基于文本挖掘技术分析治疗肺癌的中医用药规律

郭玉明[1,2],姜森[2],郑光[2],郭洪涛[1],吕爱平[2*]
(1. 上海中医药大学,上海 201203; 2. 中国中医科学院中医临床基础医学研究所,北京 100700)

[摘要] **目的**:分析治疗肺癌的常用中医用药规律,为临床应用提供参考依据。**方法**:采用一维敏感字频数统计方法等技术,统计分析常用中药用药频率及药物协同关系规律,绘制协同药物网络图,抽取其中三层进行分析讨论。**结果**:中药频数分析显示人参、黄芪等补益药物为治疗肺癌首要核心用药,中药协同关系分析显示治疗肺癌药物按照益气养阴、健脾化痰、解毒消积的规律分布。**结论**:常用中药用药规律与病因病机相符,对临床应用具有一定指导意义,文本挖掘技术可以为中医药研究提供技术支持。

[关键词] 肺癌;中医;文本挖掘;用药规律
[中图分类号] R273 [文献标识码] A [文章编号] 1005-9903(2011)16-0277-04

Regularity for Lung Cancer Treatnrent by Traditional Chinese Medicine Analyzed with Text Mining Technique

GUO Yu-ming[1,2], *JIANG Miao*[2], *ZHENG Guang*[2], *GUO Hong-tao*[1], *LV Ai-ping*[2*]
(1. *Shanghai University of Traditional Chinese Medicine*, *Shanghai* 201203, *China*; 2. *Institute of Basic Research in Clinical Medicine*, *China Academy of Chinese Medical Sciences*, *Beijing* 100700, *China*)

[**Abstract**] **Objective**: To analyze the regularity of treating lung cancer with traditional Chinese medicine (TCM) herbs, and supply a foundation for the clinical practice. **Method**: All the references searched from Chinese biomedical literature database (CBM) were analyzed with one dimension sensitive character frequencies analysis. Then a co-existed herbs network was set up. The final analysis included three of all the networks. **Result**: Ginseng (Radix Ginseng) and Milkvetch Root (Radix Astragali) were the primary tonic herbs in treating lung cancer. The regularity of prescriptions was consistent to therapeutic principle of supplementing qi and nourishing yin, strengthening the spleen and reducing phlegm, and removing the toxic and stagnation. **Conclusion**: The regularity was consistent to the etiology and pathogenesis of lung cancer. And it will be useful for the clinical application. Text mining could be used as a reliable tool of TCM research.

[收稿日期] 2011-02-28
[基金项目] 国家"十一五"科技支撑计划项目(2006BAI04A10);科技部创新方法知识体系建设项目(2008IM020900);国家自然科学基金项目(30825047,30902003);中国中医科学院自主选题项目(Z0134)
[第一作者] 郭玉明,博士生,从事临床评价研究,Tel:64014411-3401,E-mail:guoyuming_0520@126.com
[通讯作者] *吕爱平,博士生导师,研究员,从事临床评价研究,Tel:64014411-3301,E-mail:lap64067611@126.com

读 2010 年版《中国药典》[J]. 中国中药杂志, 2010, 35(16):2048.
[7] 谢培山. 基于传统的中药现代化与质量评价——继承与创新[J]. 世界科学技术——中医药现代化,2006, 8(3):8.
[8] 于江泳,余伯阳,钱忠直. 建立中药注射剂综合标准化质量控制体系的思考[J]. 中国药事, 2010, 24(1):41.
[9] 秦海林. 中药物质基础整体特征的精细表达与解析——中药指纹图谱的研究[J]. 世界科学技术——中医药现代化,2002,4(4):12.
[10] 中国药典. 一部[S]. 2010.

[责任编辑 仝燕]

[**Key words**] lung cancer; traditional Chinese medicine; text mining; regularity of medicine

肺癌,又称原发性支气管肺癌,是指原发于支气管黏膜和肺泡的实质性恶性肿瘤,是全世界最常见的恶性肿瘤之一,发病率和病死率呈逐年上升趋势。因此,为肺癌患者提供有效的治疗方法极为重要。虽然目前手术治疗、放化疗以及生物制剂的疗法不断完善,但仍未取得令人满意的效果。中医药治疗,结合手术、放化疗等手段在肺癌不同阶段可提高机体免疫力、延缓病程、大大改善患者生活质量、减轻西药治疗的副作用[1]。然而,海量治疗药物未经系统分析,很难总结其用药规律,使得临床应用及研究面临一大难题。文本挖掘[2]是从非结构化的文本中发现潜在的概念以及概念间的相互关系,是指从大量文本数据中提取出可理解的、未知的、最终可用的知识的过程。采用这种方式,对现有中医药治疗肺癌的数据进行挖掘分析,从中获取治疗的核心处方或用药规律,对于中医药治疗规律的研究提供技术支持,从而为中医药治疗肺癌起到积极的推动作用。

