·临床与预防·

糖尿病并发视网膜病变的关键影响因素分析和风险预测研究

武巾媛，安洪庆
潍坊医学院，山东 潍坊 261053

摘要：目的　探究糖尿病并发视网膜病变（diabetic retinopathy，DR）的关键影响因素，进行发病现状分析，并构建发病风险预测模型。**方法**　基于“国家人口与健康科学数据共享平台”公布的“糖尿病并发症预警数据集”，利用单因素和多因素 logistic 回归分析得到 DR 发病关键影响因素；运用熵权法、优劣解距离法（technique for order preferenceby similarity to ideal solution，TOPSIS）、联合秩和比方法（rank-sum ratio，RSR）进行发病风险分层；分别构建 logistic 回归、随机森林、支持向量机模型，使用投票、平均、加权平均三种方法进行模型融合，对模型预测效果进行评估，取得最佳预测模型。**结果**　最终提取年龄、高脂血等 14 个指标作为关键影响因素；分层结果显示未患 DR 的糖尿病患者中存在 50 人患病风险较高，约为 82.99%，需要重点关注；投票器融合模型预测效果最佳（Acc：80.18%，F1：0.786 8）。**结论**　分析得到 DR 关键影响因素，提供了治疗与预防方向；进行发病风险现状分析，划分 DR 低中高风险人群，进行风险预警；通过模型间效果对比，构建 DR 发病风险预测模型，为其临床预警提供了数据分析思路与方法。

关键词：糖尿病视网膜病变；Logistic；TOPSIS；RSR；预测模型
中图分类号：TP181；R587.2；R774.1　　**文献标志码：**A　　**文章编号：**1003-8507（2024）03-557-07
DOI：10.20043/j.cnki.MPM.202309033

Analysis of key influencing factors and risk prediction of diabetic retinopathy

WU Jin-yuan, AN Hong-qing
Weifang Medical College, Weifang, Shandong 261053, China

Abstract: Objective To explore the key influencing factors of diabetic retinopathy (DR), analyze the current situation of DR, and construct a risk prediction model. **Methods** Based on the diabetic complication early warning data set published by the national population and health science data sharing platform, the key influencing factors of DR were obtained by univariate and multivariate Logistic regression analyses. The entropy weighting method, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and the rank-sum ratio (RSR) were used to quantify the risk of DR development in patients and stratified into three levels: high, medium, and low. Logistic regression, random forest, and support vector machine models were constructed, respectively, and model fusion was performed using voting, averaging, and weighted averaging to evaluate the model predictive effect and obtain the best predictive model. **Results** Finally, 14 indexes including age and hyperlipidemia were extracted as key influencing factors. The stratification results showed that there were 50 diabetic patients without DR in this data set with a risk of about 82.99%, which was a high-risk group for DR and needed more attention. The best prediction effect was obtained from the voting machine fusion model (Acc: 80.18%, F1: 0.7868). **Conclusion** The key influencing factors of DR are analyzed, providing the direction of treatment and prevention. The low, medium, and high-risk groups of DR are classified for risk early warning. By comparing the effect between the models, the prediction model of morbidity risk of DR is constructed, providing insights for clinical early warning and data analysis.

Keywords: Diabetic retinopathy; Logistic; TOPSIS; RSR; Predictive model

糖尿病发病率日益增高，统计显示 2022 年全球已拥有 5.5 亿糖尿病患者，其中中国糖尿病患者人数超 1.4 亿，位居世界首位。糖尿病并发视网膜病变（diabetic retinopathy，DR）是糖尿病的严重并发症之一，DR 大多数早期症状不明显，目前临床诊断手段主要以荧光素眼底血管造影、超声检查或视网膜电图等眼底图像检测等眼底检查为主，所需费用较高、流程较为复杂[1]。相关研究多运用人工智能方法进行眼

基金项目：中国学位与研究生教育学会课题(2020MSA105)；中国高等教育学会 2023 年度课题(23PG0411)；山东省高等医学教育研究中心规划课题(YJKT202126)
作者简介：武巾媛(2003—)，女，本科在读，研究方向：大数据管理与应用
通信作者：安洪庆，E-mail: hongqingan01@126.com

底图像识别[2],或仅对关键影响因素进行分析[3],缺乏基于数据的监测预测手段。本研究尝试建立基于病例健康数据的预测模型。相较于图像监测手段,数据监测具有快捷易操作、费用少、可批量操作等优点,有利于疾病病程监测,达到早发现早治疗的目标。

1　资料与方法

1.1 *数据来源*　本研究数据来源于公开数据库中国人民解放军总医院国家人口健康科学数据中心数据仓储公开数据集“糖尿病并发症预警数据集”(2013 年 1 月—2017 年 12 月)[4],已通过潍坊医学院医学伦理委员的审查(编号 2022YX057)。

选择符合以下条件的 2 型糖尿病(type 2 diabetic mellitus,T2DM)患者个体作为研究样本:

(1)诊断条件:研究样本为已被临床医生明确诊断为 T2DM 的患者。确认病例时要求患者至少具备以下两个条件之一:①空腹血糖水平≥7.0 mmol/L;②随机血糖水平≥11.1 mmol/L,并伴有典型的糖尿病症状,如多饮、多尿、疲劳等。

