

医药卫生与人口健康领域科学数据引用行为与可回溯性研究^{*}

张怡然¹ 张敬晨² 孙婧雯² 王嘉阳² 张更路¹ 周伟²

(¹ 华北理工大学经济管理学院 唐山 063210 ² 中国医学科学院国家人口健康科学数据中心 北京 100730)

〔摘要〕 **目的/意义** 通过对医药卫生与人口健康领域科学数据共享机构的科学数据引用特征进行调查和分析, 为科学数据开放共享和规范引用提供借鉴和参考。**方法/过程** 以国家人口健康科学数据中心为例, 采用文献分析法, 深入分析该领域科学数据引用行为及其可回溯性。**结果/结论** 医药卫生与人口健康领域科学数据引用存在标注位置不统一、引用元素缺失等不规范引用行为, 大量开放共享的科学数据无法有效溯源, 亟须完善科学数据引用标准、加强科学数据平台管理、提高科研人员引用意识, 促进科学数据共享引用。

〔关键词〕 医药卫生与人口健康; 科学数据; 数据引用

〔中图分类号〕 R-058 **〔文献标识码〕** A **〔DOI〕** 10.3969/j.issn.1673-6036.2024.09.004

Study on Scientific Data Citation Behavior and Its Traceability in the Field of Medical and Population Health

ZHANG Yiran¹, ZHANG Jingchen², SUN Jingwen², WANG Jiayang², ZHANG Genglu¹, ZHOU Wei²

¹ School of Economics and Management, North China University of Science and Technology, Tangshan 063210, China; ² National Population Health Data Center, Chinese Academy of Medical Sciences, Beijing 100730, China

〔Abstract〕 **Purpose/Significance** By investigating and analyzing the characteristics of scientific data citation of scientific data sharing institutions in the field of medical and population health, the paper provides references and guidance for open sharing and standardized citation of scientific data. **Method/Process** Taking the National Population Health Data Center as an example, the literature analysis method is used to analyze the scientific data citation behavior and its traceability in this field. **Result/Conclusion** It is found that there are non-standard citation behaviors such as inconsistent labeling positions and missing citation elements in the scientific data in the field of medical and population health, and a large number of open and shared scientific data cannot be effectively traced. It is urgent to improve the standards of scientific data citation, strengthen the management of scientific data platform, enhance the awareness of scientific researchers, and promote scientific data sharing and citation.

〔Keywords〕 medical and population health; scientific data; data citation

1 引言

科学数据开放共享是现代科学发展的必然要求, 目前许多发达国家和国际组织在科学数据开放共享方面已经展开了一系列的研究与实践^[1]。科学数据引用将科学数据作为一种重要的引证材料, 将科学数据像参考出版物一样进行引用和说明^[2]。已

〔修回日期〕 2024-06-03

〔作者简介〕 张怡然, 硕士研究生; 通信作者: 周伟, 高级工程师。

〔基金项目〕 中国工程院重大咨询研究项目 (项目编号: 2019-ZD-23); 国家骨科与运动康复临床医学研究中心创新基金重点项目 (项目编号: 23-NCRC-CXJJ-ZD4)。

有学者^[3]对期刊科学数据引用情况进行调研,提出科学数据引用可以提高科学数据的透明度,对科学研究有积极的影响。科研文章中科学数据的规范引用,还有助于出版机构审核人员了解所用科学数据的真实性和可靠性,有利于科学过程的重现与研究成果的印证,建立数据信用^[4]。尤其是在医药卫生与人口健康领域,亟须通过规范科学数据引用来遏制数据造假、数据篡改等学术不端行为的发生,保护数据生产者的知识产权并提高科学数据复用率。

2 研究现状

近年来国际数据委员会等就如何规范科学数据引用及完善科学数据基础设施建设等问题多次举办会议讨论研究,提出了一些推荐性的科学数据引用规范^[5]。国际非营利组织、科研项目组以及数据仓储中心也陆续制定科学数据的引用标准,并逐步纳入 APA 格式、MLA 格式、Chicago 格式等论文引用规范^[6]。

