医疗大数据集成及应用平台体系构建*

王觅也 郑 涛 李 楠 张 睿

（1 四川大学华西医院 成都 610041
2 四川大学华西医院医疗信息化技术教育部
工程研究中心 成都 610041）

（四川大学华西医院 成都 610041）

师庆科

（1 四川大学华西医院 成都 610041
2 四川大学华西医院医疗信息化技术教育部
工程研究中心 成都 610041）

〔摘要〕 以四川大学华西医院为例，在介绍医疗大数据集成应用现状基础上，提出“云平台＋数据资源＋应用”的大数据集成平台构建方案，介绍平台系统设计与应用效果，推动医疗大数据应用发展。

〔关键词〕 医疗大数据；集成平台；医学术语标准；数据治理；数据共享；数据资源
〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.08.008

Building of Medical Big Data Integration and Application Platform System *WANG Miye, 1West China Hospital of Sichuan University, Chengdu 610041, China; 2Engineering Research Center of Medical Information Technology, Ministry of Education, West China Hospital of Sichuan University, Chengdu 610041, China; ZHENG Tao, LI Nan, ZHANG Rui, West China Hospital of Sichuan University, Chengdu 610041, China; Shi Qingke, 1West China Hospital of Sichuan University, Chengdu 610041, China; 2Engineering Research Center of Medical Information Technology, Ministry of Education, West China Hospital of Sichuan University, Chengdu 610041, China*

〔**Abstract**〕 Taking West China Hospital of Sichuan University as an example, the paper puts forward the building plan of the big data integration platform featuring " cloud platform + data resource + application" on the basis of introducing the integrated application status of medical big data, and introduces system design and application effect of the platform to promote the development of medical big data application.

〔**Keywords**〕 medical big data; integration platform; medical terminology standard; data governance; data sharing; data resource

1 引言

信息技术的高速发展产生海量数据，把人们带入大数据时代[1]。根据冯麟对国内外医学大数据文

〔修回日期〕 2019－06－04
〔作者简介〕 王觅也，工程师，发表论文及参编论著 3 篇。
〔基金项目〕 省科技厅支撑计划项目“基于云平台的医联体分级医疗协同模式研究与应用示范”（项目编号：2019YFS0034）。

献的分析结果，国外高频次关键词有大数据、电子病历、数据挖掘、开放数据；国内高频次关键词有大数据、物联网、医疗健康、卫生信息化等[2]。自 2012 年奥巴马政府宣布投资 2 亿美元启动“大数据研究和发展计划”开始[3]，大数据成为美国的重要战略目标，联邦政府通过各种政策和倡议鼓励健康医疗数据的使用；2015 年 9 月美国国家医疗信息技术协作办公室发布《美国联邦政府医疗信息化战略规划（2015－2020）》，明确提出“信息收集－信息共享－信息利用”的战略路线[4]。在国内，2015 年 8 月《国务院关于印发促进大数据发展行动纲要的通知》将医疗健康大数据列为大数据工程之一；2016 年 10 月国务院印发的《健康中国 2030 规划纲要》中将“推进健康医疗大数据应用”列为重点内容，强调要实现公共卫生、医疗服务、医疗保障、综合管理等应用系统的数据采集、集成共享和业务协同[5]。国内外的政策都指明大数据的整体思路：从数据采集到数据集成，再到数据应用。因此，本文提出基于医疗机构的大数据平台构建也应该是个系统工程，整合数据采集、集成以及挖掘利用的技术、工具和资源，打造从业务系统产生数据、数据平台分析数据产生知识、知识库辅助决策反哺业务系统这样一个良性循环的数据环。

2 医疗大数据集成应用

2.1 医院信息化总体情况

四川大学华西医院的全院级电子病历系统上线于 2007 年，迄今已有 11 年历史。现有业务系统超过 100 个，业务协同通过 Ensemble 消息总线实现，已积累超过 500T 的数据资源，其中医院信息系统数据量约 4T，而大部分数据资源的形态为非结构化数据。

