·疾病控制·

应用TreeNet算法建立原发性高血压早期预测模型

郁小红，钱棪梅，周晨洁，马越，唐艳超，邹玲莉

空军杭州特勤疗养中心疗养三区，浙江 杭州 310002

摘要：目的 应用TreeNet算法建立原发性高血压（EH）早期预测模型，为早期监测EH提供预测方法。**方法** 收集2014—2016年均在杭州海勤体检中心或上海亿保健康管理公司进行健康体检者的体检资料，采用TreeNet算法建立EH预测模型；采用均方根误差（RMSE）、平均绝对偏差（MAD）和决定系数（R^2）评价模型特异度，绘制受试者操作特征（ROC）曲线，评价模型预测效果。**结果** 共收集4 265人的体检资料，其中EH 224例，占5.25%。共纳入12个关联指标，按重要性由大到小依次为2015年体质指数（BMI）、BMI 2014与2015年差值（差值）、三酰甘油（TG）差值、2015年总胆固醇（TC）、2014年高密度脂蛋白胆固醇（HDL-C）、2014年TG、2014年低密度脂蛋白胆固醇、2015年体重、2014年空腹血糖、2015年TG、尿素氮差值和2015年血小板，预测精度最高为100.00%，最低为56.89%。当2015年BMI>25 kg/m^2、BMI差值>0.5 kg/m^2、TG差值为1.3~3.3 mmol/L、2015年TC为2.0~2.4 mmol/L、2014年HDL-C<0.52 mmol/L时，2016年EH发病概率显著升高。预测RMSE为0.082，MAD为0.064，R^2为0.811，ROC曲线下面积为0.788（95%*CI*：0.741～0.815），灵敏度为69.05%，特异度为66.21%。**结论** 应用TreeNet算法建立的EH预测模型有助于早期评价高风险个体。

关键词：原发性高血压；TreeNet算法；数据挖掘；预测模型

中图分类号：R544.1 **文献标识码：**A **文章编号：**2096-5087（2022）09-0923-05

Establishment of a TreeNet algorithm-based model for early prediction of essential hypertension

YU Xiaohong, QIAN Yanmei, ZHOU Chenjie, MA Yue, TANG Yanchao, ZOU Lingli

The Third Sanatorium Department, Air Force Special Service Sanatorium, Hangzhou, Zhejiang 310002, China

Abstract: Objective To create a model for early prediction of essential hypertension (EH) based on the TreeNet algorithm, so as to provide a tool for early monitoring of EH. **Methods** The health examination data were collected from individuals receiving health examinations in Hangzhou Haiqin Health Examination Center or Shanghai Yibao Health Management Co., Ltd from 2014 to 2016, and a predictive model for EH was created based on the TreeNet algorithm. The effectiveness of the model for early prediction of EH was evaluated using root mean square error (RMSE), mean absolute deviation (MAD), coefficient of determination (R^2) and receiver operating characteristic (ROC) curve. **Results** A total of 4 264 healthy examination data were collected, and the prevalence of EH was 5.25% among the participants. A total of 12 variables were included in the model, and the highest contributing variable was body mass index (BMI), followed by BMI difference, two-year BMI difference, two-year triglyceride (TG) difference, two-year total cholesterol (TC) difference, high-density lipoprotein cholesterol (HDL-C) in 2014, TG in 2014, low-density lipoprotein cholesterol (LDL-C) in 2014, body weight in 2015, fasting blood glucose in 2015, TG in 2015, urea nitrogen difference and platelet in 2015. The highest predictive accuracy was 100.00%, and the lowest was 56.89%. The risk of EH significantly increased among individuals with BMI in 2015 of >25 kg/m^2, two-year BMI difference of >0.5 kg/m^2, two-year TG difference ranging from 1.3 to 3.3 mmol/L, TC in 2015 of 2.0 to 2.4 mmol/L and HDL-C in 2014 of <0.52 mmol/L. The model presented RMSE of 0.082, MAD of 0.064, R^2 of 0.811, area under the ROC curve of 0.788 (95%*CI*: 0.741-

DOI：10.19485/j.cnki.issn2096-5087.2022.09.012

作者简介：郁小红，本科，副主任护师，主要从事健康管理和疗养工作

通信作者：邹玲莉，E-mail：575723407@qq.com

0.815), sensitivity of 69.05% and specificity of 66.21% for prediction of EH. **Conclusion** The TreeNet algorithm-based model is effective for early monitoring of high-risk individuals for EH.

