本文引用:高 远, 张仕娜, 盛博洋, 董云春, 庞瑞涵, 晏峻峰, 彭清华. 由ChatGPT引发中医智能诊断研究中数据问题的思考[J]. 湖南中医药大学学报, 2023, 43(7): 1320-1324.

由ChatGPT引发中医智能诊断研究中数据问题的思考

高 远[1,2],张仕娜[1,2],盛博洋[1,2],董云春[1,3],庞瑞涵[1,3],晏峻峰[1,2,3]*,彭清华[1,2]*

1.湖南中医药大学,湖南 长沙 410208;2.湖南中医药大学中医诊断研究所,湖南 长沙 410208;

3.湖南中医药大学信息科学与工程学院,湖南 长沙 410208

〔摘要〕ChatGPT作为现象级产品受到广泛关注,本文从数据的数量、质量、驱动和保护4个方面,通过阐述ChatGPT的数据特点,引发中医智能诊断研究中数据问题的思考。中医智能诊断研究需要建立数据共享模式和中医诊断的数据元标准,以解决中医智能诊断研究中数据的数量和质量问题。多学科团队需要基于中医诊断知识进行推理,并且需要大数据对中医诊断知识推理规则调整参数、优化算法,共同构建融合知识与数据驱动方法,实现认知智能的中医智能诊断系统,以增强计算机推理结果的逻辑性和可解释性。中医学者在数据共享的同时,也要保护中医诊断数据的版权与安全。这些数据关系人民与国家的数据安全,关系我国全民健康情况和我国卫生事业的发展趋势。

〔关键词〕ChatGPT;中医;人工智能;辨证;数据;版权;安全

〔中图分类号〕R2 〔文献标志码〕A 〔文章编号〕doi:10.3969/j.issn.1674-070X.2023.07.027

Thoughts on data issues in intelligent diagnosis research of TCM prompted by ChatGPT

GAO Yuan[1,2], ZHANG Shina[1,2], SHENG Boyang[1,2], DONG Yunchun[1,3], PANG Ruihan[1,3], YAN Junfeng[1,2,3], PENG Qinghua[1,2]**

1. Hunan University of Chinese Medicine, Changsha, Hunan 410208, China; 2. Institute of TCM Diagnostics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China; 3. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

〔**Abstract**〕As a phenomenal product, ChatGPT has received widespread attention. This paper expounds the data features of ChatGPT from aspects of data quantity, quality, driving, and protection, prompting thoughts on data issues in intelligent diagnosis research of TCM. In order to solve the problems related with data quantity and quality in intelligent diagnosis research of TCM, it is necessary to establish data sharing model and data element standard for it. Furthermore, knowledge reasoning need to be conducted based on TCM diagnostic knowledge by multidisciplinary teams; big data is needed to adjust parameters and optimize algorithms for TCM diagnostic knowledge reasoning rules; TCM intelligent diagnostic systems should be jointly built to integrate knowledge and data-driven methods. The above measures help achieve cognitive intelligence and enhance the logic and interpretability of computer reasoning results. In addition, TCM academics should protect the copyright and security of TCM diagnostic data while sharing. These data determine the data security, and concern the health situation of

〔收稿日期〕2023-03-15

〔基金项目〕湖南省研究生科研创新项目(QL20220183,QL20220184,CX20210716);湖南省教育厅重点项目(21A0250);湖南中医药大学中医学一流学科开放基金项目(2022ZYX08,2021ZYX31);湖南中医药大学研究生创新课题(2022CX02,2022CX118)。

〔第一作者〕高 远,男,博士研究生,研究方向:中医诊断与人工智能。

〔通信作者〕* 晏峻峰,女,博士,教授,博士研究生导师,E-mail:junfengyan@hnucm.edu.cn;彭清华,男,博士,教授,博士研究生导师,E-mail:pqh410007@126.com。

Chinese people and the development of health service.

