DOI 10.15302/J-SSCAE-2022.06.011

肿瘤临床大数据管理系统设计与应用

马麟[1]，包晨露[2]，李青[2]，吴静依[2]，潘虹安[2]，李鹏飞[2,3*]，张路霞[2,3]，詹启敏[3]

（1. 北京大学医学部学科建设办公室，北京 100191；2. 浙江省北大信息技术高等研究院，杭州 311215；
3. 北京大学健康医疗大数据国家研究院，北京 100191）

摘要：肿瘤是人类生命健康的重要威胁，随着我国医疗行业信息化的发展，医疗机构积累了大量的肿瘤临床数据，但因数据标准不统一、治理难度大等原因制约了数据价值的充分挖掘；应用人工智能（AI）等前沿信息技术建设肿瘤临床大数据管理系统，有助于肿瘤临床数据的深入应用、临床诊疗管理质量与效率提升。本文剖析了我国肿瘤临床数据治理与应用面临的问题及挑战，研判了肿瘤临床大数据管理体系的应用价值；针对肿瘤临床数据多来源、多模态的复杂特性，探索了AI技术应用于肿瘤临床大数据管理与科研的机制及路径；设计了包括肿瘤通用数据模型构建、临床数据采集与安全管理、标准化结构化治理、分析与建模应用、数据质量管理在内的全流程解决方案，阐述了相应系统的建设框架与技术体系；以某三甲医院肺癌临床大数据平台为案例，展示了所提方案在临床实践中的可行性及应用价值。相关研究可为丰富我国肿瘤临床大数据管理系统的建设实践、探讨领域未来重点研究方向提供参考和启示。

关键词：临床大数据；管理系统；肿瘤；人工智能；通用数据模型；自然语言处理

中图分类号：R730.1　　**文献标识码：**A

Design and Application of Clinical Big Data Management System for Oncology

Ma Lin[1], Bao Chenlu[2], Li Qing[2], Wu Jingyi[2], Pan Hong'an[2], Li Pengfei[2,3*], Zhang Luxia[2,3], Zhan Qimin[3]

(1. Academic Development Office, Peking University Health Science Center, Beijing 100191, China; 2. Advanced Institute of Information Technology, Peking University, Hangzhou 311215, China; 3. National Institute of Health Data Science at Peking University, Beijing 100191, China)

Abstract: Cancer is a serious threat to human life and health. Along with the development of medical informatization in China, healthcare institutions have cumulated a great quantity of clinical data in oncology; however, these data have not been fully explored owing to the disunity of data standards and great difficulties in data management. Hence, establishing a national clinical big data management system for oncology based on artificial intelligence could potentially promote the application of clinical data in oncology, further improving the quality and efficiency of clinical management for oncology. This study conducted an in-depth analysis of the problems and challenges of clinical data management and application for oncology and presented the significant values of an oncology clinical data management system. Considering the complexity of multi-source and multi-modal data in oncology, we explored the possible mechanisms and pathways of applying artificial intelligence to the management and research of clinical data for oncology.

收稿日期：2022-08-21；**修回日期：**2022-09-27

通讯作者：*李鹏飞，浙江省北大信息技术高等研究院高级工程师，研究方向为健康医疗大数据智能治理、分析与应用；
E-mail: pfli@aiit.org.cn

资助项目：国家自然科学基金项目(72125009)

本刊网址：www.engineering.org.cn/ch/journal/sscae

Furthermore, a full-circle solution was designed, and the construction framework and technology systems were promoted for the clinical data management system for oncology, including the development of common data models for oncology, data collection and security management, data standardization and structuring, data analysis and application, and data quality control. Besides, we validated the feasibility and benefits of the promoted system in clinical practice by taking the clinical data management for lung cancer in a tertiary hospital as an example. Finally, we proposed some suggestions on the research directions of the clinical big data management system for oncology.

