·基础理论与方法·

基于大数据的随机对照试验

许璐　王胜锋　詹思延
北京大学公共卫生学院流行病与卫生统计学系 100191
通信作者：詹思延，Email：siyan-zhan@bjmu.edu.cn

【摘要】 我国医疗领域目前已积累了海量数据，如何利用大数据开展随机对照试验日益得到关注。本研究结合国外利用大数据实施随机对照试验的成功经验，从数据来源、研究对象和研究结局确定、干预措施、随机化方法、知情同意的实施等方面进行梳理总结，以期为国内未来开展相关研究提供借鉴。

【关键词】 大数据；电子医疗数据；随机对照试验
DOI：10.3760/cma.j.issn.0254-6450.2019.06.019

Randomized controlled trial based on big data

Xu Lu, Wang Shengfeng, Zhan Siyan
Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China
Corresponding author: Zhan Siyan, Email: siyan-zhan@bjmu.edu.cn

【**Abstract**】 A large amount of data has been accumulated in Chinese medical area. Problems as how to use big data to carry out randomized controlled trials have also been increasingly noteworthy. Through learning the successful experiences in conducting randomized controlled trials on big data from abroad, this article introduces the knowledge regarding sources of data, identification of research subjects and outcomes, interventions, methods of randomization and the implementation of informed consent, *etc*., all related to big data, hoping to shed light on studies of this kind, for the years to come in China.

【**Key words**】 Big data; Electronic medical records; Randomized controlled trial
DOI:10.3760/cma.j.issn.0254-6450.2019.06.019

一直以来，传统随机对照试验（randomized controlled trial，RCT）都被视为评价干预措施的金标准。但传统RCT一方面有成本高、周期长、实施难度大等局限性，另一方面较为严格的纳入排除标准导致了研究对象的人群代表性较差，研究结果的外推常常受到质疑。为了解决传统RCT结果外推性差的问题，实效性随机对照试验（pragmatic randomized controlled trial，PRCT）的概念被提出[1]。PRCT具有以下特征：①更关注于所研究的干预措施与结局之间的关联，而不注重因果关系的解释；②研究人群更接近于一般人群[2]；③PRCT中所采取的干预措施，一般为目前已经投入临床使用的治疗方法，而不是未知的，还未在患者中推广使用的治疗措施；④与传统RCT相比，PRCT对研究对象的纳入排除标准更加宽松。但由于PRCT大体上还是沿用了传统RCT的实施方法，所以PRCT的开展仍具有一定的难度。

随着大数据时代的到来，越来越多的学者开始考虑将传统流行病学研究设计（如：队列研究、病例对照研究、RCT等）与医学大数据相结合，从大数据分析中收集真实世界的证据，大数据随机对照试验（big data randomized controlled trial，BRCT）作为一个新兴名词应运而生。如Wang等[3-4]曾提出过大数据临床试验（big-data clinical trial，BCT）这一名词，并指出BCT可以利用全部目标人群开展研究，避免了以往RCT只能抽取少数样本开展研究所带来的样本代表性较差的问题。此外，对于慢性病，利用BCT也可以实现对研究结局（如血压）的实时观测，而无需像传统RCT那样开展阶段性的随访。尽管现阶段在国内外并不能找到一个相对精准的解释对其进行定义，但目前比较主流的观点是：利用大数据来完成RCT的某一或多个环节，如研究对象的确定与招募、研究结局的确定，制订符合数据本身特点的干预措施等。将大数据元素融入RCT的优势至少包括两点：一方面可以减少资金的投入、缩短研究周期。另一方面，研究结果更具有代表性，更能反映治

