基于 Apriori 算法的高血压电子病历研究*

杨美洁 史云杰 李 准

（重庆医科大学医学信息学院 重庆 400016） （重庆医科大学附属儿童医院 重庆 400014）

〔摘要〕 采用 SQL 技术对电子病历数据进行预处理，将高血压患者按照性别以 65 岁为界，分为 <65 岁男、<65 岁女、≥65 岁男、≥65 岁女 4 组。运用 Apriori 算法研究检查检验结果与用药之间的关联规则，挖掘出 29 条强关联规则，对制定高血压临床诊疗方案提供参考依据。

〔关键词〕 高血压；电子病历；Apriori 算法；关联规则；强规则

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2016.03.013

Research on Electronic Medical Records of Hypertension Based on Apriori Algorithm *YANG Mei－jie, SHI Yun－Jie, Chongqing Medical Informatics College, Chongqing Medical University, Chongqing 400016, China; LI Zhun, Children's Hospital of Chongqing Medical University, Chongqing 400014, China*

〔**Abstract**〕 The paper uses SQL to preprocess data of Electronic Medical Records (EMR) and divides hypertension patients into 4 groups in terms of sex and the age of 65, namely male <65, female <65, male ≥65 and female ≥65. By use of Apriori algorithm, it studies the association rules between examination results and use of medicine and excavates 29 strong association rules. These provide reference for making plans for clinical diagnosis and treatment of hypertension.

〔**Keywords**〕 Hypertension; Electronic Medical Records (EMR); Apriori algorithm; Association rules; Strong association rules

1 引言

高血压是危害身体健康的常见病和多发病，是冠心病、心力衰竭、肾功能衰竭和脑中风的主要危害因素之一[1]。电子病历是医务人员在医疗活动过程中，使用医疗机构信息系统生成的文字、符号、图表、图形、数据、影像等数字化的医疗信息资料，是病历的一种记录形式[2]。目前电子病历的关联规则挖掘成为研究的热点。李准等[3]利用 Apriori 算法对冠心病电子病历的检查检验结果和用药规则之间的关系进行了研究。曾勇等[4]对脑科医院癫痫病住院电子病历进行研究，利用 Apriori 算法发现癫痫病的病因。丁卫平等[5]利用改进的快速挖掘算法－FG 算法对电子病历系统历史数据进行挖掘，结果表明该算法能有效提取隐含在电子病历系统中有用的信息，为医疗诊断提供辅助性的决策。郑银丽[6]等利用关联规则算法对医药零售业药品营销组合进行研究，提高销售业绩。

现研究重庆市某医院的高血压电子病历，对高血压患者按照性别、年龄进行分组，利用马克威软件中的 Apriori 算法分析各项检查检验指标与用药之间的关系，挖掘出不同年龄段、不同性别高血压患者的检查检验指标与用药之间的关联规则，为不同年龄段、不同性别的高血压患者的临床诊断和用药

〔修回日期〕 2015－12－18

〔作者简介〕 杨美洁，讲师，发表论文 5 篇。

〔基金项目〕 重庆医科大学医学信息学院助力计划（项目编号：2014A009）。

规则提供参考和依据。

2 资料与方法

2.1 资料来源

收集重庆市某综合医院近年主诊断为高血压的电子病历，经过数据预处理后的电子病历677 份。

2.2 研究方法

将677 份高血压电子病历备份到基本信息、检查检验、药品医嘱等3 个表中。高血压的发病与年龄、性别等因素有关[7]，因此对高血压患者按照年龄和性别进行分组。从基本信息中选取住院号、性别、年龄3 个属性进行研究。根据相关书籍、文献[7-10]并咨询相关专家，从病历数据中选取了高血压常见的75 个检查检验项目和51 种药品。数据预处理是进行关联规则挖掘的基础和前提，包括数据清洗、集成、转换[3]。本文利用SQL Server 2008 对数据进行处理，将病历中的文本数据转换成数值型数据。（1）基本信息，对基本信息中的性别、年龄进行处理。性别按照男、女分别取值为1、2；年龄以65 岁为界，低于和高于65 岁分别取值为1、2。（2）检查检验，利用SQL 技术，将检查检验的结果中高于正常值范围、正常、低于正常值范围分别取值为3、2、1，未做此检查检验项目取值0。（3）药品医嘱，将病历中的药品名称内容与已有的药品关键词进行匹配，若存在此药品名称则取值为1，否则取值为0。

