论著·智能医学与预测模型

基于两样本孟德尔随机化与可解释机器学习算法的冠心病风险预测模型构建及发病因素探究

吴 炬[1,2*],刘光阳[1,2*],明天娇[3],伍亚舟[1],李 芳[1] 陆军军医大学(第三军医大学):[1]军事预防医学系军队卫生统计学教研室,[2]基础医学院无军籍大队,重庆;[3]对外经济贸易大学国际经济贸易学院,北京

[摘要] **目的** 本研究旨在利用临床数据和基因变异数据,探究冠心病(coronary heart disease,CHD)的危险因素及其因果关联,并构建可解释的风险预测模型。**方法** 本研究为回顾性研究,从Kaggle平台和IEU OpenGWAS Project数据库分别收集临床数据和基因变异数据队列。利用SMOTEENN混合采样技术平衡数据,结合Logistic回归模型和两样本MR方法筛选危险因素并分析其因果效应。进一步基于随机森林(random forest, RF)、极限梯度提升算法(eXtreme gradient boosting, XGBoost)和神经网络(artificial neural network,ANN)构建风险预测模型,并通过SHAP算法评估特征重要性。**结果** Logistic回归分析显示,性别、年龄、教育程度、吸烟、高血压、总胆固醇、收缩压和血糖与CHD显著相关(P<0.05),而BMI、舒张压、中风和糖尿病等无显著关联。两样本MR分析表明,BMI(OR=1.014,95%CI:1.011~1.017,P=1.720×10^{-22})、年龄(OR=1.015,95%CI:1.012~1.019,P=3.455×10^{-18})、日吸烟量(OR=1.007,95%CI:1.002~1.011,P=0.005)、高血压(OR=1.057,95%CI:1.046~1.067,P=5.870×10^{-28})等11个因素与CHD呈正向因果关系;教育程度(OR=0.970,95%CI:0.960~0.979,P=1.231×10^{-9})和高密度脂蛋白(OR=0.990,95%CI:0.986~0.993,P=0.002)呈负向因果关系。机器学习模型中,SMOTEENN-RF模型表现最优,准确率达0.887,AUC为0.962。SHAP分析显示,年龄、性别、高血压、收缩压和日吸烟量对模型贡献最大。**结论** 年龄、高血压、血脂等关键因素与冠心病显著相关,并成功构建了性能优异的SMOTEENN-RF预测模型,为冠心病的风险预警与精准防治提供了可靠的方法学支持和实用工具。

[关键词] 冠心病;孟德尔随机化;可解释机器学习;预测模型;SMOTEENN

[中图法分类号] R195.1;R195.4;R541.4 [文献标志码] A

Etiological factors for coronary heart disease and construction of a risk prediction model using two-sample MR and interpretable machine learning

WU Ju[1,2*], LIU Guangyang[1,2*], MING Tianjiao[3], WU Yazhou[1], LI Fang[1] [1]Department of Health Statistics, College of Military Preventive Medicine, [2] Brigade of Non-Military Students, College of Basic Medical Sciences, Army Medical University (Third Military Medical University), Chongqing; [3]School of International Economics and Trade, University of International Business and Economics, Beijing, China

[**Abstract**] Objective To investigate the risk factors and causal relationships of coronary heart disease (CHD) by utilizing clinical data and genetic variation data, and construct an interpretable risk

[基金项目] 国家自然科学基金面上项目(82173621)

* 同等贡献者

[通信作者] 李 芳,E-mail: oucfli@163.com

2097-0927/Copyright © 2026, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议 (https://creativecommons.org/licenses/by/4.0/)

引用格式:吴炬,刘光阳,明天娇,等. 基于两样本孟德尔随机化与可解释机器学习算法的冠心病风险预测模型构建及发病因素探究[J]. 陆军军医大学学报,2026,48(02):229-241. DOI:10.16016/j.2097-0927.202507090.

Citation: WU J, LIU G Y, MING T J, *et al*. Etiological factors for coronary heart disease and construction of a risk prediction model using two-sample MR and interpretable machine learning[J]. J Army Med Univ, 2026, 48(02): 229-241. DOI: 10.16016/j.2097-0927.202507090.

prediction model for the disease. **Methods** A retrospective cohort analysis was conducted on the clinical data and genetic variation data collected from the Kaggle platform and the IEU OpenGWAS Project database, respectively. SMOTEENN mixed sampling was used to balance the data. Logistic regression models and two-sample Mendelian randomization (MR) were combined to screen risk factors and analyze their causal effects. Then risk prediction models were further constructed using random forest (RF), eXtreme gradient boosting (XGBoost), and artificial neural networks (ANN), and feature importance was assessed with the SHAP algorithm. **Results** Logistic regression analysis showed that gender, age, education level, smoking, hypertension, total cholesterol (TC), systolic blood pressure (SBP), and blood glucose were significantly associated with CHD (P<0.05), while body mass index (BMI), diastolic blood pressure (DBP), history of stroke, and diabetes had no such association. Two-sample MR analysis indicated that BMI (OR=1.014, 95%CI: 1.011 to 1.017, P=1.720×10^{-22}), age (OR=1.015, 95%CI: 1.012 to 1.019, P=3.455×10^{-18}), number of cigarettes smoked per smoker per day (OR=1.007, 95%CI: 1.002 to 1.011, P=0.005), hypertension (OR=1.057, 95%CI: 1.046 to 1.067, P=5.870×10^{-28}) were positively associated with CHD; Education level (OR=0.970, 95%CI: 0.960 to 0.979, P=1.231×10^{-9}) and high-density lipoprotein (OR=0.990, 95%CI: 0.986 to 0.993, P=0.002) were negatively associated with CHD. In the machine learning models, the SMOTEENN-RF performed best, with an accuracy rate of 0.887 and an AUC value of 0.962. SHAP analysis indicated that age, gender, hypertension, SBP, and cigarettes smoked per day contributed most to the model. **Conclusion** Key risk factors such as age, hypertension, and blood lipid levels are significantly correlated with CHD. A high-performing SMOTEENN-RF prediction model is successfully constructed, providing reliable methodological support and practical tool for CHD risk early warning and precision prevention and treatment.

[**Key words**] coronary heart disease; Mendelian randomization; interpretable machine learning; prediction model; SMOTEENN

WU Ju and LIU Guangyang contributed equally to the article. Supported the General Program of National Natural Science Foundation of China (82173621).
Corresponding author: LI Fang, E-mail: oucfli@163.com

冠心病(coronary heart disease, CHD)是由于冠状动脉粥样硬化导致管腔狭窄或闭塞,引发心肌缺血、缺氧甚至坏死的心脏疾病,是全球主要的健康负担之一[1]。据世界卫生组织统计,心血管疾病是全球死亡的首要原因,其中CHD占比尤为突出[2-3]。2021年中国心血管病报告显示,我国CHD死亡率持续上升,年住院费用高达1169.59亿元,给社会经济和公共卫生带来沉重压力。为落实《"健康中国2030"规划纲要》战略目标,深入探究CHD的危险因素及防治策略具有重要意义。

近年来,机器学习在医学领域的应用为CHD的风险预测和诊断提供了新思路。多项研究利用Logistic回归、随机森林(random forest, RF)、支持向量机等方法构建预测模型,并识别血糖、胆固醇等关键影响因素[4-9]。然而,现有研究仍存在样本量不足、混杂因素控制不充分、危险因素影响机制不明确,以及模型可解释性不足等问题。孟德尔随机化(mendelian randomization,MR)方法基于孟德尔第二定律,利用单核苷酸多态性(single nucleotide polymorphism, SNP)作为工具变量,推断暴露因素与疾病之间的因果关联。全基因组关联研究(genome-wide association studies,GWAS)为MR提供了遗传学数据支撑,用于检验非实验数据中的因果假设[10]。相较于随机对照试验,MR研究在效率、成本及伦理方面具有明显优势,尤其适用于评估潜在有害暴露对疾病发生的影响。

