

DOI: 10.13288/j.11-2166/r.2021.04.007

真实世界数据的真实性及其影响因素探讨

赵云霞¹, 邵明义², 陈晓琦², 符宇², 赵瑞霞², 李蔓¹, 王靖雯¹, 张容容¹, 邵范雷¹

1. 河南中医药大学, 河南省郑州市金水区金水东路 156 号, 450046; 2. 河南中医药大学第一附属医院

摘要 真实世界数据是产生真实世界证据的源头, 其真实性是对数据资料最基本的要求。真实世界数据具有来源广泛、数据量大、内容丰富、种类繁多、分散性强、异质性高等特点。从数据来源和数据统计分析两方面分析可能导致真实世界数据偏差的原因, 提出从病历质量、患者引导、研究人员培训、优化系统、规范数据分析等方面提高真实世界数据的真实性。

关键词 真实世界数据; 数据真实性; 真实世界研究

医疗大数据与互联网时代的到来不仅促进了真实世界研究 (real world study, RWS) 的发展, 同时也为其提供了新的思路与方法。随机对照试验 (RCT) 因其严格的试验要求及高昂的成本, 已无法全面满足现阶段快速发展的医学要求^[1], 与此同时, RWS 成为 RCT 的重要补充而备受关注^[2]。现阶段中国 RWS 广泛应用于多学科的研究, 在解决重要科学问题等方面具有巨大意义^[3-4]。真实世界数据 (real world data, RWD) 作为 RWS 进行的基本要素之一, 成为探讨 RWS 的重点。然而, 近年来 RWS 并未解决 RWD 真实性的问题, 造成了 RWS 结果的可信度不高, 推广受阻^[5]。本团队前期基于真实世界电子病历数据开展的队列研究, 深刻认识到 RWD 并非绝对真实, 而是相对真实的, 存在着“不真实”的现象。对实践中遇到的难点进行分析发现, 这些不真实的情况可以通过对数据本身的质控以及对数据处理的严格分析得以改善。基于此, 本文就引起 RWD 真实性差异的原因进行描述与总结, 并提出相应的解决策略, 以期加强同行对数据真实性的关注, 为提高数据真实性、增加研究科学性提供理论支持, 同时推进 RWS 向更精准方向发展, 使 RWS 更好地弥补 RCT 的不足。

1 RWD 与 RWS

RWD 是指与患者健康状况和/或医疗服务相关

基金项目: 国家自然科学基金 (81574101); 河南省中医药科学研究专项 (2019ZYD03); 河南省重点研发与推广专项 (科技攻关) (192102310157)

✉ 通讯作者: shmy1016@163.com; 28050827@qq.com

的, 常规采集的多来源数据^[6]。RWD 是产生真实世界证据 (real world evidence, RWE) 的源头, RWE 是临床和医疗决策、医疗器械临床评价的重要参考。随着医疗大数据时代的到来, 国内外 RWS 方兴未艾, 如孙鑫等^[7]提出建立 RWD 与研究技术规范, 以促进中国 RWD 的产生与使用; 谭婧等^[8]提出患者登记数据库构建的技术规范; Meyer 等^[9]通过 1 项 RCT 结合 RWS 证明了新辅助化学疗法在卵巢癌治疗中的应用效果; 刘园等^[10]基于 RWS 得出皮肤瘙痒患者的临床用药规律; 国际监管机构也不断探索利用 RWD 支持监管决策^[11]。由此可见, RWD 种类繁多, 在医学技术、临床用药等领域均应用广泛。

目前, RWS 在我国仍处于发展阶段, 存在诸多挑战与不足, 如将 RWD 等同于 RWE, 认为 RWS 只能用于观察性研究等^[12]。同时, RWS 本身也存在来自数据本身和研究方法的设计方面的局限^[13]。在 RWS 中, 对 RWD 的质疑最为关键。作为产生 RWE 的源头, RWD 的真实性是前提要求, 因此保证数据的真实性成为 RWS 的重点。

