

区域健康医疗数据开放共享系统设计与应用^{*}

郭 一¹ 胡盈盈¹ 胡 斐² 沈扬扬² 许 杰²

(¹ 浙江省卫生健康信息中心 杭州 310006 ² 浙江扁鹊健康科技有限公司 杭州 311100)

〔摘要〕 目的/意义 实现安全高效的健康医疗数据开放共享, 激发健康医疗数据要素价值。方法/过程 针对健康医疗数据开放共享的现状和问题, 设计一种基于数据集成治理、数据资产管理、数据产品超市、数据采样分析、在线沙箱分析的健康医疗数据开放共享系统。结果/结论 实践证明该系统能够有效实现健康医疗数据采集、分析、管理和利用。

〔关键词〕 健康医疗数据; 数据产品超市; 数据沙箱; 开放共享

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.08.014

Design and Application of the Regional Open Sharing System for Health and Medical Data

GUO Yi¹, HU Yingying¹, HU Fei², SHEN Yangyang², XU Jie²

¹ Zhejiang Health Information Center, Hangzhou 310006, China; ² Zhejiang Bianque Health Technology Co., Ltd., Hangzhou 311100, China

〔Abstract〕 **Purpose/Significance** To achieve safe and efficient open sharing of health and medical data, and to stimulate the value of health and medical data elements. **Method/Process** Aiming at the current situation and problems of open sharing of health and medical data, an open sharing system of health and medical data is designed based on data integration governance, data asset management, data product supermarket, data sampling analysis and online sandbox analysis. **Result/Conclusion** The practice has proved that the system can effectively realize the collection, analysis, management and utilization of health and medical data.

〔Keywords〕 health and medical data; data product supermarket; data sandbox; open sharing

1 引言

随着大数据时代的到来, 数据资产已经成为企业和组织最重要的资源之一, 尤其在卫生健康领域, 数据要素逐渐成为驱动医疗服务重塑提升、医疗资源优化配置、管理体系提能升级的核心动力。

2016 年《关于促进和规范健康医疗大数据应用发展的指导意见》^[1]明确提出健康医疗大数据是国家重要的基础性战略资源。2022 年《“十四五”全民健康信息化规划》^[2]更加具体地提出要构建健康医疗大数据资源要素体系, 推进健康医疗大数据应用发展, 充分释放数据价值。

健康医疗数据来源于患者临床诊疗、健康管理等多种途径, 种类多、规模大、结构多样, 如何充分利用先进的技术手段实现数据高效开放共享是一个全新的命题。健康医疗数据的开放共享, 要连接数据管理方与数据使用方两个主体, 探索建立可

〔修回日期〕 2024-04-11

〔作者简介〕 郭一, 高级工程师, 发表论文 10 余篇。

〔基金项目〕 浙江省科技计划项目重点软科学研究项目 (项目编号: 2023C25030)。

信、安全的数据开放应用环境，形成一套完整的开放共享应用模式，对健康医疗数据的治理、存储、利用和安全共享提出了更高的要求^[3-5]。本文从技术层面提出一种健康医疗数据全方位治理和开放共享的路径，通过数字化技术提升数据综合治理能力和规范化管理能力，推动数字医疗资源从孤立分散走向协同共享，促进数据“内增值、外增效”，优化便民服务和政务服务体验，推动卫生健康行政部门、医疗机构更加积极地开展健康医疗数据合规开放共享。

2 现状分析

当前，健康医疗数据使用方对数据的要求趋于精细化、标准化，数据管理方对管理的要求趋于规范化、透明化，而数据生产方产生的数据存在差异化、非标准化等问题，健康医疗数据开放共享链路参与方之间存在巨大的鸿沟，数据的整合和开放共享面临难题。一是标准规范存在差异。健康医疗领域各信息系统由于建设厂商、设计架构等差异，缺乏相对统一的数据标准规范，生产数据的质量和可用性不高。二是数据扩展存在瓶颈。健康医疗数据分散存储在不同的业务系统中，受限于系统建设的技术水平及业务规模，数据扩展的灵活性和适应性较差。三是数据安全存在隐患。健康医疗大数据分析的前提是更加规范的数据安全和隐私保护机制。随着大数据应用的拓展，数据安全和隐私保护面临严峻挑战。四是数据利用效率不高。健康医疗数据体量达到千万亿级别，海量数据的利用效率问题亟待解决，需要开发更高效的数据分析工具，更快速地为决策治理提供支持。

