·临床生物样本库·

多中心结直肠癌临床研究生物样本库信息系统的建设与管理

孔伟名[1]，吕文文[2]，姜嘉媛[2]，冯铁男[2]，俞章盛[1,2]

[1] 上海交通大学转化医学研究院，上海 200240
[2] 上海交通大学医学院临床研究中心，上海 200025

通信作者：俞章盛，E-mail：yuzhangsheng@sjtu.edu.cn

【摘要】近年来，我国结直肠癌的发病率显著升高。随着对结直肠癌复发、转移机制的深入探索，可提供全面生物医学信息的生物样本库信息系统显得尤为重要。本文阐述了采用微服务架构构建包含临床电子数据采集系统、生物样本管理系统和生物信息数据平台三个子系统的多中心结直肠癌临床研究生物样本库信息系统，同时介绍各子系统在多中心临床研究中的管理模式和特色功能；并通过统一化编码规则、结构化临床信息、标准化样本流转信息和规范化数据存储等数据标准化管理方案，保障数据质量和互联互通，从而为推进转化医学研究和精准医疗发展提供全方位生物医学信息。

【关键词】多中心；临床研究；生物样本库；信息系统
【中图分类号】R446.1；R735.3　　【文献标志码】A　　【文章编号】1674-9081(2022)04-0664-06
DOI：10.12290/xhyxzz.2021-0763

Construction and Management of Biobank Information System for Multi-center Colorectal Cancer Clinical Research

KONG Weiming[1], LYU Wenwen[2], JIANG Jiayuan[2], FENG Tienan[2], YU Zhangsheng[1,2]

[1]Institute of Translation Medicine, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

Corresponding author: YU Zhangsheng, E-mail: yuzhangsheng@sjtu.edu.cn

【**Abstract**】In recent years, the incidence of colorectal cancer in China has increased significantly. With the in-depth exploration of the mechanisms of colorectal cancer recurrence and metastasis, a biobank information system that can provide comprehensive biological information is becoming more and more important. This paper expounds on the construction of a multi-center colorectal cancer clinical research biobank information system that includes three subsystems as follows: a clinical electronic data acquisition system, a biological sample management system, and a biological information data platform using the micro-service architecture. This paper explains how these subsystems are used in the multi-center management module and their characteristic functions in clinical research. Data quality and interconnection are supported through standardized data management solutions such as unified coding rules, structured clinical information, standardized sample circulation information, and standardized data storage. Thereby, this system provides comprehensive biomedical information for the promotion of translational medicine research and the development of precision medicine.

基金项目：国家自然科学基金（12171318）；上海市自然科学基金（20JC1410100，21ZR1436300）；上海市公卫三年行动计划（GWV-10.1-XK05）；上海交大 STAR 项目（20190102）
引用本文：孔伟名，吕文文，姜嘉媛，等. 多中心结直肠癌临床研究生物样本库信息系统的建设与管理［J］. 协和医学杂志，2022，13(4)：664-669. doi：10.12290/xhyxzz.2021-0763.

【Key words】multi-center; clinical research; biobank; information system

Funding: National Natural Science Foundation of China (12171318); Natural Science Foundation of Shanghai (20JC1410100, 21ZR1436300); Shanghai Three-year Action Plan for Public Health System Construction (GWV-10.1-XK05); Sino-Shanghai-SJTU Trans-med Awards Research (20190102)

Med J PUMCH, 2022,13(4):664-669

随着居民生活水平的不断提高和饮食习惯的日益西方化，近年来我国结直肠癌的发病率明显升高，已跃居恶性肿瘤第5位[1]，大城市的增幅更快，目前结直肠癌是上海市发病和死亡数量均居第2位的恶性肿瘤[2]。虽然研究人员已在结直肠癌复发、转移机制方面进行了大量探索，但仍缺乏单细胞水平和不同时空横断面的多尺度、多层次认识，因此有必要全面系统地揭示结直肠癌转移、耐药、复发等演变过程中基因、表观、转录、蛋白、代谢等时空变化规律的全部信息（肿瘤命运全息图谱），精准阐明肿瘤命运的决定基础[3-4]。

