

医疗大数据平台建设机遇、挑战及其发展^{*}

郭 强

(国家癌症中心 国家肿瘤临床医学研究中心 中国医学科学院/北京协和医学院肿瘤医院 北京 100021)

王 丛

衡反修

(北京信诺中桥咨询有限公司 北京 100016) (北京大学肿瘤医院暨北京市肿瘤防治研究所 恶性肿瘤发病机制及转化研究教育部重点实验室 北京 100142)

〔摘要〕 分析医疗大数据平台建设面临的机遇和挑战,以肿瘤大数据平台为例阐述医疗大数据平台发展建设情况,包括建设规划、关键技术领域研究、建设实践等,为医疗卫生行业数字化、智能化转型提供支撑。

〔关键词〕 医疗大数据;肿瘤大数据;数据治理

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.01.001

Building of Medical Big Data Platform: opportunities, challenges and development GUO Qiang, National Cancer Center, National Clinical Research Center for Cancer, Cancer Hospital of Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China; WANG Cong, Beijing SINO Zhongqiao Consultant Co., Ltd., Beijing 100016, China; HENG Fanxiu, Beijing Cancer Hospital, Beijing Institute For Cancer Research, Key Laboratory of Carcinogenesis and Translational Research of Ministry of Education, Beijing 100142, China

〔Abstract〕 The paper analyzes the opportunities and challenges faced by the building of medical big data platform. Taking the cancer big data platform as an example, it expounds the building and development situation of the medical big data platform, including the building planning, researches in the key technical fields, building practice, etc., so as to provide support for the digitalization and intelligent transformation of the medical and health industry.

〔Keywords〕 medical big data; cancer big data; data governance

〔收稿日期〕 2020-12-24

〔作者简介〕 郭强,高级工程师,发表论文3篇;通讯作者:衡反修,主任。

〔基金项目〕 国家重点研发计划“数字诊疗装备研发”专项“临床辅助决策专家系统的临床验证与推广”(项目编号:2018YFC0116905)。

1 引言

当今医学与信息科学交叉融合越来越深入和广泛,互联网、物联网、生物信息技术的高速发展推动医疗数据呈爆炸式增长。医疗数据与大众健康息息相关,已上升为国家基础战略资源,其分析和再

利用成为研究热点。建设医疗大数据平台、运用大数据分析和挖掘技术可以完成过去依靠传统思维、方法、技术无法完成的任务,提升医疗服务质量、促进临床医学研究。但是医疗数据具有多源异构、非结构化数据占比高、信息涉及隐私等特点,在数据采集、治理、应用等环节存在挑战。医疗大数据平台建设风险与机遇共存,挑战与发展同在。

2 医疗大数据平台建设机遇

2.1 政策体系趋于完善

2.1.1 政策体系 2015年9月国务院发布的《促进大数据发展行动纲要》阐述了大数据发展形势和重要意义,对大数据发展做了顶层规划和统筹部署,大数据产业正式上升至国家战略层面。2016年6月国务院印发的《关于促进和规范健康医疗大数据应用发展的指导意见》要求夯实健康医疗大数据应用基础,全面深化健康医疗大数据应用。2018年7月为加强健康医疗大数据服务管理,促进“互联网+医疗健康”发展,国家卫生健康委员会印发《关于印发国家健康医疗大数据标准、安全和服务管理办法(试行)的通知》。截至2019年全国各地市相继出台大数据相关政策性文件近200个^[1]。

2.1.2 产业发展情况 顶层规划与行业性、专业性政策前后呼应为大数据产业创造了快速发展条件。2016年第1批国家试点健康医疗大数据中心在福建、江苏启动建设。2017年国家发改委批复13个国家大数据工程实验室。2018-2019年各地政府设立大数据园区109个,大部分省市建立区级以上大数据园区。2019年已有82个省市地级单位数据开放平台上线。全国开放数据集总量62 801个,是2017年的7倍左右^[1]。

2.2 新技术蓬勃发展

2.2.1 大数据技术架构体系不断完善 Hadoop项目启动标志大数据技术时代开始,而亚马逊Web服务(Amazon Web Services, AWS)商用则表明云计算正式开始改变信息时代步伐^[2]。此后大规模集群计算、服务器、处理器芯片、基础软件等方面成熟