Keywords: essential hypertension; TreeNet algorithm; data mining; predictive model

原发性高血压（essential hypertension，EH）是常见的循环系统疾病。调查显示，我国 18～<35 岁人群高血压患病率为 5%，而 65～<75 岁人群高血压患病率超过 50%[1]。体质指数（BMI）、血压、血脂和尿酸等均与高血压的发生发展密切相关[2-5]，建立高血压早期预测模型对预防高血压具有积极意义。既往高血压预测模型主要通过多种回归分析建立[2-4]，方法简便，可直观解释各影响因素的相对危险度，但对变量的具体路径和预测拐点探讨较少，难以解释复杂因素的非线性关系。TreeNet 算法是数据挖掘方法之一，将所有子项目对目标的贡献率分别建立“树”，筛选最优贡献的“树”的组合形成模型。TreeNet 算法对数据集要求宽容，可以分析复杂非线性关系数据，能够筛选目标变量高度相关因素，分析各个因素对目标变量的重要性程度，计算能力优于回归分析方法[6]。本研究采用 TreeNet 算法构建 EH 发病风险的预测模型，为 EH 的早期监测提供依据。

1　对象与方法

1.1　对象　选择 2014—2016 年均在杭州海勤体检中心或上海亿保健康管理公司进行健康体检者为研究对象。纳入标准：2014 年和 2015 年体检结果正常，2016 年体检排除 EH 和初诊为 EH 者。排除标准：首次体检<18 岁或>80 岁；体检资料不完整；患继发性高血压；妊娠或哺乳期；患其他慢性病或 3 个月内有用药记录。研究对象均签署知情同意书。

1.2　方法　通过体检人群健康档案收集研究对象 2014—2016 年体检资料[2-4]：（1）基本信息，包括性别、年龄和疾病史；（2）体格检查资料，包括身高、体重、脉搏、收缩压（SBP）和舒张压（DBP），计算 BMI；（3）实验室检测资料，包括红细胞（RBC）、白细胞（WBC）、血红蛋白（Hb）、血小板计数（PLT）、直接胆红素（DBIL）、间接胆红素（IBIL）、总胆红素（TBIL）、总蛋白（TP）、白蛋白（ALB）、球蛋白（GLB）、白球比（A/G）、谷草转氨酶（AST）、谷丙转氨酶（ALT）、总胆固醇（TC）、三酰甘油（TG）、高密度脂蛋白胆固醇（HDL-C）、低密度脂蛋白胆固醇（LDL-C）、空腹血糖（FPG）、尿酸（UA）、肌酐（Cre）、尿素氮（BUN）、癌胚抗原（CEA）、甲胎蛋白（AFP）、血清铁蛋白（SF）、糖类抗原 199（CA-199）、糖类抗原 125（CA-125）、糖类抗原 153（CA-153）、前列腺特异性抗原（PSA）、三碘甲腺原氨酸（T_3）和甲状腺素血清总甲状腺素（T_4）等。

1.3　EH 诊断标准　参考《中国高血压防治指南（2018 年修订版）》[7]，在未使用降压药物的情况下，SBP≥140 mm Hg（1 mm Hg=0.133 kPa）和（或）DBP≥90 mm Hg，且排除继发性高血压诊断为 EH。对连续 3 年的血压指标进行筛选和分析，由 2 名以上医生对诊断结果进行复核。

1.4　应用 TreeNet 算法建立 EH 预测模型

1.4.1　数据处理　对原始数据进行独立清洗，剔除缺失数据和不合理数据，确保同一变量的数据单位、计算方式一致。采用“0”和“1”分别表示观测终点（2016 年）未患 EH 和已患 EH 状态；变量尾缀“-1”和“-2”分别表示 2014 年和 2015 年检测值；D 表示 2014 年和 2015 年的差值。

