[文章编号] 1671-587X(2024)04-1044-11 DOI:10.13481/j.1671-587X.202404019

基于心脑血管疾病发病风险预测的CatBoost算法和贝叶斯网络模型分析

王爱民[1]，王凤琳[1]，黄一铭[1]，徐雅琪[1]，张文婧[1]，丛显铸[1]，苏维强[1]，王素珍[1]，高梦瑶[1]，李 爽[1]，孔雨佳[1]，石福艳[1]，陶恩学[2]

（1. 山东第二医科大学公共卫生学院卫生统计学教研室，山东 潍坊 261053；2. 山东第二医科大学基础医学院，山东 潍坊 261053）

[摘 要] **目的：**筛选影响心脑血管疾病发病的主要特征变量，基于排序前10位的特征变量构建心脑血管疾病发病风险贝叶斯网络模型，为心脑血管疾病发病风险预测提供参考。**方法：**从英国生物样本（UK Biobank）数据库中纳入315 896例参与者和相关变量，通过类别型特征提升（CatBoost）算法进行特征选择，将所有参与者按7∶3比例随机分为训练集和测试集，并基于最大最小爬山（MMHC）算法构建贝叶斯网络模型。**结果：**本研究中人群心脑血管疾病患病率为28.8%。CatBoost算法筛选的排名前10位变量分别为年龄、体质量指数（BMI）、低密度脂蛋白胆固醇（LDL-C）、总胆固醇（TC）、甘油三酯-葡萄糖（TyG）指数、家族史、载脂蛋白A/B比值、高密度脂蛋白胆固醇（HDL-C）、吸烟状态和性别。CatBoost训练集模型受试者工作特征（ROC）曲线下面积（AUC）为0.770，模型准确性为0.764；验证集模型AUC为0.759，模型准确性为0.763。临床效能分析，训练集阈值范围为0.06～0.85，验证集阈值范围为0.09～0.81。心脑血管疾病发病风险贝叶斯网络模型分析，年龄、性别、吸烟状态、家族史、BMI和载脂蛋白A/B比值与心脑血管疾病直接相关，是心脑血管疾病发生的重要风险因素，TyG指数、HDL-C、LDL-C和TC通过影响BMI和载脂蛋白A/B比值间接影响心脑血管疾病的发生风险。**结论：**控制BMI、载脂蛋白A/B比值和吸烟行为，可以降低心脑血管疾病的发病风险。贝叶斯网络模型可用于预测心脑血管疾病发病风险。

[关键词] 心脑血管疾病；CatBoost算法；贝叶斯网络；风险推理

[中图分类号] R54；R743 [文献标志码] A

CatBoost algorithm and Bayesian network model analysis based on risk prediction of cardiovascular and cerebro vascular diseases

WANG Aimin[1], WANG Fenglin[1], HUANG Yiming[1], XU Yaqi[1], ZHANG Wenjing[1], CONG Xianzhu[1], SU Weiqiang[1], WANG Suzhen[1], GAO Mengyao[1], LI Shuang[1], KONG Yujia[1], SHI Fuyan[1], TAO Enxue[2]

(1. Department of Health Statistics, School of Public Health, Shandong Second Medical University, Weifang 261053, China; 2.School of Basic Medical Sciences, Shandong Second Medical University, Weifang 261053, China)

[收稿日期] 2023-09-02

[基金项目] 国家自然科学基金项目（81803337，81872719，82003560）；国家统计局科研项目（2018LY79）；山东省科技厅自然科学基金项目（ZR2019MH034，ZR2020MH340，ZR2023MH313）；山东省教育厅高等学校青创人才引育计划项目（2019-6-156）；潍坊医学院博士启动基金项目（2017BSQD51）

[作者简介] 王爱民（2000－），男，山东省临沂市人，在读硕士研究生，主要从事卫生统计学方面的研究。

[通信作者] 石福艳，教授，博士研究生导师（E-mail：shifuyan@126.com）；陶恩学，副主任医师（E-mail：sdwftex@163.com）

ABSTRACT **Objective**: To screen the main characteristic variables affecting the incidence of cardiovascular and cerebrovascular diseases, and to construct the Bayesian network model of cardiovascular and cerebrovascular disease incidence risk based on the top 10 characteristic variables, and to provide the reference for predicting the risk of cardiovascular and cerebrovascular disease incidence. **Methods**: From the UK Biobank Database, 315 896 participants and related variables were included. The feature selection was performed by categorical boosting (CatBoost) algorithm, and the participants were randomly divided into training set and test set in the ratio of 7∶3. A Bayesian network model was constructed based on the max-min hill-climbing (MMHC) algorithm. **Results**: The prevalence of cardiovascular and cerebrovascular diseases in this study was 28.8%. The top 10 variables selected by the CatBoost algorithm were age, body mass index (BMI), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC), the triglyceride-glucose (TyG) index, family history, apolipoprotein A/B ratio, high-density lipoprotein cholesterol (HDL-C), smoking status, and gender. The area under the receiver operating characteristic (ROC) curve (AUC) for the CatBoost training set model was 0.770, and the model accuracy was 0.764; the AUC of validation set model was 0.759 and the model accuracy was 0.763. The clinical efficacy analysis results showed that the threshold range for the training set was 0.06—0.85 and the threshold range for the validation set was 0.09—0.81. The Bayesian network model analysis results indicated that age, gender, smoking status, family history, BMI, and apolipoprotein A/B ratio were directly related to the incidence of cardiovascular and cerebrovascular diseases and they were the significant risk factors. TyG index, HDL-C, LDL-C, and TC indirectly affect the risk of cardiovascular and cerebrovascular diseases through their impact on BMI and apolipoprotein A/B ratio. **Conclusion**: Controlling BMI, apolipoprotein A/B ratio, and smoking behavior can reduce the incidence risk of cardiovascular and cerebrovascular diseases. The Bayesian network model can be used to predict the risk of cardiovascular and cerebrovascular disease incidence.

