# 生物医学领域科学数据汇交管理中的挑战与启示*

张敬晨 孙婧雯 罗 葳 张 月 赵远志 周 伟

（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** 分析生物医学领域科学数据汇交管理问题并提出应对措施，促进科学数据传播力提升。**方法/过程** 基于相关国家科学数据汇交政策，以国家人口健康科学数据中心为例，从科学数据管理方角度分析生物医学领域科学数据汇交中的挑战与对策。**结果/结论** 提出要普及科学数据汇交机制，制定标准并开展生物医学科学数据质量和规范化审核，遵照相关法律法规加强数据安全和隐私保护的技术研发。

〔关键词〕 科学数据；数据汇交；生物医学

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2024.10.005

**Challenges and Implications in the Management of Scientific Data Collection in the Biomedical Field**

*ZHANG Jingchen, SUN Jingwen, LUO Wei, ZHANG Yue, ZHAO Yuanzhi, ZHOU Wei*

*National Population Health Data Center, Chinese Academy of Medical Sciences, Beijing 100730, China*

〔**Abstract**〕 **Purpose/Significance** To analyze the management problems of scientific data collection in the biomedical field and put forward countermeasures in order to improve the dissemination of scientific data. **Method/Process** Based on the relevant national scientific data collection policy, taking the National Population Health Data Center as an example, the challenges and countermeasures of scientific data collection in the biomedical field are analyzed from the perspective of scientific data manager. **Result/Conclusion** The paper puts forward some countermeasures to solve the problems, including popularizing the scientific data collection mechanism, setting standards and conducting audits of biomedical science data quality and standardization, strengthening data security and privacy protection technology research and development in accordance with relevant laws and regulations.

〔**Keywords**〕 scientific data; data collection; biomedical

## 1 引言

《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》提出要坚持共享共用，充分实现数据要素价值，提高数据要素供给数量和质量，完善数据全流程合规与监管规则体系等。《科学数据管理办法》规定政府预算资金资助形成的科学数据应当按照开放为常态、不开放为例外的原则，由主管部门组织编制科学数据资源目录，有

〔修回日期〕 2024－04－30

〔作者简介〕 张敬晨，研究实习员，发表论文2篇；通信作者：周伟，高级工程师。

〔基金项目〕 中国工程院咨询项目（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）。

关目录和数据应及时接入国家数据共享交换平台，面向社会和相关部门开放共享，鼓励社会资金资助形成的其他科学数据向相关科学数据中心汇交，统筹推进国家科学数据中心建设和发展。现代生物医学是以数据命题和数据依赖为特征的研究领域，本文从生物医学科学数据管理方视角，依据常态化数据汇交流程，结合项目汇交情况，提出目前生物医学领域科学数据汇交中的挑战，以期为生物医学领域科学数据汇交管理提供参考。

## 2 生物医学科学数据汇交

### 2.1 科学数据汇交的意义

2.1.1 提高数据质量 生物医学科学数据通常分布于不同医疗机构和生物医学科研院所，通过数据汇交可以将不同来源的数据进行整合和规范，提高数据质量和准确性[1]。

2.1.2 促进数据共享 数据汇交能够将分散的数据整合到一起，在物理存储层面形成统一数据集合，方便用户查找和使用，从而促进数据共享和重用[2]。通过生物医学科学数据汇交可以将不同来源的数据进行统一管理和发布，提供统一访问接口和元数据描述，增强数据可访问性[3]。

2.1.3 推动科学研究 数据已经成为生物医学领域科学研究的“燃料”，生物医学科学数据汇交能够为科学研究提供全面、准确、可重复利用的数据支持，有助于推动科学研究进步和发展[4]。

### 2.2 科学数据汇交机制的建立

近年来，我国政府逐渐认识到科学数据汇交的重要性，出台一系列政策措施支持科学数据汇交工作，为科学数据汇交机制的建立和发展提供有力保障。《中华人民共和国科学技术进步法》明确提出，利用财政性资金设立的科学技术研究开发机构，应当建立有利于科学技术资源共享的机制，促进科学技术资源有效利用。《科技计划形成的科学数据汇交 技术与管理规范》（GB/T 39912—2021）规定科学数据汇交的原则、管理的主体与职责、主要内容及流程。科学数据中心是促进科学数据开放共享的重要载体，要承担相关领域科学数据的整合汇交工作。国家人口健康科学数据中心是科学技术部和财政部认定的 20 个国家科学数据中心之一，负责生物医学领域科学数据汇交工作[5]。本文以国家人口健康科学数据中心为例，讨论生物医学领域科学数据汇交全流程管理及其面临的挑战。

