面向临床路径的病案首页数据挖掘分析

王春容 曾宇平 刘元铃 刘伟萍 方 鸣

（广州中医药大学第一附属医院 广州 510405） （广东省中医院 广州 510405） （广东省妇幼保健院 广州 511400） （广州中医药大学第一附属医院 广州 510405）

〔摘要〕 通过研究病案首页统计信息中各项研究变量的分布规律与特点，探讨各研究变量对是否将病历纳入临床路径的贡献。介绍数据来源和预处理过程，阐述研究方法的选择和并具体分析挖掘过程。建立研究变量 Logstic 回归方程，找出纳入临床路径概率高的病历特征。

〔关键词〕 临床路径；病案首页；数据分析；数据挖掘

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673－6036. 2016. 03. 012

Data Mining Analysis for the Clinical Pathway of Medical Records First Page Information

WANG Chun－rong, The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510405, China; ZENG Yu－ping, Guangdong Province Traditional Chinese Medical Hospital, Guangzhou 510405, China; LIU Yuan－ling, Guangdong Provincial Maternity and Child Care Center, Guangzhou 511400, China; LIU Wei－ping, FANG Ming, The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510405, China

〔**Abstract**〕 By studying the distribution rules and characteristics of research variables in statistical information on first pages of medical records, the paper discusses the contribution of each research variable to solving the problem whether medical records should be included in clinical pathways. It introduces data sources and the preprocessing process and explains the selection of research method and specific process of data analysis and mining. Then, it constructs the Logstic regression equation of research variables and finds out features of medical records which have a high probability to be included in clinical pathways.

〔**Keywords**〕 Clinical pathway; Medical records first page; Data analysis; Data mining

1 引言

临床路径（Clinical Pathway，CP）是指针对某一疾病建立一套标准化治疗模式与程序，是一个有关临床治疗的综合模式，以循证医学证据和指南为指导来促进治疗和疾病管理，最终在有限的卫生资源的条件下，通过规范诊疗过程、标准化诊疗行为，达到改善医疗质量、提高医疗效率、增进医疗效益的目的。病案首页浓缩了整份病案中最重要的内容，是病案信息的核心部分。如何充分挖掘和利用病案首页信息，探索纳入临床路径的病历特征规律，为临床及管理决策提供科学依据，是医院病案信息管理的重要课题。本文拟通过对病案首页数据的分析挖掘，探讨临床路径病历的特征与规律，从而服务于临床与管理[1]。

〔修回日期〕 2016－01－25

〔作者简介〕 王春容，硕士，技师，发表论文 4 篇；通讯作者：曾宇平。

2 资料来源与数据预处理

2.1 资料来源

通过病案统计管理系统，选择出院日期为2010.1.1－2014.12.31 共 5 年的病人病历，总共123 927份，其中纳入临床路径的病历 13 621 份，约占总病历的 10.99%，未纳入临床路径的病历109 666份，约占总病历的 88.49%，未知是否纳入临床路径的病历 640 份，约占 0.52%。本文以纳入和未纳入临床路径的 123 301 份病历为挖掘源数据，进行数据分析与挖掘，未知病历忽略不计。

2.2 数据预处理

2.2.1 选择研究变量 经初步整理发现，挖掘的源数据涵盖病人 3 部分内容，即基本情况、医疗情况、重要统计数据或管理指标，总共 338 个字段。根据各数据字段的取值特征，排除有较多空值及数据不规范的变量，选取与本次研究可能相关的 15 个字段，分别为病案号、姓名、住院天数、入院科室、出院科室、是否部分病种、是否抢救病人、是否 3 日确诊、是否月内再次住院、是否手术、临床路径病历、住院期间是否出现危重、住院期间是否出现急症、住院期间是否出现疑难、国际疾病分类（International Classification of Disease，ICD）编码等。

2.2.2 生成新的研究变量 通过公式对入院科室和出院科室进行对比，形成新的研究变量是否转科，公式为：if（入院科别 = 出院科别）then 否 else 是 endif。如果入院科室等于出院科室，则新变量是否转科取值为否，反之取值为是。同时，为统一研究疾病类别和识别的精确度，选取 ICD 码作为疾病病种的类别，截取保留小数点后 1 位，小数点后一位前的数值相同可视为同一病种，截取公式为 Substring（1，5，ICD 码），从而形成新的研究变量 ICD_ 新。