结直肠肿瘤命运全息图谱的绘制依赖于标准化收集、处理、贮存，应用健康与疾病生物体的生物大分子、细胞和组织等生物样本，以及对生物样本建立标准化操作规范、信息管理和应用系统的生物样本库。而多中心结直肠癌临床研究生物样本库信息系统可将第一线的临床信息与生物样本时空状态信息、分子信息等资料相结合，为提高生物医学研究质量、探索更安全有效的医疗方法提供宝贵的全方位生物医学信息[5]。

1 临床生物样本库发展现状与趋势

1.1 国外现状

目前，国际上运行良好的生物样本库，根据其运营主体性质可分为三类：（1）以政府机构牵头组织和运营为主导的样本库，包括英国生物样本库、韩国国家研究资源中心、新加坡组织网络和泛欧洲生物样本库与生物样本资源研究平台等；（2）以医疗机构或研究机构牵头组织和运营为主导的样本库，包括国际生物和环境样本库协会、法国国家健康和医学研究院样本库、丹麦国家生物样本库等；（3）以第三方企业或公司牵头组织和运营为主导的样本库，如加拿大Genizon生物科学公司的生物样本库对25种疾病的遗传基础进行研究，已完成了10种疾病的全基因组扫描。国际生物样本库的发展已从注重生物样本数量的1.0时代、注重生物样本质量的2.0时代进步至为个体化精准医疗提供全面循证证据的生物样本库3.0时代[6-7]。

1.2 国内现状

我国生物样本库的建设始于20世纪90年代，1994年中国科学院建立了中华民族永生细胞库，此后北京脐带血造血干细胞库、天津肿瘤医院生物样本库、复旦肿瘤医院生物样本库、上海交通大学生物样本库、南京市多中心生物样本库、首都医科大学牵头建立的北京生物样本库等相继建立，国内生物样本库凭借国家重点支持和人口资源等优势积累了丰富的生物样本[8-9]。2021年3月13日，十三届全国人大四次会议通过的《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》明确指出，“基因与生物技术”作为七大科技前沿攻关领域之一，“生物技术”作为九大战略性新兴产业之一，“基因技术”作为未来产业，推动生物技术和信息技术融合创新，构建国家生物数据中心体系[10]。现阶段，在国家政策的支持和指导下，各地区高校、医院和企业积极推进生物技术与信息技术融合创新的生物样本库建设升级工作，促进生物信息整合，推动精准医疗发展。

1.3 发展趋势

目前，临床生物样本库的发展遇到2个瓶颈。首先，为继续深入探究疾病发生发展的内在原因，研究者需要更加全面的生物信息（包括基因、表观、转录、蛋白、代谢等时空变化规律的全部信息），需要多个检测平台和医疗机构进行协作与共享；其次，以第三代测序技术为代表的新型技术不断迭代，将产生大量的分子信息数据，对生物样本库的数字化存储、交互和分析挖掘也提出了更高要求。临床生物样本库需从以贮存和管理生物样本为中心的传统生物样本库转变为以临床信息、样本信息和组学信息为中心的数据型生物样本库[11-12]。

2 结直肠癌临床研究生物样本库

2.1 项目介绍

上海交通大学医学院在上海市科学技术委员会的支持下，于2020年启动“结直肠癌转移命运机制及

过程调控研究”项目。该项目以探究肿瘤命运全息图谱（全面系统地揭示结直肠癌转移、耐药、复发等命运演变过程中基因、表观、转录、蛋白、代谢等时空变化规律的全部信息）、精准阐明肿瘤命运决定基础为目标，联合上海交通大学医学院等多家附属医院和上海交通大学医学院临床研究中心，开展结直肠癌多中心前瞻性队列研究以及多中心临床研究生物样本库信息系统建设工作。

