Statistics and Application 统计学与应用, 2026, 15(6), 34-43
Published Online June 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.156128

Hans汉斯

基于CHARLS数据库的高血压风险预测模型

冯研博，张　心，张亚宁，杨佳雪，冯玉鑫，韩　影*

华北理工大学护理与康复学院，河北　唐山

收稿日期：2026年5月12日；录用日期：2026年6月4日；发布日期：2026年6月15日

摘　要

背景：高血压是我国中老年人群常见慢性病之一，可显著增加心脑血管疾病、肾脏损害及死亡风险。基于大样本人群数据构建简便可用的高血压患病风险预测模型，有助于早期识别高风险个体并优化社区筛查。目的：基于中国健康与养老追踪调查(China Health and Retirement Longitudinal Study, CHARLS) 2020年随访数据，分析45岁及以上人群高血压相关因素，并构建高血压风险预测列线图模型。方法：提取中国健康与养老追踪调查(CHARLS)数据库2020年随访数据中45岁及以上中老年人群的社会人口学及健康相关数据。经数据清洗后共纳入19,218例，按7:3比例进行分层随机抽样，分为训练集与验证集。基于训练集进行单因素分析、多因素二元Logistic回归分析，筛选独立影响因素；采用R 4.2.1软件“rms”包构建列线图预测模型，并通过Bootstrap重抽样法进行内部验证，在验证集中进行外部验证。结果：多因素Logistic回归显示，年龄增加、女性、当前饮酒、当前吸烟、锻炼以及教育水平与高血压患病风险相关，婚姻状况在多因素模型中未达到统计学显著性。模型在训练集中的AUC/C指数为0.639 (95% CI: 0.629~0.648)，在验证集中的AUC/C指数为0.637 (95% CI: 0.622~0.651)；校准曲线显示预测风险与实际观察风险总体趋势一致，但部分风险区间仍存在偏差。结论：基于CHARLS 2020年数据构建的高血压风险预测模型具有一定区分度，可作为45岁及以上人群高血压风险初步筛查的参考工具。模型性能仍有限，未来需进一步纳入BMI、血脂、糖尿病、地区、饮食及体检指标等变量，并在独立人群中进行外部验证。

关键词

高血压，风险预测，列线图，分层随机抽样

Hypertension Risk Prediction Model Based on the CHARLS Database

Yanbo Feng, Xin Zhang, Yaning Zhang, Jiaxue Yang, Yuxin Feng, Ying Han*

College of Nursing and Rehabilitation, North China University of Science and Technology, Tangshan Hebei

Received: May 12, 2026; accepted: June 4, 2026; published: June 15, 2026

*通讯作者。

文章引用: 冯研博, 张心, 张亚宁, 杨佳雪, 冯玉鑫, 韩影. 基于 CHARLS 数据库的高血压风险预测模型[J]. 统计学与应用, 2026, 15(6): 34-43. DOI: 10.12677/sa.2026.156128

Abstract

Background: Hypertension is one of the common chronic diseases among the elderly in China, significantly increasing the risks of cardiovascular and cerebrovascular diseases, kidney damage, and death. Building a simple and applicable hypertension risk prediction model based on large sample population data is helpful for early identification of high-risk individuals and optimizing community screening. Objective: Based on the 2020 follow-up data from the China Health and Retirement Longitudinal Study (CHARLS), this study aims to analyze the factors related to hypertension among individuals aged 45 and above, and to construct a risk prediction nomogram model for hypertension. Method: Social demographic and health-related data of the elderly population aged 45 and above from the 2020 follow-up data of the Chinese Health and Retirement Longitudinal Study (CHARLS) database were extracted. After data cleaning, a total of 19,218 cases were included. Stratified random sampling was conducted at a ratio of 7:3 to divide them into a training set and a validation set. Univariate analysis and multivariate binary Logistic regression analysis were performed based on the training set to identify independent influencing factors. A nomogram prediction model was constructed using the "rms" package of R 4.2.1 software, and internal validation was conducted using the Bootstrap resampling method. External validation was performed in the validation set. Result: The multivariate Logistic regression analysis showed that age increase, female gender, current alcohol consumption, current smoking, exercise, and educational level were associated with the risk of hypertension. Marital status did not reach statistical significance in the multivariate model. The AUC/C index of the model in the training set was 0.639 (95% CI: 0.629~0.648), and in the validation set was 0.637 (95% CI: 0.622~0.651); The calibration curve showed that the predicted risk was consistent with the actual observed risk in general trend, but there was still deviation in some risk intervals. Conclusion: The hypertension risk prediction model constructed based on the 2020 data of CHARLS has a certain degree of discrimination and can be used as a reference tool for the preliminary screening of hypertension risk among people aged 45 and above. The model performance is still limited. In the future, variables such as BMI, blood lipids, diabetes, region, diet, and physical examination indicators need to be further included, and external validation should be conducted in independent populations.

