	0.0454	1.046
Fifth quintile of the standard deviation of fasting blood glucose	0.0401	1.041
Baseline HDLC	-0.0326	0.968
Second quintile of baseline LDLC	-0.0322	0.968
Fifth quintile of mean triglycerides	0.0286	1.029
Fourth quintile of mean fasting blood glucose	0.0275	1.028
Fifth quintile of mean ApoB	0.0274	1.028
Model 3		
Age	0.9854	2.679
Anti-hypertension treatment	0.1708	1.186
Female	-0.0827	0.921
Family history of ASCVD	0.0707	1.073
Baseline diabetes	0.0704	1.073
Fifth quintile of mean fasting blood glucose	0.0617	1.064
Aspirin treatment history	0.0592	1.061
Missing of education records	-0.0518	0.950

Supplementary Table S10 (Continued) β coefficients of the LASSO regression

Predictor	β coefficients	Odds ratio
Baseline systolic blood pressure	0.0454	1.046
Fifth quintile of the standard deviation of fasting blood glucose	0.0401	1.041
Baseline HDLC	-0.0326	0.968
Second quintile of baseline LDLC	-0.0322	0.968
Fifth quintile of mean triglycerides	0.0286	1.029
Fourth quintile of mean fasting blood glucose	0.0275	1.028
Fifth quintile of mean ApoB	0.0274	1.028
Model 4		
Age	0.9817	2.669
Anti-hypertension treatment	0.1681	1.183
Female	-0.0813	0.922
Family history of ASCVD	0.0699	1.072
Baseline diabetes	0.0698	1.072
Fifth quintile of mean fasting blood glucose	0.0617	1.064
Aspirin treatment history	0.0583	1.060
Missing of education records	-0.0472	0.954
Baseline systolic blood pressure	0.0440	1.045
Fifth quintile of the standard deviation of fasting blood glucose	0.0413	1.042
Second quintile of baseline LDLC	-0.0322	0.968
Baseline HDLC	-0.0306	0.970
Fifth quintile of mean triglycerides	0.0280	1.028
Fifth quintile of mean ApoB	0.0269	1.027
Fourth quintile of mean fasting blood glucose	0.0264	1.027
Model 5		
Age	0.9847	2.677
Anti-hypertension treatment	0.1703	1.186
Female	-0.0825	0.921
Family history of ASCVD	0.0705	1.073
Baseline diabetes	0.0702	1.073
Fifth quintile of mean fasting blood glucose	0.0617	1.064
Aspirin treatment history	0.0590	1.061
Missing of education records	-0.0509	0.950
Baseline systolic blood pressure	0.0452	1.046
Fifth quintile of the standard deviation of fasting blood glucose	0.0403	1.041
Second quintile of baseline LDLC	-0.0322	0.968
Baseline HDLC	-0.0322	0.968
Fifth quintile of mean triglycerides	0.0285	1.029
Fourth quintile of mean fasting blood glucose	0.0273	1.028
Fifth quintile of mean ApoB	0.0273	1.028