Keywords: clinical big data; management system; oncology; artificial intelligence; common data model; natural language processing

一、前言

世界卫生组织国际癌症研究机构（IARC）研究数据表明，2020年我国新发肿瘤确诊病例约为456.9万例，新发死亡病例约为300万例 [1]，造成了日益加深的社会负担 [2]。我国老龄化程度严重，可能会导致肿瘤发生和负担继续加重，如死亡率难以下降、预后情况不甚理想 [3]，因此需进一步加强肿瘤防治管理和研究工作 [4]。利用大数据、人工智能（AI）技术赋能健康医疗逐渐成为研究热点，在疾病监测、风险评估与干预方面具有良好价值 [5]，而在肿瘤精准诊疗、精准预后方向的作用尤其突出。AI技术在癌变部位识别 [6]、肿瘤风险预测 [7~9]、肿瘤诊断和分期判断 [10,11]、基因突变检测 [12,13]、分子靶点药物研发和设计 [14,15]、患者生存或预后情况预测 [16,17]等多个肿瘤诊疗方向均有应用实例。

肿瘤临床大数据应用的核心在于，运用AI和大数据技术开展数据整合与建模，以大规模、多来源、高维度的数据为基础开展精准诊疗医学的研究与应用，为肿瘤患者提供更全面、更准确的早期筛查诊断以及治疗和预后方案，提升肿瘤整体诊疗水平 [18~20]。发达国家在20世纪后期即开展了肿瘤医疗信息采集和数据系统建设，在21世纪初期开始推进医疗辅助决策、远程诊疗等方面的工作；一些国家建设了国家级医疗数据服务系统并逐步完善相应法律规章 [21~23]。例如，英国的国家癌症注册和分析服务覆盖了1971年至今162个英国医疗机构的医疗数据，每年采集约30万例肿瘤患者信息，是英国肿瘤诊疗和资源利用的坚实基础 [24]；法国建立了基本覆盖全部人口的医疗保健数据库（SNDS），为卫生信息监测提供了保障 [25]；美国的国家癌症数据库（NCDB）采集了3000多万例癌症患者数据（至2016年）[26]，电子病历系统得到全面普及 [27]。

我国医疗信息化工作起步较晚，在《中共中央国务院关于深化医药卫生体制改革的意见》（2009年）发布后发展加速，目前正在积极普及医疗数据信息化管理。将肿瘤临床信息与大数据、AI等新兴技术相结合，是《“十四五”卫生健康标准化工作规划》提出的重点方向；部分医院开始推进多个肿瘤病种的专病信息化建设工作，如肺结节与肺癌全程智能管理云平台 [28]、肝病与肝癌领域大数据平台 [29]；部分地区积极打通信息壁垒，建立了区域性的健康平台 [30]，但“数据孤岛”现象突出，数据结构、来源、管理规范和标准未能得到大范围认可，导致积累的患者数据开发利用程度偏低。应用数据科学技术，将多模态、多维度的临床大数据转化为医疗方案，改善肿瘤诊疗服务并为预后提供精准管理至关重要。

也要注意到，目前肿瘤患者诊疗数据采集困难，基于AI技术开展的肿瘤诊治相关研究样本量极为有限 [31~33]，因此相关研究的科学性、可靠性尚无法充分保障 [34]，其成果难以推广至特征复杂的大规模人群。因此，基于大规模肿瘤人群开展科学可靠、泛化性能强大的肿瘤辅助诊疗研究，建立统一的肿瘤医疗数据库是必要的基础。而在我国，大规模、标准化的肿瘤临床大数据管理和应用平台的探索研究仍显缺乏。本文以我国肿瘤临床大数据的发展现状及应用价值为出发点，分析应用存在的问题，基于管理与科研应用需求表述整体架构；以肺癌临床管理大数据平台为案例，针对性提出解决方案；展望未来发展方向，为肿瘤临床大数据科研布局提供基础参考。

二、肿瘤临床大数据管理系统的应用价值

（一）提高我国肿瘤临床数据信息化程度

我国医疗信息化发展迅速，但还未得到全面普及。我国现有的肿瘤临床数据信息化建设工作主要

由医疗资源相对较为丰富的大型三甲医院主导和推进，然而，除了这些大型三甲医院，我国还有大量医疗资源较为缺乏的二级医院、一级医院和社区基层医院。据国家卫生健康委员会统计，截至2021年年底，我国有两万多家二级医院和一级医院，以及九万多家基层医疗卫生机构 [35]，其中蕴含了大量宝贵的医疗数据。而中国医院协会信息管理专业委员会发布的《2019—2020中国医院信息化状况调查报告》指出，每年均有固定信息化建设预算的三级医院占93.46%，而三级以下医院中设有固定预算的只占了69.33% [36]。由于基层医疗机构缺乏丰富的医疗资源，对信息化建设的重视程度也较为欠缺，导致信息化程度相对滞后，大量基层肿瘤诊疗数据无法被获取和应用。因此，应建立各等级医院间的关联，自上而下带动并普及全国性肿瘤临床大数据信息化管理平台的建设，从而囊括我国各区域的基层肿瘤临床数据，进一步实现各级医疗机构信息化程度提升，最大程度整合全国临床数据，为后续基于大数据的临床辅助决策奠定坚实的基础。

