中国科学: 生命科学 2024 年 第 54 卷 第 6 期: 1021 ~ 1028
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述 生物样本资源专辑

大型队列研究数据资源的信息化建设

卞铮[1,2], 王翊涵[1], 鄢全意[2], 罗丹[2], 李岩[1,2*]

1. 中国科学院生物物理研究所, 北京 100101;
2. 健康医疗大数据西部研究院, 重庆 401329
* 联系人, E-mail: yanli@ibp.ac.cn

收稿日期: 2023-05-12; 接受日期: 2023-08-21; 网络版发表日期: 2024-06-13

摘要 大型队列研究在医学研究中的作用凸显, 信息化建设作为保障大型队列研究高质量开展的重要基础, 日益受到关注和重视. 本文针对队列研究数据资源的信息化建设要点进行阐述, 归纳队列研究在采集、保藏、利用数据资源过程中的全流程信息化场景, 讨论并分析信息化面临的问题与挑战, 以期为我国未来的大型队列研究的发展及参与数字中国的建设提供重要借鉴与参考.

关键词 队列研究, 大型队列, 信息化建设

队列研究是观察某些特定暴露与一种或多种结局事件的关联性研究[1], 是公认研究环境和遗传危险因素与疾病结局关联的经典流行病学研究方法, 近20年来, 随着生物技术的提高, 队列研究的范畴已经从对人群的多种表型的测量, 快速扩展到囊括基因组学、蛋白组学及其他的多组学生物信息学领域[2], 由队列研究采集样本形成的生物样本库也越来越成为生物医学研究的宝贵资源[3], 在揭示疾病的病因、发生发展、预后机制,减轻疾病负担等方面发挥巨大的作用, 成为精准医学研究的新一代推动力.

复杂性疾病的发病机制受遗传、环境暴露、生活方式等多种因素的交互作用影响[4], 小样本队列在病例病种积累、人群多样性、暴露因素收集及基因-环境交互作用等方面的效能存在不足, 因此, 各国均在积极开展具有本国特色的多民族、可持续性的大型队列研究[5]. 近半个世纪以来, 基于复杂性疾病病因学研究的大型队列研究不断建立并得以开展, 如欧洲癌症与营养前瞻性调查[6](European Prospective Investigation into Cancer and Nutrition, EPIC)、英国生物银行(UK Biobank)[7]和中国慢性病前瞻性研究(China Kadoorie Biobank, CKB)研究[8]. 这些大型队列研究的规模不仅提高统计效能, 还可弥补罕见病研究中病例数不足的劣势. 因此, 大型队列研究由于其大样本、长期随访、暴露因素收集较全、观察结局多样性等优点, 超出单纯遗传因素或环境因素的病因学研究, 在分析基因-环境因素与复杂性疾病交互作用的病因学探索中具有独特优势[9].

随着我国对自然人群健康状况及自然人群队列的关注度越来越高, 仅2016、2017年立项的国家重点研发计划就涉及20余个自然人群及疾病队列, 包括国家级大型自然人群健康队列、重大疾病专病队列和罕见病的临床队列共覆盖百万人, 队列建设规模、数量、

引用格式: 卞铮, 王翊涵, 鄢全意, 等. 大型队列研究数据资源的信息化建设. 中国科学: 生命科学, 2024, 54: 1021–1028
Bian Z, Wang Y H, Yan Q Y, et al. Information construction of data resources in mega cohort study (in Chinese). Sci Sin Vitae, 2024, 54: 1021–1028, doi: 10.1360/SSV-2023-0042

速度不断攀升[10]. 同时, 多组学和大数据技术的快速发展也在不断推动人群队列研究模式的变革. 大型队列产出的各类组学数据海量增长, 以及研究中收集的视频、图像、音频、可移动设备结果等多媒体数据的亟待处理, 促使大型队列利用新型信息化手段采集、管理、分析、使用数据成为一种必然.

本文将从队列信息化建设的基本内容和工作范畴出发, 围绕大型人群队列开展过程涉及的数据资源, 探讨信息化的应用场景及可能存在的问题, 为推动大型人群队列的数据科学化管理及信息化建设提供参考.

1 信息化建设的现状

医学研究涉及的数据繁多、关系复杂且存在大量的自然语言, 不利于后续分析, 但这些数据对于健康研究的质量至关重要. 信息化可以帮助医学研究人员更好地获取、分析和共享这些数据, 以便更准确地研究和理解人类疾病的本质[11,12]. 队列研究作为医学研究的重要组成部分, 其信息化建设必须依靠全社会、全行业的信息技术发展作为基础, 不可孤立存在.

