

BP人工神经网络在早产预测模型中的应用

徐新 谭红专 周书进 何玥 沈琳 柳祎 胡丽 王小娟 李逊

【摘要】 目的 基于BP人工神经网络的基本原理和方法,构建早产预测模型。方法 采用前瞻性队列研究方法,通过整群抽样,2010—2012年在湖南省浏阳市随机抽取怀孕妇女作为研究对象。按2:1:1将调查样本随机分为训练样本、检验样本和测试样本,分别用于建立预测模型、选择最优神经网络和模型预测评价。采用BP人工神经网络和logistic回归分析建立模型,用ROC方法评价所建立的早产预测模型的优劣。结果 整理6 270例分娩孕妇的数据,结果显示,早产265例,早产发生率为4.22%。将子宫异常及畸形、产次、妊娠胎数、妊娠期高血压、前置胎盘、胎膜早破和定期产检7个多因素分析有统计学意义的变量选入预测模型。BP人工神经网络模型预测早产的灵敏度、特异度和一致率分别为67.65%、84.87%和84.12%,ROC曲线下面积为0.795;logistic回归模型预测早产的灵敏度、特异度和一致率分别为64.71%、85.60%和84.69%,ROC曲线下面积为0.783。结论 新建立的BP人工神经网络模型实用可靠,其对早产的预测效能略优于logistic回归模型。

【关键词】 早产;神经网络;预测

Study on the application of Back-Propagation Artificial Neural Network used the model in predicting preterm birth Xu Xin¹, Tan Hongzhuang¹, Zhou Shujin², He Yue¹, Shen Lin¹, Liu Yi¹, Hu Li¹, Wang Xiaojuan¹, Li Xun¹. 1 School of Public Health, Central South University, Changsha 410008, China; 2 Liuyang Hospital for Maternal and Child Health Care

Corresponding author: Tan Hongzhuang, Email: tanhz99@qq.com

This work was supported by a grant from the National Natural Science Foundation of China (No. 30872167).

【Abstract】 Objective To establish a practical and effective model in predicting the premature birth, using the Back-Propagation Artificial Neural Network (BPANN). **Methods** This was a prospective cohort study. Data was gathered from pregnant women selected by cluster sampling method from 2010 to 2012 in Liuyang city, Hunan province and was randomly divided into training sample (to establish the prediction models), validation sample (to select the optimal network) and testing sample (to evaluate the prediction models) by ratio of 2:1:1. BPANN and logistic regression analysis were used to establish models while ROC was applied to evaluate the 'prediction models'. **Results** Among the 6 270 pregnant women, 265 premature births were seen, with the premature incidence as 4.22%. The 7 variables which entered into the forecasting model would include abnormal uterine or uterine deformity, parity, number of pregnancies, gestational hypertension, placenta previa, premature rupture of membrane and regular prenatal examination. Sensitivity, specificity, agreement rate and area under the ROC curve of BPANN were 67.65%, 84.87%, 84.12% and 0.795, respectively. However, the sensitivity, specificity, agreement rate and area under the ROC curve of logistic regression were 64.71%, 85.60%, 84.69% and 0.783, respectively. **Conclusion** The newly established BPANN model was practical and reliable, which proved that this model was slightly better than the logistic regression in the prediction of premature birth.

【Key words】 Premature birth; Neural network; Prediction

早产是一种常见的妊娠时限异常^[1]。早产影响因素较多,采用合适方法进行预测十分关键。本研

究通过大样本的前瞻性队列研究,采用BP人工神经网络构建早产的预测模型,探索一种早期有效的识别早产高危个体的方法。

对象与方法

1. 研究对象:采用前瞻性队列研究方法,通过整群抽样,从湖南省浏阳市37个乡镇中随机抽取14个

DOI: 10.3760/cma.j.issn.0254-6450.2014.09.013

基金项目:国家自然科学基金(30872167)

作者单位:410008 长沙,中南大学公共卫生学院(徐新、谭红专、何玥、沈琳、柳祎、胡丽、王小娟、李逊);湖南省浏阳市妇幼保健院(周书进)

