基于 CHARLS 数据的老年人糖尿病影响因素研究

王少锋 梁鸿

(广东东软学院，广东 佛山 528225)

摘　要：糖尿病常引发心血管疾病、视网膜病变等多种并发症，对个人健康和社会医疗体系造成巨大负担。为助力医疗机构与公众科学防控糖尿病，本文基于 CHARLS（2020）数据，采用相关性分析与逻辑回归模型探讨各类因素对糖尿病患病的影响。结果表明，吸烟和年龄增长与糖尿病风险呈正相关；而男性、农村居住、有配偶、有工资收入及南方地区居住则具有保护作用，可降低患病概率；运动得分与饮酒频率影响微弱，基本可忽略。由于类别 0（未患病）样本远多于类别 1（患病），模型倾向预测多数类，导致准确率仅为 54%；优化后类别 1 的评估指标有所改善；ROC AUC 接近 0.5，提示模型区分能力有限。据此，预防糖尿病应从戒烟限酒、规律运动、控制体重、合理膳食、调节情绪等方面入手。糖尿病的干预需整合健康教育、营养管理、运动疗法、血糖监测与药物治疗，实施综合管理。

关键词：相关性分析；逻辑回归；糖尿病

DOI：doi.org/10.70693/rwsk.v1i10.1611

1 研究背景和目的

糖尿病是一种影响全球范围内人群健康的慢性疾病。其特点是慢性高血糖、伴有胰岛素分泌不足或作用障碍，导致碳水化合物、脂肪、蛋白质代谢紊乱，造成多种器官的慢性损伤、功能障碍甚至衰竭[1]，根据 WHO 和 IDF 分类，糖尿病分为 1 型、2 型、其他特殊类型和妊娠期糖尿病，其中 2 型占 90%以上，主要表现为“三多一少”。中国健康与养老追踪调查（CHARLS 2020）数据显示，2018 年我国糖尿病患者已达 1.88 亿。患病率持续上升，2010 年为 7.51%，2018 年达 13.36%。糖尿病易引发心血管疾病、视网膜病变等并发症，给个人和社会带来沉重负担。深入分析其危险因素，对预防控制、降低发病率和改善公众健康具有重要意义。

本文基于 CHARLS 数据，通过相关性分析和逻辑回归模型探究糖尿病的危险因素，旨在为疾病预防提供科学依据。研究有助于识别关键影响因素，指导医疗单位和个人采取有效防控措施，降低糖尿病发生风险。

2 文献综述

2.1 糖尿病的流行病学研究

据《中国 2 型糖尿病防治指南（2017 年版）》资料[2]，自 20 世纪 80 年代以来，我国成人糖尿病患病率显著增加，从 1980 年的 0.67%，增加至 2013 年的 10.9%。第七次全国人口普查（2020 年）的数据显示，我国 60 岁以上的糖尿病患者约 7920 万，占了 60 岁以上人口的 18.70%[3]。

糖尿病流行病学特点以 2 型糖尿病为主要、常见类型，占 90%以上；存在地区差异性，发达地区明显高于不发达地区，城市高于农村[4]。

2.2 糖尿病的危险因素研究

糖尿病作为一种慢性病，其病因是由多种引起因素的糖代谢紊乱，因此其危险因素往往具有多样性。有学者指出合理膳食、适量运动、戒烟限酒以及良好休息会有效预防糖尿病和降低糖尿病发病率[5]。也有学者认为，空腹血糖和口服葡萄糖耐量水平与糖尿病诊断具有显著相关性，与糖尿病患病率呈现正相关[6]。此外，有研究分析，

作者简介：王少锋(1988—)，男，硕士，广东东软学院助教，研究方向为医疗信息化，慢病数据分析；
梁　鸿(2002—)，男，广东东软学院 2021 级健康服务与管理专业。
通讯作者：王少锋

