数据挖掘结课项目终期报告

糖尿病肾病患者生化检验结果分析和发病率预测

组别：11 组
组员：王永昊
董国昭
董小楠
李若松
雷丙震
学院：生命学院
专业：生物医学工程

目录

摘　要：糖尿病肾病作为一种糖尿病并发症严重危害着患者的健康。本课题针对2009-2011 年中国人民解放军总医院住院收治的糖尿病患者的生化检验数据，应用 McNemar 检验、Logistic 回归算法和 Apriori 关联算法分析糖尿病肾病的发病情况及其与各项生化检验结果之间的关系，并对通过 Logistic 回归算法建立的模型进行了评价。结果发现肌酐、尿素、葡萄糖、钠、钙、氯化物和糖尿病肾病发病的关系有统计学意义，所建立的糖尿病肾病发病率和这六种生化指标关系的评分模型的正确率与特异度较高，可以用于糖尿病患者群体的糖尿病肾病筛查和诊断参考。

关键词：糖尿病肾病；生化检测；Logistic 回归算法；Apriori 关联算法；

项目分工：

李若松：数据初步筛选及二值化处理；
董国昭、雷丙震：基于生化检测的糖尿病肾病发病预测模型（Logistic 回归算法）；
王永昊、董小楠：糖尿病患者生化指标与糖尿病肾病的关联性分析（Apriori 关联算法）；
全体组员：整合处理结果，总结结论，撰写课程报告。

数据挖掘思路：

1. **数据收集及参数整理。**根据文献及数据中心的数据进行整理，梳理出所需要的有关糖尿病患者的数据。
2. **糖尿病肾并发症的生化指标分析和筛选。**包含数据筛选、二值化处理和指标分析。
3. **基于生化检测的糖尿病肾病发病预测模型。**包含 Logistic 回归评分模型构建以及 Logistic 回归评分模型的结果评估。
4. **糖尿病患者生化指标与糖尿病肾病的关联性分析。**包含生化常规检查中的一些指标与糖尿病肾病之间的关联关系以及生化指标之间的关联关系。

1 引言

近年来，随着信息技术尤其是互联网的不断发展，医疗卫生领域的大数据分析与知识挖掘逐渐引起学术界和业界的重视[1]。越来越多的医院采用 HIS 信息系统，有研究表明，未来 10 年医疗数据将呈现高爆炸式增长[2]，其中所蕴含的各种信息值得我们深入挖掘。如果这些医疗信息在保证安全的前提下被正确使用，它将会给现代医疗水平带来很大的提高[3]。医疗卫生领域的大数据分析与挖掘还是一个比较年轻的学科[4] ，将会有很大的发展。

根据 2013 年中国卫生统计年检[5],中国糖尿病的发病率和死亡率分别达到了 2.04%和 0.38%。糖尿病（diabetes mellitus，DM）已经成为了当下中国一种主要的公众健康问题[6]，糖尿病及其多种并发症已经严重损害了人们的健康。糖尿病肾病(Diabetic Nephropathy, DN)是糖尿病最主要的慢性并发症之一，由糖尿病微血管病变引起。糖尿病肾病发展到晚期会出现严重的肾功能衰竭，是糖尿病患者的主要死因之一[7]。根据美国 2011 年报道，约 30%Ⅰ型糖尿病和 20%Ⅱ型糖尿病最终会引起糖尿病肾病, 其中最终死于肾功能衰竭者约占 50%，是慢性肾脏疾病的主要死因[8]。

由于肾病的复杂性，目前，肾穿刺活检和核医学手段是临床准确诊断肾病的主要方法。但是肾穿刺活检是一种创伤性的检查，可能出现出血、血肿、动静脉瘘、 高血压、肾脏感染等多种并发症，甚至导致死亡[9]。

本文从患者生化指标检测结果入手，通过条件 Logistic 回归算法和 McNemar 检验筛选出糖尿病肾病发病人群的生化指标与没有肾脏疾病的糖尿病患者的主要区别，并通过 logistic 回归算法建立回归方程，用生化检测的结果预估糖尿病肾病的发病情况，再以 Apriori 关联算法分析糖尿病肾病的发病情况以及各个生化检验指标之间的关系，以辅助医生分析患者的情况及预后措施。

本文数据来自临床数据中心（中国人民解放军总医院，简称 301 医院），内容为 2009、2010、2011 三年中国人民解放军总医院住院收治的全部糖尿病患者的生化检测结果。

2 糖尿病肾并发症的生化指标分析和筛选

2.1 研究的方法

2.1.1 数据集描述

从临床数据中心（301 医院）提取 2009、2010、2011 年 301 医院住院收治的糖尿病患者的生化检验结果，从中随机抽取 1039 例糖尿病肾病患者和 1039 例临床诊断未患有肾病的糖尿病患者，作为分析使用的对照样本。