疗措施在实际应用中的效果。

就具体如何发挥大数据优势开展BRCT，本研究将从数据来源与质量问题、研究对象和研究结局的确定、干预措施、随机化和知情同意几个方面重点介绍。

一、数据来源与质量问题

BRCT的数据来源广泛，但各类医疗大数据都有其自身的特点，并且数据质量上也存在一些缺点，所以研究者在数据利用过程中应将这些因素纳入考虑。

1. 数据来源及特点：①BRCT所对应的大数据不仅包括医疗保险数据、电子病历数据、区域医疗数据，也包括网络[5]、社交媒体、手机应用等来源的具有高度多样化、高容量和高速等特征的数据[6]；②BRCT最常用的数据源为医疗保险数据，主要因为其时效性强，包含了患者主要诊断、就诊、处方费用信息等，且数据结构化、标准化做得较好[7]；③尽管PRCT也使用“真实世界”数据，但PRCT所采用的数据多是为了特定研究目的而收集的，导致研究结束后相关数据的可用性有限，而BRCT所用数据一般为政府、企业等满足行政管理或商业目的而采集的，不受限于某一特定的研究目的。

2. 数据质量问题：大数据来源众多，但影响利用的重点在于数据质量，常见质量问题：①信息准确性，如对抑郁、高血压等非“硬终点”的诊断可能不准确。②缺乏某些关键混杂变量信息，导致无法控制相关混杂。如医疗大数据一般缺少吸烟、饮酒、BMI等信息。③数据完整性差，如我们通常认为药品索赔记录可以很好地反映患者的实际药物使用情况，但由于多种原因，医保数据有时候无法捕获到患者的用药记录[8]。④对于电子病历数据还存在“异地就诊”时，就诊信息难以实现在不同医院之间互通对接的问题。正是由于以上问题，利用大数据时，如何结合多方数据，建立统一、标准的数据库就尤为重要。

3. 应用注意事项：基于以上探讨的数据质量问题，利用大数据开展正式研究之前，必须要进行数据真实性的验证工作[3]。国际上很多知名杂志（如*Lancet*、*JAMA*等）在审稿时都关注利用数据库的研究是否开展了数据验证工作，*Epidemiology*从2011年起就发文要求所有利用大数据开展疗效比较研究的来稿均需提供数据真实性的验证材料[9]。Cooper等[10]也提出了利用医保数据确定研究结局时，需要再结合电子病历等医疗记录开展验证工作，得到阳性预测值等真实性评价指标。数据真实性验证过程所使用的资源和所采取的策略往往依数据库的不同而不同。目前，这一工作可通过计算机检索和人工核查等步骤完成[11]，也可结合多个数据库交叉验证。

二、研究对象和研究结局的确定

BRCT通常直接利用数据库中的诊断信息来确定研究对象与研究结局，但近年来，不断开始有研究利用网络途径来招募研究对象，为BRCT研究对象的确定提供了新的思路。

1. 直接利用数据库中诊断信息：直接利用电子病历、医保数据等数据库中的诊断信息来确定研究对象和研究结局是大多数BRCT所采用的策略，其中医保数据应用最广。诊断信息中所含有的疾病名称、疾病编码以及肿瘤形态学编码等都可以辅佐对目标疾病的锁定。如在一项探讨医保报销比例对心肌梗死患者药物依从性影响的研究中[12]，研究者直接通过ICD-9-CM（International Classification of Diseases, 9th Revision, Clinical Modification）确定目标患者，同时研究者对该识别方法的准确性进行了验证，其阳性预测值、灵敏度和特异度分别高达97%、96%和99%。

直接利用数据库中诊断信息，不仅可以降低招募患者、确定研究结局的工作复杂性和成本，同时，可以增强研究人群的代表性，减轻纳入研究对象过程中可能出现的选择偏倚。但是，也要注意部分数据库中的诊断信息仍是非结构化文本格式，这就需要采取较为复杂的文本结构化、标准化处理才能使用。

2. 利用网络途径：随着社交网站和软件的普及，网络RCT的理念也应运而生。这类RCT直接利用网络等社交平台招募研究对象，如Vandelanotte等[5]在研究视频干预对提高运动量的作用时，就融入了在社交平台发布广告的形式招募研究对象。另外，在一项探讨教育研讨对全科医生处方行为影响的研究中[13]，研究者也采用电子邮件招募来自巴黎3个县符合条件的全科医生。