鉴于此,本研究基于冠心病患者的临床和基因数据,采用SMOTEENN混合采样技术解决数据不平衡问题,联合Logistic回归与两样本MR方法筛选危险因素及其因果关联。进一步利用RF、极限梯度提升算法(eXtreme gradient boosting, XGBoost)和神经网络(artificial neural network,ANN)构建预测模型,并通过SHAP算法增强模型可解释性,旨在为CHD的早期风

险评估和精准干预提供科学依据。

1 资料与方法

1.1 数据来源

本研究为回顾性研究。通过公共数据Kaggle平台[23]收集CHD患者的临床数据资料。数据集来自马萨诸塞州佛雷明汉镇居民的一项心血管疾病风险调查,包括4 240名调查对象的15个人口学特征、行为习惯和临床检查指标等潜在疾病危险因素,以及未来10年内是否发生CHD的结局事件(表1)。

通过IEU OpenGWAS Project数据库[24]获取暴露与结局的GWAS数据,确保暴露与结局间的GWAS数据不存在人群样本重叠的可能。利用R 4.3.1软件TwoSampleMR包完成工具变量的筛选与整合。具体要求为:①基因变异与暴露相关($P<5\times10^{-8}$);②保证工具变量间的独立性,排除连锁不平衡对结果的影响($r^2<0.01$);③基因变异只通过暴露因素影响结局($P_{暴露}<P_{结局}$);④效应等位基因对齐,采用harmonise方法合并数据集,确保不同基因型或等位基因之间数据的一致性。

1.2 研究方法

1.2.1 SMOTEENN算法 SMOTEENN算法是一种结合过采样与欠采样的混合数据平衡方法。本研究采用两阶段处理策略:首先通过SMOTE算法对少数类样本进行过采样,为每个少数类样本生成插值新样本[新样本=原样本+n近邻样本差异,$n\in(0,1)$];随后应用ENN算法剔除多数类中的边界噪声样本[11-13]。

1.2.2 统计学方法 应用软件R 4.3.1和Python 3.11编程进行统计学分析。符合正态分布的计量资料以$\bar{x}\pm s$表示,采用独立样本t检验;不符合正态分布的计量资料以$M(P_{25},P_{75})$表示,采用两独立样本Mann-Whitney U检验;计数资料用例(%)表示,组间比较采用χ^2检验。$P<0.05$表示差异具有统计学意义。

1.2.3 模型构建与评价 采用随机森林多重插补方法对缺失数据进行填补。利用Logistic回归进行

表1 GWAS基本信息

Table 1 Overview of basic GWAS information

Variable	Source	Population	Sample size	ID
BMI/(kg/m^2)	UK Biobank	European	461 460	ukb-b-19953
Sex	UK Biobank	European	462 725	ukb-b-8446
Age/years	UK Biobank	European	361 194	ukb-d-age
Education	EBI GWAS Catalog	European	470 941	ebi-a-GCST90029012
Smoke	GWAS and Sequencing Consortium of Alcohol and Nicotine use	European	41 969	ieu-a-963
cigsPerDay	UK Biobank	European	337 334	ieu-b-25
BPMeds	EBI GWAS Catalog	European	152 380	ebi-a-GCST90018984
Stroke	UK Biobank	European	361 194	ukb-d-C_STROKE
Hypertension	UK Biobank	European	361 194	ukb-d-I9_HYPERTENSION
Diabetes	UK Biobank	European	337 159	ukb-a-74
Diabetes2	EBI GWAS Catalog	European	655 666	ebi-a-GCST006867
Diabetes1	EBI GWAS Catalog	European	520 580	ebi-a-GCST90014023
totChol	EBI GWAS Catalog	European	437 878	ebi-a-GCST90025953
LDL-C	UK Biobank	European	440 546	ieu-b-110
HDL-C	UK Biobank	European	403 943	ieu-b-109
SBP	EBI GWAS Catalog	European	810 865	ebi-a-GCST90000062
DBP	EBI GWAS Catalog	European	810 865	ebi-a-GCST90000059
Pulse pressure	EBI GWAS Catalog	European	810 865	ebi-a-GCST90000061
Heart rate	UK Biobank	European	463 010	ukb-b-9270
Glucose	EBI GWAS Catalog	European	400 458	ebi-a-GCST90025986
Coronary heart disease	Finngen	European	412 181	finn-b-I9_CHD

GWAS: Genome-wide association study; BMI: Body mass index; BPMeds: Taking antihypertensive drugs; LDL-C: Low-density lipoprotein-cholesterol; HDL-C: High-density lipoprotein-cholesterol; SBP: Systolic blood pressure; DBP: Diastolic blood pressure

CHD相关因素筛选,并通过两样本孟德尔随机化分析相关因素对CHD的因果效应[14-18]。将数据集按7∶3随机分为训练集和测试集,在构建预测模型时,本研究选择随机森林(random forest, RF)、神经网络(artificial neural network, ANN)、极限梯度提升算法(eXtreme gradient boosting, XGBoost)3种机器学习算法,旨在综合应对临床数据中常见的复杂性非线性关系,同时兼顾模型的预测性能、稳健性与可解释性[19-21]。通过计算准确率、精确率、召回率、F1、AUC评价指标比较各模型的预测性能,并引入SHAP算法解释各特征变量对CHD的重要程度[15-22]。见图1。

2 结果

2.1 SMOTEENN处理后基线特征

原始4 240例样本中,非CHD患者与CHD患者比例为5.58∶1。为消除类别失衡带来的结果偏倚,采用SMOTEENN混合采样获得4 138例样本,其中,CHD患者2 089例,非CHD患者2 049例,比例接近1∶1。数据以7∶3的比例随机划分训练集和测试集,训练集2 896例,其中1 479例为CHD患者;测试集1 242例,其中610例为CHD患者。两组基线特征差异无统计学意义(表2)。

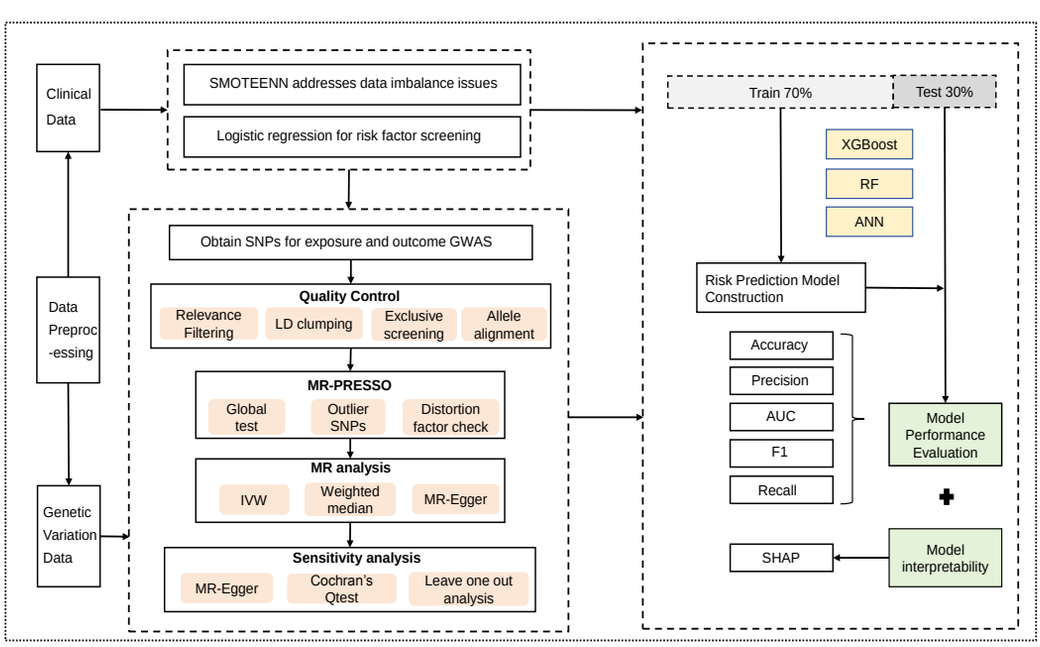
2.2 Logistic回归结果

表3显示,性别、年龄、教育程度、吸烟、服用降压药、高血压、总胆固醇、收缩压及血糖等9个变量与CHD的关联具有统计学意义(P<0.05);舒张压、BMI、日吸烟量、中风、糖尿病及心率等6个变量与CHD的关联无统计学意义。未发现模型的共线性问题(VIF<5)。