2.2 建设框架

构建多中心结直肠癌临床研究生物样本库信息系统，涉及临床数据记录、生物样本流转记录、生物样本检测数据、系统业务及用户安全数据等关键信息。充分考虑信息系统的关联性、易用性、安全性、拓展性和敏捷性等指标，其架构采用微服务架构（micro services）、浏览器/服务器模式（browser/server，B/S）。微服务架构将系统依据的业务场景和数据模型拆分为独立的服务个体，相较于整体式的面向服务架构（service-oriented architecture，SOA），微服务架构的实施模式为自下向上型，独立进行不同服务的开发、构建、部署和发布，服务间采用 Restful 通讯机制，具有敏捷迭代、灵活部署、松耦合和分布式弹性拓展等优势[13]。在 B/S 模式下，使用者可通过浏览器进行便捷访问，显著减轻了客户端安装和升级系统的负担，用户安全验证模块、网络安全防护模块和基于 HTTPS 的加密网络传输技术显著保障了系统整体的数据安全。

根据项目临床科室、生物样本库和各组学平台的网络拓扑结构业务情况分析，多中心结直肠癌临床研究生物样本信息系统可拆分为以下 3 个子系统/平台：（1）以患者为中心的临床电子数据采集（electronic data capture，EDC）系统；（2）以样本为中心的生物样本管理系统；（3）以组学数据为中心的生物信息数据平台。EDC 系统采用表单引擎构建电子病历报告表，研究者可通过浏览器远程访问 EDC 系统录入临床数据；生物样本管理子系统以互联网为通讯方式，通过微信小程序远程调用物联网标签打印机打印样本标签，并结合小程序的条码识别功能，扫码记录标签打印、样本流转和样本操作等流程信息，并将关键信息自动同步至 EDC 子系统，同时对存储设备和生物样本的状态进行监控；生物信息数据平台依托大容量的存储设施和高性能的计算节点，为生物样本检测数据提供存储服务和统计分析服务。多中心结直肠癌临床研究生物样本库信息系统平台整体架构详见图 1。

2.3 管理模式

多中心结直肠癌临床研究生物样本库信息系统管理模式详见图 2。EDC 系统为受试者分配编码，并将整个试验阶段所需收集的临床信息以电子化的方式记录至系统中，EDC 系统采用 B/S 架构便于各中心进行统一化数据录入和集中化数据核查；在基线阶段、

