

基于多案例研究的生物医学科学数据开放共享策略分析^{*}

万佳林¹ 贾晓峰² 胡志民¹

(¹ 中国医学科学院/北京协和医学院卫生健康管理政策学院 北京 100730

² 国家卫生健康委卫生发展研究中心 北京 100044)

〔摘要〕 目的/意义 提出生物医学科学数据开放共享的策略性建议, 为科学数据可持续发展提供理论支撑。方法/过程 采用多案例研究方法, 分析英国生物样本库 (UK Biobank)、美国国家生物技术信息中心 (NCBI)、全球流感共享数据库 (GISAID)、Cortellis 药物研发情报平台 (CDDI) 4 个国际典型生物医学科学数据平台的开放共享实践。结果/结论 基于数据共享前、中、后 3 个环节提出建议, 数据资源建设方面应确定数据权利和权利主体的划分, 数据开放使用方面应依据数据属性实行差异化管理, 数据生态营造方面应制定促进数据增值的机制和措施。

〔关键词〕 科学数据; 生物医学数据; 数据共享; 案例研究

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.02.004

Analysis of Open Sharing Strategies of Biomedical Scientific Data Based on the Multiple-case Study

WAN Jialin¹, JIA Xiaofeng², HU Zhimin¹

¹ School of Health Policy and Management, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100730, China; ² China National Health Development Research Center, Beijing 100044, China

〔Abstract〕 **Purpose/Significance** To put forward strategies for the open sharing of biomedical scientific data and provide theoretical support for the sustainable development of scientific data in China. **Method/Process** The paper adopts the multiple-case study method to analyze open sharing practices of four international typical biomedical scientific data platforms, including UK Biobank, National Center for Biotechnology Information (NCBI), Global Initiative on Sharing All Influenza Data (GISAID) and Cortellis Drug Discovery Intelligence (CDDI). **Result/Conclusion** Suggestions are proposed based on three processes of before, during and after data sharing. In the construction of data resources, the division of data rights and rights subjects should be determined. The open use of data should be differentiated according to data attributes. In terms of data ecology construction, mechanisms and measures to promote data value-added should be formulated.

〔Keywords〕 scientific data; biomedical data; data sharing; case study

〔修回日期〕 2023-10-19

〔作者简介〕 万佳林, 硕士研究生, 发表论文 1 篇; 通信作者: 胡志民, 贾晓峰。

〔基金项目〕 中国医学科学院医学与健康科技创新工程项目 (项目编号: 2021-12M-1-057)。

1 引言

21 世纪以来，科学数据与数据科学相互促进发展，共同推动科学数据的建设与开放使用。2022 年 12 月《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》创新性地提出数据资源持有权、数据加工使用权、数据产品经营权等“三权分置”的中国特色数据产权制度^[1]，进一步推动数据要素化发展。2023 年 3 月《党和国家机构改革方案》提出组建国家数据局，负责协调推进数据要素基础制度建设，表明中国将加快数据资源整合共享和开发利用，统筹数字经济高质量发展^[2]。

生物学是科学数据每日产出量最多的领域之一，数据具有类型多样、数量巨大、应用价值高等特点。我国生物学科学数据建设仍处于政府牵头的“量的堆积”阶段，尽管规模已相当可观，但数据质量、数据使用以及数据牵引的学术活跃度仍不足，仅依靠财政资金的建设模式难以实现科学数据中心的可持续发展。借鉴国外典型数据平台的发展经验，建设我国高质量生物学科学数据中心是一种可操作的方法，为此，本研究分析国外典型生物学科学数据平台的开放共享管理与服务，总结其成功经验，提出我国数据开放共享的策略性建议。

2 研究设计

2.1 研究方法

本研究采用多案例嵌入式研究方法。案例数量方面，多案例研究提供比单案例研究更多的数据量，有利于挖掘普遍性规律，增加结果说服力^[3]。本研究中多案例提供了更多的研究素材，便于归纳不同类型科学数据平台的共性或个性经验。案例分析层次方面，不同于整体式案例研究以揭示整体属性为目的，嵌入式案例研究通过簇群技术抽取出具次级分析单位，通过考察主分析单位和次级分析单位开展研究^[3]。本研究围绕科学数据的开放共享，选取合适的维度（次级分析单位）拆解分析所选案例，使研究框架更清晰，研究目的更聚焦。