2.2.3 数据处理与规范 经初步整理发现，选取的研究变量中部分记录含有空值和错误值，因研究变量均为分类变量，无法通过观察判断对空值和错误值进行补充或替换，且含有空值和错误值的记录较少，故对其进行过滤，不再作为数据挖掘的基础数据。过滤后的出院病人病历数量为 123 287 份，过滤病历 14 份，占总研究病历的 0.011%，故可忽略不计。另外，对 ICD_ 新进行转换，由数值范围型转换为分类变量，从而符合数据挖掘算法的需求。

3 数据分析与挖掘

3.1 研究方法的选择

Logistic 模型又称 Logistic 回归分析，主要用于通过自变量预测因变量发生的概率趋势。本次数据分析与挖掘旨在通过研究统计信息中各项研究变量的分布规律与特点，探讨各研究变量对是否将病历纳入临床路径的贡献，为临床路径病人的选择与决策提供一定数据支持[2]。通过整理发现，研究变量均为分类变量，故本次研究以“临床路径病历”作为因变量，以是否部分病种、是否抢救病人、是否 3 日确诊、是否月内再次住院、是否手术、住院期间是否出现危重、住院期间是否出现急症、住院期间是否出现疑难、是否转科作为自变量，ICD_ 新作为病种分区变量，采用 SPSS Clementine 软件中的二项 Logistic 回归分析模型对数据进行分析和挖掘。数据挖掘流程，见图 1。

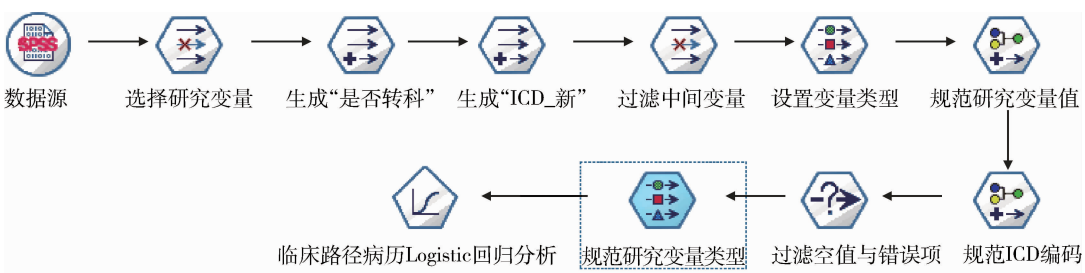

图 1 数据挖掘流程

3.2 数据分析与挖掘

3.2.1 研究变量分布规律分析 以是否纳入临床路径为标准，对纳入临床路径和未纳入临床路径的病历的研究变量及病种进行分布趋势分析，见表1，表2。由表1对比分析可知，研究病历中病种分布顺序和所占比例大致相同，研究病种分别为肿瘤化学治疗疗程、肿瘤特指医疗照顾、支气管肺炎、脑梗死、乳腺良性肿瘤、2 型糖尿病、高血压性肾衰竭、内痔、类风湿性关节炎、原发性高血压、肺特指疾患、先兆流产和特指骨坏死。由表2对比分析可知，纳入和未纳入临床路径的病历，其研究变量是否抢救病人、是否转科、住院期间是否出现急症、住院期间是否出现危重、是否3 日确诊、是否月内再次住院、是否部分病种、住院期间是否出现疑难、是否手术等研究变量的真值所占比率大致相同，分布规律相近，符合 Logstic 回归分析的数据要求。

表1 病历病种分布（前10 种）

疾病编码	疾病名称	未纳入临床路径		纳入临床路径		合计	占总病历
		百分比（%）	计数	百分比（%）	计数		百分比（%）
Z51.1	肿瘤化学治疗疗程	7.82	8 580	7.67	1 045	9 625	7.81
Z51.8	肿瘤特指医疗照顾	3.98	4 368	6.83	930	5 298	4.3
J18.0	支气管肺炎	2.07	2 266	0.26	35	2 301	1.87
I63.9	脑梗死	1.41	1 544	3.03	413	1 957	1.59
D24.x	乳腺良性肿瘤	1.34	1 469	2.84	387	1 856	1.51
E11.9	2 型糖尿病	1.11	1 214	4.16	566	1 780	1.44
I12.0	高血压性肾衰竭	1.28	1 409	1.84	250	1 659	1.35
I84.2	内痔	0.9	995	4.52	615	1 610	1.31
M06.9	类风湿性关节炎	1.08	1 179	2.97	404	1 583	1.28
I10.x	原发性高血压	1.02	1 122	0.83	113	1 235	1

