

中医症状知识库知识关联研究^{*}

谢文利

李 君

(湖北中医药大学信息工程学院 武汉 430065)

(湖北中医药大学第一临床学院 武汉 430065)

毛树松

解 丹

(湖北省中医院 武汉 430065)

(湖北中医药大学信息工程学院 武汉 430065)

〔摘要〕 介绍中医学症状研究及中医药领域知识图谱技术应用研究现状，选取《中医药学名词》作为症状知识来源，构建中医症状知识库。运用描述性统计、关联规则分析等数据挖掘方法阐述高频证型以及症状信息分布规律、症状以及症状隐性知识之间的关联关系，利用 Neo4j 构建中医症状知识图谱并分析病证－症状－属性之间的知识关联关系。

〔关键词〕 中医症状；关联规则；知识图谱；Neo4j

〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673－6036. 2023. 02. 006

Study on Knowledge Association of Traditional Chinese Medicine Symptom Knowledge Base XIE Wenli, College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, China; LI Jun, The First Clinical College, Hubei University of Chinese Medicine, Wuhan 430065, China; MAO Shusong, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan 430065, China; XIE Dan, College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, China

〔Abstract〕 The paper introduces the research status of traditional Chinese medicine (TCM) symptom and the application of knowledge graph technology in TCM field, and selects *Chinese Terms in Traditional Chinese Medicine and Pharmacy* as the source of symptom knowledge to build a knowledge base of TCM symptom. Descriptive statistics, association rule analysis and other data mining methods are used to elaborate the distribution rules of high－frequency syndromes and symptom information, as well as the association relationship between symptoms and symptom tacit knowledge. Neo4j is used to construct the TCM symptom knowledge graph and analyze the knowledge association relationship between syndromes, symptoms and attributes.

〔Keywords〕 traditional Chinese medicine (TCM) symptom; association rules; knowledge graph; Neo4j

〔修回日期〕 2022－10－04

〔作者简介〕 谢文利，硕士研究生，发表论文 2 篇；通信作者：解丹，教授，硕士生导师。

〔基金项目〕 湖北中医药大学中医药传承与创新计划“基于信息抽取技术的中医辨证传承与创新研究”（项目编号：2022SZXC012）。

1 引言

症状作为疾病的外在表象，是组成证候的基本单元，也是判断证候和病名的重要依据^[1]。研究症状之间的关系，挖掘其中的规律有利于中医学的发展。关联分析在中医药领域被广泛应用于症状间的

关系研究,多集中在用药规律研究和证素研究,包括对疾病的症状、证型以及方药进行分析,发现用药规律、配伍规律和常见方剂^[2-3],通过分析中医证素、症状体征,发现疾病的中医证候特点、强关联证素及其分布规律^[4]。

知识图谱技术逐步发展,日益广泛应用于中医药领域,可以更加有效地描述、挖掘实体之间的关系,实现中医药知识大规模存储和高效利用,为中医药知识服务奠定基础^[5]。目前已有一些中医症状知识图谱相关研究成果,主要包括:建立由疾病、证候、症状和方剂等核心概念构成的中医临床知识图谱^[6];包含疾病、证候、症状等 9 类实体的皮肤科方剂经验知识图谱^[7];基于中医医案病例数据的中医类流感“证-治”“证-症”“症-药”知识图谱^[8];基于中医专著的中医学症状词库^[9]。症状的非特异性及其复杂多重属性,在证候和病名判断中具有重要意义^[1],但在中医症状信息化研究层面很少涉及多重属性。症状的多重属性是中医临床数据高效利用的一大难题,为解决这一问题,中国中医药信息学会于 2019 年发布的团体标准《中医临床基本症状信息分类与代码》^[10]中提出了症状属性分类这一概念。针对中医临床数据进行梳理,对症状进行规范并给出症状的属性分类,如部位属性 A、颜色属性 D 等,充分展现中医临床症状的隐性知识。

当前国际广泛使用的医学术语标准包括《医学系统命名法-临床术语》(systematized nomenclature of medicine-clinical terms, SNOMED CT)、《医学主题词表》(medical subject headings, MeSH)、《国际疾病分类与代码》(international classification of diseases, ICD)、《全球医疗设备命名》(global medical device nomenclature, GMDN)等^[11]。但由于其中不包括中医药相关内容且中英文之间存在差异,难以直接运用于中医临床。国内中医术语标准研编方面已取得的相关成果主要包括:《中国中医药学主题词表》主要用于索引与编目^[12],《中医药临床术语集》借鉴 SNOMED CT 模式建立术语系统,在医学文献服务系统和卫生信息系统中均有应用^[13],但术语标准覆盖范围较为分散且开放权限受限,缺乏完整的面向临床实践的中文临床医学术语标准体系^[14]。选取一