Supplementary Figure S4 Decision curve analysis of the ML models



Supplementary Table S11 Associations between predictors and ASCVDa

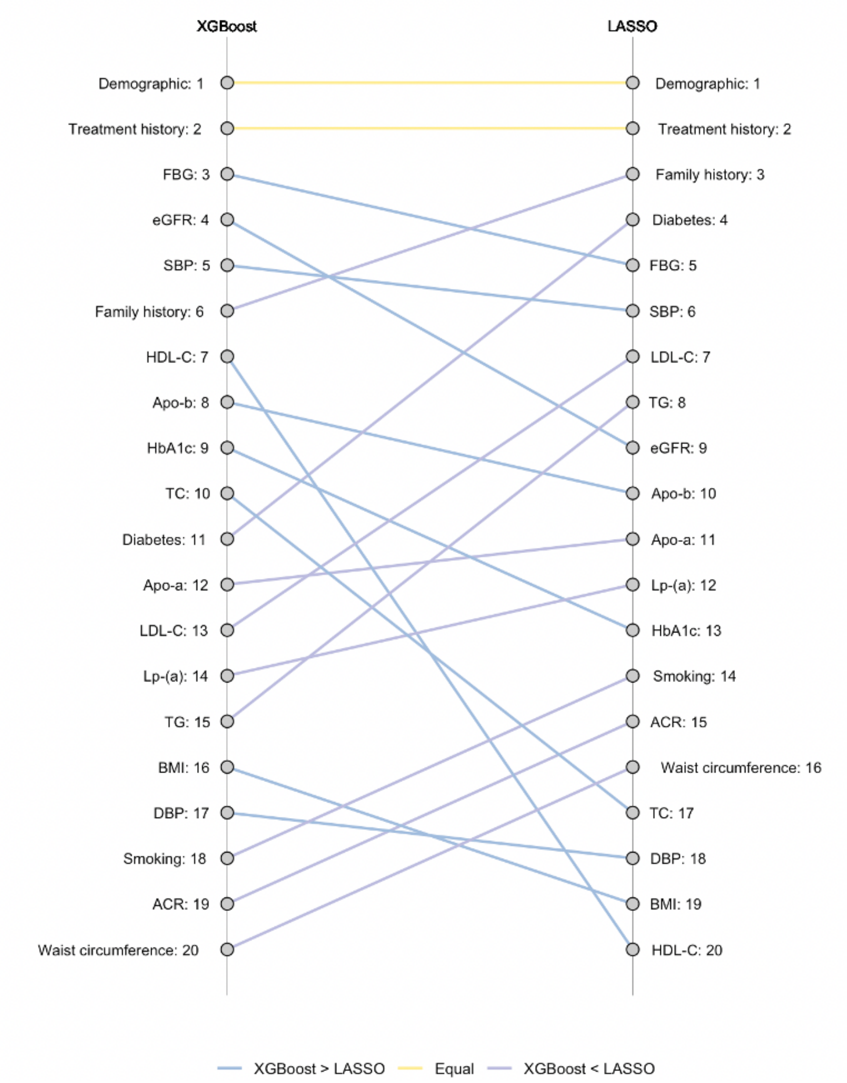
	Baseline	Repeated measurements-based Variables				
		Number of measurements	Mean	Standard deviation	Range	Difference between first and last measurements
China-PAR predictors						
Women	0.81 (0.77-0.86)					
Age	1.11 (1.11-1.11)					
Current smoker	1.12 (1.05-1.19)					
Diabetes	1.68 (1.58-1.78)					
Urban	0.93 (0.88-0.98)					
Family history of ASCVD	2.69 (2.27-3.19)					
Waist circumference	1.00 (1.00-1.00)	1.11 (1.04-1.19)	0.90 (0.79-1.02)	1.21 (1.11-1.31)	1.15 (1.07-1.24)	1.20 (1.02-1.42)
SBP	1.00 (1.00-1.01)	1.23 (1.15-1.30)	1.32 (1.20-1.46)	1.33 (1.21-1.47)	1.26 (1.17-1.36)	0.76 (0.49-1.19)
TC	1.08 (1.05-1.11)	1.43 (0.97-2.11)	1.19 (1.02-1.39)	1.31 (1.20-1.43)	1.26 (1.15-1.37)	0.73 (0.58-0.90)
HDL-C	0.75 (0.69-0.81)	1.10 (1.03-1.17)	0.76 (0.69-0.83)	1.14 (1.04-1.25)	1.23 (1.11-1.36)	0.81 (0.62-0.83)
Hypertension	1.39 (1.31-1.47)					
Other predictors						
Demography						
Education level	0.67 (0.57-0.78)					
Blood pressure						
DBP	1.76 (1.02-3.05)	1.23 (1.16-1.31)	1.31 (1.19-1.46)	1.21 (1.10-1.29)	1.19 (1.10-1.29)	2.30 (0.74-7.13)
Obesity						
BMI	0.89 (0.79-0.99)	1.17 (1.10-1.25)	1.14 (1.05-1.24)	1.21 (1.11-1.32)	1.17 (1.09-1.25)	1.35 (1.05-1.75)
Lipid metabolism						
TG	1.95 (1.15-3.32)	1.09 (1.02-1.16)	1.18 (1.05-1.33)	1.10 (1.01-1.19)	1.20 (1.07-1.34)	0.70 (0.55, 0.89)
LDL-C	0.79 (0.72-0.87)	1.11 (1.04-1.18)	0.86 (0.79-0.94)	1.31 (1.20-1.43)	1.94 (1.52-2.48)	0.84 (0.72-0.98)
Apo-A	0.77 (0.60-1.00)	0.85 (0.80-0.90)	0.67 (0.60-0.74)	1.22 (1.06-1.40)	1.47 (1.17-1.84)	0.79 (0.56-1.12)
Apo-B	0.78 (0.66-0.92)	0.85 (0.80-0.90)	0.80 (0.72-0.88)	1.30 (1.13-1.50)	1.24 (1.09-1.41)	0.79 (0.57-1.09)
Lp-(a)	0.80 (0.75-0.86)	0.82 (0.77-0.87)	0.84 (0.77-0.92)	1.32 (1.11-1.57)	0.84 (0.75-0.94)	0.72 (0.61-0.85)

