

大型综合性医院生物样本信息资源大数据科研平台的建设与应用^{*}

刘 辉 蔡宏伟 高娟娟 张琳珮 那 天 王亚文

(西安交通大学第一附属医院 西安 710061)

〔摘要〕 目的/意义 通过整合临床与生物样本信息,构建生物样本信息资源大数据科研平台,为科研人员提供一站式数据检索、集成、分析服务,建立数据治理体系,提升医院临床科研基础设施建设水平。方法/过程 采用通用数据模型和数据治理技术,通过抽取、转换、加载等步骤整合来自不同厂商的数据源,提供统一数据访问入口。结果/结论 该生物样本信息资源大数据科研平台具备多维数据筛查和快速集成分析的优势,可为临床科研提供支撑。

〔关键词〕 生物样本; 医疗大数据; 数据治理; 通用数据模型

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.01.013

Construction and Application of the Big Data Research Platform for Biological Sample Information Resources in a Large Comprehensive Hospital

LIU Hui, CAI Hongwei, GAO Juanjuan, ZHANG Linpei, NA Tian, WANG Yawen

The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China

〔Abstract〕 **Purpose/Significance** By integrating clinical and biological sample information, a big data research platform for biological sample information resources is built to provide one-stop data retrieval, integration and analysis services for researchers, and a data governance system is established, so as to improve the level of hospital clinical research infrastructure construction. **Method/Process** Common data model and data governance technology are adopted to integrate data sources from different vendors through extraction, transformation, loading and other steps to provide a unified data access portal. **Result/Conclusion** The big data research platform for biological sample information resources has the advantages of multi-dimensional data screening and rapid integrated analysis, which can provide support for clinical research.

〔Keywords〕 biospecimen; medical big data; data governance; common data model

〔修回日期〕 2023-07-17

〔作者简介〕 刘辉,博士研究生,工程师,发表论文 20 余篇;通信作者:王亚文,教授,博士生导师。

〔基金项目〕 陕西省重点研发计划一般项目(项目编号:2022SF-058)。

1 引言

生物样本是疾病研究的重要基石。近年来,对其管理与利用的监管不断加强。2022 年我国首个“生物经济五年规划”提出要强化生物资源利用平台支撑^[1]。现代生物样本库作为生物样本资源保存

的主要实体，应具备生物样本的全过程管理（“湿库”），以及相关临床数据资源的整合系统（“干库”）。整合多组学与临床数据“干库”对推动精准医学研究发展具有重要价值^[2]。医疗大数据的快速增长与科学应用逐步深入使“干库”建设尤为重要。目前国内生物样本库“湿库”信息系统建设相对成熟，“干库”建设基础较薄弱，亟须进一步加强。

整合医院多个信息系统的临床数据是“干库”建设关键环节。各临床业务系统数据标准化程度不足，系统架构、数据库等不同^[3-5]，临床医疗大数据的多源异构使“干库”建设面临巨大挑战^[6]。近年来，国内一些大型研究型医院通过信息科或大数据中心推进医疗大数据平台建设，并取得一定成果^[7-14]。目前整合临床数据和样本信息资源（“干库”建设）仍处于探索阶段，缺乏标准化的建设和

应用方案。

本文以西安交通大学第一附属医院（以下简称医院）的全院级生物样本信息资源大数据科研平台为例，介绍对生物样本及相关临床数据的整合、治理和应用案例。该平台旨在最大化挖掘战略性医疗数据资源价值，实现与实体生物样本融合，为医学临床研究提供从生物样本到临床数据的一体化科研服务。

2 生物样本信息资源大数据科研平台设计

以样本数据和临床数据为核心，建立以样本为中心的分子数据模型和以患者为中心的临床数据模型^[15]，以支持患者全生命周期浏览、科研灵感发现、人群精准检索、统计分析、科研项目管理、科研随访等多种功能，见图 1。