Keywords

Hypertension, Risk Prediction, Tree Diagram, Stratified Random Sampling

1. 引言

高血压[1] (Hypertension)是全球范围内最常见的慢性疾病之一，也是心脑血管疾病、肾脏病变及死亡的主要危险因素。根据世界卫生组织(WHO) 2023 年发布的报告[2]，全球高血压患者人数已超过 13 亿，其中近半数患者未得到有效诊断或控制。尽管近年来诊疗技术不断进步，高血压仍是公共卫生领域的重大挑战，其防治现状与临床需求之间仍存在显著差距。

当前，高血压的防治体系已逐步完善。在诊断方面，动态血压监测(ABPM)和家庭血压监测(HBPM)的普及提高了早期检出率；在治疗方面，降压药物(如 ACEI、ARB、CCB 等)的联合应用显著降低了患者的心血管事件风险[3]。此外，国际指南[4] (如 ISH 2020、ACC/AHA 2017)对血压分级和管理策略的细化，进一步规范了临床实践。近年来，研究还发现肠道菌群、免疫炎症等新机制可能参与高血压的发病，为

靶向治疗提供了潜在方向。

然而，高血压的防控仍面临多重瓶颈：诊断局限性：现行标准(如诊室血压 ≥140/90 mmHg)可能忽视隐匿性高血压或夜间高血压患者，部分人群依赖单一测量导致误诊或漏诊；治疗依从性差：约 50%患者因药物副作用、经济负担或无症状性而停药，长期控制率不足 20% [5]；个体化治疗缺失：现有指南对特殊人群(如老年人、合并代谢综合征者)的差异化方案缺乏循证依据；机制研究不足：高血压的异质性机制(如盐敏感性、交感神经过度激活)尚未完全阐明，制约了精准医学的发展。

中老年人群是高血压患病的主体人群。鉴于社区基层医疗资源有限，开发一种仅基于社会人口学信息、无需实验室检查的简易风险预测工具，对于中老年人群的高血压早期识别和分层管理具有现实意义。因此，本研究聚焦 45 岁及以上中老年人群，旨在构建并验证一种基于社会人口学因素的高血压风险预测列线图模型，以期为社区筛查提供简便易行的参考。

2. 资料和方法

2.1. 设计

回顾性观察性研究，基于公开数据库的预测模型构建与验证研究。影响因素分析采用单因素结合多因素 Logistic 回归，模型构建采用列线图法。

2.2. 数据来源

研究对象为 CHARLS 2020 年随访中年龄 45~99 岁、结局变量及核心预测变量信息完整者。结局变量为是否患有高血压。候选预测变量包括年龄、性别、婚姻状况、当前饮酒、当前吸烟、锻炼及教育水平。年龄以连续变量纳入模型；性别编码为男、女；婚姻状况分为已婚和未婚/其他；当前饮酒、当前吸烟和锻炼均按“是/否”分类；教育水平按照原始数据可用编码分为 1~4 级。

2.3. 入选标准

纳入标准：① CHARLS 2020 年随访记录；② 年龄 45~99 岁；③ 高血压诊断状态明确；④ 年龄、性别、婚姻、当前饮酒、当前吸烟、锻炼和教育水平等核心变量完整。排除标准：① 非 2020 年随访记录；② 结局变量或核心预测变量缺失；③ 年龄不在 45~99 岁范围内；④ 核心分类变量取值异常。

