省级肿瘤大数据中心的规划与建设

闫慈[1]，王鹏[1]，杨越[2]，任劲[1]，吴睿豪[1]，张茜[1]，管音[2]，孙刚[1*]

[1]新疆医科大学附属肿瘤医院，新疆，830000；
[2]神州数码医疗科技股份有限公司，北京 100000
[*]为通讯作者

基金项目：自治区创新环境（人才、基地）建设专项-科技创新基地建设（资源共享平台建设）（PT1904）

摘要
2015 年全国恶性肿瘤新发病约 392.9 万人，死亡约 233.8 万人，恶性肿瘤已成为中国乃至全球发病率和死亡率最高的疾病之一，严重威胁人类的健康。在肿瘤诊治过程中，产生海量的医疗大数据，尚无统一的肿瘤大数据中心将各医疗机构的信息和资源进行高效整合、转化和共享。本研究将以新疆医科大学附属肿瘤医院肿瘤大数据中心的构建为案例，探讨医疗大数据中心的平台架构和数据资源架构，以及数据标准化采集、分析处理、隐私安全和共享等重点问题，为我国临床数据治理和转化提供一些通用思路，同时推动肿瘤科研水平的提升，进而提高省市级恶性肿瘤的诊疗水平与防控能力。

关键词：恶性肿瘤；大数据中心；数据采集；数据处理；数据安全

Planning and construction of provincial cancer big data center

Abstract
In 2015, the incidence of malignant tumors was about 3.93 million in China, of which about 2.34 million people died. Malignant tumors have become one of the highest morbidity and mortality diseases in China and even the world, which seriously threaten human health. In the process of tumor diagnosis and treatment, massive medical big data are generated. There is no unified cancer big data center to integrate and share the information and resources of medical institutions efficiently and make full use of them. This study will take the construction of Xinjiang Cancer Big Data Center as a case to discuss the resource architecture, as well as key issues regarding data collection, analysis and processing, privacy security and sharing, to provide an general insight on clinical data governance and application. So as to improve the diagnosis and treatment level and prevention and control ability of provincial and municipal level malignant tumors.

Keywords
Cancer; big data center; data collection; data processing; data security

1.引言

2019 年 1 月，国家癌症中心发布了最新一期的全国癌症统计数据。报告显示，2015 年恶性肿瘤发病约 392.9 万人，死亡约 233.8 万人。平均每天超过 1 万人被确诊为癌症，每分钟有 7.5 个人被确诊为癌症[1]。与历史数据相比，癌症负担呈持续上升态势。近 10 多年来，恶性肿瘤发病率每年保持约 3.9%的增幅，死亡率每年保持 2.5%的增幅。

肿瘤大数据治理和应用是推进肿瘤防控的一项基础事业，中国的人口基础每年产生巨量的临床数据，数据已成为医院的重要资产，电子病历、医学影像、基因组学等海量数据的有效利用将是开展临床科研和发展医学人工智能的必备基础方法手段[2, 3]。但恶性肿瘤临床诊疗信息采集等多方面缺少规范化操作规程，尚无统一的肿瘤数据采集平台，各医疗单位之间的信息数据难以交换、共享和整合，导致无法进行大规模、有代表性的肿瘤诊断、治疗及预后相关的信息收集和分析，临床研究资源浪费情况极为严重。医院建立统一的大数据平台将有利于实现数据资产的价值转化、诊疗服务模式革新，以及创新成果的孵化。

不同的病种特点导致数据的治理流程和需求各不相同，因此数据平台的建设架构也存在差异。对于罕见病，因为发病率较低，病例数较少，适合采用病例注册登记和直报系统相结合的数据中心形式[4, 5]；在各个区域建立省级数据中心，采用相同一套登记表单和数据接口，再由各省级中心周期上传至国家中心，这样可以避免病例漏报、少报和数据质量把控不严等问题。国家罕见病中心综合全国数据来制定罕见病目录和诊疗指南。对于以提升院内数据科研利用率为需求的医院，通常建设专科疾病数据库，针对一个病种指定标准化字段目录，涵盖门诊信息、人口学信息、住院信息、病程、用药、检查等多种字段，这种仅限院内使用的大数据平台建设方案，由于接口通用，行政调动能力强，构建方式简单，已经被广泛的临床医院或科室所采纳。另外一种是针对突发性的传染性疾病，可以由政府卫生监管部门主导，建设区域性的传染病大数据防控平台，例如洪湖市 COVID-19 大数据防控平台[6]，通过整合个人每日上报数据、检测机构数据和临床医院数据，构建人群健康画像，实时监控疫情传播动态，协助疫情资源调配和防控政策制定。但对于肿瘤来说，发病率较高，数据量大，院内诊疗流程长，数据完善，同时包含不同的癌肿，对于一个医院来说，尚不足以承载这么大量数据的治理和转化，因此构建区域性或省级以上数据中心更符合我国现实国情。