KEYWORDS Cardiovascular and cerebrovascular disease; CatBoost algorithm; Bayesian network; Risk inference

心脑血管疾病是造成全球人口死亡和残疾的主要疾病之一[1-2]。相关研究[3-7]显示：全球心脑血管疾病死亡占总死亡的32.8%，1990—2019年患病人数约增加2.52亿人，且预计心脑血管疾病在未来几十年内会增加2倍，预计到2030年其治疗费用将增加至10 440亿美元。心脑血管疾病具有高发病率和高致死率等特点[8]，因此对其进行早期发病风险预测研究具有重要意义。

目前有关心脑血管疾病发病风险预测研究多采用Cox比例风险模型、Logistic回归模型、决策树和支持向量机等机器学习模型[9-13]。但上述传统的预测模型需要假设变量之间相互对立且只能基于完整数据集来预测，不适合变量之间存在复杂关系或存在缺失数据的研究。贝叶斯网络是一种灵活的建模工具，不仅能处理变量间的复杂关系和缺失数据，而且该模型的结构和参数学习可直观展示变量之间的因果关系和相关程度[14]。目前将贝叶斯网络用于心脑血管疾病早期风险预测的研究较少。本研究拟基于英国生物样本（UK Biobank）数据库，采用类别型特征提升（categorical boosting, CatBoost）算法结合贝叶斯网络的方式构建心脑血管疾病发病风险预测模型，进而直观表达和展示心脑血管疾病发病风险和各变量间的关系及相关程度，为心脑血管疾病的早期风险预测和复杂病因研究提供参考。

1 资料与方法

1.1 资料来源 本研究数据来源于UK Biobank数据库，已获得批准的项目编号为78500。该数据库在2006—2010年招募约50万名年龄40～69岁的志愿者，并持续追访，是目前全球最大的前瞻性研究数据库[15-16]。本研究的纳入和排除标准：①选择参与并回答“医生诊断的血管/心脏问题”的研究对象；②纳入可能与心脑血管疾病发生相关的16个变量，包括年龄、性别、血液生化检查、心脑血管疾病家族史、饮食习惯、吸烟和饮酒状态等；③排除因不愿回答或未参与问卷调查的志愿者。本研究最终纳入315 896例研究对象。

1.2 数据预处理 为构建离散化贝叶斯网络，本研究对纳入的连续性变量进行了离散化处理。所有

研究对象按照年龄分为4组，分别为＜45岁、45～54岁、55～64岁和≥65岁组；根据世界卫生组织关于肥胖定义标准[17]按照体质量指数（body mass index，BMI）分为4组，分别为偏瘦（＜18.5 kg·m^{-2}）、正常（18.5～24.9 kg·m^{-2}）、超重（25～29.9 kg·m^{-2}）和肥胖（≥30 kg·m^{-2}）组；《2019欧洲心脏病学会（European Society of Cardiology，ESC）/欧洲动脉硬化学会（European Atherosclerosis Society，EAS）血脂异常管理指南》[18]指出，血脂4项在不同风险下参考值范围不同，故本研究中总胆固醇（total cholesterol，TC）、低密度脂蛋白胆固醇（low-density lipoprotein cholesterol，LDL-C）、高密度脂蛋白胆固醇（high-density lipoprotein cholesterol，HDL-C）、载脂蛋白A/B比值和甘油三酯-葡萄糖（triglyceride-glucose，TyG）指数按照四分位数分别分为4组[19]；纳入研究的饮食变量参考英国国家医疗服务体系（National Health Service，NHS）指南[20]和CHUDASAMA等[21]关于寿命期望的研究将研究对象均分为2组，分别为每日蔬菜水果摄入是否满足5份及以上（是或否）组、红肉是否每周2次及以上（是或否）和油性鱼是否每周1次及以上（是或否）。其他变量均采用UK Biobank数据库的编码分类。各变量赋值情况见表1。

表1 与心脑血管疾病发生相关的主要变量和赋值情况

Tab.1 Main variables related to occurrence of cardiovascular and cerebrovascular diseases and their assignments

Variable	Assign	Variable	Assign
Age (x_1, year)	＜45=0	TC (x_{11}, mmol·L^{-1})	＜4.91=0
	45－54=1		4.91－5.64=1
	55－64=2		5.65－6.41=2
	≥65=3		≥6.42=3
Gender (x_2)	Female=0	LDL-C (x_{12}, mmol·L^{-1})	＜2.94=0
	Male=1		2.94－3.51=1
Smoking status (x_3)	Never=0		3.52－4.11=2
	Previous=1		≥4.12=3
	Current=2	HDL-C (x_{13}, mmol·L^{-1})	＜1.17=0
Drinking status (x_4)	Never=0		1.17－1.39=1
	Previous=1		1.40－1.67=1
	Current =2		≥1.68=3
Family history (x_5)	No=0	BMI (x_{14}, kg·m^{-2})	＜18.5 = 0
	Yes=1		18.5－24.9=1
Mental illness (x_6)	No=0		25.0－29.9=2
	Yes=1		≥30.0=3
Physical activity (x_7)	Low=0	Apolipoprotein A/B ratio (x_{15})	＜8.31=0
	Moderate=1		8.31－8.67=1
	High=2		8.68－9.07=2
Vegetables/Fruits (x_8, ≥5 d^{-1})	No=0		≥9.08=3
	Yes=1	TyG index(x_{16})	＜1.23=0
Red meat (x_9, ≥2/week)	No=0		1.23－1.49=1
	Yes=1		1.50－1.82=2
Oily fish (x_{10}, ≥1/week)	No=0		≥1.83=3
	Yes=1	CCVDs	No=0
			Yes=1

CCVDs：Cardiovascular and cerebrovascular diseases.