数据清洗流程如下：原始上传数据 96,628 条，筛选 2020 年随访记录后 19,395 条；删除结局及核心预测变量缺失记录后 19,349 条；进一步限制年龄 45~99 岁并校验核心变量取值后，最终纳入 19,218 例用于统计分析和模型构建。

2.4. 主要观察指标

模型的评价：区分度评估采用 C 指数(AUC)及 95% CI，一致性评估采用校准曲线及 Hosmer-Lemeshow 检验。

模型的内部验证：Bootstrap 重抽样法(1000 次)，报告经乐观校正(optimism-corrected)后的 C 指数及校准曲线。

模型的外部验证：模型在验证集的 C 指数及校准曲线及 Brier 评分评估模型泛化能力。

2.5. 统计学分析

采用 SPSS 27.0 和 R 4.2.1 软件进行统计分析，检验水准 $\alpha = 0.05$。

(1) 使用 SPSS 27.0 软件将站立时间、饮酒次数转换为有序分类变量，性别转换为无序分类变量。计量资料用均数 ± 方差进行统计描述；计数资料用例数和构成比进行统计描述(n/%)。

(2) 模型变量筛选：采用 SPSS 27.0 软件对训练集进行单因素分析，再筛选出单因素分析具有统计学意义的变量进行多因素分析，均以 $P < 0.05$ 为有统计学意义的筛选变量。

(3) 列线图模型构建：将二元 Logistic 回归分析筛选出的独立影响变量导入 Rstudio 软件，使用“rms”包构建代谢综合征列线图预测模型。

(4) 模型的评价：采用受试者工作特征(ROC)曲线及 C 指数评估。采用 Bootstrap 重抽样法(B = 1000 次，有放回抽样)在训练集中生成新样本，计算 optimism-corrected C 指数，并绘制校正后的校准曲线。在验证集中使用完整模型的预测概率计算 C 指数及校准曲线，评价模型泛化能力。

3. 结果

3.1. 高血压检出状况

共纳入 19,218 名 45 岁及以上中老年人，按照高血压诊断标准，其中非高血压 11,513 例(59.91%)，高血压 7705 例(40.09%)。详见表 1。

Table 1. Stratified sampling results of the total sample and training set, validation set
表 1. 总样本及训练集、验证集分层抽样结果

数据集	n	非高血压	高血压
总体	19,218	11,513 (59.91)	7705 (40.09)
训练集	13,452	8059 (59.91)	5393 (40.09)
验证集	5766	3454 (59.90)	2312 (40.10)

3.2. 训练集与验证集基线资料比较

Table 2. Comparison of baseline data between training set and validation set
表 2. 训练集与验证集基线资料比较

变量	分组	训练集(n = 13452)	验证集(n = 5766)	P 值
年龄，岁	均数 ± 标准差	63.49 ± 9.92	63.43 ± 9.90	0.725
性别	男	7123 (52.95)	3040 (52.72)	
	女	6329 (47.05)	2726 (47.28)	0.783
婚姻	未婚/其他	2176 (16.18)	920 (15.96)	
	已婚	11,276 (83.82)	4846 (84.04)	0.719
现在饮酒	否	8638 (64.21)	3687 (63.94)	
	是	4814 (35.79)	2079 (36.06)	0.733
现在吸烟	否	10,044 (74.67)	4277 (74.18)	
	是	3408 (25.33)	1489 (25.82)	0.487
锻炼	否	1541 (11.46)	646 (11.20)	
	是	11,911 (88.54)	5120 (88.80)	0.632
教育水平	水平 1	5841 (43.42)	2440 (42.32)	0.502
教育水平	水平 2	2919 (21.70)	1294 (22.44)	
教育水平	水平 3	2954 (21.96)	1285 (22.29)	
教育水平	水平 4	1738 (12.92)	747 (12.96)	

采用按高血压结局分层的 7:3 随机抽样后，训练集和验证集的高血压比例分别为 40.09%和 40.10%。两组在年龄、性别、婚姻、当前饮酒、当前吸烟、锻炼及教育水平方面差异均无统计学意义，提示分层随机抽样后两组基线分布较为均衡。详见表 2。