〔**Keywords**〕 ChatGPT; TCM; artificial intelligence; pattern differentiation; data; copyright; security

2022 年底，美国 OpenAI 公司发布聊天机器人程序 ChatGPT（Chat Generative Pre-trained Transformer）受到广泛关注。*Nature* 发表的论文中提到：ChatGPT 作为对话式人工智能（artificial intelligence, AI）对科学来说是游戏规则改变者[1]。ChatGPT 在对话中会保存使用者先前的对话信息，可根据之前的信息对当前问题做出即时的个性化回答。ChatGPT 这种会话方式使得人机交互变得更简单，极简的交互方式体现了极高的智能水平，将会取代传统关键词搜索式人机交互，成为互联网应用领域的新高地，尤其是在医疗领域的应用，前景广阔。

AI 技术已经成为重要的医疗辅助工具[2]。计算机模仿人类思维的智能学，可以应用于中医诊断领域，弥补中医诊断在客观化、标准化方面的劣势。中医诊断是基于四诊信息对人体的健康状态和病变本质进行辨识，做出概括性判断[3]。中医智能诊断是近年来智慧医疗领域的热点之一，ChatGPT 依靠数据、算法和算力的成功，成为数据驱动模式的现象级产品。程京院士曾指出：“无数据则无人工智能，数据是促进人工智能赋能中医药的重中之重。”[4]《“十四五”中医药信息化发展规划》基本原则指出：发挥数据作为新生产要素的关键作用，推进中医药数据的共享与安全[5]。因此，基于数据在中医智能诊断研究的基础性地位和推动性作用，本文通过阐述 ChatGPT 的数据特点，引发中医智能诊断研究中数据问题的思考。

1 数据数量问题

1.1 ChatGPT 拥有海量数据

ChatGPT 看似突如其来，实则是依赖 OpenAI 公司长期的积累。ChatGPT 通过学习大量的训练数据（包括现成文本和对话集合）模拟人的语言，根据上下文语境，使用 AI 模型生成人类可以理解的自然语言[6]，进而在人机交互过程中提供会话式服务，如同人类流畅回答各种问题。ChatGPT 基于 8000 亿个单词的语料库作为训练数据，通过训练数据学习、提取 1750 亿个参数。随着训练数据的增加、模型参数的优化，ChatGPT 处理复杂自然语言的准确性将不断提高[7]。2022 年 11 月 30 日，ChatGPT 开放公众测试，在海量跨语种的数据支持下，ChatGPT 的训练参数将进一步调整优化，下一代 ChatGPT 会有更优秀的表现。

1.2 中医智能诊断研究需要建立数据共享模式

中医智能诊断研究经历了从中医专家系统到传统机器学习再到现今大数据与深度学习探索的发展过程[8]。智能诊断研究需要全面、规范、准确地收集大量的人体四诊及其辨识结果的数据，数据包含文本、影像、音频、脉搏波等类型。中医学者往往以某位专家的临床案例或者某医院科室的病例作为研究基础，建立智能诊断系统。然而，每项研究的样本数据不同，中医学者也受到传统师承思想的束缚，许多研究的原始数据未公开，以致后续的中医学者难以重复验证，难以基于前人的研究基础展开新的研究，更难以实现多中心、大样本的研究。缺乏多中心、大样本数据是智能诊断研究所面临的一大阻碍。

智能诊断研究应跳出传统中医的经验性、模糊性的论述框架，利用 AI 技术，基于中医思维对患者进行个性化地病情判断。要实现智能诊断的精准化和高效化，需要大量的四诊及其辨识结果的数据支持，这些数据的共享将促进智能诊断研究的发展。目前，中医界尚缺乏大样本、公开的中医诊断数据库。因此，为实现优质中医药临床病例资源的共享，中华中医药学会开发建设中国中医药临床案例成果库[9]，截至 2023 年 3 月 12 日，共收录 935 篇临床案例，但还有待进一步发展。因此，开放共享的数据支持是中医智能诊断发展的重要前提，实现资源合理配置，也是推动智慧医疗发展的必要条件之一。

2 数据质量问题

2.1 ChatGPT 拥有高质量的数据标注

高质量的数据标注是支持 ChatGPT 模型训练的关键[10]。ChatGPT 通过“预训练语言模型”在大数据集上进行预训练，专业的标注人员会对 ChatGPT 生成的回答进行标注、评估和反馈，给出一个针对回答的分数或者标签，这些标注数据可以作为强化学习过程中的“奖励函数”，用来指导 ChatGPT 的参数调整，使得输出的文本符合人的认知[11]。数据标注的质量和准确性是 AI 算法有效运行的关键[10]，没有高