（二）改善肿瘤数据共享与应用程度

数据共享是肿瘤临床大数据应用的基础，尽管我国近年来已发布大量相关政策鼓励推进临床大数据的共享与广泛应用，但我国医疗共享与应用程度仍相对不足，建立全国性的肿瘤临床数据平台能够有效解决这一问题。肿瘤临床数据的结构、来源不一是造成数据整合与共享困难的原因之一。虽然我国已有较多医疗机构成功构建了肿瘤临床数据管理体系，且有部分地区已经推进了省市级或区域级肿瘤数据平台的建设 [30]，然而这些数据平台之间互不相通。其次，由于各医疗机构信息化程度不同，院内系统接口与数据结构亦均不相同，导致临床数据仍以孤岛形式存储于各数据库中。虽然我国有上百万例肿瘤患者信息，却难以整合形成类似于美国国家癌症研究所监测、流行病学和最终结果项目（SEER）及癌症基因图谱（TCGA）的大规模、多维度的医疗数据库，因此也难以支撑我国肿瘤数据科研应用工作的展开。为了消除数据孤岛，建立肿瘤各病种临床数据标准规范，是将各医疗机构的肿瘤临床数据以统一标准整合处理的基础，也是最终能够形成大规模数据库的必要条件。

有效的数据共享往往是互惠互利的合作方式，能够推动以大规模、高维度数据为驱动的研究与科研应用的发展 [37]，但肿瘤临床数据包含大量患者隐私和健康数据，涉及数据敏感性问题导致难以共享，如何保证健康医疗数据脱敏和安全使用亦是需要制定明确规定和政策进行监管。国际医学期刊编辑委员会提倡共享去标识的医疗研究数据 [38]。我国从数据安全、个人隐私保护等方面颁布了多项法律法规，《信息安全技术 – 健康医疗数据安全指南》和《医疗机构医疗大数据平台建设指南》的发布均强调了健康医疗领域数据安全的重要性，并大力推行规范化管理。但由于这些举措仍在推行阶段，尚未能达成共识。因此，肿瘤临床大数据管理系统以各类数据安全法规为基础，进一步加强健康医疗数据的安全管理和体系建设，推动肿瘤临床大数据的共享和广泛应用。

（三）加强AI在肿瘤领域的应用

“医学+AI”的跨学科交叉合作模式已成为未来的重点发展趋势，得到了多数业内人员的关注，例如，北京大学团队近期研发了“未数”健康医疗大数据智能分析平台（visdata.bjmu.edu.cn），充分利用AI等技术为各类健康医疗大数据的价值挖掘提供可行性。然而，AI的应用目前较为限制于医学影像智能识别诊断领域，近些年最热门的研究主题为“计算机视觉算法与模型”[39]。虽然医学影像对于肿瘤是有效的辅助诊疗手段，对CT影像的识别能够更准确地对高危人群和恶性肺结节进行预测 [8,40]，但其他肿瘤研究领域也应得到同等关注，例如，免疫治疗、手术预后、癌痛管理等。综上所述，AI在肿瘤诊疗领域的应用尚有许多待发掘的应用前景以及待解决的实际问题。肿瘤临床大数据管理系统的建设为AI技术应用于实际临床诊疗提供了数据基础、降低了技术门槛，有助于推动前沿技术在肿瘤诊疗领域的广泛、深度利用，从而提高肿瘤诊疗效率、改善患者预后。

三、肿瘤临床大数据管理系统框架

常规的临床大数据管理系统主要包括数据采集与储存、数据管理、可视化数据大屏，以数据获取与展示功能为主。本文描述的肿瘤临床大数据管理系统架构不仅在数据储存的基础上进行了扩充，还融

合了各类统计和AI算法，在支持数据展示的同时，提供更多数据分析应用等。该系统将肿瘤实际临床需求与前沿算法相结合，能够为数据标准化管理和数据安全等问题提供解决方案，从而提高肿瘤数据管理效率、促进数据广泛利用与深度挖掘。

肿瘤临床大数据管理系统框架包括储存层、数据管理、基本应用、算法库四个模块（见图1）。

（一）储存层

储存层的功能是储存肿瘤患者的各类肿瘤临床数据和缓存结果。首先，采集并存储肿瘤临床大数据入院诊疗相关业务的临床数据和文档，包括个人信息、住院记录、门诊记录、各类检查检验报告等。通过将这些数据转换成分析应用数据库，支持后续进行常规的人群筛选，以及肿瘤医学领域常用的数据分析与应用。储存层还支持分析结果的阶段性缓存，有效降低系统资源利用。

