·创刊40周年·

大数据时代的流行病学研究：机遇、挑战与展望

王玉琢　马红霞　靳光付　胡志斌　沈洪兵
南京医科大学公共卫生学院/南京医科大学省部共建肿瘤个体化医学协同创新中心 211166
通信作者：沈洪兵，Email：hbshen@njmu.edu.cn

【摘要】　近年来，高通量组学技术迅速发展，人群队列与样本资源逐渐累积，互联网和信息技术飞速进步，大数据开发工具不断涌现，健康医疗大数据为流行病学研究提供了丰富的数据资源和广阔的科研平台。本文将分析大数据时代流行病学研究的机遇和挑战，展望未来学科的发展方向。流行病学应抓住机遇，开辟新方向、发展新方法，推动医学基础研究成果向公共卫生和临床实践的转化应用，实现"健康中国"的美好愿景。

【关键词】　健康医疗大数据；　流行病学研究；　系统流行病学；　多学科融合

基金项目：国家自然科学基金（81820108028，81521004）

Epidemiological research in the big data era: opportunities, challenges and prospectives

Wang Yuzhuo, Ma Hongxia, Jin Guangfu, Hu Zhibin, Shen Hongbing
School of Public Health, Nanjing Medical University/Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China
Corresponding author: Shen Hongbing, Email: hbshen@njmu.edu.cn

【Abstract】　Recent years with the rapid progress in high-throughput omics techniques, the accumulation of population cohorts and biobanks, great advances in internet and information technology, and the emerging tools for big data utilization, health care big data provide abundant data resources and broad research platforms for epidemiological research. We aimed to review the opportunities and challenges of epidemiological research in the era of big data, and explore the future development of epidemiology. Epidemiology should seize the opportunities, open up new directions, develop new methods, and promote the translation of research results into public health and clinical medicine, which will eventually realize the vision of "Healthy China".

【Key words】　Health care big data;　Epidemiological research;　Systems epidemiology; Multidisciplinary integration

Fund programs:　National Natural Science Foundation of China (81820108028, 81521004)

近年来，高通量组学技术迅速发展，互联网和信息技术飞跃进步，大数据开发工具逐渐涌现，健康医疗大数据的研究价值开始受到重视。作为一门与数据息息相关的学科，流行病学正处于"大健康""大数据""人工智能"时代带来的学科发展机遇期，但在数据标准化与共享、检测技术与分析方法、法律和伦理规范与制度等方面尚存在诸多挑战。本文将分析大数据时代流行病学研究的机遇和挑战，探讨未来学科的发展方向，为大数据时代的流行病学研究提供参考。

一、大数据时代流行病学研究的机遇

1. 高通量组学技术的发展：人类基因组计划启

DOI：10.3760/cma.j.cn112338-20201203-01377
收稿日期　2020-12-03　　**本文编辑**　李银鸽
引用本文：王玉琢, 马红霞, 靳光付, 等. 大数据时代的流行病学研究：机遇、挑战与展望[J]. 中华流行病学杂志, 2021, 42(1): 10-14. DOI: 10.3760/cma.j.cn112338-20201203-01377.

动以来，以新一代测序技术和质谱技术为代表的高通量组学技术的突破，推动了基因组、表观遗传学、转录组、蛋白质组、代谢组、微生物组等生物医学组学数据的指数级增长[1]。与此同时，计算机科学、生物信息学和数理统计学的飞速发展使得组学数据分析的障碍逐渐消除。以人群为基础的基因组研究、甲基化研究、代谢组研究等组学研究方兴未艾，在探索疾病病因和寻找可能的干预靶点方面取得了重大突破，也为未来的医学研究提供了丰富的数据资源。

