

· 专题论坛 | 综述 ·

临床研究数据管理与共享最新进展

■ 李皓琳¹, 姜勇^{1, 2, 3}

基金项目

“十三五”国家重点研发计划 (2016YFC0901001; 2017YFC1310901)

作者单位

¹100070 北京

国家神经系统疾病临床医学研究中心

²首都医科大学附属北京天坛医院神经病学中心

³北京大数据精准医疗高精尖创新中心 (北京航空航天大学&首都医科大学)

通信作者

姜勇
jiangyong@ncrcnd.org.cn

【摘要】 精准医学的发展需要规范的数据管理体系和高效的数据共享来推动。我国的临床数据管理标准在国际现行规范的基础上加强了集中管控和资质监管。21世纪初期,欧美国家通过建立集成型的卒中数据库来促进数据共享,我国也在2018年提出科学数据开放汇交的指导原则。国家神经系统疾病临床医学研究中心通过制定临床研究公共数据标准、整合专病数据资源目录,搭建了符合FAIR原则——可查找 (findable)、可访问 (accessible)、可互操作 (interoperable) 和可重用 (reusable) 的脑血管病大数据管理和共享平台,以提高临床研究数据质量和促进科学数据的共享。

【关键词】 临床试验; 数据管理; 数据共享

【DOI】 10.3969/j.issn.1673-5765.2020.06.005

Advances in Clinical Research Data Management and Sharing

LI Hao-Lin¹, JIANG Yong^{1,2,3}. ¹China National Clinical Research Center of Neurological Diseases, Beijing 100070, China; ²Department of Neurology, Beijing Tian Tan Hospital, Capital Medical University, Beijing 100070, China; ³Beijing Advanced Innovation Center for Big Data-Based Precision Medicine (Beihang University & Capital Medical University), Beijing 100191, China
Corresponding Author: JIANG Yong, E-mail: jiangyong@ncrcnd.org.cn

【Abstract】 The development of precision medicine requires a standardized data management system and efficient data sharing scheme. Based on current international norms, China's clinical data management standards emphasized centralized supervision on research qualification. In the early 21st century, western countries established integrated stroke databases to promote data sharing, and China in 2008 announced principles to encourage data sharing. China National Clinical Research Center of Neurological Diseases, by establishing clinical research public data standards and integrating specialized disease data resource catalogs, built a cerebrovascular disease big data management and sharing platform which conformed to FAIR principle--findable, accessible, interoperable and reusable, to improve clinical research data quality and promote scientific data sharing.

【Key Words】 Clinical research; Data management; Data sharing

随着人口老龄化带来的慢性疾病负担加重、医疗支出不断升高、医疗人员短缺等问题的日益凸显,2016年,科技部把精准医学研究上升到了国家战略高度^[1]。在2017年的第二届大数据科学与工程国际会议上,中国科学院陈润生院士又高度总结了精准医学的本质,即大数据和医疗的结合^[1]。医疗健康产业利用大数据,可以提高诊断准确度,提高疗效,降低费用和减少不必要的资源浪费,而精准医学和大数据的发展不仅离不开数据共享来推动,更需要高效

规范的数据管理体系来辅助。

临床医学与人类的健康息息相关,该领域的研究也日渐活跃,因此对于数据的完整性和准确性具有更高的要求,同时对数据共享和重新利用的需求也愈发迫切^[2]。临床试验数据是指在临床研究中产生的一系列事实的记录^[3]。临床试验数据的质量、真实性和完整性是评价试验药物有效性和安全性的基础,而标准化的数据格式更是有助于建立医疗信息互通性,实现无缝数据交换及共享^[3]。严格实施临床数据

本文开放科学 (资源服务)
(OSID) 标识码



管理流程,有助于切实提高科研数据的真实性,促进试验进展的高效性,而积极开展数据共享有助于减少无端重复研究所造成的资金浪费,促进临床科研的深度发展,为医疗健康大数据提供高质量的数据来源。

对比国际,我国在临床数据的管理与共享方面尚处于初级阶段。因此,希望通过对国际临床数据管理与共享发展历程的总结,重点列举美国、欧洲在数据共享方面的实践经验,为我国临床数据管理和共享的发展提供借鉴。