本研究涉及患者检测的生化检验指标共 12 项，包括丙氨酸氨基转移酶、天冬氨酸氨基转移酶、尿素、γ

—谷氨酰基转移酶、肌酐、葡萄糖、血清尿酸、肌酸激酶、钙、钠、钾、氯化物。判定其指标异常的标准如下[10]：

丙氨酸氨基转移酶：检验值大于 40U/L；

天冬氨酸氨基转移酶：检验值大于 40U/L；

尿素：检验值大于 7.5mmol/L 或小于 1.8mmol/L；

γ—谷氨酰基转移酶：检验值大于 50U/L；

肌酐：检验值大于 110umol/L 或小于 30umol/L；

葡萄糖：检验值大于 6.1mmol/L 或小于 3.4mmol/L；
血清尿酸：检验值大于 444umol/L；
肌酸激酶：检验值大于 200U/L；
钙：检验值大于 2.54mmol/L 或小于 2.094mmol/L；
钠：检验值大于 130mmol/L 或小于 150mmol/L；
钾：检验值大于 3.5mmol/L 或小于 5.5mmol/L；
氯化物：检验值大于 94mmol/L 或小于 110mmol/L；

2.1.2 生化指标的分析方法

对 12 个生化指标分别进行描述性分析和条件 Logistic 回归分析，以此作为初次筛选。

描述性分析：首先，对被随机抽取的糖尿病肾病患者和未患有肾病的糖尿病患者各 1039 例中各个生化指标的发生概率进行分析，针对糖尿病肾病和非糖尿病肾病患者人群在这 12 个指标的差异进行分析。在数据集中，所有生化检验指标变量均以正常和非正常划分为二值水平，对其进行 McNemar 检验，用以观察患病人群和未患病人群危险因素的差别是否存在统计学意义。

条件 Logistic 回归分析：以患有糖尿病肾病和患有糖尿病但未见肾脏异常作为因变量（糖尿病肾病=1，未患糖尿病肾病=0），随机抽取成对的样本，所有 12 项生化指标作为自变量（正常=0，不正常=1），以 P<0.05 作为统计学显著性界限，拟合生成条件 Logistic 回归模型。对这 12 项生化指标进行多因素 Logistic 回归分析，筛选出每个对糖尿病肾病发病有统计学意义的指标。然后，将所有对糖尿病肾病发病有统计学意义的指标单独抽出，作为自变量，以 P<0.05 作为统计学显著性界限，拟合生成条件 Logistic 回归模型。其结果可以用于分析影响人群糖尿病肾病发病的生化指标。最终用 OR 值及其 95%CI 置信度衡量生化指标与糖尿病肾病发病的关联强度。OR 值的取值范围是从 0 到无限大的正数。OR>1，表示该生化指标异常与疾病为正相关，即生化指标指示疾病发生的概率增加，OR 值越大说明生化指标异常对疾病的正相关作用越大；OR<1，表示该生化指标异常与疾病为负相关，即该生化指标异常使疾病的发生下降；如果 OR 接近或等于 1，说明生化指标与所研究的疾病无关。一般 95%CI 置信度上限小于 1 时说明可能是负相关因素，相反如果下限大于 1 则说明可能是正相关因素。

统计分析应用 SPSS 软件 20.0 版本，检验水平设定为 0.05。

2.2 结果分析

2.2.1 初次筛选

被随机抽取的糖尿病肾病患者和糖尿病非肾病患者各 1039 例的条件 Logistic 回归分析的结果如表 1 所示。描述性分析的结果如表 2 所示，

在糖尿病肾病患者和糖尿病非肾病患者调查个体数相等的条件下，条件 Logistic 回归分析的结果显示，尿素、肌酐、葡萄糖、钙、钠、氯化物 6 项生化指标 P 值小于 0.05，在人群间的差异存在统计学意义。其中尿素、肌酐钙、钠、氯化物 OR 值均大于 1，葡萄糖 OR 值小于 1。

描述性分析得出结果与条件 Logistic 回归分析基本一致。在糖尿病肾病患者和糖尿病非肾病患者调查个体数相等的条件下，统计在危险因素（是=1）前提下个体患糖尿病肾病的概率。McNemar 检验结果显示，尿素、肌酐、、钙、钠、氯化物 5 项生化指标的 OR 值大于 1，葡萄糖 OR 值小于 1，与条件 Logistic 回归分析得出的结果一致。统计的结果为：

尿素不正常在糖尿病肾病患者中的发生概率为 80.08%，尿素正常在糖尿病肾病患者中的发生概率为 19.92%；

肌酐不正常在糖尿病肾病患者中的发生概率为 75.26%，肌酐正常在糖尿病肾病患者中的发生概率为 24.74%；

葡萄糖不正常在糖尿病肾病患者中的发生概率为 64.00%，葡萄糖正常在糖尿病肾病患者中的发生概率为 36.00%；

钙不正常在糖尿病肾病患者中的发生概率为 49.09%，钙正常在糖尿病肾病患者中的发生概率为 50.91%；

钠不正常在糖尿病肾病患者中的发生概率为 27.82%，钠正常在糖尿病肾病患者中的发生概率为 72.18%；

氯化物不正常在糖尿病肾病患者中的发生概率为 11.65%，氯化物正常在糖尿病肾病患者中的发生概率为 88.35%。

表 1. 以 12 种生化指标构建的条件 Logistic 回归分析

Table 1. Conditional Logistic regression of 12 biochemical indices.