3 关联规则

3.1 概述

关联规则主要反映事物之间的关联。支持度和置信度是描述关联规则的两个重要概念，主要描述规则的有用性和确定性[11]。支持度是事务集D 中事件X 和事件Y 同时发生的概率，置信度是事务集D 中事件X 发生的前提下，事件Y 发生的概率[6]。关联规则是挖掘同时满足用户给定的最小支持度和最小置信度的强关联规则[6]。本文采用马克威软件，通过Apriori 算法挖掘高血压患者的检查检验结果与用药之间的关联。最小置信度均设为80%，最小支持度从50% 逐渐下调，每次间隔10%，但不低于10%。将高于正常值范围、低于正常值范围分别简化为高和低。

3.2 <65 岁男性患者（67 人）

将数据导入马克威软件，多次测试发现最小支持度设为10% 时效果最佳。具体的强规则，见表1。共3 条强规则，规则支持度相对都较低，说明规则普遍性相对较低。出现超敏肌钙蛋白T 高和葡萄糖高的患者中有83.05% 服用了阿司匹林肠溶片。

表1 <65 岁男性患者之强规则

序号	关联规则	支持度	置信度
1	超敏肌钙蛋白T（3）Λ 淋巴细胞百分比（1）= = > > 阿司匹林肠溶片（1）	10.60	82.62
2	超敏肌钙蛋白T（3）Λ 中性粒细胞百分比（3）= = > > 阿司匹林肠溶片（1）	12.02	80.43
3	超敏肌钙蛋白T（3）Λ 葡萄糖（3）= = > > 阿司匹林肠溶片（1）	13.21	83.05

3.3 <65 岁女性患者（127 人）

将数据导入马克威软件，多次测试发现将最小支持度设为10% 时效果最佳。具体的强规则，见表2。共7 条强规则，规则支持度相对都较低，说明规则普遍性相对较低。其中肌红蛋白高、肌酐高、尿素高的人中95.21% 服用了阿托伐他汀钙片。

表2 <65 岁女性患者之强规则

序号	关联规则	支持度（%）	置信度（%）
1	肌红蛋白（3）Λ 肌酐（3） = = > > 阿托伐他汀钙片（1）	10.81	95.02
2	红细胞分布宽度 - SD（3）Λ 淋巴细胞百分比（1） = = > > 丹参酮（1）	10.91	85.31
3	淋巴细胞百分比（1）Λ 尿素（3） = = > > 泮托拉唑钠（1）	13.55	88.39
4	肌红蛋白（3）Λ 淋巴细胞百分比（1） Λ 尿素（3） = = > > 泮托拉唑钠（1）	10.81	85.61
5	肌红蛋白（3）Λ 肌酐（3）Λ 尿素（3） = = > > 阿托伐他汀钙片（1）	10.81	95.21
6	淋巴细胞百分比（1）Λ 中性粒细胞百分比（3）Λ 尿素（3） = = > > 泮托拉唑钠（1）	11.73	87.20
7	超敏肌钙蛋白 T（3）Λ 淋巴细胞百分比（1）Λ 中性粒细胞百分比（3）Λ 尿素（3） = = > > 泮托拉唑钠（1）	10.81	85.71

3.4 ≥65 岁男性患者（208 人）

将数据导入马克威软件，多次测试发现最小支持度设为10%时效果最佳。具体的强规则，见表3。共8 条强规则，规则支持度与置信度都相对较高，说明规则普遍性相对较高。其中超敏肌钙蛋白 T 高的患者80.31%的服用了呋塞米。

表3 ≥65 岁男性患者之强规则

序号	关联规则	支持度（%）	置信度（%）
1	超敏肌钙蛋白 T（3） = = > > 呋塞米（1）	30.65	80.31
2	血红蛋白（1）Λ 平均血红蛋白浓度（1） = = > > 呋塞米（1）	30.65	80.31
3	低密度脂蛋白胆固醇（3）Λ 平均血红蛋白浓度（1） = = > > 阿司匹林肠溶片（1）	30.65	80.31
4	超敏肌钙蛋白 T（3）Λ 平均血红蛋白浓度（1） = = > > 阿司匹林肠溶片（1）	30.58	80.23
5	低密度脂蛋白胆固醇（3）Λ 血红蛋白（1） = = > > 银杏达莫片（1）	30.58	80.23
6	淋巴细胞百分比（1）Λ 平均血红蛋白浓度（1） = = > > 银杏达莫片（1）	30.58	80.23
7	低密度脂蛋白胆固醇（3）Λ 超敏肌钙蛋白 T（3） = = > > 阿司匹林肠溶片（1）	30.58	80.23
8	平均血红蛋白浓度（1）Λ 呋塞米（1） = = > > 非洛地平缓释片（1）	30.58	80.23