1 临床研究数据管理的发展历程

1.1 临床数据管理的国际标准 国际上对临床试验数据的管理有着较为完备的标准,对管理规程有着较为详细的指南和条例。1977年,美国食品药品监督管理局颁布《联邦管理法典》,第一次提出了临床试验质量管理规范的概念,并对高质量数据进行了定义^[2]。1996年,国际人用药品注册技术协调会(International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, ICH)结合美国、欧盟和日本的法规,借鉴北欧国家、澳大利亚、加拿大和世界卫生组织的条例,制定了临床试验管理规范^[4-6],对试验过程的记录、源数据、数据核查等都提出了原则性的规定^[5-6]。2016年,ICH在临床试验管理规范中添加了对临床试验数据的电子记录和电子签名的规定,并在计算机系统的验证、稽查轨迹,以及文件记录的复制等方面提出了明确的要求^[5-6]。

为弥补临床数据管理具体操作规范在国际统一性上的空白,2000年临床数据管理协会发布了临床研究数据管理规范,为临床数据采集、录入、保存,不良事件报告,数据质量控制,文件归档和人员培训等环节进行了详细的规定,并给予了相应操作的最低和最高标准^[7]。该文件作为现国际通用的临床数据管理操作规范,广泛应用于北美、欧洲及日本等国家^[7]。

1.2 中国临床数据管理政策的发展与现状 相比欧美国家,我国在临床数据管理经验、数据管理的标准操作规程和临床数据的执行监管方面亟须进一步加强和完善。我国现行的临床数据管理规范是由国家药品监督管理局(National Medical Products Administration, NMPA)于2003年9月修订发布的^[8]。2016年12月,NMPA又增加了对电子数据管理系统运行与安全性的规定、对数据质量的要求和对电子数据管理系统的权限管理和稽查轨迹的要求^[9]。中国现行的临床试验管理规范是参考国际通用的标准依据中国国情制定的。总体来说,中国的临床试验管理规范注重临床研究批准权和管理权的集中管控,很看重研究开展的资格、资质和时间段;而在ICH的临床试验管理规范中,开展研究的门槛较低,但是过程中的质量监督、安全性管控和最后的生产审批过程非常严格^[2, 5, 8]。

虽然我国的《药物临床试验质量管理规范》对临床试验数据管理提出了原则性要求,但仍缺乏执行过程中的具体操作法规和技术规定。为解决我国临床药物试验规范化程度不高,实验数据管理质量良莠不齐的问题,NMPA于2016年7月颁布实施《数据管理工作技术指南》《药物临床试验数据管理与统计分析的计划和报告指导原则》和《临床试验的电子数据采集技术指导原则》,从以下六个方面提出具体操作要求:①数据管理相关人员的职责、资质和培训;②管理系统的要求;③试验数据的标准化;④数据管理工作的主要内容;⑤数据质量的保障和评估;⑥安全性数据及严重不良事件^[9-11]。

1.3 临床数据管理政策的落实 为确保数据管理流程的责任化和规范化、确保研究数据的准确性和完整性,数据管理部门需要制订数据管理计划(data management plan, DMP)内容包括项目进展中以及在项目结题后的数据收集、数据创建、数据组织、数据处理、数据存储、数

据共享和数据复用的全过程^[12]。DMP在临床研究中的重要性在近几年凸显,如美国国立卫生研究院(National Institute of Health, NIH)、美国国家科学基金会(National Science Foundation, NSF)等机构都将DMP作为必需文件纳入项目提交审核要求中^[13-14]。虽然我国并未强制规定科研单位提交DMP,但随着国内临床研究数量及资助资金的飞速增长,对研究数据质量的高要求是必然的趋势,而DMP则会成为其中至关重要的环节^[15]。

2 临床研究数据的共享

近年来,随着医学研究数量的不断增加,临床研究数据也呈海量增长趋势,但研究效率低、数据质量缺陷和研究数据孤岛化等问题也日益显现^[15]。鉴于此,科研领域逐渐认识到研究数据合理共享的重要性。不仅可以推进研究进展、促进数据再利用,同时通过深度挖掘已有数据,产生新的科学问题,可最大限度地推动科技创新进程。