2.3 孟德尔随机化分析结果

为避免因变量信息遗漏增加决策风险,本研究将15个候选暴露全部纳入初步因果推断。在工具变量筛选阶段,性别、中风、服用降压药及心率因有效SNP数据较少,无法获得稳定的因果效应估计,予以剔除。最终对BMI、年龄、教育程度、吸烟、日吸烟量、高血压、糖尿病、总胆固醇、收缩压、舒张压及血糖共11个暴露与CHD的因果关系展开分析。

通过MR-PRESSO剔除异常SNP后,总体因果效应估计变化不显著,结果稳健。表4中MR-Egger、WME和IVW结果均显示,BMI、高血压、总胆固醇、收缩压及舒张压均与CHD风险呈正向因果关系(P<0.05),基于随机效应的IVW结果显示,11个暴露均与CHD的因果关联显著(P<0.05)。需要注意的是,糖尿病与吸烟在3种方法中的效应方向不一致,提示其因果关系尚存在不确定性,有待进一步验证。教育程度表现为负向关系,其余暴露均与CHD呈现正向关系,见图4。



RF: Random forest; XGBoost: eXtreme gradient boosting; ANN: Artificial neural network

图1 技术路线

Figure 1 Technology roadmap

表2 SMOTEENN处理后的基线特征[例(%),$\bar{x}\pm s$]

Table 2 Baseline features after SMOTE-ENN processing [n(%),($\bar{x}\pm s$)]

Variable	Train (n=2 896)	Test (n=1 242)	P	Variable	Train (n=2 896)	Test (n=1 242)	P
Sex			0.085	Hypertension			0.213
Female	1 285 (44)	582 (47)		No	1 707 (59)	758 (61)	
Male	1 611 (56)	660 (53)		Yes	1 189 (41)	484 (39)	
Age/years	50.570±8.400	50.260±8.444	0.303	Diabetes			0.453
Education			0.963	No	2 748 (95)	1 197 (96)	
Below high school	999 (34)	414 (33)		Yes	148 (5)	45 (4)	
High school	964 (33)	441 (36)		TC	239.090±45.651	237.270±44.648	0.276
Vocational School	569 (20)	237 (19)		SBP	134.710±23.560	133.510±22.107	0.206
University	364 (13)	150 (12)		DBP	83.960±12.499	83.340±12.017	0.160
Smoker			0.662	BMI/(kg/m^2)	25.960±3.802	25.890±3.752	0.717
No	1 324 (46)	578 (47)		Heart rate	75.990±11.506	75.280±11.336	0.073
Yes	1 572 (54)	664 (53)		Blood glucose level	84.730±31.684	84.030±29.155	0.681
cigsPerDay	9.450±11.987	9.150±11.582	0.770	Coronary heart disease			0.318
BPMeds			0.746	No	1 417 (49)	632 (51)	
No	2 688 (93)	1 161 (93)		Yes	1 479 (51)	610 (49)	
Yes	208 (7)	81 (7)					
Stroke			0.874				
No	2 849 (98)	1 218 (98)					
Yes	47 (2)	24 (2)					

cigsPerDay: Number of cigarettes smoked per smoker per day; BPMeds: Taking antihypertensive drugs; TC: Total cholesterol; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; BMI: Body mass index

表3 潜在风险变量Logistic回归分析结果

Table 3 Results of logistic regression analysis of potential risk variables

Variable	β	SE	P	OR(95%CI)	VIF
Sex	1.174	0.105	<0.010	3.236(2.636~3.972)	1.207
Age/years	0.122	0.007	<0.010	1.130(1.114~1.146)	1.374
Education	0.237	0.048	<0.010	1.267(1.153~1.393)	1.077
Smoker	0.964	0.149	<0.010	2.622(1.960~3.508)	2.479
cigsPerDay	−0.001	0.006	0.877	0.999(0.987~1.011)	2.522
BPMeds	0.910	0.249	<0.010	2.485(1.524~4.052)	1.084
Stroke	0.885	0.532	0.096	2.423(0.854~6.872)	1.013
Hypertension	0.613	0.135	<0.010	1.846(1.418~2.404)	1.918
Diabetes	0.209	0.351	0.552	1.233(0.619~2.455)	1.536
TC	0.004	0.001	<0.010	1.004(1.002~1.007)	1.078
SBP	0.012	0.004	0.004	1.013(1.004~1.021)	3.488
DBP	0.004	0.007	0.538	1.004(0.990~1.019)	2.947
BMI	−0.004	0.014	0.770	0.996(0.969~1.023)	1.234
Heart rate	0.003	0.004	0.501	1.003(0.994~1.012)	1.100
Blood glucose	0.008	0.003	0.005	1.008(1.002~1.014)	1.561

cigsPerDay: Number of cigarettes smoked per smoker per day; BPMeds: Taking antihypertensive drugs; TC: Total cholesterol; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; BMI: Body mass index

