文本挖掘在医疗器械不良事件报告中的应用*

施雯慧 林 洁 孙志明 姚 捷 许豪勤

（江苏省计划生育科学技术研究所 南京 210036）

〔摘要〕 基于美国食品药品监督管理局公共数据开放平台上的宫内节育器不良事件报告数据，运用词频分析、主题模型等文本挖掘方法分析其中的非结构化文本变量，探讨该方法应用的可行性并指出存在的问题。

〔关键词〕 文本挖掘；宫内节育器；不良事件报告；主题模型

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.12.014

Application of Text Mining in the Adverse Events Report of Medical Instruments *SHI Wen－hui, LIN Jie, SUN Zhi－ming, YAO Jie, XU Hao－qin, Jiangsu Family Planning Research Institute, Nanjing 210036, China*

〔**Abstract**〕 Based on the data of adverse events reports about intrauterine device on the public data opening platform of the Food and Drug Administration of the USA, the paper uses the text mining methods such as word frequency analysis, topic model and so on to analyze the non－structured text variables in the data, discusses the application feasibility of this method and points out the existing problems.

〔**Keywords**〕 Text mining; Intrauterine device; Adverse event report; Topic models

1 引言

医疗器械不良事件是指获准上市的质量合格的医疗器械在正常使用情况下发生的，导致或者可能导致人体伤害的各种有害事件。通过对医疗器械使用过程中出现的可疑不良事件进行收集、报告、分析和评价，对存在安全隐患的医疗器械采取有效的控制，防止医疗器械严重不良事件的重复发生和蔓延，保障公众安全。医疗器械不良事件报告是医疗器械监测工作的重要信息来源。纵览国际协调工作组各成员国（美国、欧盟、加拿大、澳大利亚和日本）的报告表格，尽管在设计样式、填报要求上各不相同[1]，但收集的信息内容不外乎报告来源、医疗器械的种类、不良事件的表现、采取的措施等几部分，而其中不良事件情况作为关联性评价的主要依据之一，在报告表中占有较高的权重[2-3]。一般而言，不良事件陈述变量主要是描述性的文字，而非定量的结果或定性的选项，这就给后续基于大量数据的分析带来了一定困难。文本挖掘技术为处理这类非结构化数据提供了解决方法，其本质是通过自然语言处理，将文本转化为数据进行分析，目前已广泛应用于商业智能、信息检索、生物医学信息

〔修回日期〕 2016－09－22

〔作者简介〕 施雯慧，实习研究员，发表论文 6 篇；通讯作者，许豪勤，研究员，副主任医师。

〔基金项目〕 江苏省计划生育科研所科研启动基金项目“信号检测方法在宫内节育器不良事件监测中的应用研究”（项目编号：JSFP2015004）。

处理等领域。

美国医疗器械不良事件数据库收集了制造商、进口商和使用单位提交的强制报告以及个人提交的自愿报告，数据对公众开放；但当前其报告表中仅有事件描述变量（详细的文本信息），缺少简明扼要、具有总结性质的不良事件主要表现变量[4]。本文试图通过本文挖掘方法对宫内节育器相关的不良事件进行分析，以期探讨该技术在医疗器械不良事件报告中应用的可行性。

2 资料与方法

2.1 资料来源

在美国食品药品监督管理局（Food and Drug Administration，FDA）公共数据开放平台（OpenFDA）上检索数据库收录的所有与宫内节育器相关的不良事件报告，检索条件：产品编码为“HDT”（检索 FDA 产品分类数据库可知，宫内节育器的产品编码为 HDT，规定编码为 21 CFR 884.5360）或者商品名包含"Paragard","Mirena"（FDA 目前批准上市的宫内节育器仅有这两种）或者器械通用名包含"intrauterine device","iud"。