3.3. 高血压组与非高血压组基线资料比较

与非高血压组相比，高血压组年龄更高；两组在婚姻状况、当前饮酒、当前吸烟、锻炼及教育水平方面差异具有统计学意义($P < 0.05$)；性别分布差异无统计学意义。详见表 3。

Table 3. Comparison of baseline data between the hypertension group and the non-hypertension group
表 3. 高血压组与非高血压组基线资料比较

变量	分组	非高血压组(n = 11513)	高血压组(n = 7705)	P 值
年龄，岁	均数±标准差	61.70 ± 9.64	66.11 ± 9.73	<0.001
性别	男	6060 (52.64)	4103 (53.25)	
	女	5453 (47.36)	3602 (46.75)	0.411
婚姻	未婚/其他	1544 (13.41)	1552 (20.14)	
	已婚	9969 (86.59)	6153 (79.86)	<0.001
现在饮酒	否	7134 (61.96)	5191 (67.37)	
	是	4379 (38.04)	2514 (32.63)	<0.001
现在吸烟	否	8356 (72.58)	5965 (77.42)	
	是	3157 (27.42)	1740 (22.58)	<0.001
锻炼	否	1119 (9.72)	1068 (13.86)	
	是	10,394 (90.28)	6637 (86.14)	<0.001
教育水平	水平 1	4729 (41.08)	3552 (46.10)	<0.001
	水平 2	2551 (22.16)	1662 (21.57)	
	水平 3	2674 (23.23)	1565 (20.31)	
	水平 4	1559 (13.54)	926 (12.02)	

3.4. 训练集单因素 Logistic 回归分析

Table 4. Univariate Logistic regression analysis of hypertension-related factors in the training set
表 4. 训练集高血压相关因素的单因素 Logistic 回归分析

变量	B	SE	Wald χ^2	P 值	OR	95% CI
年龄	0.046	0.002	620.996	<0.001	1.047	1.044~1.051
性别(女 vs 男)	−0.014	0.035	0.160	0.689	0.986	0.920~1.057
婚姻(已婚 vs 未婚/其他)	−0.489	0.047	107.834	<0.001	0.613	0.559~0.673
现在饮酒(是 vs 否)	−0.244	0.037	43.547	<0.001	0.783	0.728~0.842
现在吸烟(是 vs 否)	−0.233	0.041	32.135	<0.001	0.792	0.731~0.859
锻炼(是 vs 否)	−0.410	0.054	56.887	<0.001	0.664	0.597~0.738
教育水平 2 vs 1	−0.149	0.046	10.443	0.001	0.861	0.787~0.943
教育水平 3 vs 1	−0.201	0.046	18.790	<0.001	0.818	0.747~0.896
教育水平 4 vs 1	−0.226	0.056	16.085	<0.001	0.798	0.715~0.891

训练集单因素 Logistic 回归结果显示，年龄、婚姻、当前饮酒、当前吸烟、锻炼和教育水平与高血压患病状态相关，性别在单因素分析中差异无统计学意义。详见表 4。

3.5. 多因素 Logistic 回归分析及预测模型构建

将年龄、性别、婚姻状况、当前饮酒、当前吸烟、锻炼及教育水平纳入多因素 Logistic 回归模型。结果显示，年龄增加与高血压患病风险升高相关；女性、当前饮酒、当前吸烟、锻炼及教育水平 3 与高血压患病状态相关；婚姻状况在多因素模型中未达到统计学显著性。需要注意的是，当前饮酒和当前吸烟在本模型中表现为 OR < 1，可能与横断面数据中的反向因果、患病后行为改变或残余混杂有关，不宜解释为保护因素。详见表 5。

Table 5. Multivariate Logistic regression analysis of hypertension-related factors in the training set
表 5. 训练集高血压相关因素的多因素 Logistic 回归分析