1 材料与方法

文本挖掘应用到生物、医学上,可以分为文本数据收集、处理、结构化分析、可视化以及评价 5 个步骤[3]。

1.1 文本数据收集 首先,登录中国生物医学文献数据库(Chinese Biomedical Literature Database, CBM)网址:http://sinomed.cintcm.ac.cn/index.jsp,在主题检索下检索自建库起至 2010 年 10 月 22 日关键词为"肺癌"的相关文献。经过检索,出现款目词、主题词、命中文献数,合并检索主题词,共得到文献58 746篇。下载文献内容包含文献流水号、标题、摘要、主题词等信息,保存为 TXT 文本。

1.2 文本数据处理 将收集来的数据,按照现在的先后顺序,整合为一个平面文件,以 ANSI 编码格式保存。然后,利用专有的文本提取工具(正申请软件著作权),对 **1.1** 中下载的非结构化的 TXT 文本数据进行信息提取,保存成格式化的、便于数据库(Access)和大型数据库(Microsoft SQL Server, SQL)处理的格式。提取信息主要是机标关键词(包括核心和非核心两种类型,以下简称关键词)。提取出来的数据,首先存入 Access 数据库,作为我们下一步数据处理的材料,然后导入 SQL 中进行下一步的挖掘分析。

1.3 一维敏感字计算 根据从 CBM 上下载的数据,包含标题、主题词、关键词和摘要,按照文章编号过滤出中药名称,排除重复出现的次数,采用一维敏感字频数统计方法计算单味中药出现的有效频数。

1.4 数据挖掘以及分析 根据 **1.2** 中生成的 Access 数据库,将"结果"数据表导入 SQL 中,以"Table_Initial"为表名称,针对"序号"和"机标关键词"进行处理。为了方便处理,将"序号"和"机标关键词"两个字段分别用 PMID(类似于 PubMed 里面的字段名)和 DescriptorName(类似于 PubMed 里面的字段名)来表示。据此方法,在同一篇文章中出现的关键词,在关键词这一抽象层面上,部分反映整篇文章的信息。并且,就某一篇具体的文献来说,相关的关键词之间存在着"共同出现"这一基本事实。这种协同出现不是随机的,而是蕴含有一定的意义[4],尤其是在以很高的频率、协同出现的关键词对,在一定的程度上,反映了全国乃至世界科研工作者对它们的重视程度。更重要的是,针对目前的文本挖掘技术来说[2-4],这些协同出现的关键词,是很好的基础素材。

基于以上分析,构造针对每一篇文献共同出现的关键词对,见表 1。经过表 1 算法的实现,得到名为 DN_pairs 的数据表。观察发现数据表 DN_pairs 存在大量相同的关键词对,这些冗余的数据,对于数据分析来说,大部分属于噪音,对此,将相同的关键词对进行合并处理,只保留它们出现的频数,见表 2。经过表 2 中算法的处理,得到名为 DN_pairs_frqcy 的数据表,在这个数据表内,所有的关键词对,都只出现一次,并且都有一个对应的频数(Frequency)。

表 1

```
USE Table_Initial
FOR each PMID
  k = Number_of_DescriptorName(PMID)
  j = 1
  FOR DescriptorNames(i) (i = 1,2,... ,k)
    DO while j≤k
        DescriptorNames _ Pair = DescriptorNames (i) +
            DescriptorNames(j)
      j = j + 1
      OUTPUT dESCRIPTORnAME_pAIR INTO
          table DN_pairs
    ENDDO
    j = 1
  ENDFOR
ENDFOR
```

表 2

```
USE table DN_pairs
k = max_line_number
DO while k≥1
GO top
  FOR DescriptorName_Pair(1)//The 1st pairs in CHD_RA
    COUNT its Frequency
  EndFor
  OUTPUT DescritorName_Pair, Frequency INTO table
          DN_pairs_Frqncy
  DELETE all DescriptorName_Pair(1) from table
          DN_pairs
  k = max_line_nutnber
ENDDO
```

1.5 数据的可视化 根据**1.4**中得到的数据表DN_pairs_frqcy,抽出不同频数的关键词对,用Cytoscape 2.7进行可视化处理,根据中药间相关频次逐步抽提,将数据分为1~5层(频数由低到高),抽取高、中、低3层进行分析。