商业化产品不断涌现。近年来以Spark和Flink为代表的新计算引擎取代了MapReduce框架,更好地满足数据实时处理需求;Greenplum框架提供较强并行数据计算性能和海量数据管理能力;Impala、Spark SQL等解决结构化数据问题。此外非结构化数据处理技术逐渐兴起。图像识别、语音识别等感知智能技术发展日益成熟,基于算法和深度学习的自然语言处理、知识图谱等认知智能稳步发展,数据整合广度和数据挖掘深度得到质的提升。随着技术发展大数据技术架构体系不断完善。

2.2.2 新一代测序技术发展为精准医疗提供支撑

自二代测序技术推出以来基因测序成本以迅速降低,最初个人全基因组测序费用高达1亿美元,目前一个100Gb数据量的人类基因组只需要1 000美元左右^[3]。随着测序技术进步、检测成本降低,临床医疗数据和组学数据融合成为必然发展趋势,生物医学研究进入大数据时代。

2.3 数据资源优势明显

我国是世界第1人口大国,拥有丰富临床研究资源^[4]。患有心血管系统疾病、肿瘤、糖尿病、呼吸系统疾病等的慢性病人规模超过3亿人^[5]。随着我国人口老龄化现象加剧,环境因素、社会经济发展及居民生活方式改变等,恶性肿瘤发病数和死亡数在未来将持续增长。庞大诊疗人群、不同流行病学和基因组学特征为医疗科研创新提供了丰富可靠的数据支撑。

3 医疗大数据平台建设面临的挑战

3.1 数据标准化

截至2018年我国已累计立项制定卫生信息标准283项,发布标准208项,但是由于卫生信息标准起步晚、起点低,标准执行缺少强制、有效的执行措施,在标准研制与应用方面尚存在问题^[6]。各厂商信息系统设计缺乏规范,各医疗单位数据质量参差不齐,因而难以达到数据互联互通、数据共享要求。中国医院协会信息专业委员会(China Hospital Information Management Association, CHIMA)

《2018-2019年中国医院信息化状况调查报告》显示,60.35%的受访医院认为医疗信息化产品集成缺乏标准、集成困难,39.71%的医院参加电子病历系统应用水平分级评价,14.61%参加互联互通标准化成熟度测评。对全国130家大型三级医院的问卷调研结果显示,其所使用医院信息系统(Hospital Information System, HIS)所属厂商共37家,电子病历(Electronic Medical Record, EMR)系统28家,医学影像存储与传输系统(Picture Archiving and Communication Systems, PACS)46家,实验室信息管理系统(Laboratory Information Management System, LIS)36家,放射信息系统(Radiology Information System, RIS)42家,病案管理系统41家。医院使用信息化产品厂家众多、结构各异,医疗数据不完整、不标准、不规范、不准确、缺乏关联等现象普遍^[7-8],亟需构建标准化数据治理流程和规范。

3.2 数据挖潜

3.2.1 现状 目前约有80%的临床数据为非结构化自由录入文书,例如入院出院记录、影像学诊断报告、病理学诊断报告、手术记录等,大量有价值信息蕴藏其中,但是病历文本上下文信息往往具有多模态、多源特点,再加上受汉字歧义性和原始信息结构不完整等因素影响,现有信息抽取、知识融合和知识加工技术难以满足大数据应用需求,缺乏高效算法和模式支撑对有价值信息的筛选和提取。近年来BERT、XLNet和ERNIE在自然语言处理领域虽然取得重要进展但依然无法较好地处理医学领域的常识和推理问题。而标注数据稀缺、标注过程复杂是阻碍人工智能(Artificial Intelligence, AI)技术发展的重要原因。医疗概念复杂、数据涉及隐私导致数据获取难度较高,数据标注工作周期长、成本高。分析公司Cognilytica在2019年的报告显示大约80%的AI项目时间花费在数据准备上,人工智能项目中只有20%的时间用于算法开发、模型训练和调整以及机器学习运算。