2.2 方法

2.2.1 数据获取 根据 OpenFDA API 查询检索式的构建方法[5]，上述检索条件可转化为如下检索 API 地址：https://api.fda.gov/device/event.json?search=(device.generic_name:intrauterine+AND+device.generic_name:device)+device.generic_name:iud+device_report_product_code:%22HDT%22+device.brand_name:Paragard+device.brand_name:mirena&limit=88)。利用 R 软件的 + jsonlite 包将返回的 JSON 格式数据进行转换，对其进行人工筛选，去除与宫内节育器不良事件无关的报告，形成可供分析的数据集。

2.2.2 一般性描述 对报告中的常规分类变量，如接收时间、报告类型、来源、事件类型及发生场所、涉及节育器商品名等进行统计。

2.2.3 文本挖掘 首先提取数据集中不良事件描述变量，利用 tm 包中的 Corpus（）函数读入 R 软件，形成语料库；然后使用 tm_map（）函数对语料库进行去除停用词（为节省空间、提高效率，在处理自然语言数据时会自动过滤掉某些字或词，被称为停用词，多为没有实际含义的功能词如介词、冠词、连词等[6]）、标点符号、多余空白、小写化、词干化等预处理；得到较合适的分词结果后，使用 DocumentTermMatrix（）函数生成文档 - 词条矩阵，转换成数据框，利用主题模型算法[7]进一步挖掘，相关分析利用 R 软件中的 topicmodels 包实现。

3 结果

3.1 检索结果

在数据库中共检索到符合条件的不良事件报告 88 例，进行人工筛选后，发现有 18 例涉及 Essure（一种绝育术工具）、Implanon（一种皮下埋植剂）或产品编码、通用名、商品名前后矛盾的，予以剔除，最后得到宫内节育器不良事件报告 70 例。

3.2 一般性描述

报告接收日期分布于 1993 - 2016 年间，其中 2015 年收到 13 例（18.57%）报告，为历年最多，见图 1。按报告类型分，见表 1。

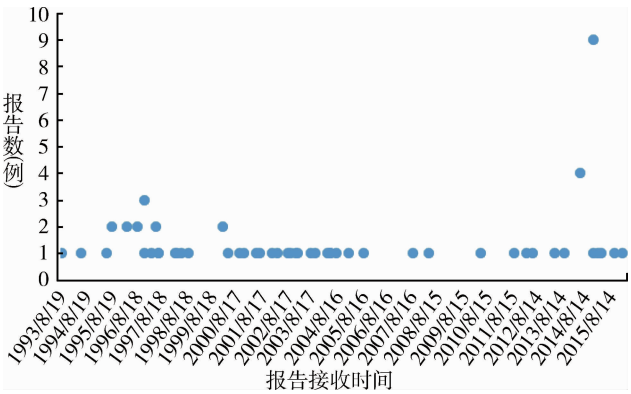

图 1 宫内节育器不良事件报告接收日期分布

表 1 宫内节育器不良事件报告分类统计

分类		频数	百分比（%）
报告接收时间	1993 – 2000 年	26	37.14
	2001 – 2010 年	21	30.00
	2011 – 2016 年	23	32.86
报告类型	仅初始报告	69	98.57
	初始报告和跟踪报告	1	1.43
报告来源	生产企业	16	22.86
	使用单位	37	52.86
	个人	17	24.29
事件类型	伤害	32	45.71
	器械故障	25	35.71
	其他	5	7.14
	不明	8	11.43
事件发生场所	家	2	2.86
	医院	22	31.43
	住院诊断机构	1	1.43
	其他	15	21.43
	不明	30	42.86
器械商品名	Dalkon Shield	3	4.29
	Lippes Loop	19	27.14
	Mirena	10	14.29
	Paragard	21	30.00
	Progestasert	1	1.43
	不明	16	22.86
报告人职业	律师	2	2.86
	生物医学工程师	1	1.43
	护士	2	2.86
	患者	11	15.71
	药师	1	1.43
	医师	6	8.57
	风险管理师	30	42.86
	其他	5	7.14
	不明	12	17.14
合计	–	70	100.00