进入21世纪以来, 信息化推动各行业的改革, 全民健康信息化是国家信息化建设的重要组成部分, 也是深化医药卫生体制改革、建设健康中国的重要支撑. 2016年6月, 国务院办公厅印发《关于促进和规范健康医疗大数据应用发展的指导意见》, 明确互联网+、大数据、云计算与医疗卫生行业的主要方向和重点区域, 2017年1月, 国家卫生计生委发布的《“十三五”全国人口健康信息化发展规划》, 将逐步形成以人口健康信息化和健康医疗大数据为基础, 有序推动人口健康信息基础资源大数据开放共享[13], 贯彻执行《关于促进和规范健康医疗大数据应用发展的指导意见》[14]和《关于促进“互联网 +医疗健康”发展的意见》[15], 加快推进医疗信息化建设.

“十三五”期间, 国家、省、市、县四级全民健康信息平台联通全覆盖, 实现公共卫生、医疗服务等六大业务应用的横纵联动, 有效整合和共享人口信息、电子健康档案和电子病历三大数据库资源, 为卫生行业的信息化发展提供有力支持[16]. “十三五”期间我国全民健康信息化建设成效显著, 但是在基础设施、共享应用、投入保障、网络安全等方面还存在短板与弱项. 2022年11月, 国家卫健委发布的《“十四五”全民健康信息化规划》[17], 着重指出应增强全民健康信息化发展的系统性、整体性和协调性, 以构建大平台、大系统、大目录为导向; 充分发挥新一代信息技术的优势, 构建基于数据驱动的生态系统, 强化区域数据汇聚应用, 推进跨部门、跨地域、跨层级、跨系统、跨业务的技术融合、数据融合、业务融合; 创新数据供给方式, 深化数据开发利用, 推进健康医疗数据资源和基础设施开放共享, 不断提高卫生健康行业治理水平.

大型队列研究由于其规模、跨度、研究范畴的要求, 是医学研究中信息化程度较高的一类研究, 在新时代信息化浪潮的推动下, 也秉承着系统平台、汇聚融合、开放共享的发展目标, 以共同构建全民健康信息化.

2 队列数据资源与信息化

队列数据资源来源于各种领域[18,19], 如因研究自发建立的测量系统、表型测量、电子健康记录、各类健康和疾病的登记系统、基因组、表观基因组、转录组、蛋白质组、代谢组和微生物测量的多维数据, 以及医学成像. 近年来, 还陆续纳入社交媒体、社会经济或行为指标、职业信息、移动应用程序或环境监测的数据[20].

大型队列由于其涉及数十万量级的样本量, 包括表型数据、组学数据、影像学数据在内的多种数据形式, 其数据具有大数据4V的部分特点, 即大数据容量大(volumn)、多样性(variety)及真实性(veracity)的特点, 为下一步的精准医学研究提供丰富的数据资源[21], 因此大型队列数据符合欧盟委员会研究与创新总局卫生局提出的健康大数据定义: “健康大数据包括在一个或多个时间点从单个个体收集到的与健康和健康状况相关的大量多样性的生物、临床、环境和生活方式信息”[22].

队列数据可以通过对相关健康医疗大数据领域的利用, 进行流行病病因学的深入探讨, 用于临床辅助诊断、治疗与决策, 提供相关科学证据用于卫生政策制定, 最终致力于人类的健康[23]. 队列研究中的数据价值, 多以数据科学的学科角度, 分析在大型队列研究中数据全生命周期下的收集、维护、处理、分析和交流(https://datascience.berkeley.edu/about/what-is-data-

science/), 但数据资源必须借助于计算机科学的信息化手段, 将研究中的各项数据资源要素进行汇总, 通过采集数据、管理数据、分析数据, 并结合医学统计学和生物信息学方法, 搭建安全、稳定、可操作的生物医学大数据平台, 通过信息交流和知识共享, 方能构建多层次精准医学知识库体系, 为推动社会发展进步提供技术支持[24].

因此, 信息化对于队列研究不仅是重要的技术手段, 也是医学行业变革的推动力, 是进入信息社会国家发展的重要组成[25].