（二）数据管理

肿瘤临床大数据管理的目标是将院内多系统多来源的数据信息以标准化、结构化的方式整合管理。首先，基于肿瘤临床数据管理与应用需求及相关国家、行业标准，构建结构化肿瘤临床通用数据模型，包括通用核心数据集和肿瘤各个病种特有的疾病相关信息。其次，从院内进行数据采集，即将院内提取的用于描述患者肿瘤病情以及诊疗状态的相关数据基于肿瘤临床通用数据模型（CDM）进行结构化治理，并将治理后的多源异构数据汇总整合。最终，对患者的个人敏感性数据进行脱敏以保证数据安全；将提取的各类数据分为可直接存储并应用的结构化信息以及非结构化信息，后续进行数据标准化治理，为数据展示和应用提供服务。

（三）基础应用

基本应用层是肿瘤临床大数据管理系统为满足多种临床诊疗应用场景需要，支持系统使用者进行数据查看、检索和挖掘分析的功能层。具体包括：面向不同应用场景的数据概览模块；基于多种质量维度的数据质量评估体系；支持复杂条件自定义、树状构造的数据检索模型；支持菜单式操作的数据可视化分析模块；基于容器化技术实现无代码统计建模与机器学习建模的数据挖掘分析模块。

（四）算法库

算法库是支持肿瘤临床大数据管理系统数据各个模块的基础架构。首先，利用真实世界肿瘤数据，通过机器学习和深度学习算法支持数据结构化与标准化模型构建。其次，基于临床与科研应用场景，将基础统计分析算法与前沿AI算法打包为功能模块，提供涵盖基础描述性分析、生存分析等生物医学领域统计分析算法，以及机器学习等训练模型和算法，实现基本服务的各类应用场景。

