应用医疗大数据分析提升临床研究可行性及效力

冯 时 刘 爽 朱 翀 郭 昊 弓孟春

（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） （中国医学科学院罕见病研究中心 北京 100730）

〔摘要〕 介绍大数据、真实世界证据、真实世界数据定义，阐述大数据分析提升临床研究的可行性，指出大数据分析具有优化临床研究招募、缩短临床研究周期、调整临床研究设计、构建复杂疾病治疗模型等效力并分析其局限性。

〔关键词〕 大数据；真实世界数据；真实世界证据；临床研究

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2019.12.003

Improving the Feasibility and Effect of Clinical Study Applying Medical Big Data Analysis

FENG Shi, LIU Shuang, Peking Union Medical College Hospital, Beijing 100730, China; ZHU Chong, GUO Hao, Digital China Health Technologies Co., Ltd., Beijing 100020, China; GONG Mengchun, Rare Diseases Research Center, CAMS, Beijing 100730, China

〔**Abstract**〕 The paper introduces the definition of big data, Real World Evidence (RWE) and Real World Data (RWD), expounds the feasibility of improving clinical study through big data analysis, points out that big data analysis has effects of optimizing clinical study recruiting, shortening clinical study period, adjusting clinical study design, building complex disease treatment model, etc., and analyzes its limitation.

〔**Keywords**〕 big data; Real World Data (RWE); Real World Evidence (RWE); clinical study

1 引言

大数据是指数据量庞大、数据结构复杂且依靠传统的方法和工具难于处理的数据集[1]。医疗领域大数据的核心特征是数据量大、种类丰富、传输速度快。对数据可靠性、医疗环节、计算技术、信息提取、数据共享等均提出挑战[2]。真实世界证据（Real World Evidence，RWE）是指在现实医疗环境中，通过分析真实世界数据（Real World Data，RWD），获知医学相关产品的用途、优点、缺点的临床证据。RWD 来源于电子健康档案、保险说明、患者登记、电子健康移动设备及应用等，与传统临床试验数据的本质区别在于数据获取的环境，即真实世界研究的数据来源于医疗机构、家庭和社区，而非存在诸多严格限制的科研场所。依据实用性临床试验原则收集的 RWD 可用于随机实验设计，将临床研究的范围拓展到进行医疗干预的时刻。RWD 分析极易发生选择偏移、信息偏移、测量错误等多种偏倚，因此数据质量难以保证。近年来越来越多

〔收稿日期〕 2019－03－20

〔作者简介〕 冯时，博士，发表论文 4 篇；通讯作者：弓孟春，博士，发表论文 20 余篇。

的大数据项目从关注数据数量转向关注数据质量[3]。通过特定技术手段，从庞杂的数据中提炼分析得到证据，是当下临床研究的新方向。

RWE 的发展现已成为各国医疗领域的前沿和热点，数据的开放与运用已成为国家综合竞争力的新标志。英国临床实践研究数据链（Clinical Practice Research Datalink，CPRD）自 1987 年起收集基础医疗记录，以此为基础已发表药物安全研究、临床指南超过 1 700 篇[4]。2009 年美国启动卫生经济和临床医学信息技术（HITECH）行动鼓励医师和医院使用电子病历系统，有力推动电子健康档案的普及，成为 RWD 的重要组成。2016 年 12 月美国颁布《21 世纪治疗法案》，要求美国食品药品管理局（Food and Drug Administration，FDA）在医疗产品审批和监管程序中纳入 RWE。此后 FDA 陆续发表声明、颁布指南阐述 RWE 的定义和特点，规范 RWE 的产生和应用，将充分发挥 RWE 在审批监管决策中的作用视作其首要战略重点。2018 年 12 月 6 日 FDA 颁布《真实世界证据方案框架》，为实现 RWE 支持药品审批决策的目标提供相对清晰的路线图。

2 大数据分析提升临床研究可行性

2.1 概述

由于前瞻性随机对照实验（Randomized Controlled Trials，RCT）在很多医疗和政策支持中的局限性，大数据的重要性越来越得到认同。而 RWD 揭示数据中真实世界的本质，为医疗决策和政策提供更加准确和有效的证据[5]。