3　信息化建设的应用

如何利用和发展信息化技术, 已经成为队列研究乃至各类医学研究开展前需要考量的首要大事, 但建成、用好信息系统需要研究管理者和工程人员的探索和实践, 秉承高标准数据收集质量及规范化管理要求是信息化建设的重要原则. 大型队列时间周期长、资源消耗多、信息体量大、专业技术强, 利用现有软件用于支持研究的可行性不大, 需开发一系列定制的信息系统同时加以集成, 以适应于研究的各个维度, 信息系统的设计难度远高于其他医学研究. 因此, 设计初期必须将信息化体系构架列为首要工作, 从设计层面出发, 保障数据的一致性、可追溯性、及时性和准确性. 由于其多中心共同参与研究, 信息系统不仅覆盖数据资源收集, 还应纳入各分中心工作内容、人员、设备、物资等管理性要素, 并实施质量控制监测和研究进展报告. 目前各国大型队列研究在采集、保藏、利用数据资源的全流程信息化应用场景如下.

3.1　电子知情同意书签署

作为人群队列招募的第一步, 信息化技术为大型人群调查的知情同意书的无纸化保存提供有力保障, 2004年我国公布《电子签名法》, 明确可靠的电子签名与手写签名具备同等的法律效力. 在大型队列的现场调查采集中, 工作人员使用身份证认证设备识别信息, 通过程序预置知情同意书模板, 在对研究对象进行充分告知后, 收集手写电子签名或可视化的数字指纹, 并联合电子认证机构, 为用户颁发数字证书, 可确保签署主体的身份真实有效. 同时, 通过区块链、时间戳等技术手段, 在电子知情同意书签署、存储过程中, 对数据文件进行上链存储, 精确记录签署时间信息, 确保原始文件不被篡改, 有效保证电子知情同意书签署的完整性和有效性[26,27]. 加州大学洛杉矶分校创新性使用电子知情同意视频进行招募工具, 研究对象观看后可以自主完成同意流程, 并于电子病历系统内推送采血管或取样管订单在后续的采血和手术中使用, 这是一种可扩展且高效的技术方案, 在大型人群研究中可用于招募[28].

3.2　问卷采集及体格检查

传统问卷收集方式是通过纸质问卷面对面访谈的形式来收集研究对象的数据, 这种收集方式不仅效率低, 需要双人二次录入, 而且问卷易丢失、不易管理, 逻辑性差. 近20年以来, 随着信息技术的发展, 医学研究已经能够借助计算机软件和数据库技术安全方便地收集和存储数据, 计算机辅助面访的电子问卷、网页端或移动设备端的自填式电子问卷, 已经成为大型人群队列开展的通用标准. 利用电子设备和信息化技术, 对队列研究现场调查中体格检查、影像检查、生理等数据, 通过计算机端口直接将数据采集到结构化数据库中, 保证大型队列研究数据的高效采集. 近年来, 数据采集范围还在发生快速变化, 可穿戴设备、集成传感器和连续监测功能越来越多的应用在队列研究的各种测量尺度[29]. 信息化手段使得各类移动端健康应用的数据接入变得更加简单易行, 这样可以更密切地与研究对象接触, 收集更详细的临床、环境和生活方式信息, 如心率和体温、身体活动和营养习惯以及睡眠和压力管理, 从而进一步分析风险暴露和疾病发作的关系[30]. 例如CKB、美国全民队列项目(All of US Research Program)、UK Biobank等大型队列, 均采取引入如电子手环、手机端应用程序等可穿戴设备的措施, 使研究可以采集研究对象的数字化健康数据, 联通多系统互作[31,32].

3.3　长期随访监测的信息化

大型队列的随访方法受其调查规模影响, 通过接触研究对象而获得自报、体检、生物样本检测等结局信息的定向随访方式不易大规模开展, 而通过个体身份识别信息(如姓名、性别、地区和身份证号码等)匹配各类监测或常规信息系统, 以获取结局信息的随访方式, 越来越成为主流趋势[33]. 此种随访方式可通过

居民死亡登记系统、公安户籍系统、民政殡葬系统等可获取死亡相关信息; 通过疾控系统的肿瘤及心脑血管监测网、医疗保险系统、医院病案信息系统、体检系统等可获取疾病诊疗信息[34]. 信息化技术的不断提升有效保障这一手段的安全实施, 通过接入方的数据接口, 使用身份识别信息作为入口参数, 加密算法对传输数据进行去标识化传输, 能够保障数据的不间断采集及处理, 以方便研究者及时进行数据分析和挖掘[35]. 如UK Biobank通过链接英国国民医疗服务系统, 获得研究对象的初级卫生保健信息、用药、疾病死亡结局等. 除外常规的结局事件获取渠道, 越来越多的研究也认识到利用电子病历及电子健康记录的重要性, 它不仅可以支持科研人员收集来自临床的原始数据, 还可以将这些原始数据转化为新的数据和信息, 加入到研究中[36].