3 数据开放共享系统架构

健康医疗数据开放共享以规范、实时、安全、稳定的数据集成治理为基础，通过全周期的数据分析利用实现数据要素向数据产品的转化。健康医疗数据开放共享系统架构设计，见图 1。数据开放共享系统依托高性能的数据交换通道和数据计算引

擎，标准化采集健康医疗数据，开展数据质控，并监控采集全链路。以数据使用需求为驱动，对采集的健康医疗数据进行编目和分类分级，从价值和数量双重维度对数据资源开展资产化评估，将有价值的数据资源开发为数据产品，构建数据产品超市。基于数据产品超市的数据开放能力，以“原始数据不出域、数据可用不可见”为原则，搭建数据采样分析工具和数据沙箱环境，实现模型迭代、参数优化等功能，赋能居民端、医生端、治理端场景建设，提升医疗服务水平和政府治理效能。

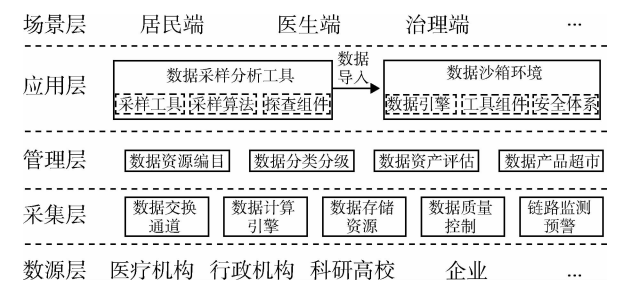


图 1 健康医疗数据开放共享系统架构

4 数据开放共享系统设计

健康医疗数据开放共享技术路线可从数据提供方、数据管理方、数据使用方 3 类参与方的活动维度拆分，包含数据集成治理、数据资产管理、数据产品上架、数据采样分析、在线沙箱处理等环节，见图 2。

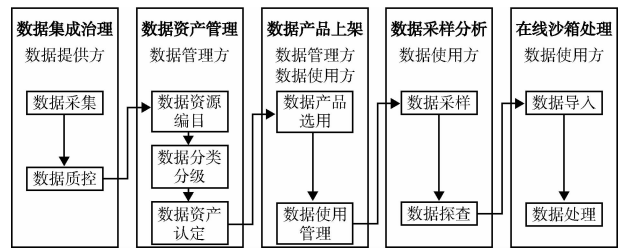


图 2 健康医疗数据开放共享技术路线

4.1 集成治理

数据集成治理是对数据标准化采集进行规划、监控和执行的管理活动，建立健康医疗数据的统一标准，提供数据采集以及质控的能力^[6]，从而提升

数据价值,促进开放共享。

4.1.1 数据采集 数据采集是从各种数据源中获取数据并将其转换为可用于分析和处理格式的过程,采集技术主要包括基于日志的增量数据同步、基于查询的批量数据迁移和基于消息的数据交互服务等,技术各有优劣,应针对不同的应用场景和数据来源选择适宜的采集技术。数据采集主要提供以下服务功能。(1) 采集通道集成。集成主流数据采集通道,多通道共用连接配置和数据模型信息,且支持集群化,多个采集引擎实例间自动故障转移,实现采集通道高可用。(2) 采集作业调度。提供可视化的采集作业配置能力,实现采集过程和结果可视化,自动、手动执行采集作业,以及采集任务统一管理。(3) 采集链路监控。全面监控数据采集源端、通道、任务和中心端等链路节点,同时通过数据对账和数据流量波动监测预警发现异常,实现统一运维监控。