(2)具备与视网膜病变相关的生化指标检查数据,如血压、体重指数(BMI)等。

(3)具备与视网膜病变相关的疾病史数据,包括高血压、高脂血症等。

1.2 *研究方法*　使用单因素 logistic 回归初步筛选因变量结局的影响因素,投入多因素 logistic 回归中矫正混杂因素的影响,筛选出 DR 关键影响因素;使用熵权 TOPSIS 联合 RSR 方法对发病现状进行分析,求得各病例的 DR 综合发病风险并进行高中低三分层;使用 R4.2.2 进行预测模型的构建,在 1 856 例研究对象中,按照 7:3 的比例随机抽取 1 301 例作为训练集剩余 555 例为测试集,分别建立 logistic 回归模型、随机森林模型、支持向量机模型,并进一步进行模型融合,对比各模型预测效果得到最优模型。

1.3 *数据处理*　剔除缺失率达到 30%以上的个变量,删除含有缺失值的病例,剩余有效样本共 1 856 例,其中共计 993 例 DR 患者,占比约为 53.5%,总样本基本情况见表 1。

表 1　变量基本情况表

Table 1　Basic information of variables

指标	特征	占比(%)	指标	特征	占比(%)	指标	Min	Max	$(\bar{x} \pm s)$
患病种类	0= 糖尿病	46.50	心肌梗死	0= 未患病	93.64	血肌酐(μmol/L)	29.6	1 300.8	110.88 ± 122.14
	1= 糖尿病并发视网膜病变	53.50		1= 患病	6.36	年龄(岁)	19	93	58.04 ± 11.02
性别	0= 男	36.21	心功能不全及心力衰竭	0= 未患病	91.81	身高(cm)	112	193	166.80 ± 7.85
	1= 女	63.79		1= 患病	8.19	体重(kg)	34	156	73.39 ± 12.74
民族	0= 汉族	94.72	心律失常	0= 未患病	93.16	收缩压(mm Hg)	91	250	139.18 ± 21.13
	1= 少数民族	5.28		1= 患病	6.84	舒张压(mm Hg)	45	145	80.82 ± 11.89
婚姻状态	0= 其他	1.56	呼吸系统疾病	0= 未患病	81.63	BMI(kg/m^2)	13.35	55.58	26.30 ± 3.70
	1= 已婚	98.44		1= 患病	18.37	空腹血糖(mmol/L)	2.13	42.19	8.37 ± 3.82
高血压	0= 未患病	27.42	下肢动脉病变	0= 未患病	79.42	糖化血红蛋白(mmol/L)	3.9	15.3	7.83 ± 1.80
	1= 患病	72.58		1= 患病	20.58	甘油三酯(mmol/L)	0.27	15.11	2.05 ± 1.64
高脂血	0= 未患病	75.97	血液病	0= 未患病	82.06	总胆固醇(mmol/L)	1.21	20.85	4.65 ± 1.42
	1= 患病	24.03		1= 患病	17.94	高密度脂蛋白胆固醇(mmol/L)	0.07	3.09	1.06 ± 0.32
动脉粥样硬化	0= 未患病	43.53	风湿免疫疾病	0= 未患病	96.12	低密度脂蛋白胆固醇(mmol/L)	0.16	17.31	2.87 ± 1.16
	1= 患病	56.47		1= 患病	3.88	纤维蛋白原(g/L)	1.08	883.8	9.38 ± 44.99
脑卒中	0= 未患病	90.30	妊娠哺乳期	0= 未患病	99.78	血尿素(mmol/L)	1.46	50.94	7.31 ± 5.16
	1= 患病	9.70		1= 患病	0.22	血清尿酸(mol/L)	11.5	1 153.4	332.05 ± 101.66
颈动脉狭窄	0= 未患病	95.15	其他内分泌疾病	0= 未患病	60.08	血红蛋白(g/L)	42	244	131.57 ± 8.89
	1= 患病	4.85		1= 患病	39.92	红细胞压积(红细胞比积测定,%)	0.14	0.74	0.39 ± 0.06
脂肪肝	0= 未患病	62.93	内分泌腺瘤	0= 未患病	95.20	血小板(10^9/L)	2	574	215.59 ± 68.69
	1= 患病	37.07		1= 患病	4.80	总胆红素(mmol/L)	1.2	278.5	10.95 ± 10.91
肝硬化	0= 未患病	97.84	多囊卵巢综合征	0= 未患病	99.89	直接胆红素(μmol/L)	0.1	226.4	3.50 ± 8.01
	1= 患病	2.16		1= 患病	0.11	总蛋白(g/L)	32.1	94.3	65.32 ± 7.57
其他慢性肝病	0= 未患病	82.70	消化系肿瘤	0= 未患病	93.10	血清白蛋白(g/L)	11.7	55.6	39.23 ± 5.96
	1= 患病	17.30		1= 患病	6.90	乳酸脱氢酶(U/L)	76.3	741.5	172.84 ± 58.76
胰腺外分泌疾病	0= 未患病	97.90	泌尿系肿瘤	0= 未患病	98.81	谷丙转氨酶(U/L)	2.7	1 076.6	24.69 ± 31.99
	1= 患病	2.10		1= 患病	1.19	谷草转氨酶(U/L)	3.6	756.8	20.66 ± 24.86
胆道疾病	0= 未患病	82.81	妇科肿瘤	0= 未患病	96.71	谷氨酰胺转移酶(U/L)	4.3	1 404.4	42.73 ± 70.03
	1= 患病	17.19		1= 患病	3.29	碱性磷酸酶(U/L)	9.5	911	75.63 ± 41.51
肾病	0= 未患病	49.19	乳腺肿瘤	0= 未患病	99.52	凝血酶原时间(s)	9.5	35.3	13.09 ± 1.22
	1= 患病	50.81		1= 患病	0.48	凝血酶原活动度(%)	0.63	164	98.97 ± 20.09
肾衰	0= 未患病	92.08	肺部肿瘤	0= 未患病	97.84	部分活化凝血酶原时间(s)	20.3	180	36.55 ± 6.34
	1= 患病	7.92		1= 患病	2.16	肿瘤标志物 CA199(U/ml)	0.6	703 2	29.67 ± 227.18
神经系统疾病	0= 未患病	93.37	颅内肿瘤	0= 未患病	99.35	间接胆红素(μmol/L)	0.4	83.7	7.44 ± 4.71
	1= 患病	6.63		1= 患病	0.65	球蛋白(g/L)	14.2	61.4	26.11 ± 4.90
冠心病	0= 未患病	66.86	其他肿瘤	0= 未患病	89.06	—	—	—	—
	1= 患病	33.14		1= 患病	10.94	—	—	—	—