2.2 大数据临床实验优越性

RCT 是当前最主要的临床研究方法。然而一旦在真实世界中对生物效应进行评估，患者可能受到并发症及其余合并用药的影响，生物效应未必等同于临床效应。另外分析 RCT 数据主要是寻找基线因素和特定临床结果之间的关系，但是很多临床试验中的患者会接受多种治疗，最终分析时需要将这些因素都纳入考量[6]。因此大数据临床实验（Big Data Clinical Trials，BCT's）的概念进入人们视野。大数据临床试验由两方面组成，一方面是收集独立个体的所有数据，另一方面是收集多个个体来代表真实世界。这里的个体未必单指患者，因为健康人群也是 BCT 研究的范围。在 BCT 背景下慢性病的治疗模式将会迎来变革。此外在流行病学方面，谷歌搭建的流感预测模型对于流感爆发的预测甚至比美国疾病控制与预防中心（Centers for Disease Control and Prevention，CDC）更加快速、精准[7]。大数据时代，BCT 研究将成为临床研究的主力军，实现对 RCT 结果及其相关大数据的客观分析，使得分析结果更加科学、准确、有效[9-10]。

2.3 RCT 与 RWE 研究共同推动临床研究

RCT 研究的最大缺点在于其结果外在效应较低。为提高内在效应，RCT 研究往往需要依据假设创造理想的实验条件，缩小实验对象的纳入范围。即使在 RCT 研究中获得正面结果，也难以真正发展形成具有普适意义的治疗方法。此外，RCT 研究往往低估药物的长期毒性，对于长期、生活质量相关参数并不敏感，研究时长、资源要求较高。而 RWE 研究可以提出问题、筛选所需的数据来源评估其优劣。可以应用合适的分析工具，在保证真实有效地的前提下寻找证据，同时保障内在和外在效应，结果更加具有普适性[8]。值得注意的是，与 RCT 相比 RWE 研究的内在效应相对较低，同时也较难以实现随机分组。因此在现阶段 RWE 研究仍不能代替 RCT，对 RWE 研究仍应采取审慎严谨的态度[10]。RWE 和 RCT 研究具有极强的互补性。RWE 研究可以帮助制定方向，为未来 RCT 研究提供假设或作为未来验证性 RCT 的基础；也可以作为 RCT 研究的后续，对于在 RCT 研究中呈现阳性结果的治疗方法，RWE 研究可以探究其长期的安全性和有效性。

2.4 以药物为中心实现大数据整合

如何收集来源可靠的数据库、电子健康档案、社交媒体中的信息，从中提取临床信息和分子数据，是一个亟待解决的问题。目前以药物为中心进

行数据整合成为一种理想的解决方法。药物通过影响特定蛋白或者通路，在起到治疗效果的同时也会导致不良反应。如果将众多患者的临床信息、药物基因组信息和不良反应信息整合就可以发现临床表型和分子信息之间的关联[11]。例如，利用 FDA 的不良事件报告系统（Adverse Event Reporting System，FAERS）数据库，研究者发现一旦阻断 β－肾上腺素通路，卵巢癌患者的死亡率随之下降，为这一临床现象的分子机制提出新思路。进一步的研究发现 Src 蛋白磷酸化通过调控 β－肾上腺素/PKA 来调控下游分子网络，促进肿瘤转移、侵袭和生长[12]。由此可见以药物作为连接纽带，大数据分析可以将临床表型和分子因素关联起来。