1.3 CatBoost模型分析 CatBoost算法是一种强大的机器学习算法，可用于分类和回归任务[22]，其原理是基于梯度提升决策树（gradient boosting decision tree，GBDT）的思想，即通过迭代训练

多个弱学习器（决策树），然后将其组合成一个强学习器，其中每一棵树都试图纠正前一棵树的预测错误，进而提高模型的预测精度[23-24]。在CatBoost模型构建中，将研究对象按7∶3比例随机分为训练集和测试集。训练集构建CatBoost模型，其最优参数基于网格搜索法[25]选择，模型参数设置如下：loss_function=MultiClass，eval_metric=WKappa，od_type=Iter，early_stopping_rounds=500。Catboost模型的预测性能通过受试者工作特征（receiver operating characteristic，ROC）曲线和混淆矩阵热图评价。Catboost模型的临床效用通过决策曲线分析（decision curve analysis，DCA）法评价[26]，DCA曲线通常包含2条参考线，分别表示所有样本均预测为阴性或阳性的极端情况。模型DCA曲线高于2条参考线范围越大则模型的临床效用越好[27]。

1.4 贝叶斯网络分析 贝叶斯网络又称因果网络，可探究各节点之间的因果关系，即在已知某些节点发生的情况下，可对未知节点发生状态进行推断。贝叶斯网络通过概率图模型探索变量之间的不确定性关系，其中概率图模型由一组变量和之间的概率关系组成，结构是一个有向无环图（directed acyclic graph，DAG），其中节点表示随机变量，节点之间的有向弧表示变量间的条件依赖关系[28]，节点方框中的数值代表各节点的先验概率[29]。

1.5 统计学分析 采用R 4.2.3统计软件和Python 3.11统计软件进行统计学分析。心脑血管疾病组和非心脑血管疾病组研究对象临床资料组间比较采用χ^2检验，检验水准为$\alpha=0.05$。采用Python 3.11统计软件中catboost包构建CatBoost模型，通过网格搜索法进行参数调优。模型的ROC曲线通过sklear包绘制，混淆矩阵热图采用seaborn包绘制，DCA曲线采用matplotlib包绘制。通过R 4.2.3统计软件中的bnlearn包以最大最小爬山（max-min hill-climbing，MMHC）算法实现贝叶斯网络的结构学习，并采用Netica软件可视化贝叶斯网络和条件概率。

2 结 果

2.1 研究对象的一般情况 本探究共纳入315 896例研究对象，其中男性150 940例，女性164 956例，平均年龄56.3岁，其中234 293例有心脑血管疾病家族史，90 818例患有心脑血管疾病。本研究对非心脑血管疾病组和心脑血管疾病组研究对象的基本情况进行比较分析，除蔬菜/水果摄入量之外，其余各变量与心脑血管疾病均有关联（$P<0.01$）。见表2。

表2 非心脑血管疾病组和心脑血管疾病组研究对象临床资料

Tab. 2 Clinical data of subjects in non-cardiovascular and cerebrovascular diseases group and cardiovascular and cerebrovascular diseases group

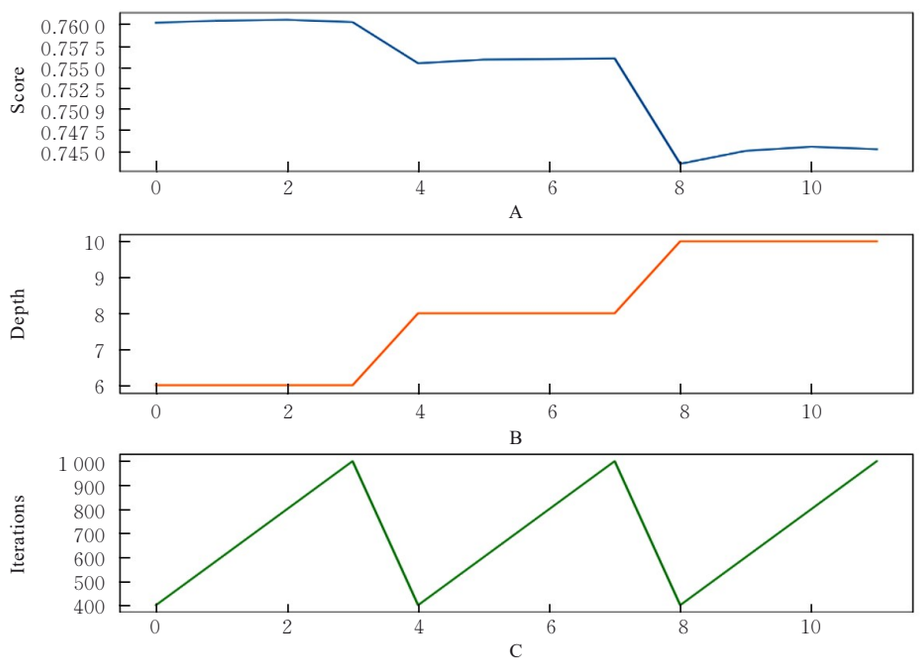
[$N=315\,896, n(\eta/\%)$]