通信作者:谭红专, Email: tanhz99@qq.com

乡镇,选取从2010年1月1日起在抽样乡镇妇幼保健机构进行初次产检且愿意参加本研究的所有孕妇为研究对象。所有对象随访孕早期(<14孕周)、孕中期(14~27孕周)、孕晚期(>28孕周)、分娩各期,直至追踪到其妊娠结局。观察终点为产后24 h,观察终止时间为2012年6月30日。排除死胎和死产以及观察终止时未能获得活产妊娠结局的研究对象。

2. 研究方法:采用面对面问卷调查和检查相结合的方法收集资料。由调查员在当地妇幼保健机构进行面对面问卷调查,内容为研究对象进行初次产前检查时获取的孕妇年龄和既往妊娠情况(孕次、产次和流产次数)。由妇幼保健机构医务人员在初次产前检查时(近似孕前)获取孕妇身高、体重和血压数据;负责各次随访时的超声检测宫颈和实验室检查,以及结合临床表现获取孕妇的妊娠胎数、子宫异常及畸形、瘢痕子宫、生殖系统炎症、妊娠期糖代谢异常、病毒性肝炎、贫血情况、胎位异常、前置胎盘、胎膜早破及妊娠结局资料。参照文献[1-3]选择17个早产可疑影响因素(潜在预测指标)。同时,为保证研究数据准确可靠,课题组在调查前对所有参与调查人员进行培训,并统一检测标准和诊断标准^[1,2]。

3. 指标定义:根据文献[1,2]进行指标的定义与赋值。早产定义为妊娠满28周至不足37周分娩者。子宫异常及畸形包括子宫肌瘤和子宫畸形。生殖系统炎症包括外阴及阴道炎症、宫颈炎、盆腔炎性疾病和宫内感染。妊娠期糖代谢异常包括妊娠期糖尿病和糖耐量异常。定期产检是指根据产前保健要求,孕期产检次数 ≥ 9 ,其中孕早期 ≥ 1 、孕中期 ≥ 3 ,且孕晚期 ≥ 5 ;不定期产检是指产检次数 ≥ 6 ,其中孕早期为0,或孕中期 < 3 ,或孕晚期 < 5 ;产检次数少是指产检次数 ≤ 5 。

4. BP神经网络预测模型构建:基于人工神经网络的基本原理^[4]和方法,构建早产预测的数学模型。

5. 统计学分析:运用EpiData 3.0软件建立数据库,按统一标准录入数据,分析前进行逻辑查错和抽样复核,以确保资料录入的准确性。运用SPSS 18.0软件进行数据分析,使用 $M(QR)$ 对偏态分布资料进行统计描述,采用非条件logistic回归分析对潜在预测变量进行初筛并建立logistic回归模型;运用MATLAB 7.1软件编程建立BP神经网络模型;应用SPSS 18.0软件绘制2个模型的ROC曲线,使用曲线下面积描述模型的预测判别能力。

结 果

1. 基本情况:进入队列的样本为6 693例,最终追踪到完整活产妊娠结局资料者6 270例(排除死胎31例、死产12例以及观察终止时未追踪到妊娠结局者380例),有效应答率为93.68%。共诊断早产265例,早产发生率为4.22%。研究对象的年龄为16~43($M=25$, $QR=4$)岁,分娩时孕周为28~42($M=39$, $QR=2$)周。

2. logistic回归预测模型建立:将样本随机分为50%的训练样本(共3 135例,早产131例)、25%的检验样本(共1 567例,早产66例)和25%的测试样本(共1 568例,早产68例)。利用训练样本,采用非条件logistic回归方法,以是否早产为应变量,以本研究收集到的17个早产的可疑影响因素为自变量进行单因素和多因素分析,筛选出有统计学意义的自变量。单因素分析结果显示,7个因素与早产的关系有统计学意义($P<0.05$, $\alpha=0.05$),单因素回归分析的变量赋值及结果见表1。将单因素分析有统计学意义的自变量进行多因素非条件logistic逐步回归分析(偏最大似然估计后退法,入选标准 $\alpha=0.05$,剔除标准 $\alpha=0.10$),最终7个变量均进入方程,多因素回归分析的结果见表2。以此分析结果建立的logistic回归模型为 $\text{logit}(P) = \ln [P/(1-P)] = 1.259X_1 + 0.602X_2 + 2.135X_3 + 1.301X_4 + 0.722X_5 + 1.660X_6 - 0.736X_7 - 2.813$,其中 P 为发生早产的概率。

