 DOI: https://doi.org/10.12345/yzlcyxzz.v7i11.22416

Analysis of Factors Influencing Workers' Blood Glucose Levels Based on Bayesian Networks

Lei Xu[1] Ling Yan[2]

1. School of Public Health, Xinjiang Second Medical College, Karamay, Xinjiang, 834000, China
2. Physical Examination Center, Karamay Central Hospital, Karamay, Xinjiang, 834000, China

Abstract

Objective: Bayesian network combined with logistic regression was used to explore the influencing factors and their relationship between workers' blood lipid levels. **Methods:** Using the occupational health database, 884 oil workers in Xinjiang were taken as the research objects, the general demographic characteristics and physical examination data of the study subjects were collected through occupational health examination, the factors influencing blood glucose level were preliminarily screened by univariate analysis logistic regression, the supervised Bayesian network was established, the importance of the influencing factors of oilfield workers' working ability was ranked by mutual information-MI, and the unsupervised Bayesian network was established. The contraindication algorithm was used to analyse the dependence between the influencing factors of blood glucose level, and the multivariate logistic regression method was used to analyse the influencing factors of blood glucose level. **Results:** The detection rate of hyperglycaemia, hyperlipidaemia and hypertension was 17.18%, 36.49% and 33.65%. The results of Naive Bayes network showed that the importance of blood glucose levels was in the order of BMI, blood lipid level, age, education level, hypertension, smoking, and shift work. The results of unsupervised Bayesian network showed that there was a direct dependence between blood glucose level and hyperlipidaemia and education level. The results of multivariate logistic regression analysis showed that after controlling for confounding factors, the age (35~45 years old group, OR=2.18, 95%*CI* 1.15~4.10; >45 years old group, OR=3.67 *CI* 1.84~7.31), smoking (OR=1.61, 95%*CI* 1.07~2.41), obesity (OR=3.14, 95%*CI* 1.80~5.47), hyperlipidaemia (OR=2.66, 95%*CI* 1.77~4.00) was related to blood glucose level ($P < 0.05$). It was a risk factor for hyperglycaemia level ($P < 0.05$), and education level (OR=0.61, 95%*CI* 0.4~0.94) was a protective factor for hyperglycaemia level ($P < 0.05$). **Conclusion:** Blood glucose level is closely related to age, education level, shift work, smoking, hyperlipidaemia and obesity, and blood glucose level is directly related to hyperlipidaemia and education level.

Keywords

blood glucose levels; influencing factors; Bayesian networks

基于贝叶斯网络的工人血糖水平影响因素分析

徐蕾[1] 闫玲[2]

1. 新疆第二医学院公共卫生学院，中国 · 新疆 克拉玛依 834000
2. 克拉玛依市中心医院体检中心，中国 · 新疆 克拉玛依 834000

摘 要

目的：利用贝叶斯网络结合logistic回归探讨工人血脂水平影响因素及相互关系。**方法：**利用职业健康数据库，以新疆地区884名石油工人为研究对象，通过职业健康体检，收集研究对象的一般人口学特征及体检资料，采用单因素分析logistic回归进行血糖水平影响因素初步筛选，建立监督贝叶斯网络，利用相互信息分析方法（mutual information-MI）对油田工人工作能力影响因素的重要性进行排序，建立无监督贝叶斯网络，使用禁忌算法分析血糖水平影响因素间的依赖关系，采用多因素logistic回归方法分析血糖水平的影响因素。**结果：**工人高血糖检出率为17.18%，高血脂检出率为36.49%，高血压检出率为33.65%。朴素贝叶斯网络结果显示：对血糖水平的重要性排序依次为：BMI、血脂水平、年龄、文化程度、高血压、吸烟、倒班。无监督贝叶斯网络结果显示，血糖水平与高血脂、文化程度之间存在直接依赖关系。多因素logistic回归分析结果发现：在控制了混杂因素后，年龄（35~45岁组，OR=2.18，95%*CI* 1.15~4.10；>45岁组OR=3.67 *CI* 1.84~7.31）、吸烟（OR=1.61，95%*CI* 1.07~2.41）、肥胖（OR=3.14，95%*CI* 1.80~5.47）、高血脂（OR=2.66，95%*CI* 1.77~4.00）与血糖水平均有关（$P<0.05$）。是高血糖水平的危险因素（$P<0.05$），文化程度（OR=0.61，95%*CI* 0.4~0.94）是高血糖水平的保护因素（$P<0.05$）。**结论：**血糖水平与年龄、文化程度、倒班、吸烟、高血脂、肥胖密切相关，血糖水平与高血脂、文化程度存在直接作用关系。

