



[DOI] 10.3969/j.issn.1001-9057.2022.11.024

<http://www.lcnkz.com/CN/10.3969/j.issn.1001-9057.2022.11.024>

· 继续教育园地 ·

临床研究数据标准化工作的思考

周晓梅 李烁 崇雨田 李新华

[摘要] 随着医学大数据时代的到来,人们对临床研究数据的管理、共享与安全提出更高要求。本研究基于临床研究数据元、数据标准发展的进程和优势分析、临床研究相关数据与技术标准的探索和实践基础,对临床研究数据标准化工作进行总结分析,并对下一步研究提出合理化建议,在数据标准化层面平衡数据共享与隐私安全保护的矛盾,在保障安全的前提下,使临床研究数据价值最大化。

[关键词] 数据标准; 数据共享; 医学数据伦理; 数据安全

[中图分类号] R211

[文献标识码] A

临床研究数据使用的增长及跨实体、跨领域共享数据技术的提升被称为“近年来最大的医疗创新”,为人口健康和精准医学带来广阔前景^[1]。统一的数据行业标准是临床研究数据资产管理与共享的关键,是发展大数据人工智能的前提。因此我们需要建立统一的临床研究数据模型标准,在数据收集、治理、分析和共享的各个环节对数据形成统一的概念,消除歧义,并使之转化成统一、易于理解且便于使用的信息,以满足临床、科研、运营管理各领域对数据分析利用的需要^[2]。临床研究数据共享是指合法合规地共享临床研究的受试者数据、统计方法、统计结果及记录等相关信息,提高数据可及性,确保临床研究的透明化,保障科研诚信^[3]。临床研究数据标准是指通过标准化活动,按照规定的程序经协商达成一致,为数据收集、治理、分析和共享的各个环节提供规范或指南。临床研究数据标准是临床研究数据交换共享、重复利用的基础^[4]。

一、临床研究数据标准化进程

1. 卫生信息交换标准(HL7标准):HL7标准是国际医学信息标准组织发布的医疗卫生信息的数据交换标准^[5],其内容涵盖临床诊疗护理、医院管理、医疗质量、临床决策辅助支持、数据安全、电子病历、移动医疗保险付费等领域,广泛应用于医疗机构间、医疗机构与患者、医疗事故行政机构、保险机构及其他机构之间的健康医疗数据交换。

2. 医学数字成像和通信(DICOM)标准:DICOM标准是美国放射学会和美国电器制造商协会组织制定的医学图像存储和传输标准,主要解决医学影像在数字化医疗设备与医院信息系统之间的信息交换和存储的规范及标准问题,在中国乃至全球均有广泛应用^[6]。

3. 美国国立卫生研究院(NIH)通用数据元(CDE)标准:

基金项目:国家重点研发计划(2018YFC1315400)

作者单位:510630 广州,中山大学附属第三医院大数据人工智能中心(周晓梅、李烁),感染性疾病科(崇雨田、李新华)

通讯作者:李新华,E-mail:lixinh8@mail.sysu.edu.cn

NIH CDE 标准用于临床研究、患者登记和其他人类学科的研究,统一的标准可降低多中心临床研究数据和电子健康记录整合的成本,促进数据融合共享。NIH 所有临床通用数据元均根据国际标准化组织/国际电工委员会(ISO/IEC)11179 元数据标准的准则创建。

4. 《国际疾病分类法》国际疾病分类(ICD)编码:ICD 编码是 WHO 依据疾病特征对疾病进行编码的分类表^[7],是临床疾病诊断规范命名的主要参考依据,全面用于病案首页疾病诊断书写与编码、医疗信息化系统的疾病诊断分类和病案统计,及其他健康问题的记录,便于临床信息系统或流行病学监控。

