网络出版时间:2018 -8 -2 09:39 网络出版地址:http://kns.cnki.net/kcms/detail/34.1065.R.20180731.1310.018.html

空腹血糖受损危险因素的 Logistic 回归及分类树分析

姚 爽[1],谢梦婷[1],邹迪莎[2],黄 芳[1],马小珺[1],江仁美[1],于 健[1]

摘要 **目的** 应用 Logistic 回归模型和分类树模型进行对比分析,探讨空腹血糖受损(IFG)患者的危险因素。**方法** 采用整体抽样的方法,对 5 952 例体检者进行调查,分别应用 Logistic 回归模型和分类树模型对 IFG 的影响因素进行探讨,并采用受试者工作特征曲线(ROC)对两模型分类效能进行检验。**结果** ① Logistic 回归模型显示:高龄、非酒精性脂肪性肝病(NAFLD)、高血压、高三酰甘油、高体重指数是 IFG 患者的危险因素,高密度脂蛋白胆固醇(HDL-C)是 IFG 患者的保护因素($P<0.05$);② 分类树模型共筛选出 5 个危险因素,包括 NAFLD、高龄、高血压、高 TG 和高 LDL-C;③ Logistic 回归模型的 ROC 曲线 Youden 指数为 41.90%,敏感性为 75.00%,特异性为 66.90%,曲线下面积(AUC)值为 0.775。分类树模型的 ROC 曲线的 Youden 指数为 43.96%,敏感性为 73.27%,特异性为 70.69%,AUC 值为 0.775,两种模型 AUC 比较差异无统计学意义。**结论** Logistic 回归模型和分类树模型均具有中等准确性,两模型曲线下面积没有明显差异性,在实际应用中可结合具体情况选择。

关键词 空腹血糖受损;危险因素;Logistic 回归模型;分类树模型;ROC 曲线

中图分类号 R 587.1

文献标志码 A **文章编号** 1000 -1492(2018)09 -1407 -05

doi:10.19405/j.cnki.issn1000 -1492.2018.09.018

空腹血糖受损(impaired fasting glucose,IFG)最早是在 1997 年由美国糖尿病协会(ADA)引入概念,IFG 和糖耐量异常(impaired glucose tolerance,IGT)都是一种处于正常和糖尿病血糖水平之间的状态,IFG 患者虽然仅是空腹血糖(fasting plasma glucose,FPG)轻度增高,但却具有同 IGT 人群一样向糖尿病发展的高危倾向,同时也存在发生大血管并发症的危险[1]。因此,及早识别发现 IFG 患者,控制 IFG 的危险因素,并积极进行早期干预,对减少糖尿病以及其并发症,降低死亡率具有重要意义。Logistic 回归是常被应用于预测结果为二分类变量的传统方法,其对单独危险因素的分析较为明确,但当多因素之间存在复杂相互关系时,会使分析误差增加。而分类树模型作为一种新的统计方法,弥补了传统统计方法的不足,对数据的类型没有严格的限制,目前被广泛应用于疾病发病风险的预测[2-3],但其在对 IFG 危险因素的分析,目前国内研究并不多见。该研究对5 952例体检者的临床资料进行分析,并构建 Logistic 回归模型及分类树模型,并比较分析两模型的准确性,为 IFG 的预防和治疗提供科学依据。

2018 -05 -11 接收

基金项目:广西壮族自治区卫生厅科研课题(编号:Z2012398);广西壮族自治区内分泌科临床重点专科建设项目

作者单位:桂林医学院附属医院[1]内分泌科、[2]肾内科,桂林 541001

作者简介:姚 爽,女,硕士研究生;

