基于人工神经网络的结直肠癌预测模型研究

吴维妙[1],徐东丽[2],李小强[1],李为希[2],李　俊[2],张　芬[2],许慧琳[2],徐望红[1],姚保栋[2]
(1.复旦大学公共卫生学院,上海 200032;2.上海市闵行区疾病预防控制中心,上海 201101)

摘　要:[目的]筛选结直肠癌危险因素,建立人工神经网络(artificial neural network,ANN)模型,为筛查人群风险分层与优化筛查方案提供科学依据,提高筛查成本效果。[方法]2012年上海市闵行区163 240名50~80岁社区居民参加结直肠癌筛查。初筛采用调查表、粪便隐血试验(FOBT)和肛门指检,阳性者进一步做结肠镜和病理检查,诊断有无结直肠癌,同肿瘤登记链接补充筛查后2年内结直肠癌诊断信息。采用单因素和多因素分析筛选变量,研究对象按1:1的比例随机分为训练集和验证集,分别用于建立和验证ANN模型。采用灵敏度、特异性、ROC曲线下面积(AUC)等指标评价模型。[结果]163 240名研究对象中新确诊结直肠癌病例363例。多因素Logistic回归筛选出年龄、性别、便秘、里急后重、血便、进行性消瘦与FOBT阳性7个变量作为ANN输入变量,其中,血便(20.8%)、年龄(18.1%)和FOBT阳性(17.1%)对结直肠癌的影响最大。训练集和验证集灵敏度分别为65.93%(95%CI:58.78%~72.43%)、60.22%(95%CI:52.95%~67.07%),特异性分别为62.07%(95%CI:61.74%~62.40%)、61.92%(95%CI:61.59%~62.26%),AUC分别为0.68(95%CI:0.64~0.72)、0.67(95%CI:0.63~0.70),验证集符合率为61.92%(95%CI:61.59%~62.26%),模型内部验证效果较好。[结论]本研究所建结直肠癌ANN预测模型总体区分度较好、诊断价值较高,但有待进一步优化和外部验证以提高其预测准确性和泛化能力。

关键词:结直肠肿瘤;预测模型;人工神经网络;贡献分析
中图分类号:R735.3　　**文献标识码**:A　　**文章编号**:1004-0242(2019)08-0621-08
doi:10.11735/j.issn.1004-0242.2019.08.A011

A Predictive Model for Colorectal Cancer Based on Artificial Neural Network

WU Wei-miao[1],XU Dong-li[2],LI Xiao-qiang[1],LI Wei-xi[2],LI Jun[2],ZHANG Fen[2],XU Hui-lin[2],XU Wang-hong[1],YAO Bao-dong[2]
(1. School of Public Health,Fudan University,Shanghai 200032,China;2. Minhang District Center for Disease Control and Prevention,Shanghai 201101,China)

Abstract:[Purpose] To establish an artificial neural network (ANN) model for prediction of colorectal cancer risk. [Methods] Total 163 240 residents aged 50~80 years in Minhang District of Shanghai participated in colorectal cancer screening in 2012. Questionnaires,fecal occult blood tests (FOBT) and anal examinations were conducted for initial screening of colorectal cancer. Further colonoscopy and biopsy were performed for subjects with a positive initial screening result. The data of identified colorectal cancer patients in 2 years after the screening were collected from cancer registry. Univariate and multivariate analysis were used to select variables for establishing ANN prediction model. The participants were randomly classified as training set and validation set by 1:1 ratio for establishing and verifying the model. Sensitivity,specificity,AUC (area under the receiver operating characteristic curve) were calculated for model evaluation. [Results] In 163 240 screening participants,363 colorectal cancer cases were diagnosed. Seven variables including age,gender,constipation,rectal tenesmus,bloody stool,progressive emaciation and FOBT results were selected as the input variables of the ANN model by the multivariate logistic regression. Bloody stools(20.8%),age(18.1%) and positive FOBT(17.1%) were the most important predictive variables for colorectal cancer. Sensitivity of the model in the training set and validation set were 65.93% (95% CI:58.78%~72.43%) and 60.22% (95% CI:52.95%~67.07%),respectively. Specificity were 62.07%(95%CI:61.74% ~62.40%) and 61.92%(95%CI:61.59%~62.26%),respectively. AUC were 0.68(95% CI:0.64~0.72) and 0.67(95% CI:0.63~0.70),respectively. The agreement rate was 61.92% (95% CI:61.59%~62.26%) in the validation set. The internal validation effect of the model was performed well. [Conclusion] The ANN predictive model for colorectal cancer established in this

收稿日期:2018-11-08;修回日期:2018-12-21
基金项目:上海市第四轮公共卫生行动计划重点学科建设课题(15GWZK0801)
通信作者:姚保栋,E-mail:yaobaodong2008@126.com

study is performed well in the risk stratification, and is of good diagnostic value. However, further optimization and external validation are needed to improve its prediction accuracy and generalization ability.
Key words: colorectal cancer; predictive model; artificial neural network; contribution analysis

