基于标准化理论的区域临床研究数据平台构建*

黄雪群[1] 陈召霞[1] 渠田田[1] 沈恩璐[1] 缪怡然[2] 李晨曦[3] 马诗洋[1] 钱碧云[4] 俞章盛[1] 冯铁男[1,5]

([1] 上海交通大学医学院临床研究中心 上海 200025 [2] 北京华慬科技有限公司 北京 100193
[3] 华东政法大学刑事法学院 上海 201620 [4] 上海申康医院发展中心临促中心 上海 200041
[5] 核工业四一六医院/成都医学院第二附属医院 成都 610057)

〔摘要〕 **目的/意义** 解决区域性临床研究数据难以高效整合的问题，推进临床研究界中的“中国证据”和“中国方案”。**方法/过程** 借鉴标准化理论提出数据标准体系，以多中心临床研究数据整合为抓手，探索区域临床研究数据平台构建和应用方法。**结果/结论** 初步形成区域临床研究数据平台的理论体系，上海地区三甲医院临床研究能力提升显著。

〔关键词〕 临床研究数据；多中心临床研究数据整合；临床研究数据平台；数据标准化
〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.05.015

Construction of a Regional Clinical Research Data Integration Platform Based on Standardization Theory

HUANG Xuequn[1], CHEN Zhaoxia[1], QU Tiantian[1], SHEN Enlu[1], MIAO Yiran[2], LI Chenxi[3], MA Shiyang[1], QIAN Biyun[4], YU Zhangsheng[1], FENG Tienan[1]

[1]Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China; [2]Beijing Huajing Technology Co. Ltd., Beijing 100193, China; [3]School of Criminal Law, East China University of Political Science and Law, Shanghai 201620, China; [4]Shanghai Shenkang Hospital Development Center Clinical Research Promotion Center, Shanghai 200041, China; [5]Nuclear Industry 416 Hospital, Second Affiliated Hospital of Chengdu Medical College, Chengdu 610057, China

〔**Abstract**〕 **Purpose/Significance** To solve the problem that regional clinical research data are difficult to integrate efficiently, and to promote “Chinese evidence” and “Chinese protocol” in the global clinical research community. **Method/Process** Based on the standardization theory, the data standardization system is proposed, and the construction and application methods of the regional clinical research data platform are explored with the integration of multi-center clinical research data as the starting point. **Result/Conclusion** The theoretical framework of the regional clinical research data platform has been preliminarily established, and the clinical research capabilities of tertiary hospitals in Shanghai have been significantly improved.

〔**Keywords**〕 clinical research data; multi-center clinical research data integration; clinical research data platform; data standardization

〔修回日期〕 2023-10-23
〔作者简介〕 黄雪群，中级，发表论文 2 篇；通信作者：冯铁男。
〔基金项目〕 上海市卫生健康委员会卫生行业临床研究专项（项目编号：20234Y0285）；上海市“科技创新行动计划”启明星项目（扬帆专项）(项目编号:23YF1421000)；成都医学院医院发展研究中心课题（项目编号：YYFZ22005）。

1 引言

临床研究，特别是研究者发起的临床研究（investigator initiate trial，IIT），是国家临床诊疗发展的原始动力[1]。为推动 IIT 的科学发展，2021 年 6 月国务院办公厅印发《关于推动公立医院高质量发展的意见》[2]，2021 年 9 月《医疗卫生机构开展研究者发起的临床研究管理办法（试行）》[3]发布，各医院越来越重视临床研究[4]。研究数据的获取与采集是高质量 IIT 的瓶颈。目前医院内数据治理已取得一定成果，院内数据基本实现互通[5]。但院与院之间的数据不能互通，难以提供更加普遍的临床研究结果。因此本文从标准化理论出发，研究上海区域内临床研究数据平台的构建，结合实践报道进行理论总结，供相关研究和实践人员参考。

2 国外区域临床研究数据平台建设

目前比较主流的研究或临床研究数据库有英国生物银行（UK Biobank）、美国国家生物技术中心的基因银行（GenBank）、弗莱明翰心脏研究（Framingham Heart Study）、健康研究网络（TriNetX）、癌症基因组图谱（The Cancer Genome Atlas，TCGA）等，见表 1。UK Biobank 是一项前瞻性队列研究，收集了来自英国各地约 50 万名招募时 40～69 岁个体的深层遗传和表型数据[6]。GenBank 是最大、使用最广泛的基因数据库[7]。Framingham Heart Study 是一个长期、持续的心血管病学研究队列。TriNetX 是较成熟的能为医疗机构提供全方位临床数据的商业服务商。美国国立卫生研究院启动的 TCGA 试点项目，创建了一个全面的癌症基因组图谱“地图”[8]。这些数据库支持大量临床研究的开展，具有代表性。

