

· 论著 · 一次研究 ·

真实世界数据质量评价指标研究



介 万^{1, 2, 3}, 姚明宏^{1, 2, 3}, 张 军⁴, 王鸣岐^{1, 2, 3}, 贾玉龙^{1, 2, 3}, 刘艳梅^{1, 2, 3}, 王雨宁^{1, 2, 3},
Larry Z. LIU^{5, 6}, 邹 康^{1, 2, 3}, 孙 鑫^{1, 2, 3}

1. 海南乐城真实世界研究院 (海南琼海 571400)
2. 四川大学华西医院中国循证医学中心 (成都 610041)
3. 四川省真实世界数据技术创新中心 (成都 610041)
4. 默沙东研发 (中国) 有限公司医学事务部 (北京 100012)
5. 默沙东有限公司V&I Outcomes Research (美国新泽西 07065)
6. 威尔康奈尔医学院 (美国纽约 10021)

【摘要】目的 构建可量化评价真实世界数据质量的指标。**方法** 计算机检索PubMed、Europe PMC、WanFang Data、CNKI文献数据库, 以及13个国际学术组织和国家/地区监管机构的官方网站, 根据事先制定的纳入排除标准筛选有关真实世界数据质量评价的文献、指南和标准。通过主题归纳和内容总结, 提炼文献和指南的核心要素, 构建指标内容, 并通过专家咨询法进行修改和完善。**结果** 最终纳入35篇文献, 结合团队经验, 构建基于前瞻性研究收集数据的数据管理流程和基于回顾性研究收集数据的数据治理流程, 并制定数据质量评价指标, 将数据的原始质量、过程质量和结果质量作为3个一级指标, 按照临床研究的时间顺序进行维度划分, 数据管理部分得到15个二级指标, 43个三级指标, 数据治理部分得到13个二级指标, 29个三级指标。**结论** 基于文献分析和专家咨询法建立的指标具有一定的科学性和可行性, 可为监管部门、申办方和学者开展真实世界数据质量评价提供一定的参考。

【关键词】 真实世界研究; 真实世界数据; 质量评价; 数据管理; 数据治理

【中图分类号】 R181.3+5 **【文献标识码】** A

Research on indicators for evaluating the quality of real-world data

JIE Wan^{1,2,3}, YAO Minghong^{1,2,3}, ZHANG Jun⁴, WANG Mingqi^{1,2,3}, JIA Yulong^{1,2,3}, LIU Yanmei^{1,2,3},
WANG Yuning^{1,2,3}, Larry Z. LIU^{5,6}, ZOU Kang^{1,2,3}, SUN Xin^{1,2,3}

1. Hainan Lecheng Institute of Real-World Study, Qionghai 571400, Hainan Province, China
 2. Chinese Evidence-based Medicine Center, West China Hospital, Sichuan University, Chengdu 610041, China
 3. Sichuan Center of Technology Innovation for Real World Data, Chengdu 610041, China
 4. Department of Medical Affairs, MSD R&D (China) Co., Ltd., Beijing 100012, China
 5. V&I Outcomes Research, Merck & Co., Inc., Rahway, NJ 07065, USA
 6. Weill Cornell Medical College, New York, NY 10021, USA
- Corresponding author: SUN Xin, Email: sunxin@wchscu.cn

DOI: 10.12173/j.issn.1005-0698.202411008

基金项目: 海南省科技厅重点研发项目 (ZDYF2024LCLH002、ZDYF2024LCLH003); 美国默克公司支持课题项目
通信作者: 孙鑫, 博士, 研究员, 博士研究生导师, Email: sunxin@wchscu.cn

【Abstract】Objective To develop an index system for the quantitative evaluation of real-world data quality. **Methods** Computerized searches were conducted across literature databases, including PubMed, Europe PMC, WanFang Data, and CNKI, as well as the official websites of 13 international academic organizations and regulatory agencies. Based on predefined inclusion and exclusion criteria, literature, guidelines, and standards related to real-world data (RWD) quality assessment, were ultimately selected. Through thematic induction and content summarization, core elements from the selected documents and guidelines were extracted to construct an initial set of indicators. These indicators were subsequently revised and refined using expert consultation methods. **Results** A total of 35 publications were included. Incorporating the research team's practical experience, this paper explored data management processes for prospective data and data governance processes for retrospective data. Corresponding data quality evaluation indicators were established accordingly. In both frameworks, data quality was categorized into three primary dimensions: raw data quality, process quality, and outcome quality. These dimensions were further organized according to the chronological sequence of clinical research. For the data management component, 15 secondary indicators and 43 tertiary indicators were developed. For the data governance component, 13 secondary indicators and 29 tertiary indicators were formulated. **Conclusions** The evaluation indicators developed through literature analysis and expert consultation method demonstrate a considerable degree of scientific rigor and feasibility. They can serve as valuable references for regulatory authorities, sponsors, and researchers involved in the quality assessment of real-world data.

【Keywords】 Real world study; Real world data; Quality evaluation; Data management; Data governance

真实世界数据 (real world data, RWD) 指从多种来源收集的与研究对象健康状况和 / 或常规诊疗及保健相关的数据^[1-2], 其核心是在实际医疗环境下形成和收集的数据^[3]。真实世界证据 (real world evidence, RWE) 是指对适用的 RWD 进行恰当和充分的分析所获得的关于药物使用情况和潜在获益风险的临床证据, 是药物有效性和安全性评价证据链的关键组成部分^[4]。

高质量 RWD 是产生高质量 RWE 的基础, 在真实世界研究中, 通常需要对 RWD 进行质量评价。数据质量是指为达到用户在健康研究、政策制定和监管方面的需求, 从数据中得到合理的结论或解释^[5-6]。数据质量的一个关键概念是“适用性”, 是指数据满足使用者需求的质量评价程度, 即从数据使用者角度出发进行数据质量评价^[7]。数据质量问题可对医疗卫生服务中临床决策和研究对象健康产生重要影响, 同时由于 RWD 来源广泛^[3], 在使用时可能面临较多问题, 例如, 单一数据库出现部分关键变量或数据记录缺失和记录不准确, 数据库之间的链接可能缺乏研究对象的唯一标识符, 难以进行数据溯源和数据链接。

中国国家药品监督管理局 (National Medical Products Administration, NMPA) 颁布实施的《用于产生真实世界证据的真实世界数据指导原则 (试行)》^[2] 将评价维度分为相关性和可靠性, 与美国食品药品监督管理局 (U.S. Food and Drug Administration, FDA) 的《真实世界证据框架》^[1] 对于数据质量评价的内容相一致。在该指导原则发布之前, 已有国内学者对 RWD 相关性和可靠性进行了阐述^[8], 但尚未见可操作的指标构建。基于已有文献、国内外相关 RWD 指南文件, 考虑研究开展的时间和指标的可操作性, 本研究通过文献筛选和专家咨询方法, 阐述 RWD 的数据管理和数据治理的流程, 以及两者的质量评价指标, 以期监管部门、申办方和学者开展真实世界研究提供参考, 获取高质量 RWD。

1 资料与方法

1.1 文献纳入排除标准

纳入标准: 研究内容涉及 RWD, 具有数据质量评价和指标介绍的文献。排除标准: ①研究方案、会议摘要、观点评论类文献; ②研究内容为新型冠状病毒、基因研究的文献。

