[文章编号]1672-3244(2013)02-0145-05

基于数据挖掘技术的老年口腔癌患者围术期并发症发生概率评估系统的建立

高宪超*,陈一铭*,俞志维,季彤
(上海交通大学医学院附属第九人民医院·口腔医学院 口腔颌面-头颈肿瘤科,
上海市口腔医学重点实验室,上海 200011)

[摘要] 目的:开发一种具有较强临床实用性的老年口腔癌患者围术期并发症发生概率评估系统,使得相关并发症发生概率的评估更加直观与高效。方法:根据 APACHEⅡ以及 POSSUM 评分体系,结合临床实际,确定核心数据项目。采用数据挖掘方法,分析各项临床数据之间潜在的逻辑关系,并以 CAG 为平台,开发评估系统。结果:历时 7 个月,整理收集并回顾性录入 513 个病例的临床数据,建立包含 49 个临床录入项目的 50×513 维数据集(数据库)。经过一系列测试后,采用随机森林算法作为核心算法,建立预测评估模型,进而开发出预测系统(软件),为系统设置自我学习功能,并计划添加数据导出与网络化功能。结论:本评估系统在老年口腔癌患者围术期并发症发生概率的预测中具有较高的临床实用性。

[关键词] 老年人;口腔癌;数据挖掘;围术期并发症
[中图分类号] R739.8 [文献标志码] A

Establishment of an evaluation system based on data mining tech aimed for predicting the rate of perioperative complications in old oral cancer patients GAO Xian-chao, CHEN Yi-ming, YU Zhi-wei, JI Tong. *(Department of Oromaxillofacial Head and Neck Oncology, Ninth People's Hospital, College of Stomatology, Shanghai Jiao Tong University School of Medicine;Shanghai Key Laboratory of Stomatology. Shanghai 200011, China)*

[Abstract] PURPOSE: To establish an evaluation system based on data mining tech aimed for predicting the rate of perioperative complications in old oral cancer patients with better clinical application. **METHODS:** Combined APACHE Ⅱ and POSSUM system as well as clinical reality, the key factors and the logical relationship were found. By using the CAG program, the evaluation system was established. **RESULTS:** Within the period of 7 months, the clinical data of 513 patients were collected, and an data sets sized as 50×513 with 49 key factors was finished. After some tests, the Radom Forest (RF) was chosen as the main algorithm of the model of the evaluation system (program). The system had the ability for self-study, and is planed to add the function of data output and internet use. **CONCLUSION:** The evaluation system is good for clinical use to predict the rate of perioperative complications in old oral cancer patients. Supported by Grant from Shanghai Municipal Bureau of Health (2009076) and the Fifth Innovation Plan for Students of Shanghai University (IAP5145).

[Key words] Old patients; Oral cancer; Data mining; Perioperative complication
China J Oral Maxillofac Surg,2013,11(2):145-149

口腔癌的发生率与年龄呈正相关性，老年人由此成为口腔癌的高发人群。手术治疗是目前治疗口腔癌最有效的方法之一。因此,准确预测围术期并发症的发生概率,对于提高患者治愈率、改善患者预后具有积极的作用。数据挖掘技术的引入,使我们可以在海量数据中找出与围术期并发症发生概率相关因素之间的潜在逻辑关系。基于此项技术,项目组开发“基于数据挖掘技术的老年口腔癌患者围术期并发

[收稿日期] 2012-11-19; [修回日期] 2013-01-20
[基金项目] 上海市卫生局资助项目(2009076);
第五期上海市大学生创新性实验项目(IAP5145)
[作者简介] 高宪超(1983-),男,硕士研究生,
E-mail:gaoxianchao2004@yahoo.com.cn;
陈一铭(1989-),男,硕士研究生,E-mail:781368531@qq.com。
* 并列第一作者
[通信作者] 季彤,E-mail:781368531@qq.com