利用网络招募研究对象，一方面将招募范围拓广，研究对象覆盖范围扩大，另一方面，招募速度快，能在短时间内找到足够数量的研究对象。但是这种方式也很容易出现无应答偏倚，一定程度影响样本的代表性，同时也可能会出现虚假信息，常常需要研究者进一步核实。

三、干预措施

BRCT不仅可以研究传统RCT、PRCT常常关注的临床治疗措施的效果[14]，还可以发挥大数据自身特点所具有的优势，研究特定的干预措施，如利用不同医保模式开展药物疗效评价、研究不同保健计划

的效果等。

1. 将不同的医保模式替代药物层面的干预措施：此类研究最直接可行，如比较某种上市后药物与对照药物在真实世界的实际效果时，直接采用随机对照试验可能存在操作层面的困难，但是可以换个思路，通过调整干预组和对照组研究对象的医保报销范围[15]，如调整干预组患者的医保计划仅报销目标疾病所有治疗药物中的试验药物，对照组患者医保计划仅报销对照药物。通过政策调整推动各个组别的大部分患者自然而然地去尽可能使用所在组别的对应药物，从而形成理想的干预组和对照组。当然如此操作还可能会存在患者依然选用非医保报销药物的情形，从而产生类似传统RCT中的“不依从”情况。但是我们可以借助工具变量等减少相应的影响[16-17]。

2. 研究不同保健计划的效果：我们可以开展关于不同医保模式的政策研究[18]，教育培训对医师处方行为的影响[13]，以及服药提醒装置对患者服药依从性的影响[19]等保健计划效果的研究。如Rand等[20]利用医保系统中的电话号码和HPV疫苗索赔记录来确定研究对象和结局，从而评估短信提示对青少年人乳头瘤病毒疫苗接种率的影响。

四、随机化方面的优势

所有传统RCT可以采用的随机化方法均可应用于BRCT，包括简单随机化、整群随机化[21-22]、分层随机化[23]、区组随机化以及多种随机化方法的结合使用[24]。但是BRCT在简单随机化以及反应变量-适应性随机化上可以很好地发挥其样本量大的优势。

1. 简单随机化：这种方法的优点在于消除了治疗组之间的随机差异，且操作简单，能够有效降低效应估计的偏倚。而对传统RCT来说，常存在样本量不够大的问题，因此在这种情况下使用简单随机化，往往会达不到完全随机化的效果[25]。而BRCT则可以充分发挥其样本量大的优势，更利于运用简单随机化方法。以DCOR试验为例[26]，通过简单随机化的方法将研究对象分到两个对比组之中，从而有效控制了治疗组之间的随机差异。

2. 反应变量-适应性随机化[27-28]：这种随机化方法不要求一定要1∶1分配，随机化比例可随着不断收集到的信息而改变，从而使研究对象被分配到不太有利治疗组的比例会随着时间不断减小。反应变量-适应性随机化方法虽然早被提出，但由于其实施相对复杂，且对样本量有一定的要求，所以传统RCT很少使用这一方法，但BRCT就可以很好地利用它的优势。

五、知情同意

BRCT同样需要符合伦理规定，但在如何知情同意方面，由于大数据的融入具有了新的特色。传统RCT均离不开知情同意的要求，但是BRCT在某些情形下可以免除知情同意。

1. 免除知情同意的情境：多数BRCT所采取的措施不会对患者造成不良影响，并且有些情况下，实施个体水平知情同意并不可行，所以很多BRCT都被伦理委员会批准免除个体水平的知情同意。如：研究保健管理策略对医疗资源消费和利用度的影响[29]和研究服药提醒装置对患者服药依从性的影响时[19]，因为这些都是提高研究对象生活质量的措施，并且研究对象可以自由选择不接受干预，所以这两项研究就被批准免除知情同意。

2. 不免除知情同意的情境：当然也有一些RCT研究，即使整合了大数据元素，也必须实施知情同意。即便如此，在知情同意的环节，除了传统的纸质知情同意，依然可以考虑借助大数据开展知情同意。如采用网页等新型方式，获取知情同意，除了形式上更加新颖，操作上也更加便捷，拓宽了知情同意的渠道。如Gabriel等[30]在评估通过互联网提供免费艾滋病毒自我检测试剂盒是否会提高艾滋病诊断率（SELPHI试验）时，就采用社交网站和app应用程序进行相关宣传和获取知情同意。