1.2 文献检索策略

计算机检索 PubMed、Europe PMC、WanFang Data、CNKI 4 个文献数据库, 以及世界卫生组织 (World Health Organization, WHO)、美国 FDA、杜克-马戈利斯卫生政策中心 (Duke-Margolis Center of Health Policy)、欧洲药品管理局 (European Medicines Agency, EMA)、欧洲药物流行病学和药物警戒网络中心 (European Network of Centers for Pharmacoepidemiology and Pharmacovigilance, ENCePP)、欧洲健康大数据工作组 (Towards European Health Data Space, TEHDAS)、日本药品医疗器械管理局 (Pharmaceuticals and Medical Devices Agency, PMDA)、中国国家卫生健康委员会、中国 NMPA、中国海南省卫生健康委员会、中国海南省药品监督管理局、中国海南自由贸易港、中国海南博鳌乐城国际医疗旅游先行区等政府网站。检索时间为建库至 2023 年 11 月 15 日。

使用主题词与自由词相结合的检索方式, 中文关键词: 真实世界数据、电子医疗记录、注册登记数据、数据质量、数据治理、数据管理; 英文关键词: real world data, RWD, Electronic Health Records, data quality, data management, data governance。以 PubMed 数据库为例, 具体检索策略见框 1。

1.3 初步构建评价指标

由 2 名研究者对检索到的文献进行独立筛选, 如遇分歧, 则与第 3 名研究者讨论后决定。将筛选获得的文献进行归纳整理, 按照研究开展的时间总结数据管理和数据治理的基本流程, 对其中

```
#1 "real world data" [Title/Abstract] OR "RWD" [Title/Abstract]
OR "real world data" [Title/Abstract] OR "Electronic Health
Records" [MeSH Terms] OR "EHR" [Title/Abstract] OR
"registry" [Title/Abstract] OR "electronic health record"
[Title/Abstract] OR "computerized medical record"
[Title/Abstract] OR "medical records systems, computerized"
[MeSH Terms]
#2 "data quality" [Title/Abstract] OR "quality assurance"
[Title/Abstract] OR "data evaluation" [Title/Abstract] OR
"quality control" [Title/Abstract] OR "data management"
[Title/Abstract] OR "data governance" [Title/Abstract]
#3 #1 And #2
```

框1 PubMed 检索策略

Box 1. Search strategy in PubMed

涉及的指标进行整理, 结合我国 RWD 质量评价的现状, 初步拟定评价指标。

1.4 专家咨询法

采用专家咨询法对初步构建的评价指标进行评审优化。专家纳入标准: ①来自医院临床流行病学科室、疾控系统、科研院校、药品监督管理局等从事真实世界研究的专家; ②具有副高级及以上职称, 本科及以上学历; ③已发表与真实世界研究相关的学术论文。将文本资料以电子邮件的形式发送给专家审阅, 请专家对初步建立的框架指标进行评价提出修订意见。经过 2 轮专家咨询, 进一步对指标进行调整和完善, 构建最终的评价指标。

2 结果

2.1 文献筛选流程及结果

初步检索文献数据库获得文献 2 787 篇, 检索相关政府网站获得指南文件 46 篇, 经筛选, 最终纳入 35 篇研究指南或文献。文献筛选流程见图 1, 纳入研究的基本特征见表 1。

2.2 数据管理的基本流程

数据管理是指根据研究目的和主要指标进行

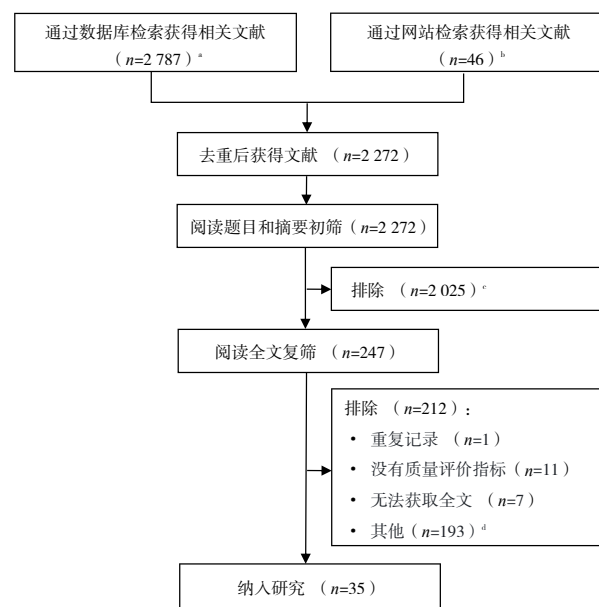


图1 文献筛选流程图

Figure 1. Flow chart of literature screening

注: a所检索的数据库及检出文献数具体如下, PubMed (n=1 961)、Europe PMC (n=802)、CNKI (n=17)、WanFang Data (n=7); b所检索的政府机构网站及检出文献数如下, WHO (n=5)、FDA (n=16)、Duke-Margolis Center (n=3)、EMA (n=4)、PMDA (n=7)、NMPA (n=11); c主要包括新型冠状病毒、基因相关文献、观点评论类文章; d主要包括与研究不相关的文献。

表1 纳入的文献特征
Table 1. Characteristics of included publications

文献名称	主要内容	发布年份	发布者	来源
Real-world evidence ^[1]	RWD和RWE的定义、应用RWE进行药物或医疗器械注册申请的指南框架、使用电子健康记录的指南文件、评价电子健康记录和医保数据用于药物和器械注册的指南及使用RWD用于提交的数据标准等	2023	FDA	网站
用于产生真实世界证据的真实世界数据指导原则（试行） ^[2]	RWD的来源及应用挑战、RWD的适用性评价（数据治理和数据管理）、RWD数据治理包含的具体内容、RWD合规性和安全性	2021	CDE	网站
建立真实世界数据与研究技术规范，促进中国真实世界证据的生产与使用 ^[3]	RWD与真实世界研究的理论框架、RWD与研究型数据库、真实世界研究设计及相关的伦理问题	2019	孙鑫，等	文献数据库
真实世界证据支持药物研发与审评的指导原则（试行） ^[4]	RWD和RWE的定义和RWD数据来源、数据的适用性、RWE支持药品监管决策	2020	NMPA	网站
Guidelines for quality assurance in multicenter trials: a position paper ^[5]	数据质量定义及包含的方面	1998	Knatterud GL, et al	文献数据库
High-quality data to empower data-driven medicines regulation in the European Union-Medicines Agency ^[6]	欧盟药品监管数据质量框架：数据质量定义、数据生命周期、数据质量维度和指标（可靠性、完整性、一致性、及时性和相关性）	2022	EMA	网站
A conceptual framework for evaluating data suitability for observational studies ^[7]	观察性研究数据的适用性分析框架	2018	Shang N, et al	文献数据库
利用好真实世界数据生产高质量真实世界证据支持药械监管 ^[8]	RWD相关概念、数据适用性的核心维度	2020	姚晨	文献数据库
特许医疗政策下真实世界数据治理的模式探索 ^[9]	RWD数据治理和数据管理流程及关键考虑	2021	贾玉龙，等	文献数据库
患者登记研究的策划与患者登记数据库构建：基于观察性设计的真实世界研究 ^[10]	登记研究的定义、类型和主要应用；登记研究的数据库构建流程，数据收集与管理过程	2018	谭婧，等	文献数据库
临床试验的电子数据采集技术指导原则 ^[11]	数据质量和真实完整性、电子数据采集的流程及各个环节的要点	2016	NMPA	网站
药物临床试验数据递交指导原则（试行） ^[12]	数据传输递交说明	2020	CDE	网站
真实世界数据体系构建的模式探索 ^[13]	RWD体系构建流程；数据构成方案说明的相关研究问题：数据变量指标和定义、数据来源、数据采集方式、数据处理流程	2022	刘艳梅，等	文献数据库
基于既有医疗数据构建研究型数据库的方法学探讨及实例解读（二）：数据治理的方法 ^[14]	基于医疗健康数据进行的数据治理流程：数据链接、数据提取、数据核查、数据清理	2023	赵国桢，等	文献数据库
医疗大数据质量评价指标体系研究 ^[15]	运用文献研究和专家咨询法构建医疗大数据质量评价指标体系和筛选完善指标	2020	李姮	文献数据库
Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research ^[16]	电子健康记录数据质量评价：完整性、准确性、一致性、合理性、及时性	2013	Weiskopf NG, et al	文献数据库
Assessment of the nursing quality indicators for reporting and evaluation (NQuIRE) database using a data quality index ^[17]	数据质量的维度和特征：数据完整性、一致性、准确性、代表性、相关性、可解释、一致性和及时性等	2020	Naik S, et al	文献数据库
Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data ^[18]	RWD（尤其是电子健康记录）数据质量评价：及时性、准确性、合理性、完整性、一致性、可比性、合规性、灵活性、相关性、可用性/易用性、安全性，以及信息丢失与退化	2020	Bian J, et al	文献数据库