四、肿瘤临床大数据管理与应用及其功能模块

肿瘤临床大数据管理系统的主要功能模块包括数据管理和临床与科研应用两方面（见图2）。

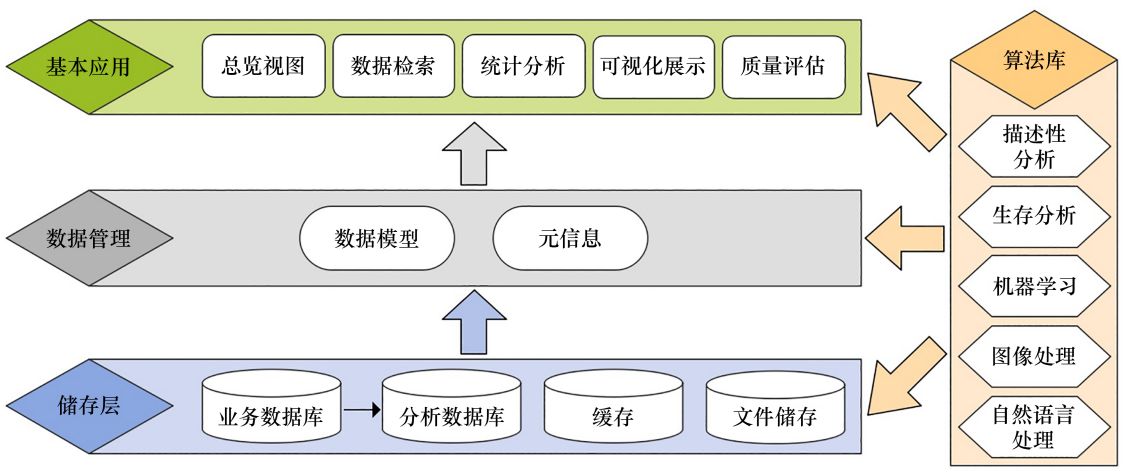

图1　肿瘤临床大数据管理系统框架

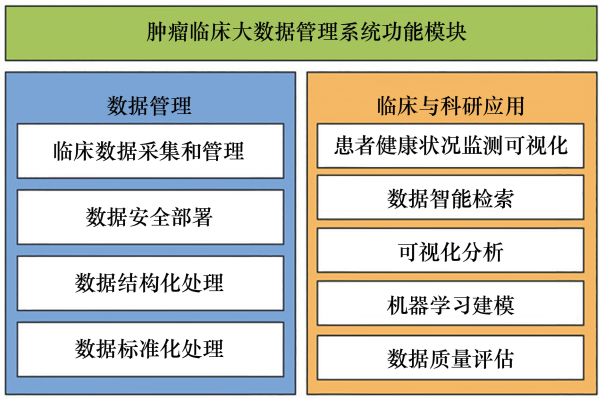

图2 肿瘤临床大数据管理系统功能模块

（一）数据管理

1. 肿瘤临床数据采集和管理

肿瘤临床数据通常储存于院内多个数据系统，如实验室信息管理系统（LIS），影像归档和通信系统（PACS），电子病历系统（EMR）等，而数据采集就是通过智能自动化数据提取技术，根据指定标识码跨系统整合相关信息。智能自动化数据提取技术能够在多个数据表中找到关联身份标识号（ID），建立跨表树形架构串联同一患者的所有数据，实现定期数据增量更新。具体过程如下：根据第十版国际疾病与相关健康问题统计分类（ICD-10）精准定位各类肿瘤患者，根据患者主表定位每例患者的唯一标识编号，找到患者每次入院对应的入院记录ID。由于患者每次入院均会产生入院记录、出院记录、病案首页等信息，若患者进行手术，还会产生手术记录以及对应的病理结果，数据提取时将每条记录编号与每次入院的记录进行匹配，能够识别同次入院的相关检查与治疗方案。与此同时，由于各种多个来源的记录中会产生重复数据，但各来源数据质量不一，缺失程度不一，应以其中缺失率最低的为基准，通过其他来源数据进行填补以保证数据完整。

2. 数据安全部署

常见的肿瘤临床数据平台有存储于医院内环境和互联网平台两种模式。若数据直接存储部署于医院院内环境，即直接于该医院院内数据库提取数据并存入院内肿瘤数据管理系统，通过数据不出院的方式保证数据安全管理。若数据存储于互联网平台，能够实现多个医疗机构肿瘤临床数据整合汇总的目的，但首先应着重考虑健康医疗数据的敏感性和隐私性。为保证肿瘤临床数据管理系统的数据安全，所有能够对患者进行标识的敏感信息，如患者的姓名、证件号码、联系方式、地址等信息，均应参照美国《健康保险流通与责任法案》（HIPPA）进行脱敏和加密处理。同时还需保证系统网络安全，降低传输过程中的安全风险。

3. 数据结构化处理

数据结构化处理是对于不同类型的非结构化数据，根据专家建议与实际临床含义，有针对性地采用不同的前沿算法进行处理的数据治理过程。病理报告、出院记录等文本数据通常包含大量临床信息，利用自然语言处理技术（NLP）将各类结构化诊疗数据从中提取。首先，基于临床需求确定被提取的结构化字段以及提取规则，采用正则表达语义分析，或者对大量文本进行实体标注或指针标注处理，再通过多种模型进行训练。经过多次核验修正，不断优化模型以提高文本识别准确率。例如，根据正电子发射断层–X线计算机断层组合系统（PET-CT）描述与结论的长文本信息，将两段文本中对同一病灶的描述进行匹配，提取病灶的数量、位置、PET倾向性等结构化的详细数据。而对于实验室检验结果、检查报告等图像类型数据，常规应用基于深度学习的图像识别技术（OCR）对图像中的结果信息进行识别，例如对肿瘤标志物的检查项目以及对应的结果进行提取。对手术视频等视频类数据，首先将其切割为图片，而后利用图像识别方法进行处理。上述对于文本、图像、视频类型非结构化数据的提取能够将无法用于分析的非结构化信息转化为可分析的结构化数据，为临床分析的可行性筑牢数据基础。

4. 数据标准化处理

各医疗机构数据结构、内容不一等问题为数据整合与共享造成了严重挑战，因此，需要制定标准数据规范，以规范化、标准化治理多源异构数据。标准化处理模块的核心是以国家出台的相关标准为基础，结合临床专家的建议和肿瘤临床数据的实际含义与内容，制定面向肿瘤临床研究与应用的通用数据模型，以确定各类数据的值域范围、参数范围和字段校验规则。

由于各医院数据结构不同，数据字段名称存在差异，甚至同一医院院内的相同检查报告由于检查时间的先后，仪器或者检查部门也均不相同，而且

相同的检查指标可能从不同的报告中提取，造成数据冗余。因此，数据标准化还应根据数据字段名称以及别名来建立各数据集与通用数据模型的映射规则，囊括各数据字段的不同结构与不同字段描述，例如经过数据提取、转换和加载（ETL）技术对识别相同数据字段的不同单位加以识别，根据不同单位，将结果自动化转换为标准单位。