5. 逻辑观测指标标识符命名与编码(LOINC)标准:LOINC^[8]提供了一套通用名称和标识码,用于标识实验室检查项目/临床观测指标的医嘱和结果。在 LOINC 数据库中,每条术语记录基于“六轴”概念表达式生成。“六轴”分别是成分/分析物名称、检查/检验属性类型、时间特征、体系/样本类型、标尺精度及检验方法类型。“六轴”的信息可确切地定义一条实验室检验项目或临床观测指标术语。

6. 临床数据交换标准协会(CDISC)标准:CDISC 标准已在临床研究尤其在临床试验中广泛应用^[9],是目前国际上公认比较成熟的标准体系。CDISC 标准涵盖临床研究方案设计、数据采集、分析、交换、递交等环节,为临床研究的全流程提供标准化的规范参考。

7. 医学系统临床术语(SNOMED CT)标准:SNOMED CT 标准是以概念为中心对疾病、临床发现、解剖结构、有机体、物质、药品、物理对象、物理力、标本等健康记录中的临床医学信息进行组织的标准数据集。SNOMED CT 标准目前已完成与 ICD-O-3、ICD-10、LOINC 等多种医学术语集交叉映射,并与 HL7、DICOM、可扩展标记语言(XML)开展了合作。

8. 优势分析:HL7、DICOM、NIH CDE、ICD、LOINC、CDISC、SNOMED CT 标准均是国际知名的医疗行业标准,每一项标准均有非常明确的编制目的和应用场景,其优势对比见表 1。

二、我国临床研究数据标准化工作遇到的问题

1. 国内医疗机构临床研究数据质量参差不齐:缺乏成熟的

表 1 HL7、DICOM、NIH CDE、ICD、LOINC、CDISC、SNOMED CT 标准的优势特色对比

标准	编制机构	初版年份	内容结构	开放程度	中文版本	主要应用
HL7 标准	HL7 委员会	1987 年	数据类型、数据字典、存储、传输标准等内容	免费	国标《健康信息学数据交换标准 HL7 临床文档结构(版本 2)》	医院数据信息传输,实现互联互通
DICOM 标准	美国放射学会和美国电器制造商协会	1985 年	数据结构、数据字典、存储、传输标准等内容	免费	国际 DICOM 标准中国委员会 DICOM3.0	医学数字图像及其信息在临床诊疗中的通信
NIH CDE 标准	NIH	2015 年	18 945 个通用数据元	免费	中国科学技术大学王学钦教授正在推进	研究人员识别和选择研究所需数据元
ICD 标准	WHO	1893 年	28 个章节,涵盖疾病分类、传统医学病症等	免费	中华医学会 ICD-11	疾病分类与代码和手术操作分类与代码等标准化
LOINC 标准	美国雷根斯基夫研究院	1994 年	94 000 多条术语,以“六轴”概念表达式呈现	免费	白求恩医院 LOINC-2020	检查/检验结果描述
CDISC 标准	临床数据交换标准协会	1997 年	定义、描述、关系、属性	免费	中国 CDISC 协调委员会正在推进	临床研究方案设计、数据采集、分析、交换、提交等全流程
SNOMED CT 标准	美国病理学会	1974 年	同义词、临床表现等 200 多种	会员制	神州医疗正在推进	临床术语表示与处理

临床研究数据行业标准是阻碍研究数据共享与医学人工智能发展的主要因素之一。2017 年,国务院在《新一代人工智能发展规划》指出,医疗人工智能的广泛应用将极大提高精准化服务水平,提升人民生活品质。在该规划的号召下,我国科研大数据平台、智能辅助诊疗工具如雨后春笋般迅猛发展。但不同层级的医疗机构由于人才、资金等不均衡发展,临床研究数据利用与科学研究水平的差距越来越大,数据质量参差不齐。部分医疗机构在做信息化规划时缺少对临床研究的考虑,未对关键研究数据进行采集与质控,导致研究数据价值较低。