2　结　果

2.1　单因素 logistic 回归初步提取影响因素　使用 R4.2.2 对进行单因素 logistic 回归分析，检验水准 α=0.1，初步提取 DR 相关指标共 49 个，见表 2。

表 2　单因素 logistic 回归结果

Table 2　Single-factor logistic regression results

序号	指标	*OR* 值(95%*CI*)	$s_{\bar{x}}$	*P* 值
1	血肌酐(μmol/L)	1.004(1.003 3 ~ 1.005 8)	0.000 6	<0.001
2	年龄(岁)	0.977(0.969 2 ~ 0.985 8)	0.004 3	<0.001
3	性别	0.977(0.808 2 ~ 1.181 4)	0.096 9	0.811
4	民族	1.461(0.961 0 ~ 2.221 6)	0.213 8	0.076
5	婚姻状态	0.809(0.384 5 ~ 1.704 7)	0.379 9	0.578
6	身高(cm)	0.999(0.988 3 ~ 1.011 5)	0.005 9	0.974
7	体重(kg)	1.006(0.999 7 ~ 1.014 2)	0.003 7	0.061
8	收缩压(mm Hg)	1.015(1.010 5 ~ 1.019 8)	0.002 3	<0.001
9	舒张压(mm Hg)	1.022(1.014 5 ~ 1.030 8)	0.004 1	<0.001
10	BMI(kg/m^2)	1.032(1.006 7 ~ 1.058 3)	0.012 7	0.012
11	高血压	1.749(1.424 1 ~ 2.149 6)	0.105 0	<0.001
12	高脂血	0.484(0.390 3 ~ 0.602 4)	0.110 7	<0.001
13	动脉粥样硬化	1.011(0.841 5 ~ 1.215 8)	0.093 9	0.903
14	脑卒中	1.944(1.401 9 ~ 2.695 9)	0.166 8	0.000
15	颈动脉狭窄	1.61(1.035 8 ~ 2.505 3)	0.225 3	0.034
16	脂肪肝	1.214(1.005 2 ~ 1.468 3)	0.096 7	0.044
17	肝硬化	0.572(0.302 1 ~ 1.084 9)	0.326 1	0.087
18	其他慢性肝病	0.903(0.709 9 ~ 1.149 0)	0.122 9	0.406
19	胰腺外分泌疾病	0.913(0.484 0 ~ 1.722 4)	0.323 8	0.778
20	胆道疾病	1.189(0.932 4 ~ 1.516 7)	0.124 1	0.162
21	肾病	4.857(3.988 9 ~ 5.914 3)	0.100 5	<0.001
22	肾衰	7.009(4.241 9 ~ 11.581 3)	0.256 2	<0.001
23	神经系统疾病	0.735(0.509 6 ~ 1.061 2)	0.187 1	0.1
24	冠心病	0.528(0.434 4 ~ 0.642 4)	0.099 8	<0.001
25	心肌梗死	0.643(0.441 5 ~ 0.937 0)	0.192 0	0.021
26	心功能不全及心力衰竭	0.962(0.690 5 ~ 1.342 1)	0.169 6	0.822
27	心律失常	0.845(0.589 8 ~ 1.212 8)	0.183 9	0.362
28	呼吸系统疾病	0.979(0.773 9 ~ 1.239 4)	0.120 1	0.862
29	下肢动脉病变	3.457(2.676 7 ~ 4.464 9)	0.130 5	<0.001
30	血液病	3.545(2.696 3 ~ 4.662 7)	0.139 7	<0.001
31	风湿免疫疾病	0.609(0.377 7 ~ 0.981 8)	0.243 7	0.041
32	妊娠哺乳期	0.868(0.122 1 ~ 6.181 2)	1.001 1	0.888
33	其他内分泌疾病	1.855(1.535 0 ~ 2.243 1)	0.096 8	<0.001
34	内分泌腺瘤	0.664(0.432 8 ~ 1.020 9)	0.218 9	0.062
35	多囊卵巢综合征	0.869(0.054 3 ~ 13.914 0)	1.415 0	0.92
36	消化系肿瘤	0.254(0.168 0 ~ 0.385 5)	0.211 9	<0.001
37	泌尿系肿瘤	0.598(0.254 4 ~ 1.405 9)	0.436 1	0.238
38	妇科肿瘤	0.444(0.260 0 ~ 0.760 0)	0.273 6	0.003
39	乳腺肿瘤	0.432(0.108 0 ~ 1.734 9)	0.708 3	0.237
40	肺部肿瘤	0.178(0.078 6 ~ 0.405 7)	0.418 7	<0.001
41	颅内肿瘤	0.287(0.077 6 ~ 1.065 5)	0.668 3	0.062
42	其他肿瘤	0.386(0.284 2 ~ 0.526 4)	0.157 3	<0.001
43	空腹血糖(mmol/L)	1.050(1.024 1 ~ 1.076 7)	0.012 8	0.0001
44	糖化血红蛋白(mmol/L)	1.286(1.216 5 ~ 1.359 5)	0.028 4	<0.001
45	甘油三酯(mmol/L)	1.026(0.970 0 ~ 1.085 4)	0.028 7	0.369
46	总胆固醇(mmol/L)	1.152(1.075 5 ~ 1.233 9)	0.035 1	0.0001
47	高密度脂蛋白胆固醇(mmol/L)	1.305(0.980 5 ~ 1.738 2)	0.146 1	0.068
48	低密度脂蛋白胆固醇(mmol/L)	1.213(1.113 1 ~ 1.322 5)	0.044 0	<0.001
49	纤维蛋白原(g/L)	1.001(0.999 1 ~ 1.003 5)	0.001 1	0.242
50	血尿素(mmol/L)	1.133(1.102 2 ~ 1.165 2)	0.014 2	<0.001
51	血清尿酸(μmol/L)	1.002(1.001 2 ~ 1.003 1)	0.000 5	<0.001
52	血红蛋白(g/L)	0.981(0.976 8 ~ 0.985 2)	0.002 2	<0.001
53	红细胞压积(红细胞比积测定,%)	0.000 4(0.000 1 ~ 0.001 9)	0.800 9	<0.001
54	血小板(10^9/L)	1.000(0.999 2 ~ 1.001 8)	0.000 7	0.467
55	总胆红素(mmol/L)	0.95(0.934 8 ~ 0.966 7)	0.008 5	<0.001