表4 11个暴露因素的因果效应估计

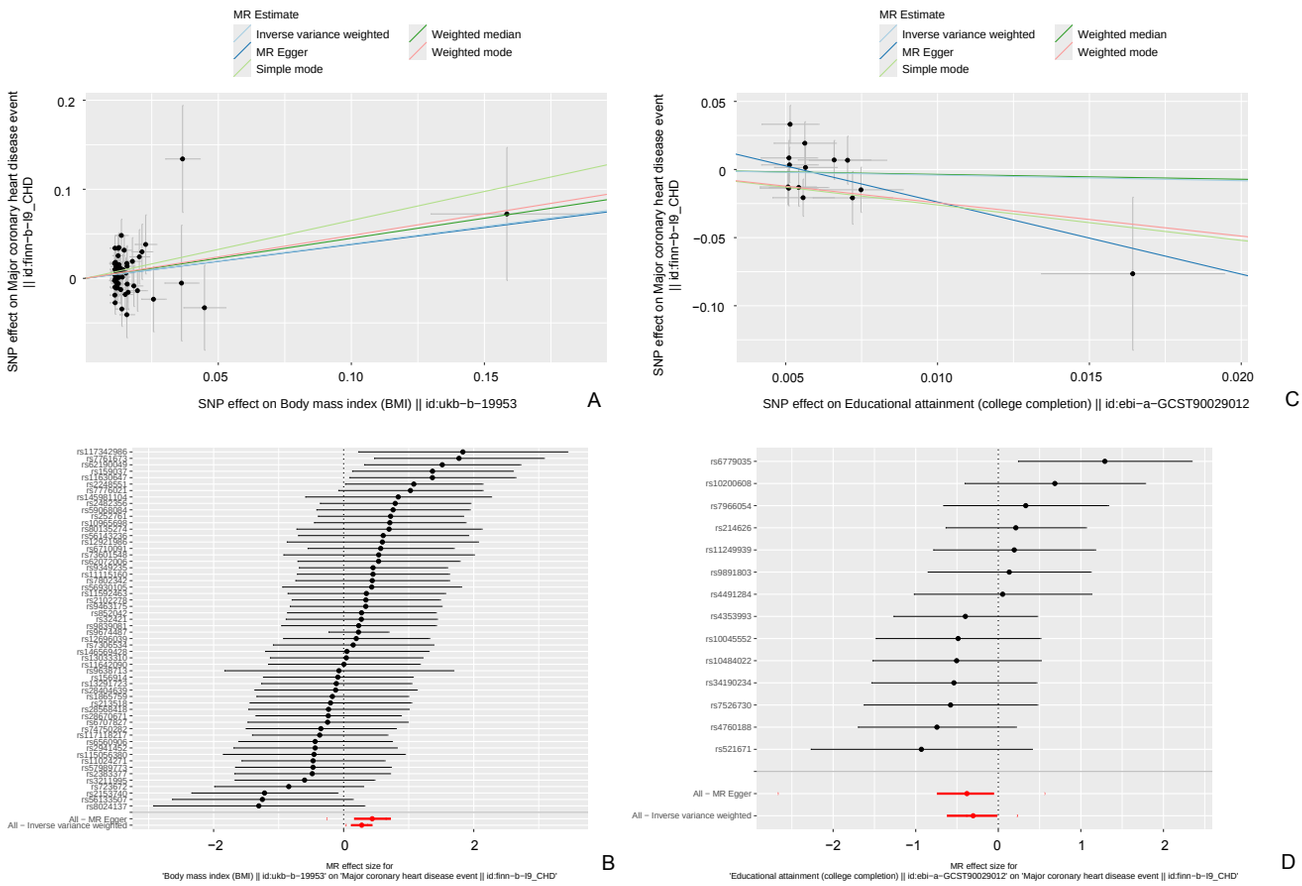
Table 4 Estimated causal effects of 11 exposure factors

Exposure	Method	β	*SE*	*OR*(95%*CI*)	*P*
	MR-Egger	0.014	0.004	1.014(1.005~1.022)	1.605×10^{-3}
BMI/(kg/m^2)	WME	0.013	0.002	1.013(1.009~1.018)	6.500×10^{-11}
	IVW	0.014	0.001	1.014(1.011~1.017)	1.720×10^{-22}
	MR-Egger	0.019	0.010	1.019(0.100~1.039)	6.015×10^{-2}
Age/years	WME	0.013	0.002	1.013(1.009~1.018)	2.259×10^{-8}
	IVW	0.015	0.002	1.015(1.012~1.019)	3.455×10^{-18}
	MR-Egger	−0.038	0.021	0.963(0.924~1.005)	8.203×10^{-2}
Education	WME	−0.032	0.007	0.968(0.955~0.982)	3.614×10^{-6}
	IVW	−0.031	0.005	0.970(0.960~0.979)	1.231×10^{-9}
	MR-Egger	−0.016	0.045	0.984(0.900~1.075)	0.724
Smoker	WME	0.018	0.011	1.018(0.996~1.041)	0.117
	IVW	0.027	0.010	1.027(1.008~1.046)	0.005
	MR-Egger	0.013	0.007	1.013(0.998~1.027)	0.099
cigsPerDay	WME	0.007	0.003	1.007(1.001~1.014)	0.030
	IVW	0.007	0.002	1.007(1.002~1.011)	0.005
	MR-Egger	0.056	0.013	1.058(1.031~1.085)	2.249×10^{-5}
Hypertension	WME	0.053	0.006	1.054(1.041~1.068)	1.101×10^{-16}
	IVW	0.055	0.005	1.057(1.046~1.067)	5.870×10^{-28}
	MR-Egger	−0.006	0.053	0.994(0.896~1.102)	0.911
Diabetes	WME	0.069	0.021	1.071(1.029~1.116)	0.001
	IVW	0.077	0.026	1.080(1.027~1.136)	0.003
	MR-Egger	0.013	0.005	1.013(1.003~1.024)	9.725×10^{-3}
TC	WME	0.010	0.002	1.010(1.005~1.015)	4.000×10^{-5}
	IVW	0.007	0.003	1.007(1.002~1.013)	6.461×10^{-3}
	MR-Egger	0.018	0.007	1.018(1.004~1.033)	1.203×10^{-2}
SBP	WME	0.018	0.003	1.018(1.011~1.025)	1.575×10^{-7}
	IVW	0.016	0.003	1.017(1.011~1.023)	5.323×10^{-8}
	MR-Egger	0.018	0.007	1.017(1.005~1.031)	7.246×10^{-3}
DBP	WME	0.019	0.003	1.019(1.013~1.026)	7.198×10^{-9}
	IVW	0.018	0.003	1.018(1.013~1.024)	9.769×10^{-11}
	MR-Egger	0.002	0.007	1.002(0.988~1.015)	0.827
Blood glucose	WME	0.007	0.003	1.007(1.001~1.013)	0.032
	IVW	0.009	0.003	1.009(1.002~1.015)	0.006

cigsPerDay: Number of cigarettes smoked per smoker per day; TC: Total cholesterol; SBP: Systolic blood pressure; DBP: Diastolic blood pressure

表5的Cochran's Q检验显示,除年龄和日吸烟量外,其余9个暴露因素的工具变量SNP间均存在异质性(P<0.05)。MR-Egger回归截距项在11个暴露中均不显著,提示不存在明显的水平多效性。漏斗图与LOOA图也未发现异质性及水平多效性(图3)。

进一步对糖尿病和总胆固醇进行了细分,分别使用1型糖尿病、2型糖尿病以及LDL、HDL的GWAS数据进行MR分析,结果显示2型糖尿病、LDL和HDL均与CHD存在因果关联。此外,尝试将收缩压和舒张压合并,采用脉压的GWAS数据进行MR分析,结果同样支持脉压与冠心病之间存在因果关联,结果见表6。敏感性分析表明,上述各变量与CHD之间均不存在水平多效性(表7)。



A: Scatter plot of BMI; B: Forest plot of BMI; C: Scatter plot of education level; D: Forest plot of education level. BMI: Body mass index

图2 BMI与教育程度的散点图及森林图

Figure 2 Scatter plot and forest plot of BMI and education level

表5 11个暴露因素的敏感性分析

Table 5 Sensitivity analysis for 11 exposure factors

Exposure	Q-test. *P*	MR-Egger intercept	intercept *P*
BMI/(kg/m^2)	1.997×10^{-5}	1.175×10^{-5}	0.879
Age/years	0.110	-1.204×10^{-4}	0.709
Education	0.012	4.640×10^{-5}	0.757
Smoker	1.782×10^{-4}	3.032×10^{-4}	0.339
cigsPerDay	0.536	-2.748×10^{-4}	0.399
Hypertension	4.318×10^{-15}	-1.036×10^{-5}	0.920
Diabetes	1.157×10^{-30}	4.347×10^{-4}	0.080
TC	2.555×10^{-56}	-1.919×10^{-4}	0.167
SBP	7.232×10^{-24}	-3.292×10^{-5}	0.784
DBP	7.012×10^{-16}	7.005×10^{-6}	0.950
Blood glucose	1.091×10^{-19}	2.021×10^{-4}	0.248