根据多案例研究的复制原则和研究目的，选取案例满足以下标准。一是生物学领域具有较大学术影响力、数据资源丰富、知名度较高的科学数据平台。二是平台收录不同类型的特色科学数据。三是在科学数据开放共享方面具有共性化或差异化的特点。基于此，选取 4 个科学数据平台作为案例分析对象：英国生物样本库（UK Biobank）、美国国家生物技术信息中心（National Center for Biotechnology Information, NCBI）、全球流感共享数据库（Global Initiative on Sharing All Influenza Data, GISAID）、Cortellis 药物研发情报平台（Cortellis Drug Discovery Intelligence, CDDI），见表 1。

表 1 所选案例基本情况

科学数据平台	主管机构	数据资源	典型特点
英国生物样本库	英国生物样本库有限公司	英国 50 万受试者的基线信息、生物样本、生活方式、健康记录等队列数据	世界上规模最大的前瞻性人类遗传队列样本库，自己产生和管理数据
美国国家生物技术信息中心	美国国立卫生研究院、美国国家医学图书馆	文献、基因、基因组、蛋白质、遗传学、化学物质共 6 类生物学数据	第三方管理的多来源数据中心，学界知名度高
全球流感共享数据库	GISAID 注册伙伴关系协会	流感病毒、呼吸道合胞病毒等病毒数据	学界牵头建立的国际性病毒数据库，为研发提供数据支撑
Cortellis 药物研发情报平台	科睿唯安信息服务有限公司	有生物活性的化学药和生物药、实验药理学数据、实验模型数据、专利等药物数据	商业化的运营模式，提供从药物研发到商业化的数据服务和技术支持

2.2 分析维度

基于文献研究，特别是以下两个标准，自行制定分析维度。一是董瑞玉等^[4]将数据共享机制分为资源整合机制、信息安全保障体系、信息访问机制、奖励激励及配套管理制度。二是孙小康^[5]分析国外数据共享政策时选取共享时间、共享范围、共享模式、共享流程、认可或引用5个维度。结合研究目的，基于数据共享前、中、后3个环节，选取资源整合、开放级别、共享方式、数据增值4个维度构建案例分析框架，见表2。

表2 案例分析分析维度

数据共享环节	分析维度	含义
数据共享前	资源整合	数据来源及数据整合机制
数据共享中	开放级别	基于数据安全需求面向不同人群的开放分级
	共享方式	数据共享具体方法、流程及相关要求
数据共享后	数据增值	数据价值的表现形式和保障机制

3 研究结果

3.1 资源整合

3.1.1 数据来源 数据资源整合发生在数据的产生、收集、加工过程中，据此可将数据来源分为内部产生数据、外部汇集数据、内部加工编辑数据和混合模式。UK Biobank 数据来源于世界上规模最大的人类遗传队列研究，数据由区域分中心收集处理后汇总到平台^[6]，属于内部产生数据的数据库。依托队列研究获得数据有助于数据来源的稳定性和可持续性，还能以平台的形式助力数据的下游研究和成果追踪。汇集外源数据的数据中心不产生原始数据，仅承担数据管理和维护职能，NCBI、GISAID 和 CDDI 均属于这类数据中心。相较 GISAID 由各研究机构和组织直接汇交，NCBI 的数据来源更多元^[7]，包括数据产生方直接汇交、与数据提供者和研究联盟的合作或协议、内部人工筛选和梳理。2020 年美国国立卫生研究院发布的《数据管理与共享政策》要求所有其资助研究产生的科学数据应汇

交到已建立完备、声誉良好的开放存储库^[8]，使其资助产生的项目数据成为 NCBI 数据来源之一。CD-DI 数据来源于临床试验、药物管线、政府文件、经销信息等渠道，同时吸纳了许多经典的药物数据库和补充信息^[9]，CDDI 没有专门的外部数据汇交，而是基于商业竞争性和客户需求，以资源整合和数据搜索为业务职能自主收集信息，如梳理行业新闻、整合政策规范、报道学术会议等^[10]。

NCBI 和 CDDI 存在不止一种数据来源：NCBI 的参考序列数据库（RefSeq）来源于对国际核苷酸序列数据库的二次筛选^[7]，是内部加工形成的精选数据库；CDDI 存在大量人工编辑、整合形成的数据情报，如 Cortellis 竞争情报数据库（CCI）由 500 多名拥有 4~7 年生化、医药学经验的专业人士编辑而成^[9]。

3.1.2 数据资源整合机制 指科学数据平台以何种方式保障数据来源稳定和可持续，4 个数据平台既有共性又有差异，见表3。

表3 数据资源整合机制

科学数据平台	数据资源整合机制
英国生物样本库	依托队列产生数据，数据产生者和管理者一体
美国国家生物技术信息中心	生物医学界普遍认可的权威性和学术影响力
全球流感共享数据库	学界认可度高，资金流多元，合作者遍布全球
Cortellis 药物研发情报平台	基于营利需求的商业化数据收集，资源丰富且合作者众多