1.4.2　模型建立　采用萨尔福德公司的 TreeNet 软件（Salford predictive modeler 8.3）建立 EH 模型。80% 的体检数据作为训练集用于建立模型，20% 的体检数据作为测试集用于检验模型预测精度。设置 EH 为目标数据，其他变量为计算数据。学习率采用自动模式，设置算法参数：学习率为 0.01，子样本比例为 0.5，每棵树的最大节点数为 6，最大树深度为 10，最小末端节点数（样本）为 10，M-回归细分为 0.99，回归损失标准采用 Huber-M。利用建立的模型直接对训练集样本进行评分，计算 EH 发病概率，然后计算预测指标的重要性，以指标在单棵树中的重要性平均值衡量全局重要性。分析预测指标与 EH 发病概率的依存关系，先规范所有预测指标的取值范围，得出 EH 发病概率集中出现的一个指标范围组合，产生一条依存曲线；重复以上步骤，产生一组依存曲线；取所有曲线的平均值，生成新的依存曲线。本文呈现预测精度前五位指标的依存曲线。

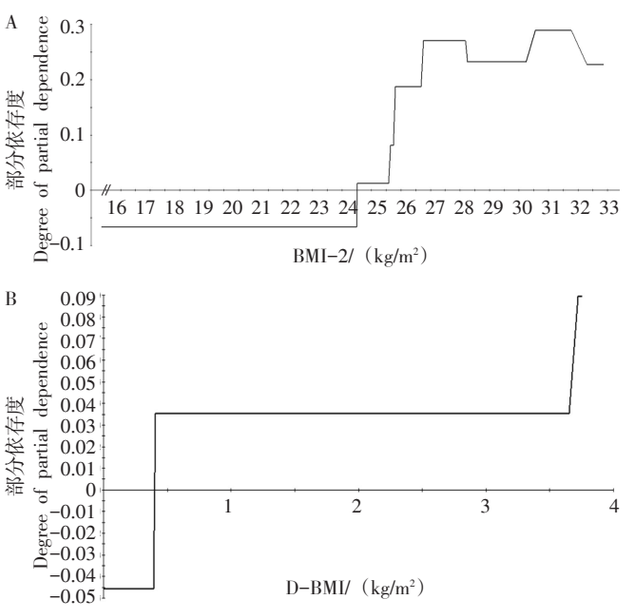
1.4.3　模型检验　采用 10 折交叉验证评估模型预测效果，采用均方根误差（root mean square error，RMSE）、平均绝对偏差（mean absolute deviation，MAD）和决定系数（R^2）评价模型特异度，RMSE 和 MAD 越接近 0、R^2 越接近 1 表示模型误差越小，特异度越好。采用 SPSS 21.0 软件绘制受试者操作特征

（receiver operating characteristic，ROC）曲线，计算曲线下面积（area under curve，AUC），评价模型预测价值。*AUC* 值越大表示模型区分度越好。将测试集样本发生 EH 定义为目标状态，通过 EH 实际患病率和模型预测准确概率评价模型的特异度和灵敏度。采用分箱法将测试集样本按照 EH 发生风险从高到低排序后分成 10 份，以每份中最终发生 EH 的概率作为统计值，比较不同风险梯度人群 EH 患病率，验证模型的预测价值。

2 结 果

2.1 EH 预测模型建立结果 收集 4 265 人体检资料，其中男性 2 692 人，年龄为（42.51±13.38）岁；女性 1 573 人，年龄为（40.59±12.38）岁。2016 年体检检出 EH 224 例，占 5.25%。训练集 3 412 人，测试集 853 人。

建立的模型纳入重要性排序前 12 位指标，分别为 2015 年 BMI、BMI 差值、TG 差值、2015 年 TC、2014 年 HDL－C、2014 年 TG、2014 年 LDL-C、2015 年体重、2014 年 GLU、2015 年 TG、BUN 差值和 2015 年 PLT，预测精度分别为 100.00%、93.54%、89.59%、89.23%、88.70%、87.24%、85.68%、82.56%、80.33%、78.97%、70.04% 和 56.89%。重要性前五位指标与 2016 年 EH 发病概率的预测关系见图 1。2015 年 BMI>25 kg/m^2，BMI 差值>0.5 kg/m^2，TG 差值为 1.3~3.3 mmol/L，2015 年 TC 为 2.0~2.4 mmol/L，2014 年 HDL-C<0.52 mmol/L 时，2016 年 EH 发病概率显著升高。



