·医学信息研究·

医疗大数据应用于真实世界研究现状及展望*

刘 爽 冯 时 郭 昊 卢瑗瑗

（中国医学科学院北京协和医院 北京 100730） （神州数码医疗科技股份有限公司 北京 100020） （空军军医大学西京消化病医院 国家消化系统疾病临床医学研究中心 西安 710032）

弓孟春 吴开春

（中国医学科学院罕见病研究中心 北京 100730） （空军军医大学西京消化病医院 国家消化系统疾病临床医学研究中心 西安 710032）

〔摘要〕 基于文献调研，分析医疗大数据应用于真实世界研究的开展情况，阐述大数据时代真实世界研究优势，包括外部真实性高、目标人群广泛、证据整体性强、获取证据高效等方面，提出医疗大数据应用于真实世界研究在基础架构和具体开展两方面面临的挑战。

〔关键词〕 大数据；医疗；真实世界研究

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.03.004

Status and Prospect of Medical Big Data Applied to Real World Study *LIU Shuang, FENG Shi, Chinese Academy of Medical Sciences, Peking Union Medical College Hospital, Beijing 100730, China; GUO Hao, Digital China Health Technologies Co., Ltd., Beijing 100020, China; LU Yuanyuan, Air Force Medical University, Xijing Hospital, National Clinical Research Center for Digestive Diseases, Xi'an 710032, China; GONG Mengchun, Center of Rare Disease, Chinese Academy of Medical Sciences, Beijing 100730, China; WU Kaichun, Air Force Medical University, Xijing Hospital, National Clinical Research Center for Digestive Diseases, Xi'an 710032, China*

〔**Abstract**〕 The paper analyzes the development of medical big data applied to real world study based on literature research, expounds on the advantages of real world study in the era of big data, including high external authenticity, broad target population, strong evidence integrity, efficient evidence acquisition, etc., proposes the challenges of applying medical big data to real world study in terms of infrastructure and concrete development.

〔**Keywords**〕 big data; medical; real world study

〔修回日期〕 2019－10－11

〔作者简介〕 刘爽，博士研究生，发表论文10余篇；通讯作者：吴开春，主任医师，教授，博士生导师，发表论文300余篇。

〔基金项目〕 国家自然科学基金资助项目“消化系肿瘤发生发展的关键分子事件及其临床意义”（项目编号：81421003），“消化道肿瘤”（项目编号：81822031）；国家重点研发计划“规范化大型胃癌队列的建立及其可用性研究”（项目编号：2017YFC0908300）。

1 引言

大数据是指因体量庞大、结构复杂而难以通过传统方式分析及处理的数据[1]，具有数量庞大、数据流高速及类型丰富 3 个核心特征[2]。医疗领域涉及的大数据主要是临床、医学影像以及包括基因组、转录组、蛋白组、微生物组、暴露组等在内的多组学数据，其广泛应用是实现医学模式转变的必要前提和核心动力[3]。近年来生物医学数据总量日渐庞大、结构趋于复杂，如何有效利用这些医疗大数据成为重要的机遇和挑战。真实世界证据（Real - World Evidence，RWE）是指通过分析多种来源的现实医疗数据而获得的证据，数据来源包括电子健康档案（Electronic Health Records，EHR）、账单、移动设备收集的健康信息等[4]。真实世界研究（Real - World Study，RWS）强调研究数据的获取环境，在研究方法和实验设计上与传统方式并无本质区别。与随机临床试验（Randomized Clinical Trials，RCT）相比，RWS 来自真实临床情景，具有证据外推性好、可用数据量大、研究易于开展的优势，是 RCT 的重要补充。基于海量真实数据，RWS 可能帮助研究者发现临床实践与现有证据之间的差距，开展人群干预研究，实现改善整体治疗及预后的目的[5]。近年来大数据在医学领域的应用逐渐深入，为真实世界研究的开展提供支持。本文旨在对当前医疗大数据在真实世界研究方面的应用现状及前景进行综述。

2 医疗大数据应用于真实世界研究的进展

在 PubMed 数据库中对医疗大数据相关文献进行检索，不限定检索年限，共检索出有价值文献 672 篇。检索策略为：（million [Title/Abstract]）AND（（（（（（（（（big data [Title/Abstract]）OR health record [Title/Abstract]）OR medical record [Title/Abstract]）OR personal health record [Title/Abstract]）OR electronic medical record [Title/Abstract]）OR personal medical record [Title/Abstract]）OR digital medical record [Title/Abstract]）OR digital health record [Title/Abstract]）OR real world evidence [Title/Abstract]）OR real world data [Title/Abstract]）。通过阅读相关领域综述补充文献 20 篇，在此基础上筛选出利用医疗大数据进行的真实世界研究。文献筛选流程，见图 1。医疗大数据支持真实世界研究的文章发表数目近年明显增加，见图 2。研究方向，见表 1。