通用数据模型（CDM）能够为多源异构数据制定标准化结构与转换映射方案以定义各类数据，达到统一处理的目的 [41,42]。为整合不同来源的数据，现已有多项研究建立了肿瘤领域CDM框架，如美国观察性医疗结果合作组织（OMOP）建立的CDM在支持数据标准化的同时，还能够降低数据的信息损失程度，是降低跨数据库研究工作技术门槛的有效工具 [42]。以此类研究为基石，肿瘤临床大数据管理系统以CDM为基准，将各医疗机构的多源异构数据映射到标准数据集中，以统一通用的规则整合多中心一体化肿瘤临床数据，为构建更大规模和更高维度的临床数据库奠定了基础，也为后续进行临床科研分析或前瞻性研究提供支持。

（二）临床与科研应用

1. 患者健康状况监测可视化

医生在常规的院内临床数据库中查阅患者信息时，往往需要花费大量时间打开多个页面甚至多个系统进行查看，既费时又难以对患者病情有完整的评估。肿瘤临床大数据管理系统应以患者个体为单位，整合所有相关肿瘤临床诊疗数据在前端页面展示，通过便捷交互协助医生随时查看患者多次入院的诊疗详情和各类检查结果（见图3）。该模块通过丰富的可视化视图展示每例患者的重要诊疗信息，以甘特图的方式展示患者的确诊时间、手术时间、化疗等治疗时间，甚至死亡时间等重要时间点，能够极大程度上节省医生查阅病历档案的时间，优化数据查阅过程，能够便于医生快速捕捉到关注的重点信息，更快捷地把握和了解患者的病情变化和诊疗方案。

2. 数据质量评估

现各医疗机构肿瘤临床数据质量层次不齐、缺失率较高、部分数据存在逻辑不一致问题，导致难以保障数据分析和应用的质量和结果可靠性，因此应纳入数据质量评估体系对数据库中存在的各种数据问题进行识别。Weiskopf和Weng [43]提出了完整性、正确性、一致性、合理性、时效性五种质量维度以及七种衡量方法。且现已有多个研究团队基于通用数据模型开发了相关质量评估工具 [44,45]，能够对数据的多种维度进行核查。应以此为基础，结合实际数据情况，为肿瘤临床数据定制质量评估体系，进行全方位评估并定位到数据存在的问题，从上述五个数据质量维度进行评分。在报告数据整体质量的同时，对存在问题的数据进行整理展示，再提出相关问题的改进建议，并对其数据质量的调整进行监管。