结直肠癌(colorectal cancer, CRC)又称结直肠癌,是我国最常见的恶性肿瘤之一,分别位居恶性肿瘤发病和死亡的第4位和第5位。我国肿瘤登记数据显示,2013年结直肠癌新发病例34.79万,标化发病率为17.45/10万,2013年全国结直肠癌死亡病例16.49万,标化死亡率为7.87/10万[1]。

目前结直肠癌的确切病因尚不完全清楚,但结直肠癌发病率高、临床前期长、具有可识别和检测的癌前病变、预后与肿瘤分期息息相关的特点使其适于通过筛查而实现结直肠癌的早发现、早诊断和早治疗,从而降低结直肠癌的发病率和死亡率[2-3]。然而,国内外研究均显示以早期诊断为目的的结直肠癌筛查项目效率不高,在大量人群中仅筛出极少部分病例,以致人群筛查的依从性不高[4-5]。鉴于结直肠癌是一种多因素疾病,若能利用年龄、性别、结直肠癌家族史等易于获得的危险因素信息建立风险预测模型,对筛查人群进行风险分层,对于提高筛查效能、降低人群筛查费用与充分发挥有限卫生资源的作用具有重要意义[6-7]。

英国、美国、日本等已开展一系列结直肠癌预测模型的研究,建模方法包括人工神经网络(artificial neural network, ANN)[8]、Cox比例风险回归[9]、Logistic回归[10-11]、决策树[12]等,呈现多元化发展,目前已建成Qcancer预测模型[13-15]、亚太地区结直肠癌筛查评分(Asia-pacific colorectal screening, APCS)评分模型[16]等,均具有地区特异性。国内结直肠癌风险预测模型的研究基于小样本的病例对照研究或队列研究为多且以分子生物标志物的研究为主[17-18],不能满足识别结直肠癌高危人群对简单、方便、低成本工具的需求。综合考虑我国结直肠癌较重的疾病负担与对简单、方便、低成本筛查工具的需求,本研究以2012年上海市闵行区社区居民结直肠癌筛查人群为研究对象,通过分析问卷调查信息、便隐血试验(fecal occult blood tests, FOBT)与肛门指检结果,构建ANN结直肠癌预测模型,为优化上海市社区居民结直肠癌筛查策略提供依据。

1 资料与方法

1.1 研究对象

2012年上海市闵行区开展了一项针对社区居民的结直肠癌筛查项目。研究对象纳入标准包括:①上海市闵行区常住居民,包括本市户籍居民和本市居住满6个月以上的非本市户籍居民;②年龄在50~80岁;③无结直肠癌史;④参加本市各类基本医疗保险和基本医疗保障;⑤了解结直肠癌筛查的目的与意义及参加筛查的获益与风险,自愿参加筛查并签署《上海市闵行区社区居民结直肠癌筛查知情同意书》。

1.2 数据收集

采用问卷调查对所有对象收集人口学特征等信息,根据风险评分标准,如症状或体征(便频、便秘、腹泻便秘交替、血便、黏液便、里急后重、大便变形、大便变细、腹部固定痛、腹胀痛、进行性消瘦、不明原因贫血)、一级亲属结直肠癌家族史、肠道疾病史(家族性腺瘤和息肉、溃疡性结肠炎、克隆病)中任何一项阳性即判为初筛阳性。此外,肛指或2次便隐血检查任一次阳性也判为初筛阳性。

初筛阳性者嘱其自行前往医院做结肠镜和病理检查,确诊有无结直肠癌。考虑到肿瘤的潜隐期,本次将筛查后2年内新诊断结直肠癌病例均视为本次筛查中新发的病例。利用研究对象身份证号与肿瘤登记进行记录联动,补充其参加筛查后2年内的结直肠癌诊断信息。高质量且全覆盖的上海市肿瘤登记系统保证了数据资料的完整性与可信度。本次结直肠癌诊断依据为国际疾病分类肿瘤学专辑第1版(ICD-O-1)(其中,C18为结肠癌、C19为直肠与乙状结肠连接处癌、C20为直肠癌)。

1.3 统计学分析

本研究统计分析均采用SAS 9.4软件完成。采

用 t 检验(连续变量)、χ^2 检验(分类变量)等方法,比较病例和非病例两组间结直肠癌可能的危险因素及 FOBT 结果的差异,将差异具有统计学意义的危险因素(P<0.15)纳入多因素 Logistic 回归模型,采用"后退法"进行变量筛选。最后在 SAS 9.4 软件的 EM 模块(SAS Enterprise Miner Workstation 13.2)中利用 ANN 方法建立和验证模型,并采用灵敏度、特异性、ROC 曲线下面积(area under the receiver operating characteristic curve,AUC)、阳性预测值(positive predictive value,PPV)、阴性预测值(negative predictive value,NPV)、阳性似然比(positive likelihood ratio,+LR)、阴性似然比(negative likelihood ratio,–LR)等指标评价模型的准确性、区分度和收益。