### 2.3 生物医学领域科学数据汇交流程

2.3.1 生物医学领域汇交计划审核 科学数据提交方根据项目研究任务编制项目数据汇交计划，梳理拟汇交科学数据详细清单，提交后由科学数据管理方确认。由于生物医学领域数据的特殊性，科学数据管理方对生物医学领域数据汇交计划的审核要点：一是项目任务书与汇交计划的数据一致性，数据提交方在项目开展中产生的数据应全部体现在汇交计划中；二是生物医学项目开展过程中产生的数据，在汇交计划中应按不同的研究对象、研究方法或研究目的等划分数据集，生物医学数据集因数据的复杂性可多维度划分，以保证数据集同质性；三是生物医学领域数据格式审核尤为重要，该领域科学数据格式繁多，在审核汇交计划时，为了保证数据的可重用性，要同时考虑原始数据和衍生数据，并以常见、通用的格式汇交。

2.3.2 生物医学领域科学数据审核 数据提交方对按汇交计划和项目实际开展情况所产生的科学数据自查后提交到科学数据管理方，科学数据管理方对元数据和规范化、标准化后的实体数据进行审核。生物医学领域科学数据审核要点如下。一是生物医学领域科学数据采集、科研项目开展由于其研究对象的特殊性和广泛性，对伦理审查和人类遗传资源等方面的要求更加严格。因此，要先审核涉及相关内容的科研项目批件，保证汇交科学数据的合理、合法、合规。二是生物医学科学数据格式多样，要审核实体数据的格式和例数，提高数据完整性和可重用性。三是生物医学领域科学数据具有隐私性，科学数据提交方应最大程度提高数据共享度，同时保证数据隐私安全。数据管理方也应关注敏感保密数据。

## 3　生物医学科学数据汇交的挑战

### 3.1　生物医学科学数据汇交机制普及度低

由于国内科学数据汇交工作开展时间不长，生物医学领域科研项目承担单位，如医院、科研院所和创新企业等对科学数据汇交要求并不完全了解。项目承担单位在接收到开展科学数据汇交工作的通知后，无法及时启动相关工作，影响整体汇交进度。总结国家人口健康科学数据中心开展科学数据汇交工作中遇到的问题，生物医学科学数据汇交普及度较低，具体体现在以下 3 方面。一是科研人员尚未充分认识数据汇交的价值。二是当前科研评价体系更侧重于成果发表而不是数据共享，科研人员缺乏了解科学数据汇交的“窗口”和参与科学数据汇交的动力。三是部分科研人员和机构对数据汇交政策的内容、目的和操作流程理解不够，忽视部分要求；科研人员缺乏技术支持和培训，不能按照规定格式和标准进行数据汇交。

### 3.2　生物医学科学数据量庞大且复杂，审核难度高

生物医学领域涉及学科类型多样、数据规模大[6]，数据审核和存储面临巨大挑战[7]。截至 2023 年 11 月 13 日，国家人口健康科学数据中心已发布的科学数据汇交项目数据集总量为 18 237 个，涉及生物学、预防医学与公共卫生学等学科大类。生物医学科学数据来源广泛，国家人口健康科学数据中心已汇交完成并发布的科学数据主要来源，见图 1。生物医学数据的复杂性直接影响数据审核，要求审核人员掌握多领域专业知识和技能。同时，生物医学领域数据标准和规范不统一，国内各医疗机构的信息存储系统结构不一致，数据异质性强，要针对每个项目的数据甄选专业软件和审核方式，数据审核困难度加大，对审核人员的专业性提出了更高要求，难以严格控制汇交数据质量，而数据质量差会增加数据共享的难度[8]。以生物组学数据为例，项目采集数据为双端测序数据，但当提交数据仅为衍生数据格式或上交数据文件每样本仅为一例时，会破坏数据完整性。

