DOI: 10.13376/j.cbls/2023172
文章编号：1004-0374(2023)12-1581-05

陈兴栋，复旦大学研究员，博士生导师，复旦大学泰州健康科学研究院执行院长，科技部第六次技术预测生物资源领域专家，国家优秀青年科学基金获得者，上海高校特聘教授 (东方学者)，上海曙光计划获得者，全国生物样本标准化技术委员会 (SAC/TC559) 委员，“福声计划”中国人群泛癌早筛前瞻性研究学术专家委员会顾问。主要研究方向为大型人群队列和人类遗传资源平台建设、重大慢性疾病的分子流行病学和遗传学研究，多次受邀参与国家部委、上海市的遗传资源咨询会议。近 5 年以第一及通讯作者在 *Journal of Hepatology*、*Alzheimer's & Dementia*、*Nature Communications* 等本领域高水平学术期刊发表 SCI 论文 60 余篇，授权发明专利 2 项，获软件著作权 9 项，出版专著 2 部，参与立项卫生健康信息团体标准 2 项。

大数据时代的大型人群队列

蒋艳峰[1,2]，高培培[1,2]，陈兴栋[1,2]*
(1 复旦大学人类表型组研究院，遗传工程国家重点实验室，上海 200433；
2 复旦大学泰州健康科学研究院，泰州 225300)

摘　要：大数据技术推动了生命组学数据的爆炸式增长，生命科学研究进入大数据时代。大型人群队列研究依托大数据技术获得了重要突破，为生命科学和精准医学研究提供了宝贵资源，推动着预防医学模式发生革命性变化。以大数据技术为导向，人群队列研究规模空前扩大、学科交叉创新层出不穷，如何充分高效地利用人群队列资源、实现队列间的互联互通与共享，也是队列研究亟待解决的难题。该文将重点分析大数据时代特征下国内大型人群队列研究的发展特点，并提出其未来发展趋势及面临的挑战。

关键词：大数据；人群队列；精准医学；表型组
中图分类号：Q811.4；R181　　**文献标志码：**A

Large-scale population cohorts in the era of big data

JIANG Yan-Feng[1,2], GAO Pei-Pei[1,2], CHEN Xing-Dong[1,2]*
(1 State Key Laboratory of Genetic Engineering, Human Phenome Institute, Fudan University, Shanghai 200433, China;
2 Fudan University Taizhou Institute of Health Sciences, Taizhou 225300, China)

Abstract: The explosion in omics data driven by big data has propelled the life science into the era of big data. Large-scale population cohort studies have made significant breakthroughs by leveraging big data technologies, thus providing valuable resources for life science and precision medicine and catalyzing revolutionary changes in the field of preventive medicine. With a focus on the big data technologies, population cohort studies have expanded on an unprecedented scale and fostered interdisciplinary innovations. However, effectively and efficiently utilizing population cohort resources, as well as achieving interconnectivity and data sharing among different cohorts, remain

收稿日期：2023-07-05；**修回日期：**2023-09-19
基金项目：上海市市级科技重大专项(2023SHZDZX02)；国家自然科学基金项目(82073637，82122060)
***通信作者：**E-mail: xingdongchen@fudan.edu.cn；Tel: 021-31246602

pressing challenges. This paper aims to analyze the development characteristics of Chinese large-scale population cohort studies in the era of big data and propose future trends and challenges.