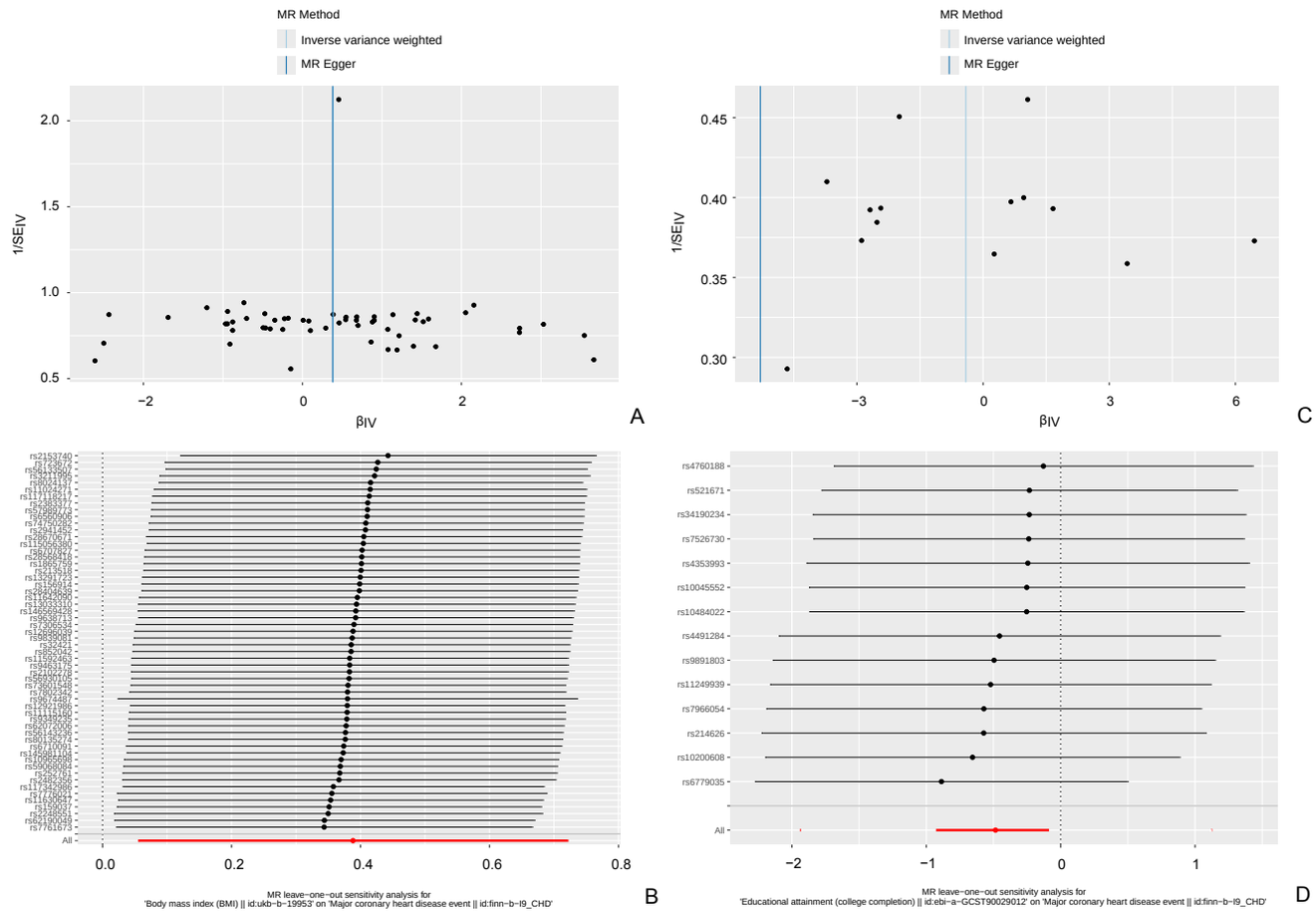
BMI: Body mass index; cigsPerDay: Number of cigarettes smoked per smoker per day; TC: Total cholesterol; SBP: Systolic blood pressure; DBP: Diastolic blood pressure

2.4 模型预测性能

表8显示,训练集与测试集的表现高度一致,提示模型未出现过拟合。测试集中,未做类别平衡的RF、XGBoost模型的各项指标均明显低于经SMOTEENN处理后的对应模型。平衡后,3种机器学习算法的预测性能全面提升,其中RF表现最优,准确率、精确率与AUC值分别达到0.887、0.885和0.962。图5 ROC曲线表明,3种模型对CHD正负样本均有出色的区分能力,RF模型整体性能最佳。

2.5 模型可解释性

图6显示,3种模型一致认为年龄是CHD的首要风险因子,而糖尿病、中风及BMI的贡献相对有限。研究以性能最优的RF模型为代表进行特征重要性可视化,SHAP值排序前5位的危险因素依次为:年龄、性别、高血压、收缩压、日吸烟量。除教育程度呈负向关联外,其余变量均与CHD呈正向关联。



A: Funnel chart of BMI; B: LOOA chart of BMI; C: Funnel chart of education level; D: LOOA chart of education level; BMI: Body mass index; LOOA: Leave one out analysis

图3 BMI与教育程度的漏斗图与LOOA图

Figure 3 Funnel plot and LOOA plot of MBI and education level

表6 4个细分因素的因果效应估计

Table 6 Estimated causal effects of 4 expansion factors

Exposure	Method	β	*se*	*OR*(95%*CI*)	*P*
Diabetes (II)	MR-Egger	0.011	0.097	1.011(0.998~1.024)	0.007
	WME	0.007	0.000	1.007(1.005~1.010)	0.001
	IVW	0.007	0.000	1.006(1.004~1.009)	0.001
LDL	MR-Egger	0.038	0.000	1.039(1.024~1.053)	0.007
	WME	0.015	0.000	1.015(1.010~1.021)	0.003
	IVW	0.014	0.000	1.015(1.007~1.022)	0.004
HDL	MR-Egger	−0.008	0.007	0.992(0.987~0.998)	0.003
	WME	−0.007	0.000	0.993(0.989~0.997)	0.002
	IVW	−0.010	0.000	0.990(0.986~0.993)	0.002
Pulse pressure	MR-Egger	0.005	0.523	1.005(0.989~1.021)	0.008
	WME	0.013	0.000	1.013(1.006~1.020)	0.004
	IVW	0.012	0.000	1.012(1.005~1.018)	0.003

LDL: Low-density lipoprotein; HDL: High-density lipoprotein

表7 4个细分因素的敏感性分析

Table 7 Sensitivity analysis for 4 expansion factors

Exposure	Q-test. *P*	MR-Egger intercept	Intercept *P*
Diabetes (Ⅱ)	0.549	-2.490×10^{-4}	0.502
LDL	0.219	-6.813×10^{-4}	0.228
HDL	1.259×10^{-24}	-7.808×10^{-5}	0.265
Pulse pressure	4.140×10^{-26}	1.138×10^{-4}	0.380