2 结 果

2.1 一般人口学特征

本研究共纳入上海市闵行区 2012 年结直肠癌筛查人群 163 240 人,其中病例组 363 人,非病例组 162 877 人,两组在婚姻状况、文化程度与职业上差异不具有统计学意义,但在年龄与性别上差异具有统计学意义(P<0.05),病例组平均年龄高于非病例组,且病例组以男性居多而非病例组以女性居多(Table 1)。

2.2 结直肠癌危险因素分析

结直肠癌危险因素通过构成比描述统计、χ^2 检验比较分析,结果显示两组在便秘、里急后重、大便变细、血便、黏液便、下腹部胀痛明显、进行性消瘦、便隐血(FOBT)检查结果上差异具有统计学意义(P<0.05),而其他危险因素的分布差异均不具有统计学意义(Table 2)。

2.3 多因素 Logistic 回归分析

将单因素分析中 P<0.15 的因素纳入多因素 Logistic 回归进行变量筛选,采用"后退法"进行逐步回归分析,共筛出 7 个变量,其中,血便与进行性消瘦的 OR 值最大,表明其在病例组的相对暴露比值最高,其次是里急后重和 FOBT 阳性(Table 3)。

2.4 模型的建立与验证

将筛出的 7 个变量作为输入神经元,是否发生结直肠癌作为输出神经元,不断调整 ANN 建模参数以使模型最优化,最终确定"7-6-6-1"四层结构模型,不同节点间连接具有经反复训练获得,使实际输出与期望输出误差最小的神经网络权值,其绝对值大小反映变量的相对重要性(Figure 1)。

年龄变量对第一隐含层的 H1、H2、H4、H6 的权值均最大,血便变量对 H3 的权值最大,进行性消瘦对 H5 的权值最大,可见不同输入层神经元对不同第一隐含层神经元影响不同(Table 4)。对第二隐含层神经元影响最大的第一隐含层神经元均是 H4,说明 H4 在两个隐含层之间作用影响最大(Table 5)。总体上输入层神经元较第二隐含层神经元对输出层神经元的影响更大(Table 6、7),其中,血便直接对输出层神经元的权值最大。根据高仁祥等[19]的贡献分析计算方法:

$$Ci=[\sum_{k=1}^{r} W_k*(\sum_{j=1}^{q} V_{kj}*U_{ji})]+W_i, i=1,\cdots p \quad (1)$$

其中,U_{ji} 为第 i 个输入对第一隐含层第 j 个隐含节点的贡献,V_{kj} 为第一隐含层第 j 个隐含节点对

Table 1 Demographic characteristics of the study population

Variables	Case group (n=363)	Non-case group (n=162877)	Z/χ^2	P
Age(years)	67.0±7.7	62.9±7.8	–10.000	<0.001
Gender				
Male	191(52.62%)	69046(42.39%)	15.506	<0.001
Female	172(47.38%)	93831(57.61%)		
Marital status				
Unmarried	2(0.55%)	328(0.20%)	4.378	0.224
Married	350(96.42%)	157362(96.61%)		
Divorce	0(0%)	887(0.54%)		
Widowed	11(3.03%)	4300(2.64%)		
Educational level				
Primary school	95(26.17%)	36120(22.18%)	3.356	0.187
High school or polytechnic school	248(68.32%)	11744(72.13%)		
College or above	20(5.51%)	9273(5.69%)		
Occupation				
Government officials	3(0.83%)	546(0.34%)	7.572	0.181
Institution staffs	17(4.68%)	8393(5.15%)		
Businessmen	175(48.21%)	74136(45.52%)		
Farmers	68(18.73%)	26722(16.41%)		
Other	93(25.62%)	48609(29.84%)		
Unemployed	7(1.93%)	4471(2.74%)		

Table 2　Univariate analysis of colorectal cancer risk factors[n(%)]