关键词

血糖水平；影响因素；贝叶斯网络

【基金项目】新疆第二医学院青年科研项目基金（项目编号：QK202216）。

【作者简介】徐蕾（1993–），男，中国江苏连云港人，硕士，讲师，从事职业流行病学研究。

1 引言

糖脂代谢紊乱相关疾病是一种以葡萄糖和脂质代谢紊乱为特征，主要表现为、血糖异常、血脂异常、高血压等临床表现的疾病[1]。糖脂代谢相关疾病在人群中发病率通常与

二型糖尿病发病率相似，对中国人群健康和社会经济产生危害[2]。糖脂代谢紊乱相关疾病其影响因素众多，但影响因素之间的相互关系需要通过一些非传统的统计学方法进行研究。贝叶斯网络是一种将变量间关系可视化为交互网络的数据分析方法，每个相互关系的结果是基于统计的条件概率，所有变量均由节点表示，在每个节点上，变量的概率分布由其与父节点的关系定义，父节点是所有传入向量所源自的根节点[3]。变量可以直接作用于结局，也可能通过其他变量对结局产生间接影响，因此贝叶斯网络可以直观地观察到变量之间的相互作用关系，可以弥补单独使用 Logistic 回归分析的不足，并初步探讨血糖水平与各个因素之间的潜在关系。

2 资料与方法

2.1 研究对象

以 2018 年参加职业健康体检的克拉玛依市某油田公司 844 名石油工人为研究对象。通过职业健康体检数据库收集研究对象的人口学资料及生理生化检测结果资料。

2.2 研究方法

2.2.1 指标测定

由克拉玛依市中心医院体检中心的工作人员对研究对象进行测量和测定，内容包括一般情况录入、身高、体重、血压测定、听力测定血常规及大生化指标测定。所有工作人员均经过统一培训并考核合格。

2.2.2 诊断标准

血糖异常的诊断标准依据中国 2 型糖尿病防治指南（2020 年版）[4]：高水平 FPG：FPG ≥ 6.1mmol/L，高血压的诊断标准依据中国高血压防治指南（2018 年修订版）[5]，收缩压（SBP）≥ 140mmHg 和（或）舒张压（DBP）≥ 90mmHg。血脂异常的诊断标准依据参照根据《中国成人血脂异常防治指南》（2016 年修订版）[6]，TG ≥ 1.7mmol/L，低 HDL-C ＜ 0.9mmol/L（男性）或＜ 1.0mmol/L（女性）。听力损失诊断标准：对性别与年龄进行校正后，以任一耳任一高频（3000Hz、4000Hz、6000Hz）的气导听阈值 >25dB 判定为听力损失，以双耳各频率气导听阈值均≤ 25dB 判定为听力正常[7]。

2.2.3 贝叶斯统计原理

贝叶斯网络通过有向无环图（DAG）来表示一组变量及条件依赖关系[8]。本研究利用数据库建立一个无监督贝叶斯网络，无监督学习方法代表了最真实的知识形式，因为对于研究中变量之间的潜在关系的探索不存在任何限制。无监督贝叶斯网络模型的目的是发现血糖水平与各个变量间的相互依赖关系。

本研究建立了一个以血糖水平为目标变量的有监督贝叶斯网络，目的是衡量各个因素对于目标变量的重要性大小[9]。在监督贝叶斯网络中，我们采用朴素贝叶斯算法构建监督贝叶斯网络，从而使互信息更容易计算，互信息是信息论的基础概念，它代表了每个变量本身所持有的信息量，互信息的定义为：目标变量的边际熵与给定目标的条件熵之间的差异正式称为互信息（mutual information-MI）[10]，血糖水平与各个因素间的 MI 值可以确定哪些因素为我们提供了最大的信息增益，即各个因素对工作能力重要性大小。