生化指标	B	S.E.	P	OR 值	OR 值的 95%C.I	
					下限	上限
丙氨酸氨基转移酶	-0.218	0.254	0.392	0.804	0.489	1.322
天冬氨酸氨基转移酶	-0.328	0.279	0.240	0.720	0.417	1.245
尿素	0.856	0.178	0.000	2.353	1.661	3.332
γ—谷氨酰基转移酶	0.143	0.168	0.393	1.154	0.830	1.604
肌酐	2.128	0.190	0.000	8.401	5.794	12.183
葡萄糖	-0.410	0.155	0.008	0.664	0.489	0.900
血清尿酸	0.406	0.196	0.380	1.501	1.023	2.202
肌酸激酶	-0.189	0.213	0.375	0.828	0.545	1.257
钙	0.327	0.153	0.033	1.386	1.027	1.871
钠	0.540	0.205	0.008	1.717	1.148	2.566
钾	0.131	0.201	0.514	1.140	0.768	1.692
氯化物	0.376	0.317	0.006	1.456	1.289	2.470

表 2. 以 12 种生化指标构建的描述性分析

Table 2. Descriptive analysis of 12 biochemical indices

生化指标		未患病（人）	指标正常/不正常占未患病%	患病（人）	指标正常/不正常占患病%	OR 值	95%C.I.	
							下限	上限
丙氨酸氨基转移酶	正常	899	86.53%	950	91.43%	0.6	0.454	0.797
	不正常	140	13.47%	89	8.57%			
天冬氨酸氨基转移酶	正常	951	91.53%	957	92.11%	0.93	0.676	1.267
	不正常	88	8.47%	82	7.89%			
尿素	正常	686	66.03%	207	19.92%	7.81	6.401	9.532
	不正常	353	33.97%	832	80.08%			
γ—谷氨酰基转移酶	正常	790	76.03%	247	50.00%	0.9	0.809	1.121
	不正常	249	23.97%	247	50.00%			
肌酐	正常	896	86.24%	257	24.74%	19.1	15.21	23.9
	不正常	143	13.76%	782	75.26%			
葡萄糖	正常	270	25.99%	374	36.00%	0.62	0.517	0.753
	不正常	769	74.01%	665	64.00%			
血清尿酸	正常	797	76.71%	656	63.14%	1.92	1.588	2.328
	不正常	242	23.29%	383	36.86%			
肌酸激酶	正常	941	90.57%	864	83.16%	1.95	1.494	2.532
	不正常	98	9.43%	175	16.84%			
钙	正常	788	75.84%	529	50.91%	3.03	2.51	3.649
	不正常	251	24.16%	510	49.09%			
钠	正常	921	88.64%	750	72.18%	1.74	1.354	2.224
	不正常	118	11.36%	289	27.82%			
钾	正常	919	88.45%	853	82.10%	1.67	1.304	2.139
	不正常	120	11.55%	186	17.90%			
氯化物	正常	1020	98.17%	918	88.35%	6.9	4.216	11.29
	不正常	19	1.83%	121	11.65%			

综合描述性分析与条件 Logistic 回归分析得出的结果进行研究。从 OR 值和 95%CI 置信区间的角度考虑，6 个生化指标中葡萄糖 OR 值小于 1，尿素、肌酐、钙、钠、氯化物五个均大于 1，说明尿素、肌酐钙、钠、氯化物五个生化指标不正常都能指示糖尿病肾病发生的概率增加；血糖不正常指示了糖尿病肾病的发生的概率降低。OR 值越大，则该生化指标正常或不正常对糖尿病肾病发生的正相关或负相关性越强，即该因素的影响作用越大。根据危险因素的 OR 值给出 5 个正相关的生化指标的影响作用从强到弱依次为：肌酐、尿素、钠、氯化物、钙。

2.2.1 生化指标的再分析

为了验证在初筛中得出的结果，从 2009、2010、2011 年 301 医院住院收治的糖尿病患者中重新抽取 1039 例糖尿病肾病患者和 1039 例临床诊断未患有肾病的糖尿病患者，作为分析使用的对照样本。将肌酐、尿素、氯化物、钠、钙、葡萄糖 6 项糖尿病肾病发病有统计学意义的指标单独抽出，作为自变量，以 P<0.05 作为统计学显著性界限，拟合生成条件 Logistic 回归模型，结果如表 3 所示。

表 3. 以 6 种生化指标构建的条件 Logistic 回归模型

Table 3. Conditional Logistic regression models of 6 biochemical indices.