3.3 不良事件描述文本词频分析

3.3.1 文档 – 词条矩阵 从分析数据集中提取不良事件描述变量，得到 71 条文本记录。由于美国医疗器械不良事件报告表中，不良事件描述文本分为“事件/问题描述”和“生产企业附加描述”两类，1 份报告可能涉及不止 1 条描述文本，因此将同份报告的描述文本合并后，得到 63 条记录。使用词频模型生成文档 – 词条矩阵，基于 63 个文档共提取出 745 个单词，其中非零元素 1 820 个、零元素 45 115 个，矩阵稀疏率达 96%。考虑到单词忽略了结构及上下文之间的联系，也无法处理否定词，因此采用二元文法（文本里 2 个单词的组合）再次生成矩阵，结果基于 63 个文档共提取出 1 840 个单词组合，矩阵稀疏率达 98%。按词频降序排列，位于前 20 位的单词和单词组合，见表 2。

表 2 文档 – 词条矩阵频次（基于单词和基于二元文法）

单词	出现频次	单词组合	出现频次
iud	139	iud remov	24
patient	111	remov iud	16
remov	79	iud insert	15
devic	48	paragard iud	12
report	47	admit hospit	9
pain	33	brand provid	9
insert	32	manufactur respons	9
state	29	patient admit	9
year	25	provid site	9
surgeri	18	respons paragard	9
uterus	18	site report	9
abdomin	17	female patient	8
physician	17	invalid data	8
sever	17	patient report	8
hospit	16	report leadership	8
string	16	abdomin pain	7
bleed	15	devic remove	7
left	15	iud iud	7
provide	15	iud t380a	7
hysterectomi	14	t380a brand	7

3.3.2 不良反应/事件词语的关联词 在结果矩阵中，可利用 findAssocs（）函数获取词条间的关系。按照词频顺序，选择前 10 个与宫内节育器不良反应相关的名词，查找相关系数≥0.6 条件下各名词的关联词条，见表 3。

表 3 前 10 位不良反应/事件词语的关联词（按相关系数降序排列）

不良反应/事件描述词	关系数≥0.6 的关联词（降序排列，前 10 位）									
	1	2	3	4	5	6	7	8	9	10
pain	alert	appear	chest	convers	estrogen	experi	fatigu	fain	gain	head
surgeri	small	acut	adenomyosi	admiss	antimesenter	appendectomi	area	assoc	aug	bacteri
uterus	inflamm	ileum	segment	bowel	section	therapi	cyst	exploratori	addomin	–
abdomin	exploratori	inflamm	therapi	emerg	vomit	ileum	bowel	laparotomi	show	colon
bleed	irregular	–	–	–	–	–	–	–	–	–
hysterectomi	chronic	cyst	salping	ovari	underw	follow	–	–	–	–
lippesloop	add	chang	cycl	knowledg	system	doubl	occur	–	–	–
paragard	arm	leadership	represent	respons	site	copper	brand	miss	manufactur	–
pelvic	unspecifi	histori	diseas	colostomi	obstruct	rectal	inflammatori	american	ampicillin	frozen
mirena	alert	appear	chest	convers	estrogen	experi	fatigu	fine	gain	head

3.4 主题模型

开始建模前，需要在文档－词条矩阵中筛选高频词并确定主题的数量。主题模型的目的是对文本进行分类，在大多数文本中出现频次都很低的词贡献并不大，但会显著地占用计算资源。通过计算得知，当前矩阵中的词频－逆向文件频率（TF－IDF）[8]中位数为 0.113 7，因此选取 TF－IDF 值≥0.12 的单词，筛选后的矩阵为 63 个文档、341 个单词。常用确定主题数的方法是计算模型的困惑度，一般困惑度越小，模型越优[9]。将主题数设置为 2～40 迭代计算困惑度，见图 2，随主题数的增加，困惑度一直呈下降趋势，这表明主题数越多，模型效果越好。而模型参数 α 和信息熵与主题数的关系，见图 3，在主题数为 6 时，信息熵达到最大值，比较此处的困惑度，困惑度的变化率越来越小，不妨认定主题数为 6 时是最优模型。