个权威且内容全面的症状术语集是构建中医症状知识库的重要前提。《中医药学名词》^[15]是由全国科学技术名词审定委员会发布实施的中医药学规范名词标准,收录条目较少,但集合全国专业学者之力编制,分类更具中医特色,有较高学术水平,对于症状的描述全面且具有权威性^[16]。

本研究参考《中医临床基本症状信息分类与代码》对《中医药学名词》所包含的症状信息进行标注,通过频次分析,发现常见证型、症状以及属性分布情况,并通过关联规则分析发现症状之间的关联关系以及症状属性之间存在的隐性关系,构建中医症状知识图谱,将病证和症状的知识属性通过图谱进行可视化,有助于发现中医临床症状之间的潜在联系。

2 资料与方法

2.1 数据来源

选取《中医药学名词》收录的内科、外科、妇科和儿科等 9 个科室的中医临床诊断术语 4 898 条作为症状知识来源,其中包括 1 207 种疾病、1 171 种证型。根据《中医诊断学》^[17]以及《中医症状学研究》^[18]构建中医临床诊断术语同义词表,共计 499 条。

2.2 数据处理

2.2.1 处理方法 参考《中医临床基本症状信息分类与代码》提出的症状隐性知识包括骨干症状、获取方式以及各属性分类,结合《中医诊断学》与《中医症状学研究》对 4 898 条中医临床诊断术语进行多轮标注^[19],标注出症状实体及其相应的骨干症状、获取方式和各属性分类信息,在标注过程中参考同义词表,对诊断术语进行归并,实现对症状实体中属性分类的规范化,避免出现同一属性的多种描述,使用 {symptom;} 标注症状实体,“-”前后分别为骨干症状和获取方式,“<>”标注各属性分类,标注示例如下。原文本为头身疼痛、喉痒咳嗽。标注后: {symptom: 疼痛-问疼痛; <A: 头-头部> <A: 身-全身>疼痛}, {symptom: 咽

痒 - 问不适; < A: 喉 - 咽喉 > 痒}, {symptom: 咳嗽 - 闻咳嗽; 咳嗽}。

2.2.2 中医症状知识库构建 症状实体包含必要类目属性和附加类目属性, 其中必要类目属性指骨干症状和获取方式, 附加类目属性是指部位、颜色等属性分类信息, 对症状实体进行抽取^[20], 得到 15 510 个症状实体, 对各科室症状数据进行去重后得到 15 459 条有效数据, 构建中医症状知识库 (以下简称“知识库”)。

2.3 技术与方法

2.3.1 描述性统计分析 数据分析中对于定性资料一般采用频数和百分比描述, 能直观反映数据情况^[21]。通过 IBM SPSS Statistics 25 进行描述性统计分析, 统计证型、症状、骨干症状、获取方式和各属性分类的频次。

2.3.2 关联规则分析 关联规则分析是指从数据库中找出隐藏知识, 并根据已知情况对未知问题进行推测^[22], 已被应用于中医证候、方剂、诊断以及名老中医经验研究等方面^[23]。本研究采用 IBM SPSS Modeler 18.0 进行关联规则分析, 通过 Apriori 算法建模发现关联程度高的症状和骨干症状。

2.3.3 知识图谱技术 知识图谱可以从多个维度对中医药事务进行描述, 发现中医药术语之间的关系^[24], 揭示临床规律, 从而完善中医临床知识体系, 推动中医临床研究的发展。Neo4j 作为面向网络的图数据库, 具有简单、灵活且容易上手的优势, 可以采用 Cypher 语言进行查询操作^[8]。本研究采用 Neo4j 构建中医症状知识图谱并对知识图谱进行查询, 实现中医症状知识库可视化, 呈现病证、症状和属性之间的关系。

3 结果

3.1 描述性统计分析

3.1.1 统计分析方法 对知识库中各科室的诊断术语数量、疾病数量、证型数量和症状实体数量进行统计, 其中内科包含 1 418 条诊断术语、547 种疾病、401 种证型和 7 128 个症状实体; 外科包含 641 条诊