续表1

文献名称	主要内容	发布年份	发布者	来源
医疗大数据质量评价指标体系构建研究 ^[19]	运用文献研究和专家咨询法从结构、过程和结果三维度数据质量评价指标体系	2019	熊兴江	文献数据库
Quality assessment of real-world data repositories across the data life cycle: a literature review ^[20]	数据质量框架：一致性、完整性和合理性等；数据质量的测量指标；数据质量的评估工具	2021	Liaw S, et al	文献数据库
Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0 ^[21]	医疗数据库的报告规范：数据来源、数据抽样、数据链接、数据清理、暴露和结局定义、随访时间、协变量等	2017	Wang SV, et al	文献数据库
The rise of big clinical databases ^[22]	临床医疗数据库的使用挑战：数据质量围绕完整性和准确性，数据的可解释性	2015	Cook JA, et al	文献数据库
Characterizing RWD quality and relevancy for regulatory purposes ^[23]	适用的RWD包括相关性和数据质量（准确性、完整性、数据处理的透明性）两个维度；数据来源、数据清理、转换和链接	2018	Duke–Margolis	网站
Data quality assessment framework to assess electronic medical record data for use in research ^[24]	电子医疗记录数据的质量评价框架：完整性、准确性、一致性、合理性和及时性	2016	Reimer AP, et al	文献数据库
Anchoring data quality dimensions in ontological foundations ^[25]	数据质量维度和指标说明：准确性、可靠性、及时性、相关性、完整性、一致性	1996	Wand Y, et al	文献数据库
Data quality assurance (DQA) ^[26]	数据质量指标：完整性和及时性、内部一致性、外部不同数据源比较和验证、数据样本量	2023	WHO	网站
A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research ^[27]	临床研究数据适用性概念框架；数据准确性、合理性、数据来源、及时性、数据样本量	2012	Kahn MG, et al	文献数据库
Directory of clinical databases: improving and promoting their use ^[28]	评价临床数据库的标准：准确性、完整性、数据库包含的变量、变量定义、编码的可靠性、主要结局指标、数据验证、代表性	2003	Black N, et al	文献数据库
Assessing the quality of clinical and administrative data extracted from hospitals: the general medicine inpatient initiative (GEMINI) experience ^[29]	评价临床数据库：完整性和合理性；通过计算的方式评价数据质量	2021	Verma AA, et al	文献数据库
Automated procedure to assess civil infrastructure data quality: method and validation ^[30]	评价数据质量完整性：数据缺失	2005	Buchheit RB, et al	文献数据库
Scaling and assessment of data quality ^[31]	评价数据质量完整性：数据缺失	2006	Evans P	文献数据库
Increasing trust in real-world evidence through evaluation of observational data quality ^[32]	观察性数据的质量评价：一致性、完整性和合理性	2021	Blacketer C, et al	文献数据库
Towards a content agnostic computable knowledge repository for data quality assessment ^[33]	通过计算的方式评价数据质量，包含准确性、完整性、一致性和及时性等维度及相应的测量描述	2019	Rajan NS, et al	文献数据库
回顾性数据库研究的概念、策划与研究数据库构建 ^[34]	回顾性数据库研究：数据源人群和研究人群、暴露和结局指标定义及编码识别、数据提取、伦理申请、数据库样本量、评估数据质量、数据清理	2018	王雯，等	文献数据库
Administrative database code accuracy did not vary notably with changes in disease prevalence ^[35]	评价疾病诊断编码的准确性	2016	Carl VW, et al	文献数据库

注：CDE. 国家药品监督管理局药品审评中心（NMPA Center for Drug Evaluation）。

患者诊疗信息的主动收集,适用于前瞻性研究设计,核心为保证数据质量、提高研究效率和控制研究成本,以确保数据满足特定研究目的的需要。数据管理过程需要前瞻性地随访和收集数据,包括但不限于以下方面^[9]:

(1) 制定数据管理计划: 主要包括明确研究目的、研究对象的纳入排除标准、随访计划、数据采集维度、数据收集方法、数据变量定义和规则、数据管理和质量控制措施等内容,以及描述和定义从研究开始到数据库锁定的数据管理流程。

(2) 设计病例报告表(case report form, CRF): 根据研究方案要求设计和记录研究中每名研究对象全部需采集的信息,统一并规范化数据采集流程、病例报告表填写标准,通过预填写完善病例报告表及填写指南。通常需要采集的数据要素包括: 人口学信息(如年龄、性别、疾病史、家族史等)、基线诊断(如主要诊断、次要诊断、病程、病情严重程度、合并症等)、干预/暴露(包括干预/暴露和对照药物的种类、干预/暴露的起始时间、单次剂量、累积剂量、实际使用频率、用药时长等)、结局指标(如临床效果指标、安全性指标、患者报告结局等)。若研究涉及经济学评价,还可采集成本相关数据,包括直接成本(如挂号费、药费、手术费、诊疗费、治疗费、材料费、护理费、检验费、检查费等直接医疗成本,由治疗产生的交通费、照护费等直接非医疗成本)和间接成本(如由于治疗产生的误工损失、因疾病带来的劳动力损失等)。

(3) 建立电子化数据采集(electronic data capture, EDC)系统: 根据病例报告表搭建 EDC 系统并进行充分的测试,系统需要兼备数据录入、数据核查质疑、系统风险管理、数据传输、轨迹留痕等功能,并通过预录入数据,完善系统功能。