质量的数据标注，就不会有如今 ChatGPT 的成功。

2.2 中医智能诊断研究需建立中医诊断的数据元标准

数据标注是开发 AI 模型预处理的一部分，中医诊断数据不仅包含了症状、体征和微观参数等信息，还包含辨病、辨证等辨识结果。数据标注需要对这些数据添加标签为 AI 模型指定上下文，帮助做出准确预测。许多企业和研究机构推出了带标注的公开数据集[10]，然而，中医诊断数据载体丰富，具有较强的主观性，难以实现汇交融合，存在同名异义、异名同义、定义内涵外延描述不规范和定量描述较随意等现象[11-13]。因此，中医诊断数据的准确、规范表达，是智能诊断研究面临的一大难题。数据元是数据的基本单元，建立中医诊断的数据元标准是智能诊断研究中数据共享和交换的基础，以方便研究人员开展规范化的数据标注工作。

3 数据驱动问题

3.1 ChatGPT 主要是以数据驱动的 AI 模式

ChatGPT 主要运用以数据驱动的深度学习、强化学习等模型开展人类思维的模拟工作。ChatGPT 支持连续多轮对话，根据输入语句，基于词语序列的概率相关性分布进行建模，预测下一个时刻不同语句甚至语言集合出现的概率分布，自动生成答案。ChatGPT 相较于以往自然语言处理模型，具有更强的生成能力。但在逻辑推理复杂的领域，ChatGPT 缺乏逻辑推理能力，回答会出现缺乏人类常识的情况。因此，增强计算机推理结果的逻辑性和可解释性是 ChatGPT 发展的方向，也对智能诊断研究具有重要启示。

3.2 中医智能诊断研究需融合知识与数据驱动的 AI 模式

中医智能诊断若仅运用数据驱动的 AI 模式，需要巨大的算力与资金支持[11]，并且缺乏逻辑推理，导致结果的可解释性较低，决策者难以信任辨识结果的可靠性。中医辨证的核心是辨证知识的表示与推理[14]，然而，知识表示与推理的方法虽然可解释性强，但存在只能解决确定性问题的局限性，难以用于具有复杂性、非线性、模糊性的中医诊断。因此，类比中医师的成长过程，需要先学习中医经典以及中医教材，掌握中医诊断知识，基于中医先验知识进行中医诊断推理训练，在临床中面对一定数量的患者，巩固强化中医诊断思维。从 AI 角度，张钹院士等[15]通过阐述 AI 的发展历程，认为知识驱动方法的 AI 模式和数据驱动方法的 AI 模式均具有局限性，未来 AI 的发展方向需要融合知识驱动与数据驱动的 AI 模式。

作为大数据时代的知识工程集大成者，知识图谱为互联网时代的数据知识化组织和智能应用提供有效的解决方案[16]。中医学者也意识到为了能对智能诊断研究获得良好的理论依据，将知识图谱应用于中医诊断领域。通过对中医知识以及临床信息构建知识图谱，从而进行知识表示，运用知识推理对知识图谱中实体关系进行推理，进而建立基于中医知识图谱的诊断模型[17]。有研究基于中医电子病历，构建融合知识图谱的多通道中医辨证模型[18]。还有研究基于中医知识图谱的端到端模型，构建中医证候诊断决策系统[19]，为中医师提供决策支持。然而，中医诊断体系具有复杂性、经验性和模糊性问题，并且中医实体关系复杂，基于知识图谱有限的推理还不能胜任中医辨识的复杂过程。如何构建中医诊断知识推理模型，成为智能诊断研究的关键问题。融合知识驱动与数据驱动的知识表示和推理的认知图谱，或许是实现人工智能从感知智能向认知智能演进的重要方法[20]。有研究实验证明，认知图谱问答模型优于其他算法模型[21]，智能诊断研究未来也能走认知图谱的技术路线[22]。但这需要中医学、数学、人工智能、复杂系统等多学科专家“多层表述，逐级定量，多次迭代，逐步近似”[23]，共同构建融合知识驱动与数据驱动方法实现认知智能的中医智能诊断系统。