断术语、96 种疾病、142 种证型和 1 100 个症状实体; 妇科包含 586 条诊断术语、115 种疾病、127 种证型和 2 091 个症状实体; 儿科包含 408 条诊断术语、113 种疾病、141 种证型和 1 563 个症状实体; 耳鼻喉科包含 503 条诊断术语、106 种疾病、112 种证型和 1 227 个症状实体; 眼科包含 486 条诊断术语、85 种疾病、117 种证型和 1 204 个症状实体; 骨科包含 468 条诊断术语、61 种疾病、32 种证型和 306 个症状实体; 皮肤科包含 282 条诊断术语、61 种疾病、72 种证型和 633 个症状实体; 肛肠科包含 105 条诊断术语、23 种疾病、27 种证型和 207 个症状实体。

3.1.2 中医症状知识库证型统计 统计知识库出现次数较多的证型, 排名前 10 位的证型 (频次) 分别是气血两虚证 (66)、阴虚火旺证 (44)、气滞血瘀证 (41)、肝气郁结证 (35)、湿热下注证 (35)、肝肾阴虚证 (33)、湿热蕴结证 (30)、脾虚湿困证 (27)、气阴两虚证 (26) 和脾肾阳虚证 (25)。

3.1.3 中医症状知识库症状、骨干症状和获取方式统计 统计知识库中的症状、骨干症状和获取方式频次, 见表 1。症状频次最多的是发热 (264), 骨干症状频次最多的是疼痛 (1 209), 获取方式频次最多的是问不适 (3 452)。发热是一个症状实体也是一个骨干症状, 发热作为骨干症状频次达到 975。

表 1 中医症状知识库基本数据频次统计前 5 位

症状	频次	骨干症状	频次	获取方式	频次
发热	264	疼痛	1 209	问不适	3 452
头晕	254	皮肤异常	979	问寒热	1 596
乏力	244	发热	975	问疼痛	1 207
头痛	206	大便异常	821	问大便	786
神疲	200	乏力	699	问渴饮	656

3.1.4 中医症状知识库各科室症状分布统计 选取知识库中的科室和症状数据, 统计各科室高频症状, 发现各科室症状分布的区别。内科症状频次较多的是乏力和头晕等, 儿科症状频次较多的是发热和咳嗽等, 耳鼻喉科症状频次较多的为耳鸣和鼻塞等, 眼科症状频次较多的为流泪和畏光等, 肛肠科症状频次较多的为肛门坠胀, 骨科症状频次最多的为活动受限, 外科症状频次最多的为发热。

3.1.5 中医症状知识库属性分类频次统计 通过 IBM SPSS Statistics 对 24 种属性分类进行描述性统计，将得到的属性分类频次导入 Excel。属性分类频次排序前 3 位的分别是：部位属性（6 771）、性质属性（2 163）和颜色属性（1 367），见图 1。

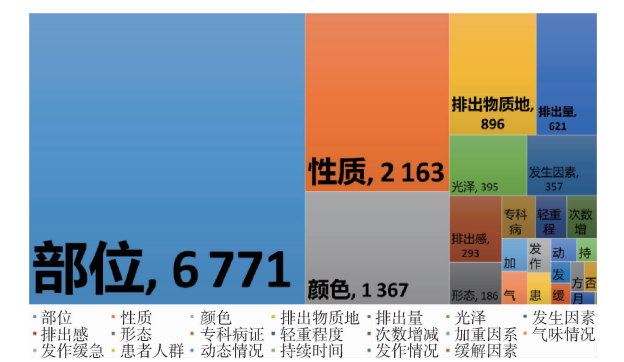


图 1 中医症状知识库属性分类频次

3.2 关联规则分析

3.2.1 症状之间的关联关系 选取知识库中的内科病证和症状数据 7 146 条，将数据导入到 SPSS Modeler，在 Apriori 模型中设置最低条件支持度为 3% ~5%，最小规则置信度为 50% ~80%，最大前项数为 2 ~5。进行多轮实验，根据结果可展示度最终选定参数为：最低条件支持度 3%，最小规则置信度 50%，最大前项数 5。将结果按照置信度由高到低排序得到 11 组症状组合，见表 2。

表 2 内科症状组合关联规则分析
(支持度≥3%，置信度≥50%)