2.1 临床数据共享的国际经验 近几十年来,临床研究数据的共享,特别是在卒中诊疗和术后恢复等领域的数据交互发展迅速。美国NIH在《NIH数据共享政策和实施指南》一文中指出在维护参与者隐私、保护机密和专有数据的同时,应尽可能广泛和自由地进行数据共享^[13]。基于此,NIH要求自2003年10月1日起,凡提交基础研究、临床研究、临床调查和NIH支持的其他类型的研究,适用于涉及人体的研究和不涉及人体的实验室研究,且项目期间任何一年直接预算费用50万美元或以上的申请人,需要提交数据共享计划,或说明不能共享的原因^[13]。英国医学研究理事会(Medical Research Center, MRC)发布的数据共享政策中明确指出“公共资助的研究数据是公共产品,是出于公共利益而产生的,应最大限度地公开获取”^[16]。自2005年起,要求由MRC资助的研究所产生的数据必须与适当的高质量元数据一起发布;所有向MRC

提交资助建议的申请人,均须在申请书内附上一份数据管理计划^[16]。国际医学杂志编辑委员会(International Committee of Medical Journal Editors, ICMJE)要求,2018年7月1日后提交到ICMJE期刊的临床试验必须包含数据共享声明^[17]。2019年1月1日后入组患者的临床试验在试验注册时必须包含一个数据共享计划,主要内容包括:脱敏的个人数据是否共享,具体是什么数据,其他相关的文档有哪些;数据共享的对象、访问方式和分析内容^[17]。

当前,国际上已经建立了很多国家级、地域性的卒中数据库,比如美国国立神经系统和卒中研究院(National Institute of Neurological Disorders and Stroke, NINDS)的Archived Clinical Research Datasets,共享了所有NINDS资助的临床研究数据^[18]。美国All of Us队列也是高质量的数据来源,该数据集涵盖了超过100万18岁及以上的美国居民,通过采集个人信息、电子病历、生物样本和可穿戴设备信息,研究健康与基因、临床治疗、生活方式及社会环境之间的关联^[19]。另外,欧洲卒中联盟(European Stroke Organization, ESO)也建立了Angels Award Database,集成并共享欧洲地区卒中患者的诊疗病例^[20]。英国的生物样本库UK Biobank同样是卒中数据的优质来源,该项目自2006年开展,截至2010年已收集了50万英国公民(40~69岁)的样本信息和近10万人的影像标本,对研究卒中、心脏病等慢性疾病发挥了重要作用^[21]。

除此之外,世界各地的研究人员也搭建了很多专项数据库用于国际卒中数据共享。例如,由美国南加州大学发起的ATLAS (Anatomical Tracings of Lesion After Stroke)数据库,存储并共享了来自世界各地的百余卒中脑部扫描图,以预测针对不同类型病患的最有效的康复疗法^[22]。世界各地的科学家可基于此进行荟萃分析,了解病灶如何影响恢复;此外,很多研究者也在摸索利用机器

学习和人工智能的方法,更加快捷地识别病灶,提高当前病灶分割方法的准确性,以帮助医师制订更适合患者的精准诊疗和康复计划^[22]。

另外,基于卒中风险和治疗结果与基因组学的关联性认知^[23],2007年国际卒中遗传学联盟(International Stroke Genetics Consortium, ISGC)成立,旨在通过整合、链接国家级生物信息存储库,例如美国的dbGAP(Database of Genotypes and Phenotypes)和欧洲基因表型库(European Genome-phenome Archive)等,完成确立影响卒中风险、预后和治疗反应的遗传因素^[24-26],寻找代表遗传风险的等位基因,开发卒中个性化治疗的方法,以及建立具有可行临床效果的全面卒中风险评估体系^[27]。