(4) 研究对象招募与随访: 应充分考虑研究对象招募条件和纳入排除标准,可适当扩大招募范围,确保人群代表性。为避免主观选择,符合纳入标准的研究对象均应连续纳入。采用统一的方式对不同来源的研究对象进行随访,提高患者随访率,可适当增加样本量避免因患者失访导致的样本不足。

(5) 数据采集和录入: 制定标准化的数据收集手册,采用统一和系统化的方法收集数据,实现数据标准化和医学编码的统一便于数据的交

流互通。通常需包含的内容有: 研究计划书、患者纳入排除标准、数据收集工具、数据字典和数据验证规则、重复记录处理、病例报告表使用手册等。

(6) 数据核查: 贯穿于数据采集和存储传输的整个流程,确保数据的准确可靠,可包括: 随机化核查(如适用,例如实效性随机对照临床试验)、违背方案核查(如适用,例如实效性随机对照临床试验)、时间窗核查、逻辑核查、范围核查、一致性核查等。数据核查应在未知试验分组的情况下进行,可通过人工或者电脑程序进行,若需要数据修订则进行详细记录。数据库锁定后,若需修改,应书面申请数据库负责人,在获得同意并进行严格记录的情况下,方可修改。数据修改过程须严格记录。

(7) 数据传输递交: 遵循相关的数据递交标准,提高数据的可读性,保证数据安全可溯^[10]。其中部分过程可贯穿整体研究过程^[11-12],如图2所示。

2.3 数据治理的基本流程

数据治理是指针对特定的临床研究问题,为达到适用于统计分析的目的而对原始数据进行的治理^[2, 13],其适用于回顾性研究设计,对于满足研究目的的常规健康医疗数据,建议以“ALCOA+CCEA”[ALCOA 指 attributable(可归因性), legible(易读性), contemporaneous(同时性), original(原始性), accurate(准确性); CCEA 指 complete(完整性), consistent(一致性), enduring(持久性), available(可获取性)]标准为基础框架,通过对不同数据源的数据提取和标准化,实现数据资源的整合利用,如图3所示,流程包括但不限于^[9]:

(1) 制定数据治理计划: 明确数据治理方案,常需要考虑数据来源、数据权属、数据链接、数据提取、数据脱敏、数据清洗、数据共享、缺失数据处理、质量控制、数据适用性评估等。

(2) 数据保密性处理: 对个人敏感信息需进行脱敏处理,防止个人信息的泄漏,提高数据安全性。

(3) 数据链接和提取: 基于现有数据库和既往数据库的变量和规则,制定数据字典,对规范化的数据进行管理,以便有效实现跨系统的数据提取和共享。通常建立患者唯一识别码,根据

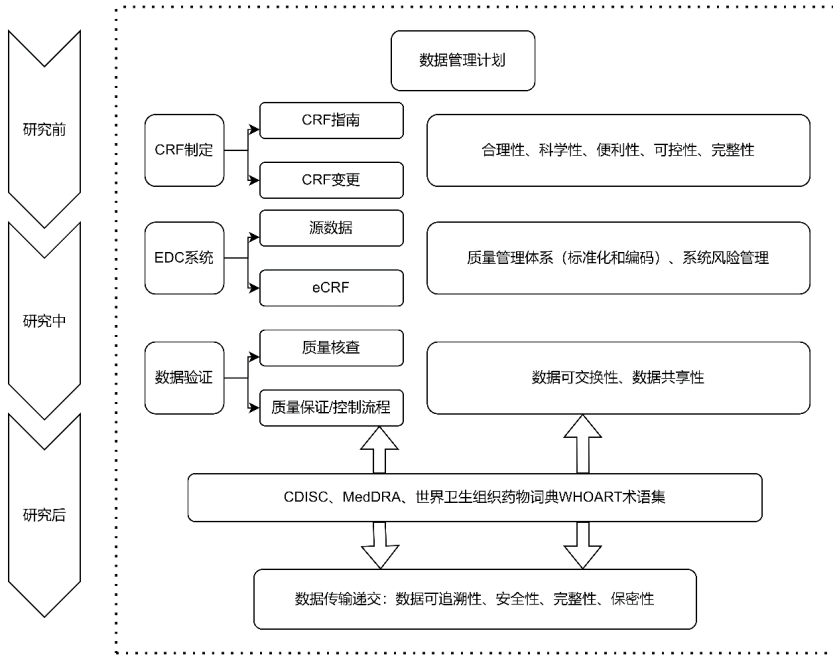


图2 数据管理的基本流程

Figure 2. The basic process of data management

注：CDISC.临床数据交换标准协会（Clinical Data Interchange Standards Consortium）；MedDRA.《国际医学用语词典》（*Medical Dictionary for Regulatory Activities*）；WHOART.世界卫生组织不良反应术语集（WHO Adverse Reaction Terminology）。

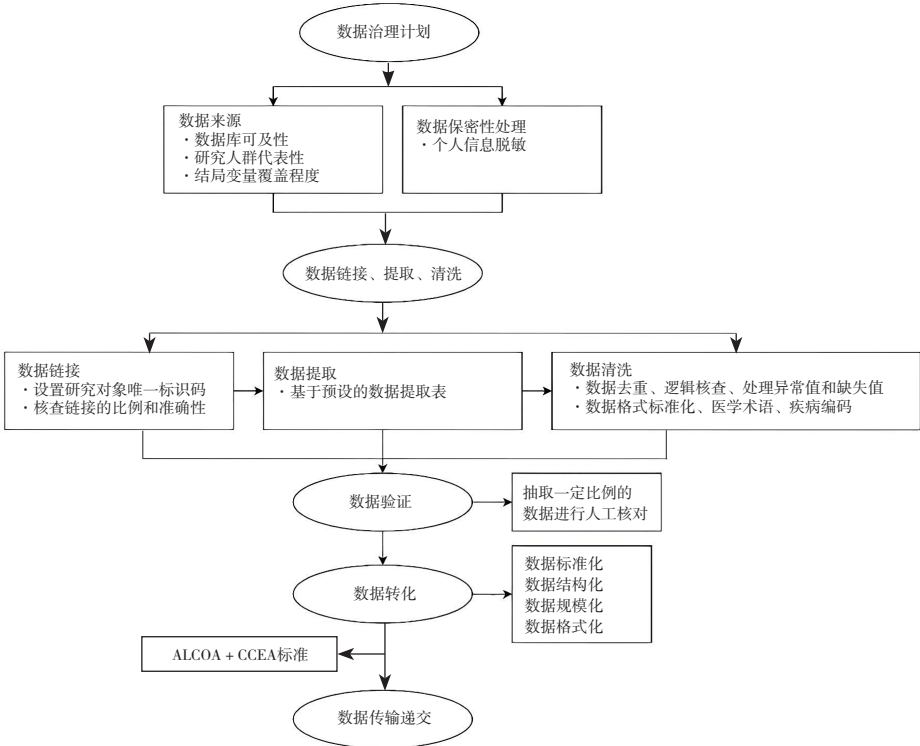


图3 数据治理的基本流程

Figure 3. The basic process of data governance

患者的唯一识别码实现多个数据库之间的数据链接；应根据提前设计的数据提取表进行数据的提取，确保数据来源的可追溯性；数据提取时，不同数据库之间的变量格式可能不同，可以根据既

