

剖宫产手术预防性使用抗菌药的数据挖掘

傅翔¹, 杨樟卫², 陈盛新³, 陈长虹⁴, 何宇涛⁴ (1. 南京军区杭州疗养院药械科, 浙江 杭州 310007; 2. 第二军医大学长海医院药学部, 上海 200433; 3. 第二军医大学药学院药事管理学教研室, 上海 200433; 4. 第二军医大学长海医院信息科, 上海 200433)

[摘要] 目的 建立、比较和评价剖宫产手术抗菌药物预防使用的分类模型, 为针对性干预打下基础。方法 应用数据挖掘软件 PASW[®] Modeler 13, 建立分类模型, 获得对抗菌药物预防性使用影响较大的变量(临床因素)。结果 由 787 例行“子宫下段剖宫产术”的病例数据获得的分类模型中, 以贝叶斯网络, logistic 回归和 CHAID 3 个模型总体较佳; 21 个变量指标中, “失血量”是对该医院剖宫产手术抗菌药物预防性应用影响程度最大的因素。结论 数据挖掘技术, 可以快速地建立反映剖宫产手术抗菌药物预防性使用的分类模型, 为药物利用调查提供了新的分析方法。

[关键词] 数据挖掘; 抗菌药物; 剖宫产

[中图分类号] R95 **[文献标志码]** A **[文章编号]** 1006-0111(2012)02-0109-06

[DOI] 10.3969/j.issn.1006-0111.2012.02.009

Data mining of antibiotic prophylactic use for cesarean section patients

FU Xiang¹, YANG Zhang-wei², CHEN Sheng-xin³, CHEN Chang-hong⁴, HE Yu-tao⁴ (1. Department of Pharmacy and Medical Appliances, Hangzhou Sanatorium of Nanjing Military Region, Hangzhou 310000, China; 2. Department of Pharmacy, Changhai Hospital, Shanghai 200433, China; 3. Department of Pharmacy Administration, School of Pharmacy, SMMU, Shanghai 200433, China; 4. Department of Information, Changhai Hospital, Shanghai 200433, China)

[Abstract] **Objective** To establish, compare and evaluate the classification models of antibiotic prophylactic use for cesarean section patients for the targeted intervention in future. **Method** PASW[®] Modeler 13 was applied to establish classification models and to get the influential variables (clinical factors) in antibiotic prophylactic use. **Result** With the data of 787 cases, the classification models were established, in which, Bayesian networks, logistic regression and CHAID were better. In 21 clinical factors, *blood loss* was the most influential variable. **Conclusion** The data mining technique was able to quickly create models reflecting the use of prophylactic antibiotics use for cesarean section, which would provide a new analysis tool for drug use survey.

[Key words] data mining; antibiotics; cesarean section

剖宫产是当前产科学中一种常见而重要的手术, 近年来, 国内剖宫产率出现迅猛增高的势头。剖宫产手术属于 II 类(清洁-污染)切口手术, 一般需预防性使用抗菌药物。

近年来, 有关剖宫产围术期抗菌药物使用调查或干预的文献时有报道, 这些研究通过了解剖宫产围术期抗菌药物的应用现状^[1-4], 反映剖宫产抗菌药物预防性使用中存在的问题。由于临床环境和临床实践的复杂多变性, 病人的药物治疗受多种因素影响, 仅仅通过对药物品种、使用频率、用药时间、药品费用数据的调查, 如忽视其他临床信息, 得出的结论可能比较片面, 干预也将缺乏针对性。

数据挖掘(data mining, DM)又称数据库中的知

识发现, 是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中, 提取隐含在其中的、事先不知道的、但又潜在有用的信息和知识的过程^[5]。分类是数据挖掘的常用模式之一, 是指把数据样本映射到一个事先定义的类中的学习过程, 由一组输入的描述属性值和相应的类标组成。

本研究利用某三甲医院 HIS 中剖宫产病人的用药数据, 结合病情、诊断、手术等可能影响抗菌药物使用的临床数据, 对剖宫产抗菌药物预防性使用情况进行分析, 基于数据挖掘中的分类技术, 利用数据挖掘软件 SPSS[®] Modeler, 建立分类模型, 挖掘影响抗菌药物使用模式的变量因素, 为针对性地加强围术期抗菌药使用的决策和干预提供依据。

1 数据来源

在某三甲医院 2009 年出院病人中, 获取进行

[作者简介] 傅翔(1972-), 男, 博士. E-mail: fqj2000@hotmail.com.

