·基础理论与方法·

应用人工神经网络预测个体患原发性高血压病危险度

周水红　聂绍发　王重建　魏晟　许奕华　李雪华　宋恩民

【摘要】 **目的**　建立个体患原发性高血压病的预测模型，评价并探讨预测个体患病的新方法。**方法**　选择 3054 名社区居民流行病学调查资料，按照年龄、性别均衡性，按4:1分为训练集(2438 名)与检验集(616 名)两部分，分别用于筛选变量、建立预测模型及对模型的检测和评价。应用人工神经网络(ANN)和 logistic 回归分析方法建立模型，用 ROC 方法评价所建立的高血压患病预测模型的优劣。**结果**　对 616 名检验集预测，ANN 模型的特异性较低，但准确性、灵敏度指标均优于 logistic 回归模型，ANN_2 的约登指数为0.8399，明显高于其他两个模型；通过 ROC 曲线下面积比较模型的预测能力：logistic 回归方法曲线下面积($Az=0.732\pm0.026$)小于 ANN 模型(ANN_2 和 ANN_1 分别为0.918 ± 0.013、0.900 ± 0.014)，即 ANN 模型有更好的预测判别效能。**结论**　初步证明在预测个体患高血压病方面，ANN 方法预测效能更优，从而为解决个体发病危险预测提供了一个新方法。

【关键词】 高血压，原发性；个体危险度；人工神经网络

The application of artificial neural networks to predict individual risk of essential hypertension　*ZHOU Shui-hong**, *NIE Shao-fa*, *WANG Chong-jian*, *WEI Sheng*, *XU Yi-hua*, *LI Xue-hua*, *SONG En-min*. **Department of Epidemiology and Health Statistics*, *School of Public Health*, *Tongji Medical College*, *Huazhong University of Science and Technology*, *Wuhan* 430030, China
Corresponding author: *NIE Shao-fa*, Email: sf_nie@mails.tjmu.edu.cn

【**Abstract**】 **Objective**　To establish models to predict individual risk of essential hypertension and to evaluate and explore new forecasting methods. **Methods**　To select data of 3054 community residents from a epidemiological survey and divided them into 4:1 (2438 cases and 616 cases) ratio in accordance with the balance of age and sex to filter variables, and to establish, test and evaluate the prediction models. Using artificial neural network (ANN) and logistic regression analysis to establish models while applying ROC to evaluate the prediction models. **Results**　Forecast results of the models applying to the test set proved that ANN had lower specificity but better veracity and sensitivity than logistic regression. In particular, the Youden's index of the ANN_2 came up to 0.8399 which was distinctly higher than the other two models. When the area was under the ROC curve of logistic regression, the ANN_1 and ANN_2 models equaled to 0.732 ± 0.026, 0.900 ± 0.014 and 0.918 ± 0.013 respectively, which proved that the ANN model was better in the prediction about individual health risk of essential hypertension. **Conclusion**　Our results showed that ANN method seemed better than logistic regression in terms of predicting the individual risk from hypertension thus supplied a new method to solve the forecast of individual risk.

【**Key words**】 Essential hypertension; Individual health risk; Artificial neural network

高血压的发病是多因素联合作用的结果，但其发生和发展具有一定规律，对高血压病进行社区干预符合成本效益，是降低心血管发病率上升的有效干预措施。因此，研究评价个体患高血压的危险度，可对其采取有效的针对性干预措施，延缓疾病的发生、降低发病的危害性。流行病学数学模型已广泛应用于流行病学研究的各个领域[1]，在研究疾病的流行特征、效果评价和疾病预测中，数学模型均有重要的作用，其中 logistic 回归分析和人工神经网络(artificial neural network, ANN)方法经常应用在相关疾病的预测研究中[2]，但对高血压患病预测方面的系统报道并不多见。本研究利用收集到的流行病学调查资料，运用 ANN 方法和 logistic 回归方法构

基金项目：国家“863”高技术研究发展计划资助项目(2006AA02Z347)

作者单位：430030 武汉，华中科技大学同济医学院公共卫生学院流行病与卫生统计学系(周水红、聂绍发、王重建、魏晟、许奕华、李雪华)；华中科技大学计算机科学与技术学院(宋恩民)