3.4 生物样本库信息管理

建立生物样本库信息化管理系统, 实现样本数字化管理和应用是大型队列信息化建设的关键一环. 生物样本库资源管理信息系统应具备灵活性和可扩展性、样本识别和追踪能力, 能根据生物样本库不断更新和变化的需求,扩展其功能和信息采集内容, 提升实验数据分析、挖掘能力[37], 包括: 采集、分装后的样本应能通过信息化管理系统追溯到原始样本并与其信息进行关联; 每份样本在信息化管理系统中应有唯一识别符号(字符编号或条形码); 样本从采集到处理、分装、运输、储存、使用后剩余样本处理等全过程都应被准确记录; 生物样本的每一次转运或位置移动都应被及时记录,信息化管理系统要能追溯到每一例样本储存位置的变更. 同时, 样本库信息化管理系统还应具有管理质量和质量控制程序及文件, 数据安全保护, 可生成样本采集、存储、使用、销毁、质量检测等管理报告, 实验室数据挖掘、分析、共享等功能. UK Biobank在样本处理中设计并使用实验室信息管理系统, 使用高速摄像对样本进行识别及分装, 保证记录样本信息、处理样本、批量指控程序中均高度自动化, 提高数据的准确性和一致性[38].

3.5 生物大数据云计算平台

大型人群队列的数据资源是流行病学结合生命科学的多维数据, 已经成为数据密集型科学. 在信息系统设计中, 结合队列的研究特点, 通过数据库模型方法形成纵向分析的数据管理, 允许对异构数据集进行灵活的纵向分析, 以实现数据集的持久化、历史化存储[39]. 考虑到生物医学研究数据的多样性、复杂性和不断增加的数量, 可以利用基于云的平台来支持各种分析模式(例如机器学习(machine learning)和深度学习(deep learning))[40], 并将数据的主副备份设计存储在两个不同的地理位置. 未来, 基于云的数据归档平台可以为管理研究数据生命周期提供动态环境, 并提供长期保存生物医学数据的能力. 大型队列数据分析所依托的云计算平台还需要可扩展性, 以支持数据摄取模式的多样性和复杂性(例如机器、软件或人工输入模式). 通过共享软件程序(例如生物信息学工具), 安全的虚拟工作空间可以集成和操作数据, 可以促进数据的“FAIR”原则: 可查找(findable)、可访问(accessible)、可互操作(interoperable)和可重用(reusable), 以满足近期和长期研究需求[41].

欧美国家注重系统使用国家级生物资源, 建立研究成果的生物分子数据库, 进行以国家级大队列为基础的大科学研究, 提供衍生生物样本信息库供研究人员查询, 从而形成重要的医学基础设施. 欧洲项目p-medicine, 创建IT基础设施, 以促进转化研究和个性化医学的发展[42]. UK Biobank于2021年9月启用基于DNAnexus提供的云端在线分析平台RAP(research analysis platform), 共享数据量超过6.2 PB, 极大的提高分析的效率与能力. 2022年3月, 美国的All of Us在云平台公布首批近10万人的全基因组数据及表型数据, 2022年, 国内的泰州队列初步完成“大型队列健康大数据平台”构建. 这些借助新型信息化技术的大型队列, 均在力图搭建高效统一的资源管理和共享服务体系, 以提升队列数据资源价值的转化力.

4 信息化建设的问题与挑战

自2016年国务院办公厅提出《关于促进和规范健康医疗大数据应用发展的指导意见》以来, 互联网与各领域的融合发展呈现出广阔前景和无限潜力, 信息化数字化已成为不可阻挡的时代潮流. 但是, 队列研究的信息化仍面临着很多的问题与挑战.

4.1　队列研究信息化建设需要顶层统筹设计

纵观国内外科研发展趋势, 队列研究这一研究形式呈现加速发展的态势, 但是由于医学研究的特殊性, 加之各研究机构对研究目的的各有不同, 导致这一开展时间长、投入耗费大的研究缺乏国家层面的引导, 缺少长期发展规划. 目前队列研究的信息化发展呈现“多小散乱”的局面, 重复建设现象严重, 难以形成有效的长期发展战略, 无益于助力数字中国的建设, 其发展方向需要国家层面的积极引导和推动.