2.2.4 统计分析方法

采用 SPSS 统计软件（23.0 版本）对数据进行分析。符合正态分布的测量数据用 x ± s 表示，单因素回归分析采用单因素多因素 logistic 回归模型进行分析，多因素回归分析采用非条件 logistic 回归模型进行分析，$P < 0.05$ 为有统计学差异。无监督贝叶斯网络及监督贝叶斯网络的建立均由 beysialab（11.5 版本）软件完成。

3 结果

3.1 研究对象基本情况分析

本次研究纳入的研究对象共 844 人，其中男性占比 59.83%，女性占比 40.17，平均年龄为（40.56 ± 7.316）岁。研究对象的血糖平均水平为 5.25 ± 1.51mmol/L。高血糖人群占比 17.18%，高血脂人群占比 36.49%，患高血压人群占 33.65%，超重人群占比 40.52%，肥胖人群占比 19.91%，患有听力损失的人占 28.55%，见表 1。

3.2 血糖水平单因素 logistic 回归分析

对各项指标进行单因素 logistic 回归分析后，结果显示，年龄、文化程度、倒班、吸烟、超重、肥胖、高血脂与高血脂水平有关（$P < 0.05$），见表 2。

3.3 贝叶斯网络的血糖水平影响因素重要性分析

对血糖水平的影响因素的重要性进行排序，以血糖水平为目标节点的监督朴素贝叶斯模型显示：BMI、血脂水平、年龄、文化程度和高血压是这 5 个因素节点与工作能力的 MI 值最大，即这 5 个影响因素是影响工作能力的主要因素，其余因素对工作能力的影响的相对重要性较小。注释：计算节点间的 MI 可以确定哪些因素为工作能力提供了最大的信息增益，贝叶斯网络图按信息增益的大小由顺时针进行排序，见图 1。

3.4 血糖水平的多因素回归分析

对上述筛选出的对血糖水平影响较大的因素进行多因素 logistic 回归分析，结果显示，年龄（35~45 岁组，OR=2.18，95%*CI* 1.15~4.10；>45 岁组 OR=3.67 *CI* 1.84~7.31），文化程度（OR=0.61，95%*CI* 0.4~0.94），吸烟（OR=1.61，95%*CI* 1.07~2.41），肥胖（OR=3.14，95%*CI* 1.80~5.47），高血脂（OR=2.66，95%*CI* 1.77~4.00）与血糖水平均有关（$P < 0.05$），见表 3。

表 1 调查对象一般情况分析

变量	人数	比例（%）
性别		
男	505	59.83
女	339	40.17
年龄（岁）		
＜ 30	208	24.64
30~45	401	47.51
＞ 45	235	27.85
文化程度		
专科以下	266	31.52
专科及以上	578	68.48
婚姻		
未婚	58	6.87
已婚	704	83.41
其他（离异及丧偶）	82	9.72
职称		
初级及以下	186	22.04
中级	253	29.98
副高及以上	405	47.98
收入（月平均）		
＜ 4000 元	344	40.76
≥ 4000 元	500	59.24
吸烟情况		
吸烟	350	41.47
不吸烟	494	58.53
饮酒		
饮酒情况	324	38.39
不饮酒	520	61.61
倒班情况		
固定白班	457	64.81
轮班	297	36.19
BMI		
正常	334	39.57
超重	342	40.52
肥胖	168	19.91
高血压		
有	284	33.65
无	560	66.35
高血脂		
有	308	36.49
无	536	63.51
高血糖		
有	145	17.18
无	699	82.82
听力损失		
有	241	28.55
无	603	71.45