通讯作者：聂绍发，Email：sf_nie@mails.tjmu.edu.cn

建不同的高血压发病风险预测模型，比较不同方法所构建模型的预测性能，探讨个体发病风险预测的方法。

对象与方法

1. 调查对象：2000 年 10 月对三峡坝区 35 岁以上的常住人口（居住 5 年以上）进行问卷调查，排除长期外出及因病无法配合调查者，同时剔除已确诊并接受治疗的原发性高血压患者以及患有心、脑、肾等疾病者。最后符合研究要求的共计 3054 名调查对象被纳入本次调查，其中高血压患者 823 例。

2. 调查内容和方法：调查内容：①问卷调查：采用入户调查方式进行问卷调查。调查内容包括人口学资料、健康知识、健康意识、健康行为、高血压相关知识等，并测量其血压、身高、体重等。②生命质量测量：采用世界家庭医师协会推荐的社区人群功能状态测定量表（COOP/WONCA 量表）中文版[3]，调查对象接受调查前 2 周的情况。

3. 原发性高血压诊断及血压分级分型：根据 1999 年 WHO/ISH 原发性高血压定义，血压测量方法及分级、分型参照《中国高血压防治指南》[4]。SBP≥ 140 mm Hg（18.7 kPa）和/或 DBP≥ 90 mm Hg(12.0 kPa)，排除继发性高血压，或既往确诊的原发性高血压者，均记为原发性高血压。

4. 统计学分析：用 Epi Data 3.1 软件平行双录入调查问卷。采用逻辑查错和区间定值查错法对所有原始数据进行详细查错，并复核。运用 SPSS 12.0 统计软件建立条件 logistic 回归模型，运用 Matlab 7.1 软件编程建立 ANN 预测模型；采用 ANN 和 logistic 回归方法建立预测模型；应用 SPSS 12.0 软件的 Syntax 编程绘制三个模型预测判别的 ROC 曲线，比较 ROC 曲线下面积，对所建立的三个预测模型进行评价分析。

由于 ANN 模型的预测预报能力与学习样本质量及信息紧密相关，故训练集的样本量应比检验集多。本次研究将 3054 名调查对象的资料按照性别、年龄组（自 35 岁起组距为 5 岁，≥65 岁者合并为一组）指标 4:1 随机分为训练集（2438 名）与检验集（616 名）两部分，每组中高血压与正常血压者的比例与原始数据保持一致。进行性别、年龄的均衡性 χ^2 检验，两组结果的差异无统计学意义。训练集和检验集分别用于筛选变量和建立预测模型，及对模型的检验和评价。

以收集到的 29 项观察指标为自变量，以研究对象是否患高血压为应变量进行分析。对自变量缺失值的处理方法：定量资料以列的算术平均值替代项目中的缺失值；定性资料用空缺属性值的所有可能的属性随机取值来填充。进行 logistic 分析时填充数据，神经网络模拟时以 NULL 替代缺失值。

结　果

1. 建立 logistic 回归预测模型：将本次研究的 3054 份调查表整理后，建立 SPSS 12.0 软件数据库，按照上述的分类标准对筛选出的 2438 名训练集进行分析。从训练集中随机抽取高血压和匹配的非高血压患者各 100 名，进行 1:1 的病例对照研究，应用条件 logistic 回归方法，进行高血压的单因素和多因素回归分析，筛选影响高血压患病的自变量。多因素回归分析结果见表 1。

表1 高血压影响因素的多因素条件 logistic 回归分析

因素	β	$s_{\bar{x}}$	Wald χ^2 值	P 值	OR 值(95% CI)
职业	0.087	0.040	4.692	0.030	1.090(1.008～1.179)
家族史	0.479	0.170	7.945	0.005	1.614(1.157～2.252)
文化程度	−0.296	0.068	18.813	0.000	0.744(0.650～0.850)
饮酒	−0.275	0.123	5.000	0.025	0.760(0.597～0.967)
蔬菜、水果摄入	−0.125	0.060	4.397	0.036	0.882(0.784～0.992)
饮食偏咸	0.136	0.045	9.310	0.002	1.146(1.051～1.250)
吃动物内脏	−0.240	0.063	14.660	0.000	0.787(0.696～0.889)
锻炼	−0.143	0.083	2.953	0.086	0.866(0.736～1.020)
BMI	0.422	0.131	10.443	0.001	1.525(1.181～1.971)
血压差	0.096	0.005	388.282	0.000	1.101(1.091～1.112)