Key words: bid data; cohort study; precision medicine; phenome

大型人群队列研究自 21 世纪以来取得了快速发展，以大数据为导向的大规模人群队列研究逐渐成为热点 [1]。自然人群是国家健康队列的重要组成部分，对代表性区域自然人群进行定期重复调查，有助于全面了解该区域人口健康状况，探索人群高发疾病与发病机制，明确疾病的发生和发展与遗传和环境因素的复杂关系，为疾病的早期预防和控制提供有力支持 [2]。大型人群队列研究在揭示疾病成因、监测多种疾病发病趋势、研究疾病自然史、实现精准医疗、建立精准医学生物大数据平台、开展表型组学研究等方面发挥着举足轻重的作用 [3]。

自人类基因组计划启动以来，高通量组学技术不断蓬勃发展，生命组学数据获得指数级增长 [4-5]。与此同时，伴随着生物信息学、表型组学以及人工智能技术的快速发展，生命科学研究进入了大样本、大数据和大发现时代 [6]。人类基因组计划完成后，表型组研究逐渐成为人类健康研究的下一个战略制高点 [7-8]。随着表型组学的兴起以及生命科学大数据时代的来临，庞大的数据资源使得人群队列研究如虎添翼，人群队列的研究模式也紧随着大数据时代的步伐悄然变革 [9]。

1 国内大型人群队列

相对于发达国家的人群队列，如美国“All of Us”百万级自然人群队列 [10]、欧洲 European Prospective Investigation into Cancer and Nutrition (EPIC，52.1 万) 和英国的 UK Biobank (UKB，50 万)[11]，我国的人群队列研究起步晚且普遍规模较小。然而，近些年来在政策支持下，我国已逐步部署建立若干大型人群队列研究项目。例如，中国疾病预防控制中心与英国牛津大学 2004 年合作启动的中国慢性病前瞻性研究 (China Kadoorie Biobank, CKB)[12]，其基线调查规模达 51.3 万，随后开展了三次重复调查，累计观察 770 万人年，收集约 130 万管血液样本以及数十万份尿液、DNA、唾液和粪便等样本；复旦大学联合多家单位在 2007 年启动的泰州人群健康跟踪调查 (Taizhou Longitudinal Study, TZL)[13] 已招募 20 万志愿者，平均随访时间 10 余年，每人产生约 1 000 个表型，累计收集血液、唾液、尿液和粪便等样本约 200 万份；上海肿瘤研究所建立的上海女性健康队列 (Shanghai Women’s Health Study)[14] 和上海男性健康队列 (Shanghai Men’s Health Study)[15] 分别对 7.5 万中老年女性和 6.2 万中老年男性进行了基线调查和多次随访。此外，还有一些针对特殊人群开展的队列研究，如 2016 年南京医科大学启动的中国国家出生队列 (China National Birth Cohort, CNBC)[16] 招募了 3 万个自然妊娠和 3 万个辅助生殖治疗家庭，其采集的生物样本除了包括常见的血液、尿液等，还有卵泡液、精浆和精子样本；华中科技大学 2008 年启动的东风 - 同济队列 (Dongfeng-Tongji Cohort, DFTJ)[17] 则招募了 3 万公司退休职工作为研究对象。

“十三五”期间国家启动了精准医学重点专项，通过京津冀、华东、华中、华南、西南、西北、东北等区域的自然人群队列共同建立了百万级自然人群大型健康队列 [18]。“健康中国 2030”中也提出了高效利用互联网和大数据技术，开展大型队列研究的建议。大型人群队列研究是探讨健康和疾病的发生、发展规律等重大生命科学问题的有效途径之一，也是逐步走向精准医学的必经之路 [19]。在大数据时代背景下，国内的大型人群队列呈现出以下特点。

1.1 队列规模增大、信息丰富多样化

大数据技术所赋能的数据收集、存储和分析等技术优势，使得大型人群队列研究已逐渐发展至数十万至百万规模。随着队列规模的逐渐扩大，基于大型队列组建生物样本库逐渐成为热点，由队列研究收集、产生的数据涵盖了基因组和各类表型信息，包括蛋白质组、代谢组、微生物组、分子影像组等多组学数据 [20-21]。队列研究的暴露也由传统的危险因素拓展为行为、认知、分子影像等多维度特征；采集的生物样本也更加全面和多样化；收集的表型、信息和样本种类也逐渐丰富和精细化，而便捷、个性化的数据采集方法 (如可穿戴设备) 也为跨时间和空间的表型信息收集提供了有利途径。这些多维、动态信息的收集将更好地帮助队列研究从多角度、全方位解析生命和健康的发展规律 [22]。