Variable	Non-cardiovascular and cerebrovascular diseases	Cardiovascular and cerebrovascular diseases	χ^2	P
Age(year)			17 796.741	<0.01
<45	29 941(9.48)	3 762(1.19)		
45－54	74 686(23.64)	17 591(5.57)		
55－64	89 276(28.26)	43 955(13.91)		
≥65	31 175(9.87)	25 510(8.08)		
Gender			3 301.340	<0.01
Female	124 833(39.52)	40 123(12.70)		
Male	100 245(31.73)	50 695(16.05)		
Smoking status			2 065.759	<0.01
Never	129 253(40.92)	45 183(14.30)		
Previous	73 069(23.13)	37 184(11.77)		
Drinking status			393.526	<0.01
Never	8 159(2.58)	3 949(1.25)		
Previous	6 746(2.14)	3 813(1.21)		
Current	210 173(66.53)	83 056(26.29)		
Family history			4 275.264	<0.01
Yes	159 655(50.54)	74 638(23.63)		
No	65 423(20.71)	16 180(5.12)		

续表

Variable	Non-cardiovascular and cerebrovascular diseases	Cardiovascular and cerebrovascular diseases	χ^2	P
Mental illness			532.280	＜0.01
Yes	72 230(22.87)	33 027(10.46)		
No	152 848(48.39)	57 791(18.29)		
Physical activity			668.927	＜0.01
Low	39 911(12.63)	19 262(6.10)		
Moderate	91 891(29.09)	37 489(11.87)		
High	93 276(29.53)	34 067(10.78)		
Vegetables/Fruits(≥5 d^{-1})			1.471	0.225
Yes	182 024(57.62)	73 616(23.30)		
No	43 054(13.63)	17 202(5.45)		
Red meat (≥2/week)			593.884	＜0.01
Yes	72 763(23.03)	33 470(10.60)		
No	152 315(48.22)	57 348(18.15)		
Oily fish (≥1/week)			558.101	＜0.01
Yes	124 776(39.50)	54 525(17.26)		
No	100 302(31.75)	36 293(11.49)		
TC (mmol·L^{-1})			11 031.944	＜0.01
＜4.91	44 548(14.10)	33 847(10.71)		
4.91－5.64	58 769(18.60)	21 358(6.76)		
5.65－6.41	61 435(19.45)	18 473(5.85)		
≥6.42	60 326(19.10)	17 140(5.43)		
LDL-C (mmol·L^{-1})			9 949.487	＜0.01
＜2.94	44 785(14.18)	33 120(10.48)		
2.94－3.51	58 282(18.45)	21 295(6.74)		
3.52－4.11	61 140(19.35)	18 714(5.92)		
≥4.12	60 871(19.27)	17 689(5.60)		
HDL-C (mmol·L^{-1})			7 270.807	＜0.01
＜1.17	47 390(15.00)	30 458(9.64)		
1.17－1.39	55 303(17.51)	24 100(7.63)		
1.40－1.67	59 873(18.95)	20 047(6.35)		
≥1.68	62 512(19.79)	16 213(5.13)		
BMI (kg·m^{-2})			17 839.780	＜0.01
＜18.5	1 350(0.43)	203(0.06)		
18.5－24.9	87 643(27.74)	17 399(5.51)		
25.0－29.9	95 984(30.38)	39 796(12.60)		
≥30.0	40 101(12.69)	33 420(10.58)		
Apolipoprotein A/B ratio			512.794	＜0.01
＜8.31	57 986(18.36)	20 441(6.47)		
8.31－8.67	56 199(17.79)	22 458(7.11)		
8.68－9.07	56 193(17.79)	23 142(7.33)		
≥9.08	54 700(17.32)	24 777(7.84)		
TyG index			8 133.568	＜0.01
＜1.23	65 110(20.61)	15 186(4.81)		
1.23－1.49	58 735(18.59)	20 715(6.56)		
1.50－1.82	54 162(17.15)	25 197(7.98)		
≥1.83	47 071(14.90)	29 720(9.41)		

2.2 CatBoost模型参数调优结果 采用Python软件sklearn包中的GridSearchCV函数进行超参数寻优，结果见图1。当树深（depth）为6，最大迭代次数（iterations）为800时，对应得分最高，此时CatBoost模型的性能最佳。



A：Variation diagram of score；B：Variation diagram of depth；C：Variation diagram of iterations.

图1 基于网格搜索法对CatBoost模型参数寻优图

Fig. 1 Parameter optimization diagrams of CatBoost model based on grid search method

2.3 基于CatBoost模型的心脑血管疾病发病风险重要变量筛选 本研究基于训练集数据，通过最优CatBoost模型对心脑血管疾病影响因素进行筛选，结果见图2。模型特征重要性排序前10位筛选变量分别为年龄（22.40）、BMI（15.20）、LDL-C（8.84）、TC（8.55）、TyG指数（7.99）、家族史（6.65）、载脂蛋白A/B比值（4.08）、HDL-C（4.02）、吸烟状态（3.80）和性别（3.64）。