生化指标	B	S.E.	P	OR 值	OR 值的 95%C.I	
					下限	上限
尿素	0.941	0.176	0.000	2.318	1.641	2.903
肌酐	2.365	0.150	0.000	10.647	7.940	14.276
葡萄糖	-0.417	0.154	0.007	0.659	0.487	0.653
钙	0.306	0.148	0.008	1.350	1.017	2.469
钠	0.568	0.203	0.005	1.764	1.186	2.624
氯化物	0.296	0.315	0.007	1.434	1.070	1.960

分析结果与从初次筛选中得到的结果基本相符。尿素、肌酐钙、钠、氯化物五个生化指标为糖尿病肾病发生的正相关指标；血糖为糖尿病肾病的发生的负相关指标。根据危险因素的 OR 值给出 5 个正相关的生化指标的影响作用从强到弱依次为：肌酐、尿素、钠、钙、氯化物，与初次筛选结果相同，验证了提出的结论。

3 基于生化检测的糖尿病肾病发病预测模型

3.1 数据来源

预测模型的数据来自临床数据中心（301 医院），内容为 301 医院 2009-2011 三年住院收治的全部糖尿病患者中生化检测结果中肌酐、尿素、钠、钙、氯化物、葡萄糖 6 种指标齐全患者的生化检验结果。共 58122 人，其中患糖尿病肾病总计 1845 人，患病率为 3.17%。对研究对象经采取随机抽样的方法，将其分为训练集和测试集，训练集共 43592 人，其中糖尿病肾病患病 1377 例，患病率 3.16%，测试集共 14350 人，其中糖尿病肾病患病 468 人，患病率 3.20%。

3.2 Logistic 回归评分模型

将训练集人群是否患有糖尿病作为因变量（1=患病，0=未患病），将肌酐、尿素、氯化物、钠、钙、葡萄糖 6 项对糖尿病肾病发病有统计学意义的指标作为自变量（1=不正常，0=正常），以 P<0.05 作为统计学显著性界限，拟合生成 Logistic 回归模型。以 Logistic 回归模型的得分 F 表示个体患糖尿病肾病的风险程度。个体的得分 F 用式（1）计算，个体的得分越高，其患糖尿病肾病的风险程度越高。

$$F = \frac{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}{1+e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \qquad (1)$$

F 代表个体糖尿病肾病患病风险的得分，Xi 代表各危险因素（生化指标）的取值，βi代表危险因素的回归系数，α代表回归方程的常数项。

使用训练集人群，建立 Logistic 回归模型见表 4。由此建立模型的方程见式（2）。

F 越大，代表患病的风险越大。

将模型应用于训练集人群，计算出训练集人群每一个人的个体得分。计算后训练集人群的得分区间为（0，0.423）。首先以 0.04 为间隔，将得分分成 11 级。然后，计算出每个分级中的糖尿病肾病实际患病概率。最后，对得分与实际概率进行函数拟合，拟合的函数见式（3）。在拟合的过程中，实际共有 8 个数据点被采纳，（0.24,0.28）、（0.32,0.36）、（0.36,0.40）共 3 个分组由于数据缺失被略去。如图 1 所示。随着 Logistic 回归模型分值的增大，患病概率呈现升高的趋势。

表 4. 以 6 种生化指标构建的 Logistic 回归模型

Table4. Logistic regression models of 6 biochemical indices

变量	生化指标	B	S.E.	P	OR 值	OR 值的 95%C.I	
						下限	上限
X_1	尿素	0.838	0.075	0.000	2.312	1.997	2.677
X_2	肌酐	2.254	0.070	0.000	9.523	8.303	10.922
X_3	葡萄糖	-0.436	0.056	0.000	0.646	0.579	0.722
X_4	钙	0.395	0.055	0.000	1.485	1.334	1.653
X_5	钠	0.400	0.088	0.000	1.491	1.254	1.773
X_6	氯化物	0.388	0.069	0.000	1.475	1.287	1.689
	常量	-4.534	0.068	0.000	0.011		

$$F = \frac{e^{(-4.531+0.849X_1+2.276X_2-0.431X_3+0.394X_4+0.412X_5+0.850X_6)}}{1+e^{(-4.531+0.849X_1+2.276X_2-0.431X_3+0.394X_4+0.412X_5+0.850X_6)}} \qquad (2)$$