多因素分析时，以 $P<0.05$ 作为选入变量的标准，$P>0.1$ 作为剔除变量的标准，采用偏最大似然估计前进法逐步回归分析，最后共筛选出 10 个影响因素。以此分析结果建立 logistic 回归模型，用于预测 616 名研究对象是否患高血压。所建立的预测模型为 $\text{logit}(P) = \ln[P/(1-P)] = 0.087$（职业）+ 0.479（家族史）− 0.296（文化程度）− 0.143（锻炼）− 0.240（吃动物内脏）− 0.275（饮酒）− 0.125（蔬菜、水果摄入）+ 0.136（饮食偏咸）+ 0.422（BMI）+ 0.096（血压差）− 5.228。其中 P 为患高血压的概率，取0.5为判别界线，即$P\geq0.5$时研究对象患高血压，$P<0.5$时研究对象不患高血压。经模型改善情况检验（$\chi^2=3.606$）和整个模型检验（$\chi^2=758.834$），该模型的分类判对率为83.6%。

2. 建立 ANN 预测模型：将以上 logistic 回归分析筛选出的 10 个自变量作为输入元，构建 ANN_1 预

测模型，网络的输入变量：BMI、血压差为连续型变量，职业、文化程度、蔬菜和水果摄入、饮食偏咸、吃动物内脏为多分类变量，家族史、饮酒、锻炼二分类变量。输出变量：是否患高血压为二分类变量。分析时构造了如图 1 所示的 3 层 BP 神经网络，其结构为：输入层含 10 个神经元；隐含层 21 个神经元，并且可调，输出层 1 个神经元，对应预测变量（即是否患高血压），传递函数为 S 型曲线。所建立模型的结构见图 1，其中隐含层只绘制了部分节点，其他节点以“·”省略。ANN_1 中隐含层节点数为 21 个神经元，目标误差取0.01，学习速率0.1，最大训练周期2000，训练中均方误差（MSE）为0.009 96，输出层的坡度 Gradient 为2.549 98/1e－010 时，即可满足训练参数中所要求的目标误差（图 2）。

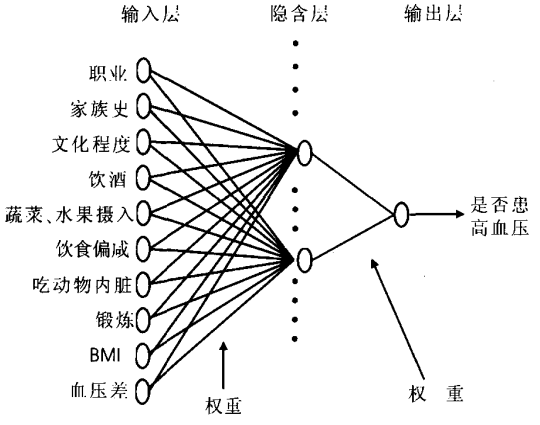

图1　模型 ANN_1 的结构示意图

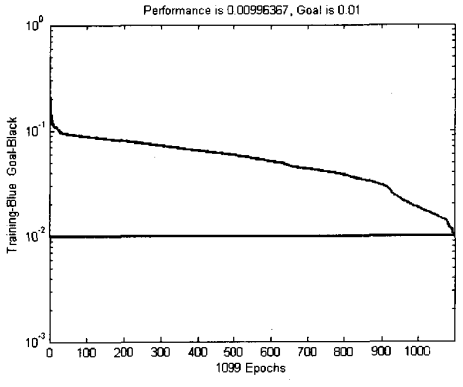

图2　神经网络 ANN_1 测试结果

结合上述 logistic 回归分析的结果和有关文献报道的高血压患病的影响因素[5-8]，选择 13 个自变量作为神经网络的输入元：年龄、BMI、收入为连续型变量，性别、饮酒、吸烟、家族史、锻炼为二分类变量，婚姻状况、饮食偏咸、文化程度、锻炼、蛋类摄入为多分类变量；是否患高血压为输出元，直接构建前馈型 BP 网络，称为 ANN_2。运算方法同 ANN_1，模型结构类似于 ANN_1。