表 1 国外主流区域临床研究数据库

数据库	数据来源	数据采集和管理	数据治理和安全	数据共享
UK Biobank	收集和存储约 50 万名英国参与者的临床、遗传、生物样本等数据	数据来自指定的采集中心，涵盖个人基本信息、试验检测、医疗历史、生活方式等内容；数据保持更新	严格遵守伦理原则和隐私保护措施，保护参与者的隐私和数据安全。数据经过匿名化和加密处理	可以访问委员会的申请程序，申请获得使用数据的权限
GenBank	GenBank 本身不采集数据，而是接收来自研究者的提交，并对这些数据进行组织、存储和公开	研究者被要求提供详细的实验和测序方法，以及相关的质量控制信息，确保数据来源可追溯，并提供质量控制信息。GenBank 的管理员审核和验证提交的数据	对提交的数据进行标准化和注释，确保数据的一致性和可理解性	数据分为公开和私密。公开数据直接通过搜索链接下载；私密数据需要通过申请获取
Framingham Heart Study	研究开始于 1948 年，最初招募了 Framingham 镇的 5 209 名成年居民作为初始队列。后来，研究逐渐扩大，招募了受试者的家庭成员和后代，形成多代家族队列	参与者接受定期（随访）的体格检查、生化指标检测、心电图、心脏超声等检查，以及面对面访谈。这些随访可以涵盖几年甚至几十年的时间跨度，还收集和保存了参与者的生物样本	研究人员整理、编码和存储采集到的数据，包括临床指标、实验结果、问卷调查数据、图像和生物样本相关信息等。数据经过质量控制和验证后，存储在数据库中，供研究人员分析和探索	通过专门的数据共享平台进行访问和获取。研究人员可以在该平台上注册并申请访问数据，通过安全的数据访问机制，确保数据的保密性和合法使用

续表 1

数据库	数据来源	数据采集和管理	数据治理和安全	数据共享
TriNetX	TriNetX 与各合作医疗机构建立数据连接，获取临床医疗数据和相关信息	TriNetX 进行数据质量检查，以识别潜在的数据问题和异常。包括检查数据的一致性、逻辑性和合理性，以及识别潜在的数据错误或异常值	对接收的数据进行清洗和标准化处理，确保数据的准确性和一致性。遵守严格的数据监管和合规性标准	研究人员可以与 TriNetX 建立合作关系，签订研究协议并获得数据访问权限。作为一个商业平台，其数据访问和共享通常需要与平台进行商业合作或达成协议
TCGA	只采集癌症患者的多维度数据，包括临床和多组学，以及影像学数据等。数据维度随技术发展更新	在数据生成阶段实施严格的质控。检查数据的一致性、完整性，注释数据，以便更好地理解和解释数据	对原始数据进行匿名化处理，匿名化后的数据只提供给研究人员。实施严格的访问控制措施，在数据传输和存储过程中采用加密技术	数据公开共享，可以通过 TCGA 官方网站（https://portal.gdc.cancer.gov/）或其他相关数据库访问和获取

3 中国区域临床研究数据平台构建现状

3.1 总体情况

目前我国医疗健康大数据较分散、数据规模不足、共享程度不高、区域整合平台范围局限，制约了医疗临床资源的最大化利用。与国外已有医疗大数据平台相比较，我国临床数据平台仍处于建设发展阶段，但平台整合与共享机制实践已有一些成绩，如上海市级医院临床信息共享平台（以下简称上海医联工程）是目前国内范围最大的联网医院临床信息共享系统，同时也是我国最大的医疗档案信息库；湖北省政府主导的医疗远程医学平台和中南大学主导的湘雅临床大数据系统建设项目都是大型的医疗信息共享平台构建研究成果[9]。此外，还有企业协同政府部门共同建设的区域性临床研究相关大数据平台，如中国心血管健康联盟倡导和神州医疗支持的全国心血管大数据平台，汇聚全国各地资源，专注于心血管领域，汇集大规模医疗数据；浪潮健康医疗大数据平台依托自主研发的平台整合医药与保险等数据；国家健康医疗大数据研究院依托山东大学等多方合作伙伴，构建健康医疗大数据产学研平台。