1.2 多学科交叉、创新型团队的组建

大数据时代对大型人群队列研究提出了更多要

求，不但需要研究人员具备扎实的理论知识和丰富的现场调查及数据分析的实践经验，还应该充分调动流行病学、表型组学、生命科学、计算生物学等多学科优势，组建多学科、多领域融合的创新团队。他们可以充分利用大数据技术对队列资源数据进行深入探索，对不同类型的组学数据进行高度整合，构建全面完善的数据分析模型，开展有价值的分析、判定和预测，帮助队列研究成果从概念走向实际应用，产生价值；通过挖掘队列数据产生新的信息知识，服务于精准医学，改善人类健康[23]。

1.3 注重合作与创新

大型人群队列研究的开展及其研究结果有助于精确追踪疾病的自然进展，其普及性和可推动性更好，而大数据则推动了各队列间更广泛的合作[24]。虽然单一队列规模已达到数十万级，但在罕见病研究和资源合理利用等方面仍显劣势，且人力、财力和时间成本高昂。目前队列间合作已成为趋势，如国际多中心人群队列 The Prospective Urban Rural Epidemiology (PURE，20.2 万) 的调查对象涵盖了 27 个国家的城市和农村社区[25]。我国也搭建了一些健康医疗大数据平台，用于队列间合作共享，如国家生物信息中心和中国队列共享平台等[26-27]。多中心大型人群队列具有广泛的研究对象、大样本量和多监测点，对其精准和精细化管理有利于提高队列的整体质量，而利用大数据技术优势，充分整合队列间的数据和样本资源，也避免了各自为营所造成的资源浪费[28]。同时，队列研究人员以创新为基本出发点和核心发力点，立足前沿领域，促进队列研究跨学科的探索与协同创新，更有利于完善促进跨学科合作的队列间的成果共享机制。

2 大型人群队列研究的发展趋势

人群队列大数据在新时代扮演着重要的基础性战略资源角色，被视为国家生物医药领域的珍贵宝库。借助大数据技术辅助驱动我国公共卫生领域的发展，可将其更紧密地与人群健康、疾病预防和管理需求相对接，释放队列研究的巨大潜力。这将促进健康产业的蓬勃发展，以满足对疾病和健康管理的多层次、个性化需求，也是我国大型人群队列研究面临的新任务和使命。

2.1 数据收集个体化和精细化

队列人群信息的收集方式将不只局限于传统的现场调查，便携式可穿戴设备的普及以及多样化移动终端的信息收集渠道，满足了人群队列研究对个性化、多时点、动态信息收集的需求。大型人群队列研究将不只局限于以组学数据为基础的数据分析与挖掘，也将加强个性化数据的收集与汇总。例如，复旦大学依托于泰州队列构建了基于表型组学系统设计的特色子队列 (脑影像队列)[29]，将基因组学、宏基因组学、代谢组学等多种生命组学技术及表型测量评估体系引入队列调查，对调查对象进行精细的表型测量和生物样本采集，系统产生了多组学数据，并进行长期随访，有助于研究影响当地居民健康的多种重大疾病的生理病理变化趋势。投入对精细化表型信息的收集及多维度、精细化表型信息的整合将有利于更加准确地对人体健康和疾病影响因素进行表征，从而实现个性化的健康监测、评估与指导。

2.2 数据分析集成化

人群队列研究中的样本、数据类型较复杂，数据格式存在差异。因此，实现数据分析集成化的前提是确保各队列建设及数据收集的标准化，并开发出相应的数据模型研究方法。目前，跨队列分析常通过数据协调将相似的变量转换为通用格式，并创建“协调数据集”用以提高多队列研究的可比性[30]。通过对数据信息的不断完善建立起系统化、智能化的数据模型，实现对数据的集成化分析。目前数据模型的创建在多个队列研究中已经得到了重视，而数据分析的集成化也将会成为人群队列大数据的未来发展趋势之一。