2.5 大数据分析拓展临床研究方向

在大数据背景下，借助回溯性分析 RWE 研究可以挖掘现有临床信息，增强对于疾病自然进程的认知，从而拓展临床研究方向。在疾病病因方面，美国帕金森病进展标志物倡议（Parkinson's Progression Markers Initiative，PPMI）利用大数据分析手段探究其致病风险因素并尝试做出预测性诊断和分类[13]。利用大数据分析方法可以挖掘已有数据库中的信息，获得新的认知。如分析癌症基因组图谱（The Cancer Genome Atlas，TCGA）中高级别浆液性卵巢癌的数据，发现遗传学改变多集中于抑癌基因失活，识别出包括 RAS/PI3K、RB、FOXM1、NOTCH 在内的数个潜在治疗靶点[14]。子宫内膜癌 TCGA 数据分析则进一步揭示疾病的分子生物学本质，可以根据预后将疾病重新分类为 POLE 超突变肿瘤、微卫星不稳定高突变负荷肿瘤、低拷贝数肿瘤和高拷贝数肿瘤[15]。在疾病治疗及预后方面，研究发现疾病治疗应答和预后并不由单基因决定，而是包括基因突变、拷贝数变异、DNA 甲基化、mRNA、蛋白及其修饰、肿瘤微环境影响等共同作用的复杂网络，因此需要对临床－分子相关的多个数据组进行分析。现有电子健康档案数据多掌握在公共卫生实体或保险公司，这些数据未得到充分利用来研发新药。REW 可以帮助寻找新药治疗靶点、验证药物安全性[16]。

3 大数据分析提升临床研究效力

3.1 概述

大数据提供实时结构化学习的机会，这将进一步推动临床实践改革。随着临床研究中组学技术应用的增加，对患者信息和已有数据进行回溯性和实时分析可以帮助做出临床决策，制定相关政策。目前大数据分析已经被用于探究真实世界，这对于临床研究设计有极大帮助[17]。

3.2 优化临床研究招募

在招募患者加入临床研究环节，可以根据预测的参数对参与者进行初步筛选。如对于临床药物试验，通过大数据分析筛选出治疗应答可能性更高的患者，或者排除应答可能性较低的患者，也可排除用药后发生不良反应风险相对更高的患者，尤其是在探究靶向治疗的临床疗效时具备极大的推广价值[18]。此外实用性临床试验（Pragmatic Clinical Trials，PCTs）的概念也进入大众视野。依据 EHRs 开展回溯性 RWE 研究对于参与者的纳入标准较为宽松，可以极大提高临床试验患者的应答率。在肿瘤药物的研发中，传统临床试验的参与者比例小于 5%，尤其是少数民族群体、老年人群、低收入人群、居住在偏僻地区的人群。而 PCTs 可以在遵循现有方法学、伦理、法律等准则的前提下发挥社区医疗的作用，让更多人参与到实验中来，对临床决策提供有力支持[19]。

3.3 缩短临床研究周期

传统的关于治疗和干预效果的临床研究所需随访周期较长，因此越来越多的研究者开始寻找新的分子标志物来替代传统分子标志物。利用蛋白质、代谢产物、表观遗传标志物等分子标志物替代传统标志物，进行小型、随访周期短的临床试验，相较于传统临床试验更加方便快捷。但是由于这些替代性分子标志物未经过足够的临床检验，临床结果和分子标记之间也没有搭建正确的关联，这一类临床研究可能难以得出有意义的研究结果。在过去几十

年发现的新型分子标志物中难以识别、确证的分子标记数量甚至超过成功验证的分子标记数量，因此应用这些替代性分子标记应注意时刻保持谨慎[20]。大数据分析为这一问题的解决提供新的思路，随着经验积累和自正，通过严格的模型推论和对实验结果的审查可以增加研究结果的可信度，从而筛选出真正有效的分子标志物[21]。

3.4 调整临床研究设计

在临床研究中可以依据收集到的患者队列信息，遵循预先设定的原则对实验设计进行进一步优化，包括但不限于调整样本数量、放弃某种治疗或剂量、改变接受治疗患者的比例、因为效果良好/不佳提前终止实验等[22]。如肺癌整合标记靶向治疗（Biomarker – integrated Approaches of Targeted Therapy for Lung Cancer Elimination，BATTLE）试验，运用适应性贝叶斯设计方法，实时留取患者生物标本，监测患者多种分子标志物水平变化，选择适当的治疗方法，以求某种或数种分子标志物水平能够反映该治疗方法的疗效[23]。体外肺灌注肺移植实验 DEVELOP – UK 调整研究纳入的参与者数量，允许由于安全、有效、无效等原因提前终止研究[24]。这种适应性的实验设计方法有诸多优点，包括减少患者接受相对无效治疗的时间，加快进度寻找有效的治疗手段，更加有效且符合伦理。