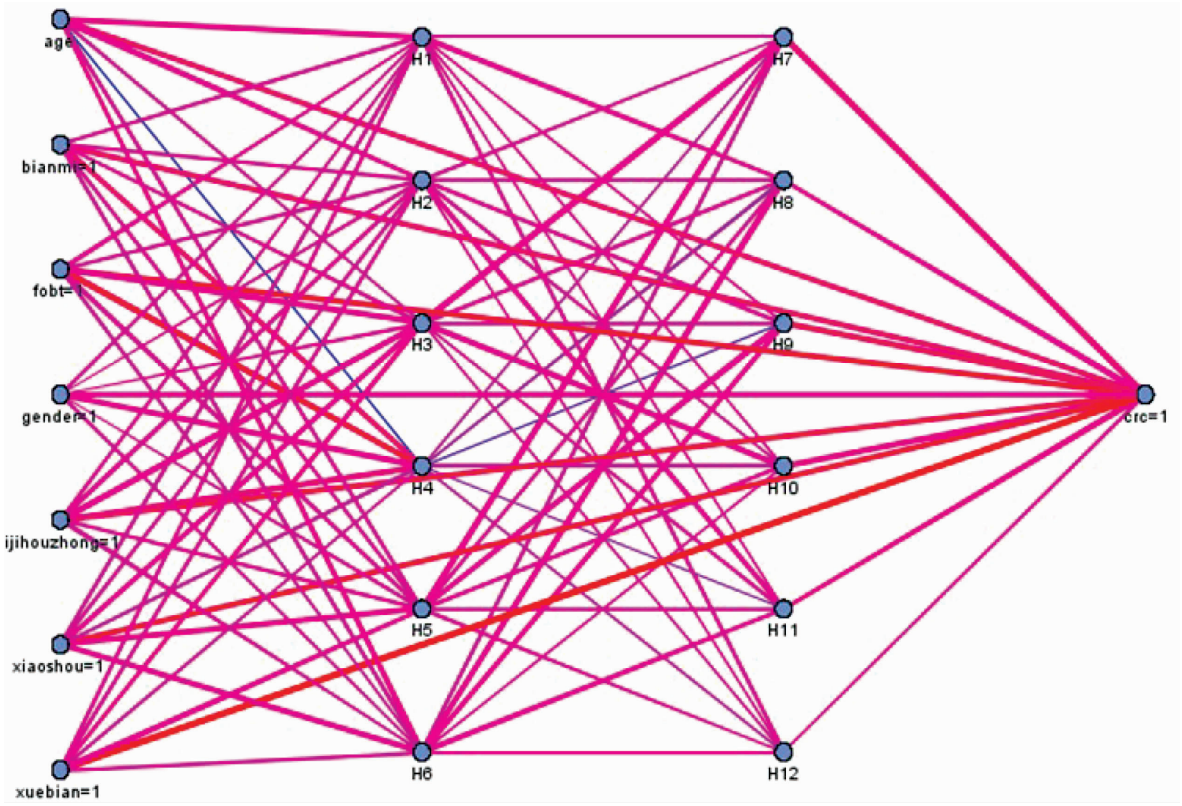
Variables	Case group (n=363)	Non-case group (n=162877)	χ^2	P
Constipation				
Yes	20(5.51)	4466(2.74)	10.381	0.001
No	343(94.49)	158411(97.26)		
Frequent stool				
Yes	7(1.93)	2260(1.39)	0.774	0.379
No	356(98.61)	160617(98.61)		
Diarrhea and constipation				
Yes	2(0.55)	669(0.41)	–	0.664△
No	361(99.45)	162208(99.59)		
Rectal tenesmus				
Yes	8(2.20)	549(0.34)	–	<0.001△
No	355(97.80)	162328(99.66)		
Slender stool				
Yes	4(1.10)	511(0.31)	–	0.029△
No	359(98.90)	162366(99.69)		
Shapeless stool				
Yes	3(0.83)	537(0.33)	–	0.120△
No	360(99.17)	162340(99.67)		
Bloody stool				
Yes	6(1.65)	255(0.16)	–	<0.001△
No	357(98.35)	162622(99.84)		
Mucous stool				
Yes	3(0.83)	188(0.12)	–	0.009△
No	360(99.17)	162689(99.88)		
Fixed abdominal pain or discomfort				
Yes	3(0.83)	439(0.27)	–	0.077△
No	360(99.17)	162438(99.73)		
Lower abdominal pain				
Yes	5(1.38)	582(0.36)	–	0.011△
No	358(98.62)	162295(99.64)		
Progressive emaciation				
Yes	2(0.55)	72(0.04)	–	0.012△
No	361(99.45)	162805(99.96)		
Unknown anemia				
Yes	1(0.28)	85(0.05)	–	0.174△※
No	362(99.72)	162779(99.95)		
Family history of CRC[a]				
Yes	0(0.00)	209(0.13)	–	0.628△
No	363(100.00)	162668(99.87)		
Familial adenomas and polyps				
Yes	0(0.00)	209(0.13)	–	1.000△
No	363(100.00)	162668(99.87)		
Other intestinal diseases[b]				
Yes	0(0.00)	62(0.04)	–	1.000△
No	363(100.00)	162815(99.96)		
Positive FOBT				
Yes	8(2.20)	876(0.54)	–	0.001△
No	355(97.80)	162001(99.46)		
Positive anal examination				
Yes	33(9.09)	16926(10.39)	–	0.417△
No	330(90.91)	145951(89.61)		

Note:△:Fisher's Exact Test; ※:the variable was with 13 missing values;[a]:including grandparents, parents and siblings;[b]:including ulcerative colitis,crohn's disease,familial carcinomatous syndrome villous or tubular adenomas.