国内外学者对科学数据引用行为、引用规范性、引用标准等方面的现状与问题进行了研究。丁楠等^[7]和屈亚杰等^[8]对社会学领域科学数据引用行为进行统计分析,发现引用不规范行为普遍存在。司莉等^[4]和邱均平等^[9]以图书情报领域作为研究对象调查科学数据引用规范程度,发现不同类型的科学数据标注方式缺乏统一性。Fenner M 等^[10]和 Cousijn H 等^[11]分别为科学数据储存库和学术出版商提供实现科学数据引用的实用路线图。王文琪等^[12]提出应加快构建规范科学数据引用机制。Force M M 等^[13]提出一套科学数据引用的索引概述,以解决科学数据引用、重用的问题。Christensen G^[14]提出可以通过发布科学数据出版政策鼓励数据作者共享科学数据。

在医药卫生与人口健康领域, Park H 等^[15]以遗传学为样本,研究其科学数据引用共享、重用现状; Florian L 等^[16]发现神经科学和分子生物学领域科学数据论文中的科学数据总量明显低于总体科学数据论文发表量的平均值。邱玉红等^[17]以美国国家生物技术信息中心为例,得出随着科学数据引用标

准和规范的不断完善,该领域科学数据规范引用行为越来越普遍的结论。

综上所述,目前我国对科学数据引用的研究主要集中于自然科学领域和社会学领域,以医药卫生与人口健康领域的科学数据引用现状为目的的研究主要集中在国外,很少见我国针对该领域科学数据引用现状及其特征的研究。因此,本研究选取我国国家级医药卫生与人口健康领域科学数据中心作为研究对象,深入调查该领域科学数据被引用现状,分析科学数据引用存在的问题,为科学数据开放共享和规范引用提供借鉴和参考。

3 资料与方法

3.1 案例选择

国家人口健康科学数据中心 (national population health data center, NPHDC) 是人口健康相关领域亚洲地区首家通过 CoreTrustSeal 全球核心储存库国际认证的数据仓储,是我国该领域具有代表性的科学数据平台。鉴于此,本文选取 NPHDC 作为科学数据来源对我国医药卫生与人口健康领域科学数据引用情况进行研究。

3.2 文献来源

采用文献研究方法,选取科学数据来源于 NPHDC 的学术文献为研究对象,共得到文献 249 篇。文献来源于以下两部分。一是 NPHDC 提供的人口健康共享杯大赛成果集。该大赛是每年定期举办的科技资源共享服务创新大赛,为参赛人员提供 NPHDC 科学数据资源,获奖文章被收录于大赛成果集。纳入 2016—2022 年使用 NPHDC 科学数据的人口健康共享杯大赛成果,共得到文献 141 篇。二是中国知网、PubMed 等文献数据库。以“国家人口与健康科学数据共享服务平台” or “国家健康人口科学数据中心” or “人口健康共享杯” or “National Population Health Data Center” 为关键词进行全文精确检索,时间范围 2016 年 1 月 1 日—2022 年 12 月 31 日,检索到文献 167 篇。经人工排查剔除未使用该平台科学数据和重复文献,最终得到有效文献

108 篇。

3.3 科学数据引用分析的维度与指标

研究科学数据引用特征，需要构建相应的分析维度。目前科学数据集引用尚未有可靠的机器标识方法，且尚未形成规范标准，因此参考国外重要科学数据管理机构的科学数据引用指南^[5]、《FORCE11 数据引用原则》^[18] 及我国《信息技术 科学数据引用》^[19] 等引用规范，人工识别并总结医药卫生与人口健康领域的科学数据集引用特征标注内容。

3.4 统计分析

采用 Excel 2019 建立数据库对引用位置、引用元素、引用行为、可回溯性等指标进行标注，使用数量与百分比进行数据描述与分析，采用柱状图和折线图绘制趋势图并进行分析。

4 研究结果

4.1 科学数据引用特征分析

4.1.1 科学数据引用位置 根据学术文献的组成结构将科学数据引用位置分为摘要、引言、文献综述、资料与方法、结果与讨论、结论、致谢、参考列表、角注和附表 10 部分。科学数据引用位置统计结果，见图 1。