此外,为了提高不同地域、不同时期的临床试验的可比性和普遍适用性,同时为将来的随机临床试验设计提供依据和经验借鉴,虚拟国际卒中试验档案(Virtual International Stroke Trials Archive)应运而生^[28-29]。这份档案涵盖了过去16年间,全球范围内的21个试验,有超过15 000例缺血性和出血性卒中患者的数据,并收录了详细的数据收集背景和方法,旨在为使用者提供匿名分析和假设检验的平台,鼓励研究人员探索新颖的分析方式,并协助其规划和制订临床试验方案^[30]。

2.2 临床数据共享的实践 我国于2018年3月出台的《科学数据管理办法》从国家角度,提出了科学数据开放为主、多层级汇交的指导原则^[31]。《国务院办公厅关于印发科学数据管理办法的通知(国办发〔2018〕17号)》要求:政府预算资金资助的各级科技计划(专项、基金等)项目所形成的科学数据,应由项目牵头单位汇交到相关科学数据中心。各级科技计划(专项、基金等)管理部门应建立先汇交科学数据、再验收科技计划(专项、基金等)项目的机制;项目/课题验收后产生的科学数据也应进行汇交^[31]。

为提高开放数据的利用率,2014年由“联

合共建数据公平港口(Jointly Designing a Data Fairport)”的学术研讨会提出了数据管理的FAIR原则,即“可查找(findable)、可访问(accessible)、可互操作(interoperable)和可重用(reusable)”,该原则已在欧盟、美国、澳大利亚等国家普及(表1)^[32-33]。

紧跟国际趋势,2019年2月11日,中国科学院发布了《中国科学院科学数据管理与开放共享办法(试行)》,其中强调科学数据应按照分等级、符合FAIR原则的形式,适时向院外用户开放共享^[34]。文件中规定,数据管理方需要对科学数据进行必要的分级分类、并明确每个级别或类别的开放共享条件^[34]。此外,文件还对国际通用的FAIR原则做出了符合中国国情的详细解读和定义:

- 可发现:每个数据集都应被赋予符合国家标准,唯一且长期不变的标识符,并配有规范的元数据描述,易于发现和定位;
- 可访问:在保障数据安全的同时,开放共

表1 FAIR指导原则^[32]

可查找:
1. (元)数据需配备全局唯一且持久的标识符
2. 使用丰富的元数据描述数据
3. (元)数据清晰明确地包含其描述的数据标识符
4. (元)数据在可搜索资源中注册或索引
可访问:
1. (元)数据可通过其标识符、使用标准化的通信协议进行检索
1.1. 协议是开放、免费且可通用实施的
1.2. 协议允许在必要时进行身份验证和授权过程
2. 即使数据不再可用,也可以访问元数据
可互操作:
1. (元)数据使用正式的、可访问的、共享的和广泛适用的语言来标识
2. (元)数据使用遵循FAIR原则的词汇
3. (元)数据包括对其他(元)数据的合格引用
可重用:
1. (元)数据要以多种准确且相关的属性进行丰富的描述
1.1. (元)数据的发布要用清晰易读的数据使用许可证
1.2. (元)数据需与详细的出处相关联
1.3. (元)数据需符合与域相关的社区标准

注:FAIR原则,即可查找(findable)、可访问(accessible)、可互操作(interoperable)和可重用(reusable)。

享的数据应公开提供稳定且易于获取的访问地址和访问方式;

- 可重用: 开放共享的数据应明确数据使用的条件和要求, 具有对数据生成及处理过程、数据质量等的详细描述信息, 符合相关的标准规范, 以便可以在不同应用中重复使用或与其他数据融合后使用^[34]。

3 脑血管病临床研究数据管理与共享的实践

为解决脑血管病临床研究数据标准不统一, 现行标准应用率低的问题, 国家神经系统疾病临床医学研究中心建立了标准的脑血管病临床研究公共数据标准, 专病分级数据元和数据元协同管理平台, 旨在形成一套统一的公共数据元及分病种的核心数据元来提高临床研究数据质量和促进科学数据的共享。为应对数据库分散所造成的信息不畅和高质量数据源紧缺的问题, 中心通过汇交脑血管病防控领域大型社区队列、临床注册登记队列、多中心临床试验数据库和电子病历系统数据, 建立了脑血管病临床研究大数据。通过建立高性能计算平台, 利用基于虚拟化的桌面云技术将数据和计算资源整合在一起, 既保证了数据安全, 又提高了数据分析效率, 确保数据使用权和所有权有效分离, 为推进数据共享提供了一种高效的模式。