于 健,女,教授,主任医师,硕士生导师,责任作者,E-mail:duduyu1623@qq.com

1 材料与方法

1.1 研究对象 选择 2012 年 7 ~ 11 月在桂林医学院附属医院体检中心体检的5 952例汉族人群为研究对象,进行横断面调查,研究对象均来自桂林市七星区、叠彩区、秀峰区、雁山区及象山区等,并获得知情同意,该研究由桂林医学院伦理委员会批准。其中,排除年龄小于 20 岁、孕妇、以及已确诊的糖尿病患者、患有恶性肿瘤等重大疾病以及严重肝肾功能不全者,共5 952例资料完整者纳入研究。研究对象年龄 20 ~ 70(45.82 ± 11.69)岁。其中男 3 074 例(51.65%),平均年龄(46.22 ± 12.04)岁;女 2 878 例(48.35%),平均年龄(45.39 ± 11.29)岁。本研究共检出 IFG 患者 404 例,平均年龄(53.04 ± 9.55)岁,其中男 244 例(60.4%),女 160 例(39.6%)。成人 IFG 的粗患病率为 6.79%(男:7.94%;女:5.56%)。

1.2 问卷调查 在研究对象进行体检时,采用问卷调查和体检的方法,并通过专业人员进行记录,获取其基本资料,包括个人基本情况(姓名、性别、年龄、民族、婚姻状况、职业、文化程度);既往史:高血压、非酒精性脂肪性肝病(nonalcoholic fatty liver disease,NAFLD)、糖尿病等;生活习惯(饮食、吸烟史、饮酒等)、家族史及运动情况等。

1.3 体格测量及测量方法 经过专业培训和考核的体检中心专业人员,采用统一标准,使用经校准的 SK-CK 超声波人体秤,时间为早晨 7:30 ~ 9:00,被

测量者免冠、赤脚、空腹、排空膀胱、着轻单衣、以"立正"姿势站立，身高测量精确至 0.1 cm，体质量测量精确至 0.1 kg。计算体重指数(body mass index，BMI)，BMI = 体重(kg) /身高(m^2)。研究者要避免饮酒、吸烟、喝茶和咖啡，避免剧烈运动，安静休息 5 ~ 10 min，呈坐位，使用水银血压计测量右臂肱动脉血压，手臂与心脏在同一水平，测量 3 次，每次间隔 30 s，取 3 次测量的平均值进行分析，数值精确到 1 mmHg。

1.4 实验室检查及测量方法 所有受检者至少禁食 10 h，清晨空腹采集肘静脉血(5 ml)，在口服 75 g 葡萄糖负荷后 120 min 再次采集血液样本。生化指标：使用罗氏 Cobas C501 全自动生化分析仪检测：FPG、口服葡萄糖耐量实验(oral glucose tolerance test，OGTT)、2 h 血糖(2-h plasma glucose，2 hPG)、尿酸(uric acid，UA)、三酰甘油(triacylglycerol，TG)、总胆固醇(total cholesterol，TC)、低密度脂蛋白胆固醇(low-density lipoprotein cholesterol，LDL-C)及高密度脂蛋白胆固醇(high-density lipoprotein cholesterol，HDL-C)。所有研究对象由固定的超声专科医师行肝胆 B 超检查。

1.5 诊断标准 NAFLD 的诊断参照中华医学会肝病分会脂肪肝及酒精性肝病学会组制定的非酒精性脂肪性肝病诊断标准[4]。高血压的诊断标准按《中国高血压防治指南 2010》定义[5]：3 次测量平均值收缩压(SBP) ≥140 mmHg(1 mmHg = 0.133 kPa)和(或)舒张压(DBP) ≥90 mmHg 为高血压或近 2 周内服降压药血压正常者，排除继发性高血压者。血脂异常诊断标准：参照 2007 年中国成人血脂异常防治指南的标准[6]：TC ≥ 6.22 mmol/L；TG ≥ 2.26 mmol/L；LDL-C ≥ 4.14 mmol/L；HDL-C < 1.04 mmol/L，出现以上任何一项即诊断为血脂异常。参照中国 2 型糖尿病防治指南[7]，IFG 组诊断标准为 6.1 mmol/L≤FPG < 7.0 mmol/L 且 OGTT 2 h 血糖 < 7.8 mmol/L。将 24.0 kg/m^2≤BMI < 28 kg/m^2 诊断为超重，BMI ≥ 28 kg/m^2 诊断为肥胖[8]。高尿酸血症诊断标准[9]：男性血尿酸≥420 μmol/L 或女性血尿酸≥360 μmol/L 者为高尿酸血症者。