第二隐含层第 k 个隐含节点的贡献,W_k 为第二隐含层第 k 个隐含节点对输出的贡献,W_i 为第 i 个输入直接对输出的贡献,Ci 为第 i 个输入对输出的总贡献。最后算得各输入对输出影响比重依次为血便 20.8%、年龄 18.1%、FOBT 阳性 17.1%、进行性消瘦 15.8%、里急后重 13.2%、便秘 10.3%、性别 4.7%(Table 8)。

2.5　模型的评价

建模中数据集分配比例为训练:验证=1:1,其确诊病例数分别有 182 例和 181 例。所建模型灵敏度和特异性大小取决于所规定阳性结果截断值(cut-off point)。当截断值取 0.45 时,灵敏度和特异性在训练集分别为 74.73%、49.18%,验证集分别为 73.48%、49.11%,验证集符合率为 49.17%。当截断值取 0.5 时,灵敏度和特异性在训练集分别为 65.93%、62.07%,验证集分别为 60.22%、61.92%,验证集符合率为 61.92%。当截断值取 0.55 时,灵敏度和特异性在训练集分别为 47.80%、76.29%,验证集分别为 42.54%、76.36%,验证集符合率为 76.29%。综合考虑模型预测准确性、精确性,本研究截断值取 0.5,即当个体结直肠癌预测概率 P>0.5 时判为阳性结果。AUC 反映筛检试验的区分度大小,本研究训练集、验证集中 AUC 分别为 0.68 和 0.67,均接近 0.7,说明所建模型区分度较好。NPV 接近 100%,而 PPV 仅 0.4%左右,主要与人群中结直肠癌患病率较低有关。而似然比的计算只涉及灵敏度与特异性,不受患病率影响,较全面反



Note: blue, red and purple connecting lines indicated negative weights, positive weights and small weights, respectively.

Figure 1 Diagram of artificial neural network (ANN) model

Table 3 Multivariate analysis of colorectal cancer risk factors

Independent variables	β	SE	Wald χ^2	*P* value	OR(95%CI)
Age(years)	0.061	0.007	89.640	<0.001	1.06 (1.05~1.08)
Gender (male *vs* female)	0.178	0.053	11.314	0.001	1.43 (1.16~1.76)
Constipation	0.264	0.118	4.960	0.026	1.70 (1.07~2.70)
Rectal tenesmus	0.842	0.190	19.593	<0.001	5.39 (2.56~11.36)
Bloody stool	1.033	0.216	22.889	<0.001	7.90 (3.39~18.42)
Progressive emaciation	0.976	0.393	6.166	0.013	7.04 (1.51~32.89)
Positive FOBT	0.655	0.180	13.187	<0.001	3.70 (1.83~7.51)

Table 4 Weights between input layer and the first hidden layer neurons

The first hidden layer neuron	Input layer neuron						
	Age	Constipation	Positive FOBT	Gender	Rectal tenesmus	Progressive emaciation	Bloody stool
H1	0.47	−0.14	0.09	−0.19	−0.03	−0.28	0.01
H2	0.21	0.01	−0.16	−0.10	0.01	−0.17	−0.10
H3	−0.30	−0.02	0.28	0.05	0.20	0.20	0.48
H4	−3.73	0.84	1.34	0.31	0.67	−0.66	−0.37
H5	−0.13	−0.04	−0.09	−0.10	0.07	0.44	0.22
H6	0.45	−0.28	−0.38	−0.20	−0.14	0.17	−0.11

映出筛检试验的诊断价值，非常稳定。本研究中−LR 为 0.6 左右，+LR 接近 2，−LR 越小，+LR 越大，筛检试验的诊断价值越高。可见，各指标在验证集中均较训练集中稍差一点，但基本接近，说明该模型的内部验证效果较好(Table 9，Figure 2)。

3 讨 论

近年来，国内外学者通过分析结直肠癌危险因素建立预测模型对筛查人群进行风险分层与预测判别的研究层出不穷，如 Yeoh 等[16]通过多因素 Logistic 回归建立的 APCS 评分模型可对筛查人群进行风险分层；陈锴等[20]通过多因素

Table 5 Weights between the first and second hidden layer neurons

The second hidden layer neuron	The first hidden layer neuron					
	H1	H2	H3	H4	H5	H6
H7	−0.36	−0.15	0.30	−0.89	0.30	−0.35
H8	0.17	0.003	−0.02	−1.23	−0.03	0.27
H9	0.07	0.03	−0.20	−2.74	0.16	0.36
H10	0.05	−0.16	0.17	−0.42	0.10	−0.07
H11	0.02	0.16	−0.05	−1.80	−0.002	0.22
H12	−0.08	0.08	−0.13	−0.91	−0.01	0.05

Table 6 Weights between input layer and output layer neurons

Output layer neuron	Input layer neuron						
	Age	Constipation	Positive FOBT	Gender	Rectal tenesmus	Progressive emaciation	Bloody stool
CRC=1	0.90	0.88	1.44	0.34	1.26	1.41	2.10

Table 7 Weights between the second hidden layer and output layer neurons

Output layer neuron	The second hidden layer					
	H7	H8	H9	H10	H11	H12
CRC=1	0.66	0.22	1.07	0.52	0.47	0.03