3.2.2 挑战 国内医学术语标准化建设虽然已取得一定成果,但由于起步较晚尚未建立通用术语标

准。信息和知识来源广泛,医学文献、临床指南、循证医学证据、临床证据之间的内容存在不一致和冲突,新的医学概念、药物和治疗方式不断出现,如何统一标准实现知识融合和存储、持续维护和更新保证知识的正确性和时效性面临挑战。

3.3 数据共享和应用安全

3.3.1 数据开放共享需要政策支持和机制创新 目前数据的归属权、使用权以及侵犯个人隐私数据所要承担的法律后果尚不明确^[9],不同区域、行业、部门之间数据壁垒广泛存在,数据公开、流动、交易等运行机制尚需完善。

3.3.2 医疗数据应用环境面临挑战 海量的数据、复杂的外部环境和技术平台的脆弱性使得医疗数据泄露风险大幅增加。2018年9月腾讯《医疗行业勒索病毒专题报告》显示,2018年7-9月在全国共有247家三甲医院信息化系统检出勒索病毒,几乎每个月都会发生3~4起重大医疗数据泄露事故。如何兼顾发展和安全、平衡效率和风险是大数据平台建设和应用过程中需要解决的问题。

4 医疗大数据平台发展:以多中心肿瘤大数据平台为例

4.1 肿瘤大数据平台建设规划

4.1.1 亟须建立肿瘤大数据平台 恶性肿瘤发病率和死亡率较高,已经成为严重威胁人类健康的重大疾病之一。当前我国尚未建立以肿瘤临床数据为基础的多中心研究网络,由于缺少跨区域、多中心的肿瘤发病与诊疗数据,肿瘤防治、临床研究、疗效评价、指南制定等具有较大盲目性^[10]。建立肿瘤大数据平台并进行科学管理、分析和利用是我国肿瘤防治领域亟须解决的问题,其建设应服从于国家“大卫生、大健康”的方针政策,构建高效、统一、协同的多中心乃至全国性的肿瘤大数据平台,实现3级预防和诊疗流程的规范化、科学化管理列为肿瘤预防与治疗数字化工作重点。肿瘤大数据平台建设应该紧紧围绕该项重点开展规划设计。

4.1.2 建设原则 应遵循“1234”原则,见图1。

1 个中心转变：肿瘤领域未来发展将从以治疗为中心向全面健康和肿瘤预防为中心方向转变，从“治病”转向“健康”，强调早诊断、早治疗、早康复，对数据获取范围、渠道和方式都将产生影响。2 个维度：一是服务对象广泛化，未来大数据平台要广泛面对不同群体管理和科研需求，包括政府、高校、医疗机构、药企、保险公司等，数据的完整性、机密性、易用性、真实性将成为平台价值的关键考量；二是技术发展多元化，5G、云计算、大数据、物联网、区块链等新技术簇的发展将构建泛在智能生态，诊疗、科研、管理模式等都将发生颠覆性变化。3 个互联互通：肿瘤大数据平台要与各医疗机构数据互通，能与区域健康信息平台业务互联，能与关键技术研究领域成果共享，平台的标准

化建设是基础也是灵魂，见图 2。4 个环节：一是数据采集环节，实现全维度自动化采集。随着数据标准化和数据采集流程规范化，手工填报或文件上传等被动数据采集方式将被淘汰，支持多方数据聚合的自动化采集技术将成为主流；二是数据治理环节，从碎片化数据治理发展为流程化、自动化治理，平台具备更加灵活的元数据配置管理，采用高可靠、可解释的治理技术和工具完成数据的结构化和标准化；三是数据赋能环节，构建知识图谱和知识库，化数据为信息、知识，形成业务智能；四是数据联动环节，实现开放共享、互动协同，医疗、医保、医药平台与肿瘤大数据平台实现业务和项目联动，提高肿瘤预防和诊疗整体水平。