区域肿瘤大数据中心建设是提升我国肿瘤临床诊疗水平的基础步骤。通过设计合理的平台架构和通用的数据模型，实现多种类型数据的存储、传输和共享；收集肺癌、食管癌、肝癌、胃癌、结直肠癌、乳腺癌、宫颈癌和鼻咽癌等常见肿瘤的病理学、细胞学、检验学和影像医学等多种类型电子病历数据，实现医疗大数据标准化采集和分析处理[7, 8]；借助病理和影像数据的人工智能分析算法，可以极大提升区域肿瘤的诊断水平和准确度，降低医疗成本[9, 10]。本研究以新疆医科大学附属肿瘤医院肿瘤大数据平台建设为案例，探索临床研究大数据平台的建设架构，以及面临的技术挑战和应对策略。此项工作对于恶性肿瘤防治事业具有提纲挈领的作用，凸显了解决区域内恶性肿瘤防控工作中瓶颈的决心，在全国肿瘤防控工作中起到应有的作用。

2.体系规划

2.1.总体架构

肿瘤大数据中心在共享机制建立的情况下，通过肿瘤大数据采集与集成系统与各级医院信息平台（集成平台）或业务系统进行业务层与数据层对接，并建立全系统协同与共享机制，总体框架见图 1。

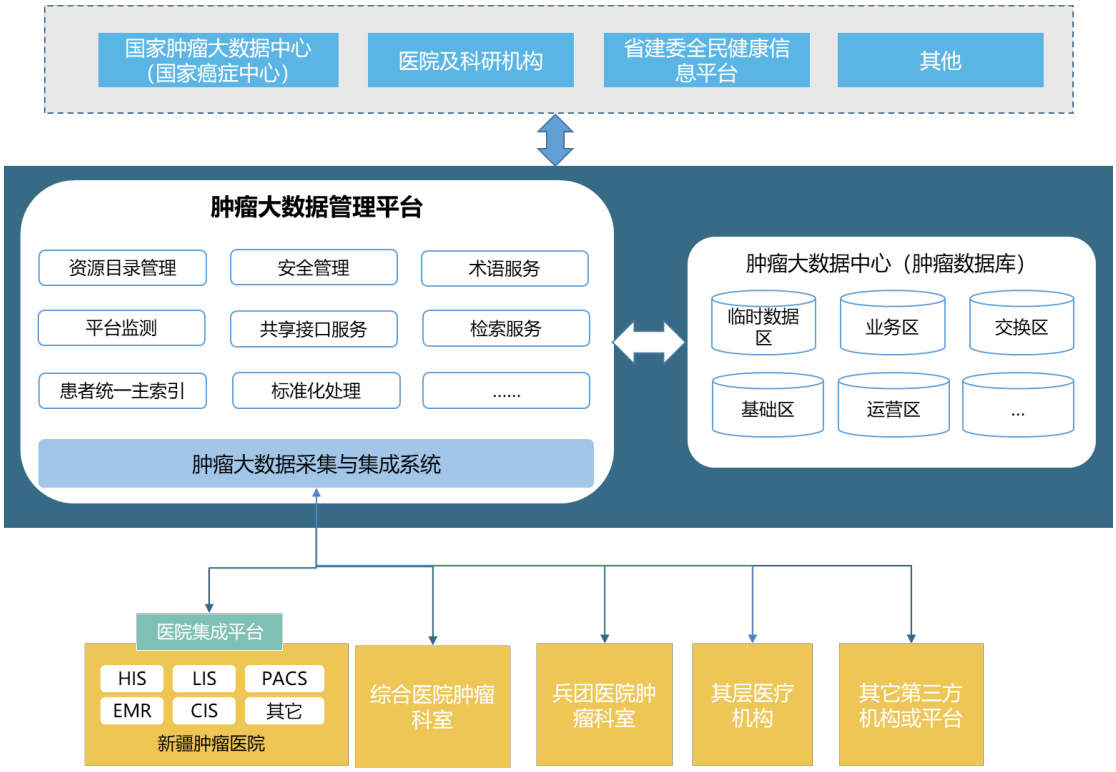

图 1 总体框架

通过各级医院集成平台收集 HIS、CIS、PACS 等不同院内系统数据并进行整合，利用肿瘤大数据采集与集成平台对接各级医院的信息平台或临床数据中心 CDR 实现数据集成，通过与 CDM（Common Data Model）数据模型相映射形成大数据中心。