Table 8 Contributions of the input variables to the output

CRC=1	Age	Constipation	Positive FOBT	Gender	Rectal tenesmus	Progressive emaciation	Bloody stool
Percentage(%)	18.1	10.3	17.1	4.7	13.2	15.8	20.8

Table 9 Model evaluation indexes in the training and validation set

Evaluation index (95%CI)	Training set (n=81621)	Validation set (n=81619)
Sensitivity(%)	65.93(58.78~72.43)	60.22(52.95~67.07)
Specificity(%)	62.07(61.74~62.40)	61.92(61.59~62.26)
AUC	0.68(0.64~0.72)	0.67(0.63~0.70)
NPV(%)	99.88(99.84~99.90)	99.86(99.82~99.89)
PPV(%)	0.39(0.32~0.46)	0.35(0.29~0.42)
−LR	0.55(0.45~0.67)	0.64(0.54~0.77)
+LR	1.74(1.57~1.93)	1.58(1.40~1.78)

Logistic 回归建立的结直肠癌早期筛查高危评分模型有助于发现结直肠癌高风险人群；Hippisley-Cox 和 Coupland[14]通过 Cox 比例风险回归建立的男、女性QCancer 预测模型分别用于识别男、女性外表健康人群中虽患结直肠癌但尚未诊断的无症状患者。本次所建 ANN 预测模型旨在为每位研究对象“量体裁衣”地给出个性化的预测结果，从而实现有效的风险分层与预测判别[21]。结直肠癌作为一种慢性非传染性疾病，影响因素众多、作用方式复杂，若将流行病学资料作为基础、利用传统线性判别函数的“刚性”方法进行疾病状态预测则存在较大局限性[21]。ANN 作为一种生物神经网络在结构、功能及某些特性上抽象形成的信息处理系统，具有自学习和识别变量间关系的能力，能以任意精度逼近任意非线性函数，个体疾病预测与辅助诊断是 ANN 医学应用中相对活跃的领域。

本次所建 ANN 模型包含 12 个神经元的两层（直连）隐含层，隐含层及其神经元数取决于多种因素，如输入层和输出层神经元数目、训练样本大小、样本噪音大小、所面对问题的复杂程度等[22]。样本量较大但结直肠癌发病率低、病例数较少，是本次网络结构选择的重要影响因素之一。采用“试错法”分别建立 ANN 模型，其中，隐含层神经元数(hidden nodes，HN)的选择综合考虑了国内外研究中提到的方法：HN=M*O/W（其中，M 为训练样本量，O 为输出层神经元数，W 为权值数）[8]、HN=L+(N+O)1/2（其中，N、O 分别为输入层和输出层神经元数，L 为 1~10 之间的整数）、HN=2N+1（其中，N 为输入层神经元数）[19]。最终，隐含层为 12 个神经元两层（直连）的 ANN 模型效果最好。该 ANN 预测模型纳入了年龄、性别、便秘、血便、里急后重、进行性消瘦及 FOBT 阳性 7 个变量。

Yeoh 等[16]在 APCS 评分模型中纳入年龄、性别、家族史和吸烟 4 个变量；叶丁等[18]的模型纳入年龄、文化程度、职业、慢性腹泻、慢性便秘、结肠息肉史、家族史 7 个变量；Hippisley-Cox 和 Coupland[14]的 QCancer 预测模型女性版纳入年龄、胃肠道癌家族史、腹痛、食欲下降、直肠出血、体重下降和贫血 7 个变量，男性版还纳入大便习惯改变与饮酒 2 个变量；Williams 等[23]有关结直肠癌预测模型的综述显示，除里急后重外的其他 6 个变量曾被纳入国内外多个预测模型中，而里急后重可能因未包含在调查问卷中而研究较少。可见，本次所纳入变量作为结直肠癌重要危险因素具有较多的文献支持。但是，结直肠癌家族史与家族性腺瘤和息肉在病例组与非病例组间分布差异不具有统计学意义，这与国内外研究已证实其为结直肠癌高危因素的结果不一致[10,14,16]，可能是该类个体注意避免吸烟、饮酒、精神创伤等其他危险因素暴露或主动体检而发现早期病变、及时治疗，