2 结果

2.1 治疗肺癌常用中药频数分析 治疗肺癌常用中药频数统计显示,频数由高至低排列,前25位中药如表3所示。人参、黄芪、艾叶、苦参、白及出现频率超过100。经过回溯文献摘要,人参、黄芪在治疗肺癌中药复方中应用广泛;艾叶多数出现在"康艾注射液"、"艾迪注射液"等中药注射剂中;苦参相关文献多数围绕苦参注射液、苦参碱、苦参素进行报道;其中4处白及表达为"白及"与肺癌治疗相关,其他均为噪音。

表3 常用中药频率及噪音分析

中药	频数/次	噪音过滤(是/否)
人参	181	否
黄芪	173	否
艾叶	159	否
苦参	137	否
白及	135	部分
鸦胆子	77	否
丹参	63	否
麦冬	58	否
白花蛇舌草	57	否
薏苡仁	52	否
天冬	40	否
甘草	36	否
姜黄	36	否
沙参	32	否
川芎	26	否
天花粉	25	否
白术	23	否
百合	23	否
青蒿	23	否
党参	21	否
苏木	21	否
莪术	21	否
半夏	21	否
生地黄	20	否
瓜蒌	19	否

2.2 治疗肺癌常用中药协同关系分析 治疗肺癌的常用中药主要包括黄芪、人参、麦冬、沙参、白花蛇舌草,人参与黄芪协同出现频数到达60次,其次为人参-麦冬、黄芪-白花蛇舌草、黄芪-沙参、沙参-麦冬。见图1。

将药物协同关系网络进一步降低层次,发现增加了薏苡仁、白术、半夏、女贞子,这4味药物单独出现频数均位列前30名,从协同关系来看,半夏和薏苡仁协同出现14次,女贞子和黄芪协同出现14次,白术和黄芪协同出现13次。见图2。

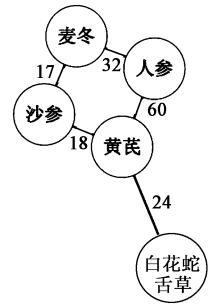

图1 肺癌用药网络频数高层图

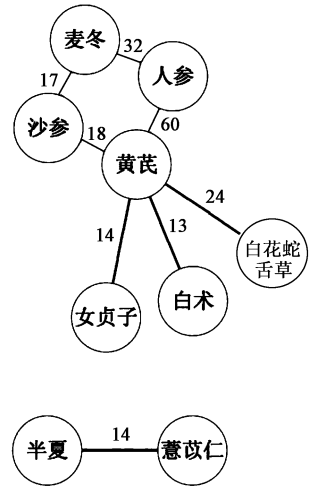

图2 肺癌用药网络频数中层图

将药物协同关系网络继续降低层次,共出现21种药物。此层较上一层增加了天冬、甘草、百合、莪术、党参、生地黄、茯苓、枸杞、浙贝母、生牡蛎、杏仁、熟地黄,其中出现中药常用药对茯苓-白术。见图3。

3 讨论

中医无"肺癌"的病名。依据肺癌的病因病机、症状、预后,可将其归为中医的肺积、咳嗽、息贲、咯血等范畴。《医宗必读》认为:"积之成也,正气不足,而后邪气踞之"。《景岳全书》:"脾肾不足及虚弱失调之人,多有集聚之病"。可见,肺癌的发生与正气虚损和邪毒入侵关系密切,正气内虚、脏腑阴阳失调是该病的主要基础。因此,治疗当以扶正培本、祛除邪毒、调理脏腑为大法。

根据文本挖掘分析所得用药规律,使用频数最高的药物为具大补元气之功效的人参,继之为补中益气的黄芪,两药均为培补正气之要药,治疗切中病因病机。经过常用中药协同关系网络频数高层分析(图1),人参、黄芪协同出现60次,同样为频数统计的最常用药物,可见这两味药是中医治疗肺癌使用的较为核心的补益药物。结合现代药理研究的结果,人参可以通过抑制炎症-肿瘤中介物质,活化过氧物酶体的增殖,激活γ受体及肿瘤生长因子-β_1(TGF-β_1),从而起