（续表）

序号	指标	*OR* 值（95%*CI*）	$s_{\bar{x}}$	*P* 值
56	直接胆红素（μmol/L）	0.804(0.761 8 ~ 0.848 6)	0.027 5	<0.001
57	总蛋白（g/L）	0.933(0.920 9 ~ 0.946 6)	0.007 0	<0.001
58	血清白蛋白（g/L）	0.912(0.896 6 ~ 0.929 2)	0.009 1	<0.001
59	乳酸脱氢酶（U/L）	1.003(1.001 3 ~ 1.004 7)	0.000 9	0.0005
60	谷丙转氨酶（U/L）	0.982(0.976 6 ~ 0.987 5)	0.002 8	<0.001
61	谷草转氨酶（U/L）	0.977(0.969 5 ~ 0.985 7)	0.004 2	<0.001
62	谷氨酰胺转移酶（U/L）	0.995(0.994 0 ~ 0.997 9)	0.001 0	<0.001
63	碱性磷酸酶（U/L）	0.996(0.994 4 ~ 0.999 4)	0.001 3	0.016
64	凝血酶原时间（s）	0.832(0.760 2 ~ 0.912 0)	0.046 5	0.0001
65	凝血酶原活动度（%）	1.002(0.997 9 ~ 1.007 0)	0.002 3	0.285
66	部分活化凝血酶原时间（s）	0.986(0.970 9 ~ 1.001 8)	0.008 0	0.082
67	肿瘤标志物 CA199（U/ml）	0.999(0.998 3 ~ 1.000 3)	0.000 5	0.188
68	间接胆红素（mol/L）	0.952(0.931 1 ~ 0.973 3)	0.011 3	<0.001
69	球蛋白（g/L）	0.972(0.954 7 ~ 0.991 2)	0.009 6	0.004

2.2 多因素 logistic 回归提取关键影响因素 将 DR 相关指标投入多因素 logistic 回归中，其中 21 个定类变量以哑变量形式代入模型进行回归分析，以第一个值为参考，检验水准 α=0.05，得到 DR 关键影响因素指标共 14 个，见表 3。