2. 人群队列研究数据和样本资源的积累：队列研究是探讨疾病致病因素和评价干预措施的有效方法，也是将基础研究转化应用到疾病防治实践的重要途径。许多发达国家已将构建具有生物样本库的超大规模人群队列作为战略布局，如英国生物样本库、丹麦国家生物样本库、美国百万老兵项目，以及美国“All of Us”研究计划[2]。我国流行病学工作者通过国内外合作，建立了一批具有生物样本库的大型人群队列，如中国慢性病前瞻性研究[3]和泰州人群健康跟踪调查队列[4]。“十三五”期间，我国依托国家重点研发计划启动了精准医学重点专项和重大慢性病防控重点专项，建立了百万级自然人群队列和重大疾病专病队列。同时，基于双生子、母婴、老年人等特殊人群的队列也逐渐涌现，如国家出生队列、中国健康与退休前瞻性队列等[5]。这将为未来开展全组学设计的流行病学研究创造条件。

3. 大数据和互联网技术的发展：当今科学技术高速发展，以物联网、互联网、移动智能、云计算、大数据为代表的信息技术广泛运用于医疗领域，医疗信息化得到全面、快速发展。在我国，国家政策已深入医疗信息化改革，各地正大力推进以电子病历为核心的医院信息化建设，积极探索并建立以电子健康档案为核心的区域医疗信息平台[6]。医疗信息化的广泛应用在科学研究、医疗服务、健康保健和卫生管理过程中形成了健康医疗大数据（health care big data）。我国于 2019 年 7 月发布的《健康中国行动（2019-2030 年）》中指出“鼓励和支持科研机构与高新技术企业深度合作，充分运用互联网、物联网、大数据等信息技术手段，开展大型队列研究……”。在我国，疾病登记、死亡登记、医院质量监测系统等数据已经应用于疾病负担估计和病因探索。我国搭建了一些健康医疗大数据平台，如国家生物信息中心和中国队列共享平台等[7]。

4. 人工智能技术的发展：健康医疗大数据的应用前景取决于对这些数据进行多层次的挖掘和全方位的整合，增进对人类健康或疾病的理解。传统的统计方法无法适应大数据分析的要求。而以机器学习和深度学习为代表的人工智能技术在大数据分析中的应用正呈现爆发式增长，在流行病学监测、疾病诊断、病因学研究、风险预测等方面显示出巨大的应用潜力[8]。例如，有研究者将深度卷积神经网络模型用于超声图像进行甲状腺癌诊断，模型的灵敏度和特异度均高于 85%[9]；机器学习算法中的特征选择技术能够从高维数据中识别相关特征，被广泛应用于组学数据和临床数据的整合分析，从而形成病因假设、预测疾病风险及预后[8]。人工智能技术的发展将为大数据时代的流行病学研究提供方法支撑。

5. 流行病学研究方法的发展：近年来，组学数据分析方法日趋成熟，数理统计、人工智能等技术的进步推动了大数据的有机整合，流行病学研究也取得了长足发展。健康医疗大数据极大地丰富了潜在危险因素图谱，有利于提出更多的病因假设。其丰富的表型组信息有助于目标遗传变异与全部健康结局的关联分析，全表型组关联研究（PheWAS）应运而生[10]。大数据拓宽了可研究的工具变量，为孟德尔随机化研究提供了新机遇。电子健康档案的普及和完善推动了时效性临床试验（PCT）的发展，PCT 在真实世界中评价预防干预措施的效果，与传统的随机对照试验互为补充，助力干预措施的效果评价[11]。大数据时代为流行病学研究提供了前所未有的机遇，将推动基础医学研究成果向公共卫生和临床实践的转化应用。

二、大数据时代流行病学研究面临的挑战

1. 我国健康医疗大数据来源、收集和管理的共享机制尚不成熟，形成数据孤岛：我国健康医疗大数据的共享管理机制尚不成熟，导致大量数据分散在不同的实验室和机构内部[12]。来源于各个机构、领域的数据往往使用不同的系统和软件平台，数据收集、存储和管理的标准规范也各不相同，多来源的数据在格式、结构、语义关联、质量等方面都存在差异。这些因素严重制约我国健康医疗大数据的整合共享与转化利用[13]。如何实现数据的标准化和规范化，将跨机构、跨领域、跨平台的数据库打通共享，是目前大数据使用面临的重大挑战之一。