2.3 建立大数据平台

队列大数据平台的建立有助于实现队列研究项目精细化的管理，促进队列资源的合理维护及管理利用，帮助实现队列资源的可持续增长。通过利用平台的数据采集、治理和管理，可以极大程度地优化队列工作模式，提高队列研究工作效率，使其工作内容有迹可循且更加规范化和系统化[6]。如复旦大学泰州健康科学研究院建立的“健康大数据共智平台”，可进行标准化表型数据的采集、融合、管理、分析与共享。建立队列人群大数据平台，对系统解析复杂疾病的病因结构，提供重大疾病风险评估和预测、早筛分类、个体化治疗及疗效监测的整套解决方案具有重要的意义。

3 大型人群队列研究面临的挑战

在大数据助力下，队列人群数据规模的不断扩大给队列研究带来了良好的发展契机，但人群队列大数据面对浩如烟海的数据信息，也将面临一定的

挑战和难题。

3.1 海量数据如何充分且高效地利用

大数据技术与人群队列研究的高度融合，推动了预防医学和生命科学的发展，为提高人类健康水平做出了巨大贡献。在队列研究的发展过程中，大量的研究数据正在源源不断地产生，如何充分利用大数据技术对数据进行探索和挖掘是队列研究面临的一项重大挑战。在实际应用中，不仅要求保证分析结果的准确性，还要有效保障时效性，从而准确地表征健康和疾病发生发展的规律。

3.2 跨学科、复合型人才缺乏

复合型、跨学科人才的缺乏是国内人群队列研究面临的共同难题。要全面发掘队列大数据的价值，具有专业背景知识的复合型人才必不可少。他们不但要掌握一定的健康或疾病背景知识，还需要熟练应用数据分析工具和计算机技术。这也催生了对于流行病学、生命科学、表型组学和计算机科学等交叉学科教育的需求。

3.3 队列间信息共享机制亟待完善

尽管跨队列合作共享已成为趋势，但由于各队列在研究设计、目的及信息收集方式上存在差异，以及数据存储格式不一致，导致资源共享面临重重阻碍。建立共享合作网络，支持跨队列互联互通是解决这一问题的关键环节。而 TB 级甚至 PB 级的庞大数据资源信息已超出现有个人计算机程序可处理的能力范围，队列研究也需要建立新的共享方式来攻克这一难题 [5]。同时，还需要建立多层次、立体化的合作策略和共享机制流程，为队列研究提供新的数据来源和合作渠道，帮助形成共享和共赢的大型人群队列数据研究体系 [31]。

总的来说，大型人群队列研究在精准医学和大数据时代中的发展前景十分广阔，基于大型人群队列开展的基因组研究、表型组研究等组学研究方兴未艾，在探索人群健康发展规律、疾病病因和寻找疾病标志物方面取得了重大突破，为精准医学和个性化的健康管理提供了丰富的资源和知识支持 [4]。伴随着大数据技术的发展和跨学科交叉的深化，大型人群队列研究正迎来新的发展特征和趋势，也面临着艰巨的挑战。

[参 考 文 献]

[1] 王笑峰, 金力. 大型人群队列研究. 中国科学: 生命科学, 2016, 46: 406-12

[2] 单广良. 京津冀自然人群队列研究的理念与实践. 中华流行病学杂志, 2021, 42: 1493-7

[3] 王慧, 陈培战, 张作文, 等. 我国人群队列研究的现状、机遇与挑战. 中华预防医学杂志, 2014, 48: 1016-21

[4] 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. 中华流行病学杂志, 2021, 42: 10-4