变量	B	SE	Wald χ^2	P 值	OR	95% CI
年龄(连续，岁)	0.044	0.002	450.436	<0.001	1.045	1.041~1.049
性别(女 vs 男)	0.113	0.047	5.834	0.016	1.119	1.021~1.226
婚姻(已婚 vs 未婚/其他)	−0.094	0.052	3.247	0.072	0.910	0.822~1.008
现在饮酒(是 vs 否)	−0.142	0.042	11.153	<0.001	0.868	0.799~0.943
现在吸烟(是 vs 否)	−0.215	0.049	19.062	<0.001	0.807	0.733~0.888
锻炼(是 vs 否)	−0.192	0.057	11.339	<0.001	0.825	0.738~0.923
教育水平 2 vs 1	0.014	0.049	0.085	0.771	1.014	0.922~1.116
教育水平 3 vs 1	0.107	0.050	4.521	0.033	1.113	1.008~1.228
教育水平 4 vs 1	0.039	0.060	0.411	0.521	1.039	0.924~1.169

3.6. 高血压风险预测模型列线图

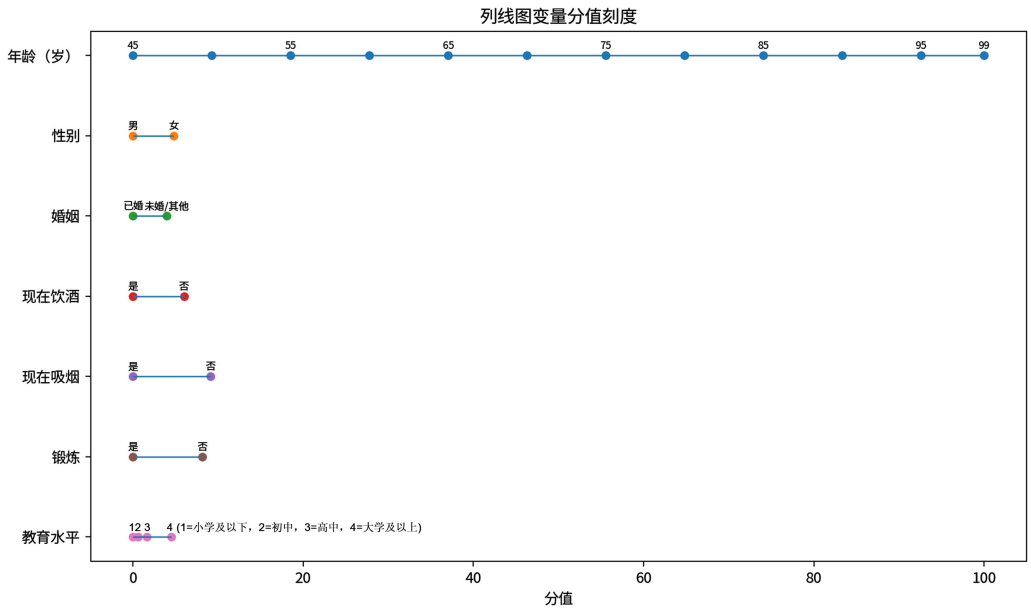

Figure 1. Nomogram for hypertension prediction model
图 1. 高血压预测模型列线图

使用 Rstudio 中的“rms”包，根据多因素分析筛选出的变量构建高血压的列线图预测模型，见图 1。列线图的应用如下：根据列线图，得出个体每个预测指标对应的分数值(Points)，得出各分数和的总分(Total Points)后，与总分相对应的预测概率为中老年人患有高血压的概率。该列线图说明了年龄、婚姻、教育、性别等独立影响因素对高血压的预测效能，根据各危险因素得分总和，可得出患高血压的概率，详见图 2。