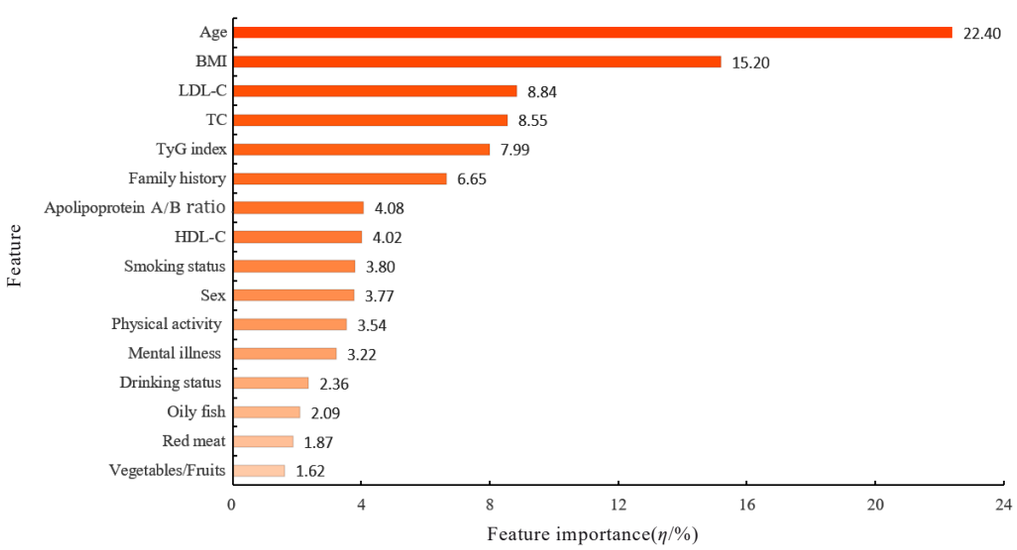
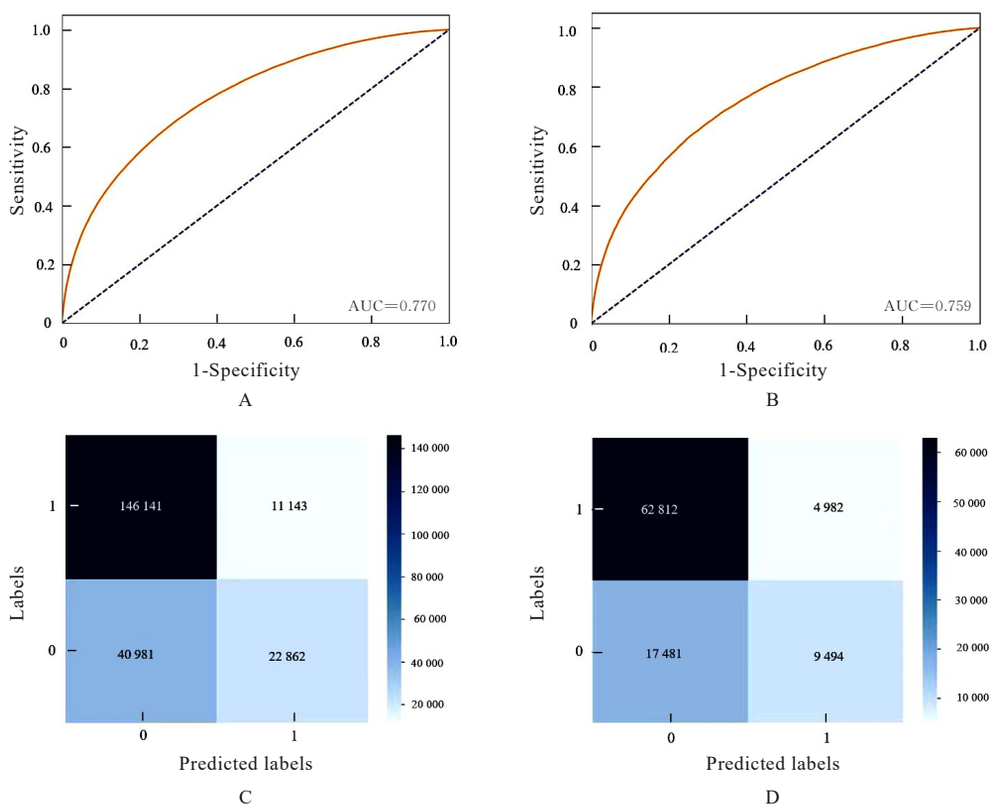


图2 基于CatBoost算法的特征变量重要性排序图

Fig. 2 Importance ranking graph of feature variables based on CatBoost algorithm

2.4 心脑血管疾病发病风险 CatBoost 模型评价

本研究分别基于训练数据集和验证数据集，以前述步骤筛选的排序前 10 位变量为参数变量，分别构建 CatBoost 模型，并绘制模型的 ROC 曲线、混淆矩阵热图和 DCA 曲线，见图 3 和 4。训练集模型的 ROC 曲线下面积（area under curve，AUC）为 0.770，模型准确性为 0.764；验证集模型 AUC 为 0.759，模型准确性为 0.763，由此可知基于 CatBoost 算法筛选的排序前 10 位变量构建的模型其预测准确度较高。临床效能分析结果显示：训练集和验证集的 DCA 曲线均具有较宽的净收益取值范围，训练阈值范围约为 0.06～0.85，验证集阈值范围约为 0.09～0.81，模型具有较好的临床实用性。



A，C：Training set；B，D：Test set.

图 3 训练集和验证集 CatBoost 模型 ROC 曲线（A，B）和预测混淆矩阵热图（C，D）

Fig. 3 CatBoost model ROC curves (A, B) and prediction confusion matrix heat maps (C, D) of training set and validation set