表2 病历各研究变量分布

研究变量	未纳入临床路径		纳入临床路径	
	百分比（%）	计数	百分比（%）	计数
是否抢救病人	2.44	2 676	2.12	289
是否转科	4.55	4 988	3.92	534
住院期间是否出现急症	5.8	6 365	4.16	567
住院期间是否出现危重	5.08	5 567	4.56	621
是否3 日确诊	5.48	6 014	6.62	902
是否月内再次住院	6.7	7 350	6.67	909
是否部分病种	11.74	12 880	8.49	1 157
住院期间是否出现疑难	10.58	11 608	11.04	1 504
是否手术	35.3	38 716	34.21	4 660

3.2.2 Logstic 回归分析 表3 可知 Logstic 回归方程显著性检验的总体情况，概率 $-p$ 值小于显著性水平 α（0.05），故应拒绝零假设，认为所有回归系数不同时为0，变量的全体与 Logit p 之间的线性关系显著，采用该模型合理。表4 可知各变量的 wald 检验概率均为.000，小于显著性水平 α（0.05），故各系数均可保留在方程中，该模型合理[3]。

表3 Logstic 回归模型系数的综合检验

类型	卡方	df	显著性
步骤	14.379	6	0.026
块	14.379	6	0.026
模型	14.379	6	0.026

表 4 影响是否纳入临床路径的病历相关因素的 Logistic 回归分析结果

变量	B	S. E.	Wald	df	显著性
是否抢救病人	-21.440	27 209.155	0.000	1	0.999
是否3 日确诊	20.902	6 780.763	0.000	1	0.998
是否手术	-1.422	18 284.244	0.000	1	1.000
住院期间是否出现急症	-20.902	6 780.763	0.000	1	0.998
住院期间是否出现疑难	-19.684	9 402.610	0.000	1	0.998
是否转科	-1.435	18 186.435	0.000	1	1.000
常量	-17.366	36 918.393	0.000	1	1.000

4 讨论

4.1 结果分析

4.1.1 指标对病历是否纳入临床路径的贡献 由上述分析结果可知，住院期间是否危重和月内是否再次住院两研究变量未纳入 Logstic 回归分析方程，即本次研究中，这两项指标对病历是否纳入临床路径贡献小。是否 3 日确诊与病历是否纳入临床路径为正相关，是否抢救病人、住院期间是否出现疑难、是否手术、住院期间是否出现急症和是否转科与病历是否纳入临床路径为负相关，且从影响因素所占比重看，是否抢救病人、住院期间是否出现疑难、是否 3 日确诊和住院期间是否出现急症这 4 个因素对病历是否纳入临床路径贡献较大。即如果病历为非抢救病历、住院期间没有出现疑难，且 3 日内能确诊，没有进行手术、没有转科、住院期间没有出现急症则病历纳入临床路径的概率大。

4.1.2 本研究的科学性及不足 临床路径的核心是将某种疾病（手术）关键性的检查、治疗、护理等活动标准化，确保患者在正确的时间、地点得到正确的诊疗服务，以期达到最佳治疗效。临床路径病历的选择，要结合病种的实际情况，选择治疗手段明确、疗效较好且结果可预测、发症少的常见病和多发病实施。本次研究的结果与病历纳入临床路径的实际情况相符，结果具有科学性。不足之处在于，由于病案首页中收集的是否并发症、疗效、是否单病种等指标数据不规范或空值较多，未纳入到研究变量范围，未能研究其与是否纳入临床路径病历的关系，故应做好这些数据指标的规范工作，为日后数据分析和挖掘提供可靠的数据保证。另外研究中涉及的 ICD 选取最新的 ICD－10 作为编码标准。其中，ICD 编码作为病种分区变量，对本次研究结果起到至关重要的作用。通过对本次研究数据的分析发现，医院病历中的疾病 ICD 编码存在不够准确、诊断书写不够规范、不能满足临床路径的需要等问题。本次研究为忽略误差，统一研究疾病类别和识别的精确度，保留了 ICD 编码小数点后 1 位，对病种以大类分类进行数据挖掘分析，得出了本次研究结果。但为更好地对研究病历的病种进行细分并进行针对性的分析与挖掘，医院应加强对病历质量的控制、重视并加强对研究病历的编码规范，务必做到病必编码，编则必准，以期为医院日后的数据分析及挖掘提供可靠、可用、准确的第一手分析挖掘素材，实现更加准确、全面、客观的数据分析与挖掘。