2 RWD 真实性的影响因素

RWD 的来源主要包括电子病历 (EMR), 电子健康档案、医保数据、出生死亡登记、公共健康监测数据等^[14]。RWD 具有来源广泛、数据量大、内容丰富、种类繁多、分散性强、异质性高等特点^[15], 如果不加以控制, 会引起对数据真实性的质疑。RWS 结果外部可推性强, 但内部真实性不高, 对其探讨要从数据来源和数据统计分析入手。

2.1 数据本身产生真实性偏差的原因

2.1.1 人为因素 课题组在基于真实世界中临床数据的原发性肝癌循证证据构建研究中,对影响 RWD 真实性的人为因素进行分析,发现其主要表现在医护人员、患者、研究人员 3 个方面。

1) 医护人员方面。医护人员作为接触患者的主要人群,对数据的产生起着至关重要的作用。在疾病诊疗过程中,由于医护人员而产生数据真实性偏差的原因有:(1) 受既往病例的影响,对复查患者的诊疗信息数据记录未能准确详实;(2) 受技术水平的限制,医生采集到的数据与患者实际情况不符;(3) 数据采集者视、触、叩、听,或望、闻、问、切的不规范操作造成获得不真实的医疗数据;(4) 由于各种原因造成的采集样本送检不及时或样本不合格,造成具体数据的偏差等。

2) 患者方面。患者作为数据产生的最根本来源,产生真实性偏差的原因有:(1) 对既往病史的回忆出现偏倚,夸大或隐瞒病史;(2) 依从性差,对于医生给出的诊疗方案并未真实执行;(3) 受试者受试状态不同,造成获取的数据并未反映患者的真实状态等。

3) 研究人员方面。研究人员作为研究数据采集的主力,在数据采集过程中产生真实性偏差的原因有:(1) 开始进行研究之前未明确标准;(2) 未进行药物临床试验质量管理规范培训,对数据的采集流程、方法、要求不明确等,造成数据采集不规范;(3) 在不同数据库导入过程中存在不规范、不合要求的现象,亦会出现数据的不真实。

2.1.2 系统因素 主要包括电子病历、电子健康档案、医保数据、出生死亡登记、公共健康监测数据 5 个方面。

1) 电子病历。电子病历作为临床研究最常见的数据来源,记载了大量患者的基础信息数据、诊疗数据、医嘱数据、理化检查数据等^[3]。其产生真实性偏差的原因有:(1) 药物剂量固定化,与病人真实服用的剂量不符;(2) 门诊处方的实名认证程度不高,出现药不对人的现象;(3) 出院带药时未进行随访造成患者具体服药天数不明确等。

2) 电子健康档案。电子健康档案可以降低患者护理成本,一方录入,多方使用,各种记录具有标准化和数字化的特点^[4]。其产生真实性偏差的原因有:(1) 涉及患者隐私导致患者隐瞒病史,避重就轻;(2) 各地区体检设备参差不齐,数据结果不准确;(3) 储存档案时出现数据丢失、错

乱,导致最终档案质量不高;(4) 慢性病、老年病患者日常体检的数据不能及时归档,导致数据丢失等^[7]。

3) 医保数据。随着居民收入的提高和健康意识的增强,医保政策得到很大普及,同时也带来了一些问题:(1) 有些住院期间的用药并非诊疗过程的必需用药,不能真实体现患者病情变化;(2) 住院天数的限制造成医保数据不能真实反映患者病情变化;(3) 因转诊等问题造成患者以急诊入院时的急诊病例书写不真实;(4) 转院后其他医院医保数据丢失;(5) 病历作假、用药作假、虚开票据等方式来套取医保基金等^[8-20]。