3.5 ≥65 岁女性患者（275 人）

将数据导入马克威软件，多次测试发现最小支持度设为10%时效果最佳。具体的强规则，见表4。共11 条强规则，规则支持度和置信度相对较高，说明规则普遍性相对较高。其中低密度脂蛋白胆固醇高、平均血红蛋白浓度低的患者100%服用了阿司匹林肠溶片。

表 4 ≥65 岁女性患者之强规则

序号	关联规则	支持度（%）	置信度（%）
1	红细胞（3） = = > >阿司匹林肠溶片（1）	36.84	100.00
2	白细胞（3） = = > >左氧氟沙星（1）	31.58	85.71
3	低密度脂蛋白胆固醇（3） = = > >阿司匹林肠溶片（1）	36.84	87.50
4	白细胞（3）Λ 嗜酸性粒细胞百分比（1） = = > > 阿司匹林肠溶片（1）	31.58	100.00
5	白细胞（3）Λ 中性粒细胞百分比（3） = = > > 阿司匹林肠溶片（1）	31.58	100.00
6	白细胞（3）Λ 葡萄糖（3） = = > > 阿司匹林肠溶片（1）	31.58	100.00
7	低密度脂蛋白胆固醇（3）Λ 淋巴细胞百分比（1） = = > > 阿司匹林肠溶片（1）	36.84	87.50
8	低密度脂蛋白胆固醇（3）Λ 平均血红蛋白浓度（1） = = > > 阿司匹林肠溶片（1）	31.58	100.00
9	低密度脂蛋白胆固醇（3）Λ 嗜酸性粒细胞百分比（1） = = > > 阿司匹林肠溶片（1）	36.84	87.50
10	低密度脂蛋白胆固醇（3）（3）Λ 嗜酸性粒细胞百分比（1） = = > > 阿司匹林肠溶片（1）	31.58	85.71
11	低密度脂蛋白胆固醇（3）（3）Λ 中性粒细胞百分比（3） = = > > 阿司匹林肠溶片（1）	36.84	87.50

4 讨论

4.1 提供用药依据

本文对高血压患者按照性别、年龄进行划分，利用 Apriori 算法对不同性别、不同年龄的高血压患者的检查检验结果和用药进行强规则挖掘。结果显示：（1）年龄 <65 岁的患者强规则的普遍性不高。其中 <65 岁男性患者大多具有超敏肌钙蛋白 T 的症状，主要用药阿司匹林肠溶片；<65 岁女性患者大多具有中肌红蛋白高、肌酐高、尿素高等症状，主要用药阿托伐他汀钙片、泮托拉唑钠。（2）≥65 岁男性患者具有低密度脂蛋白胆固醇高、超敏肌钙蛋白 T 高、平均血红蛋白浓度低等症状，主要用药阿司匹林肠溶片、呋塞米、银杏达莫片、非洛地平缓释片。≥65 岁女性患者具有低密度脂蛋白胆固醇高、平均血红蛋白浓度低、白细胞高、淋巴细胞百分比低、嗜酸性粒细胞百分比低、中性粒细胞百分比高等症状，主要用药为阿司匹林肠溶片。上述结果为医生对不同年龄段、不同性别的患者诊断和用药提供了临床实践参考，为高血压慢病管理的用药管理提供了依据。

4.2 今后重点

基于 Apriori 算法对高血压电子病历患者的检查检验结果与用药进行了关联规则挖掘。未来的工作中将基于 FP – growth 关联规则算法对高血压电子病历的检查检验结果与用药进行挖掘，比较二者结果的异同。还将利用关联规则算法对糖尿病等电子病历的检查检验结果与用药进行挖掘，为医生的临床实践和慢病管理等方面提供参考依据。