3.2 数据来源质量方面

临床研究数据包括诊疗业务数据和临床试验数据两大来源，二者在数据维度上存在显著差异，因此需要付出大量努力以实现二者的有效整合与利用。大量的诊疗数据分布在不同的业务系统中，但其数据具有人群代表性差、非结构化数据多等特点，尚不能直接用于临床研究的数据分析[10]。如上海医联工程和湖北医疗远程医学平台仅整合诊疗业务数据，数据维度不足，对临床研究支持有限。而医疗企业获得数据的主要方式是基于相关数据法律法规，利用大数据技术与当地政府和事业单位进行协作，获取与临床研究相关的居民健康数据，但同样存在数据来源质量和维度不足的问题。

3.3 数据采集和管理方面

除规范程度高的随机对照研究外，现有临床研究或健康相关数据几乎没有统一字段标准，且数据采集标准缺失严重。以诊疗服务为出发点采集的数据只能满足诊疗需求，且未形成规范的诊后随访管理，导致横向多维度数据采集和纵向时序随访数据永久性缺失。

3.4 数据治理和安全方面

大量面向疾病的语义分析模型构建后，可高效地将非结构化数据转为结构化数据[11]。目前从事数据治理的公司普遍存在缺乏临床研究专业知识、与医生或研究者沟通效率低下的问题，导致数据治理效果未能达到满足临床研究的标准。数据安全方面，我国信息技术部门数据安全意识增强，数据安全保障工作较好，但同样因为临床研究者和信息技术人员存在不同学科领域的认知差异，临床研究数

据应用受到一定限制。

3.5 数据共享模式方面

整合海量分散研究数据的目的即数据共享，通过区域管理部门的政策引导和推动、部门之间交换或企业带动的方式建立可行的数据共享模式。当前湖北医疗远程医学平台为国家政策驱动模式，但仅覆盖湖北省远程医疗服务体系。上海医联工程和湘雅临床大数据系统也仅支持本区域医院内部共享[12]。我国缺乏大型公共资源共享平台及国际合作项目等，亟待宏观性战略部署，鼓励和规范临床研究大数据的共享。

4 多中心临床研究数据驱动区域临床研究数据平台建设

4.1 统一多中心临床数据标准

开展多中心临床研究面临数据标准化问题[13]。可以基于病种确定数据源和制订数据采集标准，即确定疾病需要采集的数据维度，确定字段名标准。这样既有研究的同质化，也能够将多中心数据更好地融合[14]，研究初始阶段即可形成对数据来源、采集和存储标准的规范。

4.2 多中心数据整合与分析

临床研究数据平台不是仅支持一项临床研究，而是要支持一系列的临床研究。因此待整合的数据异质性很高，需要高效的数据治理工具规范数据，发现更深层次的疾病与表型指标的关系，以获得更全面和科学的结果，得到有效性更高的证据。

4.3 多中心数据安全与隐私保护

在整合多中心临床研究数据时，特别要注意数据的安全性和隐私保护。遵循国内外相关法律法规，采取相应的数据安全措施，依照规范的多中心共享机制，保障数据安全和患者隐私权。

5 标准化理论对区域数据整合平台建设的指导（图 1）

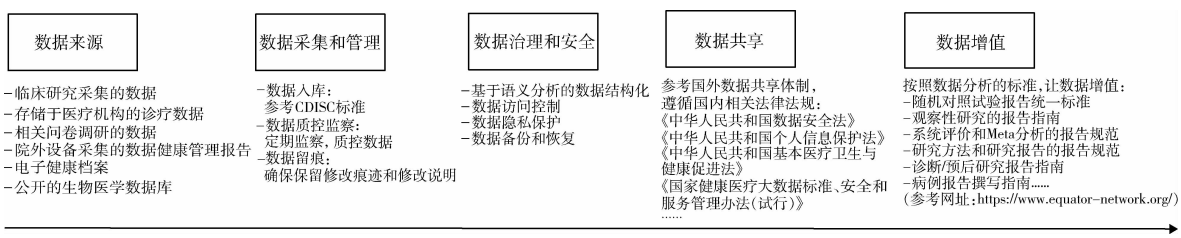

图 1 区域临床研究数据平台标准化建设内容

5.1 数据来源标准化

确保数据来源信息清晰明确。要求从数据采集的初始阶段即建立完备记录，详细注明数据源头、采集时间、采集地点以及采集方式，为数据的准确性、一致性和合法性，以及为支持高质量临床研究提供保障。