1.6 统计学处理 采用 SPSS 20.0 软件和 Medcalc(版本 15.2.2)软件进行分析，分别建立 Logistic 回归模型和分类树模型，比较分析两种方法结果的差异，以 $P < 0.05$ 为差异有统计学意义。

2 结果

2.1 Logistic 回归分析 IFG 影响因素中，年龄取平均值 45.8 为分界点，其余因素结合专业知识进行赋值，见表 1。以 IFG 有无作为因变量，各影响因素作为自变量进行单因素 Logistic 回归，结果显示：高龄、男性、高 TG、高 TC、高 LDL-C、高 BMI、高尿酸血症、NAFLD、高血压是 IFG 的危险因素，高 HDL-C 是 IFG 患者的保护因素($P < 0.05$)，见表 2。将单因素 Logistic 分析中有意义的因素作为自变量，进行多因素 Logistic 回归分析，控制混杂因素后，结果显示：高龄、NAFLD、高血压、高 TG、高 BMI 是 IFG 患者的危险因素，高 HDL-C 是 IFG 患者的保护因素($P < 0.05$)，见表 3。

表 1 IFG 患者影响因素主要变量及赋值

变量	赋值
IFG	无 = 0；有 = 1
性别	女 = 0；男 = 1
年龄(岁)	< 45.8 = 0；≥45.8 = 1
TG(mmol/L)	< 2.26 = 0；≥2.26 = 1
TC(mmol/L)	< 6.22 = 0；≥6.22 = 1
LDL-C(mmol/L)	< 4.14 = 0；≥4.14 = 1
HDL-C(mmol/L)	< 1.04 = 0；≥1.04 = 1
BMI(kg/m^2)	< 24.0 = 0；≥24.0 = 1
高尿酸血症	无 = 0；有 = 1
NAFLD	无 = 0；有 = 1
高血压	无 = 0；有 = 1

表 2 IFG 影响因素的单因素 Logistic 回归分析结果

变量	*B*	*SE*	*Wald*	*P* 值	*OR* 值	95% *CI*
性别	0.382	0.105	13.156	< 0.001	1.465	1.192 ~ 1.800
高龄	1.311	0.121	117.010	< 0.001	3.710	2.926 ~ 4.705
高 TG	1.180	0.111	113.715	< 0.001	3.254	2.619 ~ 4.042
高 TC	1.017	0.145	49.162	< 0.001	2.765	2.081 ~ 3.674
高 LDL-C	0.826	0.124	44.398	< 0.001	2.284	1.792 ~ 2.913
高 HDL-C	−0.305	0.141	4.714	0.030	0.737	0.560 ~ 0.971
高 BMI	0.984	0.112	77.627	< 0.001	2.675	2.149 ~ 3.329
高尿酸血症	0.639	0.123	26.927	< 0.001	1.894	1.488 ~ 2.411
NAFLD	1.329	0.107	152.980	< 0.001	3.776	3.059 ~ 4.661
高血压	1.300	0.105	151.937	< 0.001	3.670	2.985 ~ 4.513

OR：比值比

表3 IFG影响因素的多因素Logistic回归分析结果

变量	*B*	*SE*	*Wald*	*P*值	*OR*值	95% *CI*
性别	0.135	0.116	1.353	0.245	1.145	0.912~1.437
高龄	1.028	0.130	61.649	<0.001	2.796	2.168~3.607
高TG	0.798	0.136	34.307	<0.001	2.222	1.701~2.902
高TC	0.330	0.211	2.446	0.118	1.390	0.920~2.102
高LDL-C	0.070	0.180	0.151	0.698	1.072	0.754~1.525
高HDL-C	−0.454	0.164	7.650	0.006	0.635	0.461~0.876
高BMI	0.295	0.128	5.300	0.021	1.343	1.045~1.726
高尿酸血症	0.030	0.140	0.046	0.830	1.030	0.784~1.355
NAFLD	0.832	0.126	43.849	<0.001	2.297	1.796~2.938
高血压	0.685	0.116	34.805	<0.001	1.983	1.580~2.490