“子宫下段剖腹产术”的全部 807 例病例,排除在住院期间还接受其他手术的病例 20 例,余下 787 例为本次研究的对象。采集 787 例剖宫产手术病人的基本信息和诊断、手术、医嘱等数据,归纳为 3 类(常规、诊查、手术)21 种因素,作为分类预测变量。

2 研究方法

2.1 数据预处理 对原始数据中预测变量(描述属性)进行数据转换。每个描述属性(分类预测变量)指标数量化。对数字连续型变量,如“年龄”和“孕周”参考产科妊娠评分指标^[6]进行离散化。

2.2 目标变量(类标属性)分类标号确定 以抗菌药物预防使用作为分类挖掘的目标变量(类标号属性)。根据《剖宫产手术围手术期预防用抗菌药物管理实施细则(征求意见稿)》,总结剖宫产手术预防性使用抗菌药物的品种和方法见表 1。

表 1 剖宫产手术常用预防性抗菌药物和静脉给药方法

抗菌药物	单次剂量(g)	给药时机	给药时长
头孢唑啉	1~2	1、脐带结扎后,手术时间持续	1、手术结束后不必再用。
头孢拉定	1~2	2、若有感染高危因素者,	2、若有感染高危因素者,
头孢呋辛	1.5	时间超过 3 h,或失血量超过 1	术后 24 h 内可再用 1~3
头孢西丁	1~2	500 ml,应补充一个剂量	次,特殊情况可延长至术后 48 h。
克林霉素	0.6~0.9		
氨曲南	1~2		
甲硝唑	0.5		

由于“剂量”、“给药途径”及“给药时机”上各病例间基本一致,因此主要从“选用品种”及“给药时长”角度进行比较和判别。“选用品种”或“给药时长”两项均“符合标准”,则定义该病例抗菌药物预防性使用整体“符合标准”;“选用品种”或“给药时长”两项任一项不符合,则定义该病例抗菌药物预防性使用整体“不符合标准”,分别赋值“0”和“1”。

2.3 预测变量的初步筛选 使用特征选择算法来缩小预测变量的选择范围,识别分析中重要的字段。通过将注意力迅速集中到最重要的字段上,可以降低所需的计算量,最终获得更简单、精确和易于解释的模型。

2.4 训练样本和测试样本的确定 采用 K-折交叉验证(K>2)的原理,将样本集随机分成 3 个样本数基本相同且互不重叠的子集。依次取 1 个子集为测试集,其余 2 个子集合并为训练集合,获得 3 组训练测试数据。每种模型都采用这 3 组数据进行训练和测试。

2.5 分类模型建立与评价 采用决策树模型

(C&RT、CHAID、QUEST、C5.0 算法)、神经网络、贝叶斯网络和 Logistic 模型建立分类模型,观察各预测变量的相对重要性。

2.5.1 准确性评价 准确性是指挖掘模型与所提供数据中的属性的相关联程度,即采用这个算法所获的模式(知识)对样本进行判别分类结果的正确性,常用准确度(accuracy)或曲线下面积(AUC)衡量。

2.5.2 有用性评价 有用性反映了模型是否提供有用信息各种指标,体现了对业务的优化能力,可用提升指数(lift)衡量。

2.5.3 可靠性评价 可靠性指当提供不同的测试数据时,挖掘模型表现出稳定预测结果的能力,即对训练样本进行判别的准确率与对测试样本进行判别准确率的一致性。定义可靠因子 R(reliability)^[7]。

$$R = \frac{\text{对未知样本的分类准确度}}{\text{对已知样本的分类准确度}}$$

R 越接近 1,表示算法所获模式的可靠度高。由于很难推测未知样本的概率分布,所以实际计算中采用 r 近似于 R:

$$r = E\left(\frac{\text{对测试样本的分类准确度}}{\text{对训练样本的分类准确度}}\right)$$

3 研究结果

3.1 数据预处理结果 787 例行“子宫下段剖腹产术”的病例在围手术期间全部使用抗菌药物,对照前文所述的抗菌药物合理使用的标准,符合标准的 700 人(88.95%),不符合标准的 87 人(11.05%)。经预处理的预测变量及目标变量汇总见表 2。