LDL: Low-density lipoprotein; HDL: High-density lipoprotein

表8 模型预测结果评价

Table 8 Evaluation of model prediction

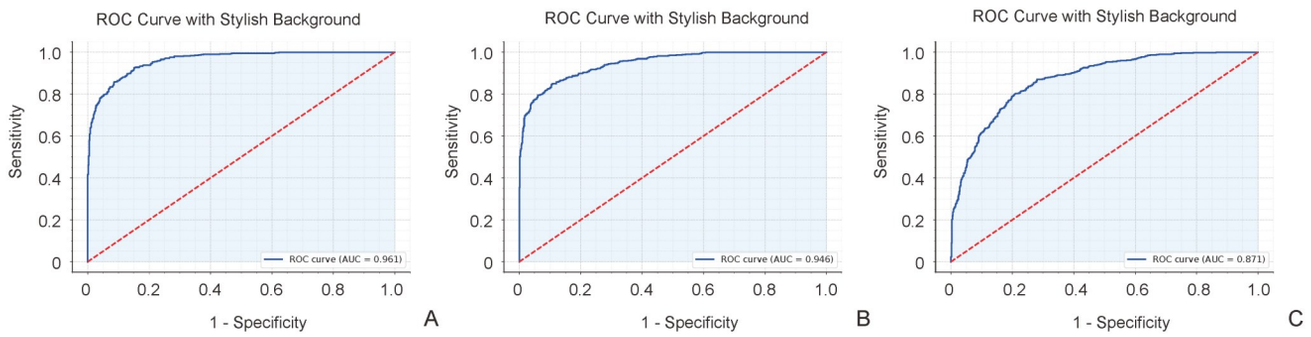
Data	Balance technology	Model	Accuracy	Precision	Recall	F1	AUC
Train	SMOTEENN	RF	0.952	0.956	0.949	0.953	0.994
		XGBoost	0.864	0.876	0.853	0.865	0.944
		ANN	0.818	0.828	0.812	0.820	0.896
	SMOTE	RF	0.934	0.967	0.899	0.932	0.985
		XGBoost	0.865	0.919	0.800	0.855	0.932
		ANN	0.782	0.808	0.738	0.772	0.866
	Null	RF	0.880	1.000	0.204	0.338	0.963
		XGBoost	0.862	0.975	0.087	0.160	0.796
		ANN	0.860	0.750	0.107	0.188	0.757
Test	SMOTEENN	RF	0.887	0.885	0.885	0.885	0.962
		XGBoost	0.866	0.866	0.859	0.863	0.946
		ANN	0.791	0.797	0.770	0.783	0.871
	SMOTE	RF	0.873	0.904	0.835	0.868	0.941
		XGBoost	0.851	0.902	0.790	0.842	0.923
		ANN	0.772	0.800	0.733	0.764	0.853
	Null	RF	0.849	0.647	0.056	0.103	0.739
		XGBoost	0.849	0.632	0.060	0.110	0.734
		ANN	0.849	0.580	0.100	0.170	0.731

RF: Random forest; XGBoost: eXtreme gradient boosting; ANN: Artificial neural network



图4 模型在测试集的预测性能

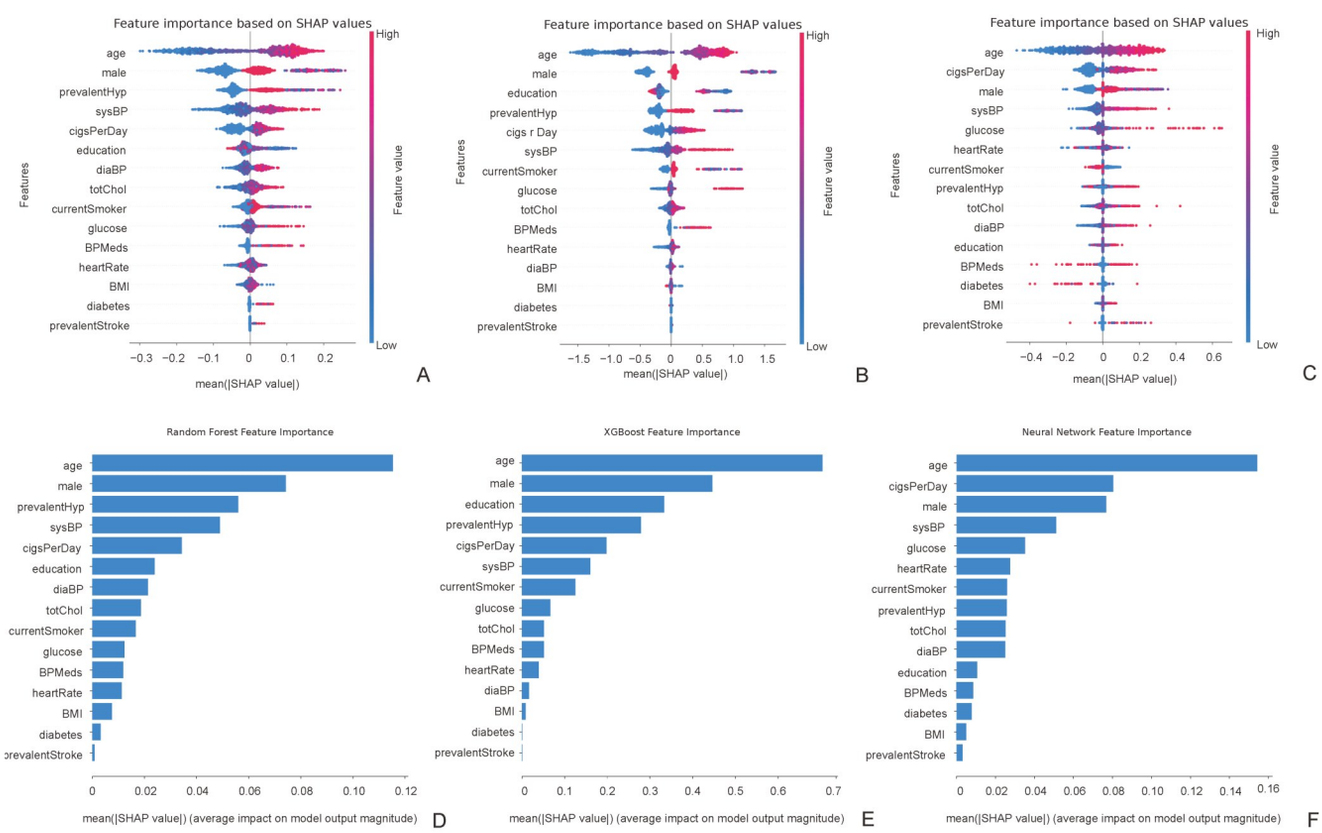
Figure 4 Prediction performance of the models on the testing set



A: RF; B: XGBoost; C: ANN. RF: Random forest; XGBoost: eXtreme gradient boosting; ANN: Artificial neural network

图5 模型ROC曲线

Figure 5 Receiver operating characteristic curves of the models



A: RF; B: XGBoost; C: ANN. RF: Random forest; XGBoost: eXtreme gradient boosting; ANN: Artificial neural network

图6 模型SHAP解释概要图和条形图

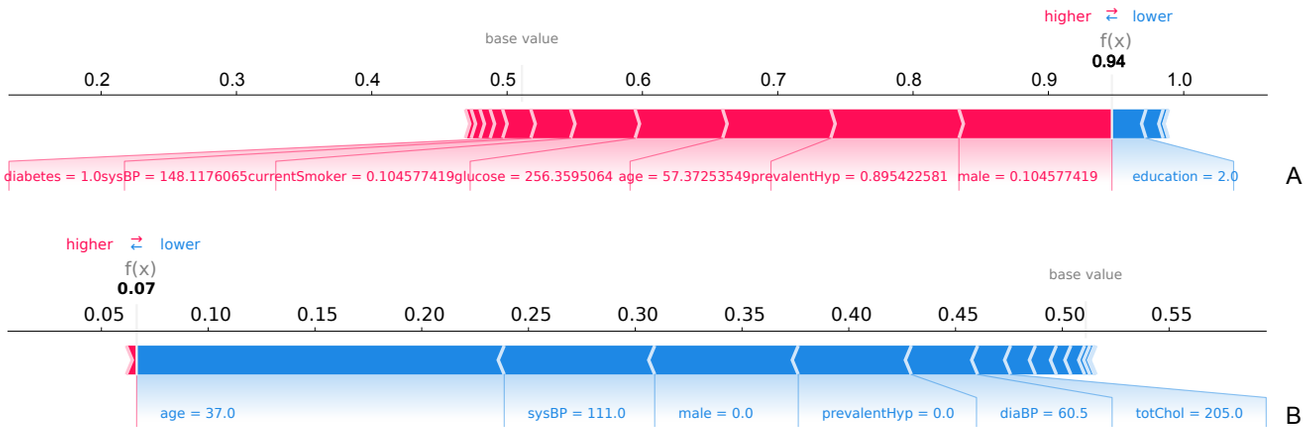
Figure 6 Model SHAP explanation summary plot and bar chart