Supplementary Table S11 Associations between predictors and ASCVD (continued)[a]

	Baseline	Repeated measurements-based Variables				
		Number of measurements	Mean	Standard deviation	Range	Difference between first and last measurements
Glucose metabolism						
FBG	3.46 (2.60-4.59)	1.06 (0.99-1.13)	1.89 (1.69-2.10)	1.72 (1.55-1.90)	2.02 (1.75-2.34)	0.48 (0.40-0.58)
HbA1c	2.43 (1.97-3.00)	0.70 (0.63-0.78)	1.98 (1.65-2.38)	1.47 (1.14-1.89)	1.66 (1.23-2.24)	0.43 (0.29-0.66)
Renal function						
eGFR	0.66 (0.60-0.72)	5.24 (2.72-10.08)	0.82 (0.77-0.88)	1.37 (1.22-1.55)	1.52 (1.35-1.72)	0.74 (0.67-0.81)
ACR	2.62 (1.93-3.54)	0.68 (0.58-0.81)	2.28 (1.72-3.03)	2.75 (1.66-4.55)	2.20 (1.31-3.69)	0.61 (0.41-0.91)
Medication						
Hyperlipidemia	1.03 (0.97-1.10)					
Hyperglycemia	0.91 (0.79-1.06)					
Aspirin	1.26 (1.18-1.34)					

[a] The association between predictors in the China-PAR model was multi-variable adjusted for each other. The other predictors were individually adjusted for predictors in the China-PAR model. All the associations were estimated in the whole study population of 215,744.

Supplementary Figure S5 The importance of predictors in the two ML models



(a)

No.	XGBoost	LASSO
1	Age	Age
2	Anti-hypertension treatment	Anti-hypertension treatment
3	Mean of eGFR	Sex
4	Mean of FBG	Family history of ASCVD
5	Standard deviation of FBG	Diabetes at baseline
6	Baseline SBP	Mean of FBG
7	Family history of ASCVD	Aspirin use
8	Mean of HDL-C	Education level
9	Mean of HbA1c	Baseline SBP
10	Mean of Apo-b	Standard deviation of FBG

(b)