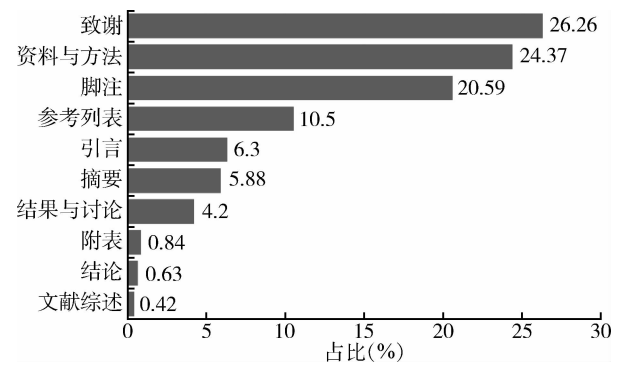


图 1 科学数据引用位置

在致谢处引用科学数据的最多，占比 26.26%，有助于促进学术共同体内部交流和互动；在资料与方法处引用的占 24.37%，以提供基础数据和介绍

研究所用数据概况；20.59% 在脚注位置引用科学数据。这 3 部分是医药卫生与人口健康领域主要的科学数据引用位置。

4.1.2 科学数据引用元素 科学数据引用元素以作者、数据集名称、出版商、标识符为必备元素。只有参考列表、脚注、附表、资料与方法部分相对全面地标注了 4 种必备元素。作者和标识符在文献中的标注率相对较低。科学数据使用者对作者的引用主要集中于参考列表和脚注位置，对出版商的引用大多集中于资料与方法 and 致谢部分，见表 1。

表 1 NPHDC 科学数据引用元素

引用位置	作者	数据集名称	出版商	标识符	合计
摘要	0	23	24	0	47
引言	1	23	27	0	51
文献综述	0	1	2	0	3
资料与方法	3	83	112	2	200
结果与讨论	0	13	20	0	33
结论	0	1	4	0	5
致谢	3	26	112	0	141
参考列表	34	45	49	50	178
脚注	17	84	96	17	214
附表	2	3	4	2	11
合计	60	302	450	71	883
占比 (%)	6.80	34.20	50.96	8.04	100

4.1.3 科学数据引用行为统计 研究人员结合科学数据引用位置与引用元素两个特征将科学数据引用行为分为规范引用、不规范引用和未引用 3 类。科学数据引用行为统计结果，见图 2，规范引用仅占 12.45%。科学数据引用行为类目，见表 2。

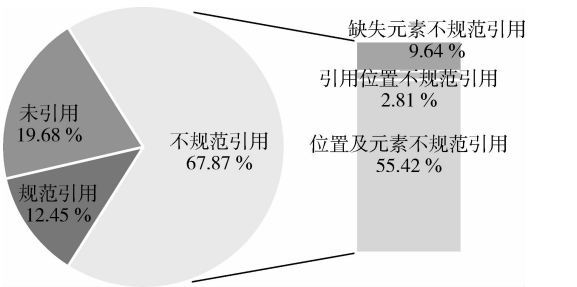


图 2 科学数据引用行为

表2 科学数据引用行为类目

引用行为	二级分类	定义	示例
规范引用		必备引用元素齐全，且在参考列表处引用	参考文献处标注“<作者>.<数据集名称>.<出版商>,<标识符>.”
不规范引用	缺失元素不规范引用	必备引用元素缺失，在参考列表处引用	参考文献处标注“<作者>.<数据集名称>.”
	引用位置不规范引用	必备引用元素齐全但未在参考列表处引用	脚注处标注“本文使用了<作者>.<数据集名称>.<出版商>,<标识符>.”
未引用	位置及元素不规范引用	必备引用元素缺失且未在参考列表处引用	正文处标注“数据来源于<作者>.”
		使用了科学数据但并未对科学数据来源做任何标注或注释	

4.1.4 科学数据引用可回溯性统计 选择解析地址与主题词两种元素对可回溯性进行统计。71.49%（178 条）的科学数据在数据仓储中回溯失败。在回溯成功的 71 条科学数据引用中有 22 条的引用标注可根据解析地址直接查询，有 49 条引用片段可根据主题词检索到相关科学数据，但存在凭数据集名称无法确定科学数据版本以及具体科学数据仓储位置的情况。