5.2 数据采集和管理标准化

按照国际公认的临床数据交换标准协会（Clinical Data Interchange Standards Consortium，CDISC）标准体系，统一数据收集的标准化格式[15]。首先，制订适用于临床研究数据的标准和规范，包括定义数据收集字段名称和测量方法、统一数据格式等关键要素。其次，采用标准化的数据监控和审核方法，确保数据在每个阶段均准确。最后，采用标准化的数据采集工具，确保不同研究场地或不同研究人员采集到的数据具有一致性。

5.3 数据治理和安全标准化

数据治理过程中，应建立数据字典，明确数据项的名称、定义、取值范围等信息，确保所有数据项在研究中使用统一标准；采用标准化表单，确保数据的采集过程符合规范，且可以方便地整理和存

储数据。为了提高数据结构化效率，要求根据不同病种构建对应的语义分析模型，形成迭代优化机制[16]。医疗数据安全的标准化，通过数据访问控制和数据访问权限管理确保数据只被授权用户或实体访问与使用；数据隐私保护确保数据处理符合法律法规和伦理准则；通过定期数据备份和恢复，最大限度地保障数据的安全性和稳定性。

5.4 合理的数据共享

参考国外数据共享模式，遵循国内相关法律法规，形成标准化的安全共享机制，包括身份验证和访问控制、数据加密、数据隐私保护、审计和监控、合规性和法规遵循、数据标准和格式、数据所有权和责任、风险管理和应急响应、数据生命周期管理以及合同或协议规定，规范医疗数据共享。

5.5 参照规范化准则推动数据增值

数据增值即产生有价值的分析报告，是体现临床研究价值的阶段。当前不同类型的研究均有规范的报告标准（www. equator - network. org），以 E-QUATOR 规范化的科学研究报告准则为参照，帮助研究者提高研究报告的准确性和规范性，强调采用公认的、规范化的准则制订和呈现研究分析结果，推动医学领域的知识共享和推广，实现数据的最大化增值。

6 区域临床研究数据治理平台的实践和构建要点

6.1 面向疾病的标准数据字段库建设

聚焦专病病种，打造区域级专病数据和生物样本库等扩展库，为开展专病临床研究提供重要标准和规范。一是建立具有迭代机制的疾病数据字段集合。在区域内选择特定疾病的一名或多名临床专家，基于其丰富的临床经验形成 1.0 版疾病数据字段集合。随着数据采集和研究发展的迭代不断完善。二是参考国际组织、学术机构、专业协会已制定的相关数据标准和数据模型，如 CDISC 标准等，构建对应的标准字段名称。同时创建数据字典，定义数据库中使用的变量和数据字段，确保每个变量的定义清晰明确，开发对应的电子数据表单，支持数据的标准化采集[17]。