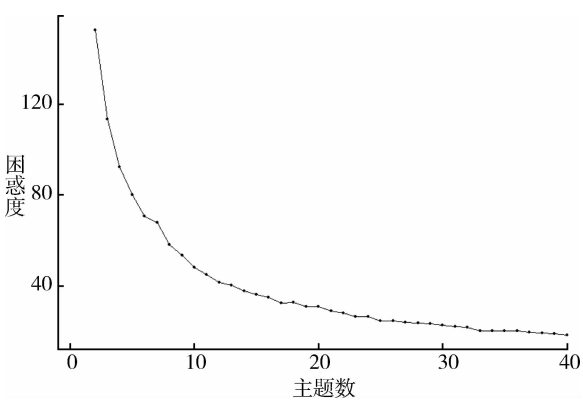

图 2 模型困惑度与主题数的关系

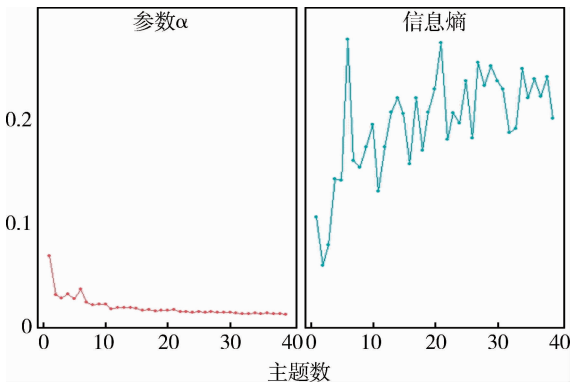

图 3 模型参数 α、信息熵与主题数的关系

4 讨论

4.1 报告总体情况

从对报告的一般性描述来看，美国与宫内节育器相关的不良事件报告数量极少，与其实际使用情况有关。Guttmacher 研究所的数据显示[10]，2012 年全美 15～44 岁女性中 6.4% 使用 IUD 避孕（约 388 万）。报告来源主要是使用单位而非生产企业，与医疗器械不良事件报告的总体来源主要是生产企业有明显差别。涉及的 IUD 种类主要有 Paragard、Lippes Loop、Mirena、Dalkon Shield 等，其中 Paragand、Mrinena 是目前 FDA 批准的在美国上市的宫内节育器，而 Lappes Loop、Dalkon Shield 则分别于 1962 年和 1971 年上市，是当时被广泛使用的宫内节育器品种，后者更因发生的一系列不良事件而引起美国国内对使用 IUD 的恐慌，导致使用人数急剧下降。

4.2 文本挖掘结果分析

对不良事件描述变量（文本结构）的词频分析显示，排名前 20 位的单词主要涉及“宫内节育器”、“病人”、“放置”、“取出”、“报告”、“描述”等，仅“疼痛”、“手术”、“子宫”、“腹部”、“出血”、“子宫切除术”与不良反应相关；前 20 位的单词组合也表明，高频出现的是“放置 IUD”、“取出 IUD”、“入院”、“生产厂家回应”等模式化描述性文字，与不良反应特异性高相关的词组较少。但基于词频的分析仍能体现出报告的总体特征：（1）“病人”一词出现频率高，表明报告人多是医务人员或从事不良事件风险管理的人员。（2）报告中多对放置节育器情况有描述，且大部分事件中取出了宫内节育器。（3）在所有不良事件中，疼痛最为常见，其次为“出血”和“子宫切除”。