3.2 特征筛选结果 经运行 Modeler 中的特征筛选节点。“serious_cond”等 6 个字段因单个类别过大被筛选;“charge_type”等 5 个字段因重要性偏低而被筛选,保留剩余的“重要”的“blood_loss”等 10 个字段作为下一步建模的预测变量。

3.3 分类模型评价

3.3.1 模型准确性比较结果 总体来看,各模型在训练中的准确率差别不大,平均都在 91% 以上,贝叶斯网络、神经网络和 Logistic 回归的平均准确率超过 92%;但在测试过程中,各模型表现的准确率有所差别,最低的为贝叶斯网络,其次神经网络模型的测试准确率也低于 90%,最高的为 CHAID,为 91.61%。

表2 变量指标与赋值

	变量代码	变量标识	值标识	变量值
常规	age	年龄(y)	<20	1
			20 ≤ and <30	2
			30 ≤ and <35	3
			35 ≤ and <40	4
			40 ≤	5
	preg_week	孕周(w)	<37	1
			37 ≤ and ≤42	2
			42 <	3
	BMI	身体质量指数	≤25	1
			25 < and ≤30	2
			30 < and ≤35	3
			35 <	4
	primipara	产次	初产	0
	charge_type	费别	经产	1
军队医改			1	
地方医保			2	
pat_adm_condition	入院病情	地方标准	3	
		危	1	
		急	2	
alergy_drugs	过敏药物	一般	3	
		无	0	
		青霉素或头孢	1	
		甲硝唑	2	
		其他	3	
诊查	attending_doctor	主治医生	医生姓名	1 ~ 16
	risk_fact	感染高危因素	分娩并发症	1
			妊娠合并症	2
			临产后的剖宫产手术	3
			2种以上上述诊断	4
			无上述诊断	0
	serious_cond	住院期间病重	无	0
			有	1
	emer_treat	住院期间抢救	无	0
			有	1
hospital_stay	住院日(d)	≤7	1	
		7 < and ≤14	2	
		14 < and ≤21	3	
		21 <	4	
		0 < and ≤7	1	
手术	bef_operation	术前住院日(d)	7 < and ≤14	2
			14 <	3
			全麻	1
	anaesthesia_method	麻醉方式	腰麻	2
			硬膜外	3
			腰硬联	4
			特	1
	operation_scale	手术综合等级	大	2
			中	3
			≤60	1
ope_time	手术时间(min)	60 < and ≤120	2	
		120 < and ≤180	3	
		180 <	4	
		≤500	1	
		500 < and ≤1000	2	
in_fluids_amount	术中液体总入量(ml)	1 000 < and ≤2000	3	
		2 000 <	4	
		≤500	1	
		500 < and ≤1000	2	
out_fluids_amount	术中液体总出量(ml)	≤500	1	

变量代码	变量标识	值标识	变量值
blood_losed	失血量(ml)	500 < and ≤1000	2
		1 000 < and ≤2000	3
		2 000 <	4
		≤200	1
		200 < and ≤400	2
		400 <	3
blood_transferred	输血量(ml)	≤200	1
		200 < and ≤400	2
		400 <	3
heal	切口愈合	甲	1
		乙	2
目标变量 anti_bact	抗菌药物预防使用	符合标准	0
		不符合标准	1

表3 各模型训练与测试样本分类平均准确率汇总比较

	训练样本分类准确率 (%)	测试样本分类准确率 (%)
贝叶斯网络	93.72	85.94
Logistic 回归	92.05	91.11
神经网络	92.40	89.10
CHAID	91.61	91.61
QUEST	91.81	91.33
C&R 树	91.87	91.49
C5	91.67	91.49

从各模型的 ROC 曲线下面积来看(表4),贝叶斯网络,Logistic 回归,神经网络平均在 0.80 以上,CHAID 平均为 0.79,接近 0.80。

表4 各模型 ROC 曲线下面积

模型	S1S2	S1S3	S2S3	平均值
贝叶斯网络	0.92	0.90	0.93	0.92
Logistic 回归	0.87	0.85	0.84	0.85
神经网络	0.86	0.83	0.83	0.84
CHAID	0.78	0.81	0.79	0.79
QUEST	0.66	0.69	0.79	0.71
C&R 树	0.66	0.69	0.65	0.67
C5	0.61	0.69	0.65	0.65