表 2 单因素 logistic 回归分析血糖水平的影响因素

变量	β	S.E	Z	*P*	OR（95%CI）
性别					
女					1.00（Reference）
男	-0.08	0.19	-0.42	0.677	0.92（0.64~1.33）
年龄分组（岁）					
＜ 35					1.00（Reference）
35~45	1.02	0.31	3.33	＜ .001	2.78（1.52~5.08）
＞ 45	1.65	0.31	5.26	＜ .001	5.19（2.81~9.58）
文化程度					
专科以下					1.00（Reference）
专科及以上	-0.94	0.19	-5.06	＜ .001	0.39（0.27~0.56）
倒班					
固定白班					1.00（Reference）
轮班	0.56	0.19	3.03	0.002	1.75（1.22~2.52）
职称					
初级及以下					1.00（Reference）
中级	0.17	0.26	0.65	0.514	1.19（0.71~1.99）
副高及以上	0.22	0.24	0.89	0.373	1.24（0.77~2.00）
婚姻					
未婚					1.00（Reference）
已婚	0.25	0.39	0.64	0.525	1.28（0.59~2.78）
离异及丧偶	0.49	0.47	1.05	0.294	1.63（0.65~4.09）
收入					
＜ 4000 元					1.00（Reference）
≥ 4000 元	-0.27	0.18	-1.46	0.143	0.76（0.53~1.10）
吸烟					
不吸烟					1.00（Reference）
吸烟	0.81	0.19	4.35	＜ .001	2.24（1.56~3.22）
饮酒					
不饮酒					1.00（Reference）
饮酒	0.31	0.19	1.62	0.105	1.37（0.94~2.00）
BMI 分组					
正常					1.00（Reference）
超重	1.02	0.25	4.15	＜ .001	2.78（1.72~4.51）
肥胖	1.72	0.26	6.57	＜ .001	5.61（3.35~9.39）
高血压					
无					1.00（Reference）
有	0.92	0.19	4.96	＜ .001	2.51（1.75~3.61）
高血脂					
无					1.00（Reference）
有	1.32	0.19	6.94	＜ .001	3.74（2.58~5.43）
听阈提高					
无					1.00（Reference）
有	0.34	0.19	1.73	0.083	1.40（0.96~2.05）

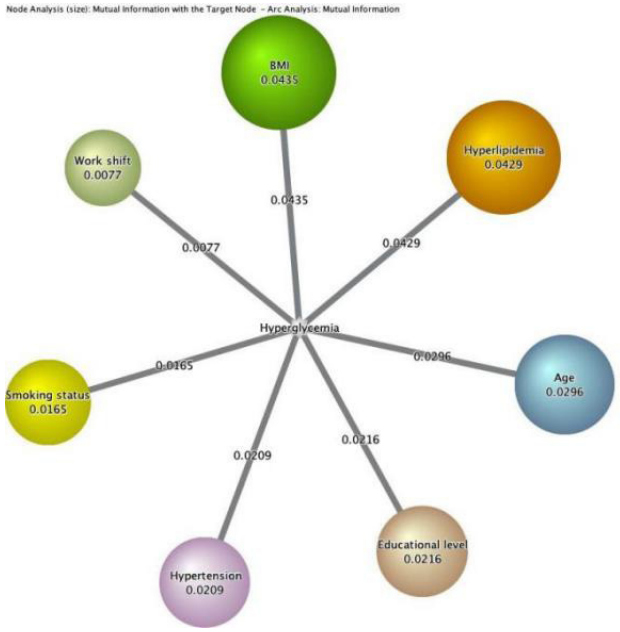

图 1 监督贝叶斯网络的影响因素重要性排序分析

注释：Hyperlipidemia 高血脂，Hyperglycemia 高血糖，Hypertension 高血压，Age 年龄，Educational level 文化程度，Work shift 轮班情况，Smoking status 吸烟情况。