表 3 多因素 logistic 逐步回归结果

Table 3 Multi-factor logistic stepwise regression results

指标	Estimate	*z* value	*OR* 值（95%*CI*）	Std.Error	*P* 值
(Intercept)	2.880 0	1.955 0	17.811(0.025 6-5.799 5)	1.473 0	0.050
血肌酐（μmol/L）	-0.001 8	-1.701 0	0.998(-0.003 8 ~ 0.000 3)	0.001 0	0.088
年龄（岁）	-0.021 6	-3.131 0	0.978(-0.035 2 ~ -0.008 1)	0.006 9	0.001
民族	0.298 0	1.115 0	1.347(-0.221 9 ~ 0.828 0)	0.267 4	0.265
体重（kg）	-0.003 9	-0.385 0	0.996(-0.023 8 ~ 0.015 7)	0.010 1	0.700
收缩压（mm Hg）	0.001 8	0.459 0	1.001(-0.005 7 ~ 0.009 3)	0.003 8	0.646
舒张压（mm Hg）	0.011 8	1.775 0	1.011(-0.001 2 ~ 0.024 8)	0.006 6	0.075
BMI（kg/m^2）	0.034 0	1.064 0	1.034(-0.027 8 ~ 0.097 4)	0.032 0	0.287
高血压	0.206 2	1.367 0	1.229(-0.089 2 ~ 0.502 2)	0.150 8	0.171
高脂血	-0.574 3	-3.898 0	0.563(-0.864 5 ~ -0.286 5)	0.147 3	<0.001
脑卒中	0.373 1	1.702 0	1.452(-0.052 6 ~ 0.807 9)	0.219 3	0.088
颈动脉狭窄	0.175 8	0.620 0	1.192(-0.374 9 ~ 0.739 1)	0.283 6	0.535
脂肪肝	0.275 4	1.958 0	1.317(0 ~ 0.551 7)	0.140 7	0.050
肝硬化	0.193 3	0.418 0	1.213(-0.718 7 ~ 1.103 8)	0.462 8	0.676
肾病	1.025 0	7.255 0	2.787(0.749 2 ~ 1.303 3)	0.141 3	<0.001
肾衰	0.916 6	2.937 0	2.5(0.323 7 ~ 1.552 2)	0.312 1	0.003
冠心病	-0.384 8	-2.580 0	0.68(-0.677 6 ~ -0.092 7)	0.149 1	0.009
心肌梗死	0.005 3	0.020 0	1.005(-0.512 9 ~ 0.517 8)	0.262 6	0.984
下肢动脉病变	0.948 3	6.031 0	2.581(0.643 2 ~ 1.260 1)	0.157 2	<0.001
血液病	0.209 1	1.009 0	1.232(-0.196 8 ~ 0.615 9)	0.207 1	0.312
风湿免疫疾病	-0.314 5	-1.011 0	0.73(-0.932 2 ~ 0.290 5)	0.311 2	0.312
其他内分泌疾病	0.442 9	3.441 0	1.557(0.191 1 ~ 0.695 9)	0.128 7	<0.001
内分泌腺瘤	-0.825 7	-3.034 0	0.437(-1.363 4 ~ -0.294 7)	0.272 2	0.002
消化系肿瘤	-0.321 5	-1.245 0	0.725(-0.838 4 ~ 0.176 6)	0.258 3	0.213
妇科肿瘤	-0.843 0	-2.404 0	0.43(-1.548 2 ~ -0.169 4)	0.350 7	0.016
肺部肿瘤	-0.982 1	-1.819 0	0.374(-2.121 6 ~ 0.013 4)	0.539 9	0.068
颅内肿瘤	-0.447 3	-0.530 0	0.639(-2.256 9 ~ 1.125 9)	0.844 1	0.596
其他肿瘤	-0.510 1	-2.429 0	0.6(-0.924 6 ~ -0.100 4)	0.210 0	0.015
空腹血糖（mmol/L）	-0.024 5	-1.317 0	0.975(-0.060 8 ~ 0.012 3)	0.018 6	0.188
糖化血红蛋白（mmol/L）	0.418 8	9.414 0	1.52(0.332 7 ~ 0.507 2)	0.044 5	<0.001
总胆固醇（mmol/L）	-0.124 5	-1.283 0	0.882(-0.315 4 ~ 0.067 0)	0.097 0	0.199
高密度脂蛋白胆固醇（mmol/L）	0.217 6	1.029 0	1.243(-0.194 6 ~ 0.635 5)	0.211 5	0.303
低密度脂蛋白胆固醇（mmol/L）	0.005 9	0.052 0	1.005(-0.218 2 ~ 0.229 7)	0.113 7	0.958
血尿素（mmol/L）	0.070 8	2.568 0	1.073(0.018 1 ~ 0.126 1)	0.027 6	0.010
血清尿酸（μmol/L）	0.000 0	-0.058 0	1(-0.001 4 ~ 0.001 4)	0.000 7	0.953
血红蛋白（g/L）	0.021 5	1.493 0	1.021(-0.006 7 ~ 0.049 8)	0.014 4	0.135

（续表）

指标	Estimate	*z* value	*OR* 值（95%*CI*）	Std.Error	*P* 值
红细胞压积（红细胞比积测定）（%）	−13.330 0	−2.600 0	0(−23.430 3 ~ −3.323 6)	5.127 0	0.009
总胆红素（mmol/L）	0.349 3	0.790 0	1.418(−0.327 5 ~ 1.801 2)	0.441 9	0.429
直接胆红素（μmol/L）	−0.392 2	−0.884 0	0.675(−1.845 1 ~ 0.289 1)	0.443 6	0.376
总蛋白（g/L）	−0.012 9	−0.098 0	0.987(−0.320 5 ~ 0.212 7)	0.131 4	0.921
血清白蛋白（g/L）	−0.020 4	−0.154 0	0.979(−0.247 4 ~ 0.288 2)	0.132 0	0.877
乳酸脱氢酶（U/L）	0.001 0	0.807 0	1.001(0.001 5 ~ 0.003 6)	0.001 3	0.419
谷丙转氨酶（U/L）	−0.004 7	−0.864 0	0.995(−0.015 2 ~ 0.006 1)	0.005 4	0.387
谷草转氨酶（U/L）	−0.010 2	−1.239 0	0.989(−0.027 3 ~ 0.004 9)	0.008 3	0.215
谷氨酰胺转移酶（U/L）	0.000 1	0.065 0	1(−0.002 9 ~ 0.002 9)	0.001 5	0.948
碱性磷酸酶（U/L）	−0.002 0	−0.814 0	0.998(−0.006 8 ~ 0.002 5)	0.002 4	0.415
凝血酶原时间（s）	−0.183 9	−2.779 0	0.832(−0.318 2 ~ −0.059 2)	0.066 2	0.005
部分活化凝血酶原时间（s）	−0.009 5	−0.798 0	0.99(−0.033 8 ~ 0.012 2)	0.012 0	0.424
间接胆红素（μmol/L）	−0.326 2	−0.738 0	0.721(−1.778 8 ~ 0.350 8)	0.442 1	0.460
球蛋白（g/L）	−0.006 2	−0.047 0	0.993(−0.233 3 ~ 0.302 4)	0.132 1	0.962

