

DOI:10.13288/j.11-2166/r.2024.12.002

慢性萎缩性胃炎数据科学体系的构建与思考

孙建慧^{1,2,3}, 徐伟超^{2,3}, 张霞¹, 孙润雪², 陈艳哲², 王绍坡²,
王玉曼¹, 刘震⁴, 杜艳茹²✉, 杨倩^{2,3}, 姜建明^{1,3}

1. 河北中医药大学, 河北省石家庄市鹿泉区杏苑路3号, 050020; 2. 河北省中医院; 3. 河北省浊毒证重点实验室; 4. 中国中医科学院广安门医院

[摘要] 以慢性萎缩性胃炎(CAG)为例, 将数据科学前沿技术引入中医药传承创新发展领域, 为开展真实世界中医专科专病临床研究提供范例。在阐述数据科学内涵及其在中医药领域应用价值的基础上, 提出构建CAG数据科学体系的思路; 梳理构建CAG数据科学体系的路径, 即通过数据获取-知识表达-知识推理完成CAG数据库建立、知识库建立及病证诊断平台开发; 分析CAG数据科学在提升数据治理能力和知识发现效率、深化知识共享水平、推进学科交叉融合、加强产学研一体化进程方面的应用前景。

[关键词] 慢性萎缩性胃炎; 数据科学; 数据获取; 知识表达; 知识推理

慢性萎缩性胃炎(chronic atrophic gastritis, CAG)是常见的消化系统疾病, 其特征是胃黏膜上皮和腺体萎缩, 腺体减少, 黏膜变薄, 或者伴有肠化生以及异型增生^[1]。我国专家共识将胃黏膜萎缩和肠化生归类为胃癌前状态^[2]。《健康中国行动(2019—2030年)》中指出^[3]: 癌症是我国重大慢性病之一, 要强化癌症的中医药预防及早期干预能力。中医药在CAG的诊治中有独特的优势与特色^[4], 积极防治CAG, 截断“炎-癌转化”进程, 既是提升中医药防治胃癌能力的关键一环, 也是消化内科领域面临的重要挑战与需求。在中医临床诊疗实践过程中产生的大量的以语言、图像、舌象、脉象为代表的多样、复杂、非结构化的数据, 尚未形成系统知识化体系, 限制了中医临床数据的开发与利用。如何提升数据治理能力, 进一步转化为知识, 破解临床工程难题, 逐渐成为医学领域的共性需求。

1 数据科学内涵及其在中医药领域中的应用价值

在信息爆炸式增长的背景下, 数据俨然成为基

础性战略资源。大数据泛指海量的数据集, 蕴含着一定量的有待被挖掘、分析、提炼的碎片化知识集, 可服务于具体应用场景, 进而产生新的可应用的数据, 以待被进一步挖掘形成新的知识。如何从大数据中挖掘、凝练并转换为可用的“知识”, “知识”又如何物化成可计算的“数据”, 实现数据与知识交互驱动形式的大数据应用, 是信息学、统计学、计算机学等多个学科领域关注的热点^[5-6]。数据科学即是在大数据时代出现的新理论、新方法、新思维、新模式和新应用^[7-8], 该体系把数据看作现实世界在数字世界的映射, 通过利用和分析数据, 发挥其聚合效应、成像原理和态势感知能力^[9-10], 以更好地挖掘和研究数据背后所蕴藏的科学问题。

近年来, 在大数据、云计算技术迅速崛起与“健康中国”战略深入推进的叠加效应下, 数据科学引起了中医药领域专家的高度重视^[11-13]。如有研究^[14]建立了精神类疾病大数据科研管理平台, 以2013—2021年为时间跨度, 纳入病例63万余例, 后续利用机器学习技术建立双相情感障碍患者复发风险预测模型, 以数据提升疾病预测效率。另有研究^[15]采集了18 954例我国上消化道肿瘤高发区自然人群舌象, 建立高发区基线-病理-舌象信息数据库, 并运用数据科学理念, 通过大数据分析探索食管癌前病变到早期食管癌发展过程中舌象的转变规律。