高血压、家族史、糖调节受损、超重以及年龄也是影响糖尿病生存率的重要因素[7]。糖尿病主要影响因素包括生

活方式、家族史、人口特征、体测指标、心脑血管状况及生化指标。性别影响尚存争议，有研究认为无关[8]，亦有指出男性患病率高于女性[9]。

2.3 逻辑回归模型在疾病研究中的应用

逻辑回归模型可综合多因素分析糖尿病风险，揭示发病规律。研究表明其在准确率、敏感度及 AUC 等方面表现良好，稳定性与泛化能力强，适用于糖尿病风险预测与辅助诊断，具有重要应用价值[10]。

3 研究方法

3.1 数据来源和研究对象

本研究使用的数据集来源于中国健康与养老追踪调查（CHARLS）2020 年的调查数据。

3.2 研究工具

使用 SPSS 软件进行相关性分析，使用 Python 中数据分析相关库建立逻辑回归模型。

3.3 数据预处理

对存在缺失值和异常值所在的行进行删除。

3.4 被解释变量

以医生是否告诉自己患有糖尿病为因变量。在是否患糖尿病中，来自问卷中 DA003 列，对应的问卷题目为“是否有医生告诉自己患有糖尿病”，答案有两个选项，分别是“有”、“无”。在将分类变量由文本转换为数值时，患有糖尿病为 1，没有患糖尿病为 0。

3.5 解释变量

以运动得分、年龄、是否吸烟、饮酒频率、性别、一般居住地、有无配偶、有无工资以及所在地为南方或者北方作为自变量。

在运动得分中，将每周有无至少做高/中/低强度运动十分钟和每周做几天高/中/低强度运动至少十分钟两种类别变量，设定高强度运动 3 分/天，中强度运动 2 分/天，低强度运动 1 分/天，最后计算合计分数，即运动得分；

在一般居住地因素中，来自问卷中 BA008 列，对应的问卷题目为“一般居住地城乡类别”，答案有四个选项，分别是“城或镇中心区”、“城乡或镇乡结合区”、“农村”、“特殊区域”。在将分类变量由文本转换为数值时，城或镇中心为 1，城乡结合地区为 2，乡村为 3，特殊区域为 4；

在工资因素中，过去一年内有工资为 1，没有工资为 0；

在所在居住地为南方或者北方中，将所得到的 PSU 表格中的 communityID 列由 Python 代码转换成对应的省份信息，再将所得到的每个省份按照其地域划分转变成南方/北方。在将分类变量由文本转换为数值时，南方为 1，北方为 0。

3.6 相关性分析

相关性分析的基本原理是通过计算相关系数来衡量两个变量之间的关联程度[11]。本文选择皮尔逊卡方检验，因为皮尔逊卡方检验适用于线性关系的测量，并且数据呈正态分布。

3.7 逻辑回归模型

本文逻辑回归模型的构建主要步骤如下。

1.数据集划分

训练集中有 8214 个样本，每个样本有 9 个特征；测试集中 2054 个样本，每个样本有 9 个特征。

2.模型优化

根据训练集里类别 0 和类别 1 的样本量比例，对类别 1 进行过采样，使得类别 0 和类别 1 的最终样本量大致

相等，避免模型数据不平衡的情况。

4 研究结果

4.1 描述性分析结果

根据 CHARLS（2020）调查数据，剔除其中年龄小于 45 岁以下的对象和缺失数据，本文最终纳入的例数有 10268 例，患有糖尿病的有 570 例，患病率为 5.55%。其中女性有 8352 例，占总体样本的 81.34%；男性有 1916 例，占总体样本的 18.66%。

4.2 相关性分析结果

本文的变量分为解释变量和被解释变量，将解释变量和被解释变量进行卡方检验。

4.2.1 不同年龄段对患糖尿病影响情况

糖尿病与年龄分段的相关性分析结果如表 4.1。

表 4.1 是否患有糖尿病*年龄分段交叉表

	年龄段 0	年龄段 1	年龄段 2	年龄段 3	总计
未患糖尿病	5000	3777	885	36	9698
患有糖尿病	237	276	57	0	570
总计	5237	4053	942	36	10268

45–60 岁糖尿病阳性率 4.53%，61–75 岁升至 6.81%，76–90 岁为 6.05%，91–107 岁为 0%（样本仅 36 人）。卡方检验显示年龄分段与糖尿病患病率显著相关（χ^2=25.318，p<0.001），表明两者存在统计学关联。