图 1 肿瘤大数据平台建设规划“1234”原则

4.2 肿瘤大数据平台相关技术领域研究

4.2.1 数据治理是核心任务 数据治理涉及战略、制度流程、技术和标准规范、应用实践等方面，而技术和标准规范是治理体系的关键。

4.2.2 医疗数据采集规范化及标准化 对于跨区域多中心的数据采集平台标准化构建，元数据管理和主索引建立尤为重要，元数据解决数据可读性和一致性问题，主索引解决数据之间关联性问题。元数据管理首先需要根据项目目标、数据范围和特点建立最小数据集，然后以此为基础创建和管理元数据，数据集和元数据的制定应严格参照国际、国内相关标准规范，如《住院病案首页数据填写质量规



图 2 肿瘤大数据平台互联互通

范》、《疾病分类编码 ICD-10》、《手术操作编码 ICD-9》、医学系统命名法-临床术语 (Systematized Nomenclature of Medicine Clinical Terms, SNOMED CT) 等。相对于元数据管理主索引的建立更为复杂,既要通过主索引进行数据精确关联匹配,又要利用身份证号、医保卡号、姓名、日期等进行模糊关联匹配,因此对于不同 HIS 厂商就诊 ID 的生成机制、不同业务系统之间的数据交换方式、患者同一或不同医院多次就诊信息的关联方法、孤立系统数据的读取方法等内容都需要进行专家论证和技术攻关。

4.2.3 医疗数据处理与知识转化 医疗数据面临诸多质量问题,阻碍数据资源价值转化,亟须探索适宜的数据治理技术,传统抽取-转换-加载 (Extract-Transform-Load, ETL) 方法与新兴人工智能技术结合,实现非结构化和半结构化数据的结构化、结构化数据的标准化、标准化数据的资产化。高质量数据变为有用信息和知识是一个循序渐进的过程,数据转化为信息,信息抽象为知识,知识应用于实践并通过数据验证和完善,互为因果,互相促进,最终实现医疗数据处理与知识转化的生态闭环,见图3。

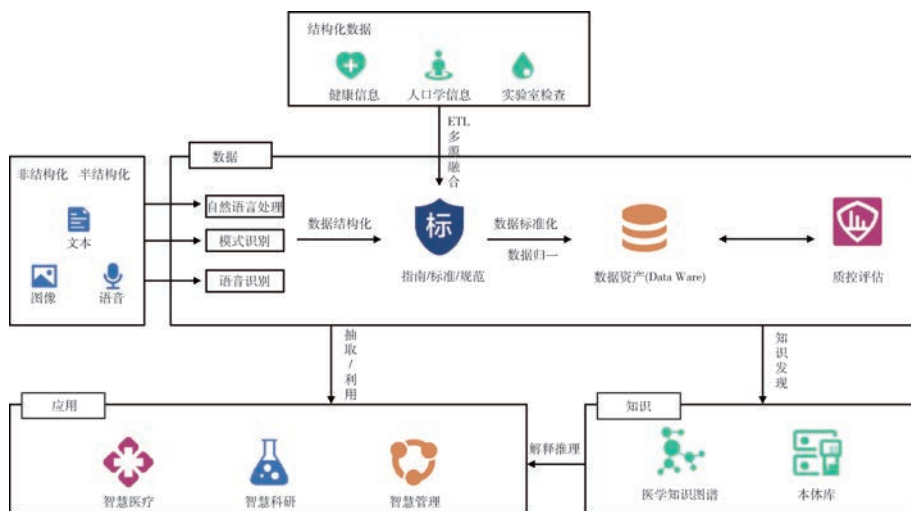


图3 医疗数据处理与知识转化模型

4.2.4 医疗数据安全 医疗大数据应用发展的基础是安全,安全管理贯穿数据的整个生命周期。在管理层面,对敏感数据进行分级分类管理,围绕个人隐私安全和重要数据安全完善大数据安全管理体系;在技术层面,一是依照纵深防御思想,完成从通讯网络到应用环境的整体防护,通过虚拟专网、强制访问控制等技术保障主机安全,建立数据完整性主动验证机制确保数据存储安全等;二是基于可信根设备、可信计算环境、可信边界、可信网络通讯构建可信安全管理中心支持下的积极主动防御框架;三是学习和推动与隐私保护、数据安全和数据流动相关的新技术应用,如区块链、多方安全计算、同态加密、联邦学习等先进方法^[11],实现数据