4.2 科学数据引用发展趋势分析

4.2.1 引用行为发展趋势 NPHDC 各年度科学数据引用行为基本统计情况，见图 3。科学数据每年引用量均在 27 篇以上。2019 年科学数据不规范引用行为达 90%；此后规范引用行为比例逐渐提高，2022 年增长到 39.39%；不规范引用逐渐减少，2022 年降至 9.09 %。



图3 NPHDC 科学数据引用行为分布

4.2.2 可回溯性发展趋势 科学数据引用可回溯性总体趋势，见图 4。2016—2022 年回溯失败的科学数据占比逐年下降。自 2018 年根据解析地址回溯成功的数量逐年增加，且增幅明显，尽管 2021

年有所下降，但同时回溯失败的科学数据量也在急剧降低。

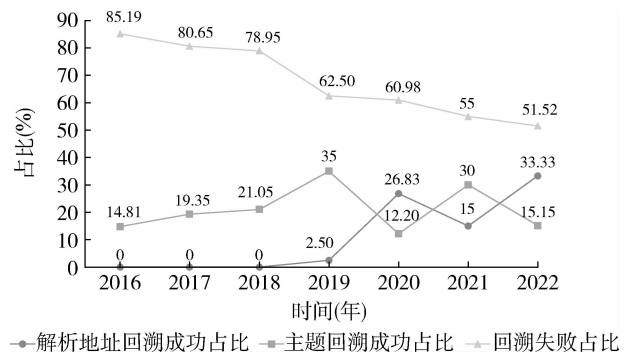


图4 NPHDC 科学数据可回溯性

5 讨论与建议

5.1 讨论

5.1.1 科学数据引用行为探讨 使用 NPHDC 科学数据的文献数量逐年递增，可见医药卫生与人口健康领域科学数据共享和重用行为已有明显增多的趋势，但该领域的科学数据引用仍存在较多不规范之处。首先，科学研究人员在进行科学数据引用时，并未单独描述出版商、数据集名称、作者等必备元素，不利于激发科学数据生产者的共享意愿。其次，现有大多数引用散落在正文中，在参考列表处的引用很少，从而难以量化评估科学数据的引用。最后，未注明 DOI 或 URL 等唯一标识符，导致很难有效溯源原始科学数据，使引用文献无从考证其结果真实性。

5.1.2 科学数据可回溯性探讨 本研究中 8.83%

的科学数据可以根据解析地址回溯成功。科学数据引用中的解析地址可以使科技文献与科学数据形成相互引用的关联环境,进一步促进科技资源的开放共享,且解析地址回溯成功的论文可以更直接快速地帮助读者成功溯源原始科学数据。但 71.49% 的科学数据引用难以获取原始科学数据出处。回溯失败的原因主要有:科学数据共享平台的引用说明不够规范、学术期刊未对科学数据引用方式进行规范要求、科研人员自身缺乏规范引用意识。

5.2 建议

5.2.1 完善科学数据引用标准 部分已出台的国家科学数据引用标准存在与医学健康科学数据难以匹配的窘境,因此应针对该领域科学数据的特性,通过出版案例研究和手册,制定专门的引用规则。定期审查和更新现有的国家科学数据引用标准,使之能够适应新的科学数据类型和研究方法。

5.2.2 加强科学数据平台管理机制 科学数据平台应建立健全更加开放透明的科学数据引用体系。首先,确保其注册的数据集有完整的元数据,并参照文献数据库提供科学数据集的标准引用格式,应至少包括作者、数据集名称、出版商、标识符 4 种必备元素,为正确引用提供便利。其次,提供教育资源,指导如何有效地管理和引用科学数据。最后,科学数据平台应提供数据使用与共享等数据影响力相关信息,以激励科研人员访问和引用科学数据,促进领域内科学数据共享。