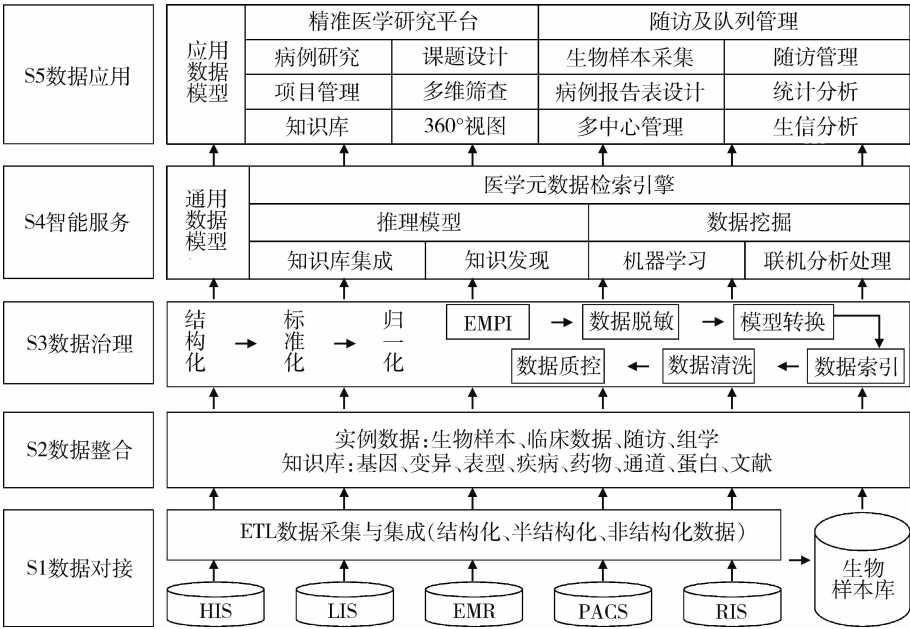


图 1 生物样本信息资源大数据科研平台总体设计

3 生物样本信息资源大数据科研平台实践

利用通用数据模型（common data model，CDM）解决医院临床数据资源多源异构和数据整合复杂性的问题。CDM 从多种电子数据库中提取特定信息的结构和框架，通过标准化的变量表单从海量数据中

快速、准确地提取所需关键信息^[16-17]。其定义了统一的数据标准，并通过数据的抽取、转换和加载（extraction - transformation - loading，ETL）过程形成标准化数据结构^[18]。CDM 的应用有效解决了数据整合难题。

本文基于 CDM 构建临床表型数据库，使用机器学习模型将医学数据映射到临床表型 CDM 中。

同时，设计标准化表型研究智能处理方法和医疗大数据挖掘与分析方法。通过 ETL 高效识别医疗数据中核心实体，并充分利用其关系构建事件图谱，实现医疗大数据的结构化和事件序列化，见图 2。关

键技术模块包括数据采集与存储、CDM 映射转换和智能挖掘、医学元数据检索引擎构建、应用数据模型（application data model，ADM）开发与使用。



图 2 临床表型通用数据模型

3.1 多源异构临床表型数据采集与存储

3.1.1 多源异构临床表型数据采集 本文构建的多模态数据库能够提取和存储患者基本信息、病历信息、病理报告和影像报告等各种类型数据，涉及文本、图像、数字和符号等多种形式。然而，临床表型数据来源为医院信息系统（hospital information system，HIS）、检验信息系统（laboratory information system、LIS）、影像存储与传输系统（picture archiving and communication system，PACS）、电子病历（electronic medical record，EMR）系统等，每个系统采集数据的标准和规范不同。为解决这一问题，将采集的数据分为历史数据和实时数据，历史数据通过厂商镜像库推送到本地中心库，而实时数据通过 ETL 工具访问厂商实时镜像库进行抽取，或采用符合医疗信息系统互操作性协议等标准协议的前置技术进行实时抽取。

3.1.2 本地原始数据存储库构建 采用前置机技术从不同厂商以数据表为维度 1:1 抽取临床数据，并以此构建本地原始数据存储库。对于基于文档的

非结构化电子病历数据，通过转换配置将其以数据文件形式存储到本地中心库。对于 PACS 中的图像非结构化数据，通过对其报告文本数据进行结构化和文档化处理，建立起报告文档数据与图像数据之间的映射连接，并将其存储到本地中心库。由此有效地管理和利用临床数据，并建立报告文档数据与图像数据之间的关联。

3.2 临床表型本体数据库整合与智能挖掘

3.2.1 患者主索引建立 根据医院业务系统汇总患者基本信息数据，建立患者主索引（enterprise master patient index，EMPI），使用患者的原始 ID 号、身份证号、护照号、住院号、门诊号、姓名、性别、出生年月、手机号等信息来判断不同 ID 或住院号的患者是否为同一人，并生成主索引号，见图 3。生物样本信息以捐献者（患者）为维度进行组织和管理，每个捐赠者赋予唯一标本号，样本赋予唯一样本编号，样本编号是构成数据信息的最小颗粒标识符。通过样本编号可以追踪和管理样本，并与临床信息关联。