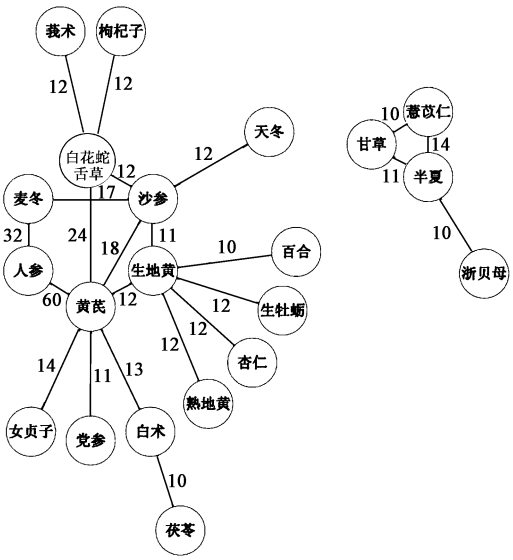

图 3　肺癌用药网络频数低层图

到抑制肿瘤及调节机体免疫的作用[5]；黄芪可以通过降低血清肿瘤坏死因子-α(TNF-α)水平、下调 Bax 蛋白的表达、抑制细胞凋亡从而起到调节免疫功能的功效[6]。药理研究结果提示两药用于肺癌的治疗可以起到调节免疫功能、抗肿瘤的作用。沙参-麦冬为一组滋阴润肺、清热生津的常用中药药对，现代医学证实沙参具有祛痰镇咳、免疫抑制以及抗肿瘤的功能[7-8]，麦冬通过提高 NK 细胞活性起到抑瘤作用[9]，麦冬多糖可以促进体液免疫和细胞免疫[10]。白花蛇舌草功专清热解毒，该药以其已被证实的免疫调节作用，并通过刺激机体的免疫系统杀伤或吞噬肿瘤细胞，广泛应用于临床调节免疫功能和治疗肿瘤等疾病[11]，与黄芪协同使用，攻补兼施。根据上述用药规律分析可见，扶正祛邪为中医药治疗肺癌的基本治则。

从常用中药协同关系网络频数中层(图 2)分析来看，与频数分析的高层相比，增加了女贞子、白术、薏苡仁以及半夏 4 味药物。其中，女贞子滋补肝肾，白术益气健脾，与黄芪协同使用，补益脏腑的虚损；薏苡仁健脾渗湿、清热排脓，半夏燥湿化痰、消痞散结，祛除痰瘀所致的积聚，这既符合痰瘀致病的病因，又解决的痰瘀这一病理产物。从常用中药协同关系网络频数低层(图 3)分析来看，在常用养阴益气和化痰散结药物的基础上，增加了天冬、百合、地黄等养阴润肺之品，以及莪术、生牡蛎等软坚消积的药物。脾为生痰之源，肺为储痰之器。茯苓-白术益气健脾，入脾、肺两经，从源头祛除痰湿，消除积聚，可以有效治疗肺癌。女贞子、莪术等多味药物分别经实验研究证实具有提高免疫、抑制肿瘤增殖作用。

综合协同中药频数分析的 3 个层次，结果显示频数最高层中药多以益气养阴为主，逐渐加入清肺解毒化痰的治疗药物，最低层加入攻积散结的药物，用药力度逐层加强，主要治则遵从扶正、祛邪，3 个层次用药的变化一定意义上反映治疗肺癌中药的常见加减规律，与正气虚损为主、邪毒入侵致病的病因相应。

通过中药用药规律的分析，及其与中医理论的印证，表明文本挖掘技术基本准确的反映了临床应用的数据本质，可以为中医药研究及临床应用提供方法学的支持；分析所得的用药规律也可以为临床提供参考依据，对中医药治疗肺癌的用药具有一定的指导价值。

[参考文献]

[1] 姜苗，左明焕，侯丽，等. 中医药治疗肺癌的优势分析[J]. 中国中医药信息杂志，2005，12(3)：3.

[2] Seifert J W. Data mining：an overview[J]. CRS Report 2004，RL31798.

[3] Mathiak B，Eckstein S. Five steps to text mining in biomedical literature. In Proceedings of the second european workshop on data mining and text mining for bioinformatics[C]. held in conjunction with ECML/PKDD in Pisa，Italy，2004，24：47.

[4] Campagna A，Pagh R. Finding associations and computing similarity via biased pair sampling[C]. 2009 ninth IEEE international conference on data mining，2009：61.

[5] Hofseth L J，Wargovich M J. Inflammation，cancer，and targets of ginseng[J]. J Nutr，2007，137：183.

[6] Zhang R，Zhang X，Ruan Y，et al. Protective effect of Radix Astragali injection on immune organs of rats with obstructive jaundice and its mechanism[J]. World J Gastroenterol，2009，15(23)：2862.

[7] 高亦珑，赵淑红. 南沙参的药理作用[J]. 中国药师，2007，10(6)：594.

[8] 林喆，赵亚，原忠. 北沙参的化学成分及药理作用研究进展[J]. 中国中医药信息杂志，2007，14(7)：91.

[9] 舒琦瑾，吴良村. 新加沙参麦冬汤抗肿瘤的实验研究[J]. 中国中医基础医学杂志，2002，8(4)：34.

[10] 周世文，徐传福. 多糖的免疫药理作用[J]. 中国生化药物杂志，1994，15(2)：143.

[11] 单保恩，张金艳，杜肖娜，等. 白花蛇舌草的免疫学调节活性和抗肿瘤活性[J]. 中国中西医结合杂志，2001，21(5)：370.

[责任编辑　何伟]