表 3 多因素 logistic 回归分析血糖水平影响因素

Variables	β	S.E	Z	*P*	OR（95%CI）
年龄分组（岁）					
＜ 35					1.00（Reference）
35~45	0.78	0.32	2.40	0.016	2.18（1.15~4.10）
＞ 45	1.30	0.35	3.69	＜.001	3.67（1.84~7.31）
文化程度					
专科以下					1.00（Reference）
专科及以上	-0.49	0.22	-2.22	0.026	0.61（0.40~0.94）
吸烟					
不吸烟					1.00（Reference）
吸烟	0.48	0.21	2.30	0.021	1.61（1.07~2.41）
BMI 分组					
正常					1.00（Reference）
超重	0.45	0.26	1.70	0.089	1.57（0.93~2.63）
肥胖	1.14	0.28	4.04	＜.001	3.14（1.80~5.47）
高血脂					
无					1.00（Reference）
有	0.98	0.21	4.70	＜.001	2.66（1.77~4.00）
常量项	-3.11	0.43	-7.30	＜.001	0.04（0.02~0.10）

3.5 血糖水平直接影响因素与间接影响因素的贝叶斯网络分析

利用无监督贝叶斯网络建立变量之间的相互关系，图 2 中血糖水平的父节点有 2 个，分别是血脂水平、文化程度，说明工作能力与这 2 个变量存在直接依赖关系，而 BMI、年龄、高血压与血糖水平无直接关系，可能是通过血脂水平以及文化程度间接对血糖水平发生作用。

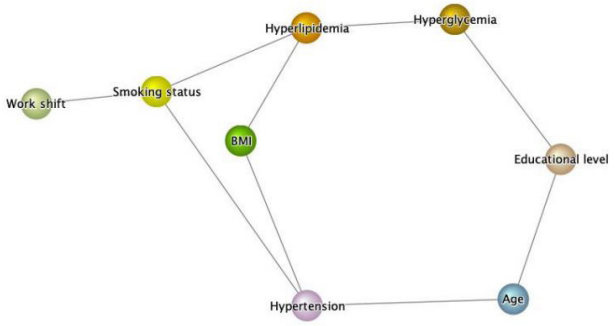

图 2 无监督贝叶斯网络的影响因素相互依赖关系分析

注释：Hyperlipidemia 高血脂，Hyperglycemia 高血糖，Hypertension 高血压，Age 年龄，Educational level 文化程度，Work shift 轮班情况，Smoking status 吸烟情况。

4 讨论

代谢性紊乱相关疾病是职业人群常见疾病。丁洋等人 [11] 对鞍钢钢铁工人重点慢性病患病现状调查的结果显示，2020 年鞍钢工人高血压、血脂异常、高血糖检出率分别为 41.3%、48.4% 和 10.7%，超重、肥胖率为 45.8% 和 18.1%。本研究的结果显示，工人高血糖检出率为 17.18%，高血脂检出率为 36.49%，高血压检出率为 33.65%。由此可见，代谢性紊乱相关疾病在职业人群中高发。职业人群的身心健康直接影响到产业的生产安全和经济效益。因此，健康体检及慢性病管理对预防和控制慢性病的发生至关重要。

职业人群健康体检每年产生大量体检数据，但数据难以得到很好的利用。贝叶斯网络模型有利于揭示血糖水平相关影响因素间的网络关系。本研究利用 logisti 回归结合贝叶斯网络将因素之间的关系可视化再对血糖水平影响因素进行分析。血糖水平受到多种因素的影响，本次研究发现，年龄是血糖水平异常的危险因素，这与吕芳等人 [12] 的研究结果一致。这也符合随年龄增长血糖值升高的全人群流行规律。其次，文化程度是血糖水平异常的保护因素，即文化程度高的人血糖水平异常发生风险较低，并且通过贝叶斯网络发现文化程度与血糖水平之间具有直接关系。其原因主要在于，文化程度影响人群对慢性病知晓度、关注度、早筛率和干预率，低文化程度的人群由于其低知晓、低早筛导致慢性病发病率升高 [13]。由于文化程度与血糖水平直接相关，所以应加强低文化程度人群对慢性病知晓率和筛查率，做到早预防、早干预。此外，吸烟也是血糖水平异常的危险因素。目前，主动吸烟及被动吸烟会增加 2 型糖尿病患病风险均已经得到证实 [14]，原因可能在于，烟草中的尼古丁会损伤胰岛细胞功能，直接改变葡萄糖稳态 [15]。