六、机遇与挑战

现阶段BRCT既有研究周期短、资金投入少、新颖、便捷等优点，但也存在数据质量差、多数据库整合困难、数据需要清理等缺点（表1），但是我们必须意识到大量的医疗大数据对于RCT的开展是一个助力机遇。BRCT的提出使得平台RCT研究的开展有了更可行的途径。在平台RCT中，不再仅仅关注某一种特定的干预措施，而是比较针对某一疾病的多种干预措施，并随着试验的开展，根据各干预措施表现出的疗效情况，增加或减少所要研究的干预措施[31-32]。利用大数据做平台RCT可以实现同时比较目前有关某一疾病的所有治疗措施的效果，并及时调整研究方案，把精力集中在更需要进一步研究的治疗方法上。

当然，我们也必须重视目前开展此类工作面对的挑战。首先是对数据质量的挑战，提高数据质量是最重要也是最基础的问题。只有有了高质量的数据，才能开展其他方法学方面的探索。因此，要强化相关机构对数据收集过程中质量的把控，实现数据

表1 大数据随机对照试验的优缺点

优点	缺点
①所需资金投入少	①数据质量差
②研究周期相对较短，弥补了传统RCT知识转换过程滞后的问题	②混杂因素相对较多且难以控制。数据库内缺少如BMI、吸烟、饮酒等变量信息，导致难以充分调整可能的混杂因素
③数据库利用率高。同一数据库可用于多项研究，也可同时研究关于某一疾病的多个干预措施	③多数据库的整合利用尚待解决。很多数据库里的患者诊断、药物等编码分类不一致，这就导致很难将多个数据库整合使用
④研究对象代表性强，研究结果外推性高	④数据清理费用较高。虽然开展BRCT理论上可以节约成本，但国内目前的现状是，各种数据源的数据质量不佳，因此，数据清理工作就可能造成较高的费用支出

信息的结构化和标准化，做好数据真实性和可靠性的验证工作。其次是对多数据库整合的挑战。尽管目前国内已有多种来源的医疗大数据，但这些数据库资源分散，还未能做到数据信息的互联互通、资源共享。因此，抓紧制订统一规范，建立医学大数据统一管理机构，使得数据的采集和利用统一化和规范化。最后是对信息安全的挑战。在数据利用过程中不应该只关注数据本身，更要注意针对患者的隐私保护问题，保证患者的信息安全[33]。所以在数据使用过程中可以使用患者信息去识别等技术，保证患者信息安全。

综上所述，随着人工智能、机器学习等新兴技术的发展，利用大数据来完成传统RCT或者PRCT的某一环节，以减轻研究负担，增大样本量，更好地回答卫生政策、卫生经济以及临床治疗等多领域的问题已是大势所趋。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] Ford I, Norrie J. Pragmatic trials[J]. N Engl J Med, 2016, 375(5): 454–463. DOI: 10.1056/NEJMra1510059.

[2] Sedgwick P. Explanatory trials versus pragmatic trials[J]. BMJ, 2014, 349: g6694. DOI: 10.1136/bmj.g6694.

[3] 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. DOI: 10.3978/j.issn.2072–1439.2013.06.24.

[4] Wang SD, Shen YX. Redefining big-data clinical trial (BCT)[J]. Ann Transl Med, 2014, 2(10): 96. DOI: 10.3978/j.issn.2305–5839.2014.10.03.

[5] Vandelanotte C, Short C, Plotnikoff RC, et al. Taylor Active-Examining the effectiveness of web-based personally-tailored videos to increase physical activity: a randomised controlled trial protocol[J]. BMC Public Health, 2015, 15: 1020. DOI: 10.1186/s12889–015–2363–4.

[6] Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data[J]. Epidemiology, 2015, 26(3): 390–394. DOI: 10.1097/EDE.0000000000000274.