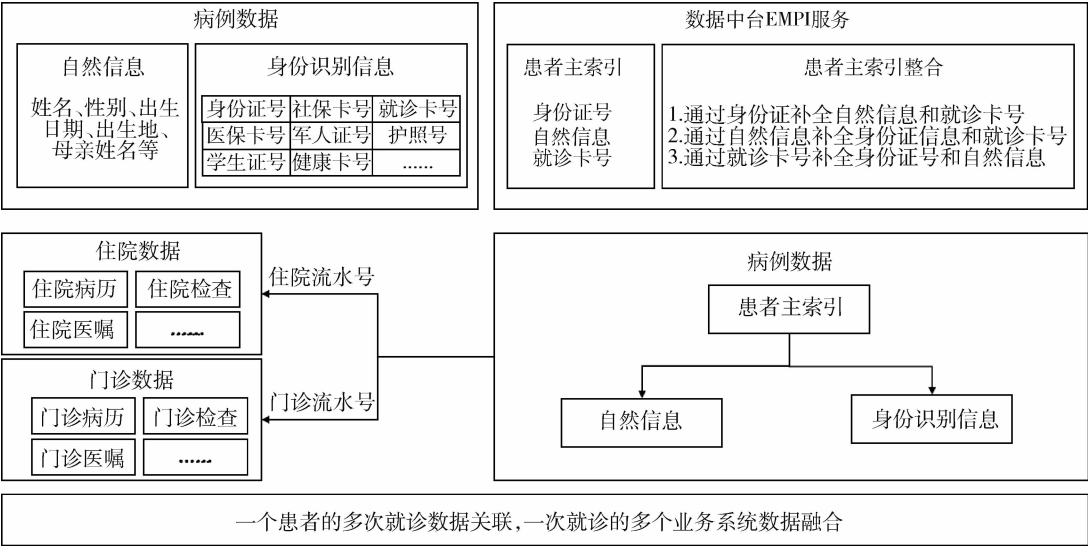


图3 EMPI 服务总体框架

3.2.2 统一 CDM 建立 通过建立统一的数据结构和标准通用模型，实现各临床业务系统数据标准化处理。模型包括两部分：结构化数据模型转换和非结构化电子病历解析入库。采用机器学习模型将结构化和非结构化的临床医疗数据映射到临床表型 CDM 上，以提高临床医疗数据的可管理性和可共享性。

3.2.3 数据清洗和质量控制 数据清洗对后续研究的模型效果和最终结论至关重要。数据清洗包括字段清洗、正则表达式清洗和复杂逻辑清洗。通过数据清洗将不同来源的数据整合到一个参考体系下。本文使用数据质控模块检验数据质量，包括完整性（检验空值率）、规范性（验证字段格式和数据是否符合规范要求）、有效性（验证业务发生时间的先后关系）、一致性（校验数据总量，确保数据一致性）、唯一性（进行重复数据校验）、关联性（校验孤儿数据，确保其与引用数据的完整性关联）和分布性（根据分布字段进行分组统计，分析数据的时间和值域分布）等。

3.2.4 数据脱敏 为保护患者隐私，根据《个人信息保护法》《信息安全技术 个人信息安全规范》等相关法律法规和标准，对临床数据中的个人信息（如姓名、身份证号等）进行脱敏处理。脱敏过程中严格控制数据的使用权限和获取范围。根据数据的敏感程度，采用加密、匿名化、去标识化等方法进行数据处理，如用“***”代替“姓名”，见表1。

表 1 临床数据脱敏字段与策略示例

患者数据	敏感等级	脱敏方式	举例说明
姓名	P0	滤除 (* 号代替)	***
联系人姓名	P0	滤除 (* 号代替)	***
街道名称	P0	仅保留城市区县	西安市碑林区 ***
完整地址名称	P0	仅保留城市区县	西安市碑林区 ***
联系人地址 (区县名称)	P1	仅保留城市区县	西安市碑林区 ***
联系人地址 (街道名称)	P1	仅保留城市区县	西安市碑林区 ***
出生日期	P1	保留年	1976
手机号码	P1	隐藏中间 6 位	156 ***** 32
联系人手机 号码	P1	隐藏中间 6 位	156 ***** 32
社保卡号	P0	去标识化	XSOFYWEHEHDUDHU
患者住院号	P0	去标识化	20RUFDHUEUEUSNU
患者就诊卡号	P0	去标识化	203UFDHUEUEUSNU
证件号	P0	隐藏末尾 8 位	6101031976 *****