往数据库中的变量格式和规则，制定数据字典，再进行提取；对于结构化程度较高的数据，如实验室检查数据可以直接采用计算机辅助提取的方式，对于非结构化数据，如从病程、主诉、既往

史中提取症状，生命体征数据，可使用自然语言处理技术实现数据结构化，再进行人工核查和计算机提取；数据提取过程中，若涉及人工提取环节，则需要考虑不同研究人员数据提取结果可能存在差异的情况^[14]。

（4）数据清洗：对提取的原始数据进行重复或冗余数据的去除，变量值的逻辑核查、异常值及缺失值的处理，明确矛盾数据处理优先级和文本信息结构化规则，整个数据清洗过程中需要保留原始数据及清理记录。

（5）数据验证：应对数据库构建中的关键环节开展数据验证，包括数据链接、数据提取及数据治理等。可随机抽取一定比例的数据与原始记录进行人工核对，评价其准确性。

（6）数据传输与存储：数据的传输和存储过程中均需要加密保护，其中涉及受试者个人敏感信息应进行匿名化处理，确保根据已有信息无法进行个人信息的重标识，建立数据治理各环节的加密技术、风险评估和应急处置操作流程。

2.4 RWD质量评价指标构建

RWD 质量评价指标构建主要考虑以下四方面因素：科学性、系统性、可操作性和可扩充性。

对检索收集到的文献资料进行梳理总结，将数据的原始质量、过程质量和结果质量作为 3 个一级指标，按照临床研究的时间顺序进行维度划分。

采用专家咨询法进行指标筛选。第 1 轮邀请 5 位专家，第 2 轮邀请 10 位专家，有 12 位专家返回了建议。结合专家意见，采纳了 1 条新增 1 个二级指标的意见（在数据管理过程质量条目下新增“数据采集”）；采纳了 1 条新增 2 个二级指标的意见（在数据治理结果质量条目下新增“数据相关性”“完整性”）；采纳了 1 条新增 2 个三级指标的意见（在数据治理原始质量伦理法规条目下新增“应符合相关法规要求”“伦理审批”）；采纳了 3 条名称修改建议（将二级指标名称“数据源链接”更改为“数据链接”，将二级指标名称“受试人群完整”更改为“研究对象纳入情况”，将二级指标名称“代表性”更改为“数据代表性”）；删除了 1 个二级指标（“数据管理的及时性”）。删除了 3 条三级指标（“预试验”“病例报告表完整”“数据库处于连续活动状态”）。

最终纳入指标在数据管理部分得到 15 个二级指标，43 个三级指标，数据治理部分得到 13 个二级指标，29 个三级指标。详见表 2、表 3。

表2 数据管理的各级指标及指标解释
Table 2. Hierarchical indicators for data management and explanations

一级指标 ^[15]	二级指标	三级指标	指标解释
原始质量	数据采集	及时采集数据 ^[16-18]	以百分比表示，可根据预先制定的随访计划，评价受试者数据信息采集率（已采集受试者信息占实际受试者数目）、研究数据采集（已采集数据点数量占所有预期采集数据点总数）来反映
		采集人员签名 ^[15]	以百分比表示，数据采集和录入均有相应的责任人姓名和记录日期
	伦理法规	数据的使用是否符合法规要求 ^[18]	达标要求为“是”
		伦理审批	达标要求为“是”，并保存完整的过程文件
		患者知情同意	可视情况申请豁免患者知情同意
	数据标准	病例报告表设计的及时性	达标要求为“是”，病历报告表是否在第一位受试者入选前获准使用
		数据采集标准 ^[15]	采用统一的数据采集标准
	数据库的合规性	启用数据库前通过测试并具备完整的测试内容	达标要求为“是”，并保存完整的记录文件，体现数据库在启用前经预先测试保证可操作
		数据库质控	达标要求为“是”，并保存完整的记录文件
		数据库是/否在首位受试者入选前获准启用	达标要求为“是”，并保存完整的记录文件
		数据库运行环境及安全备份情况测试符合要求并保留记录 ^[19]	达标要求为“是”，并保存完整的记录文件，体现数据库运行环境的安全保障
		数据库使用者在授权使用数据库前获得系统培训 ^[20]	达标要求为100%，并保存完整的记录文件
		数据盲态审核（如适用）	数据经过研究人员的盲态审核，并按照研究方案进行统计分析人群划分、核查不良事件/反应的报告与处理情况记录等

续表2

一级指标 ^[15]	二级指标	三级指标	指标解释
过程质量	项目经费	数据库锁定	提前制定数据库锁定工作程序
		基金来源	明确研究项目的基金项目来源
	数据采集	是否开展预试验	根据研究设计和目的开展预试验, 进行数据预采集, 完善数据采集的流程
		数据采集标准 ^[15]	采用相同的定义框架, 遵守相应的标准操作流程
		研究对象随访 ^[21-22]	需要长期随访的研究对象, 可分阶段制定数据核查计划, 及时处理在数据收集中可能出现的各种偏倚
	数据录入	是否采用统一的录入标准	应采用统一的录入标准, 达标要求为“是”
		评估数据录入的准确程度 ^[17]	可采用数据双录入的方式, 以百分比表示不一致的比例, 数值越小说明准确性越好
	数据存储 ^[19-20]	统一的数据存储平台, 且有密码保护	体现原始数据的保存质量、可及性和安全性
		数据的原始记录是否进行备份	达标要求为“是”, 并保存完整的记录文件
	数据链接 ^[21, 23-24]	数据链接	采用唯一标识码进行多源数据链接
		数据保密	数据链接、提取、分析是否进行数据的保密性处理方法
		数据链接准确性	同一个研究对象在不同数据库之间的链接成功率
	数据管理的完整性 ^[25-26]	研究进展阶段的数据管理文件是否完备	达标要求为“是”, 并保存完整的记录文件
		研究进展阶段的用于数据审核报告的列表及说明	达标要求为“是”, 并保存完整的记录文件
		数据溯源: 对相关文档、数据库、医学编码和编码字典等更新需要有具体记录 ^[23]	达标要求为“是”, 并保存完整的记录文件
		是否建立参与研究人员的简历, 并定期更新	达标要求为“是”, 并保存完整的记录文件
		对数据的原始文件进行核查记录, 确保数据库数据与源数据一致 ^[27]	达标要求为“是”, 并保存完整的记录文件
		数据录入系统是否具备进行数据检查、记录匹配等 ^[17]	建议为“是”, 便于后续对数据的管理和进一步的处理
		医学编码错误率 ^[28]	以百分比表示, 数值越小越好, 可计算错误数据的发生率
		违背方案核查	患者是否符合纳入排除标准; 用药计划及合并用药的规定(此点适用于实效性临床试验)等
结果质量	数据清理 ^[21, 23]	是否制定数据变量字典、明确的数据清理规则及流程	达标要求为“是”。应详细记录衍生变量的产生方式
		是否制定矛盾数据处理方法 ^[23]	建立矛盾数据处理优先级
		各个访视点的数据缺失情况、关键变量的缺失等 ^[23-24]	以百分比表示, 记录数据的缺失情况
		变量取值的逻辑核查和准确性核查 ^[21, 23, 27, 29]	定量数据的取值要在相应的取值范围内(年龄、身高等); 核查极值和离群值, 日期数据的格式, 定性数据的取值符合规定
		指标核查 ^[23]	对关键指标和非关键指标进行相应的核查; 随机抽取一定比例的数据进行核查, 计算核查一致率
		缺失值处理方法 ^[17, 21, 23, 30-31]	缺失值的比例越小越好, 若采用填补方法, 则记录填补方法和保存相应的数据操作过程
	完整性 ^[16, 25, 32-33]	研究对象纳入情况 ^[28]	参与研究的研究对象占签订知情同意书人员的比例, 达标要求为100%, 体现受试人群的完整性
		完整可分析的数据库 ^[23]	数据库锁定时, 不存在待处理的质疑数据条目、未经编码数据和未确认的缺失数据条目等
	可溯源性一致性 ^[25]	是否按照研究设计完成随访	100%完成随访, 尽量减少失访, 并要说明失访的比例和原因
		具有完整的数据稽查记录 ^[23, 27]	达标要求为“是”
	质量保障	记录受试者信息的各个临床研究管理系统的信息一致 ^[23, 27]	达标要求为100%, 若不一致, 寻找和记录原因并保存相关的处理方法
		因发现数据质量问题而产生的数据库解锁次数	建议为0次, 若发生则需详细记录解锁原因, 产生问题的数据, 解锁时间和修改内容并进行详细记录保存文件