[7] 杨羽，詹思延. 上市后大数据药品安全主动监测模式研究的必要性和可行性[J]. 药物流行病学杂志，2016，25(7)：401–404.
Yang Y, Zhan SY. Analysis of necessity and feasibility in studies of post-marketing drug safety active surveillance based on big data[J]. Chin J Pharmacoepidemiol, 2016, 25(7): 401–404.

[8] Lauffenburger JC, Balasubramanian A, Farley JF, et al. Completeness of prescription information in US commercial claims databases[J]. Pharmacoepidemiol Drug Saf, 2013, 22(8): 899–906. DOI: 10.1002/pds.3458.

[9] Hernán M. With great data comes great responsibility: Publishing comparative effectiveness research in epidemiology [J]. Epidemiology, 2011, 22(3): 290–291. DOI: 10.1097/EDE.0b013e3182114039.

[10] Cooper WO, Hernandez-Diaz S, Gideon P, et al. Positive predictive value of computerized records for major congenital malformations[J]. Pharmacoepidemiol Drug Saf, 2008, 17(5): 455–460. DOI: 10.1002/pds.1534.

[11] García Rodríguez L, Ruigómez A. Case validation in research using large databases [J]. Br J Gen Pract, 2010, 60(572): 160–161. DOI: 10.3399/bjgp10X483472.

[12] Choudhry NK, Avorn J, Glynn RJ, et al. Full coverage for preventive medications after myocardial infarction[J]. N Engl J Med, 2011, 365: 2088–2097. DOI: 10.1056/NEJMsa1107913.

[13] Le Corvoisier P, Renard V, Roudot-Thoraval F, et al. Long-term effects of an educational seminar on antibiotic prescribing by GPs: a randomised controlled trial[J]. Br J Gen Pract, 2013, 63(612): e455–464. DOI: 10.3399/bjgp13X669176.

[14] Johnson L, Shapiro M, Mankoff J. Removing the mask of average treatment effects in chronic lyme disease research using big data and subgroup analysis[J]. Healthcare (Basel), 2018, 6(4): 124. DOI: 10.3390/healthcare6040124.

[15] Choudhry NK. Randomized, controlled trials in health insurance systems [J]. N Engl J Med, 2017, 377(10): 957–964. DOI: 10.1056/NEJMra1510058.

[16] Ertefaie A, Small DS, Flory JH, et al. A tutorial on the use of instrumental variables in pharmacoepidemiolog [J]. Pharmacoepidemiol Drug Saf, 2017, 26(4): 357–367. DOI: 10.1002/pds.4158.

[17] Baiocchi M, Cheng J, Small DS. Tutorial in biostatistics: instrumental variable methods for causal inference[J]. Stat Med, 2014, 33(13): 2297–2340. DOI: 10.1002/sim.6128.

[18] Choudhry NK, Brennan T, Toscano M, et al. Rationale and design of the Post-MI FREEE trial: a randomized evaluation of first-dollar drug coverage for post-myocardial infarction secondary preventive therapies[J]. Am Heart J, 2008, 156(1): 31–36. DOI: 10.1016/j.ahj.2008.03.021.

[19] Choudhry NK, Krumme AA, Ercole PM, et al. Effect of reminder devices on medication adherence: the REMIND randomized

clinical trial[J]. JAMA Intern Med, 2017, 177(5): 624-631. DOI:10.1001/jamainternmed.2016.9627.

[20] Rand CM, Brill H, Albertin C, et al. Effectiveness of centralized text message reminders on human papillomavirus immunization coverage for publicly insured adolescents[J]. J Adolesc Health, 2015, 56 Suppl 5: S17-20. DOI: 10.1016/j.jadohealth.2014.10.273.

[21] Gravenstein S, Dahal R, Gozalo PL, et al. A cluster randomized controlled trial comparing relative effectiveness of two licensed influenza vaccines in US nursing homes: design and rationale[J]. Clin Trials, 2016, 13(3): 264-274. DOI: 10.1177/1740774515625976.