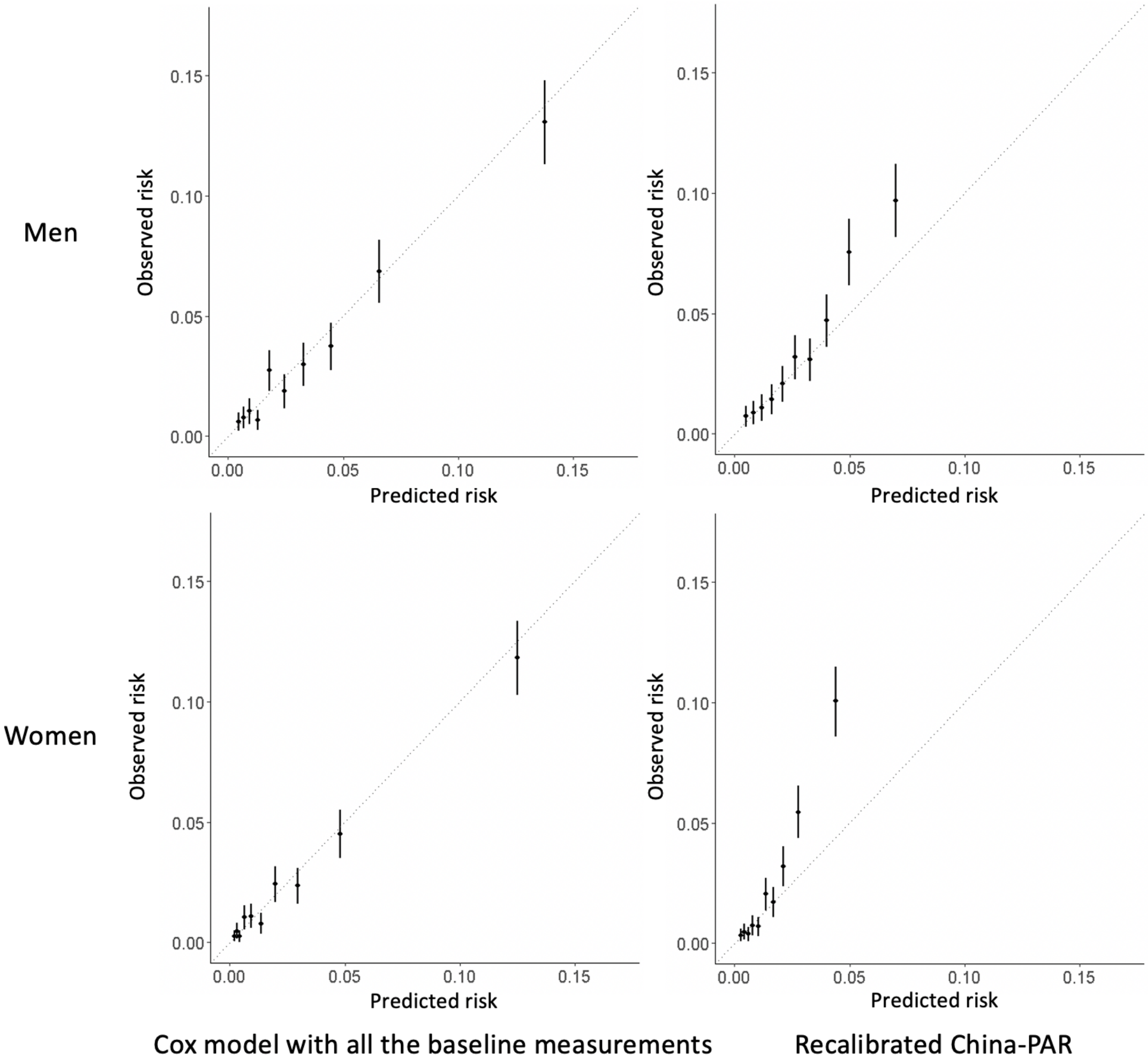
(a) The minimum rank of importance in each kind of measurement by models. (b) Top 10 irrelevant important predictors in the two ML models. Predictors with a correlation coefficient larger than 0.7 were not counted.

Supplementary Table S12 The differences of C statistics between ML models and Cox model with all the baseline measurements[a]

Sex	Model	C statistics (95% CI)	Difference in C statistics	*P*
Overall				
	Cox model with all baseline measurements	0.7861 (0.7718, 0.8007)	Reference	
	LASSO regression	0.7883 (0.7737, 0.8029)	0.00214 (-0.00088, 0.00515)	0.1654
	XGBoost model	0.7918 (0.7776, 0.8060)	**0.00563 (0.00118, 0.01009)**	**0.0133**
Men				
	Cox model with all baseline measurements	0.7614 (0.7408, 0.7820)	Reference	
	LASSO regression	0.7623 (0.7415, 0.7831)	0.00091 (-0.00340, 0.00522)	0.6788
	XGBoost model	0.7700 (0.7502, 0.7898)	**0.00860 (0.00183, 0.01536)**	**0.0128**
Women				
	Cox model with all baseline measurements	0.8040 (0.7829, 0.8251)	Reference	
	LASSO regression	0.8077 (0.7866, 0.8287)	0.00365 (-0.00057, 0.00788)	0.0904
	XGBoost model	0.8071 (0.7861, 0.8281)	0.00309 (-0.00286, 0.00903)	0.3081

[a] The results were given based on the validation set of 31,544.