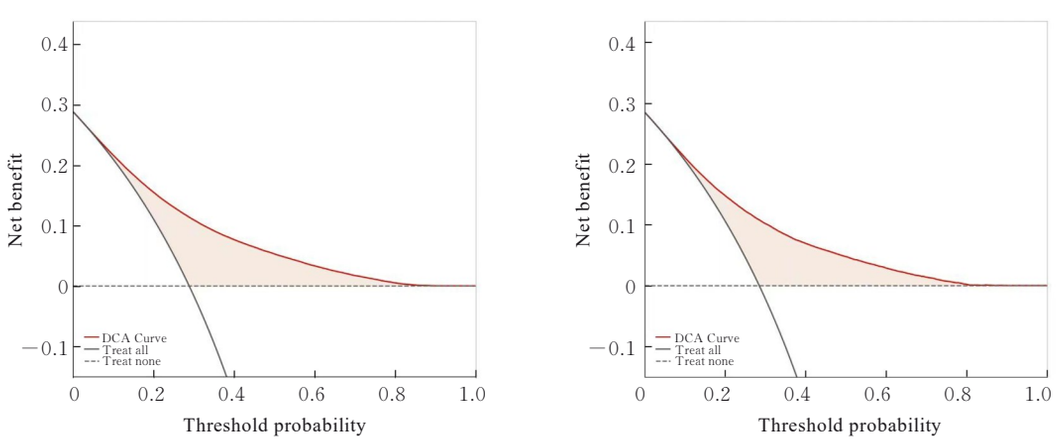
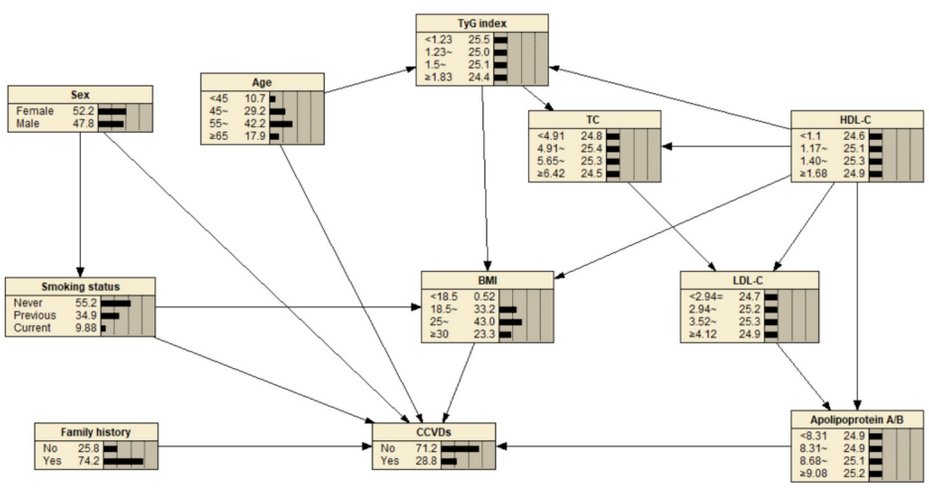


图 4 训练集（A）和验证集（B）CatBoost 模型临床效用评价 DCA 曲线图

Fig. 4 DCA curves for clinical effectiveness evaluation of CatBoost model in training set (A) and validation set (B)

2.5 心脑血管疾病发病风险贝叶斯网络模型构建 本研究采用MMHC算法和极大似然估计法，以CatBoost模型筛选的特征重要性排序前10位的变量为参数变量，构建心脑血管疾病发病风险贝叶斯网络模型（图5）。图5中节点方框的数值代表各节点的先验概率，如心脑血管疾病的患病对应数值为28.8，表示患心脑血管疾病者在总研究人群中所占比例为28.8%，为该节点的先验概率。年龄、家族史、性别、吸烟状态、载脂蛋白A/B和BMI为心脑血管疾病的父节点，直接影响心脑血管疾病的患病概率，TyG指数、TC、HDL-C和LDL-C通过影响BMI和载脂蛋白A/B比值进而间接影响心脑血管疾病患病概率。

在本研究中，当其他节点不设置先验条件时，随着年龄和载脂蛋白A/B比值检测值增加，心脑血管疾病的患病风险也逐渐增加，男性和吸烟行为会增加心脑血管疾病的患病风险，有心脑血管疾病家族史的研究对象较无心脑血管疾病家族史的研究对象患病风险增加11.7%，偏瘦、超重和肥胖人群心脑血管疾病的患病风险均高于体质量正常的人群。



CCVDs：Cardiovascular and cerebrovascular diseases.

图5 基于MMHC算法的心脑血管疾病发病风险贝叶斯网络图

Fig. 5 Bayesian network diagram of cardiovascular and cerebrovascular diseases risk based on MMHC algorithm

2.6 不同个体心脑血管疾病贝叶斯网络发病风险预测 贝叶斯网络可通过有向弧和条件概率进行风险预测，即在已知某些节点发生的情况下，预测未知节点的状态。假定某个体有家族史，可推测其患心脑血管疾病的患病风险为31.9%（表3）；年龄超过65岁且具备家族史，可推测心脑血管疾病的患病风险为47.1%（图6A）；在以上条件下，当个体BMI正常、无吸烟行为且载脂蛋白A/B比值<8.31时，心脑血管疾病的患病风险为24.5%（图6B）。