3. BP神经网络预测模型建立:以logistic回归分析筛选出的子宫异常及畸形、产次、妊娠胎数、妊娠期高血压、前置胎盘、胎膜早破和定期产检7个变量为输入层神经元,以是否早产为输出层神经元,构建BP神经网络预测模型,神经网络的变量赋值同logistic回归一致。神经网络的输入层神经元数为7,输出层神经元数为1,隐含层神经元数根据 $m = \sqrt{7+1} + a$ (a 为[1,10]之间的常数)计算而得(3~13),利用检验样本通过凑试法反复训练发现,当 $m=3$ 时,检验样本的MSE最小,因此确定隐含层的最佳神经元数为3。

4. 最佳临界点的确定:用BP神经网络模型对测试样本进行预测。ROC曲线提示,当输出值为0.06时,得到最大约登指数(0.525 2),即0.06为最佳临界点,此时预测的灵敏度为67.65%,特异度为84.87%,一致率为84.12%,即此模型可筛出目标人群中67.65%的早产患者。将不同输出值作为判别界值时测试样本的预测效果见表3。

表 1 单因素 logistic 回归分析的变量赋值与结果

因素	赋值说明	β	Wald χ^2 值	P值	OR值
年龄(岁)	哑变量,以 20~35 为参考组		2.538	0.281	
	<20=1,其他=0	1.361	1.576	0.209	3.900
	>35=1,其他=0	0.472	1.002	0.317	1.603
BMI(kg/m ²)	正常=0,异常=1	0.275	2.156	0.142	1.317
子宫异常及畸形	无=0,有=1	1.363	6.198	0.013	3.910
瘢痕子宫	无=0,有=1	0.407	3.269	0.071	1.503
孕次	1=0,≥2=1	0.152	0.684	0.408	1.164
产次	初产=0,经产=1	0.618	11.896	0.001	1.856
流产次数	无=0,有=1	0.189	1.122	0.289	1.209
妊娠胎数	1=0,≥2=1	2.131	22.562	0.000	8.423
生殖系统炎症	无=0,有=1	0.114	0.135	0.713	1.121
妊娠期高血压疾病	无=0,有=1	1.114	9.285	0.002	3.047
妊娠期糖代谢异常	无=0,有=1	0.410	0.463	0.496	1.507
病毒性肝炎	无=0,有=1	-0.341	0.646	0.422	0.711
贫血	无=0,轻度=1,轻度以上=2	0.120	0.448	0.503	1.128
胎位异常	无=0,有=1	-0.386	0.560	0.454	0.680
前置胎盘	无=0,有=1	0.770	5.956	0.015	2.160
胎膜早破	无=0,有=1	1.524	69.482	0.000	4.591
定期产检	少=1,不定期=2,定期=3	-0.771	30.548	0.000	0.463

表 2 多因素 logistic 回归分析的结果

因素	β	s_e	Wald χ^2 值	P值	OR值(95%CI)
子宫异常及畸形(X_1)	1.259	0.594	4.498	0.034	3.521(1.100~11.271)
产次(X_2)	0.602	0.191	9.941	0.002	1.827(1.256~2.656)
妊娠胎数(X_3)	2.135	0.503	18.015	0.000	8.459(3.156~22.673)
妊娠期高血压(X_4)	1.301	0.392	10.990	0.001	3.672(1.702~7.924)
前置胎盘(X_5)	0.722	0.332	4.717	0.030	2.058(1.073~3.948)
胎膜早破(X_6)	1.660	0.193	74.247	0.000	5.261(3.606~7.675)
定期产检(X_7)	-0.736	0.144	26.114	0.000	0.479(0.361~0.635)

表 3 不同输出值下 BP 人工神经网络模型的预测效果

输出值	灵敏度 (%)	特异度 (%)	约登指数 (%)	一致率 (%)
0.01	86.76	35.07	21.83	37.31
0.02	82.35	45.53	27.88	47.13
0.03	82.35	46.93	29.28	48.47
0.04	82.35	50.87	33.22	52.23
0.05	67.65	84.33	51.98	83.61
0.06	67.65	84.87	52.52	84.12
0.07	66.18	85.40	51.58	84.57
0.08	58.82	90.00	48.82	88.65
0.09	58.82	91.20	50.02	89.80