4 数据保护问题

4.1 ChatGPT 涉及数据版权保护与安全保护问题

OpenAI 公司主要通过网络爬虫技术获得超过万亿单词的公共语言数据集[11]。2022 年 11 月 30 日，ChatGPT 开放公众测试，用户在 ChatGPT 输入的内容以及用户反馈，会为下一代 ChatGPT 提供迭代训练数据。换而言之，ChatGPT 为用户提供便利的同时，也获取了用户的数据。并且 OpenAI 公司拥有的数据库和服务器均在美国，连同处于美国的亚马逊公司都警告员工不要与 ChatGPT 分享机密信息，因

为 ChatGPT 的某些回复看起来与亚马逊的内部情况十分相似[24]。《华尔街日报》记者 Francesco Marconi 就公开指责 OpenAI 公司，未经授权大量使用国外主流媒体的文章训练 ChatGPT 模型，并且从未支付任何费用[25]。换言之，ChatGPT 是通过爬取现成文本和对话集合进行数据训练并创作，回答并未溯源原始引用出处或者标注根据何处改编，并且可能会篡改作者原意，误导用户。OpenAI 公司并不否认 ChatGPT 的输出结果会侵犯他人作品的版权[26]。因此，ChatGPT 涉及数据版权保护与安全保护问题。

4.2　中医智能诊断研究的数据版权保护问题

版权，亦称“著作权”，指原创作者或组织对具有独创性并能以一定形式表现的智力成果依法享有的权利，我国的著作权保护期为 50 年。若在著作权保护期内未经作者允许擅自使用，可能会损害原创作者的合法权益[27]。2018 年 4 月国务院印发的《科学数据管理办法》指出：数据使用者应遵守知识产权规定，在工作中注明所使用和参考引用的数据[28]。多位专家针对 ChatGPT 带来的版权问题展开讨论，ChatGPT 是在受版权保护的素材里训练的语言模型，可能涉及侵犯他人的版权[29]，并且未溯源原始引用出处。但现阶段，也很难对人工智能生成物进行法律追究，未来是否对人工智能生成物采用版权保护还有待专家们的进一步讨论。

数据在共享过程中涉及大量版权问题，国内外很多高校都致力于数据版权的管理工作[30]。中医诊断数据包括了中医专家经验和临床诊疗数据，蕴含着实践经验、诊断方法、治疗方案和科研成果等方面的内容，反映中医药的独特优势和前沿成果。中医诊断数据的收集、整理和应用具有重要的意义和价值，可以推进中医药文化传播、提高人民健康水平。由于中医诊断数据承载了中华文化以及作者的智慧和心血，是一种知识产权，应该受到法律的保护。数据共享和版权保护二者之间存在冲突，但并不意味着不可调和，存在对立统一关系。数据共享和版权保护具有共同目标——促进知识创新和科技发展，实现数据生产者与使用者互利共赢[31]。

4.3　中医智能诊断研究的数据安全保护问题

数据泄露是指未经授权或许可，数据被泄露的行为。我国“十四五”大数据产业发展规划明确指出，数据是国家基础性战略资源[31]。中医药在现代医疗领域中占据重要地位。大数据时代，患者的个人诊疗信息经整合与分析后，具有极高的价值。对于个体而言，中医诊断数据可用于判断人体状态，包含生理病理特点、体质、病和证等方面的内容。随着数据的不断增长，中医诊断数据不仅可以从全局掌握我国全民健康情况，还可以预测我国卫生事业的发展趋势，甚至为非我国授权的研究提供数据支持。此外，ChatGPT 的服务器在国外，在大数据环境下，如果中医诊断数据被掌握在国外公司的手中，这些数据很难受到中国法律的保护，极易被非法使用，造成损失。这将对中医药文化的传承和发展带来不利影响，也可能危及我国的数据安全和国家利益。因此，必须加强对中医药数据的保护和管理，确保其合法合规使用，促进中医药事业的健康发展。