依词条关系提示了与文本中最常出现的每个不良反应相关词关系最为紧密的一组词语，其中与“疼痛”关系较密切的涉及身体部位的词语有“胸部”、“头部”，表明可能为使用者不适症状的描述；与“手术”相关的词语有“急性”、“子宫腺肌症”、“小肠系膜对侧”等；与“腹部”关联的词语有“剖腹探查”、“炎症”、“呕吐”等；与“出血”最相关的词语是“不规则”，表明不规则出血是较常见的不良事件；与“盆腔”关联的词语有“结肠造口术”、“炎症”、“氨苄西林”等，表明可能在描述盆腔炎等症状或导致的后果（行结肠造口术）。对报告中涉及最多的 3 种节育器也进行关联词分析，结果提示：与“Paragard”（TCu380A）关联的词语有“（节育器）横臂”、“铜”、“品牌”、“缺失”等，表明 TCu380A 较常出现横臂缺失的情况；与“Lippes Loop”关联的词语有“增加”、“改变”、“回转”、“双（double）”等，这是由于该节育器的全称为“Lippes Loop Double S”，词组 Lippes Loop 常与 Double 连用，“改变”、“回转”等可能描述节育器的形状改变；而与“Mirena”关联的词语和“疼痛”的较为一致，且这两个词互为对方的关联词，可能是常同时出现在一个文本中所致。

主题模型算法将文本分为 6 个主题，主题 1 与“无效数据”、“妊娠试验阳性”有关，主题 2 与“检查”、“含铜的”、“萨尔瓦多”有关，主题 3 与“黏连”、“慢性”、“盆腔”有关，主题 4 与“严重”、“曼月乐”、“腹部”有关，主题 5 与“白色塑料”、“效果”、“植入”、“腹腔镜检查”有关，主题 6 与“（节育器）臂”、“缺失”、“铜”、“缠绕”有关。除主题 1、3、6 中涉及的词语的主题性比较强外，其余主题词语对文本的代表性并不理想。

4.3 存在问题

在对分析文本进行预处理时发现存在如下问题：（1）使用约定俗成的缩写形式代替正常单词的情况，如用 pt/pts 表示 patient、yr/yrs 表示 year、hosp 表示 hospital、rptr 表示 reporter、rep 表示 representive 等。（2）拼写错误，如 laparoscopic 错拼成 laparascopic、copper 错拼成 cooper 等。（3）医学术语多，由于文本内容涉及不良事件/反应，一般性用语中掺杂大量医学术语。（4）同义词多，如 bleeding、hemorrhage 都表示“出血”，pain、cramp、spasm 都意为“疼痛”等。上述问题（1）（2）理论上可以通过 content_ transformer（）函数规整来解决，但人工规整工作量较大，全面性也无法得到保证；问题（3）（4）理论上可以通过引入医学术语词典解决，但目前尚未知有类似工具。

5 结语

对文本进行分类尝试的结果表明，在构建出适宜的文档－词条矩阵后，可通过训练的方式实现文本的自动分类，这对于快速了解文本的主要内容有一定帮助；但是文本挖掘技术关注的还是重要程度较高的词语，词频非常低的词语可能会在降维处理过程中被舍弃，不利于罕见的不良反应/不良事件的发现，因此需根据研究者的目的选取适宜的分析方法。

参考文献

1 张素敏，曹立亚，曾光，等. 世界各国医疗器械不良事件监测现状比较［J］. 中国医疗器械信息，2005，11（6）：52－56.

（下转第 76 页）

据挖掘等搭建医学信息服务平台，可以有效整合国内外医学信息资源，规范服务工作流程，提升医药卫生行业情报服务和决策支持的效率，形成多种功能、互为支撑的完整服务体系。通过平台的建设和推广应用，可实现情报服务人员、政府部门、科研院所、企业和社会用户之间的良性互动，充分发挥平台的信息库和智库作用，实现共赢的同时，也可为医学科技创新发展带来动力。

参考文献

1 于施洋，王建冬，童楠楠. 大数据环境下的政府信息服务创新：研究现状与发展对策［J］. 电子政务，2016，(1)：26－32.