3.2 开放级别

当前生物医学科学数据平台开放级别主要分为开放、限制、关闭3种。开放指用户可以无障碍访问数据平台；限制指外部用户满足一定条件后可访问，如注册、付费等方式；关闭指仅面向内部用户^[11]。3种开放级别并不完全独立，部分数据平台同时包含开放数据和限制数据，实行不同的管理方式。

根据开放级别，NCBI 平台收录的大部分数据库对使用者几乎无任何限制，属于完全开放式，仅

对涉及人类信息数据的受控访问数据库采取分级模型；无法进行个人识别的数据不采取任何限制措施；能进行个人识别的数据采用审核制并限时访问^[12]。UK Biobank、GISAID 和 CDDI 均属于限制级别，但开放方式和程度各异，见表 4。UK Biobank 仅面向科研人员开放，用户需要通过身份认证才能访问数据资源^[6]。GISAID 数据面向通过身份一次性认证、遵守《GISAID 数据库访问协议》条款的所有自然人，无论是否是科研工作者^[13]。CDDI 具有商业性质，面向支付注册费的用户，受众主要为医药企业、科研机构等大型群体对象。

表 4 数据平台开放级别及面向人群

科学数据平台	开放级别	主要面向人群
英国生物样本库	限制	全球科研人员
美国国家生物技术信息中心	开放、限制	全球用户（受控访问数据库除外）
全球流感共享数据库	限制	实名制且遵守数据访问协议的自然人
Cortellis 药物研发情报平台	限制	付费的注册用户

3.3 共享方式

3.3.1 共享方法 早期数据共享方法通常包括数据档案、数据飞地、数据管理人员传播、混合模式等^[5]。近年来开放存储库的建设和发展越来越普遍，期刊纷纷强调论文关联数据的提前共享，科研资助机构也陆续出台政策促进数据共享，将数据汇交到开放的公共存储库已成为数据共享的主流方式。UK Biobank、NCBI、GISAID 均属于科学数据公共存储库，生产/汇交的数据经平台审核、清洗或加工后，按发布期限要求对外开放共享。CDDI 是企业所有的私有数据平台。在生化药物类数据领域，企业出资建立非公共存储库也是常见的共享方法之一。

3.3.2 共享流程 数据共享流程主要指用户需要遵循怎样的流程才能获取数据，以及数据使用协议和要求。涉及人类数据的科学数据平台如 UK Biobank 和 NCBI 常采用“申请－审核”访问机制。UK Biobank 在用户注册并提供身份证明后，还需要

提交一份阐述研究项目的申请书，经数据访问分委会审核通过后才能获取其数据和生物样本^[6]。NCBI 绝大部分数据库可以直接访问，但其基因型和表型数据库（Database of Genotypes and Phenotypes, dbGaP）需要受控访问，采用“申请－审核”并“限定使用期限”保障数据安全^[12]。不涉及人类数据的 GISAID 通过自行制定的《GISAID 数据库访问协议》规范平台的数据共享和使用行为，并明确数据访问的流程和要求。该协议通过保证数据提供者的所有权并承认其贡献鼓励数据共享积极性，并要求数据提交方不附加任何限制^[14]，确定了“身份认证＋贡献认可”的共享机制，通过实名制为病毒数据的共享使用提供良好范式。其他数据平台也有相应的数据访问和使用协议，研究人员使用 UK Biobank 数据须与平台签订访问协议，详细说明使用数据目的等相关条款，并承诺不识别任何数据和样本提供者^[15]；dbGaP 要求数据申请者严格遵守《数据使用行为准则和协议》，包括不复制和保存数据、不主动识别个体参与者、主动通报数据泄漏事件等^[12]。CDDI 通过提供数据资源和服务获取利益，不以“无偿”“公益”的数据共享为目的，缴纳注册费后用户可享受平台提供的全研发周期的数据和技术工具，而不需要提交额外的数据使用申请。

3.4 数据增值

科学数据平台确立成果反馈机制追踪数据转化结果。UK Biobank 规定所有数据使用应出于公益目的，且研究者应将数据二次分析结果及时反馈以用于未来研究^[16]。此外，UK Biobank 还会给予受试者有限的反馈。该反馈机制确保 UK Biobank 数据资源愈加完整丰富，数据可利用性增强，确保价值链条完整可追踪。

数据平台可通过支撑科技创新和成果转化提升科学价值和经济价值。NCBI 除 dbGaP 外的 30 余个数据库完全开放共享，产生数据资源“虹吸效应”^[17]，进一步增加来源广度、提高更新速度、助推平台建设发展和数据的深度利用。GISAID 支撑数据高通量提交、突发公共卫生事件监测、疾病风险评估和流感界创新研发的下游分析等工作^[18]。