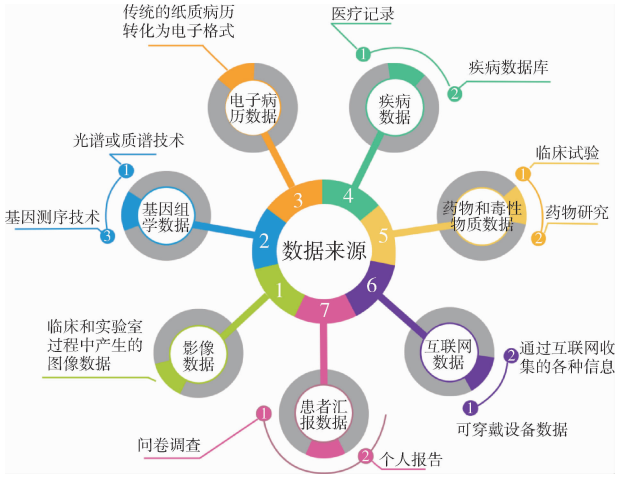


图 1　生物医学数据来源

### 3.3　数据隐私性强，泄漏风险高

生物医学数据涉及个人隐私和患者信息，须要采取严格的数据安全和隐私保护措施。国家人口健康科学数据中心已汇交完成的各类数据可能包含的隐私内容，见表 1。一方面，生物医学数据泄漏可能会对个人和社会造成不良影响[9]。另一方面，部分生物医学科学数据是对人类遗传资源进行分析和利用的结果，可为疾病预防、诊断和治疗提供重要的科学依据[10]；生物医学领域科学研究多涉及隐私和伦理，要在确保医学研究和医疗实践道德性和合法性的前提下开展科学研究[11]。因此，如何精确判断生物医学领域数据的合规、合法性，保证汇交过程中数据隐私安全成为汇交的难点。

表 1　生物医学领域科学数据隐私内容（部分）

| 数据类型 | 数据隐私内容 | 影响层面 |
|---|---|---|
| 影像数据 | 个人信息、就诊信息 | 个人 |
| 电子病历数据 | 个人信息、就诊信息 | 个人 |
| 基因组学数据 | 遗传信息、家系遗传信息、特定群体遗传信息 | 个人、群体 |
| 患者汇报数据 | 个人信息、就诊信息 | 个人 |
| 疾病数据 | 个人信息、就诊信息 | 个人 |
| 药物和毒性物质数据 | 个人信息 | 个人 |
| 互联网数据 | 个人信息、群体信息 | 个人、群体 |

## 4 应对生物医学科学数据汇交挑战的对策

### 4.1 加强数据汇交机制普及

《科学数据管理办法》的出台是国家层面构建科学数据汇交机制的重大举措。为解决生物医学科学数据汇交普及度低的问题，应积极宣传相关信息和政策要求，如科学数据汇交的必要性、生物医学领域科学数据规范化汇交要点、生物医学领域科学数据伦理等。具体可采取以下措施。一是通过各种渠道，包括学术会议、新闻稿和在线平台，向科研人员进行数据汇交政策教育和宣传，提高其对数据共享重要性的认识。二是依据国家政策，将科学数据汇交作为科技计划项目管理的重要环节，建立先汇交数据再验收项目的机制，将科学数据工作情况作为考核内容，鼓励科研人员了解并积极参与科学数据汇交。三是负责科学数据汇交的科学数据中心，如国家人口健康科学数据中心等，应制定相关领域清晰的数据汇交流程，包括数据汇交计划制定，数据实体的制备、提交、审核等，明确生物医学领域数据汇交审核要点；科研机构、数据管理机构、政府相关部门之间应通力合作，推动生物医学科学数据汇交政策的有效实施和普及。

### 4.2 完善治理标准

生物医学数据具有复杂性和多样性，数据质量问题可能会影响研究结果的准确性和可靠性，甚至导致错误的结论和决策[12]。数据治理是对数据资产的管理，如数据质量管理、元数据管理、数据安全管理等，以保证数据的可用性。为解决生物医学领域科学数据审核难度高的问题，生物医学领域科学数据汇交应建立完善的数据质量审核管理体系。通过数据质量治理流程确保数据的准确性、完整性和一致性，通过定期的数据质量评估、标化和清洗提高汇交数据的可靠性和有效性。不同来源和类型的数据规范性低、整合难度较大，应制定相应标准和规范，开发相应工具和技术来降低数据审核复杂性，并支持数据的规范化和整合[13]。首先，应充分利用科学数据管理标准，开展通用科学数据汇交工作管理[14]。科学数据管理常用标准，见表 2。其次，应细化并制定专业数据标准，使数据审核有标准可依。目前国内已制定实施部分生物医学专业科学数据标准，见表 3。应借助相关平台和工具推动生物医学领域数据标准应用[15]，如使用数据目录、元数据管理工具和数据质量管理软件，更好地管理和控制数据汇交。