4) 出生死亡登记。出生死亡登记是医学、人口学、社会学研究最重要的基础信息,其产生真实性偏差的原因有:(1) 人口流动性大,数据收集困难,导致申报、注销不及时或信息未及时整合,出现漏报信息情况,造成生存期等数据失真;(2) 患者死因不明,由家属填写造成死因判断的失真;(3) 系统不完善造成人员培训不到位导致数据可信度低等^[21-23]。

5) 公共健康监测数据。公共健康监测数据具有大、杂、疑等特点,其产生真实性偏差的原因有混杂因素多、垃圾数据的存在等^[23]。

2.1.3 经济因素 不同的经济条件,引起医疗服务水平差异而带来真实性偏差的原因有:(1) 不同的筹资水平对应不同的待遇保障标准;(2) 不同级别医疗机构用药限制(如抗生素等)造成临床用药的不真实等。

2.1.4 客观因素 外界客观因素包括在不同医院的检验、检查结果标准值的差异、不同厂家的药械检测结果不同致使数据采集过程中出现真实性差异。中国现存着现代医学与传统医学两种医学体系,中医学作为中国传统医学的重要组成部分,所产生的中医数据对患者的诊疗具有重要意义。但中医学术语具有复杂性,规范化程度不高,导致对于同一情况的界定上具有明显差异;同时,中医术语量化程度低,标准的量化数字化程度不高,在中医术语的应用时常产生客观差异,导致所产生的中医数据存在不可避免的不真实的现象^[24-25]。

2.2 数据统计分析过程中产生真实性偏差的原因

在对采集到的数据进行统计分析的过程中,仍存在着影响真实性的因素。如数据分析方法是否合理,是否将风险控制在可接受的范围等,都从方法层面影响着数据的真实性;同时,混杂因素存在于

真实世界研究的整个过程,是无法完全避免和控制的,从实际操作层面影响着数据的真实性。对于中医临床研究干预方法多样,临床数据混杂因素众多的情况,一般的统计分析方法不太适用,无法有效控制偏倚和混杂,最终导致其真实性产生偏差^[24]。

3 提高 RWD 真实性的可行性方法

3.1 提升诊疗水平,提高病历质量

医生的诊疗水平直接影响到所收集数据的真实性。医生要本着以患者为中心的诊疗态度,不断提高个人业务水平,完善相关知识储备,更新先进的诊疗方法,认真负责,提高诊疗水平^[25]。对病历书写者而言,在进入临床科室之前进行病历书写培训,明确病历书写的要素,掌握病历书写要点;定期组织开展病历书写大赛或培训等活动提高医护人员对病历质量的重视。

3.2 对患者进行正向引导

为求真实了解患者病情以给予恰当的诊疗措施,要求患者不隐瞒、不欺骗,如实陈述病情病史,对患者进行正向引导,明确利弊,积极配合治疗。针对数据采集过程中存在的回忆偏倚,可通过研究者灵活运用调查询问技巧,积极引导患者回忆;同时通过改进研究设计等方法,最大限度地降低回忆偏倚^[26]。

3.3 对研究人员进行规范化培训

在研究开始之前,对所有参与人员进行规范化培训,制定数据采集标准与指南,明确研究目的,严格按照标准进行数据的采集与录入,必要时采取预实验。在研究过程中遇到的数据问题,严格与指南标准对照,确保数据的真实性。在数据录入或制定业务质量规则时检查有效性,并定期进行质量检查,审查数据变量^[27]。

3.4 升级优化系统

与医院信息科建立适合科研的临床系统,实现系统数据结构化,并大范围推广,实现各地患者诊疗信息数据共享;在各个系统中增加备注选项,在患者需要与系统存在冲突时,记录患者真实运用的诊疗措施;增设诊疗数据追踪选项,详细记录患者诊疗数据的产生过程,包括时间、地点、方法、设备、结果、标准值等。将微信、手机应用软件等移动技术整合进入数据管理系统中,极大地提升效率,动态开放式设计,允许研究过程中不断增加新数据点,要求拍照留底,适应临床发展需求。