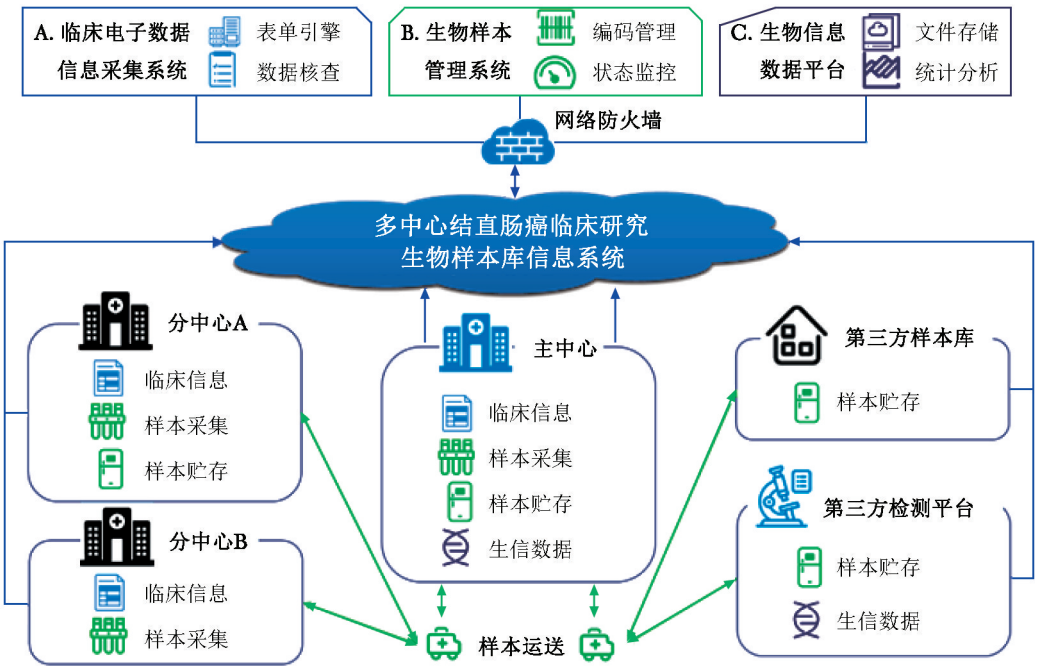

图 1 多中心结直肠癌临床研究生物样本库信息系统平台架构

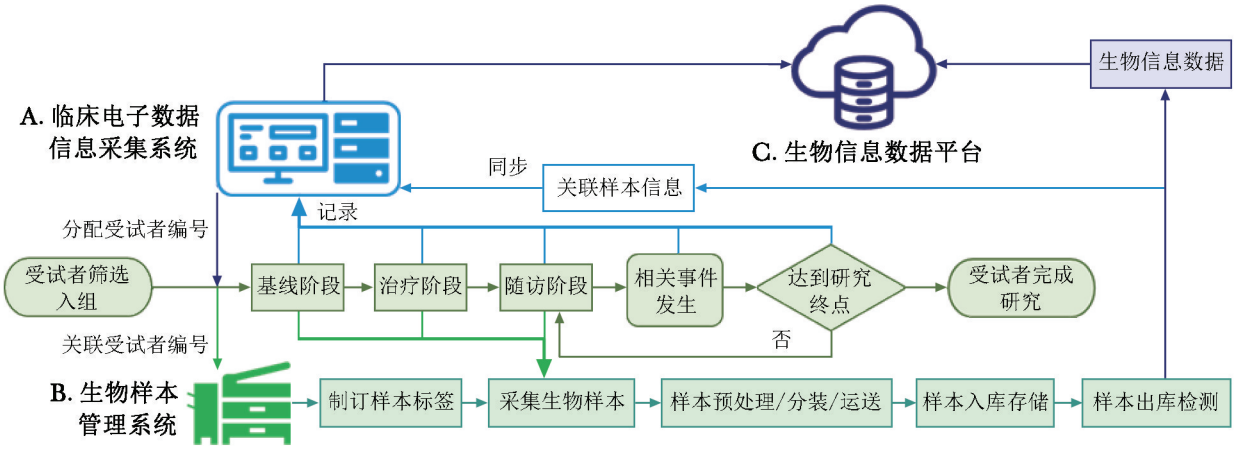

图 2 多中心结直肠癌临床研究生物样本库信息系统管理模式

治疗阶段和随访阶段，研究人员根据临床研究方案要求和样本采集的标准操作规程（standard operating procedure，SOP）获取装有目标生物样本的采集管，采集管上附有定制编码（以条形码/二维码或其他图形符号标识），此定制编码由生物样本管理系统制订和识别，样本在后续的接收、运输、采集、处理、贮存和发送等过程的识别均以采集管上的编码信息为凭证，相应的时间、空间信息变换将以互联网为通讯方式记录在生物样本管理系统中，并将样本编号及关键信息同步至 EDC 系统。同时，存放生物样本的设备温度、湿度和状态信息，亦可通过物联网记录在生物样本管理系统中，方便研究人员对生物样本的贮存状态进行管控。根据临床研究方案要求，将生物样本送至满足检测要求的机构进行检测，并将检测结果上传至生物信息平台，由平台统一存储生物信息数据，该平台与 EDC 系统互联互通，可通过受试者的临床信息检索出其对应的生物样本检测数据，亦可通过生物样本编码查找到对应的受试者信息，将临床信息与生物信息数据相结合进行统计分析，为精准医学的发展奠定数据基石。

