

浅议症状元素与复杂症状术语的处理方法

吕晓颖¹ 李敬华² 贾李蓉² 杨逢春² 李鸿涛² 刘丽红² 田野²

(1 北京中医药大学, 100029; 2 中国中医科学院中医药信息研究所)

摘要 目的: 尝试运用症状元素的概念处理寻找一种处理方法解析中医药领域中的复杂症状术语, 解决词义模糊、同义词、一词多义等问题, 进一步为医药信息的标准化提供更为清晰和条理的术语支持。方法: 以中医医案数据库中大量症状数据为基础, 以症状术语的词性及限定元素为轴对其进行清洗、分类, 提取出症状元素简化症状术语的表达。结果: 通过对中医医案数据库的整理得到症状数据 8 万余条, 并得出一个基本清晰的症状元素划分方法。结论: 运用症状元素这一概念并且根据词性和限定元素这一组划分轴, 可以有效处理复杂的症状术语, 对中医药领域的学术沟通、对中医药领域隐性知识的挖掘、发现和利用具有重要现实意义。

关键词 症状元素; 症状术语

Discussion on the Symptom Elements and Complex Symptom Terminologies

Lu Xiaoying¹, Li Jinghua², Jia Lirong², Yang Fengchun², Li Hongtao², Liu Lihong², Tian Ye²

(1. Beijing University of Traditional Chinese Medicine, Beijing 100029, China; 2. Institute of Information on Traditional Chinese Medical Sciences, Beijing 100700, China)

Abstract Objective: To find a method to deal with complex symptom terminologies of Chinese medicine by using the concept of symptom elements, in order to solve various terminology issues and support the medical information standardization. **Methods:** Based on Chinese medical records database, the symptom terminologies are categorized, cleaned and extracted to obtain symptom elements. **Results:** We obtained over 80,000 symptom related data from clinical medicine database by means of a basic categorization method. **Conclusion:** Complex symptom terminologies can be effectively handled by using symptom elements and categorization according to term attributes. This is conducive to academic communication of Chinese medicine, and tapping the potentials of Chinese medicine.

Key Words Symptoms elements; Symptom terminologies

中医是一门经验学科, 不同的学者对同一症状往往有不同的记录, 西方医学经传道士带入中国, 对中医症状的表述也产生了很多影响。目前中医对症状的描述不统一的现象非常突出, 概念范围界定不清, 常具有一个含义多种表述方式或者两种症状的表述存在交叉, 如“多食易饥”, 在《黄帝内经》中称为“消谷善饥”, 在《伤寒论》中称为“消谷喜饥”, 后世中医各家对其则有多种称谓, 如多食善饥、食欲亢进、好食易饥、贪食、多食、能食、易饥等, 它很容易与消渴病之中的消症混淆^[1]。症状是诊断的根据, 是辨证基础, 随着信息时代技术的革命性发展, 大规模的数据库建设与数据仓库形成, 数据挖掘与利用的广泛开展亟需对于症状数据统一的处理标准, 以便于在计算机系统完成对海量杂乱数据的理解, 进而从中挖掘出规律与知识, 为医生提供决策支持。本研究以中医症状元素的分类与提取开

展研究, 从中探索繁杂症状数据的处理方法, 以便与计算机科学相结合, 提高医生的诊断效率。

1 症状与症状元素

症状是指患者陈述的异常感觉和医生发现的异常现象^[2]。一般说症状是广义的症状, 包含症状和体征两个方面。“症状”是指患者主观感到的痛苦或不适, 如头痛、耳鸣、胸闷、腹胀等; “体征”是指客观能检测出来的异常征象, 如面色白、喉中哮鸣、舌苔黄等^[3]。通过历史文献的研究, 张启明等^[4]提出了症状单元的概念并将症状单元按症状部位和症状性质在表示最小内涵情况下进行分类。杨在纲等^[5]提出“三类五性”的划分方法, 并根据症状在辨证中的作用不同将中医“症”进行了“意”“性”归类。“意”包括一般意义的症、定性意义的症、定位意义的症。定性意义的症又分为一般定性意义的症和特殊定性意义的症。“性”是将症状归纳为必要性资料、特征性资料、偶见性资料、一般性资料和否定性资料^[5]。这两种划分方法都对症状及其分类进行了深入思考, 但是医生在描述患者症状时, 往往还会利用其他词汇对基本症状加以形容, 以

基金项目: 国家自然科学基金委员会青年科学基金(编号: 81001560); 中国中医科学院自主选题研究项目课题(编号: ZZ03029)