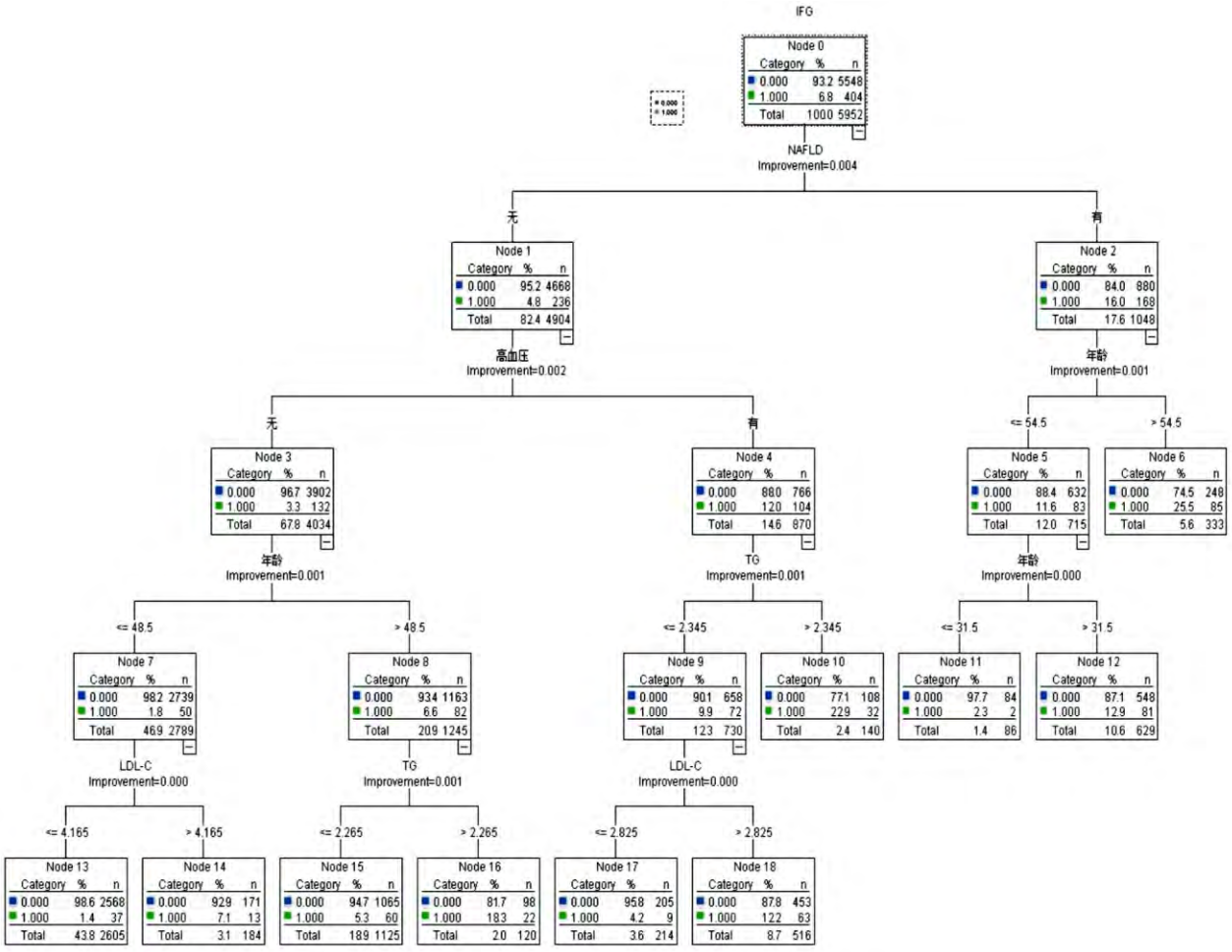

图1 IFG患者的分类树模型

2.2 构建分类树模型 利用分类树模型分析IFG的危险因素，将IFG患者赋值为1，将正常组赋值为0。将性别、年龄、TG、TC、HDL-C、LDL-C、BMI、UA、NAFLD、高血压作为IFG患者的影响因素构建分类树模型。基于根节点和子节点的最小样本大小的限制后的增长和修剪后，得出分类树模型包括4层，19个节点，以及10个终末节点。树模型提示NAFLD、高龄、高血压、高TG和高LDL-C 5个变量是IFG的危险因素。见图1。