2.2 医院大数据集成应用现状

过去 10 年的信息化建设是医院系统数字化的历程，每个专业领域都选择国内顶级的系统开发商，因此在院内形成一个个独立的应用系统，为保证业务流程的连续性，众多系统之间基于消息总线开发了诸多接口，维护量大，异常事务时有发生。因此亟需构建基于数据标准的数据中心，将医疗数据统一存储，支撑以数据为中心的业务互联互通。四川大学华西医院每天住院人数 4 500 左右，门诊量 1.3 万，急诊人数 500 左右，数据量大，对于历史数据查询和计算速度很慢，极端情况下会影响正常业务运行速度，针对大量数据的查询和计算均放在数据仓库中。但现有数据仓库是基于传统数据库架构，查询失败的情况时有发生，因此亟需打造基于 Hadoop 的数据资源存储和计算平台来解决这类问题。从现有数据资源的特征分析，超过 90% 的数据都是非结构化数据，如病历文书等文本资料、影像图像、病理图像、超声视频、手术视频等，对这类数据的分析必须借助大数据处理技术，如自然语言处理、图像识别、语音识别等，因此亟需构建一整套的数据治理方法和技术，使得医疗数据真正可以被识别、可以被分析，同时必须遵循一定的标准去加工和处理数据，使得被分析的数据是高质量、可复用、可检验的。之前对数据孤岛的定义大多针对业务系统体系，随着数据应用系统的发展和构建，孤岛现象也逐渐出现在数据应用体系中，如针对手术间资源配置的分析决策系统数据分析方法和分析结果无法被平移应用到类似的其他系统中。因此，构建统一平台，将所有数据类的分析方法、处理技术、分析模型和分析结果有机整合，使所有原始、二次加工和分析结果数据均在一个平台环境内流动，除可解决新的数据孤岛问题以外，还可将科研产出结果在平台上转化，应用于实际的临床和管理活动。

2.3 大数据集成与应用平台体系建设目标

以上诸多技术瓶颈和系统应用现状，极大影响医院大数据资源的开发和利用。因此亟需创造一个健康医疗大数据的生态体系，构建一个“云平台＋数据资源＋应用”的立体服务模式。通过此大数据生态体系建设，打破信息孤岛，搭建数据桥梁，打通产生、采集、治理、统一、交换、分析、反哺应用等整个链条的全部环节，使数据真正成为资源，发挥出应有的价值。

3 医疗大数据集成与应用平台系统设计

3.1 总体架构

医疗大数据集成及应用平台体系的核心是数据，关键技术包括数据标准化、数据治理、数据分析利用和数据安全等。平台的总体架构分为 7 个模块：基于 Hadoop 的云平台、数据集成与治理、应用支撑功能集、数据资源仓储、应用系统服务平台、数据标准管理和数据安全管理，见图 1。

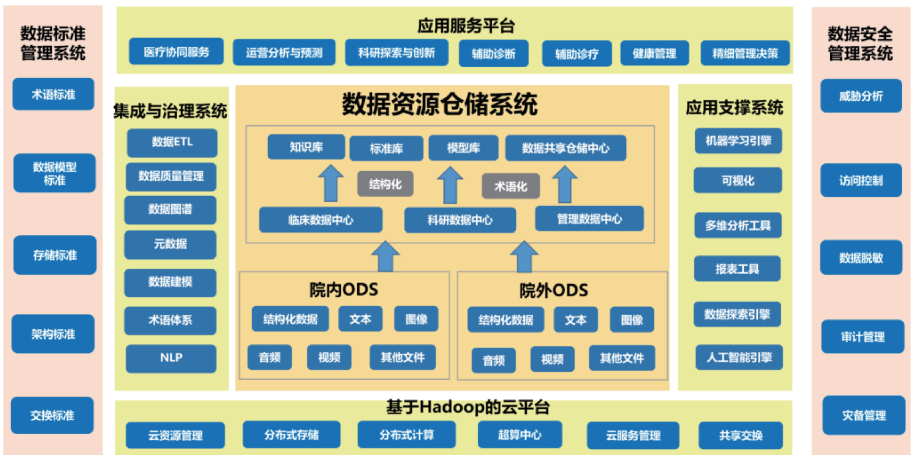

图 1 大数据集成及应用平台系统架构设计

3.2 基于 Hadoop 的云平台功能

基于 Hadoop 的云平台主要包括计算资源、存储、信息安全、网络系统及机房工程等建设，以及云服务平台。与传统数据集成平台不同的是，使用 Hadoop 技术可支持 PB 级海量数据的存储及高效计算，通过云模式，融合新型的分布式计算和传统的并行计算技术，实现数据的高效交换，资源的弹性供给，解决医院海量数据利用低效的问题。云平台的另一强大功能是支撑数据共享和服务共享能力。数据共享技术包括医疗行业数据共享标准转化、共享数据智能识别、医疗元数据管理、共享数据安全管理技术等；遵照现行的共享标准，抽象出独立的共享数据识别系统，通过共享交换引擎将外部共享数据识别并转换，整合形成院内院外的整合数据集。服务共享能力强调服务封装和服务共享，在形成数据产品的过程中所产生的知识模型、数据处理模型等，云平台可将其封装成服务，为其他应用产品所调用和共享。