国家神经系统疾病临床医学研究中心建立的国家神经系统疾病临床研究大数据管理和共享平台, 既整合了国家临床医学大数据资源目录, 还包含符合FAIR原则的专病临床试验数据库, 为研究者提供一站式服务, 使其可以更加快捷、精准和高效地查找所需要的研究信息, 对已有数据资源开展深度学习, 为临床试验提供新的发展方向。未来通过与国际卒中研究机构的合作与数据共享, 研究者可以通过平台申请国外研究数据的使用权限, 或实时跟进国内外卒中领域研究的最新进展等。该平台的优化有利于汇交临床研究数据, 开拓数据共享模式, 提高临床研究的效率和质量, 促进大数

据技术应用和发展, 加速人工智能产品的研发和应用, 推动医学精准化、个体化。

4 结语

精准医学和大数据的发展不仅需要数据共享来推动, 更需要高效的数据管理体系来规范。医疗健康产业利用大数据可以提高临床诊断的准确性, 更加精准地预测治疗方案的成本与疗效。同时通过对人口健康数据的挖掘, 可以为预测疾病爆发和推动个性化医疗的发展提供助力。此外, 对临床数据、研究数据、个人及公共健康数据等多元数据的汇交、分析、挖掘与共享, 是推进临床决策、公共卫生和健康管理工作发展的重要基础。

参考文献

- [1] 第二届大数据科学与工程国际会议精彩荟萃[C/OL]. 信通传媒: 大数据. (2017-05-25) [2020-03-01]. <http://bdse2017-j-bigdataresearch.com.cn/default9.htm>.
- [2] 李文霞, 孔德宝, 毛宗福. 临床试验数据管理概述[J]. 中国新药与临床杂志, 2014, 33 (4): 256-259.
- [3] 谢高强, 姚晨. 数据管理在临床研究中的地位和作用[J]. 北京大学学报: 医学版, 2010, 42 (6): 641-643.
- [4] World Health Organization. Efficacy: good clinical practices (GCP) [EB/OL]. (1996-05-01) [2020-03-01]. http://www.who.int/medicines/areas/quality_safety/safety_efficacy/efficacy_en/index.html.
- [5] ICH指导委员会. 药品注册的国际技术要求(中英文对照)[M]. 周海钧, 主译. 北京: 人民卫生出版社, 2007: 1-4.
- [6] ICH Expert Working Group. ICH harmonized tripartite guideline: guideline for Good Clinical Practice E6 (R1) [EB/OL]. 2005 version. <http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html>.
- [7] Society for Clinical Data Management. Good clinical data management practices[M]. Milwaukee, WIs: Society for Clinical Data Management Inc, 2007: 26-40.
- [8] 国家食品药品监督管理局. 药物临床试验质量管理规范[EB/OL]. (2003-08-06) [2020-03-01]. <http://www.nhc.gov.cn/wjw/bmgz/201105/ebb64ab0d2f84fe08e143c1d0e56244c.shtml>.
- [9] 国家食品药品监督管理局. 临床试验的电子数据采集技术指导原则[EB/OL]. (2016-07-27) [2020-03-01]. <http://www.cde.org.cn/zdzy>.