3.5 严格执行国家医保政策

执行国家医保政策是在确保患者得到个性化诊

疗的同时,限制患者的不必要诊疗;严格按照急诊住院要求,及时准确书写急诊病历。

3.6 提高数据结构化程度

运用医学词典实现自然语言向结构化语言的转换^[28],如针对中医真实世界研究,刘保延等研发“中医医疗与临床科研信息共享系统”以实现中医诊疗信息的数据化^[29]。利用人工智能技术对诊疗数据进行深入挖掘,开发出具有提取自然语言语义的智能系统,自动识别剔除极端值,结构化数据,实现数据的整合应用^[30]。

3.7 规范进行数据分析处理

数据采集完成之后,在对数据进行统计分析的过程中,需从统计方法的选择、混杂因素的控制以及缺失数据的处理等方面进行控制。不同的研究问题需要不同的统计分析方法,如根据不同治疗模式,可选择描述性分析、差异性比较、相关性分析等统计方法,根据不同影响因素可选择单因素、多因素线性回归、Logistic 回归、Cox 回归等统计方法。针对 RWD 来源广泛、数据量大、内容丰富、种类繁多、分散性强、异质性强等特点,可运用倾向性评分、多因素分析及工具变量法进行混杂因素的控制。查明数据缺失机制与缺失比例,多选择均值填补或多重填补处理缺失数据,更准确地进行结果解读和评价。

4 小结

RWD 的真实与否是 RWS 成败的关键。RWD 在 RWS 过程中的真实性决定了最终产生 RWE 等级的高低^[31]。在 RWD 演变成 RWE 的过程中,由于数据本身的性质特点与在此过程中的研究方法以及数据的采集、分析等过程存在着来自于各个方面的干扰因素等原因导致其具有真实性不高的情况。整体情况下,RWD 因被赋予法律效应而拥有相对真实性,但存在着因主观、客观等因素而表现出的在现有体制条件下的不真实,其本质理应是相对真实与绝对真实的动态平衡。在 RWS 中,可以通过严谨的研究方法设计以及控制数据本身的真实性,在主观因素和客观因素上共同控制,严格遵循 RWS 的规范规则,提高 RWD 的真实性,得到高等级的 RWE。

前瞻性的 RWS 可以在一定程度上弥补回顾性数据的不足,将相关的影响因素纳入统计范围,根据这些因素去做持续的追踪研究、分析判断,避免回顾性数据采集过程中的真实性问题,但由于条件

限制过多,操作比较困难,使用并不普遍。医疗大数据的冲击下将 RWD 的不真实部分变成小概率事件,但仍存在关键数据的真实性问题影响最终研究结论。人工智能的运用也为 RWD 的统计分析提供更可靠的方法途径,但对未知的情况无法准确分析。因此,在未来的 RWS 中,迫切需要灵活运用多学科交叉的方法,充分发挥人工智能与大数据分析等方法技术的优势,尽可能提高 RWD 的真实性,提高 RWS 的内部真实性,推进 RWS 的进一步发展。