4.2　信息化技术面临着技术创新和人才培养挑战

大型队列是一种基于数据驱动方法的新兴医学实践, 该方法考虑个体的相关医学、遗传、行为和环境信息, 通过将不同的数据集链接在一起, 揭示迄今为止未知的随机路径和相关性, 医学大数据比以往任何时候都更加精准[43]. 高通量、高分辨率数据生成技术的最新科学进展使得能够对个人健康的大数据集进行经济高效的分析. 为分析和整合如此大的信息, 需要新的计算方法, 如更快、更集成的处理器、更大的计算机存储器、改进的传感器、更复杂的新算法、方法和云计算[44], 这些方法可以通过提供有用的信息来指导未来的应用实践.

同时, 制约研究发展重要的瓶颈还包括缺乏具备生物信息学及医学方面的专业人士, 以能够使用大数据分析领域开发的最新信息技术, 处理大型队列所产生的数据[45].

4.3　数据治理是信息化建设的重点工程

队列研究的成败取决于数据库中收集到并可以利用的个体暴露数据、随访数据、临床数据及组学数据[46]. 队列数据的数据挖掘问题具有一些独特的特点, 主要困难与所涉及数据的复杂性质(异构、分层、时间序列)、缺失值导致的质量和数量差异有关, 这些特点使其与其他领域不同, 并使其更难解决.

目前几乎所有的队列研究使用的都是自行创建系统进行初始数据的收集, 同时利用已有的各医疗卫生系统进行结构化数据的收集, 但是医疗行业存在大量半结构化、非结构化的文字、影像数据, 例如病历中的病程记录和治疗记录, 这些资源对于研究疾病转归和预后具有重大意义, 但由于这些数据的结构差异, 导致无法进行利用[47].

为最大限度地利用所产生的信息, 应通过数据治理解决技术挑战, 将基因型、表型的结构化数据与半结构化和非结构化数据(例如医学成像、电子病历、生活方式、环境和健康经济数据)相结合用于科学研究.

4.4　积极推进队列研究的数据标准化工作

国家对健康医疗领域已经制定300余项标准, 但大型队列的组建形式多样, 相比于集中型队列, 由各个分中心组成的分散式队列数据标准不一致的问题突出, 数据标准化与否严重影响着后期的有效利用. 由于队列人群各具特色, 暴露和结局的操作方法、测量方法、参考标准、实施标准等不统一, 使得各队列仍相对分散, 数据结构各自独立, 不能直接进行合并使用, 限制队列共建的成效和研究效率. 在标准化过程中, 很难要求延续的队列放弃以往使用的操作流程和标准而改换用统一规定的操作流程和标准, 因此建立队列研究领域的行业标准和数据治理模型势在必行[48,49], 泰州队列针对大型人群队列常规调查的共性问题制定《大型人群队列数据集标准》[50], 也可根据已有数据情况制定队列研究最小数据集标准[51]. 队列研究还应该向其他健康领域, 如基因组标准联盟(2016年)、临床数据交换标准联盟以及ISO国际标准委员会(International Organization for Standardization)的标准化范例(如ISO TC276 WG5, 2016)进行学习借鉴.

4.5　打通信息孤岛是队列研究信息化成功的关键

信息孤岛是硬件平台、操作系统、数据库、数据标准、应用系统之间的异构所导致的计算信息系统不能互相交换. 在队列研究中, 信息来源系统由现场调查、生物样本库、体格检查、医疗卫生系统等多个软件功能所组成, 不同的服务商的底层设计逻辑不同, 核心代码不开放, 加剧信息孤岛的解决难度. 尤其涉及到队列随访阶段, 研究需要从卫健委、医疗保险等各个政务部门获得随访事件, 研究对象在不同医院就诊产生的信息隔离, 且各系统间的分散管理产生的信息孤岛, 对研究的开展造成极大的阻碍[52].

为解决信息孤岛, 创建数字化社会, 近年来各地方政府相应设立省级大数据管理局. 大数据局整合原分散于地方多个部门的相关职能, 为构建统一高效的地方政

府数据治理体系探索有益经验[53]. 省级大数据局在省域内发挥顶层设计、总体布局、统筹协调和整体推进作用, 有利于建立标准统一、上下协同、运行高效的数据治理组织体系, 有助于解决“信息孤岛”的问题.

5 信息化建设总结与展望

2023年2月, 中共中央、国务院印发的《数字中国建设整体布局规划》(http://www.gov.cn/xinwen/2023-02/27/content_5743484.htm)指出, 要畅通数据资源大循环, 构建国家数据管理体制机制, 建设数字中国, 既是国家战略, 也是未来社会和经济发展的必然. 对任何一个行业和机构而言, 都需要积极参与数字中国建设, 也需要把握相关领域的发展机遇.