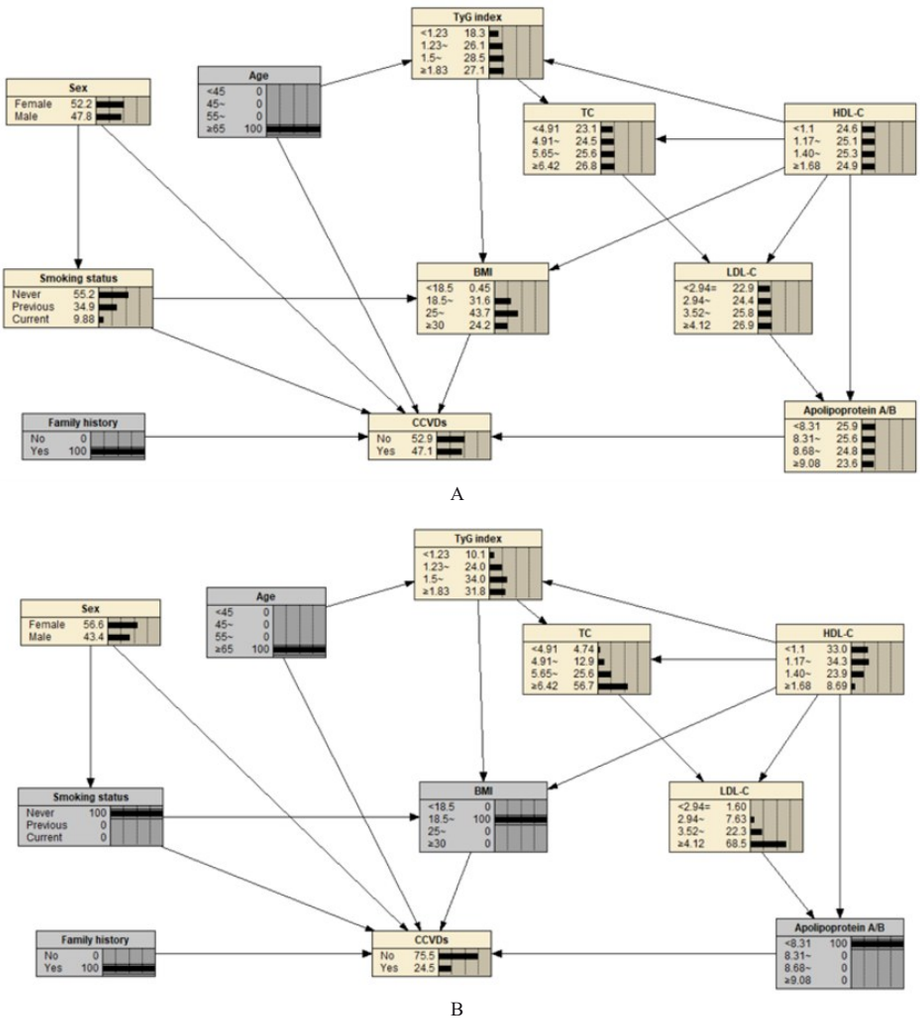
表3 心脑血管疾病独立父节点的条件概率表

Tab. 3 Conditional probability table of independent parent node of cardiovascular and cerebrovascular diseases （η/%）

Variable	Risk	Variable	Risk
Age(year)		Gender	
<45	12.5	Female	25.0
45－54	19.9	Male	33.1
55－64	33.0	BMI (kg·m^{-2})	
≥65	43.4	<18.5	22.6
Family history		18.5－24.9	18.0
Yes	31.9	25.0－29.9	28.7
No	20.2	≥30.0	44.7
Smoking status		Apolipoprotein A/B	
Never	27.1	<8.31	26.7
Previous	31.3	8.31－8.67	28.3
Current	29.9	8.68－9.07	28.6
		≥9.08	31.6

3 讨 论

本研究首先通过CatBoost算法进行心脑血管疾病特征选择，并以CatBoost模型中特征值排序前10位的变量构建贝叶斯网络模型，结果显示：CatBoost算法确定的排序前10位特征变量为年龄、BMI、LDL-C、TC、TyG指数、家族史、载脂蛋白A/B比值、HDL-C、吸烟状态和性别；在贝叶



A：When the individual was over 65 years of age and had a family history；B：When the individual had normal weight，no smoking behavior and apolipoprotein A/B＜8.31. CCVDs：Cardiovascular and cerebrovascular diseases.

图 6 个体心脑血管疾病发病的风险预测

Fig. 6 Risk prediction of incidences of individual cardiovascular and cerebrovascular diseases

斯网络模型结构中显示：年龄、性别、吸烟状态、家族史、BMI和载脂蛋白A/B比值与心脑血管疾病间存在有向弧，表示与心脑血管疾病直接相关，而TyG指数、HDL-C、LDL-C和TC与心脑血管疾病为间接相关。

本研究中患心脑血管疾病的先验概率为28.8%，当某个体有家族史时，其心脑血管疾病的发病风险概率为31.9%，风险提高了3.1%；如果某个人年龄超过65岁并且有心脑血管疾病家族史，则其心脑血管疾病的风险概率为47.1%，其患病风险大幅提高。本研究结果提示：具有心脑血管疾病家族史的人群随着年龄的增加应及时体检，进而做到心脑血管疾病的三早防治；如果BMI正常、无吸烟行为且载脂蛋白A/B比值＜8.31时，则其心脑血管疾病的风险概率为24.5%，提示平时可以通过饮食和体育锻炼等方式控制BMI、载脂蛋白A和载脂蛋白B水平，减少吸烟，进而可降低心脑血管疾病的患病风险。

综上所述，本研究构建的贝叶斯网络可以直观表达心脑血管疾病发病风险与各变量间的关系及相关程度，即使变量之间存在复杂关系或缺失医学检测数据，仍可以通过已知变量进行心脑血管疾病发病风险推断，本研究结果为心脑血管疾病的防控提供了一定的实践依据。但本研究也存在一定的局限性：在构建心脑血管疾病贝叶斯网络模型时，主要采用了横断面分类变量，构建的模型为静态贝叶斯网络模型，该模型未能反映各变量的纵向动态变化情况，后期在条件许可的情况下，可考虑构建心

脑血管疾病动态贝叶斯网络；采用的MMHC算法容易陷入局部最优[29]，因此为提高贝叶斯网络模型的普适性，可通过禁忌搜索等算法综合比较进而构建最优贝叶斯网络模型。

利益冲突声明：

所有作者声明不存在利益冲突。

作者贡献声明：

王爱民参与研究选题、研究设计、数据分析、论文撰写和修改，王凤琳、黄一铭、徐雅琪、张文婧、丛显铸、苏维强、高梦瑶和李爽参与数据整理和图表编辑，王素珍、孔雨佳、石福艳和陶恩学参与研究设计和论文修改及审校。

[参考文献]

[1] REN Q Q, LI S Y, XIAO C L, et al. The impact of air pollution on hospitalization for cardiovascular and cerebrovascular disease in Shenyang, China[J]. Iran J Public Health, 2020, 49(8): 1476-1484.