2. 健康医疗大数据存在系统误差，影响暴露-结局之间关联的真实性：相对于传统的随机抽

样，大数据通常来自具有某些特征的人群（如医院就诊人群、携带可穿戴设备人群），可能存在选择偏倚。多源异构的健康医疗大数据存在表型定义的不确定性、变量测量的不准确性和数据缺失等问题，造成信息偏倚；部分信息的测量方法和获取途径难以得知，为质量控制带来困难[14]。健康医疗大数据存在大量噪声信号，大规模地探索疾病结局有关的因素而忽略因果关系判断的准则，可能导致疾病影响程度的错误估计、触发错误警报、造成虚假关联和生态学谬误[15]。此外，基于健康医疗大数据的流行病学研究属于观察性研究，存在混杂和反向因果关联等诸多问题[16]。因此，大数据时代的流行病学研究应注意对研究质量的严格控制，实现数据收集与处理的标准化和规范化，重视因果关系判断的基本要素，确保结论的可靠性。

3. 组学检测技术有待发展完善：在大规模人群开展多组学检测依赖于成熟的高通量组学技术，其对样本和检测条件的要求不高，检测成本低廉、检测方法快速、结果稳定性和准确度高。目前基因组测序仍存在一定的错误率[17]；蛋白质组和代谢组检测平台之间的技术差异增加了数据标准化和可重复性评估的难度[18]。组学检测技术仍有待发展完善。值得一提的是，近几年迅速兴起的单细胞转录组测序不仅揭示出不同细胞类型之间的异质性，而且可以发现全新的细胞类型。今后，单细胞测序技术将逐步扩展到基因组、表观遗传学、蛋白质组等，为研究者在单细胞水平探索疾病发生发展机制打开崭新视角[19]。

4. 健康医疗大数据分析、整合与存储的关键技术需要新的突破：多组学生物统计和生物信息方法有待进一步提高。基因、蛋白质、代谢物等所构建的调控网络极为复杂，如何整合多组学信息以形成系统层面的理解是健康医疗大数据研究面临的重要挑战之一。由于每个组学数据都存在一定的偏倚和变异度，需要针对不同组学数据建立可靠的统计分析框架[17]。而对多组学数据进行整合分析时，则需要更复杂的多维统计方法，如贝叶斯模型、神经网络等[17]。健康医疗大数据来源广泛，种类繁多，包含结构化、半结构化、非结构化等不同的数据类型。然而，当前缺少统一的数据标准和规范，导致不同层次、不同类型的数据整合困难。如何对半结构化（如电子病历）、非结构化（如心电图、医学影像资料）和对流数据（实时视频、传感器数据、医疗设备监测数据）进行处理，是健康医疗大数据分析面临的又一重要挑战。此外，临床诊疗大数据、健康监测数据等非科研数据存在信息缺失（通常为非随机性缺失）和测量、记录不准确等数据质量问题；如何进行数据预处理和数据清理以保障研究结果的可靠性是大数据研究亟须解决的重要问题[20]。除了数据分析以外，健康医疗大数据的存储和分析需要巨大的计算资源。目前适用于大数据的软硬件平台，大数据存储、传输、高性能计算和云计算等技术尚不成熟[13,21]。这些因素制约着健康医疗大数据的研究、转化和应用。

5. 健康医疗大数据的法律和伦理规制已成为亟待解决的问题：健康医疗大数据在存储、共享、分析、挖掘等过程中存在个人信息权、隐私权可能受侵害的风险和数据安全的隐患[22]。如何对健康医疗大数据进行法律和伦理上的规制，已成为亟待解决的问题。我国应尽快推动个人健康信息和隐私保护相关法律法规的立法，明确电子健康信息在存储、管理、利用、交换各环节的权责归属与过程标准，促进数据安全技术的与时俱进。相关部门应从保护个人安全和公平的角度，合理界定伦理风险，把握伦理审批的界限[13]。我国在2019年颁布《人类遗传资源管理条例》，迈出了大数据安全与伦理规制的探索性尝试。2020年，国家药品监督管理局相继发布《真实世界证据支持药物研发与审评的指导原则（试行）》、《用于产生真实世界证据的真实世界数据指导原则（征求意见稿）》和《真实世界数据用于医疗器械临床评价技术指导原则（试行）》，指出“真实世界研究涉及个人信息保护应遵循国家信息安全技术规范、医疗大数据安全管理相关规定”，要求实行个人信息保护和数据安全性处理，拉开了我国监管机构从监管层面构建真实世界数据使用框架体系的序幕。