基于肿瘤病例特征、电子病历和 HL7 等医疗标准，建立包括肿瘤业务区（涵盖肺癌、食管癌、肝癌、胃癌、结直肠癌、乳腺癌、宫颈癌和鼻咽癌等）、共享交换区、基础区、科研区为核心的肿瘤大数据平台中心数据库。基于肿瘤大数据管理平台，实现与国家癌症中心、医院及科研机构、其他第三方机构等的共享协作并提供相应的服务。肿瘤大数据中心物理环境采用虚拟化+超融合技术实现，为各机构数据互联互通，开展远程协作和数据采集等工作提供基础支撑环境。

2.2.数据资源架构

肿瘤大数据中心区域根据业务类型的不同进行划分，具体分为：临时数据存储区、交换缓冲数据区、核心数据区及运营数据区（图 2）。

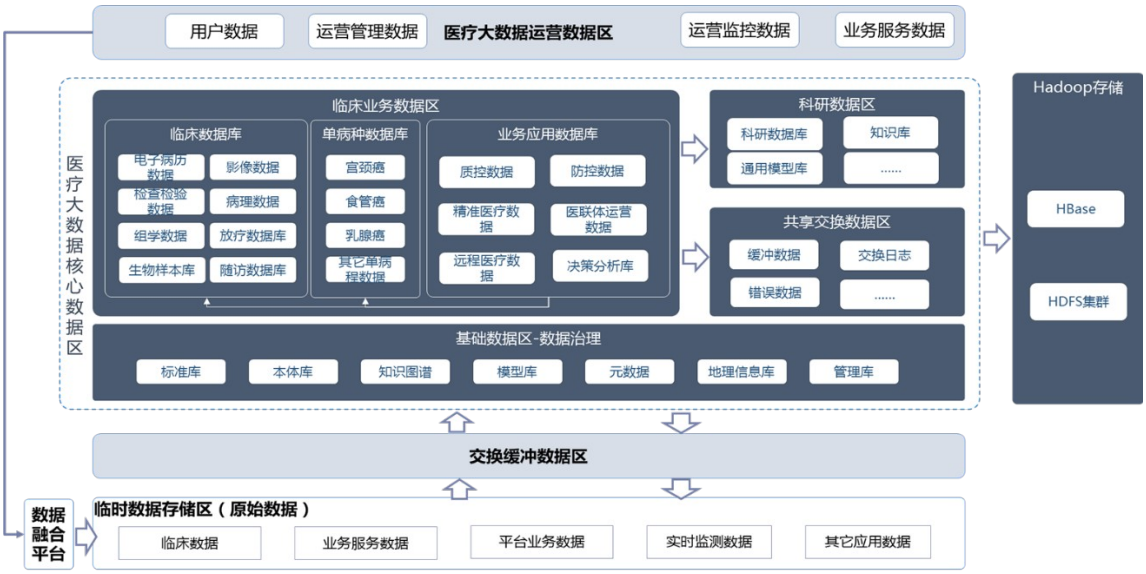

图 2 数据资源架构

临时数据存储区：存放接入的各机构原始数据，具体包含：临床数据、业务服务数据、实时监测数据等。

交换缓冲数据区：机构提供的原始数据进入平台核心数据区之前的数据交换缓冲区，防治出现数据处理错误或丢失时便于数据的朔源追踪。

核心数据区：划分为临床业务区、科研数据区、共享交换区及基础数据区四部分。

其中基础数据区是数据中心的基础，主要存放标准规范、基础数据字典、患者索引信息、平台日志用户管理信息等用来支撑平台的元数据区。

临床业务区，为临床诊疗业务应用（智能检索、诊疗辅助决策等）提供支撑，主要包含患者就诊活动中产生的医疗服务信息即临床业务数据库，按肿瘤种类进行划分的单病种数据库及业务开展的应用数据库。