Supplementary Figure S6 Calibration plots of models in sensitivity analysis[a]



[a]The results were given based on the validation set of 31,544.

Supplementary Table S13 The difference in C statistic between ML models and recalibrated China-PAR model[a]

Sex	**Model**	**C statistics (95% CI)**	**Difference in C statistics**	***P***
Overall				
	Recalibrated China-PAR model	0.7513 (0.7369, 0.7657)	Reference	
	LASSO regression	0.7883 (0.7737, 0.8029)	**0.03670 (0.02906, 0.04487)**	**<0.0001**
	XGBoost model	0.7918 (0.7776, 0.8060)	**0.04047 (0.02605, 0.05488)**	**<0.0001**
Men				
	Recalibrated China-PAR model	0.7226 (0.7017, 0.7434)	Reference	
	LASSO regression	0.7623 (0.7415, 0.7831)	**0.03975 (0.02815, 0.05135)**	**<0.0001**
	XGBoost model	0.7700 (0.7502, 0.7898)	**0.04744 (0.03512, 0.05976)**	**<0.0001**
Women				
	Recalibrated China-PAR model	0.7836 (0.7629, 0.8044)	Reference	
	LASSO regression	0.8077 (0.7866, 0.8287)	**0.02402 (0.01459, 0.03345)**	**<0.0001**
	XGBoost model	0.8071 (0.7861, 0.8281)	**0.02345 (0.01300, 0.03392)**	**<0.0001**

[a]The results were given based on the validation set of 31,544.

Supplementary Appendix The TRIPODS checklist for this study[a]

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	P1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	P3
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	P5L13-P6L3
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	P6L5-9
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	P6L13-23
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	P6L25-P7L3
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	P6L13-16
	5b	D;V	Describe eligibility criteria for participants.	P6L16-20
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	P8L3-9
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	P7L9-28
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	NA
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Supplementary Method
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	P8L19-23
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	P8L19-P9L1
	10c	V	For validation, describe how the predictions were calculated.	NA
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	P9L13-22
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	P9L21-22
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	Supplementary Table 6

Supplementary Appendix The TRIPODS checklist for this study (continued)[a]

Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Figure 2, Table 1
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	P10L14-21
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Table 1
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	Table 1
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	P11L26-27, Supplementary Table 10
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	P10L24-P11L23
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	P15L2-11
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	P13L4-P14L5, P14L19-28
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	P15L13-18
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	P14L13-17
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	NA
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	P15L22-23

[a]Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

Supplementary Method S1 Detailed description of the CHERRY study

The source data were from the CHinese Electronic health Records Research in Yinzhou (CHERRY) study, a longitudinal and population-based cohort study in China. The detailed study protocol and some findings on CVD have been previously published.[1] In short, the CHERRY study was established based on the integrated Health Information System in Yinzhou, a developed area in Eastern China. The system consists of various databases, e.g., Population Census and Registered Health Insurance Database, Health Check Database, Disease Management Database, Death and Disease Surveillance Database, and Electronic Medical Records (EMRs). Individual information about cardiovascular risk factors, clinical measurements, and outcomes from different databases were linked via a unique and encoded identifier.

Data collection on the Longitudinal measurement of cardiovascular risk factors

Regarding the traditional CVD risk factors, local GPs in Yinzhou have built up an impressive scheme on frequent health checks among adults and regular epidemiological surveys as part of primary care routine services over the 10 years after China's healthcare reform was initially launched. CHERRY then includes longitudinal measurements of risk factors related to CVD at the individual level, e.g., smoking status, alcohol use, body mass index (BMI) and other obesity risk factors, and daily physical activity. Detailed description of longitudinal measurements was published in the original study protocol. We provided the information regarding the CVD risk prediction in this study as follows:

Predictors	**Measurement Methods**
Age at baseline in years, continuous	Patients' ID numbers were derived from population census and registered health insurance database and their date of birth was then identified (which is recorded in this number as eight digits)
Education level, categorical	Education level was acquired from population census and registered health insurance database.
Body mass index in kg/m^2, continuous	BMI was calculated as weight(kg)/height(m)^2. Weight and height recorded on the same measurement date and in the same database were used in calculation. Information was mainly from the routine epidemiological survey, health checks, and disease surveillance and management system.
Blood pressure in mmHg, continuous	This was measured by either a general practitioner (population census and registered health insurance database, disease manage database) or practice nurses (health checks database).