[22] Zwarenstein M, Hux J, Kelsall D, et al. The Ontario printed educational message (OPEM) trial to narrow the evidence-practice gap with respect to prescribing practices of general and family physicians: a cluster randomized controlled trial, targeting the care of individuals with diabetes and hypertension in Ontario, Canada[J]. Implement Sci, 2007, 2: 37. DOI: 10.1186/1748-5908-2-37.

[23] Pinto D, Heleno B, Rodrigues DS, et al. An open cluster-randomized, 18-month trial to compare the effectiveness of educational outreach visits with usual guideline dissemination to improve family physician prescribing[J]. Implement Sci, 2014, 9: 10. DOI: 10.1186/1748-5908-9-10.

[24] Choudhry NK, Isaac T, Lauffenburger JC, et al. Rationale and design of the study of a Tele-pharmacy Intervention for Chronic diseases to Improve Treatment adherence (STIC2IT): a cluster-randomized pragmatic trial[J]. Am Heart J, 2016, 180: 90-97. DOI: 10.1016/j.ahj.2016.07.017.

[25] Nguyen TL, Collins GS, Lamy A, et al. Simple randomization did not protect against bias in smaller trials[J]. J Clin Epidemiol, 2017, 84: 105-113. DOI: 10.1016/j.jclinepi.2017.02.010.

[26] Peter WLS, Liu JN, Weinhandl E, et al. A comparison of sevelamer and calcium-based phosphate binders on mortality, hospitalization, and morbidity in hemodialysis: a secondary analysis of the dialysis clinical outcomes revisited (DCOR) randomized trial using claims data[J]. Am J Kidney Dis, 2008, 51(3): 445-454. DOI: 10.1053/j.ajkd.2007.12.002.

[27] 于莉莉. 临床试验中适应性设计若干问题的探索性研究[D]. 西安: 第四军医大学, 2007.
Yu LL. Exploration on several aspects of adaptive design in clinical trial[D]. Xi'an: Fourth Military Medical University, 2007.

[28] Angus DC. Fusing randomized trials with big data: the key to self-learning health care systems?[J]. JAMA, 2015, 314(8): 767-768. DOI: 10.1001/jama.2015.7762.

[29] Wennberg DE, Marr A, Lang L, et al. A randomized trial of a telephone care-management strategy[J]. N Engl J Med, 2010, 363(13): 1245-1255. DOI: 10.1056/NEJMsa0902321.

[30] Gabriel MM, Dunn DT, Speakman A, et al. Protocol, rationale and design of SELPHI: a randomised controlled trial assessing whether offering free HIV self-testing kits via the internet increases the rate of HIV diagnosis[J]. BMC Infect Dis, 2018, 18: 531. DOI: 10.1186/s12879-018-3433-x.

[31] Berry SM, Connor JT, Lewis RJ. The platform trial: an efficient strategy for evaluating multiple treatments[J]. JAMA, 2015, 313(16): 1619-1620. DOI: 10.1001/jama.2015.2316.

[32] Berry SM, Petzold EA, Dull P, et al. A response adaptive randomization platform trial for efficient evaluation of Ebola virus treatments: a model for pandemic response[J]. Clin Trials, 2016, 13(1): 22-30. DOI: 10.1177/1740774515621721.

[33] 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. DOI: 10.14236/jhi.v22i3.177.

（收稿日期：2018-11-16）
（本文编辑：李银鸽）

中华流行病学杂志第七届编辑委员会通讯编委名单

（按姓氏汉语拼音排序）

陈　曦	党少农	窦丰满	高　婷	高立冬	还锡萍	贾曼红	金连梅	荆春霞	李　琦	李十月
李秀央	林　玫	林　鹏	刘　莉	刘　玮	刘爱忠	马家奇	倪明健	欧剑鸣	潘晓红	彭晓旻
彭志行	任泽舫	施国庆	汤奋扬	田庆宝	王　丽	王　璐	王金桃	王丽敏	王志萍	武　鸣
谢　娟	解恒革	严卫丽	阎丽静	么鸿雁	余运贤	张宏伟	张茂俊	张卫东	郑　莹	郑素华
周脉耕	朱益民	祖荣强								