三、大数据时代流行病学研究的发展展望

1. 高质量人群队列的建设：随着电子健康档案的普及和完善，将电子健康档案与基于生物样本库的多组学数据以及其他常规数据（如出生登记、死亡登记、环境监测数据等）互融互通而建立大型动态队列已成为流行病学研究的热点[7]。由于研究资源和经费的限制，大规模前瞻性队列难以实现全表型的收集和全组学的检测。因此，选择具有代表性的研究对象，使用最先进的技术收集队列成员的表型组信息，进行多次动态随访，采集和储存多时点生物标本，开展多组学检测，由此构建精细队列，具有深远的研究价值[2]。

2. 严格的流行病学研究设计和质量控制：严格的流行病学研究设计和质量控制是高质量队列研究的重要保障。队列成员的多样性是研究结果外推应用于全人群的关键。美国“All of Us”研究计划将人群多样性作为队列建设的首要目标，其80%的队列成员为弱势群体[2]。队列研究应注意提高应答率和依从性，避免选择偏倚。队列信息的类型和深度应综合考虑其与健康结局的相关性、检测方法的可靠性和研究经费。队列随访方面，近年来“成人疾病胎源说”备受关注，理想的随访应实现“从摇篮到坟墓”的终身观察；随访间隔将视研究结局的变化速度，研究的人力、物力等条件而定[2]。

3. 健康医疗大数据管理和共享平台的建设：大数据是国家重要的基础性战略资源。许多发达国家很早就开始重视大数据的收集、管理、分析和应用，建立起一批健康医疗大数据中心，如美国国家生物技术信息中心（NCBI）、欧洲生物信息学研究所（EBI）以及日本DNA数据库（DDBJ）等[12]。2019年，我国成立国家生物信息中心（CNCB），承担国家生命与健康大数据汇交存储、安全管理、开放共享与整合挖掘研究体系的建设，研发大数据前沿交叉与转化应用的新方法和新技术，建成支撑我国生命科学发展、国际领先的基因组科学数据中心（NGDC），迈入了新的历史发展阶段。

4. 基于队列研究的群体基因组学和系统流行病学研究：随着高通量组学技术和健康医疗大数据的发展，系统流行病学应运而生。系统流行病学借助系统生物学、流行病学、计算数学等技术，将人体从暴露组，到基因组、表观遗传组、蛋白质组、代谢组等，再到临床表型组的各个层次有机地整合在一起进行研究，从而深入理解多层次因素间复杂的关系网络及其相互作用，实现人群层面“暴露因素-组学标志物-疾病结局”的病因学推断[23]。在此基础上，系统流行病学将构建以病因网络为基础的疾病风险预测模型，推动“精准预防”理念的实现[23]。

5. 多学科交叉、复合型人才的培养：大数据时代对流行病学专业人才提出了更高的要求，包括扎实的流行病学理论基础，丰富的流行病学研究经验，对大数据进行深入挖掘，做出有价值的分析、判定和预测的能力。流行病学研究应聚焦国家重大需求，加强学科交叉融合，打破行政壁垒，加快交叉学科人才引进与培养，促进跨学科的合作与协同创新，完善促进跨学科合作的学术评价和成果共享机制，让微观和宏观流行病学有机地结合起来。注重系统整合协同，坚持开放创新，组建多学科、多领域、多系统融合的创新大团队[24-25]。