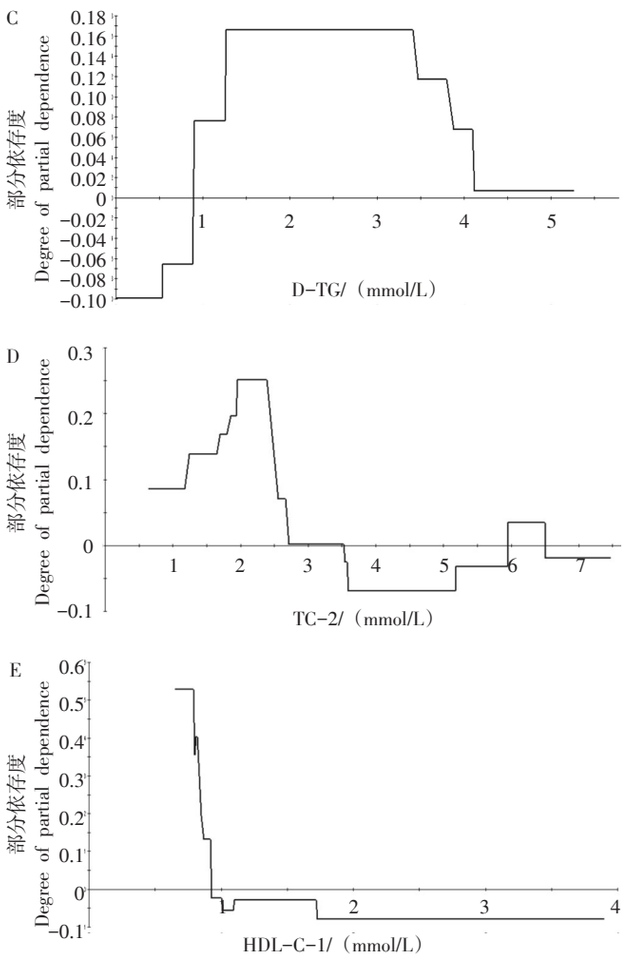


图 1 重要性前五位指标与 EH 发病概率的依存曲线

Figure 1 Dependence curve between top five important variables and EH incidence probability

2.2 EH 预测模型检验结果 训练集 RMSE 为 0.087，MAD 为 0.056，R^2 为 0.776；测试集 RMSE 为 0.082，MAD 为 0.064，R^2 为 0.811。*AUC* 值为 0.788（95%*CI*：0.741～0.815）。测试集实际确诊 EH 42 例，模型预测 EH 303 例，特异度为 66.21%，灵敏度为 69.05%。

测试集样本按照 EH 发病风险从高到低排序后，第 1 份纳入 EH 发病风险排名前 9.38% 的样本，涵盖 28.57% 的实际确诊 EH 患者，这部分人群患 EH 风险较普通人群提升 3.05 倍；第 2 份纳入 EH 发病风险排名前 18.76% 的样本，涵盖 54.76% 的实际确诊 EH 患者，这部分人群患 EH 的风险较普通人群提升 2.79 倍。见表 1。

3 讨 论

TreeNet 算法能呈现具体指标的重要性、路径和

表 1　EH 预测模型的预测效果评价

Table 1　Effectiveness of the predictive model of essential hypertension (EH)

序号 Number	样本Sample（*n*=853）			实际确诊EH Confirmed EH（*n*=42）				风险识别提升倍数 Risk increasing times
	人数*n*	构成比 Proportion/%	累积构成比 Cumulative proportion/%	例数*n*	患病率 Prevalence/%	构成比 Proportion/%	累积构成比 Cumulative proportion/%	
1	80	9.38	9.38	12	15.00	28.57	28.57	3.05
2	80	9.38	18.76	11	13.75	26.19	54.76	2.79
3	84	9.85	28.60	4	4.76	9.52	64.29	0.97
4	87	10.20	38.80	3	3.45	7.14	71.43	0.70
5	89	10.43	49.24	3	3.37	7.14	78.57	0.68
6	89	10.43	59.67	1	1.12	2.38	80.95	0.23
7	89	10.43	70.11	2	2.25	4.76	85.71	0.46
8	92	10.79	80.89	3	3.26	7.14	92.86	0.66
9	82	9.61	90.50	1	1.22	2.38	95.24	0.25
10	81	9.50	100.00	2	2.47	4.76	100.00	0.50