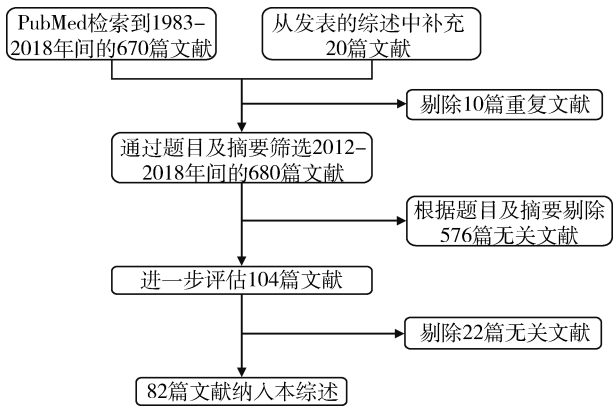

图 1 文献筛选流程

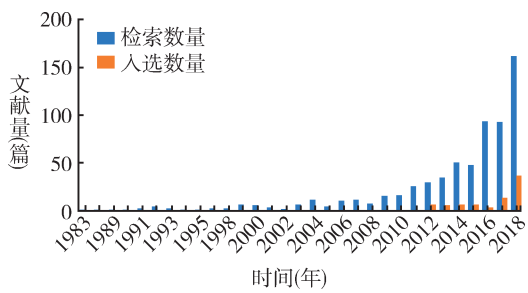

图 2 1983 –2018 年间医疗大数据（百万级）相关文献发表情况

表 1 1983 –2018 年间医疗大数据（百万级）相关真实世界研究用途汇总

研究用途	研究数量（项）
增进对疾病/状态的认识（病因、共病、药物应用情况等）	27
疾病分布（发病率、患病率等）	20
药物安全监视	10
疾病识别	9
卫生经济学	8
人群分层	4
验证既往研究成果（RCT 证据、数据库可用性等）	2
医院管理	1
疾病预测	1
治疗证据	1

3 大数据时代真实世界研究优势

3.1 外部真实性高

RCT 研究是建立因果关联、评估治疗手段安全性及有效性的金标准，通过随机、对照、盲法原则构建临床试验的理想场景，有效控制可能干扰结果的混杂因素，从而实现结论的高度内部有效性[6]。然而真实临床场景与 RCT 差异较大，RCT 的结论只适用于特定人群及环境，其外推性受到很大限制，难以指导真实世界中更为复杂的医疗决策场景。与 RCT 相比，来源于临床实践的医疗大数据为研究者提供多维度、高通量的数据支持，有助于产生外部真实性高的临床证据。

3.2 目标人群广泛

RCT 研究常通过严格的纳排标准保证研究人群的同质性，且在人群选择方面受到预期效果、组织管理、伦理等诸多方面的限制，往往会排除年龄过小或过大、病情较重、合并其他临床病症的患者[7]。这部分患者长期处于临床研究的“盲区”，缺乏有证据支持的治疗方案以及接受新型治疗试验的机会。以肿瘤学研究为例，一般状况差、既往有恶性肿瘤病史、合并器官衰竭、脑转移的患者往往会因预期效果差而被传统 RCT 拒绝[6]，RWS 可以将传统研究未能纳入的患者纳入观察范围，更为全面地评估特定药物/治疗方法的有效性和安全性，从而实现拓展适应症、指导临床决策的目的[8]。

3.3 证据整体性强

真实世界临床情景的复杂性导致研究证据与临床应用之间往往存在差距。RCT 可以证明某一药物或治疗在特定情境下的有效性，但在临床实践中仍有诸多问题存在争议。美国食品药品管理局（Food and Drug Administration，FDA）在衡量新药疗效及安全性时提出证据整体性这一概念，强调任何一项研究证据都并非孤立存在，而是建立在其他知识的基础之上[9]。RCT 是药物/治疗上市的前提，而 RWS 则可作为 RCT 的有效补充，在更广泛的人群范围、更多样的临床情景和更长的时间维度上评估其有效性和安全性，优化证据的整体性。

3.4 获取证据高效

RWS 除强调数据来源于医疗机构、家庭、社区等真实医疗环境外，在研究方法和试验设计方面并无规定[4]，可以根据数据类型和研究目的设计适宜的研究方案。现有的 RWS 主要基于 EHR 或注册数据库，多采用回顾性研究设计，无需像 RCT 一样投入大量人力、物力、财力及时间。对于临床试验而言，真实世界证据有助于确定可实现预期检验效能的最小样本量，节约临床试验开展时间，提高证据获取效率。以 TRANSFORMS、FREEDOMS 和 FREEDOMS Ⅱ 3 项研究为例，将 RWS 证据纳入分析模型后得出的样本量，比只引入 RCT 证据得出的样本量减少 40% 以上，可节省至少 6 个月的研究时长[10]。此外真实世界数据还可作为单臂研究的外部对照，有助于决策者更好地解读已有临床研究[6]。