症发生概率评估系统”并制作成单机版软件，提供包括数据录入、概率预测、自我学习在内的多项功能，以便于本学科临床数据的回顾性录入及病例的前瞻性评估。

1 材料与方法

1.1 临床资料收集

临床资料取自2006年1月—2011年12月间上海交通大学医学院附属第九人民医院口腔颌面外科收治的口腔癌患者。纳入标准：年龄>60岁，全麻手术治疗，组织病理学诊断为口腔鳞癌。

1.2 围术期并发症的判定标准及手术分级

术前并发症参照成人并发症合并量表27（Adult Comorbidity Evaluation-27，ACE-27），选取包括心血管、胃肠道、呼吸、肾脏、内分泌、免疫系统、神经、精神类疾病、伴发的恶性肿瘤、风湿性疾病等27项病变或症状。术后并发症参考Copeland[1]的定义与临床实际，添加皮瓣术后相关并发症。病例在总体上分为2大类：一类有并发症、另一类无并发症。

将POSSUM评分系统中的手术分级标准与临床实际相结合，将手术分为小、中、大及特大4级：小手术，仅口腔内病灶局部切除；中手术，口腔内病灶扩大切除合并单侧颈淋巴清扫术（unilateral neck dissection）；大手术，口腔内病灶扩大切除合并双侧颈淋巴清扫术（bilateral neck dissection）或同期皮瓣修复术（simultaneous flap repair）；特大手术，口腔内病灶扩大切除术合并显微外科血管吻合（microsurgical vascular anastomosis）或颈动脉结扎。

1.3 临床因素筛选

参照国际上通用的APACHE II以及POSSUM评分体系，并结合老年口腔癌患者的临床实际，初步选取49项临床指标（APACHE II系统18项，POSSUM系统14项，新增17项）。APACHE II系统：体温（℃）、心率（beat/min）、呼吸率（breath/min）、平均动脉压（MAP）、血清钾（mmol/L）、血清钠（mmol/L）、血清肌酐（μmol/L）、血细胞比容（%）、白细胞（10^9/L）、动脉血氧饱和度（PaO_2/FiO_2）、动脉pH值、Glasgow评分、年龄评分、慢性健康评分。POSSUM评分系统：年龄（岁）、心脏征象、脉率（bpm）、收缩压（mmHg）、心电图分类、呼吸征象、Glasgow评分、血红蛋白（g/L）、白细胞（10^9/L）、尿素（mmol/L）、术前钾（mmol/L）、术前钠（mmol/L）、手术范围、手术类别、手术种数评分、总失血量、腹腔感染、恶性肿瘤等。本课题新增临床指标：性别、吸烟、嗜酒、心功能分级（NYHA）、术前血糖（mmol/L）、术前治疗史（手术治疗及放化疗）、合并症、肿瘤大小（cm）、临床分期（TNM分期）、组织病理学分级、ASA评分、手术时间、术后第1天血糖（mmol/L）等。

1.4 统计学处理

采用SPSS17.0软件包对以上指标中的计数指标进行χ^2检验和Fisher精确性检验，计量资料进行t检验，$P<0.05$为差异具有显著性。

同时使用MedCalc软件对POSSUM、APACHE II、ASA的受试者工作特征（ROC）进行检验。

2 结果

2.1 数据库建立

经过统计和筛选，共收录513例病例（原发342例，复发171例），男女性别比为267:246，年龄60~101岁。手术部位分布：唇25例，面部皮肤17例，腮腺区6例，颊部65例，口底33例，舌122例，口咽25例，颈部49例，颌骨176例。TNM分布：Ⅰ期27.3%（140例），Ⅱ期44.5%（228例），Ⅲ期17.9%（92例），Ⅳ期10.3%（53例）。组织病理学分级：Ⅰ级28.8%（148例），Ⅱ级51.1%（262例），Ⅲ级9.2%（47例），无明确分级10.9%（56例）。手术分布：极大手术、大手术、中手术分别占20.1%、50.9%、29%。256例术后并发症患者的病情分布：感染19.9%（58例），皮瓣坏死7.2%（21例），水肿15.8%（46例），血肿8.6%（25例），下肢静脉血栓2.4%（7例），伤口裂开13.7%（40例），涎瘘5.5%（16例），肺部疾病6.9%（20例），胃肠道症状3.1%（9例），精神症状2.8%（8例），心绞痛2.1%（6例）。将以上数据录入Excel，得到1份50×513维数据集（Excel表格）。