[5] 宋菁, 胡永华. 流行病学展望:医学大数据与精准医疗. 中华流行病学杂志, 2016, 37: 1164-8

[6] 陈兴栋, 蒋艳峰, 徐萍, 等. 大型人群队列遗传资源建设与利用. 遗传, 2021, 43: 980-7

[7] Jin L. Welcome to the phenomics journal. Phenomics, 2021, 1: 1-2

[8] Ying W. Phenomic studies on diseases: potential and challenges. Phenomics, 2023, 3: 285-99

[9] 杨瑞馥. 大数据时代的预防医学研究: 数字化预防医学. 中华预防医学杂志, 2014, 48: 161-3

[10] Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. Genet Med, 2017, 19: 743-50

[11] Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. Lancet, 2015, 386: 533-40

[12] 郭彧, 余灿清, 吕筠, 等. 大型自然人群队列示范研究进展与成果. 中华流行病学杂志, 2023, 44: 1-6

[13] Wang X, Lu M, Qian J, et al. Rationales, design and recruitment of the Taizhou Longitudinal Study. BMC Public Health, 2009, 9: 223

[14] Zheng W, Chow WH, Yang G, et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. Am J Epidemiol, 2005, 162: 1123-31

[15] Shu XO, Li H, Yang G, et al. Cohort Profile: The Shanghai Men's Health Study. Int J Epidemiol, 2015, 44: 810-8

[16] 胡志斌, 杜江波, 徐欣, 等. 中国国家出生队列建设背景和设计简介. 中华流行病学杂志, 2021, 42: 569-74

[17] 何美安, 张策, 朱江, 等. 东风-同济队列研究:研究方法及调查对象基线和第一次随访特征. 中华流行病学杂志, 2016, 37: 480-5

[18] 杨景丽, 黄文雅, 黄佩瑶, 等. 中国队列研究建立和发展现状. 中国公共卫生, 2019, 35: 1393-9

[19] Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. Nat Rev Genet, 2006, 7: 812-20

[20] Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med, 2020, 26: 29-38

[21] 祁子凡, 张凤旭, 张玲. 美国精准医学计划"All of Us"百万自然人群队列设计方案的经验和启示. 中国循证医学杂志, 2021, 21: 980-5

[22] 余灿清, 李立明. 大型队列研究中的数据科学. 中华流行病学杂志, 2019, 40: 1-4

[23] 王波, 吕筠, 李立明. 生物医学大数据: 现状与展望. 中华流行病学杂志, 2014, 35: 617-20

[24] 董文斌, 雷小平. 大数据时代出生队列研究的新趋势. 西部医学, 2015, 27: 641-4

[25] Walli-Attaei M, Joseph P, Rosengren A, et al. Variations

between women and men in risk factors, treatments, cardiovascular disease incidence, and death in 27 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. Lancet, 2020, 396: 97-109

[26] Sun Y, Pei Z, Zhao H, et al. Data resource profile: China Cohort Consortium (CCC). Int J Epidemiol, 2020, 49: 1436-m

[27] CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. Nucleic Acids Res, 2023, 51: D18-28

[28] 陈柯妗, 於一凡, 刘静, 等. 多中心大型人群队列全生命周期管理理论与实践探索. 现代预防医学, 2022, 49: 2317-9, 2334

[29] Jiang Y, Cui M, Tian W, et al. Lifestyle, multi-omics features, and preclinical dementia among Chinese: The Taizhou Imaging Study. Alzheimers Dement, 2021, 17: 18-28

[30] Broderick C, Christian N, Apfelbacher C, et al. The BIOMarkers in Atopic Dermatitis and Psoriasis (BIOMAP) glossary: developing a lingua franca to facilitate data harmonization and cross-cohort analyses. Br J Dermatol, 2021, 185: 1066-9

[31] 杨羽, 赵厚宇, 詹思延. 队列数据共享的必要性与可行性. 北京大学学报(医学版), 2018, 50: 381-5