5. 模型预测结果的评价:根据建立的 logistic 回归模型和 BP 人工神经网络模型,分别对测试样本进行预测。其中 logistic 回归模型的最佳临界点确定方法与 BP 人工神经网络相同,选取最大约登指数(0.503 1)对应的输出值 0.05 为最佳临界点。BP 人工神经网络模型略优于 logistic 回归模型,见表 4。

讨 论

本研究数据来源于大样本的前瞻性队列研究。

结果显示,2010—2012 年湖南省浏阳市分娩孕妇中早产发生率为 4.22%,略低于国内其他研究的 5%~15%^[1]。

由于临床上早产的并发症发生率和死亡率均较高,围生医学领域迫切需要一个指标较少、易于操作以及能早期发现高危个体的预测模型。本研究旨在构建能优先于临床诊断的早产预测模型,首先,预测指标的选择既应考虑其临床意义,又应考虑其公共卫生学意义和实际可能性(即选择早期能常规获得的指标)。根据这些原则,本研究选择 17 个能够反映孕产妇一般情况、既往妊娠情况和此次妊娠情况的早产的可能预测指标。

为了保证预测模型简化实用,本研究采用回归分析对可疑影响因素进行初筛,从 17 个指标中筛选出 7 个与早产发生相关的影响因素,即子宫异常及畸形、产次、妊娠胎数、妊娠期高血压、前置胎盘、胎膜早破和定期产检,可作为预测指标参与预测,与其他研究相符^[5-8]。子宫异常及畸形(反映生殖器机能状

况)与产次能反映孕产妇的身体机能状态;较高的产次预示着较高的早产风险,但无法排除年龄这一混杂因素的影响。妊娠胎数、妊娠期高血压疾病、前置胎盘和胎膜早破易诱发早产,能反映此次妊娠情况。定期产检反映了产前保健情况,充分利用卫生服务,可减少不良妊娠结局的发生风险。因此,据此建立的早产 BP 人工神经网络预测模型和 logistic 回归预测模型能较准确的预测早产的发生情况,具有临床实用价值。

由于早产是多因素疾病,影响因素较多,且各因素之间可能存在相关关系,因此选择合适的方法建

表 4 2 个模型预测结果的评价指标

指 标	logistic 回归模型	BP 人工神经网络模型
灵敏度(%)	64.71	67.65
特异度(%)	85.60	84.87
约登指数(%)	50.31	52.52
一致率(%)	84.69	84.12
ROC 曲线下面积	0.783	0.795

立预测模型十分关键。而 BP 神经网络对被分析的资料特性几乎没有任何要求,并行处理的方式使其对误差信号具有一定的处理能力,同时还具有自动学习、识别变量间关系的能力^[9],其在疾病预测方面有较高的应用价值。本研究通过比较 BP 人工神经网络与 logisitc 回归方法建立的预测模型,发现 BP 神经网络预测模型较 logisitc 回归预测模型有更强的预测判别效能。

本研究建立的 BP 人工神经网络模型的灵敏度、特异度、一致率和 ROC 曲线下面积分别为 67.65%、84.87%、84.12% 和 0.795。与 Catley 等^[7]构建的早产 BP 人工神经网络模型相比,本研究增加了反映身体机能及此次妊娠情况的指标,从而使模型灵敏度与 ROC 曲线下面积较高。与 Lee 等^[8]构建的模型在灵敏度、特异度和一致率方面相近。