3.5 构建复杂疾病 – 药物治疗模型

真实世界与理论世界最大的不同在于临床患者往往不会只患有一种疾病，而是存在并发症，需要同时接受多种药物治疗，这些合并用药具有治疗效果的同时也会影响患者其他正常生理功能，从而降低其生活质量。利用大数据分析可以准确辨别药物的适应症和不良反应，评估药物安全性。将斯坦福临床数据库（Stanford Clinical Data Warehouse，STRIDE）中超过 100 万患者的药物、疾病、疾病 – 药物的频率分布结合起来，可以辨别药物与特定疾病的关系中哪些是适应症、哪些是不良反应[25]。随着对药效生物分子机制的理解加深，在电子医疗档案等大数据的帮助下，可以构建模型，模拟存在并发症情况下的人体机能，观察合并用药是否会改变抗癌药物的药效。这种建模方法可以使临床研究更加接近真实世界[16]。如果合并用药会影响药物的抗癌效果，那么就将其列为临床试验中需要排除的药物；如果在系统预测中发现患者的并发症可能会受到实验新药影响而进一步恶化，可以在实验招募环节将有此类并发症的患者排除在外。此外具体的用药剂量可根据患者的个体情况进行调整[26]。

4 大数据分析的局限性

大数据分析面临的核心问题已经不是数据量，而是多维度数据整合的方法，如机器学习、深度学习、网络分析等。而这些方法都通过黑盒子来探究多因素之间的联系，因此采用相同数据集、不同方法进行分析可能得到完全不同的分析结果，从而导致其无法真正上升到真实世界证据所需的高度。如对比 MammaPrint[27] 和 OncotypeDX[28] 两种算法对乳腺癌患者预后分析结果发现没有相同基因[29]。即使针对性地改进数据分析工具，由于变量数量极大、方法学复杂，依然极有可产生数据噪声、分析偏性、假阳性等问题[17]。因此首先需要在临床研究中寻找有力证据，证明分析结果，才能逐步推广应用于临床。另外 RWD 的重要组成电子健康档案是为计费和医疗设计的，因此往往难以从繁杂、未结构化的数据中将真正与临床相关的信息筛选出来[19]。庞杂的数据量可能使研究者们忽视研究设计的重要性。然而只有经过审慎的考虑、仔细研究设计才能充分挖掘已有的临床数据，完成高质量的临床研究，保障研究结果具有良好的可重复性和临床应用价值。同时数据的共享和开放需要公共政策倾斜，也需要研究者们的共同努力，确保可以获取完整的数据[16]。存储和分析大量的患者数据对计算能力提出较高要求，同时需要多个领域专家的参与[30]。此外由于大数据粒度及信息量大，为防止患者信息被再识别，数据安全需要格外注意[17]。

5 结语

医学研究突破的潜在方向，包括疾病的精确分

型、发病机制、预测预警、快速诊断和精准治疗等目前都与医疗大数据密切相关。而医疗大数据是真实世界证据的基石。真实世界证据为生物学家、临床医生、流行病学家及医疗卫生政策制定专家提供有效支持，使得数据驱动的决策制定成为可能并最终实现对疾病治疗、健康监测的优化，对患者产生有益影响。

参考文献

1 Toga AW, Foster I, Kesselman C, et al. Big Biomedical Data as the Key Resource for Discovery Science [J]. J Am Med Inform Assoc, 2015, 22 (6): 1126 - 1131.

2 Baro E, Degoul S, Beuscart R, et al. Toward a Literature - driven Definition of Big Data in Healthcare [J]. Biomed Res Int, 2015 (2015): 639021.

3 Rickards A, Cunningham D. From Quantity to Quality: the central cardiac audit database project [J]. Heart, 1999, 82 (Suppl 2): II18 - 22.

4 Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: clinical practice research datalink (CPRD) [J]. Int J Epidemiol, 2015, 44 (3): 827 - 836.

5 Basu A, Axelsen K, Grabowski DC, et al. Real - world Data: policy issues regarding their access and use [J]. Med Care, 2016, 54 (12): 1038 - 1044.