3.3 数据集成与治理系统功能

3.3.1 概述 完成数据采集、数据治理、数据管理等数据制备工作，需包含数据抽取、转换、装载管理、元数据管理、数据质量管理、语义解析、数据整合、迁移、建模工具和可视化等通用性功能，是将杂乱无章的数据规整为高可用数据的基础。

3.3.2 数据治理是数据管理的核心 其关键技术包括元数据管理、数据图谱、主数据管理、主索引管理、自然语言处理等。目前，常规的数据治理技术在应对现有复杂度极高的结构、半结构和非结构化医疗数据面前多少有点捉襟见肘。而在大数据集与治理系统中，构建具有图谱关系的元数据网络，由信息技术人员、标注人员、临床专家、管理专家等设定符合数据特征及相关关系的治理策略，在保证数据资源完整、统一且可追溯的前提下，利用分布式处理技术的优势可最终实现医疗数据的治理。

3.3.3 医学术语是数据治理的最有效方法 医学系统命名法－临床术语（SNOMED CT）是国际主流术语集，用其进行术语概念表达是主流方式。基于医学术语的数据治理，可以将诸多标准不一、规范不一的数据，通过术语网络，关联临床数据、药品、检验、检查等相关医疗数据，最终完整实现数据集成交互的目标。利用 SNOMED CT 术语集与自然语言处理算法，通过临床专家标注与机器训练，帮助各术语

形成医学语义网，将完整的临床资料用于数据服务和知识表达，使得临床文本可被机器识别。

3.4 应用支撑系统

集成数据应用所需的通用工具包，包括机器学习、报表开发工具、多维分析工具、实时查询工具、科研探索等，支撑上层应用。同时也需集成医疗数据处理分析的一些特殊工具，包括影像处理、语音解析、医学术语处理和医疗人工智能等。应用支撑系统更偏重于专业领域的数据处理专有能力，还有部分是在数据集成层通用能力上二次开发后的专属能力，是形成高可用数据资源的重要工具集。医疗人工智能的关键技术包括知识表示、自然语言理解、机器学习、知识获取、知识处理系统、计算机视觉、自动推理和搜索方法、智能机器人、自动程序设计、专家系统等[6]。大数据应用支撑系统将各类医疗人工智能算法和模型统一存储和管理，以健康医疗元数据为核心组织数据图谱网络，搭配医学、管理、计算机等行业专家参与的机器学习及深度学习后形成海量数据标签，在图形处理器（Graphics Processing Unit，GPU）集群环境下形成较大规模的计算能力，由不同行业专家组成的算法研发团队，在统一的环境下快速、安全、高效的分析和挖掘医疗数据的价值，形成专家知识库。

3.5 数据资源仓储系统

关键技术包括领域内的数据模型设计、立体的数据中心设计、同步与异步的数据抽取技术以及临床术语化等。数据资源层的核心是海量的健康医疗数据，基于 HL7 的领域模型构建主体框架，以实际需求出发进行合理的调整和改造，形成符合国情、行业所需的数据模型。利用数据仓库的特点，将可追溯、数据血缘图谱植入数据仓库中，完成立体的数据仓储中心建设。数据资源仓储系统中存储的数据既包括贴近业务数据源的数据，也包括以某类标准整合的类目数据，还包括以需求为宗旨的数据集市数据。既包括院内数据，也包括院外数据。通过多源异构数据整合的医疗数据资源中心建设，保障医疗数据的互联互通，消除信息孤岛；通过数据共享，实现对外信息交互。数据资源仓储系统按不同分析类目而分模块独立建模、分类存储，模块包括临床、管理、科研数据中心等，基于数据利用目的和参照信息标准进行数据建模。除存储结构化数据，还需利用自然语义算法将电子病历中的非结构化文本解析为结构化数据，便于临床科研的分析利用。除存储诊疗过程产生的数据，还需存储基因组学等组学数据，与临床症状、临床诊断等信息结合分析，开展精准医疗，最终实现对于疾病和特定患者进行个性化精准治疗的目的，提高疾病诊治与预防的效益。随着大数据技术的应用发展，临床数据资源中的文件格式多种多样，除标准的数据、文本文件，还有图形文件，如放射影像图片，可支撑临床医生基于薄层影像图片进行三维重建等。