表3 数据治理的各级指标及指标解释
Table 3. Hierarchical indicators for data governance and explanations

一级指标 ^[15]	二级指标	三级指标	指标解释
原始质量	数据可及性 ^[17, 34]	数据来源可及	是否提供数据访问链接、数据是否便于公开、是否可以用于支持监管机构核查
		伦理法规	注重数据安全和研究对象隐私信息的保护
		伦理审批 ^[34]	达标要求为“是”，并保存完整的过程文件
	数据提取	患者知情同意	可视情况申请豁免患者知情同意
		明确的纳入排除标准	根据研究方案制定的纳入排除标准，筛选符合条件的数据记录
	样本量	关键变量覆盖度 ^[22]	原始数据应包含研究相关的结局变量、暴露变量、协变量等关键变量
		源数据经数据治理后是否符合统计分析要求的样本量 ^[28, 34]	对于评价治疗结局、预后影响因素等研究问题，需评估数据库中样本量是否充分
	数据代表性	是否能够代表目标人群 ^[17, 28]	根据研究设计和目的进行评价
	项目经费	基金来源	明确研究项目的基金项目来源
	数据存储 ^[19-20]	数据存储具备密码保护	体现原始数据的保存质量、安全性
过程质量	数据链接 ^[23-24]	数据的原始记录是否进行备份	达标要求为“是”，并保存完整的记录文件
		是否链接研究相关变量	链接多源数据库通常可获得更全面、完整的变量信息
	数据清理 ^[23]	链接比例及准确性 ^[23]	提供链接的比例及验证链接的准确性，数据可链接的比例越大越好
		数据溯源 ^[23, 27]	对相关文档、数据库、医学编码和编码字典等更新需要有具体记录，达标要求为“是”，并保存完整的记录文件
		原始文件核查	达标要求为“是”，并保存完整的记录文件，确保数据库数据与源数据核对一致
		医学编码错误率 ^[28, 34-35]	以百分比表示，数值越小越好
		数据变量字典、清理规则及流程 ^[23]	达标要求为“是”，便于后续对数据的分析和利用
		是否包含研究结局 ^[23]	达标要求为“是”，数据库覆盖时长范围包含受试者预期结局，尤对探索长期用药治疗及慢性病预后相关研究
		是否包含关键变量 ^[22-24]	达标要求为“是”
		矛盾数据处理方法 ^[23]	建立矛盾数据处理优先级
结果质量	数据相关性	逻辑和准确性核查 ^[23, 27, 29]	定量数据的取值要在相应的取值范围内（年龄、身高等），核查极值和离群值，日期数据的格式，定性数据的取值符合要求
		评估变量的完整性，缺失值的处理方法 ^[17, 23, 30-31, 34]	缺失值的比例越小越好。若采用填补方法，则记录填补方法和保存相应的数据操作过程
		干预/暴露的定义 ^[28, 34]	暴露的定义应与研究问题的临床意义与理论依据相一致
		临床结局的定义 ^[28, 34]	临床结局定义需要基于相应的诊断标准、测量方法、测量工具、计算方法、测量时点、变量类型、终点事件评价机制等
		时间窗合理性	暴露和临床结局的定义需考虑时间因素，在合理的时间范围内收集相关的数据
		多源异构数据融合 ^[24]	通过身份标识符对个体水平的数据进行准确链接，以支持通用数据模型对数据源中关键变量进行整合
	完整性 ^[16, 23, 25, 32-33]	完整可分析的数据库数据 ^[23]	经治理完成的数据库在数据库锁定时，不存在待处理的质疑数据条目、未经编码的数据和未确认的缺失数据条目等，应该包含关键变量，如暴露/非暴露、协变量、混杂变量和结局变量，达标要求为不存在上述各项数据问题
	准确性 ^[16, 25]	数据与其描述的客观特征是否一致	数据值域是否在合理范围
	透明性 ^[23]	数据治理方案和治理过程清晰透明	数据分析中的暴露变量、协变量和结局变量可追溯至源数据，并反映数据的提取、清洗、转换和标准化过程；标准操作程序和验证确认文件清晰记录和存档
	质量保障指标	因发现数据质量问题而产生的数据库解锁次数	建议为0次，体现数据的质量问题，若发生则需详细记录解锁原因，产生问题的数据，解锁时间和修改内容并进行详细记录保存文件

3 讨论

本研究基于“原始-过程-结果”质量评价模型,采用文献分析的方法,根据研究设计和数据收集方式的不同,总结数据管理和数据治理的操作流程,并分别制定相应的数据质量评价指标。当前 RWD 研究逐渐广泛应用于创新药品的上市申请和上市后监测, RWD 质量直接关系到 RWE 的质量,不同机构提出了不同的 RWD 质量评价方法,但绝大多数缺乏可操作性。本研究结合国内外已有的指南文件和文献,尝试对现有的 RWD 质量进行系统和完整的指标构建,以期今后在实证研究中为 RWD 质量评价提供参考。

本研究首先构建 RWD 数据管理和数据治理的基本流程, RWD 来源广泛,相较于随机对照试验, RWD 在大多数情况下缺乏其记录、采集、存储等流程的严格质量控制,会造成数据不完整、关键变量缺失、记录不准确等问题^[2],通过针对前瞻性研究收集的数据进行数据管理,针对回顾性研究收集的数据进行数据治理过程,可跨地域、跨系统、跨领域的整合多源异构数据,增加研究样本量,记录数据操作过程可用于进行数据溯源,保证数据的准确和真实,提升数据的可利用性。在临床实践过程中,完整有效地记录数据,也是重要的改善数据适用性的方法。

通过筛选有关数据质量评价指标的文献和政府指南文件,初步确定各级指标的具体评价内容及相应的指标达标要求,运用专家咨询法对各指标进行评价。目前针对 RWD 质量评价的研究可能实施较为困难,而本研究构建的指标是从质量评价的整体出发,从数据的产生、数据管理/治理、存储、链接和质量保障等多个方面进行指标构建。各个二级指标的内容设定又是基于官方指南文件内容^[2, 6, 23],并且结合拟评价 RWD 的实际情况,这些指标均具有可操作的性质。三级指标是结合实际临床试验研究和从数据质量的需求出发进行指标设定,各个指标具有详细的达标要求描述。因此本研究构建的评价指标具有较强的可操作性和可行性。