Smoking status, categorical	Smoking status of patients were recorded in several databases, mainly including health checks database and population census and registered health insurance database. They were recorded as categorical variables (non-smoker, ex-smoker, current smoker) or continuous variables (number of cigarettes smoked per day) due to different design forms. These were combined into two categories indicating ever smoked or not.
Glycemia, continuous	The HbA$_{1c}$ (%) and FBG (mmol/l) were recorded in disease management database, health checks database and inpatient EMR.
Lipid profiles, continuous	Basic lipid profiles were measured in community laboratories or hospitals, health checks, and disease surveillance and management system. Novel markers like Apos and Lp(a) were extracted from EMR.
eGFR in ml/(min·1.73m^2)$^{-1}$, continuous	The eGFR was calculated using the CKD-EPI equation. Serum creatinine level used in the equation was recorded in health checks database and inpatient EMR.
ACR in mg/mmol, continuous	ACR was recorded in health checks database and inpatient EMR.
Family History of ASCVD, categorical	This variable was identified according to routine epidemiological survey of local GPs. The results were self-reported by the participants.
Blood pressure lowering medication, categorical	Categories of blood pressure lowering medication included angiotensin converting enzyme (ACE) inhibitors, beta-blockers, thiazide, angiotensin II receptor blockers (ARB), calcium channel blockers, and alpha-blockers. If participants were prescribed blood pressure lowering medication prior to the study index assessment, they were classified as having blood pressure lowering medication. The information was extracted from disease management database, health checks database and inpatient EMR.
Lipid lowering medication, categorical	Lipid-lowering medication includes statins, nicotinic acid, cholesterol absorption inhibitors, probucol, cholic acid chelating agent, fibrates. If participants were prescribed lipid lowering medication prior to the study index assessment, they were classified as having lipid lowering medication. The information was extracted from disease management database, health checks database and inpatient EMR.
Hypoglycaemic medication and insulin, categorical	Oral hypoglycaemic agents include biguanides, sulfonylureas, non-sulfonylurea derivatives of anisic acid, alpha-glucosidase inhibitors, thiazolidinediones, glucagon-like peptide 1 (GLP-1) receptor agonist, dipeptidyl peptidase IV (DPP-4) inhibitors and sodium glucose transporter 2 (SGLT2) inhibitors. If participants were prescribed hypoglycaemic medication or insulin before the study index assessment, they were classified as having lipid lowering medication or insulin. The information was extracted from disease management database, health checks database and inpatient EMR.
Aspirin treatment history	If participants were prescribed aspirin before the study index assessment, they were classified as having lipid lowering medication or insulin. The information was extracted from disease management database, health checks database and inpatient EMR.

Quality control

The major quality control procedures were listed as follows:

(1) The validity and reliability of the data were first checked by the Yinzhou District Centre for Disease Control and Prevention. The integration of different data sources was conducted under uniformed processes following pre-defined criteria. Especially, in the CHERRY Study, for fatal outcomes, attribution of death refers to the primary cause provided by cause-specific mortality on death certificates in the health information system. Data undergo

annual quality assessments. The description of the death certificates has been reported previously.[1] For non-fatal outcomes, multiple sources exist in the system for the outcome definition, that is, disease management database (primary care), EMRs database (hospital care), health insurance database and disease surveillance database (disease registry). We define the disease surveillance database as the gold standard.

(2) Standard data dictionary was pre-defined. Each variable was converted to the same unit and outliners were removed based on the **Supplementary Table S3**.

(3) Conflicting data across different sources in EHR-based data exists in CHERRY. Multiple records with similar but slightly different times of diagnosis for one patient may be recorded from different sources owing to varying timing accuracy. Prioritisation of sources in terms of conflicting data will be set up. Disease surveillance was considered as the gold standard. Events for one patient within a certain time range will be considered a single event; the allowed time window is disease-specific.