基金项目:国家自然科学基金(82205314); 河北省重点研发计划项目(23377701D); 河北省自然科学基金(H2022423364); 河北省“三三三人才工程”资助项目(C20231017); 河北省高等学校科学技术研究(BJK2024183); 河北省中医药管理局中医药类研究课题(2024097, 2024103)

✉ 通讯作者: szydy@163.com

数据科学旨在从海量的、实时变化的临床大数据中获取人类可理解、机器可表征与可推理的结构化知识体系,将其嵌入到计算机系统中并建立推理机制,让计算机拥有知识并能进行辅助诊断、辅助决策,据此针对具体临床场景解决实际问题^[16-17]。数据科学应用于中医药领域,能够实现中医临床数据挖掘和知识引导的互补共享,促进数据增值、知识集成及数据知识的更新与共享,为病证诊疗、健康管理、科研教学提供新型支撑与保障^[18-20],是中医药向数字化、智能化发展的必经之路。

2 构建 CAG 数据科学体系的背景及思路

随着数据密集型研究范式的兴起,各学科领域大数据驱动的知识发现模式逐渐成为科学研究特色方式^[21]。在 CAG 的临床诊疗实际中会产生海量数据,包括中医四诊信息、中医病证资料数据、中药处方数据等,这些数据蕴含的中医药诊治规律有待挖掘。例如,舌为脾之外候,舌苔由胃气熏蒸而成,可见舌象与脾胃关系最为密切,其变化也能客观反映脾胃健康状态^[22]。腹诊是对患者腹部症状和体征的诊察,是中医四诊体系中的一个子系统,通过腹诊推断 CAG 患者体内的病理变化,也是 CAG 诊断的特色手段之一^[23]。这些数据加工处理后,能够转化为可分析、可整合、可提炼且具有一定辨证规律和中医特色的知识,即数据信息知识化,隐性知识显性化。以知识为核心,将知识以机器可理解可学习的方式表达,让机器模拟人的思维方式辅助诊断、决策、研究,以 CAG 为范例,探索数据科学在中医脾胃病领域的实践,并就此推动以数据、知识为核心的数据科学体系在其他专病领域的可持续发展。

3 构建 CAG 数据科学体系的路径

3.1 数据获取-建立 CAG 病例数据库

数据获取是按照统一标准,整合多个数据源,将非结构化、半结构化数据进行标注、转化,以结构化数据形式分层、分类归属,形成不同的子数据库,综合而成数据库的过程^[24-25],是后续知识表达的基础工作。建立 CAG 病例数据库是实现数据获取的关键。首先,收集整理国家标准和行业标准,包括名词术语标准、中医临床信息分类与代码、中医临床数据元标准、中医临床数据集标准等,通过抽取脾胃领域的数据集、构建词典、标注语料等过程,建立 CAG 专病数据集标准库。其次,对符合 CAG 诊断标准的病例信息,包括文本、声音、图

像等多种非格式化数据进行处理,运用歧义切分、校正等自然语言处理技术,同时通过分词、词性标注等具体方法,既能识别出相关语境中的新词语,还能定位关键词,适当解决提取冗杂无意义词语的问题^[26]。通过识别、抽提并规范以自然语言记录的临床资料,初步形成 CAG 临床数据集,使其在不同的研究及应用场景下快速传输、交换、共享。再次,设计中医临床病例报告表,进一步规范观察数据项的类型、名称及其内涵属性,通过填报该表来记录 CAG 临床观察数据。最后,将经过规范化整理的临床资料进行结构化处理,整合 CAG 症状、舌象、脉象等子数据库信息,形成字段命名规范、统一的 CAG 临床病例报告数据库。