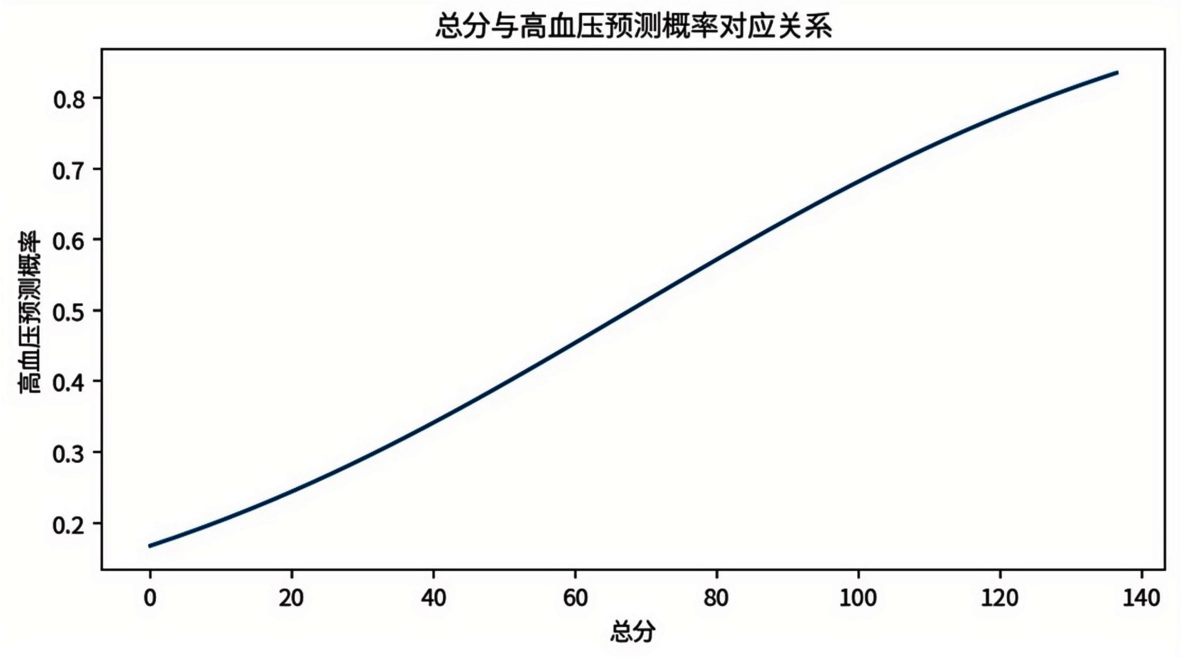

Figure 2. Correlation between the total score of the line chart and the probability of hypertension prediction
图 2. 列线图总分与高血压预测概率对应关系

3.7. 模型区分度与校准度评价

模型在训练集中的 AUC/C 指数为 0.639 (95% CI: 0.629~0.648)，在验证集中的 AUC/C 指数为 0.637 (95% CI: 0.622~0.651)。训练集和验证集 AUC 接近，提示模型在验证集中的区分度未出现明显下降，但总体区分度仍属于有限水平，详见图 3、图 4。校准曲线显示预测风险与实际观察风险总体趋势一致；验证集校准斜率为 0.989，接近 1，但 Hosmer-Lemeshow 检验 $P < 0.05$，提示在大样本条件下部分风险分层仍存在校准偏差。详见表 6。

Table 6. Evaluation metrics of training set and validation set models
表 6. 训练集与验证集模型评价指标

数据集	样本量	高血压例数(%)	AUC/C 指数(95% CI)	Brier 分数	校准截距	校准斜率	P 值
训练集	13,452	5393 (40.09)	0.639 (0.629~0.648)	0.228	0.000	1.000	<0.001
验证集	5766	2312 (40.10)	0.637 (0.622~0.651)	0.228	−0.000	0.989	0.010

4. 讨论 Discussion

本研究基于 CHARLS 数据库构建了 45 岁以上高血压风险预测列线图模型，探讨了年龄、性别、婚姻、教育等社会人口学因素与高血压的关系。结果显示，模型在训练集和验证集中的 AUC 分别为 0.639 和 0.637，说明模型具有一定区分能力，但尚不足以作为独立诊断工具。