(4) Outpatient and inpatient EMRs, containing information of patients' healthcare services, laboratory tests and medications, were directly transferred to the integrated data platform.

(5) Both individuals with and without health insurance can access the primary care and hospital services and therefore are all included in the system/study.

(6) For patients receiving care outside Yinzhou (e.g., patients might go to famous hospitals in Shanghai for certain complex surgical procedures), major non-fatal events occurred (e.g., CVD and cancer) are tracked from both disease surveillance and chronic disease management systems.

Potential biases of the data sources

Different EHR sources can introduce various potential biases that need to be considered when conducting research. We listed the major potential biases of data sources in our study as follows:

(1) Selection Bias: By requesting the valid lipid measurements, there were 215,744 Chinese participants in this analysis set from all 1.05 million adults in the original CHERRY study. This then didn't represent the entire population. However, this may reflect the clinical practice under a real-world scenario where nowadays lipid measurements were generally required even using the traditional guidelines recommended models.

(2) Healthcare Utilization Bias: EMR data source is primarily collected from individuals seeking medical care, which can lead to biased representations of certain health conditions or risk factors that are more likely to be captured

in clinical settings. This may be the case especially for novel risk factors, e.g., Lp (a) or BNP. In fact, the availability of these novel risk factors was indeed correlated with patients' health conditions and further associated with the outcome. By using machine learning algorithms to handle missingness in this situation, we are able to capture this information for CVD risk prediction.

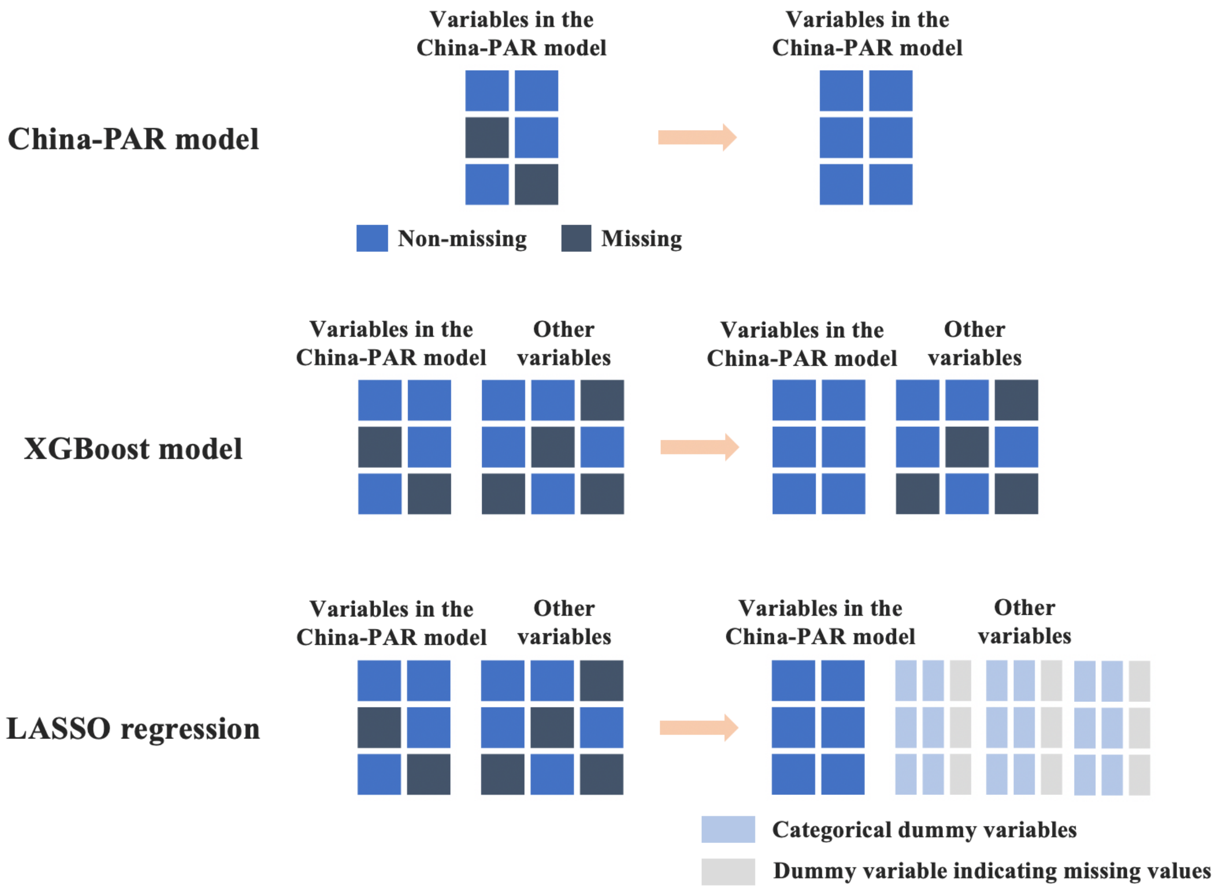
(3) Documentation Bias: Variability in data recording practices among healthcare providers can lead to inconsistencies and missing information, potentially skewing the dataset. In CHERRY, EMR information including all the laboratory test was directly copied to the integrated data platform. Repeated measurements also can help handle this problem.