为直观展示模型在临床决策中的辅助价值,分别选取典型病例进行解释。图7A病例:高龄男性,受教育程度较高,舒张压正常但收缩压与血糖均偏高,并伴高血压及吸烟史。模型输出CHD风险为94%,其中受教育程度起负向调节作用,舒张压几乎无贡献。图7B病例:年轻女性,无高血压,收缩压、舒张压及总胆固醇均正常。模型预测其CHD概率仅7%。

3 讨论

3.1 Logistic回归关联变量差异

本研究基于Kaggle公开的CHD临床数据,探讨了CHD危险因素及其预测模型构建。在危险因素分析方面,本研究采用Logistic回归和两样本孟德尔随机化(MR)相结合的方法,发现虽然舒张压、BMI、日吸烟量



A: Patients with CHD; B: Normal individuals. CHD: Coronary heart disease

图7 个体的SHAP特征贡献图

Figure 7 Individual feature contribution plot

等指标在临床指南中被广泛认可为CHD危险因素[23-28],但在本研究的Logistic回归中未显示显著关联。这一差异可能与数据分布及混杂因素控制有关。后续机器学习模型比较表明,保留这些变量仍有助于提升预测性能,提示变量在预测中的贡献可能不仅依赖于其在传统回归中的显著性,还应考虑其在不同模型中的稳健性和互补性。因此,在冠心病预测建模中,需结合多种分析方法审慎评估变量作用,避免仅依赖单一统计模型的结果进行变量筛选。

3.2 MR分析结果差异

在因果关系验证方面,MR分析揭示了多个因素与CHD的因果关联。值得注意的是,本研究对糖尿病和血脂指标进行了更细致的分层分析:将糖尿病细分为1型和2型,发现2型糖尿病与CHD存在正向因果关系;对总胆固醇进行细分后,明确LDL-C的正向作用和HDL-C的保护作用。特别值得关注的是脉压指标,研究证实其作为心血管疾病独立危险因素的价值,建议在预测模型中可考虑用脉压替代传统的收缩压和舒张压指标[29-31]。虽然Cochran's Q检验显示部分SNP存在异质性,但综合LOOA、漏斗图和MR-Egger等敏感性分析结果,认为这种异质性在可接受范围内,不影响主要结论的可靠性。

值得说明的是,尽管本研究的Logistic回归和MR分析中糖尿病整体未显示与CHD的显著关联,与部分临床观察性研究结果存在差异,但本研究认为这恰恰反映了方法学差异和人群异质性的影响。后续使用细分变量进行的MR分析为糖尿病的因果作用提供了更深入的见解,2型糖尿病表现出明确的阳性因果关联,与临床共识高度一致。这一结果不仅强化了2型糖尿病作为CHD独立危险因素的证据,也凸显了在因果推断中区分糖尿病亚型的重要性。

3.3 模型创新与不足

在预测模型构建方面,采用SMOTEENN混合采样技术处理数据不平衡问题,并比较了RF、XGBoost和ANN 3种机器学习算法的性能。结果显示,SMOTEENN-RF模型表现最优,其SHAP解释结果与临床经验高度吻合,证实年龄、性别、高血压等关键因素的重要预测价值。

尽管本研究取得了一定成果,但仍存在以下不足。研究数据的代表性可能受到样本来源地域局限性的影响,主要基于美国特定地区的临床数据可能无法完全反映其他人群的特征差异。同时,部分具有潜在临床价值的预测因子由于缺乏相应的遗传标记数据而未能纳入分析,这对建立更全面的预测体系造成了一定限制。在数据处理环节,虽然使用了采样技术来平衡数据集,但这一过程本身可能引入新的复杂性。未来研究若能整合多中心、多地域的临床数据,特别是纳入具有不同遗传背景的人群队列,并开发更精细的数据处理方法,将有望构建更具普适性的预测模型,为临床决策提供更可靠的依据。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] SHAKYA S, SHRESTHA A, ROBINSON S, *et al*. Global comparison of the economic costs of coronary heart disease: a systematic review and meta-analysis[J]. BMJ Open, 2025,

15(1): e084917. DOI:10.1136/bmjopen-2024-084917.

[2] 徐兴梅, 胡阳, 邓洁, 等. 冠心病心理障碍及其诊治进展[J]. 中国老年保健医学, 2022, 20(1): 99-103. DOI:10.3969/j.issn.1672-2671.2022.01.030.
XU X M, HU Y, DENG J, *et al*. Psychological disorders of coronary heart disease and its progress in diagnosis and treatment[J]. Chin J Geriatr Care, 2022, 20(1): 99-103. DOI:10.3969/j.issn.1672-2671.2022.01.030.

[3] AMINORROAYA A, YOOSEFI M, REZAEI N, *et al*. Global, regional, and national quality of care of ischaemic heart disease from 1990 to 2017: a systematic analysis for the Global Burden of Disease Study 2017[J]. Eur J Prev Cardiol, 2022, 29(2): 371-379. DOI:10.1093/eurjpc/zwab066.

[4] 王成龙, 何新叶, 李燕奇, 等. 冠心病危险因素分析及机器学习算法构建预测模型研究[J]. 中国循证心血管医学杂志, 2023, 15(12): 1334-1337. DOI:10.3969/j.issn.1674-4055.2023.12.12.
WANG C L, HE X Y, LI Y Q, *et al*. Analysis of coronary heart disease risk factors and construction of disease prediction models[J]. Chin J Evid Bases Cardiovasc Med, 2023, 15(12): 1334-1337. DOI: 10.3969/j.issn.1674-4055.2023.12.12.

[5] JOLOUDARI J H, HASSANNATAJ JOLOUDARI E, SAADATFAR H, *et al*. Coronary artery disease diagnosis; ranking the significant features using a random trees model[J]. Int J Environ Res Public Health, 2020, 17(3): 731. DOI:10.3390/ijerph17030731.

[6] GARAVAND A, SALEHNASAB C, BEHMANESH A, *et al*. Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms[J]. J Healthc Eng, 2022, 2022: 5359540. DOI: 10.1155/2022/5359540.

[7] NAUSHAD S M, HUSSAIN T, INDUMATHI B, *et al*. Machine learning algorithm-based risk prediction model of coronary artery disease[J]. Mol Biol Rep, 2018, 45(5): 901-910. DOI:10.1007/s11033-018-4236-2.

[8] ZHANG T X, HUANG S M, XIE P F, *et al*. Development of machine learning tools for predicting coronary artery disease in the Chinese population[J]. Dis Markers, 2022, 2022: 6030254. DOI:10.1155/2022/6030254.

[9] 黄浩东, 刘小株, 龚军, 等. 基于机器学习算法建立2型糖尿病患者冠心病辅助诊断模型[J]. 复旦学报(医学版), 2022, 49(2): 226-233, 256. DOI: 10.3969/j. issn. 1672-8467.2022.02.010.
HUANG H D, LIU X Z, GONG J, *et al*. Auxiliary diagnosis model of coronary heart disease in patients with type 2 diabetes mellitus based on machine learning algorithm[J]. Fudan Univ J Med Sci, 2022, 49(2): 226-233, 256. DOI: 10.3969/j.issn.1672-8467.2022.02.010.