3. 数据智能检索

临床研究过程中，通常需要对不同类型的人群进行分析，例如不同性别、不同肿瘤病理类型、不

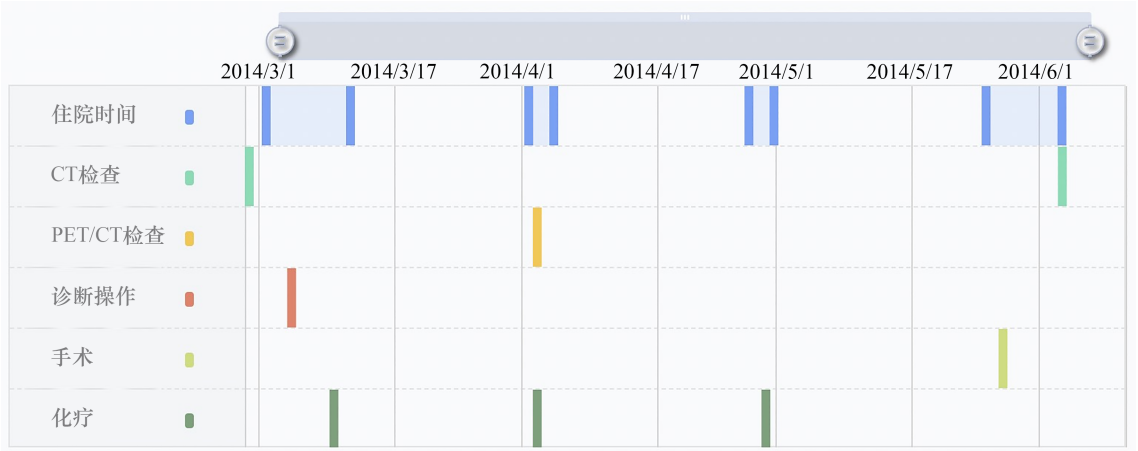

图3　健康状况监测可视化

注：图中显示为模拟数据。

同诊疗方案的患者等，但院内数据库往往无法支持患者信息筛选或无法支持复杂条件来查找特定患者人群。智能检索技术的实现能够基于实际临床筛选需求，协助临床医生和科研人员快速找到指定患者人群，是支撑后续的数据分析与应用的有效工具。该技术基于时间、检查、诊疗事件等各类实体形成数据检索模型，以“事件 / 字段 – 比较符 – 结果”实体结构以及多种变形为基础检索结构，将多个条件以 AND/OR 关系串联形成检索条件组，亦能够支持将多个条件组进行嵌套形成树状架构的关联条件。上述具有较高的自定义程度的检索条件，能够囊括大多数基础和复杂条件的组合，支持尽可能多的临床应用场景。

上述智能检索可通过两种交互方案实现。一是通过菜单式交互，由使用者选择基础检索结构中的每个元素，自主搭建树状检索结构，这种方式多层嵌套结构展示清晰，但对使用者的逻辑要求较高。二是通过检索语句输入条件，利用自然语言提取技术提取每个实体直接带入检索结构和多层嵌套逻辑，无需自行建立复杂条件，更适合低门槛使用。

4. 可视化分析

可视化分析模块通过图形和表格呈现出总体肿瘤患者的各类数据的趋势与分布情况，是挖掘数据价值的有效工具。由于肿瘤临床数据的高敏感性，将数据导出再由专业分析人员进行处理存在一定数据泄漏的风险，所以直接在线上应用数据库进行操作能够尽量规避这一问题。而且由于医生具有数据分析人员所不具备的临床专业敏感度，直接由医生在线上进行分析能够更容易发现和总结科研问题。

在线上数据库操作，肿瘤临床大数据管理系统应提供较为完整的数据处理与分析体系和低门槛的操作系统。首先，基于原数据库中的所有信息脱敏后并加以存储形成应用数据库，确保原临床数据库中的数据变化能够同步至分析数据库用于分析，但分析数据库中的任何操作均不会对原临床数据库产生影响。其次，以分析数据库为数据支撑，基于多种分析应用场景，将权威的生物统计算法打包形成功能模块，支持显著性检验、生存分析、基于不同结局的相关性分析等相关研究。第三，由于临床科研人员统计背景较弱，因此建立菜单式、无代码的交互方式为医生或者获得数据使用许可的科研人员提供便捷，降低可视化分析的使用门槛，协助研究人员更准确地把握研究重点，提供临床指导意见。

5. 机器学习建模

目前已有多项 AI 和医学的跨领域研究对于肿瘤的全流程诊疗发挥了重要作用，但相关研究均需要依靠 AI 领域专家进行算法训练与处理，数据科学领域的专家缺乏临床医学问题的敏感性，与医疗专家的合作也会存在一定的沟通差异成本和时间成本。因此，肿瘤临床大数据管理系统应以降低医生或科研人员使用门槛，促进医生对 AI 领域的涉猎为目的，提供容器化的基础机器学习建模的低代码版操作，让医护和科研人员能够直接使用应用数据库进行模型训练。机器学习能够将专业机器学习算法打包为容器，提供贴近真实临床应用的建模目的，根据实际数据状态和建模特征，实时调整预设的各种超参数，可视化展示建模的评估结果，最终将训练结果直接应用于临床数据中。

五、肿瘤临床大数据建设实例

肺癌是我国全人群发病率和死亡率最高的恶性肿瘤 [46]，以某三甲医院牵头搭建的肺癌临床大数据管理平台为例，能够囊括大多数肿瘤数据管理的应用场景，具有良好的参考意义。该肺癌平台基于本文提出的系统框架，并应用数据管理和科研应用的各类技术和解决方案，以建立国家级数据管理系统为目标，构建了肺癌临床管理大数据平台。该肺癌临床管理大数据平台的数据管理和应用流程如图 4 所示。

首先，通过每家医院内网的虚拟专用网络（VPN）管道，严格遵循“敏感数据不出院”的原则，在院内完成肺癌患者精准查询、数据库定位、智能化自动数据提取，把肺癌患者列表部署于院内系统，实现每日数据增量更新，保证肺癌平台所存储数据的及时性；而基于互联网平台部署于院外的多中心肺癌大数据平台，则整合了多个医疗机构服务器脱敏后的数据，同时对平台使用者进行了严格的权限管控，确保数据使用范围均经过授权管理，从而形成较大规模的真实世界肺癌临床数据库。获取的所有数据依据核心标准数据集和肺癌专病标准数据集，对每个字段规则的值域、单位进行识别和转化，再根据肺癌临床含义，从病史、入 / 出院记录、CT 和 PET-CT 图像以及手术视频等各类数据中