2 王正国. 数字化时代的医学革命［J］. 中国数字医学，2009，4（1）：8－11.

3 陈锐，冯占英，李焱，等. 大数据对生物医学信息服务各环节的影响研究［J］. 图书情报工作，2015，59（9）：68－72.

4 贺德方. 数字时代情报学理论与实践—从信息服务走向知识服务［M］. 北京：科学技术文献出版社，2006.

5 李国栋. 大数据下的医学文献情报创新［J］. 中国科技信息，2015，(13)：72－73.

6 刘冬云. 云计算环境下高校图书馆数字信息资源共建共享系统的构建［J］. 图书馆学刊，2014，(11)：99－101.

7 刘海龙. 关于网络环境下提高医学信息服务质量和效益的思考［J］. 健康导报·医学版，2015，20（12）：275.

8 彭晓东，王茂林. 基于高校图书馆的企业情报服务平台研究［J］. 情报理论与实践，2011，34（2）：58－61，71.

9 于彤，张竹绿，贾李蓉. 面向循证医学的知识服务平台概述［J］. 中国中医药图书情报杂志，2014，38（8）：55－57.

10 栾旭伦. 大数据环境下高校图书馆个性化信息服务系统研究［J］. 图书馆学刊，2014，(8)：118－121.

11 乔幸娟. 数据挖掘技术在数字图书馆中的应用研究［J］. 农业图书情报学刊，2014，26（12）：118－120.

12 张文惠. 数据挖掘技术提升高校图书馆水平［J］. 电脑开发与应用，2014，27（12）：49－51，54.

13 黄东流，张旭，刘娅. 基于共建共享模式的知识服务系统建设研究［J］. 情报杂志，2011，30（3）：170－175.

14 王建文. 基于图书情报系统的知识服务能力优化策略［J］. 科技创新导报，2015，(15)：190.

15 万美. 卫生信息化视角下的医学信息资源建设［J］. 医学信息学杂志，2014，35（4）：77－79.

16 邓红巧. 略论知识服务在高校图书馆中的应用［J］. 高校图书馆工作，2010，30（6）：70－72.

（上接第 65 页）

2 孟刚，潘蕾，高菁，等. 医疗器械不良事件报表质量评估方法研究［J］. 中国医疗器械信息，2008，14（2）：43－47.

3 刘斌，翟伟，马宁，等. 医疗器械不良事件报告表质量评价方法探索［J］. 中国药物警戒，2011，8（3）：165－168.

4 施雯慧，姚捷，赵燕，等. 美国医疗器械不良事件报告数据库研究［J］. 中国医药导报，2014，(28)：148－152.

5 施雯慧，陈颖，姚捷，等. FDA 公共数据开放项目中屈螺酮炔雌醇片的分析研究［J］. 中国药物警戒，2015，12（9）：552－555.

6 化柏林. 知识抽取中的停用词处理技术［J］. 现代图书情报技术，2007，(8)：48－51.

7 Blei D，Carin L，Dunson D. Probabilistic Topic Models［J］. IEEE Signal Processing Magazine，2010，27（6）：55－65.

8 Wu H，Luk R，Wong K，et al. Interpreting TF－IDF Term Weights as Making Relevance Decisions［J］. ACM Transactions on Information Systems，2008，26（3）：1－37.

9 Grün B，Hornik K. Topicmodels：an R package for fitting topic models［J］. Journal of Statistical Software，2011，40（13）：1－30.

10 Daniels K，Mosher WD，Jones J. Contraceptive Methods Women Have Ever Used：United States，1982－2010，National Health Statistics Reports，2013，No. 62［EB/OL］.［2016－07－15］. http：//www.cdc.gov/nchs/data/nhsr/nhsr062.pdf.