doi:10.3772/j.issn.1002-0470.2014.11.004

冠心病关键词在中医古文献中的分布研究①

王宏生②* 朱 丹③* 张明雪** 李 涵*** 张 颖****

(*沈阳工业大学 沈阳 110870)
(**辽宁中医药大学附属医院 沈阳 110032)
(***沈阳市健康教育中心 沈阳 110000)
(****辽宁中医药大学 沈阳 116600)

摘 要 以 700 篇中医古文献为研究对象,采用 Delphi 开发环境对中医领域专家给出的 1521 个冠心病关键词进行了词频统计分析,分析出了中医冠心病及其合并病关键词在古文献中的分布情况,以及高频关键词和最相关文献。结果表明,中医专家提取的关键词具有代表性和系统性特点,所选书目具有文献学代表性,冠心病及其常见合并病关键词词典可为科学应用中医经典文献进行文本数据挖掘提供技术支持。

关键词 冠心病,中医古文献,词频统计

0 引 言

中国的古文献是古代科技发展的记载形式,是祖先留下来的巨大财富。在众多种类的古文献中,中医古文献有着特殊的历史地位,它是当代中医理论与诊疗技术发展研究的重要依据,是中医文化的源泉,是利用率最高的一类古籍文献[1,2]。但是,目前中医文献信息管理水平不高,现有的中医文献软件在检索功能、查全率、查准率、检索时间等方面的指标均不容乐观,尤其是中医学中古今异义、异词同义、同词异义的现象更大大制约了其检索效率。故而,拟定关键词词典,在此基础上进行文献关键词的分布研究,对提高中医文献检索效率,有效地为中医临床提供文献参考,具有重要意义[3,4]。某一学科领域内文献的高频关键词反映了该领域的研究重点,通过对某一时期某一学科的关键词在领域文献中的统计分布分析可以全面地了解该学科发展的现状。本文研究了冠心病关键词在中医古文献中的分布,该项研究为中医冠心病本体设计与查询提供了技术支持,对科学应用中医经典文献进行文本数据挖掘等相关研究具有借鉴意义。

1 数据来源和关键词的确定

关键词是指能够揭示或表达文献的核心内容的具有实际意义的自然语言词汇[5,6],关键词在古文献中出现的频次高低可以确定该领域的研究重点和文献所侧重反映的具体病症[7,8]。由于关键词是一篇文献的核心内容的浓缩和提炼,因此,如果某一关键词在其所在领域的文献中反复出现,则可反映出该关键词所代表的病症是该领域的研究重点[9,10]。

1.1 数据来源

课题组选取了自秦汉至现代有重要医学价值的 700 部中医文献,内容涉及医经类(33 部)、综合医书类(80 部)、伤寒金匮类(52 部)、温病类(23 部)、诊法类(26 部)、临证各科类(198 部)、本草类(56 部)、方书(71 部)、针灸推拿类(42 部)、养生食疗外治类(17 部),医案医话类(99 部)、其它类(3 部)共 12 个门类。本文所选中医文献朝代分布见表 1。

所选书目基本涵盖了中医经典必读书目和历代医家的代表性著述,入选的文献数目与中医学成熟与发展的客观规律相符。因此,选取的书目有代表性和可信性。其中《神农本草经》是我国现存第一部本草专著[11],《本草纲目》是历代诸家本草中最具

① 国家自然科学基金(81273698)和沈阳市科技计划(F12-155-9-00)资助项目。
② 男,1964 年生,硕士,副教授;研究方向:智能信息处理;E-mail:whslike@163.com
③ 通讯作者,E-mail:489408431@qq.com;whslike@163.com
(收稿日期:2014-07-03)