2.2 临床数据挖掘以及核心模型的建立

数据挖掘技术（DM）又称数据库中的知识发现（knowledge discovery in databases，KDD），意为从大量数据中寻找潜在有价值知识（多为某种逻辑关系）的过程。数据挖掘（DM）主要分为3大步骤：数据准备、数据挖掘、结果表达与分析[2]。

目前，数据挖掘技术常用的模式[3-5]包括分类模式（classification）、聚类模式（clustering）、回归模式（regression）、关联模式（association）、偏差模式（deviation）等。实际操作中最常用的方法包括模糊

方法（fuzzy）、粗糙集理论（rough sets）、云理论（cloud）、支持向量机（SVM）、随机森林（random forest）、旋转随机森林（rotating random forest）、贝叶斯网络（Bayesian network）、朴素贝叶斯网络（Naive Bayesian networks）等。

本项目使用以下 5 种数据挖掘方法即支持向量机（SVM）、旋转随机森林（rotating random forest）、随机森林（random forest）、贝叶斯网络（Bayesian network）及朴素贝叶斯网络（Naive Bayesian networks）。之前所得的数据集以 2:1 的比例随机分配，得到训练集和测试集，而后分别使用以上 5 种数据挖掘方法，对训练集进行数据挖掘，各得到 1 个预测模型，最后将测试集中的数据代入模型，以检测模型的可靠性与准确性。

表 1 中“ALL”（全变量模式）包括 APACHE Ⅱ 系统 14 项因素、POSSUM 系统 18 项因素以及新增因素（NEW）17 项，共计 49 项。结果表明：使用全变量进行预测的准确性优于使用单一组别变量模式进行预测所得结果，并且在全变量（ALL）模式下，随机森林算法（random forest）的准确率高于其他 4 种算法。后期实验中，各种变量模式下（图 1）随机森林

表 1. 5 种模型的预测准确率
Table 1. Accuracy of five predict models

变量	预测结果准确率（%）				
	支持向量机	随机森林	旋转森林	贝叶斯网络	朴素贝叶斯网络
ALL	83.4308	89.0838	85.9649	82.2612	75.6335
New	81.6764	87.1345	83.0409	77.7778	73.4893
Possum	79.3372	82.2612	79.1423	74.0741	71.9298
Apache II	75.4386	76.4203	76.0234	73.2943	75.8285

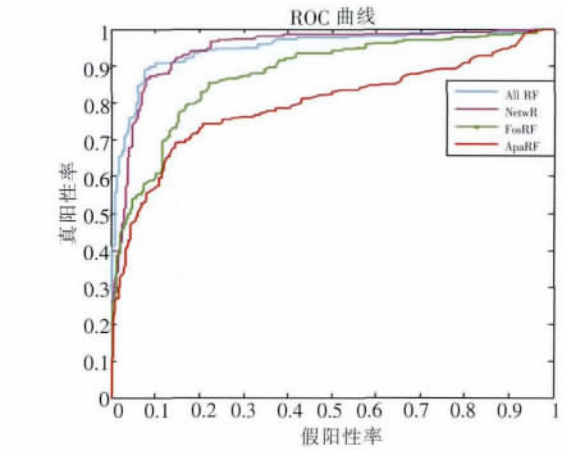

图 1. 4 种变量的随机森林模型对老年口腔癌围术期并发症的 ROC 曲线比较
Figure 1. Pairwise comparison of ROC curves of four variables based on random forest models for perioperative complications of the aged patients with oral cancer