4.2 研究的扩展与深入

临床路径应用于某个单病种，应综合考虑纳入病历所属病种的临床路径纳入标准、排除标准和可能发生的变异等，在应用的实践中不断修订和改进，使之更加合理有效，达到既控制医疗费用，又保证医疗质量的整体效果。为使研究进一步深入，在本次研究基础上，综合考虑病种的入径标准及条件，对纳入研究的病历进行分组分类，增加各类费用（如总费用、药费、检查费、治疗费等）为因变量，充分考虑每个病种的疾病分型以及其他对费用产生影响的因数，使研究方案更加灵活多样，不单纯执行一种标准，样本更加丰富准确，从而有针对性地进行分析和挖掘。控制医疗费用是实施临床路

径的目标之一。以本次研究为基础，在入径病历病种细分的基础上，以医疗费用作为研究对象，其他可能的影响因素如住院日、药占比、麻醉方式、年龄、付款方式、手术以及各类检查与治疗等为自变量，可深入分析挖掘纳入临床路径及未纳入临床路径的病历的费用差异，从而为医疗费用的控制以及医院医疗质量的深入管理提供可靠的数据支撑[4]。

5 结语

临床路径已经被证明是持续改进医疗质量、控制医疗成本、优化服务流程的有效途径，也是我国现阶段推进医疗体制改革的重要举措。电子病历作为现代医疗机构开展高效、优质的临床诊疗、科研以及管理工作所必需的重要临床信息资源，已在各级医疗机构中推广和普遍应用。在国家卫计委制定并下发《电子病历基本架构与数据标准》、《电子病历系统功能规范（试行）》等系列标准后，电子病历系统的应用得到了进一步规范并积累了大量来自于临床实践、针对患者的诊疗信息数据，使得通过大数据及数据挖掘等成熟的现代信息技术，在循证医学理论指导下科学、高效地构建、评估和管理高质量的临床路径成为可能。本研究在此背景下，立足于电子病历的病案首页数据信息，面向临床路径进行数据分析合理可行且具有重要意义[5-7]。

参考文献

1 王毅，李礼安，莫远明，等. 临床路径系统设计与应用［J］. 医学信息学杂志，2013，(10)：24-27.

2 韩萍，陆琴，藏逗. Logstic 回归方法分析儿童哮喘临床护理路径变异相关因素研究［J］. 护士进修杂志，2014，29（11）：965-967.

3 虞海燕，李劲松，曹淑真，等. 基于DeepSee 的医院药库数据挖掘［J］. 中国数字医学，2010，(10)：34-38.

4 李红，康楠，马立旭. 单病种临床路径纳入标准探讨［J］. 中国医院统计，2014，21（5）：348-350.

5 李慧玲，杨小平，宇文姝丽. 支持临床路径的电子病历系统开发设计［J］. 医学信息学杂志，2011，32（3）：13-18.

6 王毅，李礼安，莫远明等. 临床路径系统设计与应用［J］. 医学信息学杂志，2013，(23)：24-27.

7 曹洪欣，蔡海英，王侠，等. EMR 适于数据挖掘构建临床路径的数据特征分析［J］. 中国医院管理，2013，33(3)：55-58.

（上接33 页）

2 吴双兵，刘传. 网上预约挂号系统设计与实现［J］. 医学信息学杂志，2015，(1)：36-39.

3 朱妍昕，邱君瑞，徐维. 医学信息学检索与利用教学网站设计与实现［J］. 医学信息学杂志，2012，(1)：86-88.

4 王修来，吴美娟，张丽丽. 虚拟医院建设的探索与实践［J］. 医学信息学杂志，2011，(12)：23-25.

5 胡建理，李小华，周斌. 一种基于安全隔离网闸技术的医院内部网安全解决方案［J］. 医疗卫生装备，2010，(31)（7）：44-45，5.

6 张骁，李红信. 信息安全建设中的隔离网闸技术应用研究［J］. 山西师范大学学报（自然科学版），2010，(2)：43-47.

7 万美. 大数据时代的公共卫生信息安全［J］. 医学信息学杂志，2012，(12)：56-58.

8 陈德玉. 不再谈癌色变——浅谈肿瘤预防与控制新概念［J］. 医药与保健，2007，(12)：28-29.

9 赵娟，李锋，李思源，等. 生物样本库的建立与管理［J］. 现代生物医学进展，2010，(5)：34-35.

10 金爱山，韩爽，申展. 肿瘤患者随访信息平台建设与作用［J］. 医学信息学杂志，2012，(3)：25-27.