通讯作者: 李敬华, 博士, 主要研究方向为中医信息化, zingarlee@hotmail.com

更准确描述患者的异常状态。因此,也就有了本研究中对症状更为细致的分类,尽量保证症状元素的任意组合可以完整表述医生对患者症状的表达。

症状元素的概念结合了数学中集合与元素一对定义。集合是指具有共同属性的事物的总体称为“集”,而元素就是组成集的每个事物,即将症状按照一定的规则划分使其每一个症状元素都属于症状的一个分类集合。每一个症状元素都是唯一的、确定的和无序的,并且在一定的分类框架下具有最小内涵意义。

2 研究方法

2.1 症状数据来源 本研究症状描述词主要提取于中医医案数据库。中医医案数据库始建于2008年,由中国中医科学院中医药信息研究所组织全国20余家科研院所近200名人员,以数据共建共享方式,合作开发的中医医案数据库。该库是结构化的关系型数据库,以公开发表于中医期刊杂志的医案文献为加工对象,采用三级审校加工方式管理,著录内容涵盖文献题录、作者、医生、患者、症状、体征、证候、检查、诊断、治疗、预后等全部信息,现已收录1984年以来近8万条数据,并在不断增量之中。在具有一定数据规模基础上,本所进行医案数据的挖掘与利用研究,以方便研究者及公众使用医案资源并促进中医临床及基础研究。

2.2 症状数据的清洗 选择中医医案数据库已加工完毕数据8万余条,数据量大庞杂,抽取其中症状数据,建立access数据库并进行以下步骤筛选。1)对所抽取的症状描述词进行数据清洗,去除空白及缺损数

据。2)对数据库中的症状描述进行去重的操作。此处去重并不是指删除重复的症状描述词而是把具有相同意思的症状描述词进行归类。重复意思表述的症状主要有两种情况:第一种描述前后顺序导致的症状在形式上的不一致,例如,口唇干裂和唇口干裂;第二种是同一种症状的不同表述方式,例如:吐血和咯血、嗝气和打嗝等。将这些意思一致的症状进行整合,并对异名进行标注。3)对整理后的症状描述词进行频数统计,去除频数为1的症状描述词,重点对10次以上的症状描述词进行分类,频数词为2次以上的症状描述词进行补充。经过以上三步的数据筛选整理后,会得到一个相对干净的症状描述数据库,这是进一步症状元素划分的前提与基础。

2.3 症状元素的分类 本研究将从中医个案数据库中提取的所有症状从多维角度进行拆分,将任意元素组合描述临床症状,以弥补不同临床医生在表述症状时的差异。所有症状元素主要分为两类,基本的症状元素类和限定元素类。基本症状元素类分为单纯症状词和复合症状词,限定元素类不是具体意义的症状元素而是对症状元素意义的外延,与基本症状元素结合使用时可对患者临床症状的表述更为准确。具体划分维度及定义如表1。

3 症状元素的提取

将经过清洗的症状描述词进行频数统计,将在数据库中出现10次以上的症状描述按照症状元素的划分规则进行分类,最后将其他症状描述作为补充查看症状元素的分类结果是否有缺漏,部分结果如表2。

表1 症状元素分类难度及定义

			说明
基本症状元素类	单纯症状词		不能再分的症状,表达一个明确的症状现象
	复合症状词	词性	
		部位名词	可以再分的症状中的部位
		排出物名词	可以再分的症状中的排出物,单独不能表达完整症状现象
		功能名词	可以再分的症状中表示人体功能的名词,单独不能表达完整症状现象
		动作名词	可以再分的症状中表示人体动作的名词,单独不能表达完整症状现象
		性质描述词	可以再分的症状中用来描述症状性质的词语,单独不能表达完整症状现象
		舌象描述词	舌质及舌苔的性质描述词
		脉象描述词	脉象的性质描述词
	数词		表示数目多少或顺序多少的词
限定元素	频率词		表示与次数,频率有关的副词
	时间词		以十二地支来表示一昼夜十二时辰的变化,运用占天色纪时、地支纪时与今序数纪时等
	程度词		表示程度的副词
	方位词		表示方向或位置的词
	否定词		表示否定意义的词
	诱因		导致某种情况发生的原因

表 2 症状元素提取结果(部分)