$$y = 31.374x^4 - 13.577x^3 + 1.039x^2 + 1.120x - 0.016 \qquad (3)$$

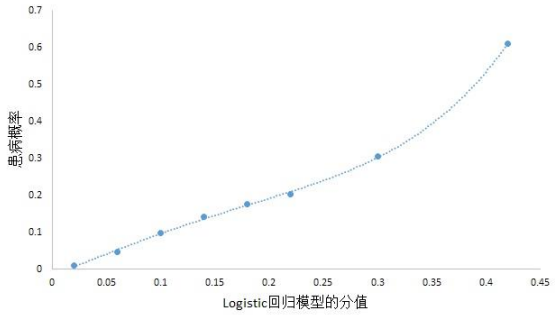

图 1. Logistic 回归模型患病概率预测拟合曲线

Figure 1. Morbidity fitting curve of Logistic regression model

3.3 Logistic 回归评分模型的结果评估

3.3.1 模型的结果对比

评价模型优劣最直接的方式就是看模型对糖尿病肾病患者和非糖尿病肾病患者的区分能力，即本模型的得分在两种群体的差异。将已建立 Logistic 回归评分模型应用到测试集当中，计算出测试集每一个个体在 Logistic 回归评分模型下的个体得分，即式（3）中的 F，比较糖尿病肾病患者和非糖尿病肾病患者的得分情况。结果见表 5。结果显示，糖尿病肾病患者在 Logistic 回归评分模型中的得分明显高于非糖尿病肾病患者。经过卡方检验，两者的得分差异具有统计学意义。

表 5. 测试集 Logistic 回归评分模型分值的比较

Table 5. Comparison of test's Logistic regression model scores

类别		糖尿病肾病患者	糖尿病非肾病患者
不同分组的得分情况	均值	0.133	0.028
	标准差	0.094	0.051
	最大值	0.424	0.424
	最小值	0.005	0.005
显著性检验	卡方值	14354	
	P	0.00	

3.3.2 预测效果的评价方法

针对 Logistic 回归评分模型的评价需要使用以下参数：

TP：真阳性，即患者被正确判断出患病的人数；

TN：真阴性，即非患者被正确判断出未患病的人数；

FP：假阳性，即非患者被错误判断患病的人数；

FN：假阴性，即患者被错误判断未患病的人数。

关于以上参数的说明见表 6。

表 6. 评价方法参数说明

Table 6. parameters description of evaluation methodology

		预测类别	
		阴性	阳性
实际类别	阴性	TN	FP
	阳性	FN	TP

本文使用的评价方法包括：

1.正确率（Accuracy，以下简称 A）：样本中所有个体被正确判断患病或未患病的可能性，见式（4）；

2.灵敏度（Sensitivity，以下简称 Se）：糖尿病肾病患者被正确判定为患者的可能性，见式（5）；

3.特异度（Specificity，以下简称 Sp）：非糖尿病肾病患者被正确判定为非糖尿病肾病患者的可能性，见式（6）；

4.约登指数：灵敏度+特异度-1；

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Se = \frac{TP}{TP+FN} \quad (5)$$

$$Sp = \frac{TN}{FP+TN} \quad (6)$$

使用累积分布曲线和约登指数判断本模型得分的临界阈值。

3.3.3 累积分布曲线

累积分布曲线可以表明在本模型下糖尿病肾病患者和糖尿病非肾病患者的分布规律，还可以直观的确定临界阈值。图 2 为本模型的累积分布曲线，可以发现患者的曲线与非患者曲线在 0.05 附近相交。

3.3.4 预测效果的评价

表 7 列出了本模型得分在不同临界分值下的预测结果。结果显示，当临界分值为 0.03 时，约登指数最大，为 0.737，故视 0.03 为临界阈值，即视分值大于 0.03 为患病，小于 0.03 为未患病。

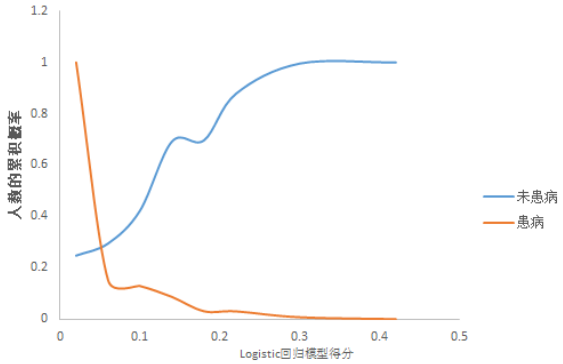

图 2. Logistic 回归模型累积分布曲线

Figure2. Cumulative distribution curve of Logistic regression model

表 7. Logistic 回归评分模型测试集预测效果

Table 7. Forecast effect of Logistic regression model test set

得分临界阈值	灵敏度	特异度	约登指数
0.02	0.859	0.805	0.664
0.03	0.850	0.896	0.746
0.04	0.752	0.896	0.648
0.05	0.731	0.899	0.630

当 Logistic 回归评分模型的临界阈值为 0.03 时，对模型进行评价。此时模型的预测情况见表 8。

表 8. Logistic 回归评分模型评价交叉表

Table 8. evaluation cross table of Logistic regression model test set

		预测情况	
		未患病	患病
实际情况	未患病	12460	1422
	患病	70	398

经式（4）、（5）、（6）计算，模型的评价结果为：
正确率（Accuracy）： A=89.60%；
灵敏度（Sensitivity）： Se =85.04%；
特异度（Specificity）： Sp=89.76%；
约登指数=0.746.