6 Hudis CA. Big data: are large prospective randomized trials obsolete in the future? [J]. Breast, 2015, 24 (Suppl 2): S15 - 18.

7 Wang SD, Shen Y. Redefining Big - data Clinical Trial (BCT) [J]. Ann Transl Med, 2014, 2 (10): 96.

8 De Lusignan S, Crawford L, Munro N. Creating and Using Real - world Evidence to Answer Questions About Clinical Effectiveness [J]. J Innov Health Inform, 2015, 22 (3): 368 - 373.

9 Wang SD. Opportunities and Challenges of Clinical Research in the Big - data Era: from RCT to BCT [J]. J Thorac Dis, 2013, 5 (6): 721 - 723.

10 Maissenhaelter BE, Woolmore AL, Schlag PM. Real - world Evidence Research Based on Big Data: Motivation - challenges - success factors [J]. Onkologe (Berl), 2018, 24 (Suppl 2): 91 - 98.

11 Rodriguez - Esteban R. A Drug - Centric View of Drug Development: how drugs spread from disease to disease [J]. PLoS Comput Biol, 2016, 12 (4): e1004852.

12 Armaiz - Pena GN, Allen JK, Cruz A, et al. Src Activation by β - adrenoreceptors is a Key Switch for Tumour Metastasis [J]. Nat Commun, 2013 (4): 1403.

13 Dinov ID, Heavner B, Tang M, et al. Predictive Big Data Analytics: a study of parkinson's disease using large, complex, heterogeneous, incongruent, multi - source and incomplete observations [J]. PLoS One, 2016, 11 (8): e0157077.

14 Cancer Genome Atlas Research N. Integrated Genomic Analyses of Ovarian Carcinoma [J]. Nature, 2011, 474 (7353): 609 - 615.

15 Cancer Genome Atlas Research N, Kandoth C, Schultz N, et al. Integrated Genomic Characterization of Endometrial Carcinoma [J]. Nature, 2013, 497 (7447): 67 - 73.

16 Singh G, Schulthess D, Hughes N, et al. Real World Big Data for Clinical Research and Drug Development [J]. Drug Discov Today, 2018, 23 (3): 652 - 660.

17 Hernandez I, Zhang Y. Using Predictive Analytics and Big Data to Optimize Pharmaceutical Outcomes [J]. Am J Health Syst Pharm, 2017, 74 (18): 1494 - 1500.

18 Taglang G, Jackson DB. Use of " Big Data" in Drug Discovery and Clinical Trials [J]. Gynecol Oncol, 2016, 141 (1): 17 - 23.

19 Khozin S, Blumenthal GM, Pazdur R. Real - world Data for Clinical Evidence Generation in Oncology [J]. J Natl Cancer Inst, 2017, 109 (11): djx187.

20 Yetley EA, DeMets DL, Harlan WR. Surrogate Disease Markers as Substitutes for Chronic Disease Outcomes in Studies of Diet and Chronic Disease Relations [J]. Am J Clin Nutr, 2017, 106 (5): 1175 - 1189.

21 Parast L, Cai T, Tian L. Evaluating Surrogate Marker Information Using Censored Data [J]. Stat Med, 2017, 36 (11): 1767 - 1782.

22 Pallmann P, Bedding AW, Choodari - Oskooei B, et al. Adaptive Designs in Clinical Trials: why use them, and how to run and report them [J]. BMC Med, 2018, 16 (1): 29.

23 Liu S, Lee JJ. An Overview of the Design and Conduct of the BATTLE Trials [J]. Chin Clin Oncol, 2015, 4 (3): 33.

24 Fisher A, Andreasson A, Chrysos A, et al. An Observational Study of Donor Ex Vivo Lung Perfusion in UK lung transplantation: develop - UK [J]. Health Technol Assess, 2016, 20 (85): 1 - 276.

25 Liu Y, Lependu P, Iyer S, et al. Using Temporal Patterns in Medical Records to Discern Adverse Drug Events from Indications [J]. AMIA Jt Summits Transl Sci Proc, 2012 (2012): 47 - 56.

