·专题二:双清论坛“全维度数据与智能诊疗的前沿与挑战”·

基于高质量临床研究,建立医疗健康多维度大数据

姜 勇 孟 霞 王拥军*

首都医科大学 附属北京天坛医院/国家神经系统疾病临床医学研究中心,北京 100070

[摘 要] 医疗健康大数据在维度、广度和深度都正快速增长。与西方发达国家相比,我国医疗健康多维度大数据的高质量数据源较少、数据规模较小。以高质量临床研究为基础,链接电子病历等临床诊疗数据和卫生行政数据为补充是目前建立医疗健康多维度大数据的最佳方式。建议国家出台战略规划,在现有高质量社区队列和专病队列的基础上,优化布局,保证对大型队列建设的持续投入;制定可操作的临床研究数据共享管理办法,确保通过数据共享激励数据收集者的未来工作积极性,改变未来临床研究模式,加速数据科学和人工智能的发展。

[关键词] 多维度大数据;精准医学;临床研究;数据共享

近十几年来,医疗健康大数据在维度、广度和深度都正快速增长。随着电子病历和电子健康档案的广泛应用,大型临床研究的不断开展使得医疗健康大数据的数量快速增长[1]。近几年,随着医学“分辨率”革命,基因组、蛋白组、代谢组等多组学技术在医疗健康领域广泛应用,移动医疗、物联网和可穿戴设备快速发展,临床检测技术和多模态高分辨医学影像技术的发展应用,新的生物标记物不断发现[2],这些都使得医疗健康大数据的内涵越来越丰富、医疗健康大数据的维度持续增长。全基因组测序技术的发展进步和测序费用的不断下降,使得临床研究大数据量迅速进入了 PB 甚至 ZB 时代[3]。

医疗健康多维度大数据的快速发展为医学研究的发展带来了新的机遇。大数据分析方法和技术日新月异的发展,特别是机器学习和深度学习技术的发展使多维度数据分析有了越来越多的分析挖掘工具[4,5]。云计算、高性能计算技术和发展为大数据分析提供了强大的计算能力,分布式存储的发展为临床研究大数据的存储提供了经济、高效、稳定的支撑。

大规模、多维度、高质量的数据源是大数据研究的前提。如何快速建设我国的医疗健康多维度大数据? 2019 年 5 月 11~12 日,国家自然科学基金委员会第 232 期双清论坛“全维度数据与智能诊疗的前沿与挑战”在上海召开,本次论坛的主题之一就是大规模人群队列建设及面向疾病精准诊疗的智能分析。会议达成共识,应继续建立并规范重大慢病高质量和大数据量的专病队列数据库,创建数据共享平台并制定管理方案[6]。

王拥军 首都医科大学附属北京天坛医院院长,国家神经系统疾病医疗质量控制中心主任,国家神经系统疾病临床医学研究中心副主任,中华医学会神经病学分会主任委员,中国卒中学会常务副会长,*Stroke and Vascular Neurology* 和《中国卒中杂志》主编。长期从事脑血管病精准诊疗、规范化防控策略的相关临床研究。在 *New England Journal of Medicine* 和 *JAMA* 等国际顶级杂志发表 SCI 期刊论文 240 余篇,多项成果受到国内、国际同行的广泛关注与高度评价。入选“北京学者”、“万人计划”科技创新领军人才、“北京市高层次创新创业人才支持计划杰出人才”,获得国家科技进步二等奖、首批全国创新争先奖章等多个奖项,带领团队获得首批科技部重点领域优秀创新团队称号。

姜勇 医学博士,副研究员,首都医科大学附属北京天坛医院国家神经系统疾病临床医学研究中心大数据中心负责人。主要负责国家神经系统疾病临床研究大数据平台的建立,开展脑血管病流行病学及临床研究和大数据分析方法学研究。担任中国卫生信息学会健康医疗大数据国际合作与交流专委会常委,中华预防医学会慢病分会委员,中华预防医学会流行病学分会委员。