有影响的医药学巨著[12],《黄帝内经素问》、《黄帝内经灵枢》是医学从哲学及其它学科中开始分离的标志[13,14]。

表 1　中医古文献朝代分布表

朝代	文献数
汉	11
魏晋南北朝	11
隋唐	16
宋	54
金	18
元	32
明	129
清	366
近现代	50
朝代不详	13

1.2　关键词的确定

由辽宁中医药大学专家提供中医冠心病领域的关键词[15],这些关键词是在教学、科研和临床实践中总结出来的,保证了统计结果的权威性、可参考性。关键词共有 1521 个,其中包括主关键词 437 个、候选关键词 1010 个和心病病机 74 个。

2　计算机统计分析方法

本文采用 Delphi 作为统计分析工具。Delphi 是由 Borland 公司推出的一个集成开发环境,其使用的核心是由传统 Pascal 语言发展而来的 Object Pascal,它以图形用户界面为开发环境,透过 IDE、VCL 工具与编译器,配合连结数据库的功能,构成一个以面向对象程序设计为中心的应用程序开发工具。统计分析的具体过程如下:

(1) 将选取的 700 篇中医古文献统一保存为纯文本文档格式。并将格式规范为"标号-文献名",标号从 000 到 699。由于数据处理中,Excel 只能存储 256 列,故将文献《000-神农百草经》--《249-周慎斋遗书》、《250-奇经八脉考》--《499-金匮要略方论》、《500-金匮要略心典》--《699-名老中医之路》分 3 段进行统计处理。

(2) 将文献[15]中医专家提供的"冠心病关键词词典" Excel 文件的主关键词列读入关键词表。

(3) 将中医文献读入数据库。

(4) 设计统计处理算法,将关键词在全部文献出现的词频和中医文献出现关键词的词频进行统计,结果在 Excel 文件"中医文献标注结果"中输出。

(5) 重复(2)、(3)、(4)步骤,依次对候选关键词和心病病机的词频完成统计。

(6) 用 Excel 对三类关键词的词频统计结果进行整理排序,对所有关键词进行分析,同时生成统计直方图。

3　关键词词频统计结果及分析

3.1　关键词在全部文献出现的词频统计结果及分析

中医专家提取到的 1521 个不同的冠心病及其合并病关键词在全部文献中的出现总次数为 1113916 次,每个关键词的词频范围为 0 ~ 85022,平均词频为 732,这说明其词频变幅较大,充分反映了医家对中医疾病表现的记述各有侧重的特点,而词频范围很宽也表明中医专家选取的关键词较全面地反映了冠心病及其合并病的临床表现(关键词词频按分段数统计结果直方图见图 1)。