5　结语

ChatGPT 作为新一代人工智能技术，将对医疗领域带来巨大影响。中医师诊断过程中需要对望、闻、问、切四诊信息进行分析，融合中医知识与临床案例相结合的思维方式，进行辨识。因此，中医智能诊断研究也需要基于中医诊断规则与数据迭代，对影像、音频、文字、脉搏波等四诊信息进行多模态信息融合。OpenAI 发布 ChatGPT-4 是大型多模态 AI 模型（接受图像和文本输入、文本输出），ChatGPT 的更新迭代速度十分迅速，已经超乎想象。ChatGPT 的数据特点是成功因素之一，中医智能诊断研究可以从 ChatGPT 的成功获得启迪。因此，中医智能诊断研究需要建立数据共享模式和中医诊断的数据元标准，以解决中医智能诊断研究的数据数量和质量问题。多学科团队需要基于中医诊断知识进行知识推理，并且需要大数据支持对诊断知识的推理规则调整参数、优化算法，共同构建融合知识与数据驱动方法，实现认知智能的智能诊断系统，以增强计算机推理结果的逻辑性和可解释性。

ChatGPT 为人们带来便利的同时，数据版权与安全保护问题也需要引起足够的思考。一方面，中医诊断数据包括中医专家诊断经验和临床诊疗数据，蕴含着实践经验、诊断方法、治疗方案和科研成果等方面的内容，承载了中华文化以及作者的智慧和心血；另一方面，数据是国家基础性战略资源，关系人民与国家的数据安全，反映我国全民健康情况和我

国卫生事业的发展趋势。因此,未来需要培养中医学者的数据保护意识,加强数据安全人才队伍建设,加大经费投入数据保护,建立健全数据保护的法律法规,加强对中医药数据的管理和监管,共同促进中医药信息化发展,为中医药现代化插上腾飞的翅膀。

参考文献

[1] VAN DIS E A M, BOLLEN J, ZUIDEMA W, et al. ChatGPT: Five priorities for research[J]. Nature, 2023, 614(7947): 224-226.

[2] DUAN Y Y, LIU P R, HUO T T, et al. Application and devel opment of intelligent medicine in traditional Chinese medicine[J]. Current Medical Science, 2021, 41(6): 1116-1122.

[3] 李灿东. 中医诊断学[M]. 4 版. 北京: 中国中医药出版社, 2016: 202-205.

[4] 林静怡, 李诗翩, 郭 义, 等. 人工智能助力中医药发展现状、问题及建议[J]. 世界中医药, 2022, 17(6): 864-867.

[5] 国家中医药管理局. "十四五"中医药信息化发展规划[EB/OL]. (2022-11-25)[2023-02-18]. http://www.natcm.gov.cn/guicaisi/zhengcewenjian/2022-12-05/28427.html.

[6] 冯志伟, 张灯柯, 饶高琦. 从图灵测试到 ChatGPT: 人机对话的里程碑及启示[J]. 语言战略研究, 2023, 8(2): 20-24.

[7] 齐 旭, 刘 晶, 宋 婧. 没有百亿参数的大模型, 不敢奢谈 ChatGPT[EB/OL]. (2023-02-24)[2023-02-25]. https://www.sohu.com/a/645375330_121134737.

[8] 文志华, 夏帅帅, 刘东波, 等. 中医智能辨证诊断技术的演进与问题探讨[J]. 世界科学技术(中医药现代化), 2021, 23(11): 4298-4304.

[9] 中医药发展研究.中国中医药临床案例成果库第一批案例成功入库[EB/OL].(2020-02-26)[2022-01-05]. https://mp.weixin.qq.com/s/RUAqoEL0zQlI-Hfbd_AwyA.

[10] 蔡 莉, 王淑婷, 刘俊晖, 等. 数据标注研究综述[J]. 软件学报, 2020, 31(2): 302-320.

[11] 李新龙, 黄培冬, 朱 爽, 等. 智能化挖掘中医临床诊疗数据面临的问题和挑战[J]. 中华中医药杂志, 2022, 37(12): 6962-6965.