参考文献

- [1] 张俊华,孙鑫,李幼平,等.循证中医药学的现在和未来[J].中国循证医学杂志,2019,19(5):515-520.
- [2] 庞博,郑文科,张俊华,等.中医药真实世界研究的应用场景[J].世界中医药,2019,14(12):3115-3118,3122.
- [3] 王佳,杨柱,龙奉玺,等.真实世界研究在恶性肿瘤研究中的应用[J].中国医药导报,2019,16(34):46-49.
- [4] 张冬,张俊华,孙凤,等.真实世界研究与中医药大数据[J].世界中医药,2019,14(12):3119-3122.
- [5] 金鑫瑶,郑文科,张俊华,等.推进真实世界研究的透明化[J].世界中医药,2019,14(12):3106-3110.
- [6] 杨松,马龙腾,张菁菁.中国临床医学真实世界研究施行规范[J].解放军医学杂志,2018,43(1):1-6.
- [7] 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.
- [8] 谭婧,彭晓霞,舒啸尘,等.患者登记数据库构建技术规范[J].中国循证医学杂志,2019,19(7):771-778.
- [9] MEYER LA, CRONIN AM, SUN CC, et al. Use and effectiveness of neoadjuvant chemotherapy for treatment of ovarian cancer[J]. J Clin Oncol, 2016, 34(32):3854-3863.
- [10] 刘园,袁海龙,徐俊,等.基于真实世界研究的皮肤瘙痒患者的临床用药规律分析[J].空军医学杂志,2019,35(5):451-454.
- [11] MAKADY A, HAM RT, ANTHONIUS B, et al. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies[J]. Value Health, 2017, 20(4):520-532.
- [12] 孙鑫,谭婧,唐立,等.重新认识真实世界研究[J].中国循证医学杂志,2017,17(2):126-130.
- [13] 吴一龙,陈晓媛,杨志敏,等.真实世界研究指南(2018年版)[M/OL]. (2018-08-03) [2019-12-20]. <https://max.book118.com/html/2018/1023/8001100005001130.shtm>.
- [14] 张强,邵明义,刘奕兵,等.基于真实世界中医临床数据构建原发性肝癌临床疗效评价方法探索[J].中医杂志,2019,60(18):1567-1571.
- [15] 杨薇,谢雁鸣,庄严.基于 HIS “真实世界”数据仓库探索上市后中成药安全性评价方法[J].中国中药杂志, 2011, 36(20):2779-2782.
- [16] ANDRIKOPOULOU E, SCOTT P, HERRERA H, et al. What are the important design features of personal health records to improve medication adherence for patients with long-term conditions? A systematic literature review[J]. BMJ open, 2019, 9(9):e028628.
- [17] 高斌,马海燕.居民健康档案管理存在的问题与对策[J].中国卫生事业管理,2010,27(4):281-283.
- [18] WANG Z. Data integration of electronic medical record under administrative decentralization of medical insurance and healthcare in China: a case study[J]. Isr J Health Policy Res, 2019, 8(1):24-26.
- [19] 仇雨临,吴伟.城乡医疗保险制度整合发展:现状、问题与展望[J].东岳论丛,2016,37(10):30-36.
- [20] 余舒晴.完善我国医保制度的对策探析[J].劳动保障世界,2019,(27):41.
- [21] 马文军.完善死亡登记工作的几点建议[J].华南预防医学,2011,37(6):1-2.
- [22] 伍晓玲,张学高,赵飞,等.我国生命登记信息共享与应用实践[J].中国卫生信息管理杂志,2019,16(2):180-184.
- [23] 孟润堂,罗艺,宇传华,等.健康大数据在公共卫生领域中的应用与挑战[J].中国全科医学,2015,18(35):4388-4392.
- [24] 符宇,邵明义,燕树勋,等.真实世界研究与中医临床疗效评价[J].中医杂志,2019,60(7):546-550.
- [25] 王晓丽,万生芳,魏昭晖,等.真实世界大数据下中医临床研究特点探析[J].世界最新医学信息文摘,2019,19(6):244,246.
- [26] 赵瑞霞,邵明义,符宇,等.基于真实世界中医电子病历数据的原发性肝癌疗效评价难点及解决策略[J].中医杂志,2019,60(23):2009-2012.
- [27] 宋观礼,张润顺,刘保延,等.真实世界中医临床诊疗信息数据化实施与质量控制[J].中医杂志,2015,56(3):198-201.
- [28] STEVEN SC, 黄文湧.流行病学研究中的回忆偏倚[J].国外医学(社会医学分册),1991,8(3):124-125,131.
- [29] MAISSENHAELTER BE, WOOLMORE AL, SCHLAG PM. Real-world evidence research based on big data: Motivation-challenges-success factors [J]. Onkologie (Berl). 2018, 24(Suppl 2):91-98.
- [30] 刘保延.真实世界的中医临床科研范式[J].中医杂志,2013,54(6):451-455.
- [31] 金昌晓,计虹,席韩旭,等.大数据科研分析平台在临床医学研究中的应用探讨[J].中国数字医学,2019,14(2):37-39.
- [32] 张俊华,郑文科,张伯礼,等.真实世界研究[J].世界中医药,2019,14(12):3101-3105.