症状描述分类		举例
单纯症状词		腹水、黄疸
复合症状词	部位名词	鼻、舌<边、尖、中、根><舌质、舌苔>、肠、耳
	动词	堵塞、吐[流、咯]血(衄)、痒(瘙痒)、煽动、停滞
	排出物名词	白带、清涕、月经<经行、经质、经量>、大便、汗、气
	功能名词	表情、睡觉(寐、眠)、梦、记忆力(记性)、精神(神识、神志、情怀、情志)、视物(视力)
	性质描述性	齿印(齿痕)、红(潮红、赤)、燥、紫(紫滞)、软、硬
	舌象描述词	胖、绛、腻、润、厚
	脉象描述词	数、细、涩、紧、缓、沉
限定元素	数词	0~9、一至九、十、百、千、万、亿
	频率词	频频(频作、频多)、时有(时作)、延长、缩短、次、反复、
	时间词	夜、日、午后、后(延后、延期、推后、延迟、后错)、前(提前)、不定期、前期、后期、中期、中后期、周期不定、晨、五更
	程度词	多(增多、增加)、时多时少、少(减少)、正常、恶、差(减退)、略(微、稍、偏)、喜(偏好)、尚可、渐、异常、中等、欠、较、轻、重
	方位词	后、前、左右、固定不移、游走不定
	否定词	不、否、不思(不欲、不想)、不行(不下、未行、未解、不畅、不爽)、不安(不宁)、拒按、无味(乏味、不香、不馨)、无华(少华)、无
	诱因	气候

4 讨论

本研究根据症状在医案中的实际出现情况,探索了症状元素与复杂症状术语的处理方法,将 8 万条医案数据进行了症状元素分析,症状元素方法可以完成复杂症状自由语言的语义抽取和分类,在一定程度上实现语义多维度理解,但是症状术语数量极其庞大,本研究仅仅限于高频症状的症素抽取,更多更细的数据分析将有助于该方法的进一步完善。

参考文献

- [1] 赵金铎. 中医症状鉴别诊断学[M]. 北京: 人民卫生出版社, 1985: 45.
- [2] 朱文锋. 中医诊断学[M]. 北京: 中国中医药出版社, 2002: 132.
- [3] 黄碧群. 中医症状标准化的重要性[J]. 中华中医药杂志(原中国医药学报), 2011, 26(3): 429-432.
- [4] 张启明, 王义国, 张磊, 等. 内涵最小的独立症状[J]. 北京中医药大学学报, 2010, 33(1): 5-10.
- [5] 王卓飞, 杨在纲. 浅议“症的分类”在辨证中的重要性[J]. 甘肃中医, 2004, 17(10): 3-4.

(2012-06-12 收稿)

欢迎订阅 2013 年《世界中西医结合杂志》

《世界中西医结合杂志》(CN 11-5511/R, ISSN 1673-6613)是经国家新闻出版总署批准创刊,由中国科学技术协会主管,中华中医药学会主办的国家级期刊,面向国内外公开发刊。本刊为中国科技核心期刊,精品科技期刊示范项目。由中国科学院院士、著名医学家、国家最高科技奖获得者吴孟超任名誉主编,著名中医学家、中国中医科学院资深研究员路志正教授任主编。辟有述评、国医传承、博士论坛、文献研究、临床研究、实验研究、思路与方法、临证经验、药物研究、医案医话、医技与护理、教育研究、海外中医药动态、综述、政策研究、标准规范指南等栏目。

本刊现已被《中国科技论文统计源期刊》、《中国学术期刊综合评价数据库统计源期刊》、《中国期刊全文数据库全文收录期刊》、《中文科技期刊数据库》、《中国学术期刊(光盘版)》、波兰《哥白尼索引》IC 源期刊、美国《乌利希期刊指南》收录。

本刊为月刊,大 16 开本,每期 112 页,每月 28 日出版。每期定价 12 元,144 元/年。全国各地邮局均可订阅,国内邮发代号 80-538;国外向中国图书贸易总公司订阅,代号 M7968。有意购买者可直接汇款到编辑部邮购。

欢迎订阅,更欢迎您对本刊提出宝贵意见。

订阅方式:本刊采用邮发和自办发行两种发行方式。2012、2013 年杂志可在全国各地邮局订阅,国内邮发代号:80-538,亦可直接向本刊编辑部订阅。若需要 2006 年至 2011 年过刊,亦可直接向编辑部订阅。

汇款方式:邮局汇款:100101,北京市朝阳区北四环东路 115 号院 6 号楼 109 室,《世界中西医结合杂志》编辑部收,联系地址:北京市朝阳区北四环东路 115 号院 6 号楼 109 室,《世界中西医结合杂志》编辑部,邮编:100101,电话/传真:010-64822253。网址: <http://www.zhongxi.org>。E-mail: sjzxyjh@126.com。