**表 2 科学数据管理常用标准**

| 序号 | 标准名称 | 标准号 | 发布时间（年） |
|---|---|---|---|
| 1 | 《数据论文出版元数据》 | GB/T 42813—2023 | 2023 |
| 2 | 《人类生物样本中医信息基本数据集》 | GB/T 42465—2023 | 2023 |
| 3 | 《科技资源核心元数据》 | GB/T 30523—2023 | 2023 |
| 4 | 《科技计划形成的科学数据汇交 通用代码集》 | GB/T 39908—2021 | 2021 |
| 5 | 《科技计划形成的科学数据汇交 通用数据元》 | GB/T 39909—2021 | 2021 |
| 6 | 《科技计划形成的科学数据汇交 技术与管理规范》 | GB/T 39912—2021 | 2021 |
| 7 | 《科技平台 用户元数据》 | GB/T 39913—2021 | 2021 |
| 8 | 《国家卫生与人口信息概念数据模型》 | WS/T 672—2020 | 2020 |
| 9 | 《信息技术 科学数据引用》 | GB/T 35294—2017 | 2017 |
| 10 | 《科技平台 元数据汇交业务流程》 | GB/T 32845—2016 | 2016 |
| 11 | 《科技平台 元数据汇交报文格式的设计规则》 | GB/T 32846—2016 | 2016 |
| 12 | 《科技平台标准化工作指南》 | GB/Z 30525—2014 | 2014 |
| 13 | 《科技平台 通用术语》 | GB/T 31075—2014 | 2014 |
| 14 | 《科技平台 元数据标准化基本原则与方法》 | GB/T 30522—2014 | 2014 |
| 15 | 《科技平台 元数据注册与管理》 | GB/T 30524—2014 | 2014 |

表 3 生物医学领域科学数据标准（部分）

| 序号 | 标准名称 | 标准号 | 发布时间（年） |
|---|---|---|---|
| 1 | 《药学数据集 药理学》 | T/BIA 11—2023 | 2023 |
| 2 | 《药学数据集 药物毒理学》 | T/BIA 13—2023 | 2023 |
| 3 | 《药学数据集 化学》 | T/BIA 15—2023 | 2023 |
| 4 | 《骨科疾病诊疗数据集 严重肢体损伤》 | T/BIA 10—2022 | 2022 |
| 5 | 《重症医学数据集 患者数据》 | DB11/T 1866—2021 | 2021 |
| 6 | 《慢性阻塞性肺疾病临床研究通用标准数据》 | T/GDPHA 026—2021 | 2021 |
| 7 | 《肺结核监测基本数据集》 | T/CHATA 011—2021 | 2021 |
| 8 | 《心血管疾病研究通用标准数据集》 | T/GDPHA 030—2021 | 2021 |
| 9 | 《脑血管疾病研究通用标准数据集》 | T/GDPHA 031—2021 | 2021 |
| 10 | 《血液管理信息基本数据集》 | DB11/T 486—2021 | 2021 |
| 11 | 《糖尿病临床研究通用标准数据集》 | T/GDPHA 029—2021 | 2014 |
| 12 | 《电子病历基本数据集》 | WS 445. 1—2014 | 2014 |

最后，积极利用人工智能技术辅助完成数据汇交管理，大幅度提高数据汇交管理效率。可依据标准自动为数据集生成元数据，如标签、描述和关键词，有助于提高数据的可发现性和可搜索性，便于用户快速找到所需数据，同时降低元数据审核难度。在数据质量检测方面，人工智能可辅助识别和纠正数据中的错误、不一致和缺失，提高数据质量，确保数据的准确性和完整性。

### 4.3 保护隐私与安全

积极开展数据安全治理，确保数据安全与合规性。在数据汇交过程中，可以通过加强数据安全措施，如加密、访问控制和审计日志，以保护数据不被未授权访问或泄漏，确保数据汇交遵守相关法律法规和行业特定的合规要求。为保证生物医学领域数据的合法合规性，科学数据管理方相关人员需要持续学习生物医学伦理知识，实时关注相关法律法规，并依法依规开展数据汇交管理，保证审核专业性，见表4。

表 4 生物医学数据隐私与数据安全部分相关法律法规

| 序号 | 相关法律法规 | 发布时间（年） |
|---|---|---|
| 1 | 《涉及人的生命科学和医学研究伦理审查办法》 | 2023 |
| 2 | 《中华人民共和国个人信息保护法》 | 2021 |
| 3 | 《中华人民共和国医师法》 | 2021 |
| 4 | 《中华人民共和国数据安全法》 | 2021 |
| 5 | 《中华人民共和国刑法》 | 2020 |
| 6 | 《中华人民共和国生物安全法》 | 2020 |
| 7 | 《中华人民共和国民法典》 | 2020 |
| 8 | 《中华人民共和国人类遗传资源管理条例》 | 2019 |
| 9 | 《中华人民共和国基本医疗卫生与健康促进法》 | 2019 |
| 10 | 《中华人民共和国精神卫生法》 | 2012 |