科研数据区，基于临床业务区建立科研通用数据模型，过滤出适合进行科研研究的人群信息，建设知识库服务与科研分析研究。

共享交换数据区，基于临床业务区分离出需要机构之间共享交换的信息单独进行隔离，便于提高共享协作的效率，同时可按不同主题进行划分及记录共享交换的日志。

运营数据区：存放用户数据、运营管理数据、运营监控数据及业务服务数据，支撑大数据中心的运营监控。

3.肿瘤大数据管理

肿瘤大数据中心的建设是开展临床与科研应用的基础，影响应用效果的核心是数据的管理与处理，因此良好的数据管理与治理的实现是一切的重中之重，肿瘤大数据管理平台技术架构如图 3 所示。

图 3 数据管理技术架构

基于上述数据管理技术架构，数据处理流程如下图 4 所示：

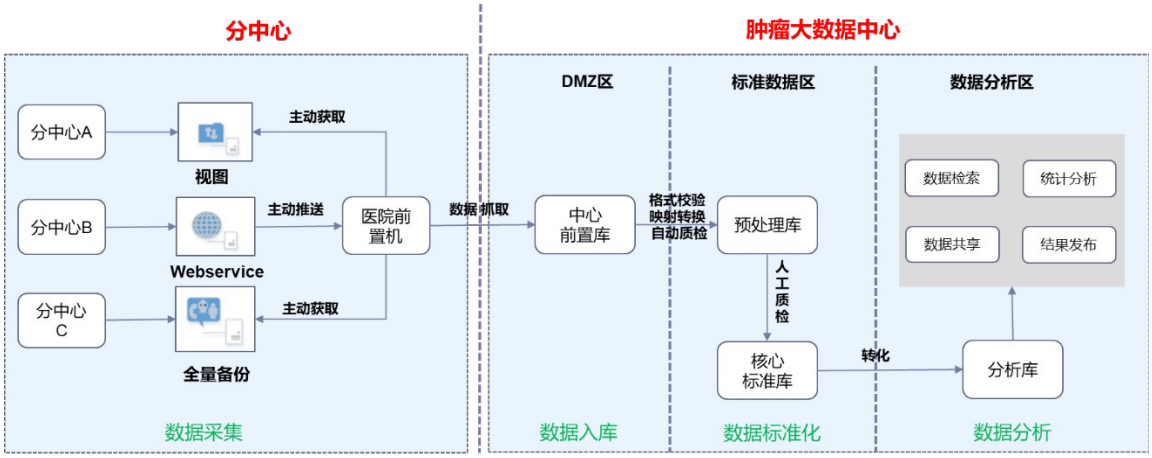

图 4 数据处理流程

3.1.异构复杂数据采集与管理

数据采集

基于各医疗机构所处位置、使用数据库类型的不同、数据模态的多样性，针对这种复杂异构的情况，我们充分利用政务云资源，在云端架设大数据集群环境，在云端采取备份数据库、Webservice 服务或视图等多种方式主动获取各机构的增量数据，不抛弃任何业务数据，从而保证数据的完整性、准确性及与原生产系统的一致性，不会导致数据失真。另外，数据获取时根据业务类型的不同实行实时和定时两种方式进行。同时支持将分析成果返还区内医疗机构进行应用展现。

通过中间表视图，医院临床业务系统中的数据以符合通用数据模型需求的格式呈现出来。中间表是指连接通用数据模型和医院临床业务系统数据内容的数据库表。数据库中的视图，是一个虚拟表，其内容由查询定义，可以提供和真实表相同的数据内容和字段。中间表视图能够在几乎不增加数据库负担的情况下，实时、准确的将临床、科研、管理所需的临床数据从不同的临床业务系统中，以满足 OMOP 通用数据模型需求的方式查询显示出来。

Kettle 是数据抽取过程中常用的 ETL 工具。Kettle 通过可视化的方式，提供了便捷高效的数据提取方式。通过在 Kettle 中配置输入输出数据库的接口和数据表并设置互相映射的字段，中间表视图中的临床数据被完整的抽取到通用数据模型的数据表中。

数据校验

在数据采集的过程中，势必会由于人工或者医院业务系统本身的原因，而导致抽取至数据中心的数据并非完全是正确的、可用的、一致的。为了解决该问题，提高数据中心存储信息的准确性，本方案特设置了数据质量校验平台。该平台通过设置一系列数据检验规则，对采集的数据进行校验，从而提高数据中心的数据质量。

通过数据质量校验平台，可以进行原始数据关联性检查、原始数据的分布检查、清洗后数据与原始数据一致性检查，并且在此过程中引入了美国最权威的临床数据质控标准，以及 14 大类共共计 1200+的数据检测规则。我们对数据完整性、一致性、准确性、及时性四个方面进行校验，定期迭代优化数据采集流程。