共享互通。

4.3 肿瘤大数据平台建设实践

4.3.1 项目概况 国家癌症中心(中国医学科学院肿瘤医院)2017年启动国家肿瘤数据中心及肿瘤大数据平台项目,在全国范围内开展肿瘤大数据获取、存储、处理与分析关键技术研究,目前已合作签约97家医院。平台采用安全套接层(Secure Sockets Layer, SSL)技术、隔离区(Demilitarized Zone, DMZ)策略构建安全传输网络,在保证数据安全传输的同时实现多种类型数据的存储、传输和共享,见图4。

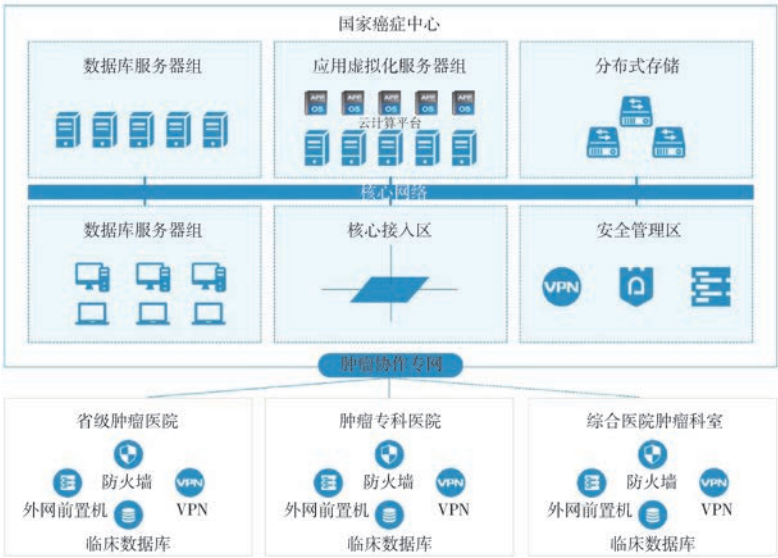


图 4 平台数据采集网络架构

4.3.2 数据治理 通过对比国内外权威数据库和病历标准并结合实际数据情况，平台对数据结构进行整理及扩充，形成覆盖肿瘤治疗流程的患者通用数据模型（Common Data Model，CDM）。在数据治理过程中以 CDM 为基础建立了一套从数据入库、关联、清洗到脱敏、标准化、结构化、质控的标准化处理流程。在数据解析环节针对电子病历等非结构化数据采用模板归纳技术实现对不同厂商、不同

文本格式的快速解析；在数据清洗环节利用 BERT + LSTM + CRF 技术实现命名实体识别和关系抽取，根据患者多维度信息重构患者唯一标识，完成患者不同医院多次诊疗信息的关联与合并；在数据质控环节对照国际、国内以及行业相关标准完成医学术语体系构建，通过深度学习、机器学习技术实现电子病历、药品、检查检验、病理诊断等数据的自动化映射和数据归一，见图 5。