指标评价可为卫生监管部门、医疗企业和医疗机构研究者对 RWD 质量评价提供参考。目前 NMPA 已经发布相关的指导原则,本研究在指导原则的基础上进一步将研究内容细化,制定较完

整可行的指标,加强 RWD 质量,使监管部门、医疗申办方或数据管理机构和医疗机构研究者达到全面、系统审查和管理临床试验数据质量的目的,同时也便于在不同的试验中心或申办机构之间进行数据质量对比,提供高质量数据。

该评价指标实际应用时遇到的具体问题或者三级指标是否仍需进行细化和补充等问题,可能还需要进一步的研究探索解决。同时专家咨询法也存在一定的专家意见主观性和方法本身具有一定的不足之处,可以进一步结合其他的研究方法确定各个评价指标,完善指标。

真实世界研究近年来得到了快速发展,特别是在电子健康记录系统的普及和完善、大数据分析能力的提升等方面取得了显著进展。随着跨机构和跨地区的数据共享机制的建立,真实世界研究将更加深入和广泛。本研究基于分析文献和政策指南,在了解当前 RWD 质量评价的现状和所面临的问题的基础上,明确了 RWD 管理或治理基本流程,再根据文献总结和专家咨询法构建一套关于 RWD 质量的评价指标,对 RWD 研究的开展具有一定的理论和实践意义。

利益冲突声明: 作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

- 1 U.S. FDA. Real-world evidence[EB/OL]. (2023-05-10) [2023-09-20]. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>.
- 2 国家药监局药审中心. 关于发布《用于产生真实世界证据的真实世界数据指导原则(试行)》的通告(2021年第27号)[EB/OL]. (2021-04-13) [2023-10-19]. <https://www.cde.org.cn/main/news/viewInfoCommon/2a1c437ed54e7b838a7e86f4ac21c539>.
- 3 孙鑫,谭婧,王雯,等. 建立真实世界数据与研究技术规范,促进中国真实世界证据的生产与使用[J]. 中国循证医学杂志, 2019, 19(7): 755-762. [Sun X, Tan J, Wang W, et al. Developing technical guidance for real-world data and studies to achieve better production and use of real-world evidence in China[J]. Chinese Journal of Evidence-Based Medicine, 2019, 19(7): 755-762.] <https://kns.cnki.net/kcms2/article/abstract?v=IU01-AKyIMB5ageovNFcG7GldfPMMCR12QD1DBq8vV20k6-jORe66EAdVrzMg0OeuH81NrWjTNP2pN379E8NlFm61drLnc41Z1aFC3ZZ59U7wNIUmH33BBdZ4ehh3Zd9Le9FyEpEzuGvJatzRGRztVBbSb86-EbjamZtaCzRalb2xj6jgy-mw=&uniplatform=NZKPT&language=CHS>.
- 4 国家药品监督管理局. 关于发布真实世界证据支持药物研发

- 与审评的指导原则（试行）的通告（2020 年第 1 号）[EB/OL]. (2020-01-03) [2023-12-05]. <https://www.nmpa.gov.cn/xxgk/ypgtg/ypgtg/ypgtg/20200107151901190.html>.
- 5 Knatterud GL, Rockhold FW, George SL, et al. Guidelines for quality assurance in multicenter trials: a position paper[J]. *Control Clin Trials*, 1998, 19(5): 477-493. DOI: [10.1016/s0197-2456\(98\)00033-6](https://doi.org/10.1016/s0197-2456(98)00033-6).
- 6 European Medicines Agency. High-quality data to empower data-driven medicines regulation in the European Union—Medicines Agency[EB/OL]. (2022-10-10) [2023-09-20]. <https://www.ema.europa.eu/en/news/high-quality-data-empower-data-driven-medicines-regulation-european-union>.
- 7 Shang N, Wend C, Hripesak G. A conceptual framework for evaluating data suitability for observational studies[J]. *J Am Med Inform Assoc*, 2018, 25(3): 248-258. DOI: [10.1093/jamia/ocx095](https://doi.org/10.1093/jamia/ocx095).
- 8 姚晨. 利用好真实世界数据生产高质量真实世界证据支持药械监管[J]. *中国食品药品监管*, 2020(2): 22-27. [Yao C. Use real world data to produce high-quality real world evidence to support drug device regulation[J]. *China Food & Drug Administration Magazine*, 2020, (2): 22-27.] DOI: [10.3969/j.issn.1673-5390.2020.02](https://doi.org/10.3969/j.issn.1673-5390.2020.02).
- 9 贾玉龙, 姚明宏, 刘艳梅, 等. 特许医疗政策下真实世界数据治理的模式探索[J]. *中国循证医学杂志*, 2021, 21(12): 1373-1380. [Jia YL, Yao MH, Liu YM, et al. Exploring the patterns of real-world data governance in the context of special healthcare policy[J]. *Chinese Journal of Evidence-Based Medicine*, 2021, 21(12): 1373-1380.] DOI: [10.7507/1672-2531.202108147](https://doi.org/10.7507/1672-2531.202108147).
- 10 谭婧, 程亮亮, 王雯, 等. 患者登记研究的策划与患者登记数据库构建: 基于观察性设计的真实世界研究[J]. *中国循证医学杂志*, 2018, 17(12): 1365-1372. [Tan J, Cheng LL, Wang W, et al. Real world study using observational designs: plan for patient registry study and development of registry database[J]. *Chinese Journal of Evidence-Based Medicine*, 2018, 17(12): 1365-1372.] https://kns.cnki.net/kcms2/article/abstract?v=IU01-AKyIMDw_SXDvj-SY7mauihSjVKP8jsvFdTmtkmfhFG7NMbzWVkuKFfDMkiA5Nadh0Hw6EQp9q6shy-v2kLAy1VNw_IEwtjbqjB2ijf5CsedKn0WTgcq0oJGSHMCVoj4KKezhUPrWiLbbF2pmhPsDMapTJPVoe7xRzpqF0uTGpjonG20EA=&uniplatform=NZKPT&language=CHS.
- 11 国家药品监督管理局. 关于发布临床试验的电子数据采集技术指导原则的通告（2016 年第 114 号）[EB/OL]. (2016-07-27) [2023-09-22]. <https://www.nmpa.gov.cn/xxgk/ypgtg/ypgtg/ypgtg/20160729184001958.html>.
- 12 国家药监局药审中心. 关于发布《药物临床试验数据递交指导原则（试行）》的通告（2020 年第 16 号）[EB/OL]. (2020-07-20) [2023-09-22]. <https://www.nmpa.gov.cn/xxgk/ypgtg/ypgtg/ypgtg/20200720171201514.html>.
- 13 刘艳梅, 任燕, 贾玉龙, 等. 真实世界数据体系构建的模式探索[J]. *中华流行病学杂志*, 2022, 43(3): 418-423. [Liu YM, Ren Y, Jia YL, et al. Exploration of construction mode of real-world data system[J]. *Chinese Journal of Epidemiology*, 2022, 43(3): 418-423.] DOI: [10.3760/ema.j.cn112338-20210824-00675](https://doi.org/10.3760/ema.j.cn112338-20210824-00675).
- 14 赵国桢, 闫世艳, 郭玉红, 等. 基于既有医疗数据构建研究型数据库的方法学探讨及实例解读（二）: 数据治理的方法[J]. *中国中医药信息杂志*, 2023, 30(9): 17-21. [Zhao GZ, Yan SY, Guo YH, et al. Methodological discussion and case interpretation of construction research database based on existing medical data (II): methods of data governance[J]. *Chinese Journal of Information on TCM*, 2023, 30(9): 17-21.] DOI: [10.19879/j.cnki.1005-5304.202210065](https://doi.org/10.19879/j.cnki.1005-5304.202210065).
- 15 李姮. 医疗大数据质量评价指标体系研究[D]. 武汉: 华中科技大学, 2020. DOI: [10.27157/d.cnki.ghzku.2020.003499](https://doi.org/10.27157/d.cnki.ghzku.2020.003499).
- 16 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research[J]. *J Am Med Inform Assoc*, 2013, 20(1): 144-151. DOI: [10.1136/amiajnl-2011-000681](https://doi.org/10.1136/amiajnl-2011-000681).
- 17 Naik S, Voong S, Bamford M, et al. Assessment of the nursing quality indicators for reporting and evaluation (NQuIRE) database using a data quality index[J]. *J Am Med Inform Assoc*, 2020, 27(5): 776-782. DOI: [10.1093/jamia/ocaa031](https://doi.org/10.1093/jamia/ocaa031).
- 18 Bian J, Lyu T, Loiacono, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data[J]. *J Am Med Inform Assoc*, 2020, 27(12): 1999-2010. DOI: [10.1093/jamia/ocaa245](https://doi.org/10.1093/jamia/ocaa245).
- 19 熊兴江. 医疗大数据质量评价指标体系构建研究[D]. 武汉: 华中科技大学, 2019. DOI: [10.27157/d.cnki.ghzku.2019.002068](https://doi.org/10.27157/d.cnki.ghzku.2019.002068).
- 20 Liaw S, Guo JGN, Ansari S, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review[J]. *J Am Med Inform Assoc*, 2021, 28(7): 1591-1599. DOI: [10.1093/jamia/ocaa340](https://doi.org/10.1093/jamia/ocaa340).
- 21 Wang SV, Schneeweiss S, Berger ML, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0[J]. *Value Health*, 2017, 20(8): 1009-1022. DOI: [10.1016/j.jval.2017.08.3018](https://doi.org/10.1016/j.jval.2017.08.3018).
- 22 Cook JA, Collins GS. The rise of big clinical databases[J]. *Br J Surg*, 2015, 102(2): e93-e101. DOI: [10.1002/bjs.9723](https://doi.org/10.1002/bjs.9723).
- 23 Duke-Margolis Center for Health Policy. Characterizing RWD quality and relevancy for regulatory purposes[EB/OL]. (2018-10-01) [2023-11-10]. <https://healthpolicy.duke.edu/publications/characterizing-rwd-quality-and-relevancy-regulatory-purposes-0>.
- 24 Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research[J]. *Int J Med Inform*, 2016, 90: 40-47. DOI: [10.1016/j.ijmedinf.2016.03.006](https://doi.org/10.1016/j.ijmedinf.2016.03.006).
- 25 Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations[J]. *Commun ACM*, 1996, 39(11): 86-95. DOI: [10.1145/240455.240479](https://doi.org/10.1145/240455.240479).
- 26 World Health Organization. Data quality assurance (DQA)[EB/OL]. (2023-02-14) [2023-12-05]. <https://www.who.int/data/data-collection-tools/health-service-data/data-quality-assurance-dqa>.
- 27 Kahn MG, Raebel MA, Glanz JM, et al. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research[J]. *Med Care*, 2012, 50