4 医疗大数据应用于真实世界研究的挑战(表2)

表2 医疗大数据应用于真实世界研究的主要挑战

类别	过程	具体挑战
基础架构	信息化部署	部署基础设施：电子病历系统、生物样本库
		构建临床研究网络
		将新型数据持续整合到现有信息部署体系
	医学信息学技术	数据规范采集：语义标准化、信息互操作性、文本信息提取
		数据融合分析：数据调用、数据整合
		信息安全和隐私保护
研究开展	研究设计	提出具有切实临床意义的研究问题
		选取相关性好、体量足够大的数据集
	研究实施	保证数据质量，减少研究偏倚
		实现不同研究单位间的数据共享
	研究发表	结果模糊报道、选择性发表

4.1 基础架构

4.1.1 信息化部署 医疗大数据的处理包括采集、标准化、存储、调用、融合分析等多个维度[11]，合理部署信息化基础设施是应用医疗大数据

的前提。高度结构化的电子病历系统和严格质控下的生物样本库作为基础设施，是临床表型数据和多组学数据的基础。在此基础上多维度原始数据通过机器学习等方法进行融合分析，由数据集转化为新知识，从而为临床及基础研究提供源源不断的高质量数据。另一方面，由于单个中心逐渐难以满足RWS 对数据体量的要求，构建临床研究网络近年来应用日益广泛。然而识别适宜的合作单位、构建高效运转的数据存储及共享体系较为复杂。此外随着新生物标志物的发现和患者自报结局（Patient - Reported Outcomes，PRO）的提出，新的数据类型不断产生，需要持续整合到现有的信息部署体系[12]。

4.1.2 大数据处理对医学信息学技术提出更高要求 单纯的数据意义有限，关键是要开发一套标准的提取、转化、加载架构，使医疗大数据能够进行整合分析[13]。首先，数据规范采集是后续共享、集成的基础，涉及语义标准化（基于医学本体系统）、信息互操作性（基于信息交换标准）、文本信息提取（基于自然语言处理）、多组学数据整合分析等多方面的技术要求。其次，对多维度数据进行融合分析是进一步利用的关键，需要在数据调用（基于搜索引擎和跨库检索）和整合（基于机器学习、多组学分析等）技术方面实现突破。此外信息安全和隐私保护也是医疗大数据应用中的重要问题，涉及去识别化处理和存储安全两个维度，要求在个体隐私保护和数据价值挖掘之间实现平衡[14]。

4.1.3 医疗大数据尤其是基因组学数据应用对伦理学提出挑战[15] RWS 相关伦理审查原则和规范仍在发展之中，但项目开展之前应接受独立的伦理审查，将此作为整体研究质量的重要指标[16]。RWS 开展过程中面临的主要伦理问题包括：如何处理临床实践与科学研究之间的关系，评估患者参加研究的风险与获益，以恰当的方式获取患者的知情同意，保证数据获取与传输过程中的信息安全等[17]。其中知情同意是多维度数据提取及分析的基础，是RWS 规范开展的关键，需根据具体研究设计在恰当时机使患者充分知情大数据研究的潜在伦理风险。

4.2 研究开展

4.2.1 研究设计 合理开展研究是充分利用医疗大数据的关键。RWS 开展过程中需要尽可能控制真实世界中的各种偏倚，从而提高研究结论的内部有效性。研究设计阶段的关键在于提出具有切实临床意义的研究问题，选择相关性好且体量足够大的数据集进行验证[12]。然而真实世界数据来源于预先设计的数据采集系统，未必与待研究问题直接相关，且数据完整度及准确性难以预先核实，这为 RWS 的研究设计带来困难。

4.2.2 研究实施 该阶段的主要挑战是真实临床情景中存在复杂多样且难以控制的偏倚，包括选择、信息、实施、测量、失访偏倚等[18]。如何保证研究数据质量，尽量减少偏倚，是 RWS 实施过程中的一大挑战。此外如何实现不同单位间的数据共享也是研究实施过程中的切实问题。

4.2.3 研究发表 主要问题在于结果的模糊报道和选择性发表。据报道仅有少数上市后研究得以发表[19]，而大量临床问题未经真实世界数据检验或是未能公之于世。未来可能需要实行更为严格的研究报告制度，保证研究设计、数据分析和结果解读过程的透明度[20]。学术界有必要对 RWS 使用的数据及方法进行深入研究，从而发现现有研究开展过程中的潜在缺陷，更好地指导研究者利用真实世界数据[5]。

5 结语

随着大数据在医学领域应用逐渐深入，类型丰富、用途广泛的真实世界研究进入快速发展阶段。与随机临床对照试验相比，真实世界研究具有外部真实性高、目标人群广泛、证据整体性强、获取证据高效的优势，是传统研究形式的有效补充。现阶段基于医疗大数据的真实世界研究在基础架构和具体开展方面仍面临诸多挑战，需要在制度建设和技术手段两个层面进一步寻求突破。

参考文献

1 Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare [EB/OL]. [2018 - 05 - 10]. https://doi.org/10.1515/jib - 2017 - 0030.