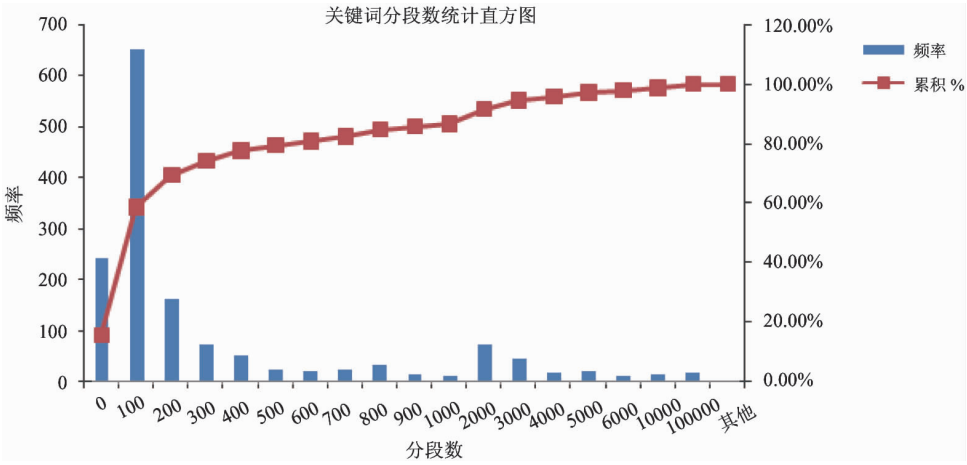

图 1　关键词词频按分段数统计结果直方图

分析冠心病病名与 5 个核心症状的代表性关键词词频结果(见表 2)可知,这 18 个冠心病代表性关键词占关键词总数的 1.2%,其频次达 73968 次,占频次总数的 6.6%,其平均词频数为 4109,为整体水平的 5.6 倍。这个结果表明,中医专家选取的关键词集中反映了对冠心病认识的核心内容,是具备概括性或代表性的概念,冠心病病症代表性关键词的提取与文献联系密切。此外,我们对词频大于 10000 的关键词进行了统计,按词频由大到小顺序排列为泻、惊、喘、汗出、膈、气虚、恶寒、湿热、腹痛、头痛、咳嗽、阴虚、血虚、身热、下利、悸、腹满,统计结果见表 3。可以看出,这些高频词或用于判断证候属性,或是有助于辨证的全身症状,这也在另一个侧面体现了关键词库代表性与全面性的统一,证实了关键词库的构建较为成功。

表 2　冠心病病症代表性关键词词频统计结果

关键词类别	关键词	频率
病名	胸痹	1165
症状 1	心痛	6314
症状 2	悸	10083
	心跳	815
	怔忡	1863
	心忪	354
	心慌	168
	心怔	72
	心忡	57
症状 3	气急	2226
	气促	944
	呼吸急促	35
症状 4	喘	37245
症状 5	短气	4594
	少气	4461
	气少	2441
	气短	1124
	呼多吸少	7

表 3　词频大于 10000 的关键词词频统计结果

关键词	词频	关键词	词频
泻	85022	头痛	14085
惊	40594	咳嗽	13217
喘	37245	阴虚	12976
汗出	27096	血虚	12157
膈	26515	身热	12077
气虚	22322	下利	12043
恶寒	21027	悸	10083
湿热	16908	腹满	10003
腹痛	14715		

3.2　中医文献出现关键词的词频统计结果及分析

每篇文献的关键词词频可以反映该篇文献与中医冠心病的相关程度。中医古文献出现关键词的词频统计(见图 2)。在 700 篇文献中,1521 个关键词出现的词频总数为 1303091 次,每篇文献涵盖关键词的词频范围为 0 ~ 35786,平均词频为 1861。这些关键词在文献中出现的词频范围波动较大,但除《十二经补泻温凉引经药歌》之外,均有命中关键词,体现出中医专家选取的关键词与各篇文献均有相关,也在一定程度上反映出冠心病及其合并病症状多样,在各类中医文献中均能找到相关条目,为临床辨治提供启发。