（random forest）模型的运用及全变量模式下运用 5 种模型所得到的 ROC 曲线（图 2）均验证了此观点。

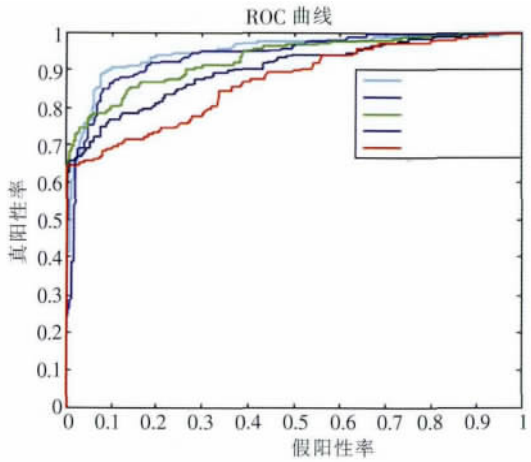

图 2. 5 种模型的全变量 ROC 曲线比较
Figure 2. ROC curves for the support vector machines, random forest, rotation forest, Bayesian network, naive Bayesian network algorithms models for perioperative complications of the aged patients with oral cancer

根据以上研究及验证结果，项目组最终选取随机森林模型作为核心模型。

2.3 统计学分析发现

手术时间、手术大小、术中失血量、肿瘤大小、临床分期、术后第 1 天血糖等因素与围术期并发症的发生有较高的相关性（$P<0.05$），并且 POSSUM 较其他 2 种评分系统对围术期并发症的预测效果更为显著（表 2、3，图 3）。

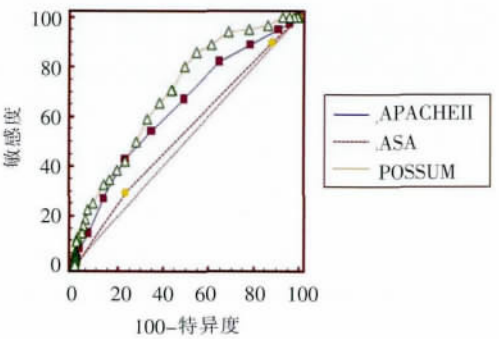

图 3. ASA、POSSUM 和 APACHEII 对老年口腔癌围术期并发症的 ROC 曲线下面积
Figure 3. Comparison for the area under the ROC curve of the ASA, POSSUM, and APACHEII

2.4 评估系统（软件）的开发

确立核心模型后，选择 CAG 程序作为开发平台，开发出一款具有传统 Windows 操作界面的老年口腔癌患者围术期并发症发生概率评估系统（软件）。软件中所有 49 项变量均以可填写空格或下拉菜单形式，呈现在软件的操作界面中（图 4），最终评

表 2. 老年口腔癌围术期并发症危险因素分析
Table 2. Analysis of the relative risk factors of perioperative complication of the-aged patients with oral cancer

因素	无并发症组	有并发症组	t 值	P 值
术前心率(次/分)	75.25±10.42	76.55±12.13	1.298	0.132
术前收缩压(mmHg)	149.14±22.62	146.38±23.36	1.358	0.843
术前白细胞计数($\times10^9$/L)	6.04±1.86	6.46±2.14	2.289	0.063
术前血红蛋白(g/L)	127.00±15.44	128.03±16.93	0.713	0.712
术前血钠(mmol/L)	143.17±8.106	135.20±6.00	1.136	0.095
术前血钾(mmol/L)	3.30±0.57	3.34±0.62	0.628	0.410
术前血糖(mmol/L)	5.13±1.01	5.23±1.10	0.942	0.957
术前尿素(mmol/L)	5.01±1.53	5.52±8.78	0.888	0.054
术后第 1 天体温(℃)	36.69±0.47	36.89±0.57	4.376	0.062
术后第 1 天心率(次/min)	80.19±10.92	82.32±10.89	2.208	0.381
术后第 1 天呼吸率(次/min)	19.38±2.38	19.20±2.33	0.868	0.412
术后第 1 天白细胞($\times10^9$/L)	11.89±4.32	13.07±4.09	2.916	0.075
术后第 1 天血钠(mmol/L)	135.58±4.59	135.56±4.70	1.136	0.095
术后第 1 天血钾(mmol/L)	3.36±0.69	3.37±0.58	0.011	0.844
术后第 1 天血细胞比容	0.35±0.04	0.33±0.04	5.303	0.065
术后第 1 天血清肌酸酐(umol/L)	72.53±20.32	74.97±23.55	0.986	0.064
术后第 1 天血糖(mmol/L)	6.83±1.01	8.23±1.10	0.982	0.042
肿瘤大小(cm)	2.65±1.54	3.49±1.86	5.526	0.000
失血量(mL)	239.73±231.40	606.29±356.31	13.084	0.001
手术时间(h)	3.19±2.06	5.74±2.69	11.996	0.001