4.1.2 质控 数据质量的量化直接影响数据资源的价值评估。数据质量监控基于统一的数据采集标准,量化采集数据的可用性,同时结合问题整改和闭环反馈等措施,持续提高数据质量。数据质量监控主要提供以下功能。(1) 数据质控报告查询。可按核验周期、数据来源、量化规则、质量问题等多角度查询数据质量评价报告,支持查看数据质量变化趋势、导出报表,包括综合评估结果和异常数据明细,便于管理者和技术人员及时、全面了解数据质量,保障数据生产和传输过程中的完整、准确、一致、及时。(2) 数据质量闭环管理。建立数据质量提升的控制流程,通过运维工单自动记录、发布数据质量问题,通过消息提醒通知数据生产方及时整改数据,并跟踪问题整改情况,实现数据质量的全程监控和自动预警。

4.2 资产管理

数据资产是经过资产化处理的数据资源,可以为行业内部和外部的数据使用者提供价值。数据资产的特点是同一资产可以供多个使用者使用。要实现资产的复用,首先要对数据资源编目,基于数据分类分级管理数据资产,以便数据使用者能够了解

数据资产明细以及支持的应用场景。

4.2.1 资源编目 数据资源编目对健康医疗数据登记、归类,是对已采集的数据资源描述、组织和管理的过程,是数据开放共享的基础准备工作。按照一定分类架构和标准,序化形成的健康医疗数据资源目录能在一定程度上解决信息资源来源、渠道、管理、应用等问题^[7-9]。数据资源编目主要包含元数据管理和层次结构管理两部分内容。(1) 元数据管理。维护数据资源的台账属性,数据资产化率的提高取决于数据台账属性的完整程度,核心数据台账属性包括标识、发布者、名称、摘要、关键词、数据字段、发布日期、版本、分类、维护和更新频率、限制信息、维护信息、安全限制分级等。(2) 层次结构管理。依据业务主题域构建分类框架,组织管理数据资源,形成不同的类目,例如统计类、建设类和调查类数据集等,产生目录层级,并制定相应的编码规则,然后与元数据关联,帮助用户理解数据的上下文,提高数据的可用性和可管理性。

4.2.2 分类分级 数据分类分级是按数据的内涵、重要程度以及对数据主体可能造成的损害和影响级别,对数据资源划分和分类的过程,目的是对健康医疗数据进行安全保护,为数据资源使用权限的分配提供依据,提高数据资源的安全性。依据 GB/T 39725—2020《信息安全技术 健康医疗数据安全指南》的要求,健康医疗数据可以分为 5 级 6 类,包括个人属性数据、健康状况数据、医疗应用数据、医疗支付数据、卫生资源数据、公共卫生数据。

4.2.3 资产认定 并非所有的数据都是数据资产,只有能够产生价值的数据资源才被认定为数据资产。数据资产认定应基于需求驱动,建议由业务部门发起,类比无形资产的评估方法,综合归纳已有的数据估值方法,形成权威的数据资产价值评估模型或参考模型^[10],并从价值和质量双重维度评价,确认数据资源达到资产化标准后,再用于开发数据产品,进入数据产品上架流程。(1) 数据资产评估。采用定量和定性分析结合的方法,分析数据资产的业务价值和应用场景,核验数据质量(完整性、准确性、一致性、时效性等),并评估数据使

用的潜在风险,确定数据的适用性,制定适当的数据使用管理制度。(2)数据产品开发。基于数据资产评估的结果,根据不同场景的数据应用需求,提取、清洗和整合原始数据,开发具体的数据产品,包括数据结构设计、实体关系模型、数据增量变更、业务维度视图等内容,为用户提供能解决特定业务问题的有价值的结果数据。

4.3 产品上架

经过资产评估后的数据产品上架到数据产品超市,对其开展集中管理。数据管理方通过数据产品超市面向最终用户提供直观的数据产品浏览功能。用户可以检索和申请所需数据产品,经过管理者审批获取对特定数据产品的访问权限,确保数据使用的安全性和合规性。这一过程使数据的管理、共享和使用更加透明、高效。