2. 缺乏通用(公认)的临床研究数据标准:数据元是临床研究数据标准的基础。经世界各地研究者的不懈努力,目前代表性的临床研究数据元标准有 CDISC 标准、NIH CDE 标准等^[10],但国际上仍未建立一套成熟的通用临床研究数据元^[11]。制定数据标准主要存在以下难点:(1)不同疾病的病因复杂多变,存在很多难以量化的描述;(2)不同语言、不同区域的医学术语未统一,未能对术语给予统一的精确定义,造成交流合作的障碍;(3)病历书写格式与习惯的不统一。

3. 数据伦理与法律问题:数据隐私保护的伦理与法律法规体系已形成。临床研究数据涉及患者个人隐私,关系到患者的切身利益,每个环节都要保证数据隐私安全^[11]。国家相关部门相继出台《中华人民共和国人类遗传资源管理条例》、《民法典》、《中华人民共和国生物安全法》、《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》等法律法规,以应对数字技术高速发展大环境下数据安全的治理难题,为我国临床研究的安全发展保驾护航。

但对于大多数研究人员,法律法规条文专业性较强,学习门槛高。经过收集和处理的临床数据,究竟在储存、使用和后续共享中应遵循什么规则,常让研究人员在实际的临床研究工作中感到困惑^[12]。我国临床研究数据伦理原则与法律体系正在逐步完善,需要进一步普及,保证落实执行。

4. 数据管理制度问题:数据管理制度的合理性与贯彻落实可减少数据泄露事件的发生。根据美国 Verizon 公司发布的《2021 数据泄露调查报告》显示,医疗健康行业数据泄露 61%

由外部因素造成,39% 是内部因素(违规)。内部造成数据泄露的主要行为包括错误交付(36%,包括电子文件和纸质文件)、错误发布(20%)、配置错误(20%)及其他(24%)。内部恶意泄露行为的比例自 2019 年起逐年降低,但因内部人员对数据库与系统更熟悉,数据可得性更高,且数据规模逐年扩大,单次内部恶意泄露的危害更大,需要建立责权分明、责任到人的数据隐私和安全保护管理制度。

5. 临床研究数据安全技术问题:临床研究数据保护离不开强大的安全保护工具^[13]。国家卫生健康委员会在《2016 三级综合医院评审标准考评办法(完整版)》中首次明确规定了重要业务系统须达到等级保护(以下简称等保)三级标准才满足三级医院评审标准中对于网络安全的要求。重要业务系统一般指医院信息系统(HIS)、实验室信息管理系统(LIS)、影像归档和通信系统(PACS)及电子病历系统(EMR)四大系统,不包括网络办公系统、门户网站系统等。国家卫生健康委员会 2018 年发布的《国家健康医疗大数据标准、安全和服务管理办法(试行)》^[14]和《互联网医院管理办法(试行)》,分别规定承载健康医疗大数据的平台必须通过等保(未规定级别),三甲医院一般以等保三级为主,承载互联网医院的平台必须通过等保三级测评。现存的问题是,大部分三甲医院的非核心业务系统未通过等保三级测评,技术层面的数据安全保护措施不足,基层医疗机构的数据安全保护措施不完善^[15]。

三、临床研究数据标准化工作的反思与建议

国务院《关于构建更加完善的要素市场化配置体制机制的意见》已将数据列为第五大生产要素,由此可见数据潜在的巨大价值。以临床研究数据为驱动的医学人工智能是全球医疗行业的发展战略,临床研究数据标准则是临床研究数据资产管理与共享利用的核心。临床研究数据共享可最大程度地激发数据的潜力,其中最大的难题是如何保证数据隐私安全,既要共享,又要安全。在数据共享与隐私安全保护的博弈中找到平衡点,在保证安全的前提下,使临床研究数据价值最大化。针对此目标本文提出以下建议:

1. 继续推进临床研究数据行业标准建立:统一临床研究数据行业标准是数据区域共享的重要基础。没有统一的临床研究数据标准,就难以实现不同医疗机构、不同社区、不同地市等区域临床研究数据的高效共享。通过统一的行业标准对临床研究数据资产进行梳理,明确定义、格式、收集数量、收集范围等,能大大提升数据质量与数据的可靠度,降低数据治理难度。海量研究数据可通过统一标准汇聚,提升数据分析价值与利用率。标准制订部门可据此继续完善临床研究数据存储、传输、分析等技术标准,支持临床研究数据挖掘和分析运用平台的开发。临床研究数据标准与平台技术标准双管齐下,规范临床研究数据采集、存储、传输与分析等每个环节,为国家科研大数据平台建设奠定基础,缩小城乡医疗机构临床研究数据质量差距,整体提升临床研究的规模、质量与效率。

2. 加强临床研究数据隐私安全保护:临床研究数据隐私安全保护是数据共享的最大难题。临床研究数据的高价值也意味着其存在较高的泄露风险。临床研究数据的泄露分为外部(入侵、勒索软件等)攻击和内部(无意或恶意)泄露。

针对外部(入侵、勒索软件等)攻击问题,医疗机构需建立强大的入侵监测与预警系统。常规操作有系统与网络监控、流量监控、审计分析等,发现异常立即预警并切断连接。临床研究数据属于共享交换较为频繁的数据,因此需建立多道保护屏障。临床研究数据脱敏加密是非常有效的保护手段,高级加密标准和差分隐私方法可实现临床研究数据的加密脱敏^[16],其破解难度大、成本高,可有效保障数据安全;还可采用区块链技术,守护数据存储与传输过程,相当于给临床研究数据加上一道密码锁。区块链具有很好的不可篡改性、可追溯性、可共享性,我们可基于区块链的医疗隐私数据共享机制,采用身份验证、智能合约、共识机制等技术,实现多源性数据产权保护,从而推进细粒度可追溯敏感数据的权限管理与高效共享^[17]。

针对内部管理(主要是信息部门),医疗机构须建立完善的临床研究数据资产管理制度,降低职权滥用的风险。(1)制定权责分明的临床研究数据资产管理制度,明确数据查询、数据转换、数据分析、数据与分析结果导出等各个环节的标准流程与相关责任人,减少误操作。(2)完善数据库账号密码管理,专人负责,定期检查、更新。(3)数据库与系统监控,全部操作均需留痕,责任可追溯。(4)数据资产管理流程嵌入科研数据管理系统,保证操作规范且可审计。(5)定期进行数据隐私安全教育与数据安全相关法律法规学习,起到警示教育作用。(6)定期审计数据库、系统操作日志,及时发现并制止违法违规操作,对表现优秀者进行奖励,对违规人员进行处罚。

3. 继续完善数据隐私安全保护法律体系:临床研究数据隐私安全还需相关伦理与法律制度的保护,使数据全生命周期管理有规可循、有法可依。近年来,我国陆续颁布数据安全相关法律法规,建立对个人数据进行保护的法律体系框架,有效遏制不法分子的侵害。但还有一些问题尚需解决,如原始数据的归属、数据去标识化、数据应用与收益、合作成果的归属及安全保护分工等问题暂无明确规定,需规范文件继续补充解释。

执法机关还需做好数据隐私安全相关法律法规的普法工作。随着《中华人民共和国人类遗传资源管理条例》、《民法

典》、《中华人民共和国生物安全法》、《中华人民共和国个人信息保护法》和《中华人民共和国数据安全法》等法律法规的相继出台,我国数据隐私安全保护法律体系已成型,后续关键在于贯彻实施。很多研究者只知法规其名,但未能深入学习了解其内容与实施方法。因此,执法及相关监管部门应多举行普法活动,贯彻落实数据安全的各项法规,使人民知法守法。