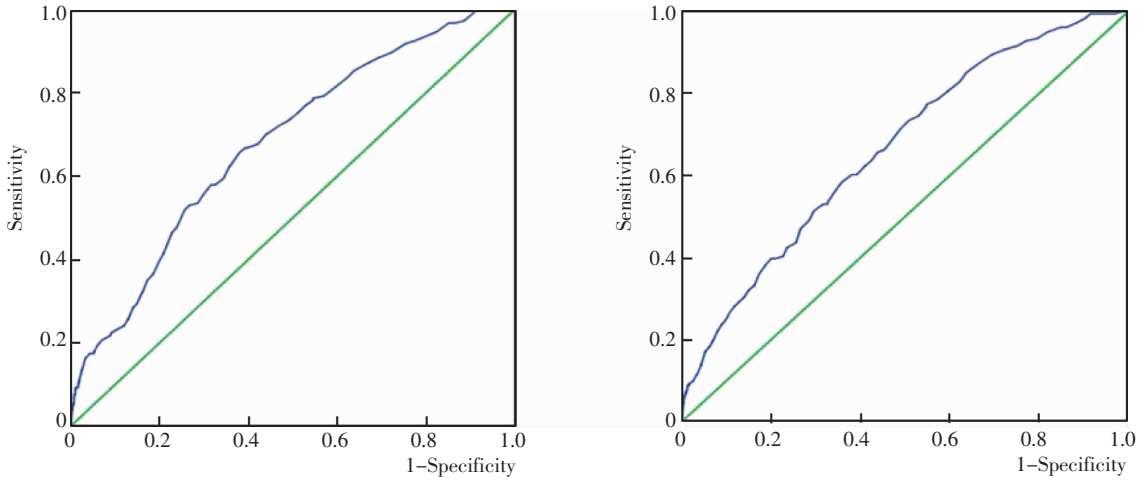


Figure 2　The receiver operating characteristic (ROC) curve in the training set(left) and validation set(right)

降低发病率。进一步算得各输入变量对结直肠癌结果的影响比重，发现血便、年龄与 FOBT 阳性对研究结果的影响最大。

本次所建模型训练、验证集的灵敏度、特异性均在 60%左右，可能是由于模型所纳入变量信息不够全面，缺少体力活动、吸烟、饮酒等行为生活方式、精神创伤、糖尿病史等结直肠癌可能危险因素。训练集、验证集的AUC 均为 0.7 左右，模型区分度较好。训练集、验证集的 NPV 接近 100%而 PPV 由于结直肠癌的人群患病率低且该研究中信息收集不够全面而仅为 4%左右。训练集、验证集的+LR、–LR 分别为 2.0 和 0.6 左右，说明该模型具有较高的诊断价值。训练、验证集的模型效果接近，说明模型的内部验证效果较好，在该人群中体现出了较好的泛化能力。

本研究的优势主要是样本量大，并且仅纳入年龄、性别、便秘、血便、里急后重、进行性消瘦和 FOBT 阳性 7 个变量，通过简单的信息收集即可对个体的发病状态进行判别，大大简化结直肠癌的初筛内容，适用于大规模的结直肠癌筛查且利于提高筛查人群的依从性。此外，仅经 ANN 判为阳性者才需要做结肠镜检查，利于降低结直肠癌筛查成本。最后，本研究具有明确的研究对象纳入标准，避免了研究对象选择的随意性，减小选择偏倚。调查员均是经过统一培训、考核的专业人员也在一定程度上减少了调查和测量偏倚，诊断标准为国际公认的 ICD-O-1，考虑到结直肠癌的潜隐期，同肿瘤登记系统进行记录联动补充漏诊病例，保证了结局信息的完整性与准确性。

本次研究的局限性主要是调查问卷未能收集到完整的结直肠癌危险因素信息，尤其是缺少行为生活方式、精神创伤等重要因素，影响模型的预测准确性，有待进一步收集数据后完善和优化模型。其次，本研究基于横断面调查，收集的信息仅能反映筛查当时个体的疾病与暴露状况，除性别、年龄外，难以确定先因后果的时相关系，而仅能用于判别个体筛查时的疾病状态而无法预测个体之后的发病风险。同时，调查对象均为幸存者而无法调查死亡的人可造成幸存者偏倚，由于回答不准确或对过去的疾病史回忆不清可造成信息偏倚。此外，本研究中病例数较少可能导致模型对病例信息的训练、学习不够充分而影响模型效果。最后，该模型还有待进一步在外部人群中进行外部验证以提高模型的外推和泛化能力。

综上所述，本次研究基于年龄、血便、FOBT 阳性等 7 个因素建立的结直肠癌 ANN 预测模型判别个体疾病状态的区分度较好，诊断价值较高，适用于上海市闵行区 50~80 岁人群的结直肠癌初筛，但由于模型中未纳入体力活动、饮食习惯、精神创伤等结直肠癌可能重要危险因素，从而影响了模型预测的准确性，在实际应用中应考虑这些因素对预测结果的影响。

参考文献：

[1] Du LB, Li HZ, Wang YQ, et al. Analysis of colorectal cancer incidence and mortality in China, 2013[J]. Chinese Journal of Oncology, 2017, 39(9): 701–706.[杜灵彬，李辉章，王悠清，等.2013 年中国结直肠癌发病与死亡分析