4.3.1 产品选用 为了便于用户快速定位所需数据产品,应基于搜索引擎搭建数据产品超市门户,提供强大的检索功能、友好的展示界面以及详细的样例信息,旨在为用户提供一站式数据产品选用服务,使用户在海量数据中迅速找到并申请使用符合其需求的数据产品,提高数据产品的可知性和可用性。数据产品选用主要提供以下功能。(1)数据产品检索。使用关键词、标签、分类分级等检索条件查找特定的数据产品,支持复合检索和全文检索。(2)数据产品查看。查看数据产品的元数据信息和使用说明,包括名称、描述、来源、数据量、发布时间、更新频率、更新方式、数据结构、关联关系、数据索引等。(3)数据样例预览。预览和下载数据产品的样例数据,便于直观了解数据的内容和格式,评估数据的质量和可用性。

4.3.2 使用管理 加工后的数据产品可能仍包含重要的业务信息和敏感信息,因此数据产品的使用需要建立申请和审批流程。数据使用方发起数据使用申请,并根据数据最小化原则选择适当的去标识化处理措施。数据去标识化是在不影响数据统一特性和分析准确性的前提下,通过数据变形方式处理原始数据中的敏感信息,进而降低数据敏感程度和减少个人隐私风险的一种处理技术^[11]。数据使用管

理主要包含使用审批和隐私保护配置两部分内容。

(1)数据使用审批。使用方可以提交使用申请,查看审批状态和历史申请记录。申请时需说明应用场景、使用范围、使用时限、使用量和隐私保护策略,支持提交临时和长期数据使用申请。管理者可以对数据使用申请进行通过、拒绝、撤销授权等审批处理,这个过程建议由信息部门和业务部门协同处理,审批结果及时通知用户。(2)隐私保护配置。提供多种数据去标识化处理策略配置功能,并基于数据分类分级给出符合共享安全要求的默认配置建议。去标识化策略包括脱敏、加密、散列值、令牌化、泛化等,以满足不同场景使用需求。

4.4 采样分析

健康医疗数据开放共享过程中应提供数据采样分析功能,以满足各种利用需求。数据采样是从大规模数据集中抽取一个相对小的子集,进行分析、研究或测试的过程。目的是减少数据处理的复杂性,节省计算资源,提高分析效率。数据采样可以在不明显增加采集成本的条件下尽可能提高数据的质量,科学合理地抽样采集,减少不必要的数据采集,帮助简化问题,提高模型的训练速度。

为确保采样结果仍然能够代表原始数据的特征和分布,在不丢失重要信息的前提下减少分析数据量,数据采样时应选择合适的采样方法和采样比例。系统的采样功能支持简单采样、系统采样、分层采样、聚类采样等多种算法,数据使用者可以根据使用目的和应用场景选择合适的算法,配置采样参数,如采样间隔和分层维度等,以使样本数据的分布更加均衡,更能反映整体数据特征。

对于抽取的样本数据,使用探查组件获取数据的全方位特征,探查结果可以为进一步采样分析提供重要决策信息。例如,若样本数据中存在大量数值缺失,用户可以考虑删除字段或者重新抽取一批样本;若数据分布呈现明显的偏斜,用户可以结合样本的分层特征针对性地处理数据。

4.5 在线沙箱处理

数据沙箱是通过构建一个安全虚拟的环境,将

数据应用、存储和个人分开,通过对双方的通信限制保证应用安全性^[12],目的是保护敏感数据。构建数据沙箱计算分析环境,能实现健康医疗数据所有权和使用权分离,确保分析安全可控以及使用合规,并隔离不同用户的数据处理过程。

数据沙箱需要在专用的安全空间内搭建,采用虚拟化私有网络技术,将沙箱与数据中心的网络完全隔离,防止沙箱内的操作影响生产环境。此外,还要综合使用防火墙、入侵检测与防御系统等安全工具,阻止外部恶意软件的入侵和传播。