参 考 文 献

- [1] Huang YP. Premature delivery[M]//Le J. Obstetrics and Gynecology. 7th ed. Beijing: People's Medical Publishing House, 2008: 86-89. (in Chinese)
黄引平. 早产[M]//乐杰. 妇产科学. 7 版. 北京:人民卫生出版社, 2008: 86-89.
- [2] The Obstetric Group of the Branch of Obstetrics and Gynecology of Chinese Medical Association. The recommended guidelines of clinical diagnosis and treatment of premature (Draft) [J]. Chin J Obstet Gynecol, 2007, 42(7): 498-500. (in Chinese)
中华医学会妇产科学分会产科学组. 早产的临床诊断与治疗推
- [3] Guo LJ, Ye RW, Wang GX, et al. Birth weight distribution among premature infants and related social factors [J]. Chin J Epidemiol, 2009, 30(12): 1243-1247. (in Chinese)
郭丽君, 叶荣伟, 王桂霞, 等. 江苏、浙江省 4 县(市)早产儿体重分布状况及相关社会因素 [J]. 中华流行病学杂志, 2009, 30(12): 1243-1247.
- [4] Wang W, Xu W, Zheng YJ, et al. Study on a back propogation neural network-based predictive model for prevalence of birth defect [J]. Chin J Epidemiol, 2007, 28(5): 507-509. (in Chinese)
王伟, 许伟, 郑亚军, 等. 基于 BP 神经网络的围产儿出生缺陷患病率预测 [J]. 中华流行病学杂志, 2007, 28(5): 507-509.
- [5] Creasy RK, Gummer BA, Liggins GC. System for predicting spontaneous preterm birth [J]. Obstet Gynecol, 1980, 55(6): 692-695.
- [6] Courtney KL, Stewart S, Popescu M, et al. Predictors of preterm birth in birth certificate data [J]. Stud Health Technol Inform, 2008, 136: 555-560.
- [7] Catley C, Frize M, Walker CR, et al. Predicting high-risk preterm birth using artificial neural networks [J]. IEEE Trans Inf Technol Biomed, 2006, 10(3): 540-549.
- [8] Lee KA, Chang MH, Park MH, et al. A model for prediction of spontaneous preterm birth in asymptomatic women [J]. J Womens Health (Larchmt), 2011, 20(12): 1825-1831.
- [9] Li YY, Li R, Shi L, et al. Study on application of back-propagation neural network model in screening for type 2 diabetes Mellitus in community population [J]. J Environ Occup Med, 2008, 25(4): 329-332. (in Chinese)
黎衍云, 李锐, 施亮, 等. 逆反馈神经网络在社区糖尿病筛查中的应用 [J]. 环境与职业医学, 2008, 25(4): 329-332.

(收稿日期: 2014-03-24)

(本文编辑: 万玉立)

中华流行病学杂志第六届编辑委员会成员名单

总编辑 李立明

副总编辑 乌正赓 曲成毅 王滨有 姜庆五 何 耀 詹思延

编辑委员 按姓氏拼音排列

毕振强(山东)	曹广文(上海)	曹务春(北京)	陈 坤(浙江)	陈维清(广东)	董柏青(广西)
段广才(河南)	龚向东(江苏)	顾东风(北京)	何 耀(北京)	贺 雄(北京)	胡永华(北京)
姜庆五(上海)	阚 飙(北京)	李 辉(北京)	李敬云(北京)	李立明(北京)	梁万年(北京)
刘殿武(河北)	刘 民(北京)	刘天锡(宁夏)	陆 林(云南)	栾荣生(四川)	吕 繁(北京)
马文军(广东)	毛宗福(湖北)	孟 蕾(甘肃)	米 杰(北京)	潘凯枫(北京)	潘先海(海南)
乔友林(北京)	曲成毅(山西)	瞿世和(新疆)	沈洪兵(江苏)	时景璞(辽宁)	孙瑞华(北京)
谭红专(湖南)	唐耀武(北京)	汪 华(江苏)	汪 宁(北京)	王滨有(黑龙江)	王建华(天津)
王克安(北京)	王 鸣(广东)	王声湧(广东)	王擷秀(天津)	吴 凡(上海)	吴先萍(四川)
吴尊友(北京)	武阳丰(北京)	乌正赓(北京)	项永兵(上海)	肖东楼(北京)	徐 飏(上海)
许汴利(河南)	闫永平(陕西)	严延生(福建)	杨维中(北京)	叶冬青(安徽)	于普林(北京)
于雅琴(吉林)	曾 光(北京)	詹思延(北京)	张国刚(湖南)	张建中(北京)	张孔来(北京)
张顺祥(广东)	赵 冬(北京)	赵仲堂(山东)	庄 辉(北京)		

名誉总编辑 郑锡文

顾 问 魏承毓 吴系科 施偲元 俞顺章

秘 书 王 岚(北京)