2.3　指标方向判断　规定数值越大 DR 发病风险越高的指标为正向指标（危险因素），反之为负向指标（保护因素）。查阅相关文献并根据多因素 logistic 回归结果中 estimate 的值判断指标方向，便于归一化处理。确定年龄[5-7]、凝血酶原时间、冠心病[8]、高脂血[9]、红细胞压积、内分泌肿瘤、妇科肿瘤、其他肿瘤[10]8 个指标为 DR 发病的负向指标，HbA1c[11-13]、其他内分泌疾病、肾病[14]、肾衰、下肢动脉病变、血尿素[15]6 个指标为其正向指标。

2.4　发病现状分析　熵权法指标权重分析。记原始矩阵为 X_{ij}，根据极差法归一化公式得到归一化矩阵 Z_{ij}，正向指标 $Z_{ij}=\frac{X_{ij}-\min(X_j)}{\max(X_j)-\min(X_j)}$，负向指标 $Z_{ij}=\frac{\max(X_j)-X_{ij}}{\max(X_j)-\min(X_j)}$。对 Z_{ij} 进行运算得到其概率矩阵 $p_{ij}=\frac{Z_{ij}}{\sum_{i=1}^{n}Z_{ij}}$，对于每个指标，计算其信息熵 $e_j=\frac{1}{\ln n}\sum_{i=1}^{n}p_{ij}\ln(p_{ij})(j=1,2,\cdots,m)$与信息效用值 $d_j=1-e_j$，将其信息效用值归一化得到每个指标对 DR 发病风险的贡献权重 $W_j=\frac{d_j}{\sum_{j=1}^{m}d_j}(j=1,2,\cdots,m)$，指标权重计算结果见表 4。

TOPSIS 联合 RSR 方法进行发病风险分层。利用如下公式求得各病例到正理想点 1 的距离 $d_i^{+}=\sqrt{\sum_{j=1}^{m}\omega_j^2(1-n_{ij})^2}$、到负理想点 0 的距离 $d_i^{-}=\sqrt{\sum_{j=1}^{m}\omega_j^2(n_{ij}-0)^2}$。求得与正理想解的贴进度 $C_i=\frac{d^-}{d^++d^-}$。用各 C_i 值替代 RSR 值，按照 *Ci* 值升序排列，计算累计秩次、平均秩次以及百分比数 P（P=R/n，最后一项用 1/4n 矫正），并计算对应概率单位 Probit 值。根据《常用分档数及对应概率单位表》，依据概率单位的统计学中的正态分布原则，按照 Probit 值将结果分为患病风险高（Probit≥6）、中（4≤Probit<6）、低（Probit<4）三个档次，分档情况见表 5。

表 4　指标权重

Table 4　Indicator weights

指标	方向	权重	权重排序
肾衰	正	0.361 0	1
下肢动脉病变	正	0.225 1	2
其他内分泌疾病	正	0.130 7	3
肾病	正	0.096 4	4
冠心病	负	0.057 3	5
高脂血	负	0.039 1	6
血尿素（mmol/L）	正	0.037 5	7
其他肿瘤	负	0.016 5	8
糖化血红蛋白（mmol/L）	正	0.014 4	9
年龄（岁）	负	0.007 5	10
内分泌腺瘤	负	0.007 0	11
妇科肿瘤	负	0.004 8	12
红细胞压积（红细胞比积测定）（%）	负	0.002 3	13
凝血酶原时间（s）	负	0.000 3	14
权重总和		1	

表 5　分档结果

Table 5　Grading results

分档	总人数	患病人数	患病率（%）	平均 *Ci*
高	294	244	82.99	0.538 9
中	1 267	706	55.72	0.246 7
低	295	43	14.58	0.081 2

使用 R 4.2.2，将 *Ci* 值作为因变量、Probit 值作为自变量拟合线性回归方程 *Ci*=0.149 2，Probit−0.478 8，方差分析 F=14 958(on 1 and 1 854 DF)，R^2=0.889 7，拟合效果良好，P<0.001，认为三档间的 *Ci* 值差异具有统计学意义[16]。低档 *Ci* 均值为 0.081 2、中档为 0.246 7、高档为 0.538 9，差异较大，说明高中低三档发病风险存在较大差异。根据分档结果，未患 DR 的 863 名 T2DM 中，有 252 人 DR 发病率约为 14.58%，为低风

险人群;561 人患病风险约为 55.72%，为中风险人群;50 人患病风险约为 82.99%,为高风险人群,需要重点关注。

2.5　单个预测模型　对于 logistic 模型,采用默认的参数设置,包括 L2 正则化(默认设置)和逻辑回归损失函数;对于随机森林模型,由图 1 可知,ntree 数量为 300 时,模型错误率稳定在 27%左右,因此设置决策树的数量 ntree=300 以平衡模型的复杂性和性能;对于 SVM 模型,采用默认的参数设置,包括核函数类型（kernel)、正则化参数 C 以及核函数的其他参数(例如,多项式核的次数)。