3 结直肠癌临床研究生物样本库的特色

3.1 规范的临床电子数据采集系统

为规范临床试验电子数据采集技术的应用，促进临床试验电子数据的真实性、完整性、准确性和可靠性符合《药物临床试验质量管理规范》和数据管理工作相关规定的原则要求，2016 年 7 月，国家食品药品监督管理总局发布了《临床试验的电子数据采集技术指导原则》[14]。多中心临床研究 EDC 系统应达到指导原则中规定的软件、硬件、人员、系统环境及使用的基本要求；并制订数据管理计划，包括试验启动阶段、进行阶段和结束阶段相关电子数据采集系统的功能使用要求及说明。范德堡大学（Vanderbilt University）开发的电子数据采集系统 REDCap（Research Electronic Data Capture）可供全球学术、科研及公益机构使用，该项目采用 REDCap 作为临床电子数据采集系统，符合 21 CFR Part 11（Title 21 Code of Federal Regulations Part 11）、FISMA（the Federal Information Security Management Act）、HIPAA（Health Insurance Portability and Accountability Act）和 GDPR（General Data Protection Regulation）规范要求[15]。

3.2 便捷的生物样本管理系统

多中心临床研究使生物样本管理系统面临巨大的考验，由以往单实验室的单点业务模式转变为多家医疗机构、多个贮存中心、多个检测平台相互关联的网络拓扑结构业务模式，需更加详细和精准地记录生物样本的时间、空间和状态信息。得益于当前智能手机和条形码/二维码识别技术的快速发展和普及，无需额外购置专用的便携式扫码枪即可实现对条形码/二维码的识别，使得条形码/二维码识别工具的应用门槛大大降低；研究人员可使用智能手机扫描采集管上的条形码/二维码，对采集管的编号进行识别[16]。

目前，对于采集管上条形码/二维码的设置存在两种方案，一种是采集管在出厂时即已印制条形码/二维码，另一种是研究者在使用时将打印好的条形码/二维码标签粘贴于采集管外部。为兼容不同平台的贮存设备和多种型号的样本采集管，本项目采用在采集管外部粘贴条形码/二维码标签的方案，该方案可在标签上标注生物样本的关键信息，方便进行识别和操

作，可兼容多种型号的采集管，适配性较好，但需注意选取满足贮存环境要求的标签纸和打印设备（如-80 ℃低温冰箱条件下，应使用能够耐受-80 ℃低温的标签纸和全树脂基碳带的热转印标签打印机）。同时，为方便识别标签，应规范其粘贴位置。本项目自主研发的生物样本管理系统，通过微信小程序远程调控物联网标签打印机打印样本标签，并结合微信二维码的识别功能，扫码记录标签打印、样本流转、样本操作等流程信息，将关键信息自动同步至 EDC 子系统，同时对贮存设备和生物样本状态进行监控。

3.3 超大容量存储和高效分析的生物信息平台

本项目以探究肿瘤命运全息图谱、精准阐明肿瘤命运决定基础为目标，其中组学数据检测结果相关文件占用较大存储空间。以单细胞测序为例，单个生物样本测序结果占用的存储空间一般为 100 GB 以上，此类文件的存储、交互及挖掘对生物信息平台提出了较大挑战。在文件存储方面，生物信息平台采用分布式文件存储服务器，支持弹性扩容、多副本、自动纠错和容灾恢复等[17]。针对生物信息分析工具种类繁多、部分工具安装和使用较为繁杂的情况，采用支持通用工作流语言（common workflow language，CWL）和工作流描述语言（workflow description language，WDL）的生物信息数据分析平台 Galaxy[18]。该平台提供基于 WEB 端的可视化操作界面，允许研究人员通过浏览器访问系统页面后自定义分析流程和命令，且附带多种常用的生物信息分析工具，可显著减轻研究人员工作量，提高分析效率[19]。Galaxy 同时支持超级计算中心的 SLURM（simple linux utility for resource management）集群管理器和作业调度系统，可将复杂任务分配于集群运行，突破常规分析软件单节点计算资源受限的瓶颈。