[2] YOU Q, SHAO X Y, WANG J P, et al. Progress on physical field-regulated micro/nanomotors for cardiovascular and cerebrovascular disease treatment[J]. Small Methods, 2023, 7(10): e2300426.

[3] BENJAMIN E J, MUNTNER P, ALONSO A, et al. Heart disease and stroke statistics-2019 update: a report from the American heart association[J]. Circulation, 2019, 139(10): e56-e528.

[4] MENSAH G A, ROTH G A, FUSTER V. The global burden of cardiovascular diseases and risk factors: 2020 and beyond[J]. J Am Coll Cardiol, 2019, 74(20): 2529-2532.

[5] DISEASES AND INJURIES COLLABORATORSGBD. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019[J]. Lancet, 2020, 396(10258): 1204-1222.

[6] MELA A, RDZANEK E, PONIATOWSKI Ł A, et al. Economic costs of cardiovascular diseases in Poland estimates for 2015-2017 years[J]. Front Pharmacol, 2020, 11: 1231.

[7] QIAO W J, ZHANG X Y, KAN B, et al. Hypertension, BMI, and cardiovascular and cerebrovascular diseases[J]. Open Med, 2021, 16(1): 149-155.

[8] STROKE COLLABORATORSGBD. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019[J]. Lancet Neurol, 2021, 20(10): 795-820.

[9] BOYD C, BROWN G, KLEINIG T, et al. Machine learning quantitation of cardiovascular and cerebrovascular disease: a systematic review of clinical applications[J]. Diagnostics, 2021, 11(3): 551.

[10] CHEUNG C Y, XU D J, CHENG C Y, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre[J]. Nat Biomed Eng, 2021, 5(6): 498-508.

[11] AZMI J, ARIF M, NAFIS M T, et al. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data[J]. Med Eng Phys, 2022, 105: 103825.

[12] KELSHIKER M A, SELIGMAN H, HOWARD J P, et al. Coronary flow reserve and cardiovascular outcomes: a systematic review and meta-analysis[J]. Eur Heart J, 2022, 43(16): 1582-1593.

[13] ZHENG P F, CHEN L Z, LIU P, et al. Identification of immune-related key genes in the peripheral blood of ischaemic stroke patients using a weighted gene coexpression network analysis and machine learning[J]. J Transl Med, 2022, 20(1): 361.

[14] BIEDERMANN A, TARONI F. Bayesian networks and probabilistic reasoning about scientific evidence when there is a lack of data[J]. Forensic Sci Int, 2006, 157(2/3): 163-167.

[15] BYCROFT C, FREEMAN C, PETKOVA D, et al. The UK Biobank resource with deep phenotyping and genomic data[J]. Nature, 2018, 562(7726): 203-209.

[16] 黄夏璇，黄 韬，杨 瑞，等. UK Biobank数据的应用介绍[J]. 中国循证医学杂志，2022，22(9)：1099-1107.

[17] UCHAI S, ANDERSEN L F, THORESEN M, et al. Does the association between adiposity measures and prefrailty among older adults vary by social position? Findings from the Tromsø study 2015/2016[J]. BMC Public Health, 2024, 24(1): 1457.

[18] MACH F, BAIGENT C, CATAPANO A L, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk[J]. Eur Heart J, 2020, 41(1): 111-188.

[19] PIZZI N J. Fuzzy quartile encoding as a preprocessing method for biomedical pattern classification[J]. Theor Comput Sci, 2011, 412(42): 5909-5925.

[20] JAYAWARDENA R, SOORIYAARACHCHI P. The inside story of fruits; exploring the truth behind conventional theories[J]. Diabetes Metab Syndr, 2021, 15(6): 102085.

[21] CHUDASAMA Y V, KHUNTI K K, ZACCARDI F, et al. Physical activity, multimorbidity, and life expectancy: a UK Biobank longitudinal study[J]. BMC Med, 2019, 17(1): 108.

[22] 苗丰顺, 李 岩, 高 岑, 等. 基于CatBoost算法的糖尿病预测方法[J]. 计算机系统应用, 2019, 28(9): 215-218.

[23] HANCOCK J T, KHOSHGOFTAAR T M. CatBoost for big data: an interdisciplinary review[J]. J Big Data, 2020, 7(1): 94.

[24] 胡建锦, 熊 伟, 方陆明, 等. 基于距离相关系数和Catboost方法的森林蓄积量估测[J]. 中南林业科技大学学报, 2023, 43(5): 27-35.

[25] PRABU S, THIYANESWARAN B, SUJATHA M, et al. Grid search for predicting coronary heart disease by tuning hyper-parameters[J]. Comput Syst Sci Eng, 2022, 43(2): 737-749.

[26] ROUSSON V, ZUMBRUNN T. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies[J]. BMC Med Inform Decis Mak, 2011, 11: 45.

[27] 唐 末. 基于循证医学及机器学习的中医药影响早中期结直肠癌预后模型研究[D]. 北京: 中国中医科学院, 2022.

[28] 钟 璐, 薛付忠. 基于贝叶斯网络不确定性推理的肺癌风险预测模型[J]. 山东大学学报(医学版), 2023, 61(4): 86-94.

[29] 王旭春, 宋伟梅, 潘金花, 等. MMPC-Tabu混合算法的贝叶斯网络模型在高脂血症相关因素研究中的应用[J]. 中国卫生统计, 2022, 39(3): 345-350, 355.