序号	前项	后项	支持度 (%)	置信度 (%)
1	目眩	头晕	3.95	85.11
2	恶寒	发热	3.53	69.05
3	神疲	乏力	5.05	68.33
4	气短	乏力	4.88	56.90
5	畏寒	肢冷	3.11	56.76
6	无汗	恶寒	3.45	56.10
7	恶寒	无汗	3.53	54.76
8	耳鸣	头晕	4.21	54.00
9	无汗	发热	3.45	53.66
10	无汗	头痛	3.45	53.66
11	肢冷	畏寒	3.45	51.22

对频次前 10 位的症状进行分析，绘制内科症状关联网络图，见图 2。

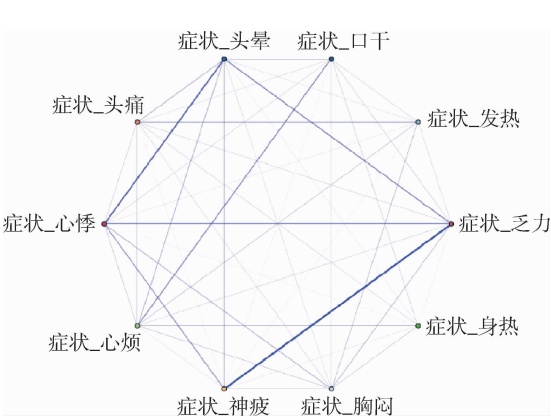


图 2 内科症状关联网络

气血两虚证是知识库中出现频次最多的证型，因此选取知识库中包含气血两虚证的病证和症状数据共 418 条。设置最低条件支持度为 10%，最小规则置信度为 80%，最大前项数为 5，得到 11 组症状组合，见表 3。统计频次前 10 位的症状，绘制气血两虚证症状关联网络，见图 3。

表 3 气血两虚证症状组合关联规则分析
(支持度≥10%，置信度≥80%)

序号	前项	后项	支持度 (%)	置信度 (%)
1	头晕、乏力	神疲	15.15	90.00
2	目眩、神疲	头晕	12.12	87.50
3	目眩、心悸	头晕	12.12	87.50
4	少气	懒言	10.61	85.71
5	气短、头晕	心悸	10.61	85.71
6	头晕、乏力、心悸	神疲	10.61	85.71
7	目眩	头晕	18.18	83.33
8	乏力	神疲	45.45	83.33
9	乏力、心悸	神疲	24.24	81.25
10	失眠	心悸	15.15	80.00
11	气短	心悸	22.73	80.00

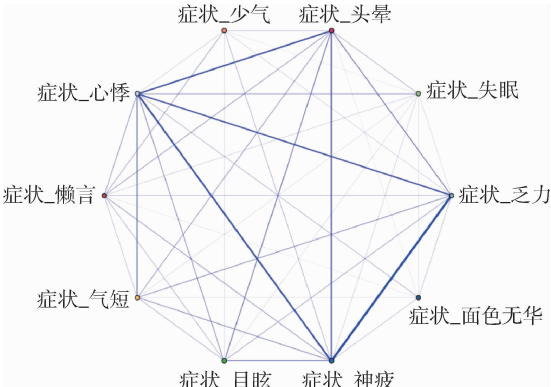


图 3 气血两虚证症状关联网络

3.2.2 骨干症状之间的关联关系 选取知识库中的病证以及骨干症状数据,去除重复得到 13 871 条数据,将数据导入 SPSS Modeler,选取 Apriori 模型,设置最低条件支持度为 10%,最小规则置信度为 50%,最大前项数为 5,得到 6 组骨干症状组合,见表 4。选择频次排名前 10 位的骨干症状,绘制骨干症状关联网络,见图 4。

表4 骨干症状组合关联规则分析
(支持度 $\geq 10\%$, 置信度 $\geq 50\%$)

序号	前项	后项	支持度 (%)	置信度 (%)
1	少神	乏力	10.91	62.15
2	汗异常	发热	11.39	61.33
3	口干	发热	19.90	60.73
4	少神	皮肤异常	10.91	55.52
5	小便异常	大便异常	19.90	54.50
6	小便异常,大便异常	口干	10.84	52.70