使用准确率(Acc)、精确率(P)、召回率(R)、F1 分数、假正率(FPR)、受试者工作曲线下面积(AUC)对模型预测效果进行评估,结果见表 6。支持向量机模型的预测准确率要优于另外两个模型,同时其具有更高的召回率,在临床应用中误把 DR 患者诊断为正常的概率更小。使用 trControl 函数进行十折交叉验证,准确率平均值分别为 0.759 2、0.752 7、0.762 3,支持向量机模型预测效果更为稳定。

2.6　模型融合　投票，即单模型预测结果中占多数的为最终预测结果;平均,即将单模型预测结果为 1 的概率值取平均,大于 0.5 则分类为 1。加权平均,即单模型预测准确率降序排序，依次为支持向量机、随机森林、logistic,降序编秩分别为 3、2、1,秩和为 6,分别赋权为 1/2、1/3、6/1，对预测概率进行加权平均得到最终概率。三模型预测效果对比见表 7,可知三个融合模型各方面预测效果相较于单模型均有所提高,投票器模型的准确与 F1 值最高，综合预测效果最好。

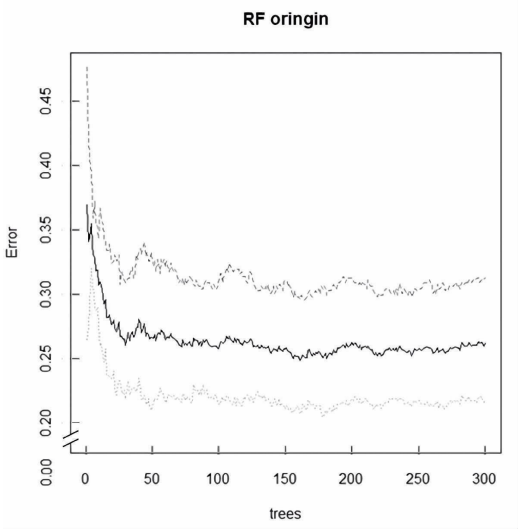

图 1　随机森林训练过程图

Figure 1　Random forest training process diagram

表 6　单模型预测效果对比

Table 6　Comparison of single-model prediction results

模型	Acc(%)	P(%)	R(%)	FPR(%)	F1	AUC	TenFold-ACC
Logistic	78.56	80.34	71.31	15.15	0.755 6	0.831 7	0.759 2
随机森林	78.37	80.53	70.54	14.81	0.814 7	0.752 1	0.752 7
支持向量机	79.46	79.46	71.71	13.80	0.814 8	0.764 5	0.762 3

表 7　融合模型预测效果对比

Table 7　Comparison of prediction effects of combining models

模型	Acc(%)	P(%)	R(%)	FPR(%)	F1
投票	80.18	78.68	78.68	18.52	0.786 8
平均	79.82	81.47	73.26	14.48	0.771 4
加权平均	79.46	81.58	71.09	14.14	0.765 4

3　讨　论

DR 的发生受多种因素共同作用,本研究使用单因素和多因素 logistic 回归确定 14 个指标为 DR 关键影响因素:年龄、高脂血、肾病、肾衰、冠心病、下肢动脉病变、其他内分泌疾病、其他肿瘤、糖化血红蛋白、血尿素、红细胞压积、凝血酶原时间、内分泌腺瘤、妇科肿瘤;运用熵权 TOPSIS 联合 RSR 方法进行现状分析。将 1 856 条有效数据按照 DR 患病风险划分为三层，确定未患 DR 的 T2DM 患者中存在 50 人 DR 患病风险较高,约为 82.99%,需要重点关注;进行 logistic、随机森林、支持向量机预测模型对比融合,三者的投票器模型预测效果最佳,准确率达到 80.18%,同时其他指标表现均较优。

本研究存在一定局限性。一是数据集中未收录患者糖尿病病程、遗传、环境等方面指标,所研究关键影响因素仅限于生化与其他病症指标。二是本研究所提取有效指标均为时点指标,不能反应指标波动情况对 DR 的影响,仍需进一步探索。本研究预测结果存在不确定性:(1)样本不确定性:样本数据的选择可能会对模型的泛化性能产生一定程度的不确定性,进一步研究可以包括更多地区、不同来源的样本,以增加模型的稳健性和泛化能力。(2) 特征不确定性:模型的性能和效果通常受到所选特征的质量和相关性

的影响。虽然我们已经通过单因素和多因素回归分析筛选了关键影响因素，但特征选择本身可能存在不确定性。未来的研究可以探索其他可能影响 DR 的因素，以提高模型的预测性能。(3)模型不确定性：模型的构建和融合过程中，不同机器学习算法和融合方法可能导致不同的结果。在研究中，我们使用了 logistic 回归、随机森林和支持向量机等多个模型，并进行了模型融合。虽然投票器模型在预测效果上表现最佳，但不同模型之间的选择也会引入一定程度的不确定性。(4)预测不确定性：模型的预测结果通常伴随着一定的不确定性，这取决于测试集的分布以及模型的性能。本研究使用十折交叉验证方法来评估模型的稳定性和预测不确定性，后续研究中可通过增加数据量、改进特征工程、使用更复杂的模型、改进数据质量控制等提高模型预测效果稳定性。

利益冲突声明　本研究不存在任何利益冲突

参考文献

[1] 孟倩丽，张良，谢洁．几种糖尿病相关眼病的诊断治疗规范[J]．眼科新进展，2022，42(4)：253-261.
Meng QL, Zhang L, Xie J. Diagnosis and treatment of several diabetes-related eye diseases [J]. Recent Advances in Ophthalmology, 2022, 42(4): 253-261.