为解决生物医学数据隐私和安全问题，要依据法律法规加强数据隐私和数据安全技术研发，包括数据加密、访问控制、审计监控等技术[15-17]。如使用人工智能技术进行自动化文档审核，检查数据

汇交过程中的文档是否符合预设的合规性标准；又如人工智能识别敏感信息、不合规用语或潜在的版权问题，保护数据汇交过程中的隐私安全。

## 5 结语

数据汇交是实现科学数据共享的关键步骤，对于提高科学研究的效率和质量具有重要意义。为解决生物医学领域数据汇交过程中汇交机制普及度低、审核困难、数据隐私性高等问题，本文提出相应对策，以期为保障生物医学领域科学数据汇交的安全顺利开展、提高数据质量提供参考。未来随着数据汇交和共享政策的逐步推行，更多创新方法和技术将被应用于数据汇交过程，科学数据汇交范围的进一步扩大会加快国家生物医学领域科学数据资源建设的速度，应尽快梳理国家数据战略资源，提高数据质量，促进数据共享，推动科学研究的进步和发展。

**作者贡献**：张敬晨负责提出研究思路、数据收集与图表制作、论文撰写与修订；孙婧雯负责数据收集、论文审核与修订；罗葳负责图表制作、论文审核与修订、研究监督；张月、赵远志负责图表制作、论文审核与修订；周伟负责提出研究思路、研究监督。

**利益声明**：所有作者均声明不存在利益冲突。

## 参考文献

1 MÜLLER A, CHRISTMANN L S, KOHLER S, et al. Machine learning for medical data integration [J]. Studies in health technology and informatics, 2023, 302 (1): 691 – 695.

2 王伟英. 不动产登记数据整合和档案整理的基础——不动产登记档案关联关系梳理 [J]. 档案管理, 2018 (3): 50 – 51.

3 GIEREND K, FREIESLEBEN S, KADIOGLU D, et al. The status of data management practices across german medical data integration centers: mixed methods study [J]. Journal of medical internet research, 2023, 25 (1): e48809.

4 唐源, 吴丹. 国外医学科学数据共享政策调查及对我国的启示 [J]. 图书情报工作, 2015, 59 (18): 6 – 13.

5 PHDA. National population health data center population health data archive [J]. Chinese medical sciences journal, 2022, 37 (1): 102.

6 王雪艳, 井艳玲, 赵爱芳. 基础医学科学数据汇交管理及数据特征 [J]. 基础医学与临床, 2022, 42 (5): 852 – 856.

7 文昱琦. 面向多维生物医药数据整合挖掘的人工智能关键算法与应用研究 [D]. 北京: 中国人民解放军军事科学院, 2023.

8 PACKER M. Data sharing in medical research [J]. British medical joumal, 2018, 360 (8141): 272.

9 WILLIAMS C M, CHATURVEDI R, CHAKRAVARTHY K. Cybersecurity risks in a pandemic [J]. Journal of medical internet research, 2020, 22 (9): e23692.

10 FEDERER L M, LU Y L, JOUBERT D J, et al. Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff [J]. Plos one, 2015, 10 (6): e0129506.

11 于丽娟, 冯诚. 总结成果继往开来——《中国医学伦理学》特别专题《医学科学数据共享与使用的伦理要求和管理规范》应用和进展研讨会纪实 [J]. 中国医学伦理学, 2023, 36 (1): 112 – 114.

12 TOGA A W, DINOV I D. Sharing big biomedical data [J]. Journal of big data, 2015, 2 (1): 1 – 12.

13 GREENBERG J, WHITE H C, CARRIER S, et al. A metadata best practice for a scientific data repository [J]. Journal of library metadata, 2009, 9 (3/4): 194 – 212.

14 陶毅, 苏爽, 赵正宜, 等. 基于元数据的计量科学数据汇交系统研究 [J]. 中国科技资源导刊, 2022, 54 (2): 1 – 12, 92.

15 李雪凝, 刘丰源, 王凌, 等. 多源通用数据标准管理平台的设计和应用 [J]. 计算机应用与软件, 2018, 35 (5): 62 – 66.

16 JONES M, JOHNSON M, SHERVEY M, et al. Privacy – preserving methods for feature engineering using blockchain: review, evaluation, and proof of concept [J]. Journal of medical internet research, 2019, 21 (8): e13600.

17 WIRTH F N, MEURERS T, JOHNS M, et al. Privacy – preserving data sharing infrastructures for medical research: systematization and comparison [J]. BMC medical informatics and decision making, 2021, 21 (1): 1 – 13.