2.3 Logistic回归与分类树模型的受试者工作特征曲线（receiver operator characteristic curve，ROC）比较 根据Logistic回归和分类树模型所得的预测变量作为测试变量，IFG组作为状态变量进行绘制ROC曲线。Logistic回归模型的ROC曲线Youden指数为41.90%，敏感性为75.00%，特异性为66.90%，曲线下面积（area under curve，AUC）值

为 0.775。分类树模型的 ROC 曲线的 Youden 指数为 43.96%，敏感性为 73.27%，特异性为 70.69%，AUC 值为 0.775，两种模型均具有中等准确性，两种模型 AUC 比较差异无统计学意义（$P=0.9610$），说明两模型没有明显差异性，见图 2、表 4。

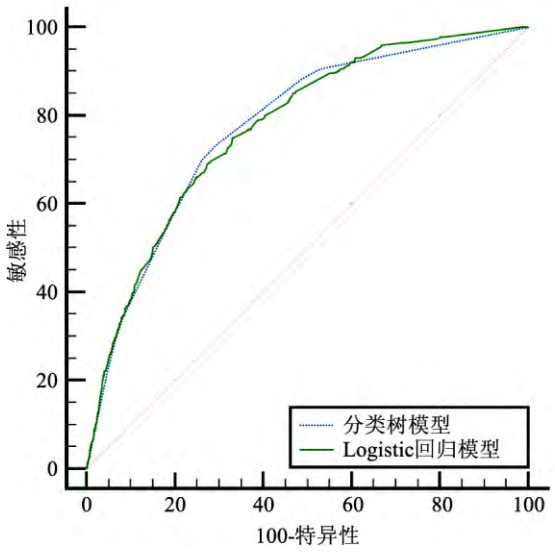

图 2　Logistic 回归模型与分类树模型的 ROC 曲线比较

表 4　Logistic 回归模型与分类树模型的 ROC 曲线参数比较

参数	Logistic 模型的 ROC 曲线	分类树模型的 ROC 曲线
敏感性(%,95% *CI*)	75.00 (70.5~79.1)	73.27(68.7~77.5)
特异性(%,95% *CI*)	66.90 (65.6~68.1)	70.69(69.5~71.9)
AUC (95% *CI*)	0.775 (0.764~0.785)	0.775(0.764~0.786)
Youden 指数(%)	41.90	43.96

3　讨论

糖尿病的发展是一个缓慢的过程，从血糖升高到出现一系列的临床症状，需要数年时间。IFG 是正常糖代谢与糖尿病之间的中间状态，是糖尿病和心血管疾病的危险因子之一[10]。探讨 IFG 危险因素，加强预防和保健工作，对于避免和减少糖尿病的发生具有重要意义。

控制混杂因素后，多因素 Logistic 回归结果显示：高龄、NAFLD、高血压、高 TG、高 BMI 是 IFG 患者的危险因素，高 HDL-C 是 IFG 患者的保护因素，与之前研究[11-12]结果一致。高龄人群 IFG 发病率高，有以下两点原因：第一，随着年龄的增长，肌肉组织逐渐老化、减少，机体储存和利用葡萄糖的能力下降；第二，随着年龄增长，β 细胞对内源性胰岛素刺激因子的反应性降低，胰岛素分泌减少，引起血糖升高[13]。我国已进入老龄化社会，据有关统计，2012 年我国 60 岁及以上老年人口数量为 1.94 亿，老龄化水平已达到 14.3%。因此，加强对高龄人群血糖的筛查，对改善其生活质量有着不可忽视的作用。NAFLD 是一种无过量饮酒史肝实质细胞脂肪变性和脂肪贮积为特征的临床病理综合征。NAFLD 的细胞脂肪堆积，会引发并加剧胰岛素抵抗，同时 NAFLD 患者的血液和肝脏中的游离脂肪酸（FFA）、肿瘤坏死因子-α（TNF-α）、纤溶酶原激活物抑制物-1（PAI-1）、瘦素[14-15]等表达增加，都可促进胰岛分泌胰岛素，形成高胰岛素血症和胰岛素抵抗，继而导致血糖升高。