注：P0 可以找到具体人；P1 可以找到某一类人；匿名化使用 * 号代替，去标识化使用加密后的编码代替。

3.2.5 数据整合 将患者历次就诊的临床数据整合在一起，包括就诊、诊断、医嘱、检验、检查、手术记录、病案首页、出入院记录、病程记录等。

按照患者就诊维度进行组织,脱敏处理后形成完整样本临床病历信息数据,并建立索引提供高效的访问接口。

3.2.6 “干湿”两库融合 生物样本信息的最小维度是样本,但是样本信息记录是以捐赠者为基础的。通过捐赠者的身份证号、姓名、住院号、门诊号等关联临床数据,实现样本和临床数据的整合。在生物样本登记入库时,会将相应项目负责人标识与样本关联起来,在大数据平台界面端,项目负责人只能访问和检索自己入库的样本信息及相关临床数据。通过访问权限管理,有效地保护捐赠者隐私信息,同时也可防止未经授权的访问和滥用,确保数据安全。

3.3 医学元数据检索引擎构建

为实现多模态多粒度医疗元数据的高效索引,采用 Elasticsearch 的 REST API^[19]来定义元数据格式,实现快速上传和高效索引构建。为实现同步检索,利用专业语料库构建各类型索引数据的语义关联网络,并将其存储于传统关系型数据库中,以提升系统面向语义、支持复杂查询的资源发现能力。

3.4 应用数据模型开发与使用

通过 ADM 构建智能检索、临床科研等数据服务应用层,主要包含以下功能。一是数据整合和智能搜索,可实现对所有临床表型数据的多维检索。二是课题设计和队列建立,支持科研人员便捷设计课题,快速构建研究队列。三是病例研究,展示脱敏后的患者 360°视图信息。四是文献和知识库检索,查询最新科研文献、关联不同领域知识库。五是统计分析,对课题数据进行基线统计、生存曲线等分析。六是科研项目管理,支持添加和删除课题成员、分配角色和权限以及课题数据共享。

4 应用效果及讨论

4.1 应用效果

4.1.1 多维筛查 本文建设的全院级生物样本信息资源大数据科研平台可提供数据检索、研究变量

集成、方案设计、数据统计与生物信息学分析技术支持等服务。使用平台的多维筛查界面可快速完成患者入组检索。基于本文所设计的通用数据模型,通过组合诊断名称、手术名称、检验指标及样本保藏情况等多个维度的筛选选项,可以精准地找到符合条件的患者。例如,根据研究者提供的患者入组条件(主要诊断是胰腺炎并发腹腔感染或腹腔出血;做过腹腔动脉造影手术;本次住院期间血红蛋白小于 80 g/L;留存有胰腺组织或血液等生物样本;2013 年 1 月 1 日—2023 年 1 月 1 日期间在肝胆外科就诊过的住院患者),平台可以多维筛选获得入组患者信息并导入项目,继而进行纳排分组,创建基线指标和研究变量。最终生成研究人群的数据子项内容开展后续研究。

4.1.2 科研数据服务 2019 年以来该平台已为医院 40 个临床科室提供 500 多项科研数据服务。通过使用该平台,科研人员工作效率和数据质量显著提高。以心血管内科某团队为例,其承担了一项 1 000 例患者的临床研究项目,需要进行建立纳排标准、筛选入组人群、收集所需研究变量数据等工作,既往需 4 名研究生耗费 3 个月时间才能完成,应用平台仅需 1 名工作人员在 2 日内即可完成。基于该平台,每年发表文章数量从 3 篇增加到 10 余篇,所需人力从 4 人减少到 1 人,花费时间占比从 50% 减少到 5%。同时,因该平台集成了全院 HIS 内网中的临床数据,科研人员可以更加全面地获取和分析临床数据,从而提高科研数据的准确性和科研成果的可信度。

截至 2022 年底,该平台数据检索与集成分析功能已支撑科研人员发表论文 70 余篇,论文发表在 *European Heart Journal*^[20]、*JAMA Internal Medicine*^[21-22]和 *Nucleic Acids Research*^[23]等高质量临床研究期刊。

4.2 存在问题

该平台可以显著提高临床科学研究的效率和质量,但应用实践中仍存在以下问题。一是生物样本出库后的实验回库数据和临床随访数据缺失。二是尚未实现确保数据安全前提下的院外数据共享。三

是缺乏合理有效的匿名化标准以确保数据安全。四是在利用该平台进行科学研究时,存在难以获得知情同意的问题。研究者需要获得患者联系方式以便征求个人知情同意,但如果没有事先获得同意,则无法获取联系方式。有待探索其他合规的方法解决该矛盾,以确保符合相关法规和伦理要求。