5.2.3 提高科研人员科学数据引用意识 为确保研究的透明度、可重复性,科研人员应使用统一和公认的科学数据引用格式,遵守数据提供者设定的任何版权或使用许可条件。引用不同类型的科学数据时,要尽可能在参考文献处标注科学数据的唯一标识符,以便读者直接访问原始科学数据。同时,可以通过共享自己的二次数据到公共存储库或数据期刊来促进科学数据引用。

6 结语

本研究发现医药卫生与人口健康领域科学数据

引用存在标注位置不统一、引用元素缺失等不规范引用行为,科学数据引用标准的制定和统一仍需完善。同时,科学数据的不规范引用导致大量开放共享的医药卫生与人口健康科学数据无法有效溯源。虽然科学数据可回溯性整体呈现不断上升趋势,但仍需通过完善科学数据平台管理机制以及加强科研人员引用意识来规范科学数据引用。

作者贡献: 张怡然负责数据采集与分析、论文撰写;张敬晨负责数据采集、整理与管理、实验方法设计、论文审核与修订;孙婧雯、王嘉阳负责数据整理与管理、论文审核与修订;张更路、周伟负责提供指导、论文审核与修订。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 盛小平,武彤. 国内外科学数据开放共享研究综述 [J]. 图书情报工作, 2019, 63 (17): 6-14.
- 2 BROWN R F. The importance of data citation [J]. Bioscience, 2021, 71 (3): 211.
- 3 GAWNE L, KRAJCIK C, ANDREASSEN H N, et al. Data transparency and citation in the journal gesture [J]. Gesture, 2019, 18 (1): 83-109.
- 4 司莉,伍丹. 我国图书情报领域科学数据引用行为调查与分析 [J]. 情报科学, 2020, 38 (2): 11-16.
- 5 黄如花,李楠. 国外科学数据引用规范调查分析与启示 [J]. 图书馆学研究, 2016 (10): 2-9.
- 6 张静蓓,田野,吕俊生. 科学数据引用规范研究进展 [J]. 图书与情报, 2014 (5): 100-104.
- 7 丁楠,杨柳,丁莹,等. 我国社会学期刊论文数据引用行为研究 [J]. 图书与情报, 2014 (6): 88-93.
- 8 屈亚杰,王亚男. 社会科学领域科学数据的引用现状与特点分析 [J]. 数字图书馆论坛, 2017 (6): 25-31.
- 9 邱均平,肖博轩,徐中阳,等. 国内外图书情报领域数据引用特征的多维度分析 [J]. 情报理论与实践, 2022, 45 (9): 44-50, 36.
- 10 FENNER M, CROSAS M, GRETHE J S, et al. A data citation roadmap for scholarly data repositories [J]. Scientific data, 2019, 6 (1): 28.
- 11 COUSIJIN H, KENALL A, GANLEY E, et al. A data citation roadmap for scientific publishers [J]. Scientific data, 2018, 5 (1): 1-11.

(下转第 43 页)