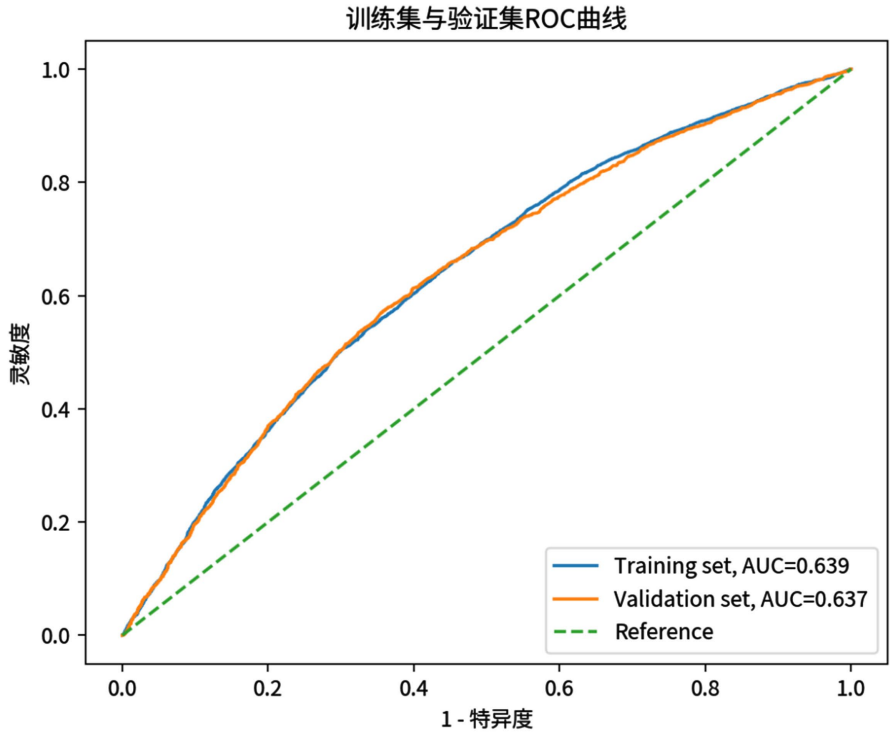

Figure 3. ROC curves of training set and validation set
图 3. 训练集与验证集 ROC 曲线

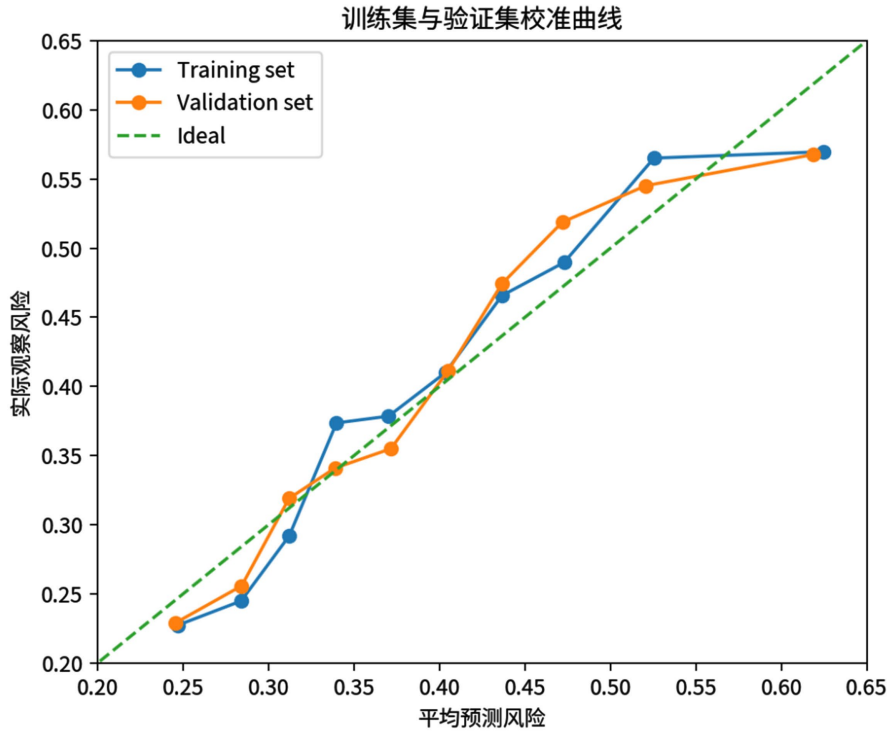

Figure 4. Calibration curves of the training set and validation set
图 4. 训练集与验证集校准曲线

多因素分析显示，年龄是高血压患病状态最稳定的相关因素。随着年龄增加，血管弹性下降、动脉硬化进展以及代谢异常累积均可能增加高血压发生风险[6]。本研究中女性在多因素模型中的 OR 略高于男性，提示在控制年龄、婚姻及生活方式等因素后，性别差异仍可能影响高血压风险，但该结果仍需结合绝经状态、肥胖、激素水平和生活方式等更多变量进一步验证[7]。