[12] 王天芳, 李灿东, 朱文锋. 中医四诊操作规范专家共识[J]. 中华中医药杂志, 2018, 33(1): 185-192.

[13] 肖晓霞, 晏峻峰, 刘东波, 等. 中医临床症状数据元提取探析(英文)[J]. 数字中医药(英文), 2018(1): 37-46.

[14] 韦昌法, 晏峻峰. 从知识表示与推理方法探讨中医数字辨证发展[J]. 中华中医药杂志, 2019, 34(10): 4471-4473.

[15] 张 钹, 朱 军, 苏 航. 迈向第三代人工智能[J]. 中国科学(信息科学), 2020, 50(9): 1281-1302.

[16] 王 萌, 王昊奋, 李博涵, 等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展, 2022, 59(9): 1947-1965.

[17] 韦昌法, 罗丽琴, 晏峻峰. 中医数字辨证配套医案智能采集与分析系统构建研究[J]. 湖南中医药大学学报, 2020, 40(1): 70-74.

[18] 叶 青, 张素华, 程春雷, 等. 融合知识图谱的多通道中医辨证模型[J]. 科学技术与工程, 2022, 22(21): 9190-9198.

[19] YANG R, YE Q, CHENG C L, et al. Decision-making system for the diagnosis of syndrome based on traditional Chinese medicine knowledge graph[J]. Evidence-Based Complementary and Alternative Medicine, 2022, 2022: 1-9.

[20] 徐 菁.《人工智能之认知图谱》重磅发布[EB/OL].(2020-08-28)[2022-02-25]. https://www.aminer.cn/research_report/5f48541f3c99ce0ab7bca8fc?download=false.

[21] 袁 满, 张维罡, 李明轩. 基于认知图谱的智能问答系统推理模型研究[J]. 吉林大学学报(信息科学版), 2021, 39(5): 589-595.

[22] 唐恒安, 唐三歌, 唐志书. 论中医思想体系的第二次抽象[J]. 中华中医药杂志, 2022, 37(9): 4897-4902.

[23] 余振苏, 倪志勇. 人体复杂系统科学探索[M]. 北京: 科学出版社, 2012: 63.

[24] 网易科技报道.亚马逊警告员工不要向 ChatGPT 分享机密,包括正在写的代码[EB/OL].(2023-01-25)[2023-02-25].https://www.163.com/tech/article/HRUIJJNU00097U7T.html.

[25] Francesco Marconi. ChatGPT is trained on a large amount of news data from top sources that fuel its AI[EB/OL].(2023-02-15)[2023-02-25].https://twitter.com/fpmarconi/status/1625867414410825728?cxt=HHwWgMC4_ZLznpAtAAAA.

[26] 康添雄. ChatGPT 产生版权纠纷的可能与不可能[EB/OL].(2023-02-20)[2023-02-25]. https://column.chinadaily.com.cn/a/202302/20/WS63f31b7ba3102ada8b22fc56.html.

[27] 中国人大网.中华人民共和国著作权法[EB/OL].(2020-11-19)[2023-02-20].http://www.npc.gov.cn/npc/c30834/202011/848e73f58d4e4c5b82f69d25d46048c6.shtml.

[28] 国务院办公厅. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL].(2018-04-04)[2023-02-20].https://most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2018/201804/t20180404_139023.html.

[29] 赵新乐, 朱丽娜. ChatGPT 爆火, 带来哪些版权问题? [EB/OL].(2023-02-16)[2023-02-25].https://ie.bjd.com.cn/5b165687a010550e5ddc0e6a/contentApp/5b16573ae4b02a9fe2d558f9/AP63ed97e4e4b03a6b6edc4a1c?isshare=1.

[30] 尹梦茹. 高校科学数据的数字版权管理研究[J]. 图书馆研究, 2022, 52(2): 1-8.

[31] 工业和信息化部. 工业和信息化部关于印发"十四五"大数据产业发展规划的通知[EB/OL].(2021-11-30)[2023-02-25]. https://twitter.com/fpmarconi/status/1625867414410825728?cxt=HHwWgMC4_ZLznpAtAAAA.

(本文编辑 匡静之)