4.2.2 不同运动得分分段对患糖尿病影响情况

糖尿病与运动得分分段的相关性分析结果如表 4.2。

表 4.2 是否患有糖尿病*运动得分分段交叉表

	分段 0	分段 1	分段 2	分段 3	总计
未患糖尿病	3771	3638	1244	1045	9698
患有糖尿病	229	221	67	53	570
总计	4000	3859	1311	1098	10268

运动得分 0–10 分组糖尿病阳性率 5.73%，11–21 分组为 5.73%，22–32 分组为 5.11%，33–42 分组为 4.83%。卡方检验显示运动得分分段与糖尿病患病率无显著关联（χ^2=2.041，df=3，p>0.05），差异无统计学意义，表明两者无显著相关性。

4.2.3 不同性别对患糖尿病影响情况

糖尿病与性别的相关性分析结果如表 4.3。

表 4.3 是否患有糖尿病*性别交叉表

	性别女 0	性别男 1	总计
未患糖尿病	7887	1811	9698
患有糖尿病	465	105	570
总计	8352	1916	10268

女性糖尿病阳性率 5.6%，男性为 5.5%。卡方检验显示性别与糖尿病患病率无显著关联（χ^2=0.023，p=0.88），差异无统计学意义，表明两者独立，性别可能不影响患病风险。但现有研究提示女性或具代谢保护作用[6]，需结合其他证据进一步分析。

4.2.4 是否吸烟对患糖尿病影响情况

糖尿病与吸烟的相关性分析结果如表 4.4。

表 4.4 是否患有糖尿病*是否吸烟交叉表

	不吸烟 0	吸烟 1	总计

未患糖尿病	9344	354	9698
患有糖尿病	536	34	570
总计	9880	388	10268

不吸烟者糖尿病阳性率 5.4%，吸烟者为 8.8%。卡方检验显示吸烟与糖尿病患病率显著相关（χ^2=7.933，p=0.005<0.05），差异具有统计学意义，拒绝原假设，表明吸烟与糖尿病呈正相关，是其危险因素之一。

4.2.5 饮酒频率对患糖尿病影响情况

糖尿病与饮酒频率的相关性分析结果如表 4.5。

表 4.5 是否患有糖尿病*饮酒频率交叉表

	从来不饮酒	一个月<1 次	一个月 > 1 次	总计
未患糖尿病	7411	883	1404	9698
患有糖尿病	446	45	79	570
总计	7857	928	1483	10268

从不饮酒者糖尿病阳性率 5.7%，每月饮酒少于一次为 4.8%，多于一次为 5.3%。卡方检验显示饮酒频率与糖尿病无显著关联（χ^2=1.250，p=0.535>0.05），差异无统计学意义，表明两者独立。尽管饮酒可能影响患者生活质量[9]，但本数据中未发现其与患病率显著相关，需结合其他研究进一步分析。

4.2.6 有无配偶对患糖尿病影响情况

糖尿病与有无配偶的相关性分析结果如表 4.6。

表 4.6 是否患有糖尿病*有无配偶交叉表

	无配偶 0	有配偶 1	总计
未患糖尿病	1725	7973	9698
患有糖尿病	119	451	570
总计	1844	8424	10268

无配偶者糖尿病阳性率 6.4%，有配偶者为 5.3%。卡方检验显示两者无显著关联（χ^2=3.489，p=0.062>0.05），差异无统计学意义，表明有无配偶与糖尿病患病率无显著相关性，在本数据中未发现其对患病风险有显著影响。

4.2.7 有无工资对患糖尿病影响情况

糖尿病与有无工资的相关性分析结果如表 4.17。

表 4.7 是否患有糖尿病*有无工资交叉表

	无工资 0	有工资 1	总计
未患糖尿病	7511	2187	9698
患有糖尿病	476	94	570
总计	7987	2281	10268

无工资收入者糖尿病阳性率 6.0%，有工资收入者为 4.1%。卡方检验显示两者显著相关（χ^2=11.441，df=1，p=0.001<0.05），差异具有统计学意义，表明是否有工资收入与糖尿病患病率存在显著关联，有收入者患病风险较低。