本模型的训练集和验证集 AUC 均约为 0.64，提示模型区分度有限。与包含 BMI、血脂、血糖、血压实测值、家族史和地区环境因素的预测模型相比，本研究主要使用人口学和生活方式变量，变量获取简便但信息量有限，这可能是模型区分度不高的重要原因。校准曲线显示预测风险与实际观察风险总体趋势一致，说明模型可用于粗略风险分层，但在具体个体层面的风险概率解释仍需谨慎。

尽管模型性能有限，本研究仍具有一定应用价值。模型所需变量易于获得，适合在社区初筛或健康管理场景中作为低成本风险评估工具。对于高分个体，可建议进一步进行标准化血压测量、生活方式评估和必要的临床检查。

4.1. 研究局限与改进方向

本研究存在以下局限。第一，本研究基于横断面数据，只能用于患病风险预测，不能推断各变量与高血压之间的因果关系。第二，本次模型未纳入 BMI、腰围、血脂、血糖、糖尿病、饮食、睡眠、城乡地区及环境暴露等重要变量，可能限制模型区分度。第三，当前饮酒、当前吸烟和锻炼等变量较为粗略，未能反映剂量、持续时间及累积暴露。第四，验证集来自同一数据库的随机划分样本，属于内部验证或拆分样本验证，尚不能替代独立外部验证。第五，最终可分析数据中未包含志愿者活动或慈善活动变量，因此修订模型删除了该变量，后续如能获取完整社交活动数据，可进一步探索其与高血压风险的关系。

4.2. 临床与公共卫生应用价值

尽管存在局限性，本模型为高血压早期筛查提供了低成本、易操作的工具。例如，社区医生可通过列线图快速识别高风险个体，优先进行健康干预(如压力管理、生活方式指导)。此外，模型提示社会行为因素的重要性，呼吁公共卫生政策关注教育公平与家庭支持对慢性病防控的潜在影响。

4.3. 结论

本研究基于 CHARLS 2020 年数据，纳入 19,218 例 45 岁及以上人群，采用 7:3 分层随机抽样建立训练集和验证集，并构建了高血压风险预测模型。模型在训练集和验证集中的 AUC 分别为 0.639 和 0.637，具有一定区分度但总体性能有限。该模型可作为中老年人群高血压风险初步筛查的参考，但仍需纳入更多关键预测因子并开展独立外部验证。

基金项目

华北理工大学校级大创项目(X2024055)。

参考文献

[1] 马丽媛, 王增武, 樊静, 胡盛寿. 《中国心血管健康与疾病报告 2021》关于中国高血压流行和防治现状[J]. 中国全科医学, 2022, 25(30): 3715-3720.

[2] 赖丽珊. 《2022 年世界卫生统计报告》复杂信息结构英汉翻译报告[D]: [硕士学位论文]. 广州: 广东外语外贸大学, 2024.

[3] 余振球, 陈云. 《ISH 2020 国际高血压实践指南》解读[J]. 中国乡村医药, 2020, 27(23): 24-25.

[4] GBD 2021 Diseases and Injuries Collaborators (2024) Global Incidence, Prevalence, Years Lived with Disability (YLDs), Disability-Adjusted Life-Years (DALYs), and Healthy Life Expectancy (HALE) for 371 Diseases and Injuries in 204

Countries and Territories and 811 Subnational Locations, 1990-2021: A Systematic Analysis for the Global Burden of Disease Study 2021. *The Lancet*, **403**, 2133-2161.

[5] 胡盛寿, 韩雅玲, 蔡军, 等. 中国高血压健康管理规范(2019) [J]. 中华心血管病杂志, 2020, 48(1): 10-46.

[6] Oishi, E., Hata, J., Honda, T., Sakata, S., Chen, S., Hirakawa, Y., *et al*. (2021) Development of a Risk Prediction Model for Incident Hypertension in Japanese Individuals: The Hisayama Study. *Hypertension Research*, **44**, 1221-1229. https://doi.org/10.1038/s41440-021-00673-7

[7] 吴月惟, 綦苗苗, 孔月琼, 等. 海南省居民高血压患病风险预测模型建立与验证[J]. 中国预防医学杂志, 2023, 24(2): 93-101.