方向，更直观了解生理指标与目标疾病发生之间的变化关系，得到预测指标的预警数值，有助于临床建立更直观的疾病筛查路径，因此在疾病的预测中具有较好的使用度。本研究采用 TreeNet 算法建立的 EH 早期预测模型的 *AUC* 值为 0.788（95%*CI*：0.741～0.815），特异度为 66.21%，灵敏度为 69.05%。分箱法验证结果显示，模型在风险高分段具备良好的识别能力，可靠性较好。

建立的 EH 预测模型共纳入 12 个 EH 危险因素，预测精度从高到低依次为 2015 年 BMI、BMI 差值、TG 差值、2015 年 TC、2014 年 HDL-C、2014 年 TG、2014 年 LDL-C、2015 年体重、2014 年 GLU、2015 年 TG、BUN 差值和 2015 年 PLT，其中 BMI、脂质代谢指标是相对重要的预测指标。

多项研究表明，BMI 可用于预测 EH。张宇宁等[8]研究发现，超重患者高血压发病率显著高于体重正常者。高仲淳等[9]研究发现，人群高血压患病率随 BMI 的升高呈上升趋势。超重患者因胰岛素抵抗提高中枢交感神经活性，促进细胞内钙潴留和钠排泄，导致机体血压升高[10-12]。对重要性排名前五位的指标作进一步依存分析发现，BMI 是 EH 的重要预测指标，且当 BMI>25 kg/m^2 时，若 BMI 差值超过 0.5 kg/m^2，EH 发病概率升高。BMI 是敏感的预测因素，应重视个体 BMI 的增长趋势。

脂质代谢与 EH 发生相关。有研究发现，血脂异常患者的代谢改变会产生具有独特氧化脂质特征的 TG 脂蛋白，促进动脉粥样硬化，引发心血管疾病[13]。血脂水平升高可促进超氧化物产生，造成血管内皮损伤，引起多种代谢调控及功能紊乱[14]，可能是血脂代谢紊乱促进高血压的主要原因。而改善脂质分布的营养和药物手段有助于维护血管内皮功能，降低高血压发病率[15-16]。本研究结果显示，TG、HDL-C、LDL-C 和 TC 水平均可能影响 EH 的发病概率，TG 差值为 1.3~3.3 mmol/L、2015 年 TC 为 2.0~2.4 mmol/L、2014 年 HDL-C<0.52 mmol/L，2016 年 EH 发病概率显著升高。这提示在 EH 预防和监测中应重视脂质代谢指标的作用，不仅要关注指标本身的阈值，更要关注指标变化，早期开展血脂干预推迟 EH 的发生。

参考文献

[1] 赵冬．中国成人高血压流行病学现状［J］．中国心血管杂志，2020，25（6）：513-515.
ZHAO D.Current epidemiology of adult hypertension in China［J］. Chin J Cardiovasc Med，2020，25（6）：513-515.

[2] 李禄伟，黄倩，施佳成，等．基于三种统计学方法构建的超重及肥胖人群高血压发病预测模型的分析比较［J］．现代预防医学，2021，48（11）：2061-2065.
LI L W，HUANG Q，SHI J C，et al.Screening risk factors and interaction analysis of hypertension in overweight and obesity population based on three statistical models［J］.Mod Prev Med，2021，48（11）：2061-2065.

[3] 王定坤，杨杉．基于 COX 比例风险模型分析心力衰竭影响因素［J］．电脑知识与技术（学术版），2021，17（24）：33-35.
WANG D K，YANG S.Analysis of influencing factors for heart fail-

ure based on COX proportional hazard model［J］.Comput Knowl Technol，2021，17（24）：33-35.

［4］ 付菲，彭映辉，徐肇元，等．急性心肌梗死患者心力衰竭风险预测模型研究［J］．中国心血管杂志，2021，26（6）：525-530.
FU F，PENG Y H，XU Z Y，et al.Study on risk prediction model of heart failure in patients with acute myocardial infarction［J］. Chin J Cardiovasc Med，2021，26（6）：525-530.