本研究还发现肥胖、高血脂是血糖水平异常的危险因素，这与顾晓美等人 [16] 的研究结论一致。本研究的无监督贝叶斯网络发现，高血脂可以直接影响血糖水平，原因在于脂肪组织释放的脂肪因子与糖脂代谢调节、胰岛素抵抗等密

切相关[17]。此外，肥胖可以通过影响血脂水平间接影响血糖水平。通过贝叶斯网络，本研究对影响血糖水平的因素进行了重要性排序，监督贝叶斯网络结果显示，BMI、血脂水平、年龄、文化程度和高血压对血糖水平的影响较为重要，提示职业人群代谢性疾病的预防控制和管理应该着重从这些相对重要的因素入手，特别是容易改变的影响因素。

综上，本研究利用贝叶斯网络结合 logistics 回归方法对工人血糖水平影响因素进行了初步研究，后续仍然需要对贝叶斯网络在职业人群健康方面的应用进行深入研究，为今后职业人群疾病防控提供科学依据。

参考文献

[1] 杜文静,顾浩琰,吴珊.噪声对糖脂代谢影响的研究进展[J].环境与职业医学,2023,40(10):1212-1217.

[2] 曹歆祎.代谢性疾病相关综述[C]//中国营养学会第十五届全国营养科学大会论文汇编,东南大学,2022:1.

[3] Gerassis S, Abad A, Saavedra Á, et al. Women's Occupational Health: Improving Medical Protocols with Artificial Intelligence Solutions[C]//Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2018: 1193-1199.

[4] 中华医学会糖尿病学分会.中国2型糖尿病防治指南(2020年版)[J].中华糖尿病杂志,2021,13(4):315-409.

[5] 中国高血压防治指南(2018年修订版)[J].中国心血管杂志,2019,24(1):24-56.

[6] 诸骏仁,高润霖,赵水平,等.中国成人血脂异常防治指南(2016年修订版)[J].中国循环杂志,2016,31(10):937-953.

[7] 姚玲莉,焦洁,张磊,等.男性噪声作业工人血糖水平与职业性噪声性听力损失的关系研究[J].医药论坛杂志,2024,45(5):502-506.

[8] An H, Xu L, Liu Y, et al. Study on a Bayes evaluation of the working ability of petroleum workers in the Karamay region, Xinjiang, China[J]. Frontiers in Psychology, 2022, 13: 1011137.

[9] Conrady S, Jouffe L. Introduction to bayesian networks & bayesialab[J]. Bayesia SAS, 2013.

[10] Conrady S, Jouffe L. Bayesian networks and BayesiaLab: a practical introduction for researchers[M]. Bayesia USA, 2015.

[11] 丁洋,李冰哲,鄢上书,等.2020年鞍钢钢铁工人重点慢性病患病现状调查[J].职业卫生与病伤,2022,37(3):133-137.

[12] 吕芳,郭雯婕,张延玲,等.2015—2019年某铁路局职工空腹血糖水平分析[J].环境卫生学杂志,2024,14(6):534-538.

[13] 刘梦冉,焦莹莹,张思婷,等.2018年中国四省育龄女性心血管代谢性危险因素流行特征[J].卫生研究,2024,53(1):1-7+29.

[14] Aulinas A, Colom C, García Patterson A, et al. Smoking affects the oral glucose tolerance test profile and the relationship between glucose and HbA1c in gestational diabetes mellitus[J]. Diabetic Medicine, 2016,33(9):1240-1244.

[15] 文茂琼.被动吸烟暴露情况对孕妇妊娠期糖尿病发病的临床影响分析[J].医学理论与实践,2024,37(15):2668-2670.

[16] 顾晓美,覃玉,陈路路,等.江苏省老年人群代谢综合征的流行情况及其与肥胖的关联[J].江苏预防医学,2022,33(3):260-264.

[17] 宁冬平,朱惠娟,阳洪波,等.脂肪因子与肥胖及代谢综合征的相关研究进展[J].医学综述,2018,24(18):3653-3657.