4.2.8 居住地类别对患糖尿病影响情况

糖尿病与居住地类别的相关性分析结果如表 4.8。

表 4.8 是否患有糖尿病*居住地类别交叉表

	市中心 1	城乡结合 2	乡村 3	特殊区域 4	总计
未患糖尿病	2419	1161	6111	7	9698
患有糖尿病	154	77	339	0	570
总计	2573	1238	6450	7	10268

市中心居民糖尿病阳性率 6.0%，城乡结合部为 6.2%，乡村为 5.3%，特殊区域为 0%。卡方检验显示居住地与糖尿病无显著关联（χ^2=3.465，p=0.325>0.05），差异无统计学意义，表明两者独立。尽管部分研究提示居住地可能影响患病风险[5,10]，但本数据未发现显著相关性。

4.2.9 南北方对患糖尿病影响情况

糖尿病与南北方的相关性分析结果如表 4.9。

表 4.9 是否患有糖尿病*南北方交叉表

	北方 0	南方 1	总计
未患糖尿病	4397	5301	9698
患有糖尿病	287	283	570
总计	4684	5584	10268

北方居民糖尿病阳性率 6.1%，南方为 5.1%。卡方检验显示地区与糖尿病患病率显著相关（χ^2=5.451，p=0.020<0.05），差异具有统计学意义，表明所在省份与糖尿病风险存在显著关联，北方地区患病率更高。

4.3 逻辑回归分析

4.3.1 解释变量影响情况

将解释变量和被解释变量进行逻辑回归分析，在模型构建阶段有调整不同类别权重和采取过采样的方法。逻辑回归模型系数和优势比见表 4.19。

表 4.10 逻辑回归模型系数表

特征变量	系数	优势比（OR）	OR 排序
是否吸烟	0.6328	1.8829	1
运动得分	0.0048	1.0048	3
饮酒频率	-0.1019	0.9031	6
年龄	0.0153	1.0154	2
性别	-0.0528	0.9486	5
居住地类别	-0.1248	0.8827	7
有无配偶	-0.0286	0.9718	4
有无工资	-0.3041	0.7378	9
南北方	-0.2438	0.7837	8

逻辑回归显示：吸烟 (OR=1.8829) 与糖尿病风险正相关；有工资收入 (OR=0.7378)、南方居住 (OR=0.7837)、有配偶（OR=0.9718）为保护因素，降低患病风险。运动得分（OR=1.0048）、饮酒频率（OR=0.9031）、年龄（OR=1.0154）、性别（OR=0.9486）和居住地城乡类别（OR=0.8827）影响微弱，接近 1，提示相关性较弱或不显著，需结合其他研究进一步验证。

4.3.2 模型性能分析

逻辑回归模型优化前后性能对比如下：优化前，模型准确率达 94%，但类别 1 召回率为 0，所有样本均被预测为类别 0，存在严重偏向。优化后，准确率降至 54%，但类别 1 召回率提升至 45%，精确率升至 6%，表明模型开始识别患病样本，F1 分数有所改善。宏平均和加权平均指标显示整体表现略有提升。然而，类别 0 召回率从 1.00 降至 0.55，假阳性显著增加。ROC AUC 值保持 0.53，接近随机水平，说明模型区分能力仍差。OR 值未变，提示特征权重未调整。结果表明，尽管过采样缓解了数据不平衡问题，模型对少数类识别能力有所提高，但整体性能仍不理想，需进一步优化特征选择与阈值设定。

5 讨论与建议

5.1 糖尿病危险因素的分析讨论

在本文中，按照逻辑回归模型系数表，是否吸烟和年龄是影响糖尿病患病率最重要的特征，其 OR 值大于 1，表明吸烟和年龄增加会增加糖尿病患病的可能性；男性、农村居住、有领工资、有配偶和南方居住会减少糖尿病患病的可能性；饮酒频率和运动得分对糖尿病患病率的影响非常小。