3.2 知识表达-建立 CAG 病例知识库

知识表达是从庞大的数据库中挖掘有价值的隐藏关系并且加以分析,将知识转化为便于计算机存储、复制、分析和共享的可计算形式^[27-29],有利于后续知识推理,是实现大数据知识智能化的关键环节。建立 CAG 病例知识库则是知识表达的具体体现。在 CAG 专病数据库的基础上,根据病名、证候、中药分类、药名编码等相关信息关联,抽取临床特征信息,对数据库中的临床数据进行中医临床知识编码。即对数据库中的信息项按照其相应的属性进行代码化数据重构,形成编码化知识表达,以降低知识交流成本,提升学习效应。运用假设检验、直线回归、多重线性回归、逻辑回归、聚类分析等数据分析方法,将 CAG 与四诊特点、疾病史、体质特征、治疗方案、预后情况等编码化的实体信息之间进行联系,用节点网络图谱的形式来呈现层次关系,挖掘 CAG 症状、证候、治疗、效果知识资源之间的潜在联系,形成图谱式知识表达,以揭示真实世界中医临床行为的事实和规律。在经过深度数据挖掘与知识融合后,集中且清晰的知识结构能帮助医生、研究人员快速理清庞杂的知识点之间的联系和区别,极大提升 CAG 临床数据的治理和利用能力。

3.3 知识推理-开发 CAG 病证诊断平台

知识推理是基于已知知识和现有逻辑,对样本数据进行训练,识别错误知识或推理新知识的机制^[16]。以知识表达的结果为输入源,以自然语言处理、语义网、跨模态学习等技术手段,挖掘知识间的关联模式,推断知识间的潜在关系,预测知识间在未来空间是否会发生联系,是实现智能化的关键技术。建立 CAG 病证结合辅助诊断-决策-研究

软件是数据获取-知识表达-知识推理的实践产物。在 CAG 病例知识库基础上, 建立并优化智能模型。采用逻辑规则算法, 对 CAG 病证诊断原始语料以公式程序代入, 利用知识图谱中存在的结构化信息的逻辑, 模拟专家的逻辑推理能力, 并结合专家先验知识辅助推理。采用深度学习算法-卷积神经网络 (convolutional neural networks, CNN), 按阶层结构对输入信息进行平移不变分类, 对 CAG 知识图谱中的知识之间的关系进行学习, 通过知识之间的序列路径建立预测模型进行相关推理。通过让计算机“学习”专家的医疗知识, 模拟专家的思维和诊断推理, 基于患者原始医疗数据进行病证识别与病情分析, 从而给出可靠的诊断和治疗方案, 辅助医生进行诊断、决策工作。CAG 病证结合辅助诊断-决策-研究软件与医院临床电子病例系统关联后, 能够自动分析医院电子病历数据和网络临床研究数据, 自动识别临床研究和特定患者之间的潜在匹配, 并将匹配关联度推荐给医生和患者。对临床研究病例大数据实现自动获取、互用共享、多维呈现、高速计算和深度分析等各类实验活动, 可进一步满足辅助研究的要求。

4 CAG 数据科学体系的应用前景与展望

面向大数据时代深层次学科知识发现需求, 结合数据科学体系建设, 构建以 CAG 为例的病种研究新范式, 符合“数字中国”“健康中国”^[3, 30] 要求。湖仓一体、人工智能、区块链等数字技术在中医药领域的应用, 对推动应用新四诊技术以最合理的成本获得最大化的健康价值、应用生物化学技术表述中药治病机理和有效成分等方面具有广阔的实践前景和行为驱动力。但是随着数据海量式增长, 技术迭代式更新, 创新人才交叉式涌现, 以 CAG 为例的数据科学体系仍有广阔的发展空间。

4.1 不断提升数据治理能力和知识发现效率, 助力学科高水平建设

数据资源是战略性资源, 也是重要的生产要素。在 CAG 数据科学体系的构建中, 数据资源的质量把控是首要步骤。建立数据元标准规范, 是高效实现数据采集、整合、存储、共享的前提, 也是提高知识发现效率的关键。目前学者们在中医临床术语、诊疗标准与规范方面做出了巨大努力, 但中医语言具有复杂性、多样性、文化性等特色。在对经典术语进行修饰、量化的过程中, 借助计算机程序实现导推机制, 既能改变中医概念表达的模糊

性, 体现中医辨证论治的思维逻辑, 又能保留中医特色, 使标准、量化的名词术语与经典术语在不同场景应用中互相转化, 实现“取之于临床, 用之于临床”的模式, 促进相关学科的传承创新发展。