6.2 参考工业级标准的多功能信息系统构建（图 2）

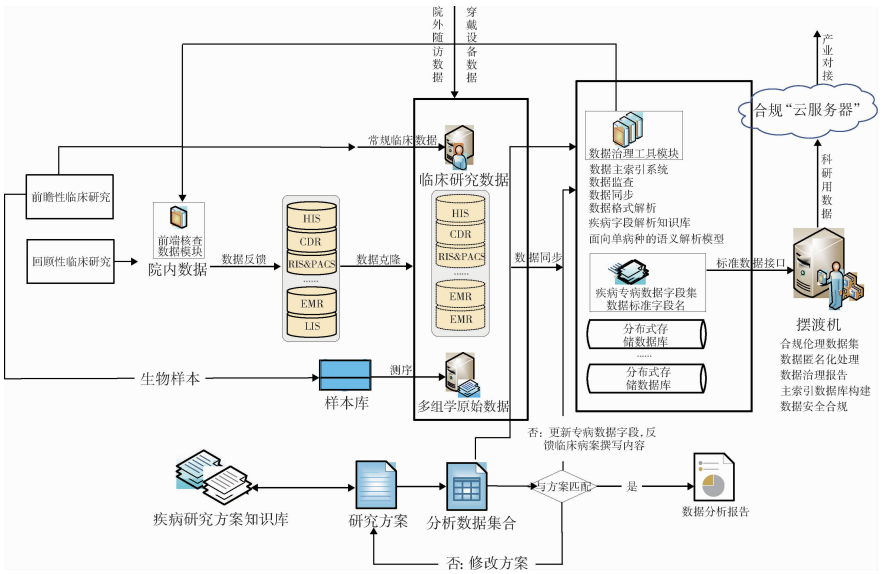

图 2 系统架构体系

注：医院信息系统（hospital information system，HIS）；临床数据仓库（clinical data repository，CDR）；放射科信息系统（radiology information system，RIS）；影像存储与传输系统（picture archiving and communication system，PACS）；电子病历（electronic medical record，EMR）；检验信息系统（laboratory information system，LIS）

基于业务流程设计信息化系统。首先，同步作为区域数据源头的院内数据，形成院内数据的克隆镜像，同时监测原始数据库的日志文件，形成增量同步机制；其次，对克隆镜像进行数据治理，构建语义解析模型，打造规范的、有索引的结构化数据，形成专病库建设的基础；最后，打通标准数据接口，通过“摆渡机”汇集数据。在研究层面形成疾病研究方案知识库，通过评估研究方案和数据一致性，不仅能验证数据集的完备程度，还为数据补全提供依据；更新语义解析模型，结合病历数据录入操作，形成前端数据核查模块，在不影响医生撰写病历的前提下，补全数据维度。此系统还可进一步整合前瞻性临床研究的数据，纳入多组学原始数据、院外随访数据和可穿戴设备数据，形成维度更全面、质量更高的临床研究数据整合体系，见图 2。系统中核心功能模块均选择工业级信息系统，即完成过全球监管部门认可项目的信息系统，包括电子数据捕捉系统、临床研究项目管理系统等。

6.3 数据采集规则和质控规则建立

科研立项后，评估研究方案和采集数据方案，规范数据的采集规则，包括数据来源、采集方式、测量标准等。除了采用系统内置或用户自定义的核查规则外，建立完善的数据质控体系和标准化的操作流程。通过整合信息化工具和引入外部人力服务，对采集的数据进行质量控制，实现数据的过程化管理，确保数据质量。

6.4 多中心临床研究数据整合

多中心临床研究均要按照上述规范和标准推进，大量高质量数据汇集在平台上，逐步形成高质量的区域临床研究数据平台，同时形成标准提升和工具优化的迭代体系。

6.5 科研立项政策支持和行政推动

为推进多中心临床研究数据融合，区域内医学院曾尝试依托医学院整合附属医院的方式，即通过医学院牵头建立临床研究大数据管理平台，整合附属医院的临床研究数据。但由于医学院对附属医院不具备完全的管理能力，对医院的政策和经费支持力度不足，数据整合工作受阻。因此，应由三甲医院上一级管理单位牵头，通过科研立项的方式，提供可落实的政策和经费支持，促进研究者通过发布课题开展规范性更高的临床研究[18]。

6.6 共享体系建设

国际主流研究或临床研究数据库的共享模式已形成相应标准和操作流程[19]。我国区域临床研究数据平台建设目前包含数据产生阶段、数据管理阶段、数据共享阶段，数据采集过程暂未完成，因此数据增值阶段还缺乏实践案例，见图 3。