收稿日期:2019-06-27;修回日期:2020-11-19

* 通信作者,Email: yongjunwang111@aliyun.com

1 医疗健康多维度大数据的发展现状

1.1 国际医疗健康大数据建立情况

欧美发达国家政府非常重视医疗健康大数据的建设，特别是基于医学科学研究的多维度大数据建设。这些高质量、大规模人群队列和疾病队列的建设，均由政府资助、组织支持建立和运行。2006 年英国政府发起的英国生物样本库（UK Biobank）项目是英国迄今以来规模最大的健康研究项目之一。在 2006 年至 2010 年期间，英国生物样本库完成了 50 余万名 40～69 岁志愿者的个人健康信息和血液、尿液、唾液等生物样本的采集。收集的数据涉及生活方式、疾病史和社会人口学指标，还进行了认知功能和听力测试[7]。2014 年，UK Biobank 启动了全球最大的医学影像研究计划，计划采集 10 万人的心脏、脑、腹部核磁、双能 X 线吸收（DXA）和颈动脉超声影像，以期望建立临床信息、多模态影像和基因等信息的多维度数据库[8]。2019 年 9 月，英国政府宣布已与四家全球领先制药公司及一家慈善机构达成战略合作，投资 2 亿英镑，支持对共计 50 万名参与者的全基因组测序项目[9]。2018 年 5 月，美国在 NIH 的支持下，启动了“All of US”精准医学队列，计划收集至少 100 万人的电子健康记录、生物样本，并通过可穿戴设备获取健康报告和信息[10]。所有参与者将进行全基因组测序。

1.2 我国医疗健康多维度大数据的建设情况

与西方发达国家相比，我国医疗健康多维度大数据的高质量数据源较少，数据规模较小。中国慢性病前瞻性研究项目（China Kadoorie Biobank，CKB）是中国医学科学院与英国牛津大学联合开展的慢性病国际合作项目，是一项超大规模自然人群队列。CKB 项目于 2004 年启动，共收集了 51 万份 30～79 岁居民的信息和生物样本，主要探讨环境、个体生活方式、体格和生化指标、遗传等众多因素对复杂慢性病发生、发展的影响[11]。2007 年，复旦大学和泰州市政府共建了泰州队列，计划收集 10 万份 30～80 岁人群的健康信息和生物样本，目的是研究主要慢性病的发病率、死亡率及行为、环境、基因的影响因素。目前总样本扩大到 20 万份，正在开展表型组学与暴露组学的相关研究[12]。

在专病队列建设方面，北京天坛医院国家神经系统疾病临床研究中心开展了中国国家卒中登记研究Ⅲ（CNSR Ⅲ），建立了超过 1.5 万人的脑血管病精准队列，目标为评价缺血性脑血管病相关危险因素，探索病因及发病机制分布，探索包括影像学特征的 TIA/卒中风险预测模型，覆盖全国 201 家医院[13]。基线收集了超过 5 000 个临床表型、高分辨影像和生物样本，完成了 1.1 万人全基因组测序，并进行长期随访，建成了脑血管病专病多维度数据。在其他疾病领域，面向精准医学的大型队列也陆续启动建立。

2 未来医疗健康多维度大数据发展方向

为推动数据科学的发展，近年来欧美发达国家启动了国家层面的数据科学战略规划，出台政策以支持医疗健康大数据的建立和共享。2017 年欧洲启动了 BigData@Heart 计划，将队列研究、电子病历、医疗质量改进注册登记研究、临床试验数据和影像数据整合在一起为新药物研发和个体化医疗提供基础[14]。2018 年 6 月 4 日，美国国立卫生研究院（NIH）发布了《数据科学战略计划》（NIH Strategic Plan for Data Science），强调支持高效的生物医学研究数据基础设施，促进数据资源生态系统的现代化，开发和推广高级数据管理、分析和可视化工具，加强生物医学数据科学的人才队伍建设，制定适当的政策以促进管理和可持续发展。其中重点提到了充分利用现有的临床研究资源，建立符合 FAIR 原则，即可查找（Findable）、可访问（Accessible）、可互操作（Interoperable）和可重复使用（Reusable）的共享数据库[15]。

十三五期间，我国部署了“精准医学研究”重点研发计划，在现有队列的基础上，建立了一系列面向精准医学研究的一般人群队列和专病队列[16]。然而，以项目支持的方式建设多维度大数据的队列力度远远不够，缺少长期、充足、稳定的资金支持，需要不断申请新的资金来源来支持，耗费了科研人员大量的时间和精力。此外，与欧美发达国家相比，国内队列的标准化、规范化和系统化水平亟待提高，现有队列的科学管理、质量控制和数据共享机制亟待完善，大型队列的随访、外部数据的获取、队列可持续发展等问题亟待解决。

3 建设健康医疗多维度大数据的具体建议

3.1 整合临床研究数据，建设高质量大数据来源

高质量的临床研究设计基于明确的临床问题，具备清晰的研究问题和科学假设，明确地纳入排除标准，有代表性的病例选择方法和准确的研究结局终点。研究方案经过多轮专家论证，科学性可以得到充分保障；数据采集的内容多参考国际或国内相关大型研究的标准，多中心研究数据有统一的采集标准。在组织实施层面，临床研究实施有良好的全