在我国全面迈向数字中国的进程中, 大型队列研究作为精准医学的推动力, 信息化建设对于支撑队列研究建设、运行、管理, 规范化全流程开展采集、保藏、利用数据资源具有极为重要的意义及作用. 大型队列研究应着眼长期, 做好信息化规划, 直面问题与不足, 根据研究目标需求, 协调好现状与未来变化的关系以及资源和技术之间的关系. 在目前信息新技术层出不穷的时代, 大型队列研究的管理者应结合近年出台的《中华人民共和国数据安全法》及《中国人民共和国个人信息保护法》, 在遵守数据隐私安全与伦理保护对生物医学要求的前提下, 着眼于打造未来生物医学大数据应用及流通平台, 动态微调信息化建设战略方向, 继续保持信息系统的核心能力, 在更加健全的法律制度保障下, 助力健康领域数据资源资产化、数据资产知识化、知识成果商品化, 推进医学研究决策科学化, 参与健康中国、数字中国的整体建设.

参考文献

1 Wang X F, Jin L. Large population-based cohort studies. Sci Sin Vitae, 2016, 46: 406–412 [王笑峰,金力.大型人群队列研究. 中国科学:生命科学, 2016, 46: 406–412]

2 De Souza Y G, Greenspan J S. Biobanking past, present and future: responsibilities and benefits. AIDS, 2013, 27: 303–312

3 Dong E D, Hu H, Yu W H. A fundamental role of biobank in biomedical research. Sci Sin Vitae, 2015, 45: 359–370 [董尔丹, 胡海, 俞文华. 生物样本库是生物医学研究的重要基础. 中国科学: 生命科学, 2015, 45: 359–370]

4 Chakravarti A, Little P. Feature nature, nurture and human disease. Nature, 2003: 412–413

5 Willett W C, Blot W J, Colditz G A, et al. Merging and emerging cohorts: not worth the wait. Nature, 2007, 445: 257–258

6 Gonzalez C A. The European Prospective Investigation into Cancer and Nutrition (EPIC). Public Health Nutr, 2006, 9: 124–126

7 Collins R. What makes UK Biobank special? Lancet, 2012, 379: 1173–1174

8 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. Int J Epidemiol, 2011, 40: 1652–1666

9 Sun D J Y, Lv J, Li L M. Mega cohort: a powerful tool for etiologic research on complex human diseases in 21st Century (in Chinese). Chin J Dis Control Prev, 2013, 17: 66–71 [孙点剑一 , 吕筠 , 李立明 . 流行病学超大规模队列研究: 开启 21世纪人类复杂性疾病病因研究的钥匙 . 中华疾病控制杂志, 2013, 17: 66–71]

10 Yang J L, Huang W Y, Huang P Y, et al. Established and on-going cohort srudies in China: a literature study (in Chinese). Chin J Public Health, 2019, 35: 7 [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: 7]

11 Liu N, Chen M. Applying themes and related data sources research of healthcare big data (in Chinese). Chin Med, 2016, 11: 6–9 [刘宁, 陈敏. 医疗健康大数据应用主题及相关数据来源研究. 中国数字医学, 2016, 11: 6–79]

12 Dhar V. Data science and prediction. SSRN J, 2012, doi: 10.2139/ssrn.2086734

13 National Health and Family Planning Commission. Notice of the national health and family planning commission on issuing the national population health informatization development plan for the 13th Five Year Plan (in Chinese). ZHONGHUA RENMIN GONGHEGUO GUOJIA WEISHENG HE JIHUASHENGYU WEIYUANHUI GONGBAO, 2017, 36–43 [国家卫生和计划生育委员会. 关于印发“十三五”全国人口健康信息化发展规划的通知. 中华人民共和国国家卫生和计划生育委员会公报, 2017, 36–43]

14 General Office of the State Council. Guiding opinions of the general office of the state council on promoting and standardizing the development of health and medical big data applications (in Chinese), ZHONGHUA RENMIN GONGHEGUO GUOWUYUAN GONGBAO, 2016. 24–28 [国务院办公厅. 关于促进和规范健康医疗大数据应用发展的指导意见, 中华人民共和国国务院公报, 2016. 24–28]