3. logistic 回归、ANN_1 和 ANN_2 模型预测模型预测：根据建立的 logistic 回归预测模型、ANN_1 和 ANN_2 模型，分别对 616 名检验集进行预测，得出的预测判别结果见表 2。

表2　三个模型的预测判别结果

实际情况	logistic 回归预测		ANN_1		ANN_2		合计
	高血压	正常	高血压	正常	高血压	正常	
高血压	85	83	160	8	161	7	168
正常	19	429	68	380	54	394	448
合计	104	512	228	388	215	401	616

4. 模型预测结果的评价：运用已经建立的预测模型，对检验集数据进行预测判别，结果如表 3 所示。网络模型 ANN_1 和模型 ANN_2 准确性、敏感性方面均优于 logistic 回归模型；网络模型特异性较低，ANN_2 的约登指数为0.8399，明显高于其他两个模型。

表3　三个模型预测结果评价指标

指标	logistic 回归	ANN_1	ANN_2
准确性（%）	83.55	87.00	90.00
敏感性（%）	50.53	95.24	96.15
特异性（%）	95.68	84.81	87.84
约登指数（%）	46.22	80.05	83.99

5. ANN 与 logistic 回归模型预测能力比较：ROC 曲线如图 3 所示。通过计算 ROC 曲线下面积可以看出，ANN_2 的 ROC 曲线下面积 $Az = 0.918 \pm 0.013$，明显高于其他两个的预测判别能力（ANN_1 模型的 ROC 曲线下面积为 $Az = 0.900 \pm 0.014$，logistic 回归模型的 ROC 曲线下面积为 $Az = 0.732 \pm 0.026$）。

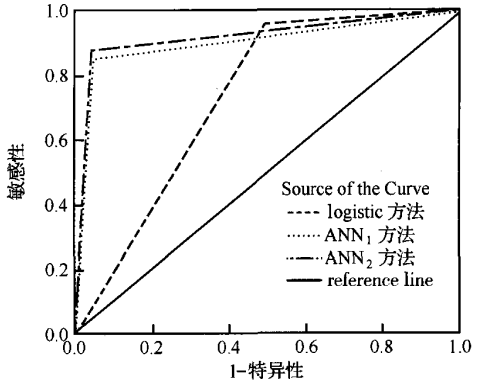

图3　logistic、ANN_1 和 ANN_2 回归模型的 ROC 曲线

讨 论

ANN 能为每名研究对象"量体裁衣"地给出一个特定的预测结果[9,10],并且 ANN 的预测能力及准确性与其输入变量包含的信息量有关,故本次研究选用了 ANN_1 和 ANN_2 两个模型并比较两者的预测能力。目前,ANN 在慢性非传染性疾病个体危险度预测中的应用还不成熟,尚有一些问题需要解决。本研究应用 ANN 方法时采用了一层隐含层,即减少了计算量,又可以防止过度拟合。对隐含层节点数目的考察,节点数为 21 时,模型的校正误差最小;当隐含层的节点数少于或超过 21 时,模型的校正误差均呈下降趋势,故确定隐含层节点为 21 个神经元。

本研究中应用 logistic 回归进行高血压危险因素筛选时,饮酒因素是保护因素,与目前的研究结果不一致,可能与研究对象报告时的偏倚有关。因为饮酒与血压之间呈一种 J 型关系,大量长期饮酒是高血压的独立危险因素,但是轻度饮酒对血压无不良影响,而本次研究调查中没有区分饮酒量。

研究中利用条件 logistic 回归分析筛选出的危险因素建立预测模型,并随机抽取训练集样本、检验集样本各 100 名进行模拟以验证模型的预测可靠性。结果模型的预测结果和实际情况十分吻合(MSE<0.01)。所建立的三个模型对检验集数据的预测判别结果:ANN 模型的准确性、敏感性均优于 logistic 回归模型;ANN_2 的准确性和敏感性均高于另外两个模型,而且预测判别能力是三个模型中最强的,其约登指数明显高于另两个模型。