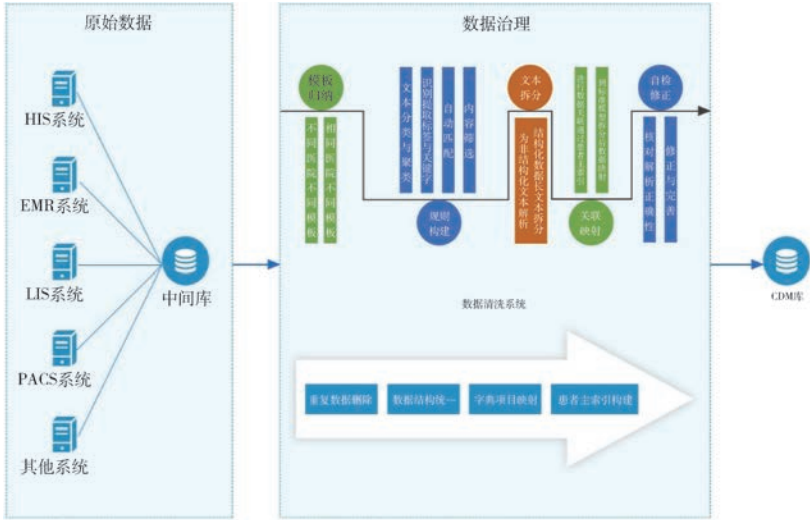


图 5 平台数据治理

4.3.3 实际应用 平台下一步将进入实际应用阶段，随着协作网络不断扩大、数据不断增多，肿瘤

标准化术语体系构建会日趋完善，数据治理工具会更加成熟，平台将为肿瘤领域的规范化诊疗和行业

创新提供强有力的数据支撑。

5 结语

近年来政府积极为大数据平台发展提供宏观动力、政策及法规环境,以大数据、人工智能、云计算为代表的先进数据与计算技术为大数据平台建设提供了技术保障,同时医疗卫生行业信息标准逐步完善、各医疗机构互联互通水平逐年提升,医疗大数据平台建设机遇大于挑战、已知大于未知。医疗大数据平台建设既要抓住政策和行业机遇又需结合自身特点丰富技术体系,做到长远谋划、与时俱进,为医疗卫生行业数字化、智能化转型提供支撑。

参考文献

- 1 陈军君,端木凌,吴红星. 中国大数据应用发展报告(2019) [M]. 北京:社会科学文献出版社,2019.
- 2 刘汪根,孙元浩. 大数据 3.0——后 Hadoop 时代大数据的核心技术 [J]. 数据与计算发展前沿,2019,1 (1): 94-104.
- 3 Sara Goodwin, John D. McPherson, W. Richard McCombie. Coming of Age: ten years of next-generation sequencing technologies [J]. Nature Reviews Genetics, 2016

(17): 333-351.

- 4 国家卫生健康委. 2018 年我国卫生健康事业发展统计公报 [EB/OL]. [2020-04-29]. http://www.gov.cn/guoqing/2020-04/29/content_5507528.htm.
- 5 孙可欣,郑荣寿,张思维,等. 2015 年中国分地区恶性肿瘤发病和死亡分析 [J]. 中国肿瘤,2019,28 (1): 1-11.
- 6 金小桃. 健康医疗大数据 [M]. 北京:人民卫生出版社,2018.
- 7 傅昊阳,徐飞龙,范美玉. 论医院健康医疗大数据治理及体系构建 [J]. 中国中医药图书情报杂志,2019,43 (3): 1-5.
- 8 阮彤,邱加辉,张知行,等. 医疗数据治理——构建高质量医疗大数据智能分析数据基础 [J]. 大数据,2019 (1): 12-24.
- 9 郭强,王乐子,母健康,等. 医疗数据信息安全政策研究 [J]. 医学信息学杂志,2020,41 (1): 20-25.
- 10 王倩. 肿瘤防控亟须大数据平台建设 [EB/OL]. [2018-10-24]. <http://www.chinahightech.com/html/chany/swyy/2018/1024/495752.html>.
- 11 梅宏. 大数据发展现状与未来趋势 [EB/OL]. [2019-10-31]. http://www.cbdio.com/BigData/2019-10/31/content_6152541.htm.

2021 年《医学信息学杂志》征订启事

《医学信息学杂志》是国内医学信息领域创刊最早的医学信息学方面的国家级期刊。主管:国家卫生和计划生育委员会;主办:中国医学科学院;承办:中国医学科学院医学信息研究所。中国科技核心期刊(中国科技论文统计源期刊),RCCSE 中国核心学术期刊(武汉大学中国科学评价研究中心,Research Center for Chinese Science Evaluation),美国《化学文摘》、《乌利希期刊指南》及 WHO 西太区医学索引(WPRIM)收录,并收录于国内 3 大数据库。主要栏目:专论,医学信息技术,医学信息研究,医学信息组织与利用,医学信息教育,动态等。读者对象:医学信息领域专家学者、管理者、实践者,高等院校相关专业的师生及广大医教研人员。

2021 年《医学信息学杂志》国内外公开发行,每册定价:15 元(月刊),全年 180 元。邮发代号:2-664,全国各地邮局均可订阅。也可到编辑部订购:北京市朝阳区雅宝路 3 号(100020)医科院信息所《医学信息学杂志》编辑部;电话:010-52328672,52328686,52328687,52328670。

《医学信息学杂志》编辑部