3.2.数据标准化处理

基于电子病历和信息集成平台系统收集了包含病理学、细胞学、检验学和影像医学等多种类型电子病历数据，数据的标准化处理是进行数据分析与应用研究的关键，数据标准化处理即原始数据进入数据缓冲区后经历结构化处理、通用数据模型映射、术语绑定、数据质控的一系列过程[11]。

对于非结构化文本数据，我们采用先进的自然语言处理技术（NLP，Natural language processing）和语义分析技术，通过文本预处理、人工标注、机器学习、模型构建及模型应用 5 个步骤进行处理，实现非结构化数据的结构化。

构建多源异构临床样本与生命组学的通用数据模型 CDM(Common Data Model), 完成从区内不同医院/队列的数据标准向恶性肿瘤临床与科研大数据采集与共享平台的数据标准转化与映射。抽取来自区内不同医院/队列数据，构建原始数据存储库。研究基于质量控制数据抽取与筛选和基于角色控制的数据安全与访问方法。

数据模型选择上，我们构建多源异构临床样本与生命组学的国际通用数据模型 OHDSI OMOP CDM(Common Data Model)，完成从区内不同医院/队列的数据标准向肿瘤大数据中心平台的数据标准转化与映射。OHDSI 是一个开放的全球医疗科研协作网络，聚焦于医学数据标准化、医疗产品安全监控、比较有效性研究、个性风险预测、数据特征化、质量改进。目前数据网络涵括来自 19 个国家和地区的 12 亿条病人记录，超过 100 个数据库。该模型以人为中心构建生命全息视图，相关数据表的定义及表结构字段的规范均以国际标准为前提，把健康医疗数据转化成研究用的数据模型，便于快速而有效的医疗大数据分析。支持多中心、跨区域、跨国家的多中心科研，允许对不同的观测数据库进行系统分析，实现数据互联互通。

对于术语绑定，采用世界上最复杂、最丰富的一体化中文医学术语本体技术 SNOMED CT，该术语库涵盖 40,000+疾病信息、20,000+药品信息及 500+检查检验项目信息，与 ICD10 体系相比，疾病或者项目的描述与分类更加细化，更适合于后期基于数据中心进行临床科研研究[12, 13]。

数据质控采用自动与人工兼并的方式进行，主要聚焦于数据完整性、规范性、一致性、准确性、唯一性、及时性方面，通过质控规则的设置、数据稽查、质量报告、问题分配管理实现质控过程的闭环管理。

3.3.数据存储与传输安全

基于庞大的数据资源，采用 Hadoop 分布式+关系型数据库相结合方式存储，支持对 PB 级数据量的快速处理、大规模数据的秒级检索，采用 SSL（Secure Sockets Layer，安全套接层）加密，分离密钥和加密数据，使用过滤器和数据备份等方式，构建数据的存储安全策略。采用 DMZ（Demilitarized Done）隔离区策略，保证数据传输的安全。

为了对医院数据进行安全保护，需要进行数据隐私处理。首先需要对原始数据进行整理，

产生中间表视图。根据这些临床数据大致可以分为两种类型，结构化的数据和非结构化的数据，结构化的数据首先会对个人信息进行脱敏处理，并且对原始数据进行二次编码，防止其逆向回溯源数据。使用成熟的医学数据脱敏算法识别信息中有关病人隐私的重要信息，如姓名，身份证号，生日等信息，进行脱敏处理。非结构化的数据首先根据这些文本类型选择自然语言处理，再根据中文本体库的建设和通用数据模型做数据的合并和映射，最后将这些数据脱敏后导入到数据仓库中去。最后，对脱敏后的数据使用预定的数据标准化规范进行标准化。脱敏后的数据可最大限度的方便医学研究，避免用户隐私信息的泄漏。

同时构建人员管理层面的数据安全管控制度，严格控制数据共享和传输操作。涉及多中心研究的情况，需由需求发起方提出数据使用书面申请并签字留存，该书面申请涵盖数据大小、数据范围、使用目的、使用场景、使用时间等内容，院方管理人员审批通过后，经过技术人员对数据进行脱敏，然后由技术方评估数据需求的安全性，达到要求后审批通过才可用于使用，并对申请表和审批表进行留存备份，该数据达到使用时限后需清除院外拷贝防止超需使用或泄漏风险。该流程从制度层面强化了数据的安全和共享机制，责任分担到人。