5 结语

通过构建和应用生物样本信息资源大数据科研平台,实现了数据整合与治理、一站式数据检索与集成,以及实体生物样本与信息资源的融合等目标。该平台为大型综合性医院临床科研提供了高效、高质量的数据支持,成为不可或缺的重要平台,推动医疗科研工作不断创新和发展。然而,为进一步提升平台的可靠性和充分挖掘生物样本库的数据资源,仍要持续改进和完善。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 张乃心, 赵庆辉, 汤红明, 等. 临床生物样本库建设的思考 [J]. 中国研究型医院, 2023, 10 (1): 11-15.
- 刘艳红, 叶庆. 精准医疗时代下生物样本库的建设与发展特点 [J]. 协和医学杂志, 2021, 12 (2): 254-259.
- 国家卫生健康委统计信息中心. 医院数据治理框架、技术与实现 [M]. 北京: 人民卫生出版社, 2019.
- 姬卫东, 李琳, 张振, 等. 互联互通背景下医疗数据治理面临的问题与对策 [J]. 中国数字医学, 2021, 16 (11): 6-11.
- 张胜发, 罗葳, 马玉环, 等. 关于加快健康医疗大数据向科学数据转化的思考与建议 [J]. 医学信息学杂志, 2022, 43 (11): 8-13.
- 周琳, 王飞, 赵浩宇. 医院数据存在问题与管理对策研究 [J]. 重庆医学, 2022, 51 (24): 4300-4303.
- 罗辉, 薛万国, 乔岫. 大数据环境下医院科研专病数据库建设 [J]. 解放军医学院学报, 2019, 40 (8): 713-718.
- 唐颐信息技术. 临床数据与样本资源库 [EB/OL]. [2023-05-07]. [http://www.eternal-health.cn/part-](http://www.eternal-health.cn/part-ner-301-1.html)

- ner-301-1.html.
- 王奕. 复旦大学附属肿瘤医院数据平台实践总结 [J]. 中国信息界 (e 医疗), 2014, 7 (8): 55-56.
- 吴正一, 崔迎慧, 陆耀, 等. 以临床数据仓库为核心的医院大数据平台构建 [J]. 中国医院管理, 2015, 35 (11): 13-15.
- 殷亦超, 高炬, 何萍. 研究型医院的临床大数据管理应用与实践探索 [J]. 中国数字医学, 2019, 14 (2): 34-36.
- 匡亚岚, 李春漾, 应志野. 医疗机构自主可控大数据科研平台建设方案 [J]. 医学信息学杂志, 2022, 43 (3): 66-70.
- 吴燕秋, 黎美秀, 丁元杰, 等. 面向临床科研的全院级医疗大数据平台建设 with 数据治理实践探索 [J]. 中华医学科研管理杂志, 2021, 34 (2): 81-86.
- 刘畅, 董宁欣. 基于医院数据中心的临床科研平台构建 [J]. 中国医疗设备, 2023, 38 (4): 107-111.
- 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 (1): 81-86.
- Observational Health Data Sciences and Informatics. The OMOP Common Data Model [EB/OL]. [2023-05-07]. <https://www.ohdsi.org/data-standardization/>.
- 张弘政, 刘迷迷, 李琳, 等. 基于通用数据模型的健康医疗大数据平台数据治理研究 [J]. 医学信息学杂志, 2022, 43 (6): 2-7, 13.
- 何家双, 肖晓旦. OMOP CDM 在临床科研中的应用思考 [J]. 中国数字医学, 2016, 11 (3): 72-74.
- 张文亮. Elasticsearch 全面解析与实践 [M]. 北京: 机械工业出版社, 2021.
- ZHAO Y X, LI G L, WANG S, et al. Multiple lipomata of the tricuspid valve and papillary muscle: case report [J]. European heart journal case reports, 2021, 5 (7): ytab260.
- LI G L, QUAN X P, YAN Y. Abnormal electrocardiograms after diarrhea and syncope in the emergency department [J]. JAMA internal medicine, 2021, 181 (12): 1645-1646.
- LI G L, TANG M Y, YAN Y. Typical electrocardiogram pattern unmasks atypical cause of muscle weakness [J]. JAMA internal medicine, 2021, 181 (11): 1503-1504.
- LIU B, LI S S, LIU Y, et al. Bacteriophage Twort protein Gp168 is a β -clamp inhibitor by occupying the DNA sliding channel [J]. Nucleic acids research, 2021, 49 (19): 11367-11378.