- Suppl(0): S21–S29. DOI: [10.1097/MLR.0b013e318257dd67](https://doi.org/10.1097/MLR.0b013e318257dd67).
- 28 Black N, Payne M. Directory of clinical databases: improving and promoting their use[J]. *Qual Saf Health Care*, 2003, 12(5): 348–352. DOI: [10.1136/qhc.12.5.348](https://doi.org/10.1136/qhc.12.5.348).
- 29 Verma AA, Pasricha SV, Jung HY, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the general medicine inpatient initiative (GEMINI) experience[J]. *J Am Med Inform Assoc*, 2021, 28(3): 578–587. DOI: [10.1093/jamia/ocaa225](https://doi.org/10.1093/jamia/ocaa225).
- 30 Buchheit RB, Garrett JH, Mcneil S, et al. Automated procedure to assess civil infrastructure data quality: method and validation[J]. *J Infrastruct Syst*, 2005, 11(3): 180–189. [https://doi.org/10.1061/\(ASCE\)1076-0342\(2005\)11:3\(180\)](https://doi.org/10.1061/(ASCE)1076-0342(2005)11:3(180)).
- 31 Evans P. Scaling and assessment of data quality[J]. *Acta Crystallogr D Biol Crystallogr*, 2006, 62(Pt 1): 72–82. DOI: [10.1107/S0907444905036693](https://doi.org/10.1107/S0907444905036693).
- 32 Blacketer C, Defalco FJ, Ryan PB, et al. Increasing trust in real-world evidence through evaluation of observational data quality[J]. *J Am Med Inform Assoc*, 2021, 28(10): 2251–2257. DOI: [10.1093/jamia/ocab132](https://doi.org/10.1093/jamia/ocab132).
- 33 Rajan NS, Gouripeddi R, Mo P, et al. Towards a content agnostic computable knowledge repository for data quality assessment[J]. *Comput Methods Programs Biomed*, 2019, 177: 193–201. DOI: [10.1016/j.cmpb.2019.05.017](https://doi.org/10.1016/j.cmpb.2019.05.017).
- 34 王雯, 刘艳梅, 谭婧, 等. 回顾性数据库研究的概念、策划与研究数据库构建 [J]. *中国循证医学杂志*, 2018, 18(2): 230–237. [Wang W, Liu YM, Tan J, et al. Concept and planning of retrospective database study and development of research database[J]. *Chinese Journal of Evidence-Based Medicine*, 2018, 18(2): 230–237.] https://kns.cnki.net/kcms2/article/abstract?v=IU01-AKyIMBtkQzmSIstN2XKSUGIZj2PiwWaVjqZR0_We0Jf1ZQbLDXcglndexLoz1i0LoD6U7iKgsQnzfgHRTMs7kqZqWg_Jq6v8njV0hYjQ0-AfQDVklvmst_larST5lWf6UMfQ91FwOFyHW3xOvvW5dLmC7UMU1arfRktNEBO39HryLDxFg==&uniplatform=NZKPT&language=CHS.
- 35 Carl VW, Shane E, Peter CA. Administrative database code accuracy did not vary notably with changes in disease prevalence[J]. *J Clin Epidemiol*, 2016, 79: 86–89. DOI: [10.1016/j.jclinepi.2016.05.009](https://doi.org/10.1016/j.jclinepi.2016.05.009).

收稿日期: 2024 年 11 月 05 日 修回日期: 2026 年 01 月 08 日

本文编辑: 沈静怡 杨 燕