4.2 努力深化知识共享水平, 创新医联体建设模式

知识能够辅助研究者进行更高层次的思考和决策, 知识的累积和应用层次决定了智慧程度。通过技术手段整合四诊信息数据、影像数据、文献数据等各类临床大数据, 利用人工智能技术进行深度挖掘, 开发基于“数据-知识双驱动决策”的 CAG 病证诊断平台, 其最本质的目的是深化知识共享水平, 在临床、科研多领域发挥价值。但因数据兼容性、隐私性, 知识共享性、应用性等因素, 目前诊断平台尚无法大规模推广应用, 尤其是在医联体相关基层医疗机构。因此, 利用人工智能技术提升数据治理能力, 规范数据网络开放共享水平, 研发融合名老中医临证思维的专科专病辅助诊疗平台, 能够打通医联体内部之间的信息通道, 从数据共享向知识共享转化, 快速强化中医辨证施治思维, 进一步推进医疗资源、专家资源下沉, 创新多元主体协同参与的医联体建设模式。

4.3 持续推进学科交叉融合, 培育多学科交叉人才

在临床实践中发现新知识、提出新观点、构建新理论是一项复杂的系统工程, 仅仅靠单一的中医学科很难发展、突破。多学科交叉融合是科技创新和理论创新的源泉。以 CAG 数据科学体系构建过程为例, “需求+技术”双牵引是在研究实践中催生的多学科交叉范式, 中医学、现代医学、计算机科学、信息学、数学等不同专业的研究者通过内部联动、思维碰撞, 能够掌握多学科基础知识和跨学科研究方式, 扩充创新人才队伍。同时, 通过打破院系壁垒、学科壁垒, 设置与中医学交叉密集领域的学科基础课程, 采用有组织的形式代替集体灌输、独立自学等传统形式, 创新交叉课程授课方式与培养计划, 为学科建设提供思路与范例。

4.4 加强产学研一体化进程, 促进中医药知识成果转化

科研成果转化是产学研体系中的重要环节, 能够将潜在的社会价值、经济价值转化为生产力。CAG 数据科学体系构建的研究目的之一就是通过数据、知识与技术的融合, 辅助临床诊断、临床决策、临床研究, 改善临床结局, 助力循证研究。规则算法、深度学习等技术在临床研究中已取得一定突破, 但因政策不明、成果转化服务体系不完善等

因素无法有效转化应用。在后续的研究中,应明确成果应用场景、技术形态,以市场为导向,预估受众群体,进一步完善成果转化的政策,简化操作的可行性,以激励性政策增加转化动力,提供完整的转化产业链,促进中医药知识成果转化,从而使中医药知识成果惠及群众,以满足时代发展和民众日益增长的医疗保健需求。

5 小结

中医药领域中的数据科学是复杂多变的信息环境与中医临床大数据结合的产物。以 CAG 为例,构建以数据为基础、知识为驱动、技术为手段、人才为核心、应用为向导的数据科学体系,推广多学科交叉前沿研究成果,形成应用示范效应,将有望为数据科学体系在中医数字化领域的深入研究和创新应用提供进一步的有益探索和尝试。