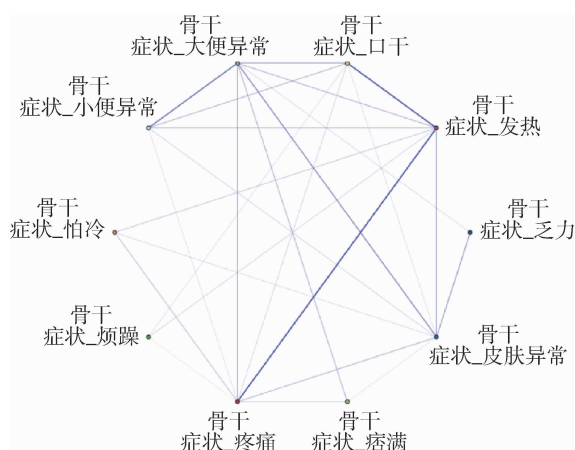


图 4 骨干症状关联网络

3.3 知识图谱构建与应用

3.3.1 知识图谱构建 将知识库的病证、症状、骨干症状、获取方式以及各属性存入 Excel，去除重复得到 15 450 条数据，通过 Python 程序将数据转化为三元组并存入 Neo4j 进行可视化，得到“病证-症状-属性”知识图谱。总计 7 948 个实体节点，

32 228 条关系，其中包含 6 类实体，分别为科室、病证、症状、骨干症状、获取方式和属性；5 类关系，分别为所属科室、症状、骨干症状、获取方式和属性。

3.3.3.2 节点、关系查询 (1) 节点查询。通过“MATCH (n: ‘病证’) where n. ‘病证’ = ‘恶露不下气血两虚证’ RETURN n”语句对“恶露不下气血两虚证”病证节点进行查询, 见图 5。“恶露不下气血两虚证”出现的科室为妇科, 主要症状有 7 个, 分别为面色苍白、恶露不下、恶露量少、耳鸣、头晕、心悸和小腹隐痛。(2) 关系查询。通过“MATCH (n: ‘骨干症状’) — (m) where n. 骨干症状 = ‘疼痛’ RETURN n, m”语句查询骨干症状为疼痛的节点并显示骨干症状为疼痛的所有症状节点, 查询到 567 个症状节点, 见图 6。