（下转第 76 页）

者解答任何关于图书馆的问题。不仅实现了微博平台的全部功能，而且没有字数限制。（2）微信的公众号码可细分为服务号和订阅号，相对于订阅号，服务号群发信息的时候所有关注的手机用户会像接收短信一样收到信息，沟通更加方便快捷。微信提供的服务是一对一的，形式不限于文字，还可通过生动活泼的视频、图片、语音等，拉近图书馆与用户之间的距离，提高公众的关注度与普及度，在一定程度上也起到免费的宣传推广作用。（3）提供更专业、更个性化的服务，如查新、查引等。

综上所述，不论从平台的管理上还是和用户的互动上，微信比其他两种平台形式具有更多优势。但也不能忽略微博和博客发布信息的作用。现代图书馆员应根据自身的需求，选择合适医院临床和科研发展的手段为广大读者服务。

4 结语

信息资源共享已成为21 世纪图书馆事业发展的一大主题[4]。面对海量文献信息资源和复杂多变的信息需求，任何图书馆单凭自己的力量难以全面搜集信息资源，满足读者的不同信息需求。通过建立现代服务平台，可以把兄弟图书馆汇集在一起，进行区域协作，开展馆际互借、协调采购等资源共享模式，加强彼此间的业务交流，促进图书馆服务水平的整体提升。医院图书馆也应对其应用进行探索，根据自身需求选择、提供适宜的服务内容，才能更好地适应信息化社会的需要。

参考文献

1 孙丽，王丽伟. 图书馆2.0 服务模式构建［J］. 医学信息学杂志，2011，32（8）：69－72.

2 百度百科. 微信［EB/OL］.［2015－10－09］. http：//baike. baidu. corn/link? url = 6 － D1VF6rWVBTIr jjjmeZMxHOkYCFgaaDU8pSMFyKwx6nLPpX － mpXXi5fpbIVVEGCzH9dsthGq5LousUF00JkVq.

3 贺青，钟方虎，于丽，等. 微博客与医学图书馆服务［J］. 医学信息学杂志，2010，30（8）：76－78.

4 钟方虎，贺青，于丽. 微博在图书管中的应用［J］. 中华医学图书情报杂志，2012，21（3）：31－32.

（上接第61 页）

5 结语

本文以高血压电子病历为研究对象，运用SQL技术对数据进行预处理，将高血压的患者按照性别、年龄分类，通过Apriori 算法分析各类检查检验结果与用药之间的关联，挖掘出27 条不同年龄和性别的检查检验结果与用药之间的强规则，可以为医生为不同年龄和不同性别的高血压患者的临床用药和慢病管理提供参考依据。

参考文献

1 刘力生. 中国高血压防治指南2010［J］. 中华高血压杂志，2011，（8）：701－743.

2 谢栋梁. 电子病历初探［J］. 医学信息（上旬刊），2010，（6）：1781－1782.

3 李准，冯思佳，杨美洁，等. 关联规则技术在冠心病电子病历中的应用［J］. 医学信息学杂志，2015，（1）：58－62.

4 曾勇. 关联规则在脑科电子病历挖掘中的应用［J］. 医学信息学杂志，2014，（10）：55－58.

5 丁卫平，祁恒，董建成，等. 基于关联规则的电子病历挖掘算法研究与应用［J］. 微电子学与计算机，2007，（3）：69－73，76.

6 郑银丽，相秉仁，赵国明. 关联规则技术在医药零售业药品营销组合中的应用［J］. 医学信息学杂志，2011，（4）：55－58.

7 汪晓芬，和渝斌，张源波，等. 高血压发病机制及诊疗研究的新认识［J］. 中国循证心血管医学杂志，2015，（2）：281－282.

8 王婷婷，吴迪. 我国H 型高血压的研究现状［J］. 医学综述，2015，（6）：1042－1044.

9 王立华，马葵芬，黄庆君，等. 6 种基本药品治疗原发性高血压的经济学分析［J］. 中国医药指南，2011，（21）：259－261.

10 姜虹，毕宪初，周跃. 基本药品目录实行前后社区高血压用药情况分析［J］. 上海医药，2014，（2）：41－42.

11 Han J，Kamber M，Pei J. Data Mining：concepts and techniques［M］. 3rd ed. Burling：Morgan Kanfmann Pubeishers，2011.