[10] LARSSON S C, BUTTERWORTH A S, BURGESS S. Mendelian randomization for cardiovascular diseases: principles and applications[J]. Eur Heart J, 2023, 44(47): 4913-4924. DOI:10.1093/eurheartj/ehad736.

[11] WANG H Y, LI X K, YUAN Y, *et al*. Association of machine learning ultrasound radiomics and disease outcome in triple negative breast cancer[J]. Am J Cancer Res, 2022, 12(1): 152-164.

[12] NONG X Z, HUANG S Y, HE F C, *et al*. Can a robust machine learning-based integrated framework benefit universal water quality evaluation applications? Comparative view in the whole coastal regions of China[J]. Mar Pollut Bull, 2025, 220: 118399. DOI: 10.1016/j.marpolbul.2025.118399.

[13] XU Y, YE W, SONG Q Y, *et al*. Using machine learning models to predict the duration of the recovery of COVID-19 patients hospitalized in Fangcang shelter hospital during the *Omicron* BA. 2.2 pandemic[J]. Front Med, 2022, 9: 1001801. DOI:10.3389/fmed.2022.1001801.

[14] 王晓丽, 施天行, 彭德荣, 等. 两种机器学习算法构建老年冠心病患病风险评估模型的效能比较研究[J]. 中华全科医学, 2021, 19(4): 523-527. DOI:10.16766/j.cnki.issn.1674-4152.001852.
WANG X L, SHI T X, PENG D R, *et al*. Comparative study on the effectiveness of two machine learning algorithms in constructing risk assessment models of coronary heart disease in the elderly[J].Chin J Gen Pract, 2021, 19(4): 523-527. DOI:10.16766/j.cnki.issn.1674-4152.001852.

[15] SHI Y J, XIANG Y, YE Y X, *et al*. A framework for detecting causal effects of risk factors at an individual level based on principles of Mendelian randomisation: applications to modelling individualised effects of lipids on coronary artery disease[J]. EBioMedicine, 2025, 113: 105616. DOI:10.1016/j.ebiom.2025.105616.

[16] LIN Z T, DENG Y Q, PAN W. Combining the strengths of inverse-variance weighting and Egger regression in Mendelian randomization using a mixture of regressions model[J]. PLoS Genet, 2021, 17(11): e1009922. DOI: 10.1371/journal.pgen.1009922.

[17] BURGESS S, DAVEY SMITH G, DAVIES N M, *et al*. Guidelines for performing Mendelian randomization investigations: update for summer 2023[J]. Wellcome Open Res, 2023, 4: 186. DOI:10.12688/wellcomeopenres.15555.3.

[18] BOWDEN J, DAVEY SMITH G, BURGESS S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression[J]. Int J Epidemiol, 2015, 44(2): 512-525. DOI: 10.1093/ije/

dyv080.

[19] KIM S, LEE J, NAM S H. Artificial intelligence applications in intracerebral hemorrhage care: implications for clinical and nursing practice - a narrative literature review[J]. Front Rehabil Sci, 2025, 6: 1620335. DOI: 10.3389/fresc.2025.1620335.

[20] IKRAM R M A, WANG M, MOAYEDI H, *et al*. Application of smart technologies for predicting soil erosion patterns[J]. Sci Rep, 2025, 15: 26479. DOI: 10.1038/s41598-025-12125-0.

[21] WEN J, SHI X W, LIU Y, *et al*. Malnutrition mediates the association between handgrip status and asthma risk: an observational and prospective cohort study from multiple European countries[J]. Front Nutr, 2025, 12: 1555888. DOI:10.3389/fnut.2025.1555888.

[22] MA C S, LEI M. Time series prediction for monitoring cardiovascular health in autistic patients[J]. Front Psychiatry, 2025, 16: 1623986. DOI: 10.3389/fpsyt.2025.1623986.

[23] RAHMAN F, AL RIFAI M, BLAHA M J, *et al*. Relation of diastolic blood pressure and coronary artery calcium to coronary events and outcomes (from the multi-ethnic study of atherosclerosis)[J]. Am J Cardiol, 2017, 120(10): 1797-1803. DOI:10.1016/j.amjcard.2017.07.094.

[24] KIM M S, KIM W J, KHERA A V, *et al*. Association between adiposity and cardiovascular outcomes: an umbrella review and meta-analysis of observational and Mendelian randomization studies[J]. Eur Heart J, 2021, 42(34): 3388-3403. DOI:10.1093/eurheartj/ehab454.

[25] 陆丕能, 孙宁玲, 陆鋆, 等. 吸烟量与冠心病关系的病例对照研究[J]. 中华流行病学杂志, 2002, 23(4): 297-300. DOI:10.3760/j.issn: 0254-6450.2002.04.015.
LU P N, SUN N L, LU J, *et al*. Case control study on the association of coronary artery disease and cigarette smoking[J]. Chin J Epidemiol, 2002, 23(4): 297-300. DOI: 10.3760/j.issn: 0254-6450.2002.04.015.

[26] LALO R, ZEKJA I, KAMBERI F. Association of cardiovascular disease risk and health-related behaviors in stroke patients[J]. Int J Environ Res Public Health, 2023, 20(4): 3693. DOI:10.3390/ijerph20043693.

[27] 洪天配, 母义明, 纪立农, 等. 2型糖尿病合并动脉粥样硬化性心血管疾病患者降糖药物应用专家共识[J]. 中国介入心脏病学杂志, 2017, 25(7): 361-371. DOI: 10.3969/j.issn.1004-8812.2017.07.001.
HONG T P, MU Y M, JI L N, *et al*. Chinese expert consensus on anti-diabetic agents in treating patients with type 2 diabetes and atherosclerotic cardiovascular disease[J]. Chin J Interv Cardiol, 2017, 25(7): 361-371. DOI:10.3969/j.issn.1004-8812.2017.07.001.

[28] BÖHM M, REIL J C, DEEDWANIA P, *et al*. Resting heart rate: risk indicator and emerging risk factor in cardiovascular disease[J]. J Tird Mil Med Univ, 2015, 128(3): 219-228. DOI:10.1016/j.amjmed.2014.09.016.

[29] 刘涛, 陈勇, 赖成民, 等. 脉压与中老年冠心病关系的探讨[J]. 第三军医大学学报, 2006, 28(15): 1628-1629. DOI:10.3321/j.issn: 1000-5404.2006.15.028.
LIU T, CHEN Y, LAI C M, *et al*. Relationship between pulse pressure and coronary heart disease in middle-aged and elderly people[J]. J Tird Mil Med Univ, 2006, 28(15): 1628-1629. DOI: 10.3321/j. issn: 1000-5404.2006.15.028.

[30] JANKOWSKI P, WEBER T. Pulse pressure and cardiovascular risk in diseased patients[J]. J Hum Hypertens, 2016, 30(5): 293-294. DOI: 10.1038/jhh.2015.83.

[31] ASSMANN G, CULLEN P, EVERS T, *et al*. Importance of arterial pulse pressure as a predictor of coronary heart disease risk in PROCAM[J]. Eur Heart J, 2005, 26(20): 2120-2126. DOI:10.1093/eurheartj/ehi467.

[32] GRIMES D A, SCHULZ K F. An overview of clinical research: the lay of the land[J]. Lancet, 2002, 359(9300): 57-61. DOI:10.1016/S0140-6736(02)07283-5.

(**收稿**:2025-07-28;**修回**:2025-11-10)

(编辑 邓强庭)