[J]. 中华肿瘤杂志,2017,39(9):701-706.]
[2] Navarro M,Nicolas A,Ferrandez A,et al. Colorectal cancer population screening programs worldwide in 2016:an update[J]. World J Gastroenterol,2017,23(20):3632-3642.
[3] Issa IA,Noureddine M. Colorectal cancer screening:an updated review of the available options [J]. World J Gastroenterol,2017,23(28):5086-5096.
[4] Carter JV,Roberts HL,Pan J,et al. A highly predictive model for diagnosis of colorectal neoplasms using plasma microRNA:improving specificity and sensitivity [J]. Ann Surg,2016,264(4):575-584.
[5] Yuan P,Gu J. Meta-analysis of the compliance of colorectal cancer screening in China,2006~2015[J]. China Cancer,2017,26(4):241-248.[袁平,顾晋. 2006~2015 年中国大肠癌筛查人群依从性的 Meta 分析 [J]. 中国肿瘤,2017,26(4):241-248.]
[6] Hong SY. Study on the strategy and the compliance in colonoscopy screening [D]. Shanghai:The Second Military Medical University,2012.[洪尚游. 大肠癌结肠镜筛查的策略及依从性研究[D].上海:第二军医大学,2012.]
[7] Peng L,Weigl K,Boakye D,et al. Risk scores for predicting advanced colorectal neoplasia in the average-risk population:a systematic review and meta-analysis[J]. Am J Gastroenterol,2018,113:1788-1800.
[8] Ahmed FE. Artificial neural networks for diagnosis and survival prediction in colon cancer [J]. Mol Cancer,2005,4:29.
[9] Steffen A,MacInnis RJ,Joshy G,et al. Development and validation of a risk score predicting risk of colorectal cancer[J]. Cancer Epi Bio Prev,2014,23(11):2543-2552.
[10] Sekiguchi M,Kakugawa Y,Matsumoto M,et al. A scoring model for predicting advanced colorectal neoplasia in a screened population of asymptomatic Japanese individuals [J]. J Gastroenterol,2018,53(10):1109-1119.
[11] Jeon J,Du M,Schoen RE,et al. Determining risk of colorectal cancer and starting age of screening based on lifestyle,environmental,and genetic factors [J]. Gastroenterology,2018,154(8):2152-2164.
[12] Kinar Y,Kalkstein N,Akiva P,et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts:a binational retrospective study [J]. J Am Med Inform Assoc,2016,23(5):879-890.
[13] Usher-Smith JA,Harshfield A,Saunders CL,et al. External validation of risk prediction models for incident colorectal cancer using UK Biobank[J]. Br J Cancer,2018,118(5):750-759.
[14] Hippisley-Cox J,Coupland C. Identifying patients with suspected colorectal cancer in primary care:derivation and validation of an algorithm[J]. Br J Gen Prac,2012,62(594):29-37.
[15] Collins GS,Altman DG. Identifying patients with undetected colorectal cancer:an independent validation of QCancer (Colorectal)[J]. Br J Cancer,2012,107(2):260-265.
[16] Yeoh KG,Ho KY,Chiu HM,et al. The Asia-Pacific colorectal screening score:a validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic Asian subjects[J]. Gut,2011,60(9):1236-1241.
[17] Yu XZ. Application of CLINPROT technology in early diagnosing colorectal cancer and liver metastases[D]. Shanghai:Fudan University,2010.[于新哲. 应用 CLINPROT 技术建立早期诊断结直肠癌及肝转移预测模型[D]. 上海:复旦大学,2010.]
[18] Ye D. Biomarker screening and risk assessment model of colorectal cancer in Han Chinese [D]. Hangzhou:Zhejiang University,2018.[叶丁.中国汉族人群结直肠癌分子标志物筛选及风险评估模型研究[D]. 杭州:浙江大学,2018.]
[19] Gao RX,Zhang SY,Liu B. Variable selection approach based on neural network analysis [J]. Journal of Systems Engineering,1998,13(2):34-39. [高仁祥,张世英,刘豹. 基于神经网络的变量选择方法 [J]. 系统工程学报,1998,13(2):34-39.]
[20] Chen K,Chen JF,Guan XM,et al. Establishment of high risk scoring model for early colorectal cancer screening[J]. China Medicine and Pharmacy,2016,6(8):27-30. [陈锴,陈锦锋,关小明,等.结直肠癌早期筛查高危评分模型的建立[J]. 中国医药科学,2016,6(8):27-30.]
[21] Guo YR,Li YQ,Wang GS,et al. Application of artificial neural network to predict individual risk of type 2 diabetes mellitus[J]. Journal of Zhengzhou University(Medical Sciences),2014,49(2):180-183. [郭亦瑞,李玉倩,王高帅,等. 人工神经网络模型在 2 型糖尿病患病风险预测中的应用[J]. 郑州大学学报(医学版),2014,49(2):180-183.]
[22] Guo QC. Research on application of artificial neural network[M]. First edition. Changchun:Jilin University Press,2016.43.[郭庆春.人工神经网络应用研究[M].第一版.长春:吉林大学出版社,2016.43.]
[23] Williams TG,Cubiella J,Griffin SJ,et al. Risk prediction models for colorectal cancer in people with symptoms:a systematic review[J]. BMC Gastroenterol,2016,16(1):63.