参考文献

- [1] 中华医学会消化病学分会, 中华医学会消化病学分会消化系统肿瘤协作组. 中国慢性胃炎诊治指南(2022年, 上海)[J]. 胃肠病学, 2023, 28(3): 149-180.
- [2] 国家消化系统疾病临床医学研究中心(上海), 国家消化道早癌防治中心联盟, 中华医学会消化病学分会幽门螺杆菌学组, 等. 中国胃黏膜癌前状态及病变的处理策略专家共识(2020)[J]. 中华消化内镜杂志, 2020, 37(11): 769-780.
- [3] 健康中国行动推进委员会. 健康中国行动(2019—2030年)[EB/OL]. (2019-07-15)[2024-02-25]. https://www.gov.cn/xinwen/2019-07/15/content_5409694.htm?eqid=be683b7200051b8100000006645718a2.
- [4] 魏玮, 杨洋. 慢性萎缩性胃炎诊治现状及中医药治疗优势[J]. 中医杂志, 2016, 57(1): 36-40.
- [5] 许鸿本, 路熹雅, 刘怡馨, 等. 构建知识与数据双驱动中医诊疗系统的可行性及路径分析[J]. 中医杂志, 2022, 63(9): 801-805.
- [6] 贾君枝, 崔西燕, 任明. 数据与知识双驱动的知识组织系统构建框架研究[J]. 情报理论与实践, 2023, 46(10): 157-162.
- [7] TANWEER A, STEINHOFF J. Academic data science: transdisciplinary and extradisciplinary visions [J]. Soc Stud Sci, 2024, 54(1): 133-160.
- [8] 朝乐门, 卢小宾. 数据科学及其对信息科学的影响[J]. 情报学报, 2017, 36(8): 761-771.
- [9] 张清华, 高渝, 申秋萍. 数据科学: 从数字世界到数智世界[J]. 数据采集与处理, 2022, 37(3): 471-487.
- [10] 彭苏元, 张腊, 杨超, 等. 健康数据科学驱动的中西医融合研究[J]. 中国科学基金, 2023, 37(1): 85-91.
- [11] 宋淑洁, 毛树松, 张盼, 等. 基于大数据知识工程的中医临床研究新模式构建研究[J]. 时珍国医国药, 2022, 33(9): 2297-2299.
- [12] 王永炎, 王忠, 王燕平. 中医临床医学迎接大数据科学时代的刍议[J]. 中医杂志, 2022, 63(1): 2-4.
- [13] 胡文, 侯政昆, 刘凤斌, 等. 关于大数据时代的中医药临床研究的思考[J]. 世界科学技术-中医药现代化, 2019, 21(8): 1656-1661.
- [14] 咸春艳, 朱雪泉, 王刚. 精神疾病医疗大数据平台的构建及应用[J]. 中国医院管理, 2023, 43(12): 64-66.
- [15] 于然, 姜彦妮, 陈冬梅, 等. 基于舌象大数据的早期食管鳞癌中医证候转化规律分析[J]. 中医肿瘤学杂志, 2024, 6(1): 57-62.
- [16] 董文波, 孙仕亮, 殷敏智. 医学知识推理研究现状与发展[J]. 计算机科学与探索, 2022, 16(6): 1193-1213.
- [17] 杨涛, 王欣宇, 朱垚, 等. 大语言模型驱动的中医智能诊疗研究思路与方法 [J]. 南京中医药大学学报, 2023, 39(10): 967-971.
- [18] 赵汉青. 基于数据驱动的中医精准诊疗模型研究[D]. 北京: 中国中医科学院, 2019.
- [19] 杨善林, 丁帅, 顾东晓, 等. 医疗健康大数据驱动的知识发现与知识服务方法 [J]. 管理世界, 2022, 38(1): 219-229.
- [20] 林思雨, 周海涛. 人工智能融入高校教学科研管理的前景、风险和策略 [J]. 高校教育管理, 2023, 17(6): 21-30, 39.
- [21] 胡正银, 刘蕾蕾, 代冰, 等. 基于领域知识图谱的生命医学学科知识发现探析[J]. 数据分析与知识发现, 2020, 4(11): 1-14.
- [22] YUAN L, YANG L, ZHANG S, et al. Development of a tongue image-based machine learning tool for the diagnosis of gastric cancer: a prospective multicentre clinical cohort study [J]. EClinicalMedicine, 2023, 57: 101834. doi: 10.1016/j.eclim.2023.101834.
- [23] 刘启泉, 杜艳茹, 王志坤. 中医腹诊在胃肠病诊疗中的应用[J]. 中医杂志, 2006, 47(3): 224-225.
- [24] 李君, 周亚娜, 李晓东, 等. 中医临床大数据知识工程的应用价值及发展前景[J]. 时珍国医国药, 2023, 34(3): 769-770.
- [25] 郑庆华, 刘欢, 龚铁梁, 等. 大数据知识工程发展现状及展望[J]. 中国工程科学, 2023, 25(2): 208-220.
- [26] 赵君珂, 张振宇, 蔡开裕. 基于自然语言处理的医学实体识别与标签提取 [J]. 计算机技术与发展, 2019, 29(9): 18-23.
- [27] 肖勇, 张魁, 沈绍武. 等. 中医临床大数据知识工程技术平台架构与功能设计[J]. 时珍国医国药, 2023, 34(6): 1518-1520.