- (6): 404 – 405.
- 2 陈多, 李芬, 朱碧帆, 等. 基于大数据的智慧信息管理平台在社区健康管理中的应用进展 [J]. 中国卫生资源, 2021, 24 (6): 725 – 729.
- 3 彭慧珍, 郭志杰, 鲍勇. 社区卫生服务中心电子健康档案利用情况的调查研究 [J]. 中国全科医学, 2009, 12 (9): 752 – 754.
- 4 王丽, 王俊星, 李燕明, 等. 社区护士应用信息化手段实施健康教育的实践 [J]. 中国护理管理, 2015, 15 (5): 528 – 533.
- 5 韩国元, 冷雪忠. 新医改十年: 中国医疗卫生信息化多元政策主体协同治理研究 [J]. 北京航空航天大学学报 (社会科学版), 2023, 37 (4): 1 – 13.
- 6 蒋文秀, 张冬梅, 张芮, 等. 整体性治理下我国县域医共体信息化建设现状分析 [J]. 中国医院管理, 2023, 43 (1): 57 – 60.
- 7 陈慧, 沈明辉, 李佳圆, 等. 基于政策工具的我国卫生信息化建设政策分析 [J]. 现代预防医学, 2020, 47 (17): 3151 – 3153.
- 8 张韦, 何东, 张研, 等. 政策工具视角下我国远程医疗国家层面政策分析 (1997—2019) [J]. 中国卫生政策研究, 2020, 13 (6): 56 – 64.
- 9 于勇, 喻明. 中国卫生信息化政策文本量化分析 [J]. 医学与社会, 2022, 35 (12): 114 – 119.
- 10 李志杰, 李伟, 卫平民, 等. 基于政策一致性指数模型的我国艾滋病防治政策量化评价 [J]. 中国艾滋病性病, 2023, 29 (6): 644 – 649.
- 11 翟运开, 郭柳妍, 赵栋祥, 等. 基于 PMC 指数模型的远程医疗政策评价 [J]. 信息资源管理学报, 2022, 12 (2): 112 – 122.
- 12 兰娅菲, 韩朦, 陈颖, 等. 国家中医药产业政策评价研究——基于 PMC 指数模型 [J]. 中国卫生事业管理, 2022, 39 (4): 280 – 286.
- 13 杨玉红. 基于 PMC 指数模型的知识产权政策评价研究 [D]. 大连: 大连理工大学, 2022.
- 14 杨赐然, 毛宗福, 崔丹. 新医改背景下公立医院改革政策评价 [J]. 中国卫生政策研究, 2022, 15 (4): 36 – 44.
- 15 张治国, 向晨. 基于 PMC 指数模型的我国医疗保障基金使用监管政策文本量化评价 [J]. 社会保障研究, 2023 (4): 1 – 14.
- 16 李祥飞, 张振, 于佳卉, 等. 基于 PMC 指数模型的我国医保支付方式改革政策量化与评价分析 [J]. 中国医疗管理科学, 2023, 13 (1): 33 – 39.
- 17 郭强, 刘冬梅. 中国农业农村科技服务政策量化评价 [J]. 中国科技论坛, 2020 (8): 148 – 158.
- 18 成全, 董佳, 陈雅兰. 创新型国家战略背景下的原始性创新政策评价 [J]. 科学学研究, 2021, 39 (12): 2281 – 2293.
- 19 王艳军, 郑建中, 张爱莲, 等. 我国区域卫生信息化建设政策分析 [J]. 中国数字医学, 2016, 11 (8): 47 – 49.
- 20 徐晓敏, 陈庆锬, 陈荃, 等. 基于 PEST 分析法基层卫生信息化的发展研究进展 [J]. 实用心脑血管病杂志, 2020, 28 (4): 14 – 18.

(上接第 31 页)

- 12 王文琪, 尤楠, 尹鑫, 等. 地学数据标注和引用现状与存在的问题初探 [J]. 遥感技术与应用, 2016, 31 (6): 1174 – 1180.
- 13 FORCE M M, ROBINSON N J. Encouraging data citation and discovery with the data citation index [J]. Journal of computer-aided molecular design, 2014, 28 (10): 1043 – 1048.
- 14 CHRISTENSEN G, DAFOE A, MIGUEL E, et al. A study of the impact of data sharing on article citations using journal policies as a natural experiment [J]. Plos one, 2019, 14 (12): e0225883.
- 15 PARK H, WOLFRAM D. An examination of research data sharing and reuse: implications for data citation practice [J]. Scientometrics, 2017, 111 (1): 443 – 461.
- 16 FLORIAN L, BIELZA C, HILL S L, et al. Data publications correlate with citation impact [J]. Frontiers in neuroscience, 2016, 10 (1): 419.
- 17 邱玉红, 焦红, 杨波. 多元数据出版模式下生物医学领域科研人员数据引用行为研究 [J]. 图书情报工作, 2022, 66 (16): 92.
- 18 Data Citation Synthesis Group. Joint declaration of data citation principles [EB/OL]. [2024 – 03 – 20]. <https://force11.org/info/joint-declaration-of-data-citation-principles-final/>.
- 19 原国家质量监督检验检疫总局, 国家标准化管理委员会. 信息技术 科学数据引用 [S/OL]. [2023 – 07 – 27]. <http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=A495CA355BAF00D962AA8DD84C3B2C16>.