在安全空间内,用户可以导入外部数据并进行完全自主地处理。数据沙箱中的各种数据分析操作依赖特定的处理引擎完成,引擎需支持主流的数据处理语言和语法表达式。系统基于数据处理引擎提供数据分析任务和数据处理模块的配置管理功能。通过使用模块化的数据处理组件,用户可以对数据进行清洗、筛选、排序、数值计算、字符串处理、数据格式化、数据令牌化以及聚合分析等操作,以满足不同的数据处理需求。

用户与沙箱的人机交互借助安全开发环境和在线分析工具完成,以严格控制并记录用户在沙箱中的操作行为,确保数据的隐私和安全,具体措施包括用户访问控制、安全审计和追溯、数据存储加密、数据复制限制、数字水印等。

5 应用案例与成效

浙江基于健康医疗数据开放共享系统建设汇集了丰富的健康医疗数据资源,并为数据使用方提供了一个安全可靠的数据支撑环境。数据集成治理方面,通过建设省域健康数据高铁,实现省、市、县、乡、村 5 级医疗卫生机构信息系统互联互通,累计覆盖二级及以上公立医疗机构 338 家,覆盖率 100%;覆盖基层医疗卫生机构 1 413 家,覆盖率 100%。数据整体质量提升 25%,采集时效从天级提升到秒级。数据资产管理方面,初步建立省、市、县 3 级数据资源目录,涵盖医卫资源、医疗运营、公共卫生、人口发展、政务服务、互联网服务 6 大类目的 715 个一级目录和 7 694 个二级类目,实

现资源目录化、目录可视化。数据分类分级为数据的开放共享奠定了坚实基础。实际应用方面,浙江省卫生健康委员会联合浙江省数据局、杭州市数据资源管理局以及信息企业,以“原始数据不出域、数据可用不可见”为原则,依托数据开放共享系统能力,集成数据、算法、算力优势,创新研发医学人工智能大模型,已应用于基层辅助诊断、健康咨询、健康宣教等领域。基于数字医疗资源整合与数据开放共享,全省开展医学检查检验结果互认共享改革(“浙医互认”改革),大大提升公众看病就医的获得感和满意度。

6 结语

健康医疗数据开放共享是健康医疗产业转型升级的重要方向。建设健康医疗数据开放共享系统,整合数字医疗资源,可以有效解决资源分散、孤岛林立、安全保护薄弱、分析手段落后等问题,在精准医疗、疾病预防、公共卫生管理等方面为医疗卫生机构提供有力支撑,促进医疗质量快速提升、医疗资源优化配置和健康产业稳健发展。

本研究提出的区域健康医疗数据开放共享系统具备一定的创新性和可推广性,但仍存在一些不足。一是数据产品超市的上架对象目前仅考虑了数据类产品,对于数据模型类和数据服务类产品的研究有待深入;二是尚未考虑针对在线沙箱处理结果的数据产品转化,沙箱运行结果的数据价值未能得到充分释放。下一步将深入研究数据模型和数据服务的产品转化体系,设计沙箱运行结果的数据利用途径。

作者贡献:郭一负责架构设计和论文修订;胡盈盈负责论文修订;胡斐负责论文撰写;沈扬阳负责资料收集;许杰负责现状分析。

利益声明:所有作者均声明不存在利益冲突。

参考文献

- 1 关于促进和规范健康医疗大数据应用发展的指导意见 [EB/OL]. [2023 - 06 - 24]. http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.

(下转第 95 页)

2

赵子建. 基于步态分析的运动障碍疾病诊断恢复辅助系统研究 [D]. 杭州: 浙江大学, 2021.

3

张洪超. 基于惯性传感器的人体动作捕捉与识别方法研究 [D]. 郑州: 河南工业大学, 2021.

4

赵杰. 动态捕捉五大技术种类及光学式运动捕捉数据的实用技巧 (上) [J]. 电视字幕 (特技与动画), 2009, 15 (2): 30-32.

5

吴文辉. 光学运动捕捉系统标定机器人的研究与设计 [D]. 镇江: 江苏科技大学, 2021.

6

冉冀. 基于惯性传感器的人体动作捕捉与控制 [D]. 武汉: 华中科技大学, 2017.