(4) Self-report bias: Similar to other epidemiological study, self-report bias may occur in the registration database and disease management database when individuals provide inaccurate information about themselves, especially for the lifestyle risk factors.

Finally, our study is based on regional data which cannot represent the Chinese population nationwide. However, as the aim of our study is to demonstrate the cardiovascular predictive value of repeated measurements when machine learning models were used, this may have limited influence in the conclusion of this research.

Supplementary Method S2 Imputation strategies to handle missing values in this study

Missing values of the predictors from the China-PAR model (including sex, age, smoking, settings, systolic blood pressure, total cholesterol, high-density lipoprotein cholesterol, waist circumference, family history of ASCVD, and anti-hypertension medication) were imputed using multiple-imputation by chain equations (MICE) and five imputation sets were generated, which was a widely used approach in handling the missing values when analyzing the EHR data such as QResearch in the UK.[2] The continuous predictors were imputed using linear regression, which included all other predictors in the China-PAR model as predictors. The categorical predictors were imputed using logistic regression with the same predictors. The imputation was iterated five times, and no interaction terms were set in the imputation models. All the validations were performed in each imputation set, and the results were then pooled together based on Rubin's rules.[3] Under the consideration of informed presence, predictors other than those included in the China-PAR model were not imputed to leverage the potential information from the absence of the records. Missing values were kept unchanged in developing the XGBoost model because it could accommodate incomplete data by separating the missing values of a specific predictor into the left and right leaf nodes. However, to utilize the missing pattern of data in the construction of the LASSO regression model, continuous predictors with missing values were categorized based on the quintiles of their unique values, with missing values as a separate group. The strategies to handle missing values in different models were illustrated in the following figure.


China-PAR model
Variables in the China-PAR model
Variables in the China-PAR model
Non-missing
Missing
XGBoost model
Variables in the China-PAR model
Other variables
Variables in the China-PAR model
Other variables
LASSO regression
Variables in the China-PAR model
Other variables
Variables in the China-PAR model
Other variables
Categorical dummy variables
Dummy variable indicating missing values

The strategies to handle the missing values in different models. In the China-PAR model, all the variables were multiple-imputed based on chain-equation. In the XGBoost model, variables from the China-PAR model were multiple-imputed using the same approach and other variables were kept the same with missing data since the XGBoost algorithm can accommodate missing values. In the Lasso regression, variables from the China-PAR model were handled with the same approach and other variables were categorized with a special dummy variable indicating missing of them.

Supplementary Reference

1. Lin H, Tang X, Shen P, et al. Using big data to improve cardiovascular care and outcomes in China: a protocol for the CHinese Electronic health Records Research in Yinzhou (CHERRY) Study. *BMJ open* 2018; **8**(2): e019698.

2. 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**.

3. Harel O, Mitchell EM, Perkins NJ, et al. Multiple imputation for incomplete data in epidemiologic studies. *American journal of epidemiology* 2018; **187**(3): 576-84.