3.3.2 模型可用性比较结果 由表5可见,贝叶斯网络,神经网络, Logistic 回归提升指数较高,分别为 5.92,5.50,5.03。

表5 各模型提升指数

模型	S1S2	S1S3	S2S3	平均值
贝叶斯网络	6.15	6.12	5.49	5.92
神经网络	6.35	5.18	4.97	5.50
Logistic 回归	5.16	5.10	4.84	5.03
CHAID	3.85	4.64	4.40	4.30
QUEST	3.75	4.19	4.40	4.11
C&R 树	3.75	4.21	3.51	3.82
C5	2.92	4.21	3.51	3.55

3.3.3 模型可靠性比较结果 由表6和表7可知,7种模型的可靠因子都在 0.92 以上;决策树的4种模型可靠因子与1最为接近,均在 0.99 和 1 之间;其次是 Logistic 回归,为 0.989 8;神经网络为 0.964 4;贝叶斯网络最低,为 0.921 1。

表6 各模型可靠因子

	S1S2-S3	S1S3-S2	S2S3-S1	平均值
贝叶斯网络	0.921 1	0.899 6	0.930 3	0.921 1
Logistic 回归	0.979 4	0.985 6	1.004 4	0.989 8
神经网络	0.940 3	0.974 8	0.978 2	0.964 4
CHAID	0.999 6	0.996 1	1.004 5	1.000 0
QUEST	0.993 1	0.988 8	1.002 3	0.994 7
C&R 树	0.993 1	0.994 1	1.000 4	0.995 9
C5	0.999 6	0.994 1	1.000 4	0.998 0

表7 各模型的 |r-1|

	S1S2-S3	S1S3-S2	S2S3-S1	平均值
贝叶斯网络	0.078 9	0.100 4	0.069 7	0.083 0
Logistic 回归	0.020 6	0.014 4	0.004 4	0.013 1
神经网络	0.059 7	0.025 2	0.021 8	0.035 6
CHAID	0.000 4	0.003 9	0.004 5	0.002 9
QUEST	0.006 9	0.011 2	0.002 3	0.006 8
C&R 树	0.006 9	0.005 9	0.000 4	0.004 4
C5	0.000 4	0.005 9	0.000 4	0.002 2

综合准确性、可用性和可靠性,从中选择并使用贝叶斯网络,logistic 回归和 CHAID3 个模型。

3.3.4 变量重要性选择结果 变量重要性图表显示预测变量的相对重要性,并且考虑到预测变量的交互性和相关性。在本研究中,变量重要性表示变量对剖宫产围手术期抗菌药物使用情况影响的程度。

贝叶斯网络、logistic 回归和 CHAID 模型的变量重要性如图 1~图 3 所示。贝叶斯网络模型所选择的 10 个变量字段中,除了“attending_doctor”(主治医师)和“anaesthesia_method”(麻醉方式)变量重要性较低,其余变量重要性差别不大,因此模型的说服

力不够强。

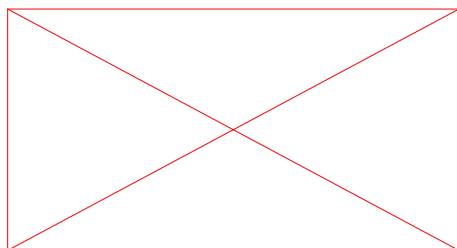


图1 贝叶斯网络模型变量重要性

Logistic 回归建模采用前进法,得出模型的变量重要性见图2,共有3个变量,以“blood_loss” (失血量)重要性最高,其次为“attending_doctor”和“anaesthesia_method”,其余变量重要性较低。最终 Logistic 模型中包含了“blood_loss(失血量)”和“anaesthesia_method”两个变量,“blood_loss”的影响更明显。结合发生比可知,失血量在200和400 ml,其发生比为失血量200 ml以下的3.678倍,即失血量多更容易引起抗菌药物的使用超出标准。