2 Austin C, Kusumoto F. The Application of Big Data in Edi-

cine: current implications and future directions [J]. Journal of Interventional Cardiac Electrophysiology, 2016, 47 (1): 51 - 59.
3 弓孟春, 陆亮. 医学大数据研究进展及应用前景 [J]. 医学信息学杂志, 2016, 37 (2): 9 - 15.
4 Sherman RE, Anderson SA, Dal Pan GJ, et al. Real - World Evidence - what is it and what can it tell us? [J]. N Engl J Med, 2016, 375 (23): 2293 - 2297.
5 Booth CM, Karim S, Mackillop WJ. Real - World Data: towards achieving the achievable in cancer care [J]. Nature Reviews Clinical Oncology, 2019, 16 (5): 312 - 325.
6 Khozin S, Blumenthal GM, Pazdur R. Real - World Data for Clinical Evidence Generation in Oncology [J]. J Natl Cancer Inst, 2017, 109 (11): djx197.
7 Grapow MT, Von Wattenwyl R, Guller U, et al. Randomized Controlled Trials Do Not Reflect Reality: real - world analyses are critical for treatment guidelines![J]. The Journal of Thoracic and Cardiovascular Surgery, 2006, 132 (1): 5 - 7.
8 Beaver JA, Ison G, Pazdur R. Reevaluating Eligibility Criteria - Balancing Patient Protection and Participation in Oncology Trials [J]. N Engl J Med, 2017, 376 (16): 1504 - 1505.
9 Sherman RE, Davies KM, Robb MA, et al. Accelerating Development of Scientific Evidence for Medical Products within the Existing US Regulatory Framework [J]. Nat Rev Drug Discov, 2017, 16 (5): 297 - 298.
10 Martina R, Jenkins D, Bujkiewicz S, et al. The Inclusion of Real World Evidence in Clinical Development Planning [J]. Trials, 2018, 19 (1): 468.
11 Toga AW, Foster I, Kesselman C, et al. Big Biomedical Data as the Key Resource for Discovery Science [J]. Journal of the American Medical Informatics Association (JAMIA), 2015, 22 (6): 1126 - 1131.
12 Maissenhaelter BE, Woolmore AL, Schlag PM. Real - World Evidence Research Based on Big Data: motivation - challenges - success factors [J]. Der Onkologe: Organ der Deutschen Krebsgesellschaft eV, 2018, 24 (Suppl 2): 91 - 98.
13 Denney MJ, Long DM, Armistead MG, et al. Validating the Extract, Transform, Load Process Used to Populate a Large Clinical Research Database [J]. International Journal of Medical Informatics, 2016 (94): 271 - 274.
14 Xia W, Heatherly R, Ding X, et al. R - U Policy Frontiers for Health Data De - identification [J]. Journal of the American Medical Informatics Association (JAMIA), 2015, 22 (5): 1029 - 1041.
15 Fiore RN, Goodman KW. Precision Medicine Ethics: selected issues and developments in next - generation sequencing, clinical oncology, and ethics [J]. Current Opinion in Oncology, 2016, 28 (1): 83 - 87.
16 李洪, 魏来, 郭晓蕙, 等. 真实世界研究伦理审查初探 [J]. 中国循证医学杂志, 2018, 18 (11): 1198 - 1202.
17 Wang S, Liu B, Xiong N, et al. Discussion of Solutions to Ethical Issues in Real - World Study [J]. Frontiers of Medicine, 2014, 8 (3): 316 - 320.
18 Visvanathan K, Levit LA, Raghavan D, et al. Untapped Potential of Observational Research to Inform Clinical Decision Making: American Society of Clinical Oncology research statement [J]. Journal of Clinical Oncology, 2017, 35 (16): 1845 - 1854.
19 Spelsberg A, Prugger C, Doshi P, et al. Contribution of Industry Funded Post - marketing Studies to Drug Safety: survey of notifications submitted to regulatory agencies [J]. BMJ (Clinical Research ed), 2017 (356): j337.
20 Berger ML, Sox H, Willke RJ, et al. Good Practices for Real - World Data Studies of Treatment and/or Comparative Effectiveness: recommendations from the joint ISPOR - ISPE Special Task Force on real - world evidence in health care decision making [J]. Pharmacoepidemiol Drug Saf, 2017, 26 (9): 1033 - 1039.