7

约瑟夫·霍金. Unity5 实战 使用 C#和 Unity 开发多平台游戏 [M]. 蔡俊鸿, 译. 北京: 清华大学出版社, 2016.

8

雷全智. 基于 UNITY 技术的三维课件开发与实现 [J]. 南昌师范学院学报, 2018, 39 (6): 25-28.

9

汤健, 欧长伟, 谢能刚, 等. 基于 Unity3D 康复机器人数字孪生模型的实时驱动 [J]. 安徽工业大学学报 (自然科学版), 2023, 40 (1): 65-69, 88.

10

张沛业. 基于多传感器信息融合的人体姿态解算算法研究 [D]. 长春: 吉林农业大学, 2021.

11

芮国胜, 张洋, 史特, 等. 基于卡尔曼增益的 Duffing 系统状态预测算法 [J]. 宇航学报, 2012, 33 (8): 1144-1149.

12

戴欢, 赵卫华, 王威, 等. 基于惯性动作捕捉的人体运动姿态分析 [J]. 科技广场, 2014 (8): 13-18.

13

杨波. 基于 MEMS 惯性传感器动作捕捉系统与轨迹追踪的研究设计 [D]. 成都: 电子科技大学, 2016.

14

YAN L, SIGAL L, BLACK M J. Realanad simulated upper body tracking with annealing particle filter and belief propagation for human - robot interaction [J]. International journal of humanoid robotics, 2011, 8 (1): 127-146.

15

CHAN K L. Video - based gait analysis by silhouette chamfer distance and kalman filter [J]. International journal of image and graphics, 2008, 8 (3): 383-418.

16

赵正旭, 戴欢, 赵文彬, 等. 基于惯性动作捕捉的人体运动姿态模拟 [J]. 计算机工程, 2012, 38 (5): 5-8.

17

柳金辉. 运动捕捉技术在三维动画设计中的应用及发展前景 [J]. 济宁学院学报, 2024, 45 (1): 90-95.

18

伍孝彬. 基于数字化的芭蕾基训教学演示系统研究 [J]. 大观 (论坛), 2024 (1): 149-151.

19

雷舒慧. 下肢康复运动生理参数动态检测系统设计 [D]. 重庆: 重庆大学, 2022.

(上接第 88 页)

2

“十四五”全民健康信息化规划 [EB/OL]. [2023-06-24]. <http://www.nhc.gov.cn/guihuaxxs/s3585u/202211/49eb570ca79a42f688f9efac42e3c0f1.shtml>.

3

魏明月, 陈敏, 胥婷, 等. 临床科研场景下医疗数据安全开放共享平台设计 [J]. 中国数字医学, 2021, 16 (7): 27-32.

4

李雪迎, 沙若琪, 姚晨, 等. 面向真实世界数据的临床研究数据治理模式选择 [J]. 中国循证医学杂志, 2020, 20 (10): 1150-1156.

5

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.

6

黄乙中. 浅谈数据治理建设方案 [J]. 轻工科技, 2020, 36 (1): 65-67.

7

孟群, 胡建平, 董方杰, 等. 我国健康医疗大数据资源目录体系建设研究 [J]. 中国卫生信息管理杂志, 2017, 14 (3): 387-391.

8

陈敏, 黄竹青, 秦健. 湖北省健康医疗大数据资源目录体系建设研究 [J]. 中国医院管理, 2018, 38 (12): 67-68.

9

李岳峰, 胡建平, 张学高. 中国健康医疗大数据资源目录体系与技术架构研究 [J]. 中国卫生信息管理杂志, 2019, 16 (3): 249-256.

10

张志刚, 杨栋枢, 吴红侠. 数据资产价值评估模型研究与应用 [J]. 现代电子技术, 2015, 38 (20): 44-47, 51.

11

王卓, 刘国伟, 王岩, 等. 数据脱敏技术发展现状及趋势研究 [J]. 信息通信技术与政策, 2020 (4): 18-22.

12

王俊. 基于安全沙箱的医院移动终端数据安全防护 [J]. 中国数字医学, 2018, 13 (3): 96-98.