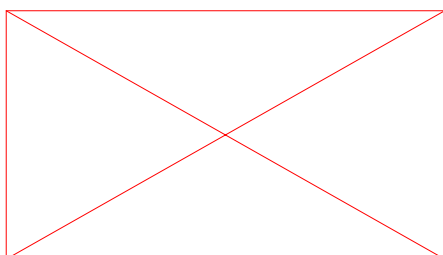


图2 Logistic 回归模型变量重要性

CHAID 只选择了3个变量(图3),其重要性由高到低分别为,“blood_loss”、“anaesthesia_method”和“hospital_stay”。由CHAID生成的决策树结构见图4,由于研究更为关心的是抗菌药物使用“不符合标准的”,表明当手术失血量大于2(大于400 ml),有85%可能引起围手术期抗菌药物预防性使用超过规定范围。综合3种模型分析,“blood_loss”是剖宫产手术抗菌药物预防性应用影响程度最大的因素。

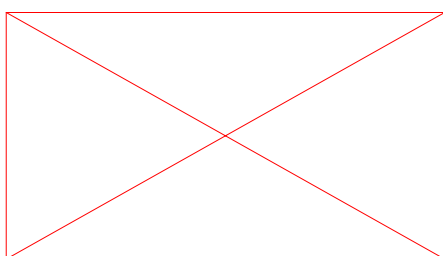


图3 CHAID 模型变量重要性

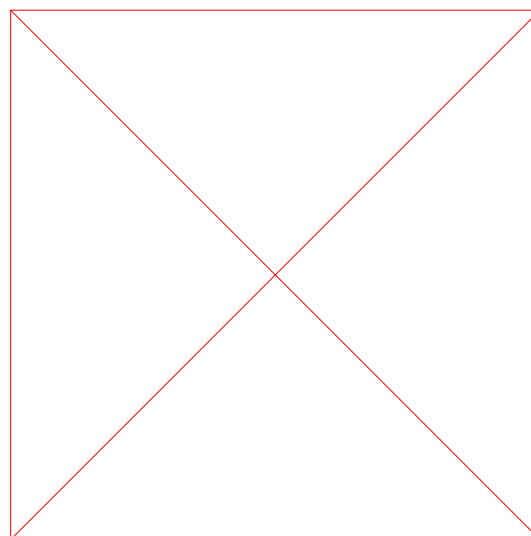


图4 CHAID 模型生成的决策树

以上结果可以提示,利用数据挖掘中的“分类”技术,可以快速地建立具有一定准确性、可用性和可靠性的,反映剖宫产手术抗菌药物预防性使用的分类模型,从中区分出可能出现“不符合标准”使用抗菌药物的病例。建立的模型提示:在没有进一步人为教育或干预措施下,在样本医院剖宫产手术中,失血量较多的患者更有可能在预防性使用抗菌药物时,在用药品种或者用药时间上,超出正常合理用药的需求。如果有下一步的针对剖宫产手术抗菌药物使用的教育或干预计划,对于失血量较多的病人应重点关注。

4 讨论

4.1 住院病人的药物使用情况与其他临床数据密不可分,即使意识到这点,由于临床数据的复杂性、数据来源的多样性以及分析方法的缺乏,目前国内在结合具体病例诊断、手术等临床数据的前提下开展药物利用分析或研究的报道仍罕见。本研究是利用数据挖掘技术,针对住院病人的用药数据进行分析利用的探索性研究。是对医疗机构住院病人相关药物利用信息分析方法尝试与创新,为药物利用研究的深入开展提供新的思路,有助于从当前海量的医疗数据中获取有用的知识。

4.2 当前,对抗菌药物使用的分析,已经从最初的使用品种、金额、限定日剂量等以“药物”为中心的总体性指标向以“病人”为中心的合理性研究转变。由于病情的复杂性以及医疗环境的特殊性,对药物的使用作出是否合理的评价应该全面而且慎重,并且需要足够的专业知识。因此,本研究在设定抗菌药物预防性使用的分类时,没有直接对

合理性作出结论,而代之以“符合”或“不符合”标准来进行分类。

4.3 本研究建立的模型发现“失血量”因素是影响该医院剖宫产围手术期抗菌药物使用的重要因素,提示可能由于失血较多,影响了病人的生理病理情况或医生对病情的判断,使医生倾向于超标准地使用抗菌药物。不同的模型还表明其他因素,如麻醉方式、住院时间、主治医师等对抗菌药物的使用有影响。这些“发现的知识”需要利用专业知识进一步筛选,评价和解释。通过分类模型建立以及变量重要性的获得,便于对抗菌药物使用容易出现“不符合标准”的病例进行重点监测和及时干预。