四、总结

大数据时代为流行病学提供了丰富的数据资源和广阔的科研平台。流行病学应抓住机遇，开辟新方向、发展新方法。今后有望在流行病学、计算机科学、临床医学、生命科学等学科的共同参与下，推动医学基础研究成果向公共卫生和临床实践的转化应用，提高疾病防控技术水平，促进健康发展，实现“健康中国”的美好愿景。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] 张国庆,李亦学,王泽峰,等. 生物医学大数据发展的新挑战与趋势[J]. 中国科学院院刊, 2018, 33(8): 853-860. DOI: 10.16418/j.issn.1000-3045.2018.08.013.
Zhang GQ, Li YX, Wang ZF, et al. New challenges and trends in bio-med big data[J]. Bull Chin Acad Sci, 2018, 33(8): 853-860. DOI:10.16418/j.issn.1000-3045.2018.08.013.
[2] Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare[J]. Nat Med, 2020, 26(1): 29-38. DOI: 10.1038/s41591-019-0727-5.
[3] Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up[J]. Int J Epidemiol, 2011, 40(6): 1652-1666. DOI: 10.1093/ije/dyr120.
[4] Wang X, Lu M, Qian J, et al. Rationales, design and recruitment of the Taizhou Longitudinal Study[J]. BMC Public Health, 2009, 9: 223. DOI: 10.1186/1471-2458-9-223.
[5] Zhao Y, Hu Y, Smith JP, et al. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS) [J]. Int J Epidemiol, 2014, 43(1): 61-68. DOI:10.1093/ije/dys203.
[6] 项高悦,曾智,沈永健. 我国智慧医疗建设的现状及发展趋势探究[J]. 中国全科医学, 2016, 24: 2998-3000. DOI:CNKI:SUN:QKYX.0.2016-24-031.
Xiang GY, Zeng Z, Shen YJ. Present Situation and Development Trend of China's Intelligent Medical Construction[J]. Chin Generl Prac, 2016, 24: 2998-3000. DOI:CNKI:SUN:QKYX.0.2016-24-031.
[7] 杨淞淳,吕筠,李立明. 流行病学研究新进展[J]. 中华流行病学杂志, 2020, 41(1): 1-5. DOI: 10.3760/cma.j.issn.0254-6450.2020.01.001.
Yang SC, Lyu J, Li LM. New progress in epidemiological research [J]. Chin J Epidemiol, 2020, 41(1): 1-5. DOI: 10.3760/cma.j.issn.0254-6450.2020.01.001.
[8] Schwalbe N, Wahl B. Artificial intelligence and the future of global health[J]. Lancet, 2020, 395(10236): 1579-1586. DOI:10.1016/S0140-6736(20)30226-9.
[9] Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study[J]. Lancet Oncol, 2019, 20(2): 193-201. DOI:10.1016/S1470-2045(18)30762-9.
[10] Bush WS, Oetjens MT, Crawford DC. Unravelling the human