［5］ 刘仕俊，袁寒艳，姜彩霞，等．杭州市老年高血压患者血压控制的影响因素研究［J］．预防医学，2021，33（7）：660-664.
LIU S J，YUAN H Y，JIANG C X，et al.Influencing factors for blood pressure control in elderly patients with hypertension in Hangzhou［J］.Prev Med，2021，33（7）：660-664.

［6］ PADMAJA B，RAMA PRASAD V V，SUNITHA V N.TreeNet analysis of human stress behavior using socio-mobile data［J/OL］.J Big Data，2016，4（1）［2022-07-08］.https://doi.org/10.1186/s40537-016-0054-3.

［7］ 中国高血压防治指南修订委员会，高血压联盟（中国），中华医学会心血管病学分会，等．中国高血压防治指南（2018年修订版）［J］．中国心血管杂志，2019，24（1）：24-56.
Writing Group of 2018 Chinese Guidelines for the Management of Hypertension，Chinese Hypertension League，Chinese Society of Cardiology，et al.2018 Chinese guidelines for the management of hypertension［J］.Chin J Cardiovasc Med，2019，24（1）：24-56.

［8］ 张宇宁，郑浩，梁洁，等．老年人体重指数与血压水平及高血压患病率的相关性［J］．中国老年学杂志，2021，41（20）：4333-4335.
ZHANG Y N，ZHENG H，LIANG J，et al.Relationship between aged people body mass index and blood pressure and prevalence rate of hypertension［J］.Chin J Gerontol，2021，41（20）：4333-4335.

［9］ 高仲淳，邹波，蓝恭赛，等.20～59岁成年人体质指数随年龄变化轨迹与高血压发病的关系研究［J］．中国全科医学，2021，24（8）：954-958.
GAO Z C，ZOU B，LAN G S，et al.The relationship between trajectory of body mass index based on age and the incidence of hypertension in adults aged 20 to 59 years［J］.Chin Gen Pract，2021，24（8）：954-958.

［10］ KAMPMANN U，MATHIASSEN O N，CHRISTENSEN K L，et al.Effects of renal denervation on insulin sensitivity and inflammatory markers in nondiabetic patients with treatment-resistant hypertension［J/OL］.J Diabetes Res，2017［2022-07-08］.https://doi.org/10.1155/2017/6915310.

［11］ D'ELIA L，STRAZZULLO P.Excess body weight，insulin resistance and isolated systolic hypertension：potential pathophysiological links［J］.High Blood Press Cardiovasc Prev，2017，25（7）：1377-1389.

［12］ TAHERI A，MIRZABABAEI A，SETAYESH L，et al.The relationship between Dietary approaches to stop hypertension diet adherence and inflammatory factors and insulin resistance in overweight and obese women：a cross-sectional study［J/OL］.Diabetes Res Clin Pract，2021，182［2022-07-08］.https://doi.org/10.1016/j.diabres.2021.109128.

［13］ RAJAMANI A，BORKOWSKI K，AKRE S，et al.Oxylipins in triglyceride-rich lipoproteins of dyslipidemic subjects promote endothelial inflammation following a high fat meal［J/OL］.Sci Rep，2019，9（1）［2022-07-08］.https://doi.org/10.1038/s41598-019-45005-5.

［14］ 丁存涛，周亚群，孙希鹏，等．糖脂代谢对原发性高血压病人血管内皮功能的影响［J］．首都医科大学学报，2017，38（3）：401-405.
DING C T，ZHOU Y Q，SUN X P，et al.Effects of glucose and lipid metabolism on vascular endothelial function in patients with essential hypertension［J］.J Cap Med Univ，2017，38（3）：401-405.

［15］ LANDI F，MARTONE A M，SALINI S，et al.Effects of a new combination of medical food on endothelial function and lipid profile in dyslipidemic subjects：a pilot randomized trial［J/OL］.Biomed Res Int，2019（6）［2022-07-08］.https://doi.org/10.1155/2019/1970878.

［16］ DJINDJIĆ B，RADOVANOVIĆZ L，KOSTIĆ T，et al.The changes of oxidative stress and endothelial function biomarkers after 6 weeks of aerobic physical training in patients with stable ischemic coronary disease［J］.Mil Med Pharm J Serbia，2017，74（11）：1060-1065.

收稿日期：2022-03-09 **修回日期**：2022-07-08 **本文编辑**：吉兆洋