4.4 数据挖掘进程需要不断的循环和深入。本研究数据来源于1家医院1年的病例,得出的结论也较为粗浅。为了增强模型的说服力,有必要采用多个样本医院的数据加以综合。此外,仅通过“失血量”和“麻醉方式”来预测分类也是不全面的,还应考虑其他因素,使模型更加完善。

【参考文献】

- [1] 王敬花. 剖宫产围术期抗菌药物的使用干预与分析[J]. 中国医院用药评价与分析, 2010, 10(8): 688.
- [2] 张奕, 李润萍, 孟繁星. 我院剖宫产手术预防感染应用抗菌药物的合理性分析[J]. 实用药物与临床, 2010, 13(3): 218.
- [3] 姜涛. 785例产妇产后抗菌药物应用的合理性分析[J]. 中国医院药学杂志, 2010, 30(10): 884.
- [4] 孟现民, 丁天然, 张莉, 等. HBsAg阳性孕产妇剖宫产术抗菌药物预防用药调查与分析[J]. 中国药物应用与监测, 2010, 7(1): 29.
- [5] 傅翔, 陈盛新, 杨樟卫. 数据挖掘在合理用药信息分析中的应用[J]. 药学实践杂志, 2009, 27(6): 411.
- [6] 李小毛, 段涛, 杨慧霞主编. 剖宫产热点问题解读[M]. 北京: 人民军医出版社, 2008.
- [7] 叶晨洲, 杨杰, 耿道颖. 应用数据挖掘技术从大脑胶质瘤病例中获取诊断知识[J]. 生物医学工程学杂志, 2002, 19(3): 426.

[收稿日期] 2011-10-08

[修回日期] 2011-12-29

(上接第108页)

长范围内检测凝胶反应过程中的浊度(透光度)变化而确定供试品中内毒素含量的方法。待测样品中内毒素与鲎试剂的C因子反应, 激活凝固酶切断凝固蛋白原中特定位置的精氨酸链, 肽链凝固产生凝胶, 根据测定反应液到达事先设定的浊度(透光度)值所需要的时间的对数值与细菌内毒素浓度的对数值成反比关系而进行定量测量的一种分析方法。

3.2 细菌内毒素与鲎试剂的反应是由一系列酶促放大作用产生的, 其干扰因素主要包括供试品的酸碱度、金属离子浓度、非特异性鲎试剂激活物、对仪器检测光源的吸收干扰等因素。本实验干扰试验结果发现: 用动态浊度法定量测定人凝血因子Ⅷ(1→100)稀释液中细菌内毒素, 其回收率均在50%~200%, 表明此条件下不存在干扰因素, 且8批样品均在规定的限度0.5 EU/IU以下, 判为合格。

3.3 鲎试剂动态浊度法与传统的家兔检查法相

比, 具有操作简单、准确、检测灵敏度高、可定量等优点, 避免了由于家兔个体差异所致的假阳性或假阴性结果, 能直观地反映样品中细菌内毒素多少, 在试验的反应过程中可以动态观察并分析干扰情况。经过对比试验, 样品的细菌内毒素定量检测结果与家兔热原检测结果基本一致, 说明使用细菌内毒素动态浊度法测定Ⅷ因子制品中细菌内毒素是可行的。

【参考文献】

- [1] Van DK, Van Der Bom JG, Lenting PJ, et al. Factor VIII half-life and clinical phenotype of severe hemophilia A [J]. Haematologica, 2005, 90(4): 494.
- [2] 赵艳华, 马平, 卜凤荣. 凝血因子分子生物学研究进展[J]. 生物技术通讯, 1999, 10(2): 158.
- [3] 国家药典2010版. 三部[S]. 2010: 95.
- [4] 劳海燕, 罗宇芬, 林秋晓. 动态比浊法用于检测生脉注射液细菌内毒素的研究[J]. 广东药学报, 2002, 18(1): 21.

[收稿日期] 2011-12-22

[修回日期] 2012-03-06

欢迎订阅2012年《药学实践杂志》

本刊网址: www.yxsjzz.cn; yxsj.smmu.edu.cn