用于评价 Logistic 回归评分模型的测试集共 14350 人，其中糖尿病肾病患病 468 人，患病率为 3.20%。从评价结果来看，Logistic 回归评分模型的正确率与特异度分别为 89.60%和89.76%，处于比较高的水平，可以用于糖尿病患者群体的糖尿病肾病的筛查工作。

4 糖尿病患者生化指标与糖尿病肾病的关联性分析

4.1 Apriori 关联算法分析

4.1.1 关联模型的构建

以 2010、2011 两年 301 医院住院收治的全部糖尿病患者的生化检测结果为数据来源。调查的生化检测指标包括尿素、肌酐、葡萄糖、钙、钠、氯化物共六项指标。选取了 41850 例数据作为关联分析的训练集，其中包括患糖尿病肾病患者 1676 例，占训练集人群的 4.00%，未患糖尿病肾病的糖尿病患者 40174 例，占训练集人群的 96.00%。

本研究使用了 IBM SPSS Modeler 数据挖掘软件，利用关联分析中的 Apriori 算法构建该关联模型。模型选择“尿素含量是否异常”、“肌酐含量是否异常”、 “葡萄糖含量是否异常”、“钙含量是否异常”、“钠含量是否异常”、 “氯化物含量是否异常”等六个生化指标作为关联模型的前项，选择“是否患有糖尿病肾病”作为后项。设置模型的最大前项数为 6，即寻找 6 个生化指标所有可能出现的组合。

模型执行流程如图 3 所示。

将数据库中的变量类型及名称规范化，并删除模型不需要的多余变量，以此节省内存及运算时间；最终经过模型执行，得到实验结果。

4.1.2 实验结果

为寻找到所有可能的关联规则，本次实验中将最小支持度设置为 0.0%，最小置信度设置为 1.0%。关联模型执行后，共寻找到 63 条关联规则，其中有 14 条规则的置信度大于 25%。

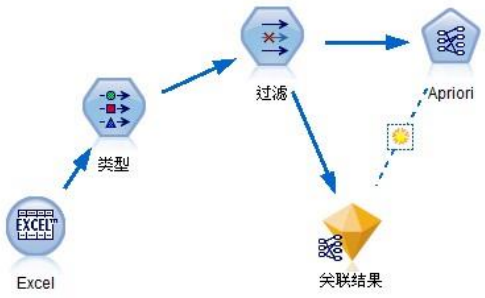

图 3. 生化指标与是否患糖尿病肾病的关联模型

Figure 3. biochemical indices and diabetic nephropathy correlation model.

表 9. 生化指标与是否患糖尿病肾病的关联模型的规则：后项=糖尿病肾病患病

Table 9. The rules of biochemical indices and diabetic nephropathy correlation model: consequent=suffering from Diabetic nephropathy

前项数	置信度>1.0%的频繁集数	置信度>25%的频繁集个数	最大置信度%	最小置信度%
6	1	0	24	24
5	6	3	30.93	14.54
4	15	4	32.77	10.45
3	20	5	29.53	8.68
2	15	2	26.22	5.61
1	6	0	18.15	3.56

发现的 63 条关联规则整理如表 9 所示，其中当前项数为 4 时，置信度最高，最大置信度=32.77%；而当前项数为最大值 6 时，该组规则的最大置信度=24%；当前项为 5 时，最大置信度=30.93%；当前项小于等于 3 时，最大置信度随前项数目减少而变小，分别为 29.53%、26.22%、18.15%。由以上可知，当前项数为 4 和 5 时，关联规则的最大置信度较大。并且以上实验结果证明，糖尿病肾病患病概率并非与生化指标异常个数呈简单线性关系。

4.1.3 寻找有意义的关联规则

在前项个数确定的条件下，由其形成的组合数目往往较大，而我们则需要在众多的规则结果中寻找到有意义的规则，从而才真正符合我们的研究目的。

为了寻找有意义的关联规则，我们将得到的 63 条关联规则，按照前项个数进行分组，共分为 6 个组。置信度的含义为前项发生的条件下，后项发生的概率；因此，我们可以将每组中置信度最大的规则视为有意义的规则，如表 10 所示，共 6 条。这 6 条规则的意义在于，在前项数固定的前提下，出现了这些指标异常的组合较其他组合具有更大的发病可能性。

表 10. 生化指标与是否患糖尿病肾病的关联模型的有意义规则：后项=糖尿病肾病患病

Table 10. The meaningful rules of biochemical indices and diabetic nephropathy correlation model: consequent=suffering from Diabetic nephropathy