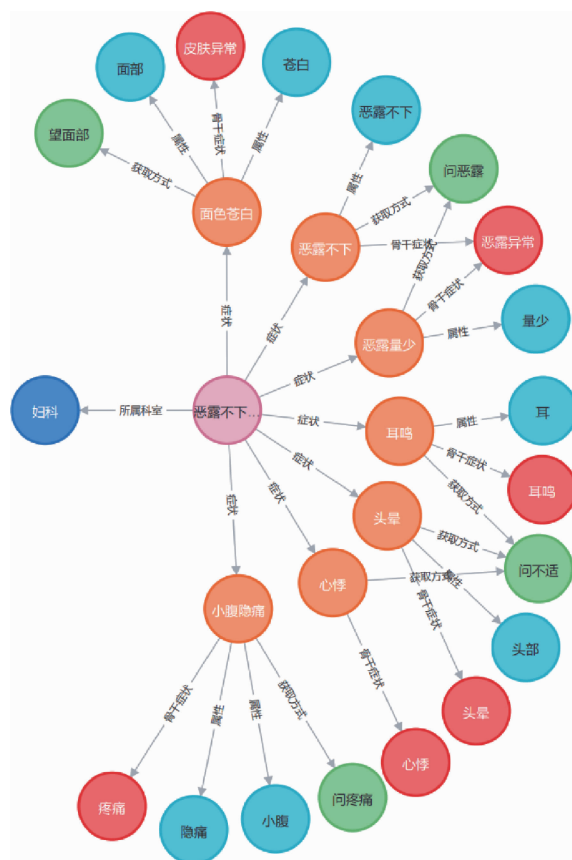


图5 症状节点查询结果

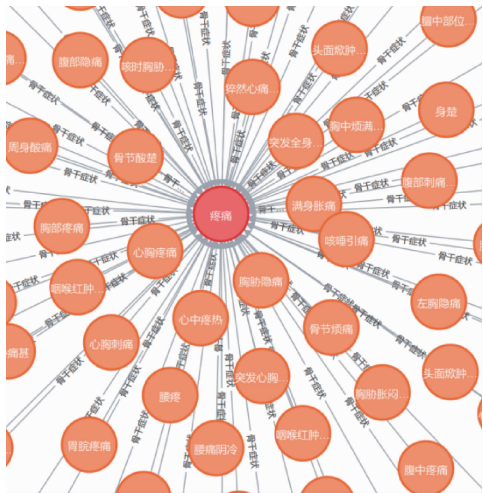


图6 属性节点与关系查询结果

4 讨论

4.1 概述

本研究选取《中医药学名词》中中医临床诊断术语 4 898 条，对其中包含的症状进行标注，并抽取症状实体以及相应的属性信息。对知识库中的数据进行描述性统计分析，通过 Apriori 模型对症状、骨干症状进行关联规则分析，发现强关联关系，通过 Neo4j 将知识库进行可视化，呈现“病证-症状-属性”之间存在的关联关系。

4.2 知识库统计结果分析

对知识库进行描述性统计以分析其中存在的规律。发现气血两虚证、阴虚火旺证和气滞血瘀证等证型出现频次较高。症状频次较高的有发热、头晕、乏力、头痛、神疲等，说明这些证型和症状在临床较为常见。对症状分科室进行统计，发现各科室高频症状也存在一定区别，反映出专科专病的病证特点。通过描述性统计分析发现知识库中的病症知识临床覆盖范围广泛，所蕴含的规律与临床具有一致性。充分证明了知识库的实用性和可行性，同时也体现出标注和抽取方法的正确性。

4.3 症状之间关系分析结果

症状是证型的重要组成，分析症状之间的关系

可为中医临床辨证提供依据与参考。选取知识库中内科症状数据进行关联规则分析，发现目眩与头晕、恶寒与发热等症状之间存在强关联关系，在内科疾病诊断时，存在强关联关系的症状同时发生的概率较高，可辅助临床医生进行诊断治疗。选取气血两虚证的症状进行关联规则分析，发现头晕、乏力、心悸、神疲等症状之间存在强关联关系，与气血两虚证的临床症状相符^[25]。说明关联规则分析的结果适用于临床，可为专科专病以及中医临床辨证提供参考。

4.4 对应症状和症状属性间关系分析

通过 Neo4j 将知识库可视化，直观反映病证之间的关系。构建中医症状知识图谱，梳理症状复杂属性并将中医临床复杂的隐性知识显性化，反映病证基本信息的同时将其中包含的属性信息进行展现。通过查询某一病证或症状节点，可找到对应症状以及症状属性之间的关系，为临床辨证论治提供参考。

5 结语

本研究建立了中医症状知识库，在后续研究中可加入治法方药等内容从而建立中医临床诊断知识库。目前知识图谱的应用主要为检索与查询，后续还应增加对中医辨证、临床决策支持等智能系统的应用。

参考文献

- 1 严石林, 沈宏春, 王浩中, 等. 中医症状的辨证意义及辨识方法 [J]. 云南中医学院学报, 2011, 34 (4): 1-3.
- 2 王哲睿, 李文娟, 王安安, 等. 基于数据挖掘的原发性胆汁性胆管炎中医用药规律研究 [J]. 中国中医药信息杂志, 2022, 29 (9): 53-58.
- 3 张庆, 轩扬, 马恒毅. 关联规则在中医药方剂挖掘中的应用 [J]. 医学信息学杂志, 2018, 39 (11): 69-72.
- 4 张洋, 雷蕾, 何建成. 基于数据挖掘的原发性高血压中医证素及症状/体征分布规律研究 [J]. 中国中医药信息杂志, 2019, 26 (1): 99-104.
- 5 王松, 李正钧, 杨涛, 等. 中医药知识图谱研究现状及