3.6 应用服务平台

是大数据集成及应用体系的门户窗口，应用越丰富，数据应用产品越多，证明这个体系越成功。数据应用的需求是无止境的，从类型上来说，包括医院精细化管理、临床研究创新、卫生经济研究、医疗业务协同、临床智能辅助决策等。调用云平台各种服务和数据资源平台的高可用数据，开发应用产品，将产出成果再次通过云平台反哺回医疗业务活动中。面向医疗服务的数据产品包括医疗协同服务、患者统一视图、慢病管理系统等；面向临床科研的数据产品包括专科病种库及随访、科研探索系统等，结合人工智能技术还可以开发疾病风险预测、医学影像辅助诊断、临床辅助诊疗、虚拟助理系统等；面向医院管理的数据产品包括医院运行管理指标监控、医疗质控预警、医院资源消耗预测、监管数据交换与上报系统等。

3.7 数据标准管理系统

参考和应用的标准包括多方面：数据交换、数据全局术语、数据存储模型以及安全标准等。采用 HL7、Dicom 标准以及原卫计委颁发的数据集标准作为信息系统数据交换的标准。以 SNOMED CT 医学术语集为核心，医学一体化语言系统（Unified Medical Language System，UMLS）为辅助，将全局

数据进行串联形成语义网。针对数据资源层的数据存储和数据模型的标准，主要根据医疗数据的特点有针对性地纳入相应的标准，以便更好地与核心 SNOMED CT 对接。其中，数据模型重点参照 HL7 的核心部分。针对药品相关的数据模型分类的设计，重点参考和遵循 ATC 标准。对于检验数据，以 LOINC 作为数据的编码系统来组织相关数据；对于检验数据，以 DICOM 协议作为支撑数据交换的标准，遵循 RadLex 作为放射术语的桥梁。数据安全标准方面，兼顾《大数据安全标准化白皮书》、《个人信息安全规范》、《信息系统安全等级保护基本要求》等国内标准，以及参考美国的《健康保险携带和责任法案》少量内容和欧盟与 2018 年最新的《通用数据保护条例》（GDPR）等国际标准。

3.8 数据安全管理系统

既包括存储、数据、系统和网络的安全，同时还要配合相应管理制度确保安全合理和有效执行。整体安全设计的理念是：越往底层，数据安全与应用耦合性越低；越往应用上层，数据安全与应用场景交互的耦合性越高。具体要求为：存储在物理介质上的所有数据，根据分类和用途，进行物理级别的加密；对集中在数据资源中心的数据，基于数据本身进行数据分级分类的安全管控；系统安全管理方面，包括系统权限、双验证授权、警卫岗权、紧急控制权等管控；应用服务安全方面，可分为角色权限、服务授权、访问控制、部署安全和数据销毁等；网络及环境安全方面，包括网络安全、流量监控、数据审计和环境监控等；安全管控制度方面，包括但不限于敏感数据、规范流程、日志审计、权限复核管理等。

4 应用效果

4.1 概述

四川大学华西医院利用超过 1 年的时间设计该平台体系，于 2018 年下半年启动该平台体系的落地建设工作，与多家 IT 服务厂商共同打造该平台。迄今为止，7 个模块的系统建设工作均在紧张进行中，已有部分系统初见成效。整个平台体系是一个持续建设的宏大工程，预计 1 年周期能够打造初版平台成型，随着应用需求的新增和技术的更迭，平台也将日益强健。该平台的建设将使医院信息化水平再上一个新的台阶，极大地提升医院大数据资源应用的整体能力。

4.2 数据治理

海量的医疗数据结合 SNOMED CT 术语集，基于医学自然语言处理技术和医学知识图谱技术，实现住院病案的临床表征、临床诊断、治疗手段等信息进行精准化提取，形成结构化数据集。经过合理的数据治理为不同需求的用户提供具有完整统一的、可追溯的健康医疗数据的生态体系。用户既可以关注数据的本身，也可以关注数据与数据间的血缘关系；既可以使用元数据属性寻求更多的数据标签，也可以使用关系辨别主数据的耦合性。打通院内系统与系统间的孤岛或烟囱，实现医院与医院间数据概念层的整合统一。

4.3 大数据资源利用

4.3.1 医院管理数据资源中心 目前通过大数据治理，已形成 3 大数据资源中心。医院管理数据资源中心主要以管理需求为目标组织数据，围绕医院资源、收支、绩效、成本效益、医保、合理用药、医疗质量、护理管理等方面整合数据，为医院全面运营管理提供全方位数据，为精细化管理和管理决策提供有力的数据支撑。已经落地的典型案例包括医院运营指标集系统，在线核心指标接近 300 个[7]，下一步计划将这些指标与业务过程实时交互，通过机器学习方法发现指标异常的原因，提供更优的路径和方法及时进行管理干预。