4 数据标准化管理方案

本项目涉及多个医疗中心、生物样本库、组学平台的数据采集和信息传输，因此加强数据管理、制订数据标准化管理方案至关重要。统一化编码规则、结构化临床信息、标准化样本流转信息和规范化数据存储等数据标准化管理方案，可保障生物样本库的整体数据质量及各子系统间数据的高效互通互联。

4.1 统一化编码规则

项目制定了统一化数据编码规则，以生物样本编号为例，统一化的生物样本编号作为生物样本的唯一识别编号，由生物样本管理系统根据受试者编号、样本类型、时间节点和采集序号进行组合生成，作为各子系统间传输生物样本信息时的唯一标识编号，可高效便捷地定位生物样本所关联的临床信息、状态信息和组学数据。

4.2 结构化临床信息

EDC 系统的表单引擎功能可对所采集的临床信息进行结构化处理，对于非结构化和自由文本类型的临床信息，尽可能对其拆分和细化，使其转化为可结构化或可量化的变量，同时参考国际 CDISC（Clinical Data Interchange Standards Consortium）标准对变量和选项进行命名，使临床信息便于收集、传输和统计分析。

4.3 标准化样本流转信息

研究人员可通过生物样本管理系统微信小程序扫描采集管上的条形码/二维码，在移动端实时录入标准化的生物样本流转信息，包括样本类型、计量单位、预采集量、采集量、采集管质量、操作人员、操作时间、操作状态、存储位置、存储时间、存储环境、送检时间、送检平台和检测结果等信息，其中操作状态包括采集、运送、接收、贮存、检测和废弃。系统将样本流转信息同步至 EDC 系统，研究人员可直接在 EDC 系统查看受试者的临床信息及其对应的生物样本实时流转信息，同时系统可根据标准操作手册对样本流转信息中的异常情况（如送样时间超时、样本采集量不足等）进行实时预警。

4.4 规范化数据存储

检测平台将检测后的组学数据文件与检测信息（关联的生物样本、使用的设备、方法、操作人员、操作时间、文件校验码等信息）一同上传至生物信息平台，研究人员可通过生物信息平台该组学数据文件的统一资源定位符（uniform resource locator，URL）获取其数据文件。平台将根据上传的检测信息，解析关联的受试者和样本，并将组学数据文件的 URL 传至 EDC 系统中对应的受试者记录相关字段内，以便于研究人员检索数据时，可以快速匹配受试者相关的临床信息和组学数据文件。

5 小结

在我国全面保障和提升全民健康的过程中，生物样本库得以快速发展。多中心结直肠癌临床研究生物样本库信息系统的建设，突破了单点式生物样本数据库数据孤岛和资源限制的瓶颈，将以贮存和管理生物样本为中心的传统生物样本库，转变为全面收集、高效管理和深度挖掘临床研究中的临床信息、样本信息

和组学信息等各种资源，形成一体化聚合数据库为中心的数据型生物样本库。多中心结直肠癌临床研究生物样本库信息系统的建设经验和管理模式值得推广至更多涵盖生物样本信息的多中心临床研究，通过整合全方位的生物医学信息，使之最大限度地服务于转化医学研究和精准医疗。

作者贡献：孔伟名负责样本管理程序开发和生物信息平台搭建及论文撰写；吕文文负责数据标准化管理方案制订；姜嘉媛负责 EDC 系统的构建；冯铁男负责项目调研分析和设备管理；俞章盛负责系统框架设计、项目整体管理及论文审校。

利益冲突：所有作者均声明不存在利益冲突

志谢：感谢上海交通大学基础医学院在设备、平台、专业技术人员、生物样本管理流程和生物信息数据存储/分析等方面提供的支持和帮助；感谢上海交通大学医学院附属瑞金医院、仁济医院等多家附属医院在研究方案设计和数据标准化管理方案中提供的支持和帮助。

参 考 文 献

[1] Sun D, Li H, Cao M, et al. Cancer burden in China: trends, risk factors and prevention [J]. Cancer Biol Med, 2020, 17: 879-895.