15 General Office of the State Council. Opinions of the General Office of the State Council on Promoting the Development of “Internet plus Medical Health” (in Chinese). ZHONGHUA RENMIN GONGHEGUO GUOWUYUAN GONGBAO, 2018. 9–13 [国务院办公厅. 国务院办公厅关于促进“互联网+医疗健康”发展的意见. 中华人民共和国国务院公报, 2018. 9–13]
16 Jin X T. Health and Healthcare Big Data (in Chinese). Beijing: People’s Medical Publishing House, 2017 [金小桃. 健康医疗大数据. 北京: 人民卫生出版社, 2017]
17 National Health Commission of the people’s epublic of China. Notice on issuing the national health informatization plan for the 14th Five Year Plan (in Chinese). ZHONGHUA RENMIN GONGHEGUO GUOJIA WEISHENG JIANKANG WEIYUANHUI GONGBAO, 2022. 7–18[国家卫生健康委员会. 关于印发“十四五”全民健康信息化规划的通知. 中华人民共和国国家卫生健康委员会公报, 2022. 7–18]
18 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst, 2014, 2: 3
19 Baro E, Degoul S, Beuscart R, et al. Toward a literature-driven definition of big data in healthcare. Biomed Res Int, 2015, 2015: 1–9
20 Fernández-Luque L, Bau T. Health and social media: perfect storm of information. Healthc Inform Res, 2015, 21: 67
21 Yu C Q, Li L M. Data science in large cohort studies (in Chinese). Chin J Epidemiol, 2019, 40: 1–4 [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]
22 Auffray C, Balling R, Barroso I, et al. Making sense of big data in health research: towards an eu action plan. Genome Med. 2016, 8: 71
23 Borges do Nascimento I J, Marcolino M S, Abdulazeem H M, et al. Impact of big data analytics on people’s health: overview of systematic reviews and recommendations for future studies. J Med Int Res. 2021, 23: e27275
24 General Office of the CPC Central Committee, General Office of the State Council. Outline of the National Informatization Development Strategy (in Chinese). ZHONGHUA RENMIN GONGHEGUO GUOWUYUAN GONGBAO, 2016. 6–16 [中共中央办公厅、国务院办公厅印发《国家信息化发展战略纲要》. 中华人民共和国国务院公报, 2016. 6–16]
25 Yang Q F. Information 2.0+: the information system in the cloud computing era (in Chinese). Beijing: Publishing House of Electronics Industry, 2013 [杨青峰. 信息化2.0+: 云计算时代的信息化体系. 北京: 电子工业出版社, 2013]
26 Zheng J N. The application and regulation of the informed consent principle in information collection (in Chinese). Oriental Law, 2020, 74: 198–208 [郑佳宁. 知情同意原则在信息采集中的适用与规则构建. 东方法学, 2020, 74: 198–208]
27 Peng Y, Shi L. Application of electronic signature technology in hospital information management (in Chinese). Chin Med Equipment J, 2019, 40: 36–39+55 [彭滢, 石磊. 电子签名技术在医院信息管理中的应用. 医疗卫生装备, 2019, 40: 36–39+55]
28 Lajonchere C, Naeim A, Dry S, et al. An integrated, scalable, electronic video consent process to power precision health research: large, population-based, cohort implementation and scalability study. J Med Internet Res, 2021, 23: e31121
29 Yaman H, Yavuz E, Er A, et al. The use of mobile smart devices and medical apps in the family practice setting. J Eval Clin Pract, 2016, 22: 290–296
30 Zheng Y L, Ding X R, Poon C C Y, et al. Unobtrusive sensing and wearable devices for health informatics. IEEE Trans Biomed Eng, 2014, 61: 1538–1554
31 Li W, Sun X H, Xu P, et al. Analysis of All of Us research program’s construction model and characteristics (in Chinese). World Sci Tech R & D, 2022, 44: 265–274 [李伟, 孙学会, 徐萍, 等. 美国All of Us队列项目建设模式与特点分析. 世界科技研究与发展, 2022, 44: 265–274]
32 Chen Z M. Population Biobank Studies: A Practical Guide. Berlin: Springer, 2021
33 Lv J, Li L M. Follow up in cohort study (in Chinese). Chin J Dis Control Prev, 2019, 23: 373–375 [吕筠, 李立明. 队列研究随访之我见. 中华疾病控制杂志, 2019, 23: 373–375]
34 Chinese Preventive Medicine Association. Technical specification of long-term follow-up for end point in large population-based cohort study (T/CPMA 002-2019) (in Chinese). Chin J Epidemiol, 2019, 40: 748–752 [中华预防医学会. 大型人群队列终点事件长期随访技术规范(T/CPMA 002-2019). 中华流行病杂志, 2019, 40: 748–752]
35 Miller A R, Tucker C. Health information exchange, system size and information silos. J Health Econ, 2014, 33: 28–42
36 Poba-Nzaou P, Uwizeyemungu S, Dakouo M, et al. Patterns of health information exchange strategies underlying health information technologies capabilities building. Health Syst, 2022, 11: 211–231
37 Wang W Y, Zhou J M, Cai Z Z. Biobank information management and sample annotation for usability (in Chinese). Chin J Clin Lab Manage (Electronic Edition), 2017, 5: 24–29 [王伟业, 周君梅, 蔡珍珍. 生物样本库的信息化管理与信息应用. 中华临床实验室管理电子杂志, 2017, 5: 24–29]
38 Elliott P, Peakman T C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and