3.4 数据共享

采用基于中心交换系统(企业服务总线为核心)与可定制的前置交换软件系统来实现区内各医疗机构间临床及组学数据的交换与共享。基于协作网络共享平台，定义临床表型数据共享标准规范，为肿瘤医学数据的共享提供完善的标准，通过统一的接口管理方式对接口标准进行封装，对接口申请单位采用统一的标准进行流程审核、接口调用进行系统监控。最终实现数据共享、统一数据标准，促进各队列的业务协同及研究成果的共享。

4.总结与展望

大数据时代，使得利用这些海量的医疗数据造福肿瘤患者、攻破肿瘤难题成为可能，但是合理利用医疗大数据产生的结果，仍是我们面临的一道难题。肿瘤临床数据包括电子病历、医学影像、临床检验等多种类型，这些数据多处于归档状态，又分散存储于不同的业务系统中，而这些数据多为半结构化和非结构化数据，具有多源异构、多模高维的特点，依靠传统的数据分析、处理技术无法满足实际需求，数据归档和处理存在难度。研究和分析肿瘤疾病的发病模式和影响因素，为早期筛查、早期诊断和药物研发提供重要依据，为医护人员提供临床指导，为患者提供最佳诊疗方案，为肿瘤精准医疗的实施奠定基础。

本研究以新疆医科大学附属肿瘤医院肿瘤大数据科研平台建设为案例，探讨了大数据架构和作业流程以及在数据采集、标准化、共享和安全保护等方面的挑战和应对措施，让大数据在肿瘤防治、临床诊断、科学研究、指南制定等方面落地赋能。医疗大数据平台的构建必然面临诸多挑战，需以传统临床经验为基础，以高新技术为依托，制定合理长远的计划，让大数据在医疗等领域的应用成为现实。

致谢：本研究得到自治区创新环境（人才、基地）建设专项-科技创新基地建设（资源共享平台建设）基金（编号：PT1904）资助。

参考文献

[1] 郑荣寿，孙可欣，张思维，等. 2015年中国恶性肿瘤流行情况分析[J]. 中华肿瘤杂志，2019,41(1):19-28.DOI:10.3760/cma.j.issn.0253-3766.2019.01.008.

[2] 周连茹，程颖，王坤. 医疗卫生信息数据中心的规划与建设[J]. 信息系统工程，2009(08):128-129.

[3] 沈炜, 李婕. 医院数据中心的建设与管理方法初探[J]. 医学信息, 2005(03):185-187.

[4] Feng S, Liu S, Zhu C, et al. National Rare Diseases Registry System of China and Related Cohort Studies: Vision and Roadmap.[J]. Human gene therapy, 2018,29(2).

[5] 冯时, 弓孟春, 张抒扬. 中国国家罕见病注册系统及其队列研究:愿景与实施路线[J]. 中华内分泌代谢杂志, 2016,32(12):977-982.

[6] Gong M, Liu L, Sun X, et al. Cloud-Based System for Effective Surveillance and Control of COVID-19: Useful Experiences From Hubei, China[J]. J Med Internet Res, 2020,22(4):e18948.DOI:10.2196/18948.

[7] 高东平, 李伟, 秦奕, 等. 肿瘤大数据中心信息系统建设初探[J]. 中国数字医学, 2018,13(03):19-22.

[8] 宋波, 朱甜甜, 于旭, 等. 医疗大数据在肿瘤疾病中的应用研究[J]. 中国数字医学, 2017,12(08):35-37.

[9] Hamilton P W, Bankhead P, Wang Y, et al. Digital pathology and image analysis in tissue biomarker research[J]. METHODS, 2014,70(1):59-73.DOI:10.1016/j.ymeth.2014.06.015.

[10] Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities[J]. MEDICAL IMAGE ANALYSIS, 2016,33(SI):170-175.DOI:10.1016/j.media.2016.06.037.

[11] Wang X, Williams C, Liu Z H, et al. Big data management challenges in health research—a literature review[J]. Briefings in Bioinformatics, 2019,20(1):156-167.DOI:10.1093/bib/bbx086.

[12] 王奕, 李芳. 基于SNOMED的嵌入式电子病历模板的设计方法[J]. 计算机应用与软件, 2008(02):223-224.

[13] 谢雪娇, 张黎黎, 奈存剑, 等. 国外医学术语标准开发方法及对我国的启示[J]. 中华医学图书情报杂志, 2019,28(11):16-21.