[2] 蔡淳，贾伟平．人工智能在糖尿病全程健康管理的应用与挑战[J]．中国科学基金，2021，35(1)：104-109.
Cai C, Jia WP. Application and challenges of artificial intelligence in diabetes mellitus health management [J]. China Science Foundation, 2021, 35(1): 104-109.

[3] 张菲菲．2 型糖尿病并发视网膜病变的相关危险因素分析[J]．临床医药文献电子杂志，2019，6(42)：25-26, 28.
Zhang FF. Analysis of risk factors associated with retinopathy complicating type 2 diabetes mellitus [J]. Electronic Journal of Clinical Medicine Literature, 2019, 6(42): 25-26, 28.

[4] 佚名．青少年健康主题数据库 [EB/OL]．[2023-12-25]. https://doi.org/10.12213/11.A0031.202107.209.V1.0.
Anonym. Database on adolescent health topic [EB/OL]. [2023-12-25]. https://doi.org/10.12213/11.A0031.202107.209.V1.0.

[5] 赵林林，赵志远，牛铭云，等．糖尿病病人并发肾病危险因素的 Meta 分析[J]．循证护理，2021，7(12)：1563-1570.
Zhao LL, Zhao ZY, Niu MY, et al. Meta-analysis of risk factors for kidney disease in diabetic patients [J]. Evidence-based Nursing, 2021, 7(12): 1563-1570.

[6] 李景兰，王培红，陈文倩，等．全视网膜光凝治疗后糖尿病视网膜病变进展的危险因素分析[J]．解放军医学院学报，2022，43(10)：1025-1030.
Li JL, Wang PH, Chen WQ, et al. Risk factors for progression of diabetic retinopathy after panretinal photocoagulation [J]. Academic Journal of Chinese PLA Medical School, 2022, 43(10): 1025-1030.

[7] 李翔，邓颖，李新宇，等．DR 患者黄斑中心凹下脉络膜厚度与年龄及病程的相关性分析 [J]．国际眼科杂志，2021，21(10)：1773-1777.
Li X, Deng Y, Li XY, et al. Correlation analysis of foveal choroidal thickness, age and course of disease in patients with diabetic retinopathy of different stages[J]. International Eye Science, 2021, 21(10): 1773-1777.

[8] Yang GR, Li DM, Li L. Comparison of coronary heart disease and stroke in association with diabetic retinopathy in adults with diabetes using a National survey [J]. Diabetes, Metabolic Syndrome and Obesity : Targets and Therapy, 2020, 13: 5079-5084.

[9] Chou YY, Ma J, Su X, et al. Emerging insights into the relationship between hyperlipidemia and the risk of diabetic retinopathy [J]. Lipids in Health and Disease, 2020, 19(1): 241.

[10] 宋亚男，武惠韬，应俊，等．基于机器学习算法探讨糖尿病视网膜病变的风险因素 [J]．解放军医学院学报，2021，42(9)：906-912, 992.
Song YN, Wu HT, Ying J, et al. Risk factors analysis of diabetic retinopathy based on machine learning[J]. Academic Journal of Chinese PLA Medical School, 2021, 42(9): 906-912, 992.

[11] Poshtchaman F, Dehnabi A, Poshtchaman Z, et al. HbA1C, proliferative and non-proliferative retinopathy in diabetic patients [J]. Medicina Clínica Práctica, 2023, 6(3): 100371.

[12] 汤春梦，李文和，杨超超，等．血红蛋白糖化指数与 2 型糖尿病患者视网膜病变的相关性研究[J]．中国糖尿病杂志，2021，29(5)：349-352.
Tang CM, Li WH, Yang CC, et al. Association between hemoglobin glycation index and retinopathy in patients with type 2 diabetes mellitus[J]. Chinese Journal of Diabetes, 2021, 29(5): 349-352.

[13] 丁庭庭，钟兴，杜益君，等．2 型糖尿病患者葡萄糖范围内时间和平均血糖波动幅度与糖尿病视网膜病变相关性的研究[J]．中国糖尿病杂志，2021，29(6)：443-447.
Ding TT, Zhong X, Du YJ, et al. Correlation of time in range and mean amplitude of glycemic excursions with diabetic retinopathy in type 2 diabetes mellitus[J]. Chinese Journal of Diabetes, 2021, 29(6): 443-447.

[14] Liu ZX, Li XL, Wang YL, et al. The concordance and discordance of diabetic kidney disease and retinopathy in patients with type 2 diabetes mellitus: A cross-sectional study of 26,809 patients from 5 primary hospitals in China [J]. Frontiers in Endocrinology, 2023, 14: 1133290.

[15] Zhong JB, Yao YF, Zeng GQ, et al. A closer association between blood urea Nitrogen and the probability of diabetic retinopathy in patients with shorter type 2 diabetes duration [J]. Scientific Reports, 2023, 13(1): 9881.

[16] 王璐瑶，曾智．基于 TOPSIS 法和 RSR 法的江苏省病床使用效率研究[J]．现代医院管理，2022，20(4)：9-12.
Wang LY, Zeng Z. Study on hospital bed efficiency in Jiangsu province based on TOPSIS and RSR [J]. Modern Hospital Management, 2022, 20(4): 9-12.

收稿日期：2023-09-03