(下转第 311 页)

- tion study [1]. *Trials*, 2014, 15: 264. doi: 10.1186/1745-6215-15-264.
- [2] BROWNE RH. On the use of a pilot sample for sample size determination [J]. *Stat Med*, 1995, 14(17): 1933 – 1940.
- [3] KIESER M, WASSMER G. On the use of the upper confidence limit for variance from a pilot sample for sample size determination [J]. *Biom J*, 1996, 38(8): 941 – 949.
- [4] JULIOUS SA. Sample size of 12 per group rule of thumb for a pilot study [J]. *Pharm Stat*, 2005, 4(4): 287 – 291.
- [5] HERTZOG MA. Considerations in determining sample size for pilot studies [J]. *Res Nurs Health*, 2008, 31(2): 180 – 191.
- [6] SIM J, LEWIS M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency [J]. *J Clin Epidemiol*, 2012, 65(3): 301 – 308.
- [7] WHITEHEAD AL, JULIOUS SA, COOPER CL, et al. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable [J]. *Stat Methods Med Res*, 2016, 25(3): 1057 – 1073.

Determining Sample Size of Pilot Trials in Traditional Chinese Medicine

ZHANG Ying¹, WANG Junhui², HU Yeyin³, SHAO Baoyi⁴, FU Zhanju⁵, LIU Jianping¹

1. School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing, 100029; 2. Guang'anmen Hospital, China Academy of Chinese Medical Sciences; 3. Dongzhimen Hospital, Beijing University of Chinese Medicine; 4. Li Ka Shing Faculty of Medicine, The University of Hong Kong; 5. School of Economics and Management, Beijing Jiaotong University

ABSTRACT A pilot trial is an important prerequisite to ensure the success of high-quality clinical trials, and it can also provide references for future sample size estimate of larger full-scale trials. This study aimed to explain the purpose and significance of conducting a pilot trial in TCM. By fully considering the characteristics of TCM clinical trials, and taking randomized controlled trials (RCTs) as an example, this study discussed how to determine the sample size of pilot trials from the perspectives of feasibility, precision, minimization of total sample size and standardized effect size.

Keywords clinical research methods; sample size estimation; pilot trial; traditional Chinese medicine clinical study
(收稿日期: 2019-09-23; 修回日期: 2020-07-22)

编辑: 柴倩云

(上接第 306 页)

Validity of Real-World Data and the Risk Factors

ZHAO Yunxia¹, SHAO Mingyi², CHEN Xiaoqi², FU Yu², ZHAO Ruixia², LI Man¹, WANG Jingwen¹, ZHANG Rongrong¹, SHAO Fanlei¹

1. Henan University of Chinese Medicine, Zhengzhou, 450046; 2. The First Affiliated Hospital of Henan University of Chinese Medicine

ABSTRACT As the basis of conducting real-world studies, the validity of real-world data (RWD) should be highly valued. RWD was characterized by the universality of sources, the abundant of data, the richness of content, and the diversity of types, and was highly dispersed and heterogenized. The risk factors that might lead to bias of RWD were analyzed from the aspects of data source and data statistical analysis. To improve the RWD validity, the quality of medical records should be improved; patients' education and researchers' training are needed; the system should be perfected and the data analysis should be standardized.

Keywords real-world data; data validity; real-world studies

(收稿日期: 2019-12-22; 修回日期: 2020-05-07)

编辑: 柴倩云