（下转第 22 页）

知识相关概念、属性及联系的抽取。以本体表达陈述性知识，以语义规则描述过程性知识，实现对临床决策知识的精确化、标准化、合理化表达。

4.1.3 实现信息交互标准化与规范化 为实现以临床数据中心为核心的医院资源库之间信息交互标准化与规范化，参考 HL7 等国内外通行标准，制定具有医院特色的术语字典及交互标准。

4.2 不足与完善方法

4.2.1 诸多知识库难以有效分类导致查询效率低 针对该问题，结合临床应用特点，在 InfoButton 管理器中建立合理的多维度且统一的知识库分类标准，应用于知识查找、展示排序、优先级设置等方面的管理，得到较好的优化和完善。

4.2.2 在不同应用环境中相同知识元素导致知识搜索效果不尽相同 在 InfoButton 管理器中逐步完善搜索数据的提取、转化优化功能，同时参照一体化医学语言系统（Unified Modeling Language System，UMLS），在全院逐步建立一套完整的术语服务技术体系，以解决临床信息数据多样性对数据交互带来的难题。

4.2.3 难以兼顾不同用户个性化知识需求 临床医师、护士、药师、科研工作者因为操作习惯、知识库偏好、应用需求不尽相同，统一的知识展示对于各式各类的临床应用场景并不适宜。为解决此问题，建立用户个性化知识库管理模型，为不同用户群体设置一系列知识库管理参数，以调整并满足该用户群体的知识需求，同时收集用户偏好信息，优化偏好模型，提供个人用户级的知识库偏好设置管理功能。

5 结语

InfoButton 模型在临床智能决策中的应用实践是医院从信息化迈向智能化发展道路中的积极探索和有效尝试。未来随着模型应用的逐步深入和不断优化，将拓展医院智慧医疗、管理和服务，利用人工智能、大数据等技术整合现有信息资源，实现诊疗流程的精细化、智能化、规范化，其具有广阔的发展前景以及良好的经济和社会效益。

参考文献

1 王雪梅，刘莉，李敬东，等. 国内外知识按钮的应用研究［J］. 中国数字医学，2017，11（8）：82－85.

2 李艳，吴梦佳，张士靖，等. 语义互操作标准在临床决策支持系统中的应用［J］. 医学信息学杂志，2017，38（10）：57－61.

3 王剑，王忠民. 基于信息集成平台的 Infobutton 信息模型建设方案设计与实现［J］. 中国数字医学，2017，12（9）：56－58.

4 雷健波，王飞，胡建平. 无线一键通：一种基于决策现场的移动临床决策支持方案的研究［J］. 中国卫生信息管理杂志，2011，8（3）：61－64.

5 何毅，王曙光，刘文浩. InfoButton 在国家人口与健康科学数据共享平台的应用研究［J］. 中国数字医学，2016，11（1）：80－83.

6 Guilherme Del Fiol, Peter J Haug, James, J Cimino, et al. Effectiveness of Topic－specific InfoButtons: a randomized controlled trial［J］. J Am Med Inform Assoc, 2008（15）: 752－759.

7 葛小玲，孙利，薛颜，等. 基于 CDR 的临床决策支持系统设计及应用初探［J］. 中国数字医学，2014，9（8）：2－4，34.

（上接第 17 页）

26 Collins B. Big Data and Health Economics: strengths, weaknesses, opportunities and threats［J］. Pharmacoeconomics, 2016, 34（2）: 101－106.

27 Slodkowska EA, Ross JS. MammaPrint 70－gene Signature: another milestone in personalized medical care for breast cancer patients［J］. Expert Rev Mol Diagn, 2009, 9（5）: 417－422.

28 Malo TL, Lipkus I, Wilson T, et al. Treatment Choices Based on OncotypeDx in the Breast Oncology Care Setting［J］. J Cancer Epidemiol, 2012（2012）: 941495.

29 Fan C, Oh DS, Wessels L, et al. Concordance among Gene－expression－based Predictors for Breast Cancer［J］. N Engl J Med, 2006, 355（6）: 560－569.

30 Mayo CS, Matuszak MM, Schipper MJ, et al. Big Data in Designing Clinical Trials: opportunities and challenges［J］. Front Oncol, 2017（7）: 187.