受试者工作曲线(ROC 曲线)是应用比较广泛的评价两种或两种诊断方法的诊断水平的标准方法[11],ROC 曲线可以直观的观察敏感性和特异性之间的关系,曲线下面积越大其诊断试验的准确度越大。本研究借助于 ROC 曲线评价三种模型的预测效果。通过 ROC 曲线下面积比较:logistic 回归预测模型 ROC 曲线下面积小于 ANN,且 ANN_1 和 ANN_2 曲线下面积比较差异不显著,提示 ANN 预测模型较 logistic 回归预测模型有更好的预测判别效能,即 ANN 较常规方法更能把握数据的内在规律。

本研究所建立的两个 ANN 模型之间预测效能差异非常小,虽然一个是 10 维的输入,一个是 13 维的输入,但从 ANN_1 和 ANN_2 两个模型预测的结果来看,其差异不大,也说明 ANN 方法处理数据的能力强大,能够自动提取有用的信息资料。

尽管 ANN 网络方法已在肿瘤、糖尿病等慢性病及药学研究等领域显出其独特的优越性,但直到目前为止,在高血压患病预测方面的文献还较少见,本研究通过比较 ANN 预测方法与 logistic 回归预测方法的预测结果,证实了 ANN 模型预测影响因素与高血压发病关系有较强的预测效能。

研究中借助现况研究的数据资料,对高血压个体发病风险进行了探讨,由于资料的地区局限性及对入选变量的限制,加上本次主要研究了环境、遗传、生活习惯等因素,所得到的预测模型存在一定的缺陷,其推广仍需要进一步的论证。

参 考 文 献

[1] 周宝森,汪宁. 理论流行病学//李立明. 流行病学. 5 版. 北京:人民卫生出版社,2004:133-136.

[2] 李士勇. 模糊控制·神经控制和智能控制论. 哈尔滨:哈尔滨工业大学出版社,1996:72-108.

[3] Weel C van, Zahn CK, Touw-Otten MM, et al. Measuring functional health status with the COOP/WONCA charts A manual [M/O]. The Netherlands: the World Organization of Family Doctors (WONCA), the Northern Centre for Health Care Research (NCH), University of Groningen, (ISBN 90 72156 33 1): 1995 (2001-11) [2006-02-01]. http://www.globalfamilydoctor.com/publications/coop-woncacharts/COOP-WONCACHARTS.HTM? refNum=5674.

[4] 中国高血压防治指南修订委员会. 中国高血压防治指南(2005 年修订版). 北京:中国高血压防治指南修订委员会,2005.

[5] 罗雷,栾荣生,袁萍. 中国居民高血压主要危险因素的 Meta 分析. 中华流行病学杂志,2003,24(1):50-53.

[6] Wildman RP, Gu DF, Kristi R, et al. Are waist circumference and body mass index independently associated with cardiovascular disease risk in Chinese adults? Am J Clin Nutr, 2005, 82: 1195-1202.

[7] Fagard RH, Cornelissen VA. Effect of exercise on blood pressure control in hypertensive patients. Eur J Cardiovasc Prev Rehabil, 2007, 14(1): 12-17.

[8] Dickinson HO, Mason JM, Nicolson DJ, et al. Lifestyle interventions to reduce raised blood pressure: a systematic review of randomized controlled trials. J Hypertens, 2006, 24(2): 215-233.

[9] Zhu J, Zhu XD, Liang SX, et al. Prediction of radiation induced liver disease using artificial neural networks. Jpn J Clin Oncol, 2006, 36(12): 783-788.

[10] Baldassarre D, Grossi E, Buscema M, et al. Recognition of patients with cardiovascular disease by artificial neural networks. Ann Med, 2004, 36(8): 630-640.

[11] Gary LG, Ruyun J. Receivor operating characteristic curve analysis of clinical risk models. Ann Thorac Surg, 2001, 72: 323-326.

(收稿日期:2008-02-14)
(本文编辑:张林东)