近年来，诸多针对 IFG 研究主要集中在危险因素的调查上，通常采用的是 Logistic 回归模型对影响因素进行筛选。本文中应用 Logistic 回归计算出各危险因素的 *OR* 值：高龄人群发病风险是低龄人群的 2.796 倍，NAFLD 患者发病风险是无 NAFLD 者的 2.297 倍，高 TG 患者发病风险是 TG 正常者的 2.222 倍，高血压患者发病风险是血压正常者的 1.983 倍，高 BMI 人群发病风险是 BMI 正常人群的 1.343 倍。这些结果可帮助研究者更直观地了解各影响因素的作用及大小，但是，此方法易受到共线性的影响，无法估计各个自变量间可能存在的相互作用。

分类树分析是一种非参数回归模型，利用递归分型将人群分为不同的亚群，近年来在国外大量应用。分类树模型在处理变量之间的相互作用中，可有效的处理数据中自变量缺失的问题，不仅能把缺失值归到众数的范畴，还可以将其设置为单独的一个分类，使结果可以不受变量的共线性的影响，并以树形图的形式显示结果，模型直观、明确、清晰、有层次。本研究应用分类树模型对 IFG 患者影响因素的信息进行深入挖掘，共筛选出 5 个 IFG 危险因素，包括 NAFLD、高龄、高血压、TG 和 LDL-C，与 Logistic 回归模型分析的结果基本相符，另外，树形图还可显示各因素之间的交互作用，本模型提示，高龄和有 NAFLD 人群患病率为 25.5%，高于低龄和有 NAFLD 人群的患病率（11.6%）。但是，与其他统计分析方法一样，分类树自身也存在缺点：第一，分类树对各因素的单独效应的定量解释不如 Logistic 模型明确，在本研究中使用 Logistic 模型得出 *OR* 值，可明确的判断出影响 IFG 患者的危险因素和保护因素。第二，在样本容量小的情况下模型稳定性差，因本研究数据量大，因此分类树模型稳定性良好。

最后，应用 ROC 曲线来评估两模型的准确度，结果显示：两模型曲线均位于参考线上方，分类树模型的 ROC 曲线更为平滑，说明分类树模型在稳健性

上更有优势，两者曲线下面积差异无统计学意义（$P>0.05$）。但因两种分类方法各有利弊，具体结论要因实际情况和具体的数据情况来表现与判定。提示在临床研究上，可结合实际应用情况，进行优势互补，综合比较。

然而，目前本研究也有一些局限性：首先，所有包括成人受试者来自一个体检中心，因此并不能代表所有桂北地区人群。其次，需要进一步的前瞻性研究来调查危险因素和 IFG 患者的因果关系。

参考文献

[1] 甘华葵，丁 燕，袁平宗，等. 内江市 20～60 岁机关从业人员空腹血糖受损及相关因素横断面的调查研究[J]. 中国糖尿病杂志，2013，21(3)：204－6.

[2] 刘建平，程锦泉，张仁利，等. 应用分类树模型构建缺血性脑卒中发病风险的预测模型[J]. 中国慢性病预防与控制，2012，20(3)：254－8.

[3] 蔡晓楠，张丹丹，严亚琼，等. 应用分类树模型构建耐多药结核病发病风险模型[J]. 中华疾病控制杂志，2016，20(1)：91－5.