genome-phenome relationship using phenome-wide association studies[J]. Nat Rev Genet,2016,17(3):129-145. DOI:10.1038/nrg.2015.36.
[11] Shortreed SM, Cook AJ, Coley RY, et al. Challenges and Opportunities for Using Big Health Care Data to Advance Medical Science and Public Health[J]. Am J Epidemiol, 2019,188(5):851-861. DOI:10.1093/aje/kwy292.
[12] 鲍一明,薛勇彪. 生命与健康大数据现状和展望[J]. 中国科学院院刊, 2018, 33(8): 861-865. DOI: 10.16418/j.issn.1000-3045.2018.08.014.
Bao YM, Xue YB. Current Status and Prospect of Life and Health Big Data[J]. Bull Chin Acad Sci, 2018, 33(8): 861-865. DOI:10.16418/j.issn.1000-3045.2018.08.014.
[13] 王胜锋,宁毅,李立明. 健康医疗大数据互联互通模式的经验与挑战[J]. 中华流行病学杂志,2020,41(3):303-309. DOI: 10.3760/cma.j.issn.0254-6450.2020.03.005.
Wang SF, Ning Y, Li LM. Experience and challenge on interoperability of big data in health care[J]. Chin J Epidemiol, 2020, 41(3): 303-309. DOI: 10.3760/cma. j. issn.0254-6450.2020.03.005.
[14] 赵一菊,王声湧. 大数据:医学科学的机遇与挑战[J]. 伤害医学:电子版,2019,8(2):1-4. DOI:10.3868/j.issn.2095-1566.2019.02.001.
Zhao YJ, Wang SY. Big Data: Opportunity and Challenge for Medical Science[J]. Shang Hai Yi Xue, 2019, 8(2): 1-4. DOI: 10.3868/j.issn.2095-1566.2019.02.001.
[15] 宋菁,胡永华. 流行病学展望:医学大数据与精准医疗[J]. 中华流行病学杂志,2016,37(8):1164-1168. DOI:10.3760/cma.j.issn.0254-6450.2016.08.022.
Song J, Hu YH. Medical big data and precision medicine: prospects of epidemiology[J]. Chin J Epidemiol, 2016, 37(8): 1164-1168. DOI: 10.3760/cma. j. issn. 0254-6450.2016.08.022.
[16] Smith GD, Ebrahim S. Data dredging, bias, or confounding [J]. BMJ, 2002, 325(7378): 1437-1438. DOI: 10.1136/bmj.325.7378.1437.
[17] Karczewski KJ, Snyder MP. Integrative omics for health and disease[J]. Nat Rev Genet, 2018, 19(5): 299-310. DOI: 10.1038/nrg.2018.4.
[18] Mathe E, Hays JL, Stover DG, et al. The Omics Revolution Continues: The Maturation of High-Throughput Biological Data Sources[J]. Yearb Med Inform, 2018, 27(1): 211-222. DOI:10.1055/s-0038-1667085.
[19] Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics[J]. Nat Rev Genet, 2020 DOI: 10.1038/s41576-020-0265-5.
[20] 代涛. 健康医疗大数据发展应用的思考[J]. 医学信息学杂志,2016,37(2):1-8. DOI:10.3969/j.issn.1673-6036.2016.02.001.
Dai T. Health and Medical Big Data Development Perspective[J]. J Med Intel,2016,37(2):1-8. DOI:10.3969/j.issn.1673-6036.2016.02.001.
[21] 宁康,陈挺. 生物医学大数据的现状与展望[J]. 科学通报,2015,60(5):534. DOI:10.1360/N972014-00895.
Ning K, Chen T. Big data for biomedical research: Current status and prospective[J]. Chin Sci Bull, 2015, 60(5): 534. DOI:10.1360/N972014-00895.
[22] 王波,吕筠,李立明. 生物医学大数据:现状与展望[J]. 中华流行病学杂志, 2014, 35(6): 617-620. DOI: 10.3760/cma. j. issn.0254-6450.2014.06.001.
Wang B, Lyu J, Li LM. Big data in biomedicine: status quo and perspective[J]. Chin J Epidemiol,2014,35(6):617-620. DOI:10.3760/cma.j.issn.0254-6450.2014.06.001.
[23] 黄涛,李立明. 系统流行病学[J]. 中华流行病学杂志,2018,39(5): 694-699. DOI: 10.3760/cma. j. issn. 0254-6450.2018.05.031.
Huang T, Li LM. Systems epidemiology[J]. Chin J Epidemiol, 2018,39(5):694-699. DOI:10.3760/cma.j.issn.0254-6450.2018.05.031.
[24] 毛琛,王岚,李立明. 我国流行病学学科发展70年的历程与成就[J]. 中华流行病学杂志,2019,40(10):1173-1179. DOI:10.3760/cma.j.issn.0254-6450.2019.10.001.
Mao C, Wang L, Li LM. Historical perspective of progress and achievement on epidemiology in the past 70 years in China[J]. Chin J Epidemiol, 2019, 40(10): 1173-1179. DOI: 10.3760/cma.j.issn.0254-6450.2019.10.001.
[25] 沈洪兵. 抓住机遇、明确目标、加强流行病学学科建设[J]. 疾病控制杂志, 2004, 8(4): 317-317. DOI: 10.3969/j.issn.1674-3679.2004.04.013.
Shen HB. Seize the opportunity, define the goal and strengthen the construction of epidemiology discipline[J]. Chin J Dis Cont Prev, 2004, 8(4): 317-317. DOI: 10.3969/j.issn.1674-3679.2004.04.013.