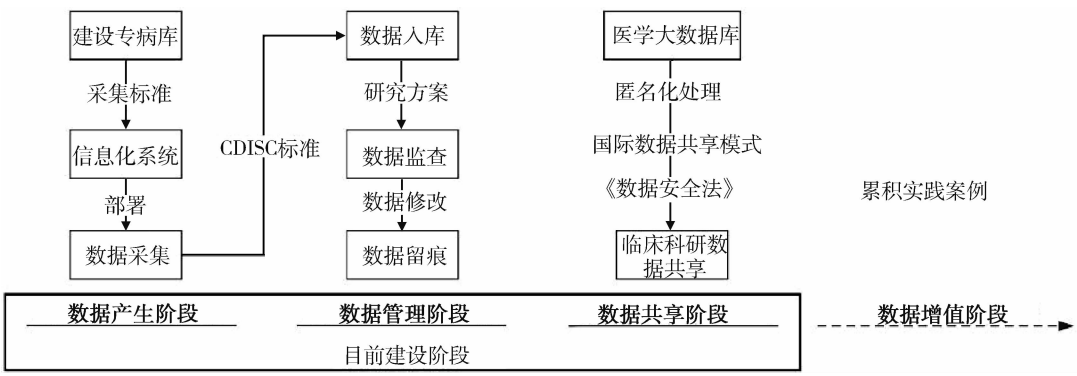

图 3 区域临床研究数据平台建设

7 结语

通过制订数据标准、建立专病队列数据库和生物样本库等扩展库、借鉴相关法律法规和标准规范，初步形成具有实践性的区域临床研究数据平台。但是，要进一步发挥该平台增值链上的潜在科研数据价值，还要不断支持符合国家发展的关键性临床研究项目立项，即加大对该数据平台的实际效用验证。因此，后续研究重点是加强平台运行，进一步论证和优化该平台对高质量临床研究的支持。不断扩大临床数据积累，有效整合全国优质资源，建立完善的数据共享机制和医企对接平台。这些平台不仅促进了生物医药产业的对接与协作，更为中国生物医药的发展提供了坚实有力的“中国证据”，为推进“中国方案”提供了有力支撑。

利益声明：*所有作者均声明不存在利益冲突。*

参考文献

1 李奕萱，谢丽，钱碧云．研究者发起的临床研究项目监管体系：现状与进展［J］．中国新药与临床杂志，2020，39（3）：146－150.

2 国务院办公厅关于推动公立医院高质量发展的意见［EB/OL］．［2023－06－04］．https：//www.gov.cn/zhengce/zhengceku/2021－06/04/content_5615473.htm.

3 国家卫生健康委员会．医疗卫生机构开展研究者发起的临床研究管理办法（征求意见稿）［J］．中国实用乡村医生杂志，2021，28（4）：5.

4 《关于抓好推动公立医院高质量发展意见落实的通知》印发［J］．中医药管理杂志，2022，30（3）：1.

5 邓军增．医院健康医疗数据治理探讨［J］．医学信息学杂志，2021，42（8）：4.

6 BYCROFT C，FREEMAN C，PETKOVA D，et al. The UK Biobank resource with deep phenotyping and genomic data［J］. Nature，2018，562（7726）：203－209.

7 LERAY M，KNOWLTON N，HO S L，et al. GenBank is a reliable resource for 21st century biodiversity research［J］. Proceedings of the national academy of sciences，2019，116（45）：22651－22656.

8 TOMCZAK K，CZERWIŃSKA P，WIZNEROWICZ M. Review The Cancer Genome Atlas（TCGA）：an immeasurable source of knowledge［J］. Contemporary oncology/współczesna onkologia，2015（1）：68－77.

9 于广军，杨佳泓，郑宁，等．上海市级医院临床信息共享项目（医联工程）的建设方案与实施策略［J］．中国医院，2010（10）：9－11.

10 岳和欣，湛永乐，边峰，等．临床队列研究的数据标准与共享［J］．中华流行病学杂志，2021，42（7）：1299－1305.

11 傅昊阳，徐飞龙，范美玉．论医院健康医疗大数据治理及体系构建［J］．中国中医药图书情报杂志，2019，43（3）：5.

12 马灿．国内外医疗大数据资源共享比较研究［J］．情报资料工作，2016，37（3）：63－67.

13 周晓梅，李烁，崇雨田，等．临床研究数据标准化工作的思考［J］．临床内科杂志，2022，39（11）：790－792.

14 肖辉，刘坤，杨章衡，等．临床数据中心建设方法探讨［J］．中国数字医学，2012，7（11）：70－72.

15 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.

16 徐维，曹洪欣，邱君瑞．电子病历用于临床研究的元数据概念及语义建构［J］．情报学报，2011，30（6）：668－672.

17 吴燕秋，梁公文，王天兵．面向临床研究的医院真实世界数据治理与建议［J］．医院管理论坛，2021，38（11）：4.

18 程晓华，舒展，徐文炜，等．新形势下研究者发起的临床研究立项管理要点［J］．医药导报，2022，41（2）：266－269.

19 WANG Z，JENSEN M A，ZENKLUSEN J C. A practical guide to the cancer genome atlas（TCGA）［M］. New York：Humana Press，2016.