[4] 中华医学会肝脏病学分会脂肪肝和酒精性肝病学组. 非酒精性脂肪性肝病诊断标准[J]. 中华肝脏病杂志，2003，2：8.

[5] 中国高血压防治指南修订委员会. 中国高血压防治指南 2010[J]. 中华高血压杂志，2011，19(8)：701－43.

[6] 中国成人血脂异常防治指南制订联合委员会. 中国成人血脂异常防治指南[J]. 中华心血管病杂志，2007，35(5)：393－4.

[7] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2013 年版)[J]. 中国糖尿病杂志，2014，6(7)：447－98.

[8] 中华人民共和国卫生部疾病控制司. 中国成人超重和肥胖症预防控制指南[M]. 北京：人民卫生出版社，2006：2－4.

[9] 中华医学会内分泌学分会. 高尿酸血症和痛风治疗的中国专家共识[J]. 中华内分泌代谢杂志，2013，29(11)：913－20.

[10] 李 雷，杨荣礼，李平静，等. 空腹血糖受损与糖耐量受损患者血管内皮功能的对比研究[J]. 临床心血管病杂志，2012，28(7)：546－8.

[11] 吴云涛，邢爱君，吴寿岭，等. 空腹血糖受损人群自然转归及其影响因素的观察[J]. 中国糖尿病杂志，2013，21(8)：728－30.

[12] 张秀英，周翔海，罗樱樱，等. 不同糖耐量状态血清脂联素水平的比较及其相关因素的研究[J]. 中国糖尿病杂志，2013，21(10)：896－9.

[13] 葛国兴，钟亚萍，倪桂娜. 2009 年－2011 年绍兴部分人群空腹血糖变化趋势调查[J]. 中国卫生检验杂志，2012，22(7)：1683－6.

[14] Gnacinska M，Malgorzewicz S，Lysiak-Szydlowska W，et al. The serum profile of adipokines in overweight patients with metabolic syndrome[J]. Endokrynol Pol，2010，61(1)：36－41.

[15] 李枫林，张 宝，管石侠，等. 非酒精性脂肪肝病大鼠 IL-1β、IL-18、TNF-α 水平的变化[J]. 安徽医科大学学报，2016，51(3)：351－4.

Logistic regression analysis and classification tree analysis of risk factors for impaired fasting glucose

Yao Shuang[1]，Xie Mengting[1]，Zou Disha[2]，et al

([1]*Dept of Endocrinology*，[2]*Dept of Nephrology*，*The Affiliated Hospital of Guilin Medical College*，*Guilin* 541001)

Abstract ***Objective*** To analyze the risk factors of impaired fasting glucose (IGF) patients using the methods of Logistic regression and classification tree analysis and to compare the result of these two methods. ***Methods*** 5 952 cases receiving physical examination were investigated by the method of cluster sampling. The influence factors of IFG were analyzed by Logistic regression and classification tree analysis，then the corresponding performance of two models are examined by receiver operating characteristic (ROC) curve. ***Results*** ① Logistic regression model showed that advanced age，nonalcoholic fatty liver disease (NAFLD)，hypertension，high triacylglycerol (TG) and high body mass index (BMI) were risk factors，while high high-density lipoprotein cholesterol (HDL-C) was the protective factor of IFG($P<0.05$). ② Five factors were selected by classification tree，including NAFLD，advanced age，hypertension，high TG，high low-density lipoprotein cholesterol (LDL-C). ③ Youden's index of ROC curve in Logistic regression was 41.90%，sensitivity was 75.00%，specificity was 66.90%，the area under curve(AUC) value was 0.775. Youden's index of ROC curve in classification tree was 43.96%，sensitivity was 73.27%，specificity was 70.69%，AUC value was 0.775. There was no significant difference in AUC between the two models. ***Conclusion*** Logistic regression models and classification tree models have moderate accuracy. There is no significant difference in the area under the curve between the two models，which can be selected according to the actual situations.

Key words impaired fasting glucose; risk factor; Logistic regression; classification tree model; ROC curve