数据平台积极促进科学数据的共享使用,营造数据良好生态。GISAID 建立了一套独特的病毒数据共享机制以鼓励数据的快速共享,通过《GISAID 数据库访问协议》明确数据访问和使用规范,非匿名化提交、访问和使用数据有效保障科学家识别并承认数据提供者,为数据贡献者提供使用数据的额外保护,确保其固有权利(如知识产权)不会丧失。协议还要求数据使用者积极寻求与数据提供者的合作,增强数据提供者与使用者之间的信任,巩固对各自互补贡献的尊重^[19]。

4 策略性建议

4.1 共享前——数据资源建设:确定数据权利和权利主体的划分

科学数据的确权问题尚未达成共识,开放共享过程中数据权利主体的博弈阻碍数据开放共享。科学数据共享是将数据从持有者转移到使用者的过程,核心问题是数据所有权和使用权的分离^[20]。当前国际上的典型做法是依靠外部或内外混合型渠道获取资源,由数据平台专职行使建设权和/或管理权,而不享有使用权,表明数据持有者和建设者角色或可重叠,但从共享角度必须将数据持有/建设权与数据使用权分离,“自产自用”的模式本质上还是对数据资源的垄断。在数据资源建设阶段应以书面形式对数据确权分权,划分存在哪些数据权利并明确各权利主体,避免后续职责边界不清,造成数据管理效率低下,阻碍数据流通和使用。

4.2 共享中——数据开放使用:依据数据属性实行差异化管理

分类开放使用的模式可有目的地推动科学数据的精细化管理,依照“谁投入、谁贡献、谁受益”的数据要素收益分配原则^[1],可将公共和企业投资产生的数据分为公共数据、企业数据、涉及个体的个人数据,3 种数据属性分别为公益性数据、产业性数据和保密性数据。具体管理方式:一是公益性数据应在不涉及保密原则时尽早面向公众开放共享,体现其公益目的和普惠性;二是产业性数据依

法由投资企业持有、使用、获利,同时鼓励企业在达成一定营利目的后自愿公益性共享;三是保密性数据可分级管理,不可识别的数据可经匿名化后共享,易识别的数据可采用去识别化、受控访问或不对外等方式。

4.3 共享后——数据生态营造:制定促进数据增值的机制和措施

生物医学科学数据的价值来源于数据共享流通和开发利用,当前科学数据平台通过追踪成果转化、创建数据共享环境、支撑产品和技术研发、认可数据贡献等手段促进科学数据共享和数据增值,营造良好的数据共享使用生态。数据价值转化过程中存在较多机制体制问题,从政策角度有以下 4 个切入点:一是建立合适的数据引用机制,承认数据提供者在数据采集过程中付出的努力和贡献,保证数据提供者权利不受侵犯;二是制定数据共享使用的激励或补偿措施,如给予数据提供者优先发表相关成果的保护机制,促进数据二次利用,形成数据和科研相互支撑的良性循环;三是明确科学的数据利益分配方式,以政府为主导,同时积极探索市场化分配,兼顾个体效益和公众效益的平衡;四是出台科学数据共享的赏罚制度,尤其是科学数据不良使用行为的追责体系,监督科学数据的共享行为。

5 结语

本研究基于国际上 4 个典型的生物医学科学数据共享平台案例进行研究分析,总结其在开放共享方面普遍性和差异性的经验,并提出中国科学数据开放共享方面可供借鉴的策略性建议:确定数据权利和权利主体的划分,剥离数据所有权和使用权;依据数据属性和特征对科学数据进行个性化管理和利益分配;制定促进数据增值的机制和措施,规范数据引用行为,完善赏罚制度。未来生物医学科学数据的开放共享仍应平衡处理好数据保护与数据共享使用的关系,尽可能使科学数据开放共享效益最大化,营造数据开放共享的良好生态。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

1 新华网. 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 [EB/OL]. [2023 - 09 - 01]. http://www.news.cn/politics/zywj/2022-12/19/c_1129220019.htm.

2 新华网. 中共中央 国务院印发《党和国家机构改革方案》[EB/OL]. [2023 - 09 - 01]. http://www.news.cn/politics/zywj/2023-03/16/c_1129437368.htm.

3 YIN R K. 案例研究: 设计与方法 [M]. 第 5 版. 周海涛, 史少杰, 译. 重庆: 重庆大学出版社, 2017.

4 董瑞玉, 冯占英, 张晓梅, 等. 基于大数据应用的医学图书馆服务定位 [J]. 医学信息学杂志, 2017, 38 (1): 75 - 78.