4.3.2 临床数据资源中心 主要以患者为核心组织数据，归集基本资料、家庭信息、家族患病史、过敏史、月经史、生育史、历史诊疗记录、体格检查、检查检验记录、检查影像数据、病程记录、医嘱记录、费用记录、手术记录、诊断信息、随访信息、组织标本信息、生物信息等。通过 NLP 算法和临床术语集将病历文本资料转化为结构化内容二次

标化存储，为数据互联互通和挖掘利用打下基础。大数据在临床上的应用主要包括临床决策支持、患者行为管理、合理用药、疾病风险预测和慢病规范治疗监控等，近期计划落地的典型案例包括基于时间轴的患者统一视图、专科视图和病历相似度匹配查询等应用。患者统一视图以患者为中心，从时间维度、诊疗事件维度、主要疾病和健康问题维度等3 个维度构成的立体视图，进行全生命周期的纵向临床记录浏览，关注患者的整体健康状况和临床信息。病历相似度匹配通过诊断、检验结果、主要体征和症状等多种条件进行组合的语义级查询，方便比较相近患者的临床记录，从而为医生诊断提供参考。

4.3.3 科研数据资源中心 主要以病种为核心组织数据，构建队列数据集，除集成临床过程产生的信息，还通过随访系统整合病例的随访资料，支持临床科研的回顾性和前瞻性两种类型的研究等。在大数据时代为医生提供全量数据集，以及已经处理好的各种病种特征和分析方法，便于医生在数据中发掘医学知识，发现数据规律。近期计划落地的典型案例包括通过科研项目引领的方式，构建国内甚至国际重大疾病、罕见病的病种数据库，为疑难疾病的科学研究和治疗手段评估等夯实数据基石。开发科研探索平台，将数据资源、分析方法和模型作为可选功能集成共享，支持医生进行探索性知识发现。

4.4 大数据共享

健康医疗大数据的协作和共享，从共享内容来看，主要是数据资源和数据服务；从共享范围来看，体现在单个机构内、跨机构之间，乃至行业之间的共享。当前阶段集成范围主要针对院内数据的治理和共享，已初步实现对不同业务系统之间的数据共享支持和数据服务支持。下一阶段，在平台内计划纳入医联体内其他医疗机构的数据，进行数据标准化、数据治理和数据服务，实现医联体内的信息共享、业务协同、数据共研，同时，基于各种大数据资源，打造由政府主导、行业关联、机构联盟、科室结合、居民参与的医疗服务模式，以互联网、移动、物联设备等为载体，通过云平台将所有角色进行有机串接，在确保安全的前提下最终实现跨机构、跨行业的全面协作和共享。将不同的工具和技术形成服务引擎，以云平台的形式进行部署，实现技术资源和数据资源的共享。

5 结语

医疗大数据的价值开发还处于初级阶段，无论是管理还是临床，由经验决策向数据决策转型是必然的趋势。大数据应用的丰富与拓展，离不开集成平台的建设。只有通过平台将医疗信息治理成为高质量高可用的数据，再结合数据挖掘方法和共享技术，才能实现医疗数据资源的高效利用，响应国家政策，研发系列技术产品，开展综合性应用示范，以推动医疗卫生行业大数据应用和带动医疗卫生大数据产业的发展。

参考文献

1 陈国青，吴刚，顾远东，等. 管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向［J］. 管理科学学报，2018（7）：1－10.

2 冯麟，雷罗，罗爱静. 基于文献的国内外医学大数据研究［J］. 医学信息学杂志，2015，36（5）：15－21.

3 迪莉娅. 我国大数据产业发展研究［J］. 科技进步与对策，2014（4）：56－60.

4 舒婷，梁铭会. 美国联邦政府医疗信息化战略规划（2015－2020）内容简析［J］. 中国数字医学，2015，10（2）：2－4.

5 新华社. 中共中央 国务院印发《“健康中国2030”规划纲要》［DB/OL］.［2018－10－25］. http://www.gov.cn/zhengce/2016－10/25/content_5124174.htm.

6 曹艳林，王将军，陈璞，等. 人工智能对医疗服务的机遇与挑战［J］. 中国医院，2018，22（6）：25－28.

7 王觅也，郑涛，张睿，等. 医院运行指标集管理系统构建［J］. 医学信息学杂志，2016，37（6）：26－31.

8 于广军，杨佳泓. 医疗大数据［M］. 上海：科学技术出版社，2015.