4. 结合实践建立长期高效的数据标准反馈迭代机制:临床研究数据标准需求来自科研大数据平台的数据共享与挖掘需要。国家、区域医学大数据人工智能平台建设是临床研究数据标准化工作的推进剂,大数据平台建设、临床研究与数据标准制订工作需相互推进,相辅相成。利用现有临床研究数据标准建设大数据平台,挖掘与分析科研数据,再通过这一过程中获得的经验进一步完善优化临床研究数据标准,反复迭代,理论与实践同步进化,方能大力发展医学研究,提高全民健康水平。

四、结论

临床研究数据标准是临床研究数据资产管理的基础。本课题组临床研究数据标准化工作现已取得阶段性成果,部分病种已形成团体标准,临床研究数据标准的深度与广度需进一步挖掘,同时伦理与法律法规、数据处理与数据安全技术标准也要齐头并进,为区域间、国际临床研究数据交换提供支持,推进科研数据交流与利用。

参 考 文 献

- [1] 张振,周毅,杜守洪,等. 医疗大数据及其面临的机遇与挑战[J]. 医学信息学杂志, 2014, 35(6): 1-8.
- [2] 李娜,邵玉峰,苏菲,等. 终末期肝病模型系统对人工肝治疗重型肝炎短期疗效的预测[J]. 中华肝脏病杂志, 2011, 19(6): 469-470.
- [3] 张海洪. 临床研究数据共享:从道德责任到负责责任的实践[J]. 医学与哲学, 2020, 41(16): 20-23.
- [4] 刘丹红,张林,杨皓,等. 医学语言与临床数据标准化概述[J]. 中国卫生信息管理杂志, 2014, 11(1): 14-17.
- [5] 张庆,高上凯. HL7 标准及其在我国的应用前景[J]. 中国医疗器械信息, 2001, 7(4): 27-31.
- [6] Panykh OS. Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide [M]. Heidelberg: Springer, 2012.
- [7] O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD data accuracy[J]. Health Serv Res, 2005, 40(5 Pt 2): 1620-1639.
- [8] Forrey AW, McDonald CJ, DeMoor G, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results [J]. Clinical chemistry, 1996, 42(1): 81-90.
- [9] 陆芳,高蕊,唐旭东,等. 临床研究中的数据管理标准 CDISC 及其应用前景[J]. 中国新药杂志, 2011, 20(24): 2400-2404.
- [10] 任慧玲,李晓瑛,邓盼盼,等. 国际医学语体系进展及特色优势分析[J]. 中国科技术语, 2021, 23(3): 18.
- [11] 胡凯,刘丽华,徐勇勇,等. 医学数据资源共享中数据元著录的标准化方法研究[J]. 中国康复理论与实践, 2006, 12(4): 3.
- [12] 贺宁,胡汝为. 临床研究大数据流动与共享的法律规制研究[J]. 科技与法律, 2019(6): 49-58.
- [13] 刘逸敏,程传苗,邢茂迎. 隐私数据保护技术在医院信息管理领域应用探讨[J]. 中国数字医学, 2008, 3(7): 53-55.
- [14] 国家卫生健康委员会. 关于印发国家健康医疗大数据标准、安全和服务管理办法(试行)的通知[EB/OL]. (2018-07-12) [2020-10-27]. <http://www.nhc.gov.cn/guihuaxs/s10741/201809/758ec2f510c74683b9c4ab4fbb646557.shtml>
- [15] 孙瑜. 基于医院关键信息基础设施安全防护技术的研究与实践[J]. 网络安全安全, 2019, 10(4): 1-6.
- [16] Jeong E, Oh S, Kim H, et al. Communication-efficient on-device machine learning: federated distillation and augmentation under non-iid private data[J]. arXiv, 2018, 1811: 11479.
- [17] 高东平,王士泉,李伟,等. 基于区块链技术的医疗大数据平台架构研究[J]. 中国数字医学, 2019, 14(1): 29-32.

(收稿日期:2022-01-06)

(本文编辑:余晓曼)