序号	前项个数	置信度最大的频繁集	置信度%	支持度%	提升度
1	6	6 个因素	24	0.48	6.00
2	5	钠、氯化物、肌酐、钙、尿素	30.93	0.56	7.72
3	4	钠、肌酐、钙、尿素	32.77	0.85	8.18
4	3	钠、肌酐、钙	29.53	0.96	7.37
5	2	钠、肌酐	26.22	1.77	6.55
6	1	肌酐	18.15	16.47	4.53

由上一节分析可知，指标异常个数的多少并不与患病率呈明确的简单线性相关关系，置信度最高的组合前项数为 4 而并非最大值 6，但是从前项数为 4 到前项数为 1 来看指标异常个数增多，患病率仍会上升。如规则 3—规则 6。规则 3：在具有 4 个生化指标异常的筛查个体中，钠、肌酐、钙、尿素同时异常时患病概率较其他个体最高，为 32.77%。规则 4：在具有 3 个生化指标异常的筛查个体中，钠、肌酐、钙同时异常时患病概率较其他个体最高，为 29.53%。规则 5：在具有 2 个生化指标异常的筛查个体中，钠、肌酐同时异常时患病概率较其他个体最高，为 26.22%。而当前项数为 6 和 5 时则不遵循线性规律，规则 1：在具有 6 个生化指标异常的筛查个体中，钠、氯化物、肌酐、钙、尿素、葡萄糖同时异常时患病概率较其他个体最高，为 32.77%。规则 2：在具有 5 个生化指标异常的筛查个体中，钠、肌酐、钙、尿素同时异常时患病概率较其他个体最高，为 32.77%。

4.2 生化指标间的关联分析

4.2.1 关联模型的构建

关联模型仍然使用与 3.1 中相同的数据集，即选取 41850 例数据作为关联分析的训练集，其中包括患糖尿病肾病患者 1676 例，占训练集人群的 4.00%，未患糖尿病肾病的糖尿病患者 40174 例，占训练集人群的 96%。

基于 SPSS Modeler 挖掘软件，仍使用关联分析中常用的 Apriori 算法构建该关联模型。模型选择前项、后项均为“尿素”、“肌酐”、“葡萄糖”、“钙”、“钠”、“氯化物”等 6 项生化指标。设置模型的最大前项数为 6。

模型执行流程如图 4 所示：

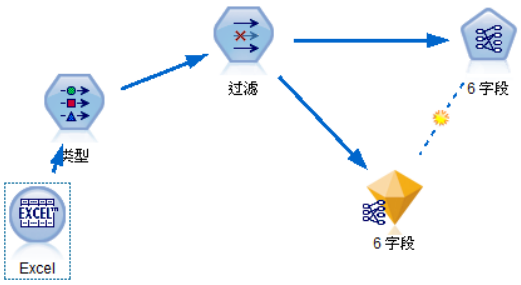

图 4 生化指标之间的关联模型

Figure 4 correlation model of biochemical indices

4.2.2 实验结果

最小支持度设置为 0.0%，最小置信度设置为 1.0%，在执行完关联模型后，共寻找到 186 条关联规则，其中有 111 条规则的置信度大于 50%，且支持度均大于 0.5%。若将最小支持度设置为 0.5%，最小置信度设置为 80%，则共寻找到 31 条关联规则。

4.2.3 寻找有意义的关联规则

在最小支持度为 0.5%，最小置信度为 80%的条件下，模型产生 31 条频繁关联规则。根据后项结果将其分为两类，即后项为尿素，如表 11 所示；后项为葡萄糖，如表 12 所示。

当后项为尿素时，按照置信度从高到低的顺序对以上规则进行排列，如表 11，结果显示，病人生化常规检查中钠、氯化物、肌酐、葡萄糖 4 项指标异常时，同时尿素指标异常的可能性最大，异常概率为 93.60%；概率其次的为氯化物、肌酐、葡萄糖 3 项指标异常，概率为 92.67%。

表 11. 生化指标之间的关联规则 后项=尿素

Table 11. association rules of biochemical indices: consequent=urea

序号	前项	支持度%	置信度%
1	钠、氯化物、肌酐、葡萄糖	0.90	93.60
2	氯化物、肌酐、葡萄糖	2.38	92.67
3	钠、氯化物、肌酐	1.11	91.83
4	钠、氯化物、肌酐、钙、葡萄糖	0.52	91.74
5	氯化物、肌酐、钙、葡萄糖	1.30	91.56
6	钠、肌酐、葡萄糖	1.38	90.83
7	钠、氯化物、肌酐、钙	0.62	90.77
8	氯化物、肌酐	3.44	90.69
9	氯化物、肌酐、钙	1.89	90.53
10	钠、肌酐、钙、葡萄糖	0.79	89.39
11	肌酐、钙、葡萄糖	5.10	88.33
12	钠、肌酐	1.77	88.24
13	钠、肌酐、钙	0.96	87.84
14	肌酐、钙	7.12	86.84
15	肌酐、葡萄糖	11.38	86.25
16	肌酐	16.47	84.47