流程的质量控制，可以很好地保证研究数据的质量。随着全基因组测序、多模态高分辨影像、冷冻电镜等高精度检查技术手段的应用，数据的维度也越来越高。临床研究数据采集方式不断改进，数据采集的效率也越来越高。基于以上临床研究的特点，建立基于高质量临床研究的医疗健康大数据是目前建立医疗健康多维度大数据的最佳方式。

临床研究大数据缺点主要表现在：传统的临床研究都是由研究者申请项目经费开展的，由于研究经费和组织实施的限制，样本量相对较小。另外临床研究的数据标准不一致，无法汇交成更大的数据样本。美国 NIH 提出，未来应通过政府持续足量投入专项经费资助专门的机构研究大数据，将考核标准定为对外提供数据共享和支持其他研究的数量和质量等；建立统一的数据标准，统一临床数据公共数据元(CDE)，建立数据共享机制，通过推动数据共享来促进高质量数据库的建立[15]。

3.2　合理利用以电子病历为代表的临床诊疗数据

电子病历数据是在临床实践中产生的用于管理目的的信息。它具有数据量大、覆盖人群广、代表性好、增长速度快、及时性高等特点。通过建设高质量的、符合科研需求的专病结构化电子病历，与院内 HIS、LIS、PACS 系统信息无缝对接，收集和保存生物样本，链接可穿戴设备数据和患者报告数据，定期开展随访，是建立起真实世界研究大数据的理想模式。

以电子病历为核心的医院临床诊疗数据是真实世界医疗健康大数据的重要来源，也是近年来医疗健康大数据研究的主要领域。英国、加拿大等信息化发展较早的发达国家，已经开展了许多基于电子病历数据、医疗保险数据和药物不良反应监测等大数据分析的研究[1]；我国也在该领域开展了一些探索性的研究，通过对社区健康档案、医院电子病历数据和生命登记数据等数据源的链接，建立基于区域卫生信息平台的真实世界研究数据库[17]。

然而基于电子病历的医疗大数据存在一些缺陷：(1) 医院数据采集的目的是根据医疗实践的管理需要，并非用于科研。这导致了某些数据并不真实、可靠[18]，如住院死亡率、药物不良反应等数据。(2) 真实世界中医生对患者诊断、治疗的个体化使得不同的患者电子病历数据的稀疏性和数据完整性往往不能满足高质量临床研究的需要。(3) 随访数据的内容优先，应答率和准确性不高。(4) 基于单个医院的研究，其电子病历数据量、代表性往往不能满足大型真实世界研究的需求，研究结果受病人来源的影响很大。(5) 多中心研究中，各医院使用的电子病历标准不一致，结构化程度参差不齐，数据编码各不相同，数据汇交要耗费大量的人力。(6) 不同医院的影像数据也存在不同的质量问题：包括分辨率低、不同模式下不同的切片采集方向、对比度差、有限的视场以及不同模式扫描的不对准等。(7) 多中心的研究往往需要各分中心医院上传数据至项目牵头医院，很多医院担心泄露医院和患者隐私而不愿意参加。这些都是真实世界研究不能广泛开展的原因。

3.3　促进卫生行政数据的开放共享和利用

以医院病案首页、医疗保险、死亡登记为代表的卫生行政数据库和公共卫生数据库是医疗健康大数据的重要来源。卫生行政数据库和公共卫生数据库是临床研究大数据的有益补充，可为大型队列的临床结局提供线索。国际上医疗信息化发展比较早的国家如英国、韩国等开展了很多基于卫生行政数据库链接的研究。我国在部分地区也进行了一些初步的探索，但在数据可及性方面需要有关部门制定政策，在保证患者隐私和数据安全同时，更加积极地开放共享数据。

结合国际上的经验来看，理想的建立医疗健康多维度大数据的方式是基于高质量临床研究大数据，经患者授权链接电子病历数据，补充收集的临床诊疗信息，链接物联网、可穿戴设备的客观数据，定期开展随访研究，并利用卫生行政数据补充必要的经济数据和结局信息。

4　结　语

建立医疗健康多维度大数据是医疗大数据研究发展的重中之重。依托国家临床医学研究中心联盟、各临床中心及协同研究网络等专业技术机构，通过政府的持续投入资助高质量临床研究，开放共享卫生数据和科研数据，可加快促进医疗健康大数据的发展，推动精准医学和智慧医疗的进步。

参　考　文　献

[1] Hemingway H, Asselbergs FW, Danesh J, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. European Heart Journal, 2018, 39: 1481—149.