5.2 与既有研究的对比

5.2.1 逻辑回归模型性能对比

在本实验中，该模型效果在预测类别 0 时的表现上比类别 1 的好，并且该模型的 AUC 值只有 53%，而李婷等人的实验研究的逻辑回归模型的 AUC 值达到了 82%[10]，马文彬等人研究的逻辑回归模型的 AUC 值达到了 90.3%[12]、郑家浩等人研究的逻辑回归模型的 AUC 值也达到了 90.4%[13]。可能是因为本文的类别 0 的样本数量比类别 1 太多，导致数据不平衡，已经在模型构建阶段有调整不同类别权重和采取过采样的方法，但仍旧不理想。

5.2.2 危险因素的对比

主要提取几个常见的因素进行对比。

在吸烟方面，本文提示吸烟量的增加会使糖尿病患病的可能性增加，而王金虹[5]、郝家乐[14]、苏银霞[15]的研究也表明戒烟对维护健康和降低糖尿病发病率至关重要；

在年龄方面，本文提示年龄与糖尿病患病率呈正比，苏银霞[15]、ZHE L[9]、Asadi[8]、彭若萱[16]的研究也表明年龄的增长会导致糖尿病患病率的增长；

在饮酒频率方面，本文提示与糖尿病患病率呈弱负相关，而王金虹[5]、郝家乐[14]、苏银霞[15]的研究以及医学常识表明饮酒会提高糖尿病患病率，因此有可能是本文患有糖尿病的样本数与没有患糖尿病的样本数严重不平衡导致的；

在运动方面，本文提示运动对糖尿病患病率的影响非常小，几乎可以忽略不计，而王金虹[5]、王旭[7]的研究表明运动可以有效预防糖尿病，有可能是本研究样本量不足造成的，或者可能是运动得分的计算方案需要进一步优化；

在性别方面，本文的相关性分析提示，女性患糖尿病的几率（0.055%）比男性（0.054%）大 0.001%，而 ZHE L[9]的研究表明女性身份是糖尿病患病的保护因素，而 Asadi[8]、王海霞[17]的研究提示性别与糖尿病患病率之间并没有显著相关性，因此可能是女性样本量比男性多，女性样本量占总体样本的 81.34%；

在居住地方面，本文认为在农村居住的糖尿病患病的可能性会低于在城市居住，而有的研究也发现，我国城市地区的糖尿病患病率显著高于农村地区，可能是农村体力活动比城市更多；

在所在地为南方或者北方方面，本文的相关性分析提示，所在地为南方或者北方与糖尿病患病率之间存在显著相关性，结合逻辑回归分析和交叉表，发现居住在南方的人与糖尿病患病率呈负相关，表明南方人的糖尿病患病率比北方人低，可能是因为北方口味较重，南方口味较淡，并且南方饮食习惯更多以水稻为主食，北方更多以小麦为主食。

5.3 预防与干预措施的建议

基于研究结果，建议从生活方式入手预防糖尿病：戒烟限酒，因吸烟显著增加患病风险；注意地域差异，南方居住者风险较低，可能与以大米为主食和较北方清淡的饮食习惯有关；有配偶者风险较低，提示家庭支持的重要性。糖尿病的干预应遵循早期发现、长期管理的原则，综合实施教育、营养、运动、监测和药物治疗。

6 结语

本研究基于 CHARLS 数据，结合相关性分析与逻辑回归模型探讨老年人糖尿病的影响因素。结果显示，吸烟和年龄增长显著增加患病风险，而男性、南方居住、有工资收入等具有保护作用；运动与配偶状况影响微弱，饮酒呈弱负相关，与部分既有研究存在差异，可能源于数据不平衡或样本偏差。尽管过采样优化提升了少数类识别能力，但模型整体区分度仍较低（AUC=0.53），提示需进一步优化特征选择与建模方法。研究创新在于引入南北方地域差异分析，弥补了现有文献的不足。然而，受限于数据不平衡及变量广度，结论存在一定局限。未来可构建更全面问卷，纳入更多临床与行为变量，并尝试随机森林、神经网络等模型进行比较，以提升预测性能。本研究为糖尿病风险因素识别提供了实证参考，也为个性化健康管理策略的制定奠定了基础。

参考文献：

曾刚. 糖尿病一体化管理模式的需求评估研究[D].四川大学，2005.
王富军，丁海霞. 《中国老年 2 型糖尿病胰岛素抵抗诊疗专家共识(2022 版)》解读[J].河北医科大学学报，2024，45(11)：1241-1246.
中国 2 型糖尿病防治指南(2017 年版)[J].中国实用内科杂志，2018，38(04)：292-344.
杨敏，柳洁. 中国糖尿病防治现状[J].中国医学创新，2014，11(07)：149-151.