当后项为葡萄糖时，按照置信度从高到低的顺序对以上规则进行排列，如表 12，结果显示，病人生化常规检查中钠、氯化物、钙、尿素 4 项指标同时异常时，同时葡萄糖指标异常的可能性最大，异常概率为 87.08%。概率其次的为同时钠、钙、尿素 3 项指标异常，概率为 85.78%。

表 12. 生化指标之间的关联规则 后项=葡萄糖

Table 12. association rules of biochemical indices: consequent= glucose

序号	前项	支持度%	置信度%
1	钠、氯化物、钙、尿素	0.92	87.08
2	钠、钙、尿素	1.53	85.78
3	钠、氯化物、钙	1.32	85.12
4	钠、氯化物、肌酐、钙、尿素	0.56	84.75
5	钠、氯化物、尿素	1.62	84.66
6	钠、钙	2.49	84.08
7	钠、氯化物、肌酐、钙	0.62	83.85
8	钠、尿素	2.77	83.71
9	钠、肌酐、钙、尿素	0.85	83.33
10	钠、氯化物	2.32	82.29
11	钠、氯化物、肌酐、尿素	1.02	82.20
12	钠、肌酐、钙	0.96	81.89
13	钠	4.58	81.20
14	钠、氯化物、肌酐	1.11	80.65
15	钠、肌酐、尿素	1.56	80.40

对以上结果进行进一步统计，整理得结果如表 13 所示。后项为尿素的关联规则共 16 条，其中肌酐异常作为前项共出现 16 次，其余项均出现 8 次；后项为葡萄糖的关联规则共 15 条，其中钠异常作为前项共出现 15 次，而钙，氯化物，尿素作为前项均出现 8 次。以上分析结果说明，肌酐异常时尿素异常的最关联因素，而钠异常是葡萄糖异常的最关联因素。

表 13. 生化指标之间的有意义关联规则

Table 13. The meaningful association rules of biochemical indices

前项	后项为尿素		后项为葡萄糖	
	出现次数	次序	出现次数	次序
尿素	——	——	8	2
肌酐	16	1	7	5
葡萄糖	8	2	——	——
钙	8	2	8	2
钠	8	2	15	1
氯化物	8	2	8	2

5 结论

本文主要从以下几个方面探讨了糖尿病肾病与生化检测指标的关系:

1.筛选与糖尿病肾病发病有关的生化检测指标：从 301 医院 2009、2010、2011 三年住院收治的糖尿病患者中随机抽取 1039 例糖尿病肾病患者和 1039 例未有肾病的糖尿病患者，对其进行描述性分析和条件 Logistic 回归分析，找出与糖尿病肾病发病有关的生化检测指标：肌酐、尿素、氯化物、钠、钙、葡萄糖，并从原总样本中重新随机抽取 1039 例糖尿病肾病患者和 1039 例未有肾病的糖尿病患者，对其肌酐、尿素、氯化物、钠、钙、葡萄糖 6 项生化指标进行分析，结果与前次基本相同；发现尿素、肌酐钙、钠、氯化物五个生化指标为糖尿病肾病发生的正相关指标；血糖为糖尿病肾病的发生的负相关指标。根据 OR 值给出 5 个正相关的生化指标的影响作用从强到弱依次为：肌酐、尿素、钠、钙、氯化物。

2．构建 Logistic 回归评分模型：将 301 医院 2009、2010、2011 三年住院收治的全部糖尿病患者中，生化检测结果中肌酐、尿素、钠、钙、氯化物、葡萄糖 6 种指标齐全的患者随机抽取分为训练集和测试集。使用训练集人群，拟合生成 Logistic 回归模型并对训练集每一个个体建立评分方程，得到方程见式（2）。计算发病率并拟合成函数见式（3），建立 Logistic 回归评分模型。应用测试集人群对模型进行测试，结果其中糖尿病肾病患者平均得分为 0.133，高于非糖尿病肾病患者的平均得分 0.028。针对测试集人群应用累积分布曲线和约登指数，确定判断本模型得分的临界阈值为 0.03 分，此时模型的灵敏度为 85.04%，约登指数为 0.746。从评价结果来看，Logistic 回归评分模型的正确率与特异度分别为 89.60%和 89.76%，处于比较高的水平，可以用于糖尿病患者群体的糖尿病肾病的筛查工作。

3.关联性分析：研究了生化常规检查中的一些指标与糖尿病肾病之间的关联关系，以及生化指标之间的关联关系。通过使用 Apriori 算法构建了关联模型，并分别分析了糖尿病肾病患病与生化指标间的关联性和以尿素和葡萄糖为代表的生化指标之间的相互关联性。以糖尿病肾病为后项得到 63 条关联规则，以生化指标互为前后项得到 186 条规则，并有 111 条规则的置信度大于 50%，且支持度均大于 0.5%。从不同的角度对得到的结果进行了详尽的分析，并试图寻找到更具有意义的关联规则。