urine. Int J Epidemiol, 2008, 37: 234–244
39 Weiß J P, Hübner U, Rauch J, et al. Implementing a data management platform for longitudinal health research. Stud Health Technol Inf. 2017, 243: 85–89
40 Navale V, Bourne P E. Cloud computing applications for biomedical science: a perspective. PLoS Comput Biol, 2018, 14: e1006144
41 Navale V, McAuliffe M. Long-term preservation of biomedical research data. F1000Res, 2018, 7: 1353
42 Christ-Neumann M L, Escrich A, Anguita A, et al. Usability on the p-medicine infrastructure: an extended usability concept. Ecancermedicalscience. 2014, 8: 399
43 Espinal-Enríquez J, Mejía-Pedroza R, Hernández-Lemus E. Computational approaches in precision medicine. In: Mukesh Verma, Debmalya Barh, eds. Progress and Challenges in Precision Medicine. Amsterdam: Elsevier. 2017, 233–250
44 Mamoshina P, Vieira A, Putin E, et al. Applications of deep learning in biomedicine. Mol Pharm, 2016, 13: 1445–1454
45 Camacho D M, Collins K M, Powers R K, et al. Next-generation machine learning for biological networks. Cell, 2018, 173: 1581–1592
46 Kaiser J. Population databases boom, from iceland to the U.S. Science, 2002, 298: 1158–1161
47 Borangíu T, Purcarea V. The future of healthcare—information based medicine. J Med Life. 2008, 1: 233–237
48 Yang W Z. Promoting the establishment of group standards in public health areas for China (in Chinese). Chin J Epidemiol, 2019, 40: 2 [杨维中. 推动我国公共卫生领域的团体标准建设. 中华流行病学杂志, 2019, 40: 2]
49 Shan G L. Principles and practice on cohort study of general population in Beijing, Tianjin and Hebei province (in Chinese). Chin J Epidemiol, 2021, 42: 1493–1497 [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: 1493–1497]
50 Chen X D, Jiang Y F, Xu P, et al. Construction and utilization of human genetic resources in large population cohorts (in Chinese). Hereditas (Beijing), 2021, 43: 980–987 [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: 980–987]
51 Liu S W, Zhang P, Li H, et al. Minimum dataset standard for cohort study of high-risk population of stroke based on regional big data platform: a consensus (in Chinese). Prev Med, 2021, 33: 1189–1198 [刘世炜, 张培, 李辉, 等. 基于区域大数据平台的卒中高危人群队列研究最小数据集标准专家共识. 预防医学, 2021, 33: 1189–1198]
52 Qu Y M, Jiang Y. The sources and application of big data in healthcare (in Chinese). Chin J Epidemiol, 2015, 36: 1181–1184 [曲翌敏, 江宇. 健康大数据的来源与应用. 中华流行病学杂志, 2015, 36: 1181–1184]
53 Zhang K. Institutional setting and functional allocation of provincial big data bureaus: an empirical analysis based on the new round of institutional reform (in Chinese). Electron Gov, 2019, 198: 113–120 [张克. 省级大数据局的机构设置与职能配置: 基于新一轮机构改革的实证分析. 电子政务, 2019, 198: 113–120]

Information construction of data resources in mega cohort study

BIAN Zheng[1,2], WANG YiHan[1], YAN QuanYi[2], LUO Dan[2] & LI Yan[1,2]

1 Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;
2 Western Institute of Health Science, Chongqing 401329, China

Mega cohort (short for MC) research plays a prominent role in medical research. Information construction of MC, which is the basis to ensure the quality, has a higher position and grasps more eyesight. This paper expounds the key points of informatization construction of cohort study data resources, summarizes the whole process informatization scenarios of cohort study in data collection, storage and utilization, and discusses and analyzes the problems and challenges faced by informatization. This paper will provide a reference or consult for the future development of MC research in China, as well as the participation in the construction of Digital China.

cohort study, mega cohort, information construction

doi: 10.1360/SSV-2023-0042