5 孙小康. 精准医学大数据汇交与共享政策研究 [D]. 北京: 北京协和医学院, 2018.

6 UK Biobank. About us [EB/OL]. [2023 - 07 - 26]. <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us>.

7 NCBI. About NCBI [EB/OL]. [2023 - 08 - 01]. <https://www.ncbi.nlm.nih.gov/home/about>.

8 KOZLOV M. NIH issues a seismic mandate: share data publicly [J]. Nature, 2022, 602 (2): 558 - 559.

9 Clarivate. Cortellis drug discovery intelligence [EB/OL]. [2023 - 08 - 03]. <https://clarivate.com/products/biopharma/research-development/pre-clinical-intelligence-analytics/>.

10 汤森路透. 临床试验进展: 生物标志物助推临床开发决策 [J]. 药学进展, 2014, 38 (4): 8.

11 吴思竹, 李赞梅, 崔佳伟, 等. 基于全球研究数据注册仓储 Re3data.org 的医学科学数据仓储建设 [J]. 中华医学图书情报杂志, 2018, 27 (9): 20 - 31.

12 TRYKA K A, HAO L, STURCKE A, et al. NCBI's database of genotypes and phenotypes: dbGaP [J]. Nucleic acids research, 2014, 42 (D1): D975 - D979.

13 GISAID. About us [EB/OL]. [2023 - 07 - 20]. <https://gisaid.org/about-us/mission/>.

14 SHU Y, MCCAULEY J. GISAID: Global initiative on sharing all influenza data - from vision to reality [J]. Euro-surveillance, 2017, 22 (13): 30494.

15 王天一, 田国祥, 谢新雅, 等. 如何利用 UK Biobank 申请研究数据和生物样本 [J]. 中国循证心血管医学杂志, 2018, 10 (12): 1450 - 1453.

16 李新, 杜昕, 马长生. 从生物样本库到开放数据库 [J]. 转化医学杂志, 2014, 3 (6): 327 - 329, 367.

17 付磊, 李金斌, 王戎, 等. 国外人口健康领域科学数据共享平台建设特点及启示 [J]. 中国科技资源导刊, 2017, 49 (5): 89 - 94.

18 KHARE S, GURRY C, FREITAS L, et al. GISAID's role in pandemic response [J]. China CDC weekly 2021, 3 (49): 1049 - 1051.

19 ELBE S, BUCKLAND - MERRETT G. Data, disease and diplomacy: GISAID's innovative contribution to global health [J]. Global challenges, 2017, 1 (1): 33 - 46.

20 张旺, 程慧平. 科学数据开放共享策略机制及优化路径研究 [J]. 情报杂志, 2020, 39 (5): 154 - 161.

(上接第 19 页)

10 CHI Y, HE D, JENG W. Laypeople's source selection in online health information - seeking process [J]. Journal of the association for information science and technology, 2020, 71 (12): 1484 - 1499.

11 吴肃然, 李名荟. 扎根理论的历史与逻辑 [J]. 社会学研究, 2020, 35 (2): 75 - 98.

12 REIFEGERSTE D, BLECH S, DECHANT P. Understanding information seeking about the health of others: applying the comprehensive model of information seeking to proxy online health information seeking [J]. Journal of health communication, 2020, 25 (2): 126 - 135.

13 DOLNIČAR V, GROŠELJ D, HRAST M F, et al. The role of social support networks in proxy internet use from the intergenerational solidarity perspective [J]. Telematics and informatics, 2018, 35 (2): 305 - 317.

14 徐中阳, 楼海萍. 在线健康社区用户隐私披露行为影响因素研究 [J]. 医学信息学杂志, 2022, 43 (4): 40 - 45.

15 朱志鹏, 杨华凤, 金玉琴, 等. 移动社交媒体用户健康信息吸收规避模型构建 [J]. 医学信息学杂志, 2023, 44 (7): 30 - 37.

16 万文智, 宋小康, 赵宇翔, 等. 在线健康信息替代搜索行为的影响因素探究: 基于扎根理论的实证 [J]. 情报资料工作, 2020, 41 (6): 88 - 94.

17 王锦, 韩毅. 用户体验视角下心理健康科普短视频的信息效用及其影响因素研究——基于 B 站大学生用户的扎根探索 [J]. 情报科学, 2023, 41 (3): 74 - 80.

18 顾晨昱, 陈素白. 焦虑但难以逃离: 网络疑病症视角下的健康信息茧房研究 [J]. 现代情报, 2023, 43 (4): 51 - 63.