- do?method=largePage&id=273.
- [10] 国家食品药品监督管理局. 数据管理工作技术指南[EB/OL]. (2016-07-27) [2020-03-01]. <http://www.cde.org.cn/zdyz.do?method=largePage&id=271>.
 - [11] 国家食品药品监督管理局. 药物临床试验数据管理与统计分析的计划和报告指导原则[EB/OL]. (2016-07-27) [2020-03-01]. <http://www.cde.org.cn/zdyz.do?method=largePage&id=270>.
 - [12] 王伟, 华夏, 王建梅. 国内外科学数据管理与共享研究[J]. 科技进步与对策, 2013, 30 (14) : 126-129.
 - [13] National Institute of Health. NIH data sharing policy and implementation guidance[EB/OL]. (2003-03-05) [2020-03-01]. https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.
 - [14] National Science Foundation. Dissemination and sharing of research results[EB/OL]. (2011-01-18) [2020-03-01]. <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>.
 - [15] 张崑, 应峻. 临床研究数据管理策略[J]. 复旦学报(医学版), 2017, 44 (1) : 122-126.
 - [16] Medical Research Council. Data sharing policy[EB/OL]. [2020-03-01]. <https://mrc.ukri.org/documents/pdf/mrc-data-sharing-policy/>.
 - [17] TAICHMAN D B, SAHNI P, PINBORG A, et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors[J]. JAMA, 2017, 317 (24) : 2491-2492.
 - [18] National Institute of Neurological Disorders and Stroke (NINDS). Archived Clinical Research Datasets[DB/OL]. [2020-03-01]. <https://www.ninds.nih.gov/Current-Research/Research-Funded-NINDS/Clinical-Research/Archived-Clinical-Research-Datasets>.
 - [19] National Institute of Health. All of Us research program operational protocol[EB/OL]. (2018-03-28) [2020-03-01]. https://allofus.nih.gov/sites/default/files/aou_operational_protocol_v1.7_mar_2018.pdf.
 - [20] European Stroke Organisation (ESO). Angels Awards Database[DB/OL]. [2020-03-01]. <https://eso-stroke.org/eso-stroke-quality-improvement-project/>.
 - [21] UK biobank data on 500, 000 people paves way to precision medicine[J]. Nature, 2018, 562 (7726) : 163-164.
 - [22] LIEW S L, ANGLIN J M, BANKS N W, et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations[J]. Sci Data, 2018, 5: 180011.
 - [23] MARIAN A J, BELMONT J. Strategic approaches to unraveling genetic causes of cardiovascular diseases[J]. Circ Res, 2011, 108 (10) : 1252-1269.
 - [24] DICHGANS M, MALIK R, KÖNIG I R, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants[J]. Stroke, 2014, 45 (1) : 24-36.
 - [25] YADAV S, COTLARCIUC I, MUNROE P B, et al. Genome-wide analysis of blood pressure variability and ischemic stroke[J]. Stroke, 2013, 44 (10) : 2703-2709.
 - [26] HOLLIDAY E G, MAGUIRE J M, EVANS T J, et al. Common variants are associated with large artery atherosclerotic stroke[J]. Nat Genet, 2012, 44 (10) : 1147-1151.
 - [27] FALCONE G J, MALIK R, DICHGANS M, et al. Current concepts and clinical applications of stroke genetics[J]. Lancet Neurol, 2014, 13 (4) : 405-418.
 - [28] BOGOUSLAVSKY J, KASTE M, OLSEN T S, et al. Risk factors and stroke prevention[J]. Cerebrovasc Dis, 2000, 10 (suppl 3) : 12-21.
 - [29] HATANO S. The WHO collaborative study on the control of stroke in the community[J]. WHO Bull, 2006, 54: 541-553.
 - [30] ALI M, BATH P M, CURRAM J, et al. The virtual international stroke trials archive[J]. Stroke, 2007, 38 (6) : 1905-1910.
 - [31] 张丽丽, 温亮明, 石蕾等. 国内外科学数据管理与开放共享的最新进展[J]. 中国科学院院刊, 2018, 33 (8) : 774-782.
 - [32] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Sci Data, 2016, 3: 160018.
 - [33] European Commission Directorate General for Research & Innovation. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020[EB/OL]. (2016-07-26) [2020-03-01]. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.
 - [34] 中国科学院. 中国科学院科学数据管理与开放共享办法(试行)[EB/OL]. (2019-02-11) [2020-03-01]. http://www.cas.cn/tz/201902/t20190220_4679797.shtml.

(收稿日期: 2020-03-15)



【点睛】 本文总结回顾了临床数据管理与共享的发展历程, 重点列举美国、欧洲的数据共享案例, 介绍了国家神经系统疾病临床医学研究中心在解决数据管理标准不一和数据共享不畅等方面的对策, 为我国临床数据管理和共享的发展提供经验性借鉴。