估结果以百分比提供给使用者。考虑到数据挖掘模型的运算精度和准确性与数据库容量呈正相关,评估软件添加了自我学习功能:参与评估的新患者的临床数据将会被添加入数据库,再次使用该软件时,评估系统将基于更新后的数据库进行全程运算。

表 3. 老年口腔癌围术期并发症的危险因素分析(χ^2 检验、Fisher 精确性检验)
Table 3. Chi-square and Fisher's exact test for the risk factors of the perioperative complication of the -aged patients with oral cancer

因素	无并发症	有并发症	P 值
原发			
是	169	165	
否	84	95	0.242
基础疾病			
有	29	150	
无	224	110	0.000
吸烟			
有	47	64	
无	206	196	0.060
酗酒			
是	32	35	
否	221	225	0.444
术前手术治疗			
是	75	84	
否	178	176	0.289
术前放射治疗			
是	14	31	
否	239	229	0.008
术前化疗			
是	14	30	
否	239	230	0.011
手术大小			
中手术	119	30	
大手术	125	136	
特大手术	9	94	0.000
临床分期			
I 期	105	35	
II 期	107	121	
III 期	29	63	
IV 期	12	41	0.000
病理学分级			
I 级	84	64	
II 级	125	137	
III 级	23	24	0.205

图 4. 软件截图
Figure 4. Screenshot of the program

3 讨论

“基于数据挖掘技术的老年口腔癌患者围术期并发症发生概率评估系统”的建立，不仅使得相关并发症的发生概率预测变得更加直观，同时也在该学科范畴内首次建立了大样本老年口腔癌患者临床数据库。

数据挖掘的基础是数据库，数据库的 2 大核心要素是样本容量与样本质量。由于原始数据仅局限于本院，数据库的规模较小，因此，新病例预测的精度受到一定影响，初次评估值低于理论预期。为了有效克服上述缺陷，后期试验中，将尝试与国内多家医院合作，共同采集病例数据，扩大数据库规模；进一步完善自我学习功能；为软件添加“库中数据导出”功能，以便使用者阶段性分析数据库中的数据成分，进行相关临床及流行病学分析；建立网络化临床数据收录与预测平台，使得不同地区、不同国家的医务工作者均可使用这款软件。

随着数据库规模的不断扩大和软件功能的不断完善，该评估系统的预测精度将会进一步提高，并发症发生概率的评估也将会变得更加科学与直观。

利益冲突声明：无。

[参考文献]

[1] Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical adult[J]. Br J Surg, 1991, 78(3): 355–360.

[2] Houston AL, Chen H, Hubbard SM, et al. Medical data mining on the internet: research on a cancer information system[J]. Artif Intell Rev, 1999, 13(5–6): 437–466.

[3] Hart JW, Micheline K. 数据挖掘概念与技术 [M]. 范明, 译. 北京: 机械工业出版社, 2001.

[4] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. Proceedings of the 20th International Conference on Very Large Data Bases[C]. Santiago: Morgan Kaufmann,1994:487–499.

[5] Agrawal R, Sharer JC. Parallel mining of association rules [J]. IEEE Trans Knowl Data Eng, 1996, 8(6): 962–969.