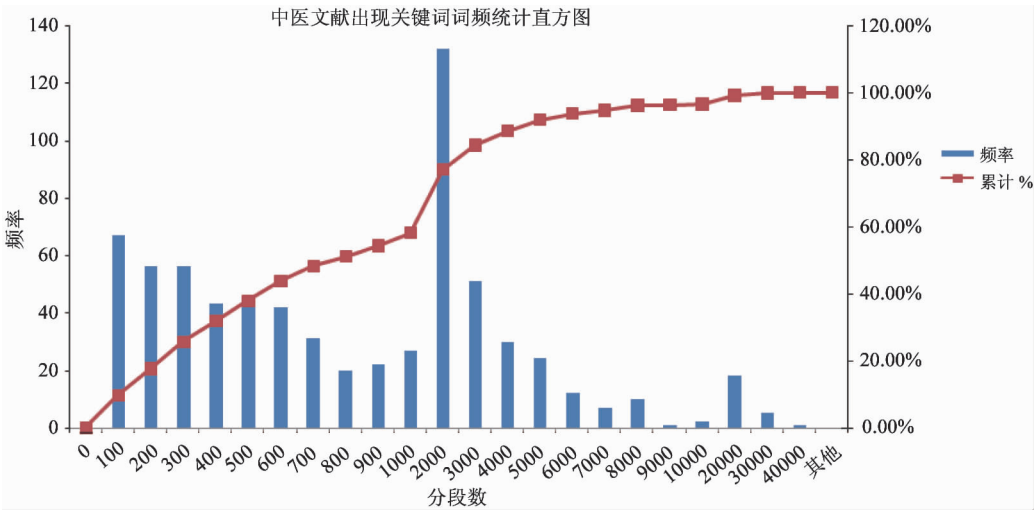

图 2　中医文献出现关键词的词频统计

表 4　关键词词频大于 5000 的中医古文献统计结果

书名	关键词词频	属类	年代
普济方	35786	方书类	明
古今医统大全	28242	综合医书类	明
圣济总录	26339	方书类	宋
医学纲目	23070	综合医书类	明
医宗金鉴	21618	综合医书类	清
景岳全书	20517	综合医书类	明
医学入门	19283	综合医书类	明
太平圣惠方	18622	方书类	宋
冯氏锦囊秘录	16952	综合医书类	明
本草纲目	16777	本草类	明
续名医类案	16595	医论医案类	清
杂病广要	15893	临证各科类(内科)	清
医述	14562	综合医书类	清
张氏医通	14227	综合医书类	清
奇效良方	13888	方书类	明
幼幼新书	12475	临证各科类(儿科)	宋
中医临证经验与方法	11489	医论医案类	现代
疡医大全	10753	临证各科类(外伤科)	清
伤寒论纲目	10491	伤寒金匮类	清
保婴撮要	9310	临证各科类(儿科)	明
外台秘要	8752	综合医书类	唐
女科证治准绳	7604	临证各科类(妇产科)	明
寿世保元	7574	综合医书类	明
重订通俗伤寒论	7358	伤寒金匮类	清
类证治裁	7356	综合医书类	清
类经	7309	医经类	明
医学衷中参西录	7298	综合医书类	民国
备急千金要方	7217	方书类	唐
订正仲景全书伤寒论注	7112	伤寒金匮类	清
玉机微义	7048	综合医书类	明
中国百年百名中医临床家——李翰卿	6843	综合医书类	现代
世医得效方	6715	方书类	元
本草品汇精要	6591	本草类	明
顾松园医镜	6307	综合医书类	清
伤寒论辑义	6221	伤寒金匮类	清
本草述钩元	6065	本草类	清
证类本草	6048	本草类	宋
伤寒溯源集	5941	伤寒金匮类	清
万病回春	5560	综合医书类	明
症因脉治	5477	综合医书类	清
临证指南医案	5468	医论医案类	清
针灸大成	5441	针灸推拿类	明
千金翼方	5405	方书类	唐
伤寒大白	5271	伤寒金匮类	清
古今医鉴	5270	综合医书类	明
医碥	5254	临证各科类(内科)	清
济阴纲目	5144	临证各科类(内科)	明
证治汇补	5062	综合医书类	清

其中关键词词频在 5000 以上的文献共有 48 篇(见表 4)。从表 4 中可看出,这些文献涵盖了医经、本草、方剂、临床各科、伤寒金匮、温病、医论医话等书目属类,较有代表性地体现了文献学的各类属性特点。其中,高频关键词出现最多的书籍以明清时期的综合医书类、内科临证类书目居多,这与冠心病及其常见合并病的疾病属性相关,并体现了明清时期内科临床医学成熟与发展的特点。大型本草、方书中关键词出现频次亦较多,这与宋金元时期方书编著与理论升华的历史特点亦较为契合。此外,这些关键词不单纯出现在内科专著中,在妇科、儿科为主的专著中也有体现,可见中医学据象辨证的特点,也提示我们在未来可有侧重地进一步科学利用这些中医经典书目。