- [28] 李君, 李晓东, 张盼, 等. 中医临床肝病数据知识化及病例知识库构建研究 [J]. 中西医结合肝病杂志, 2023, 33(5): 394-397, 401.
- [29] 杜建, 孔桂兰, 李鹏飞, 等. 可计算医学知识的基本概念与实现路径 [J]. 情报学报, 2021, 40(11): 1221-1233.

- [30] 中国共产党中央委员会, 中华人民共和国国务院. 数字中国建设整体布局规划 [EB/OL]. (2023-02-27) [2024-02-25]. https://www.gov.cn/zhengce/2023-02/27/content_5743484.htm?eqid=ad458d2c0015b361000000036497e00d.

Construction and Thinking of Data Science System of Chronic Atrophic Gastritis

SUN Jianhui^{1,2,3}, XU Weichao^{2,3}, ZHANG Xia¹, SUN Runxue², CHEN Yanzhe², WANG Shaopo², WANG Yuman¹, LIU Zhen⁴, DU Yanru², Yang Qian^{2,3}, JIANG Jianming^{1,3}

1. Hebei University of Chinese Medicine, Shijiazhuang, 050020; 2. Hebei Hospital of Traditional Chinese Medicine; 3. Hebei Key Laboratory of Turbidity Toxin Syndrome; 4. Guanganmen Hospital, China Academy of Chinese Medical Sciences

ABSTRACT Taking chronic atrophic gastritis (CAG) as an example, the frontier technologies in data science have been introduced into the inheritance, innovation and development of traditional Chinese medicine (TCM), providing reference for conducting real-world clinical research on specialized diseases of TCM. This paper put forward the construction of CAG data science system by elaborating the connotation of data science and its application value in TCM, and discussed the path to build CAG data science system, namely through "data acquisition-knowledge expression-knowledge reasoning" to establish CAG database, knowledge base and develop diagnosis platform differentiating diseases and syndromes. Besides, this paper analyzed the prospects of CAG data science in improving data governance ability and knowledge discovery efficiency, deepening the level of knowledge sharing, promoting interdisciplinary integration, and strengthening the integration process of industry, academia and research.

Keywords chronic atrophic gastritis; data science; data acquisition; knowledge expression; knowledge reasoning

(收稿日期: 2023-11-20; 修回日期: 2024-03-01)

[编辑: 果 彤, 贾维娜]

广告

欢迎订阅 2024 年《中医杂志》

《中医杂志》(ISSN 1001-1668, CN 11-2166/R)是由中华中医药学会和中国中医科学院主办的全国性中医药综合性学术期刊。创刊以来始终坚持“以提高为主, 兼顾普及”的办刊方针, 是我国中医药界创刊早、发行量大、具有较高权威性和学术影响力的国家级中医药期刊之一, 是中文核心期刊和科技核心期刊、中国精品科技期刊、首届国家期刊奖获得者和中国期刊方阵双奖期刊、新中国 60 年有影响力的期刊、中国百强科技期刊、中国百种杰出学术期刊, 中医药科技期刊分级目录 T1 级期刊。荣获第五届中国出版政府奖期刊奖, 第二届、第三届、第四届中国出版政府奖期刊奖提名奖。

《中医杂志》主要栏目中当代名医和临证心得分别介绍名老中医和临床医生辨证用药治疗疑难病的经验; 临床研究介绍中医药治疗的新方法、新成果; 学术探讨、思路与方法、循证中医药、病例讨论、综述、百家园等栏目, 提供最新学术观点、研究成果与治疗方法, 成为学习中医药、研究中医药, 不断提高临床及研究水平的良师益友。

《中医杂志》为半月刊, 每月 2 日和 17 日出版, 每期定价 40.00 元, 全年 960 元。国内读者可以到全国各地邮局办理订阅手续 (邮发代号: 2-698), 也可以与本刊读者服务部联系邮购, 邮购免邮费。电话: 010-64035632, 010-64089195。国外发行: 中国国际图书贸易集团有限公司 (北京 399 信箱, 邮编: 100044, 代号: SM140)。

本社地址: 北京市东城区东直门内南小街 16 号, 邮政编码: 100700。

网址: <http://www.jtcm.net.cn>。