[2] 吴春晓，顾凯，庞怡，等. 2016 年上海市恶性肿瘤发病和死亡情况与 2002—2016 年的变化趋势分析 [J]. 中国癌症杂志，2021，31：879-891.
Wu CX, Gu K, Pang Y, et al. Analysis of the current status of cancer incidence and mortality in Shanghai, 2016 and trends of 2002—2016 [J]. Zhongguo Aizheng Zazhi, 2021, 31: 879-891.

[3] Jung G, Hernández-Illán E, Moreira L, et al. Epigenetics of colorectal cancer: biomarker and therapeutic potential [J]. Nat Rev Gastroenterol Hepatol, 2020, 17: 111-130.

[4] Xie YH, Chen YX, Fang JY. Comprehensive review of targeted therapy for colorectal cancer [J]. Signal Transduct Target Ther, 2020, 5: 1-30.

[5] 刘世建，傅启华，王伟，等. 临床生物样本库发展的机遇与挑战 [J]. 第二军医大学学报，2017，38：265-269.
Liu SJ, Fu QH, Wang W, et al. Opportunity and challenge of clinical biobank in China [J]. Di-er Junyi Daxue Xuebao, 2017, 38: 265-269.

[6] Simeon-Dubach D, Watson P. Biobanking 3.0: evidence based and customer focused biobanking [J]. Clin Biochem, 2014, 47: 300-308.

[7] Quinlan PR, Gardner S, Groves M, et al. A Data-Centric Strategy for Modern Biobanking [J]. Adv Exp Med Biol, 2015, 864: 165-169.

[8] 郭丹，杨文航，徐英春. 临床生物样本库信息系统建设与发展 [J]. 协和医学杂志，2018，9：81-86.
Guo D, Yang WH, Xu YC. Construction and Development of the Information System of Clinical Biobank [J]. Xiehe Yixue Zazhi, 2018, 9: 81-86.

[9] 郜恒骏，杜莉利，张小燕，等. 生物样本库发展的现状、机遇与挑战 [J]. 协和医学杂志，2018，9：172-176.
Gao HJ, Du LL, Zhang XY, et al. Status, Opportunities and Challenges of Biobanks [J]. Xiehe Yixue Zazhi, 2018, 9: 172-176.

[10] 中华人民共和国中央人民政府.《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》 [EB/OL]. (2021-03-13) [2021-11-30]. http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

[11] Müller H, Dagher G, Loibner M, et al. Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management [J]. Curr Opin Biotechnol, 2020, 65: 45-51.

[12] Glicksberg BS, Johnson KW, Dudley JT. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring [J]. Hum Mol Genet, 2018, 27: R56-R62.

[13] Williams CL, Sica JC, Killen RT, et al. The growing need for microservices in bioinformatics [J]. J Pathol Inform, 2016, 7: 45.

[14] 国家食品药品监督管理总局. 总局关于发布临床试验的电子数据采集技术指导原则的通告 [EB/OL]. (2016-07-29) [2021-11-30]. https://www.nmpa.gov.cn/directory/web/nmpa/xxgk/ggtg/qtggtg/20160729184001958.html.

[15] Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners [J]. J Biomed Inform, 2019, 95: 103208.

[16] Uzun V, Bilgin S. Using QR Code Technology to Reduce Self-Administered Medication Errors [J]. J Pharm Pract, 2021, 34: 587-591.

[17] O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics [J]. J Biomed Inform, 2013, 46: 774-781.

[18] Leipzig J. A review of bioinformatic pipeline frameworks [J]. Brief Bioinform, 2017, 18: 530-536.

[19] Chappell K, Francou B, Habib C, et al. Galaxy Is a Suitable Bioinformatics Platform for the Molecular Diagnosis of Human Genetic Disorders Using High-Throughput Sequencing Data Analysis. Five Years of Experience in a Clinical Laboratory [J]. Clin Chem, 2022, 68: 313-321.

（收稿：2021-11-30 录用：2021-12-29 在线：2022-06-23）
（本文编辑：李玉乐）