4　结　论

本文采用 Delphi 开发环境,对 700 篇中医古文献进行了领域专家给出的 1521 个关键词的词频统计分析,得出了中医冠心病关键词在古文献中的分布情况、最高频关键词与最相关文献。统计结果分析表明,中医专家提取的关键词具有代表性和系统性特点,可作为文本挖掘的索引结构,所选文献属类全面,有较好的代表性,可作为进一步文本挖掘的中医文献数据库。该项研究为中医冠心病本体设计与查询提供了技术支持,可提高查阅者的查阅速度,降低查阅者选择文献的盲目性,使得查阅者可通过直观输入关键词而得到较全面的检索结果。这样使查阅者能够有更多的时间只关注查找到的信息,而不必再花时间和精力去考虑信息得到的过程,从而为科学应用中医经典文献提供技术支持。

参考文献

[1] 范晓艳. 中医古文献研究的价值. 甘肃中医学院学报, 2002,19(2):61-63

[2] 朱素兰. 医学文献检索系统的现在与将来. 天津,国际医学图持馆科学人会,2001. 353-358

[3] 杨彼德. 中文古籍数字化保存保护:合作构想. 北京,中文善本古籍保存保护国际研讨会,2001. 17-34

[4] 冯广义. 中医古籍的整理研究与中医学发展. 中医药导报, 2007,13(9):10-14

[5] 黄河胜,王华,陈志武. 用词频分析法看国内药学研究趋势. 中国科技期刊研究,2006,17(6):1110-1113

[6] 潘和平,曹红院,孙业桓等. 25 种预防医学与卫生学类核心期刊 2004 ~ 2006 年关键词词频分析. 中国科技

期刊研究,2008,19(2):207-211
[7] 陈玉琪,汤勃,张云辉等. 2008 - 2011 年《传染病信息》论文关键词统计分析. 传染病信息,2012,25(5):295-297
[8] 马费成,张勤. 国内外知识管理研究热点——基于词频的统计分析. 情报学报,2006,25(2):163-171
[9] 任淑敏,胡丽美,王倩飞等. 从《中国行为医学科学》载文关键词词频探析行为医学领域的研究热点. 中华医学图书情报杂志,2008,5(17):78-81
[10] 殷蜀梅,张智雄,吴振新. 一种从医学文本中实现自动关键词抽取和筛选的技术方法. 现代图书情报技术,2008(8):31-36
[11] 钟赣生,李少华.《神农本草经》的药物成就. 中华中医药杂志,2006,21(7):390-392
[12] 杨东方.《本草纲目——引据古今医家书目》辩证. 北京中医药大学学报,2009,32(9):590-593
[13] 黄利兴.《黄帝内经》的历史评价与读法. 医学与哲学,2013,34(11):71-72
[14] 钱会南.《黄帝内经太素》在中医理论体系框架形成中的作用. 安徽中医药大学学报,2014,33(1):1-3
[15] 辽宁中医药大学. 冠心病关键词词典, 2014

Study on distribution of coronary heart disease keywords in Chinese medicine ancient literature

Wang Hongsheng[*], Zhu Dan[*], Zhang Mingxue[**], Li Han[***], Zhang Ying[****]
([*] Shenyang University of Technology, Shenyang 110870)
([**] Affiliated Hospital of Liaoning University of Traditional Chinese Medicine, Shenyang 110032)
([***] Shenyang Health Education Center, Shenyang 110000)
([****] Liaoning University of Traditional Chinese Medicine, Shenyang 116600)

Abstract

The statistical analysis of the frequency of 1521 coronary heart disease keywords given by experts was conducted in 700 pieces of ancient literature of Chinese medicine by using the tool of the Delphi development environment, and the distribution of the keywords of the coronary heart disease and its combined diseases in ancient literature, as well as the high-frequency keywords and the most relevant literature, were obtained. The result indicates that the keywords Chinese medicine experts extracted are representative and systemic, and the selected bibliography is philology representative. The coronary heart disease and its common keyword dictionary can provide technical support in scientific application of the classical literature of Chinese medicine for text data mining.

Key words: coronary heart disease, Chinese medicine ancient literature, word frequency statistics