- 发展趋势 [J]. 南京中医药大学学报, 2022, 38 (3): 272 - 278.
- 6 于彤, 李敬华, 朱玲, 等. 中医临床知识图谱的构建与应用 [J]. 科技新时代, 2017 (4): 51 - 54.
- 7 张雨琪, 李宗友, 王映辉, 等. 赵炳南、朱仁康皮肤科流派用方经验知识图谱构建 [J]. 中国中医药图书情报杂志, 2021, 45 (2): 1 - 5.
- 8 赵燕华, 李悦, 张建勋. 中医类流感知识图谱的构建 [J]. 中华医学图书情报杂志, 2021, 30 (5): 24 - 30.
- 9 李明, 周强, 罗晓兰, 等. 中医症状术语标准及其分类体系研究 [J]. 中华中医药杂志, 2021, 36 (8): 4838 - 4842.
- 10 杨帆, 邓文萍, 孙静, 等. 中医临床基本症状信息分类与代码: T/CIATCM 020—2019 [S]. 北京: 中国中医药信息学会, 2019.
- 11 牟冬梅, 张艳侠, 黄丽丽, 等. 基于 SNOMED CT 和 FCA 的医学领域本体构建研究 [J]. 情报学报, 2013, 32 (6): 653 - 662.
- 12 储戟农, 彭莉, 崔蒙, 等. 论《中国中医药学主题词表》在中医药信息学中的作用 [J]. 国际中医中药杂志, 2014, 36 (12): 1060 - 1063.
- 13 周霞继, 刘保延, 郭玉峰, 等. 中医临床术语集在临床科研信息采集系统中的应用 [J]. 中国数字医学, 2010, 5 (10): 60 - 62.
- 14 任慧玲, 郭进京, 孙海霞, 等. 医学术语标准化研究的思考 [J]. 医学信息学杂志, 2018, 39 (5): 2 - 7.
- 15 全国科学技术名词审定委员会. 中医药学名词 [M]. 北京: 科学出版社, 2005.
- 16 崔昶旭, 朱建平, 洪梅. 中医学核心术语初探——以《中医药学名词》中医基础理论部分为例 [J]. 中华中医药杂志, 2021, 36 (3): 1560 - 1563.
- 17 朱文峰. 中医诊断学 [M]. 北京: 中国中医药出版社, 2007.
- 18 张启明. 中医症状学研究 [M]. 北京: 中国古籍出版社, 2013.
- 19 XIE W L, MAO S S, XIE D. Corpus construction for TCM clinical symptom based on information coding standard [C]. New York: Association for Computing Machinery, 2021.
- 20 DU Z Z, TANG D X, XIE D. Automatic extraction of clinical symptoms in traditional Chinese medicine for electronic medical records [C]. Houston: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020.
- 21 黄桥, 黄笛, 靳英辉, 等. 临床研究中常用的统计方法和常见问题 [J]. 中国循证心血管医学杂志, 2017, 9 (11): 1288 - 1293.
- 22 李文林, 赵国平, 陆建峰, 等. 关联规则在名医临证经验分析挖掘中的应用 [J]. 南京中医药大学学报, 2008 (1): 21 - 24.
- 23 邓宏勇, 许吉, 张洋, 等. 中医药数据挖掘研究现状分析 [J]. 中国中医药信息杂志, 2012, 19 (10): 21 - 23.
- 24 贾李蓉, 刘静, 于彤, 等. 中医药知识图谱构建 [J]. 医学信息学杂志, 2015, 36 (8): 51 - 53, 59.
- 25 郑筱萸. 中药新药临床研究指导原则 [M]. 北京: 中国医药科技出版社, 2002.

(上接第 14 页)

- 3 国务院办公厅. 关于促进和规范健康医疗大数据应用发展的指导意见 [EB/OL]. [2022-05-23]. http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.
- 4 宋保振. “数字弱势群体”权利及其法治化保障 [J]. 法律科学 (西北政法大学学报), 2020, 38 (6): 53 - 64.
- 5 国务院办公厅. 关于切实解决老年人运用智能技术困难实施方案的通知 [EB/OL]. [2022-05-23]. http://www.gov.cn/zhengce/content/2020-11/24/content_5563804.htm.
- 6 张新宝. 隐私权的法律保护 [M]. 北京: 群众出版社, 2004.
- 7 张宇清. “互联网+医疗服务”体系下患者隐私权保护研究 [J]. 医学信息学杂志, 2019, 40 (1): 6 - 11.
- 8 曾琼. 论患者隐私权保护中的权利冲突及其协调 [J]. 法商研究, 2009, 26 (6): 85 - 92.
- 9 郭进京, 张雪, 林鑫, 等. 国内患者隐私泄露情形及隐私保护现状分析 [J]. 医学信息学杂志, 2020, 41 (2): 21 - 28.
- 10 郝晓霞. 论电子医疗的患者信息保护 [D]. 济南: 山东大学, 2018.
- 11 蒋言斌, 李响. 我国医疗大数据患者隐私权保护及其模式选择 [J]. 医学与法学, 2018, 10 (1): 1 - 7.
- 12 牟朗宇. 大数据时代下个人隐私权法律保护的问题研究 [D]. 成都: 西南交通大学, 2017.