[2] Simpkins AN, Janowski M, Oz HS, et al. Biomarker application for precision medicine in stroke. Translational Stroke Research, 2019, 11(1): 615—627.

[3] Acosta JN, Brown SC, Falcone GJ. Genetic underpinnings of recovery after stroke: An opportunity for gene discovery, risk stratification, and precision medicine. Genome Medicine, 2019, 11: 58.

[4] Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 2019, 20: e262—e273.
[5] Wu O, Winzeck S, Giese AK, et al. Big data approaches to phenotyping acute ischemic stroke using automated lesion segmentation of multi-center magnetic resonance imaging data. Stroke, 2019, 50(7): 1734—1741.
[6] 国家自然科学基金委员会. 第 232 期双清论坛"全维度数据与智能诊疗的前沿与挑战"在上海召开. (2019-05-23)/[2020-06-27]. http://www.nsfc.gov.cn/publish/portal0/tab445/info75971.htm.
[7] Bycroft C, Freeman C, Petkova D, et al. The UK biobank resource with deep phenotyping and genomic data. Nature, 2018, 562(7726): 203—209.
[8] Littlejohns TJ, Holliday J, Gibson LM, et al. The UK biobank imaging enhancement of 100, 000 participants: Rationale, data collection, management and future directions. Nature Communications, 2020, 11: 2624.
[9] Biobank U. UK biobank leads the way in genetics research. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-leads-the-way-in-genetics-research-to-tackle-chronic-diseases-1.
[10] Denny JC, Rutter JL, Goldstein DB, et al. The "all of us" research program. New England Journal of Medicine, 2019, 381(7): 668—676.
[11] Chen ZM, Chen JS, Collins R, et al. China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. International Journal of Epidemiology, 2011, 40(6): 1652—1666.
[12] Wang XF, Lu M, Qian J, et al. Rationales, design and recruitment of the Taizhou longitudinal study. BMC Public Health, 2009, 9: 223.
[13] Wang YJ, Jing J, Meng X, et al. The third china national stroke registry (cnsr-iii) for patients with acute ischaemic stroke or transient ischaemic attack: Design, rationale and baseline patient characteristics. Stroke and Vascular Neurology, 2019, 4(3): 158—164.
[14] Anker S, Asselbergs FW, Brobert G, et al. Big data in cardiovascular disease. European Heart Journal, 2017, 38(24): 1863—1865.
[15] National Institutes of Health. Nih strategic plan for data science. (2029-09-14)/[2020-06-27]. https://datascience.nih.gov/strategicplan
[16] 科学技术部. 科技部关于发布国家重点研发计划精准医学研究等重点专项 2016 年度项目申报指南的通知. (2016-03-10)/[2020-06-27]. https://www.dcmst.org.cn/tz3/232-v-2016.
[17] Lin H, Tang X, Shen P, et al. Using big data to improve cardiovascular care and outcomes in china: a protocol for the chinese electronic health records research in yinzhou (cherry) study. BMJ Open, 2018, 8(2): e019698
[18] Lenzer J. Big data's big bias: Bringing noise and conflicts to us drug regulation. BMJ, 2017, 358: j3275.

Establishing Multidimensional Healthcare Big Data Based on High-Quality Clinical Research

Jiang Yong Meng Xia Wang Yongjun*

Beijing Tiantan Hospital, Capital Medical University China National Clinical Research Center for Neurological Diseases, Beijing 100070

Abstract Big data on healthcare is growing rapidly in dimension, breadth and depth. Compared with western developed countries, China has fewer high-quality data sources and smaller data scale for multidimensional big data on health care. Based on high-quality clinical research, linking clinical diagnosis and treatment data such as electronic medical records and health administration data as supplement is the best way to establish multidimensional big data of healthcare. It is suggested that strategic plan should be formulated to optimize the layout based on the existing high-quality community and patient cohorts. National continuous sufficient investment should be ensured to maintain large cohorts study. Reasonable data management and sharing protocol for clinical research should be developed to ensure that data collectors are motivated to share data, to innovate clinical research practice, and to accelerate the development of data science and artificial intelligence in the future.

Keywords multi-dimensional big data; clinical research; precision medicine; data sharing

(责任编辑 姜钧译 仲斌演)

* Corresponding Author, Email: yongjunwang111@aliyun.com