王金虹，张晓薇，马斌. 逻辑回归与关联分析膳食习惯对慢性代谢疾病的影响[J].电子技术与软件工程，2021(20)：175-178.
王朝霞, 杨慧平, 苏伟, 等. 空腹血糖筛查糖尿病高危人群的研究[J/OL].临床医学研究与实践, 2019, 4(8)：103-105.
王旭. 面向糖尿病就诊与随访信息的数据挖掘与分析[D].上海第二工业大学. 2022.
Asadi F, Fallahzadeh H,Rahmanian M, Determining the factors related to diabetes type II with mixed logistic regression[J/OL].International journal of epidemiologic research, 2016.
ZHE L. Analysis of influential factors for abnormal glucose metabolism with cumulative odds logistic regression[J/OL].Chinese Journal of Public Health, 2010.
李婷，孙媛媛，李雪玲，等. 基于机器学习分类算法的糖尿病辅助诊断研究[J/OL].电脑知识与技术，2024，20(10)：27-29.
徐昌成. 成员合作关系的多样性与网络结构对知识创造的影响研究[D].华中科技大学，2011.
马文彬，王克，于滨，等. 基于体检数据的糖尿病风险预测模型对比研究[J/OL].现代信息科技，2020，4(23)：72-75.
郑家浩，王爱民，于滨，冯超南，纪俊. 基于体检数据机器学习分析的糖尿病风险预测模型[J].青岛大学学报(工程技版)，2021，36(02)：36-41.
郝家乐. 我国老年人糖尿病患病情况影响因素分析及危险等级评估[D].首都经济贸易大学，2022.
苏银霞，王燕，王婷婷，马琦，马艳，王志强，姚华. 新疆维吾尔族 2 型糖尿病危险因素 Logistic 回归分析[J].新疆医科大学学报，2014，37(10)：1253-1256+1260.
彭若萱. 延边地区不同民族教师糖代谢异常及其影响因素分析[D].吉林大学，2019.
王海霞. 氯氮平所致精神分裂症患者血糖异常与 miRNAs 的表达变化研究[D].昆明医科大学，2022.

Influencing Factors of Diabetes in the Elderly: Evidence from CHARLS 2020

Shaofeng Wang, Hong Liang

(Neusoft Inistitute Guangdong, Foshan, Guangdong 528225, China)

Abstract: Diabetes often leads to various complications such as cardiovascular diseases and retinopathy, imposing a significant burden on individual health and the healthcare system. To support medical institutions and the public in the scientific prevention and control of diabetes, this study analyzes the influencing factors of diabetes based on CHARLS (2020) data using correlation analysis and logistic regression modeling. The results indicate that smoking and increasing age are positively associated with the risk of diabetes, while being male, residing in rural areas, having a spouse, receiving wages, and living in southern China have protective effects and reduce the likelihood of developing diabetes. Physical activity score and alcohol consumption frequency show minimal impact and are nearly negligible. Due to the substantial imbalance between non-diabetic (class 0) and diabetic (class 1) samples, the model tends to predict the majority class, resulting in an accuracy of only 54%. After optimization, evaluation metrics for class 1 show improvement; however, the ROC AUC remains close to 0.5, indicating limited discriminative ability. Therefore, diabetes prevention should focus on smoking cessation, moderate alcohol consumption, regular physical activity, weight control, balanced diet, and emotional regulation. Effective diabetes management requires a comprehensive approach integrating health education, medical nutrition therapy, exercise intervention, blood glucose monitoring, and pharmacological treatment.

Keywords: Correlation analysis, Logistic regression ; Diabetes