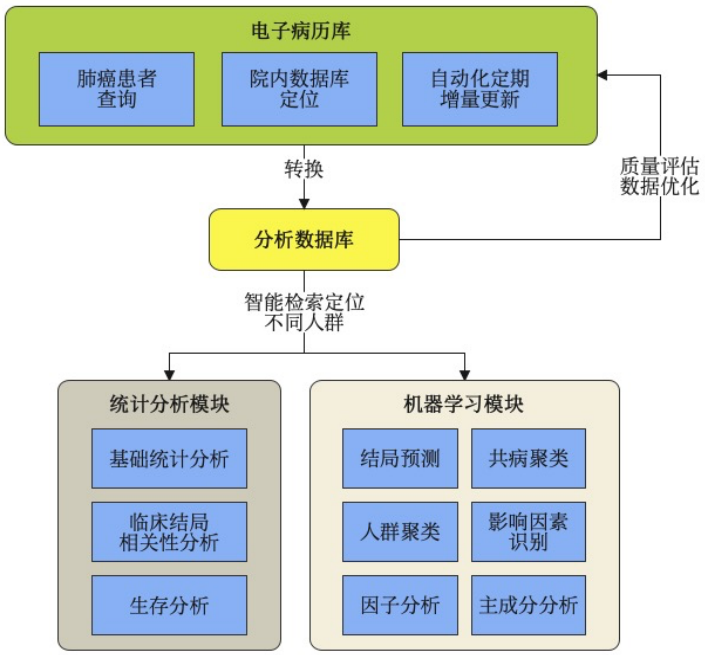

图4　肺癌临床管理大数据平台流程图

采用多种AI算法提取肺癌临床特征，结合多组学肺癌信息，建立完整且通用的电子病例库，支持医护人员随时查阅相关患者的全部肺癌诊疗记录。

其次，将电子病例库转换为分析数据库能够支持多种应用需求，包括进行患者人群筛选、统计分析、机器学习建模、数据质量评估。以进行肺癌患者的5年生存率进行分析为例，基于平台所采集的所有肺癌患者数据，选择生存分析算法，选择是否死亡为临床结局进行计算，绘制Kaplan-Meier生存曲线和构建Cox比例风险模型，还可对患者数据进行基础统计学描述以及与指定结局相关性分析定位与生存率相关的变量。若根据现有字段无法筛选出指定的分析人群，例如需根据患者年龄转换为年龄段数据，可以通过函数实现新字段的计算再放入筛选条件实现人群划分。若需要对现存活患者的5年生存率进行预测，即可通过现有特征变量以及合适的建模方案和超参数进行结局预测建模训练，并将预测结果直接应用于临床患者管理。若该平台的肺癌患者数据质量较差，对分析和建模造成严重干扰，可率先对数据质量进行评估打分，根据评估结果优化数据内容。

综上所述，通过多种线上低门槛功能的使用，该肺癌临床管理大数据平台的应用能够极大程度提升肺癌数据的应用效率，使得临床医生和科研人员能够直接介入AI领域，是临床医学和AI技术的重要桥梁。

六、结语

本文以当前肿瘤临床大数据管理与科研应用现状和目前面临的痛点问题为切入点，阐述了数据管理系统的实现价值，提出了该系统的整体框架并对其进行描述。本文基于实际应用场景和需求，以肺癌为例，讨论了如何应用数据管理和临床与科研应用的各类技术来构建完整的肿瘤临床数据管理系统的功能模块，为真实世界肿瘤数据科研提供了解决方案。本文描述的方案结合了医疗信息化和AI，分析了如何充分利用新兴技术来赋能肿瘤临床数据管理与科学研究，实现最大程度临床数据挖掘，赋能健康医疗，改善全民肿瘤诊治服务。

本文基于部分肿瘤临床数据的使用场景进行讨论，未能穷尽和探讨所有复杂的真实世界数据的应用场景。新兴技术发展日新月异，肿瘤临床大数据管理系统的未来发展应结合现实问题和需求，在以下方面谋求突破：① 根据国家数据标准，与肿瘤领域专家的建议相结合，形成全国性通用的各肿瘤病种数据标准规范，由国家推广至各级医疗机构，为建立国家级肿瘤临床大数据管理平台奠定基础；

② 未来应加强跨学科专业人才的培养，建立AI领域与肿瘤医学的交叉合作网络，进一步挖掘肿瘤临床数据价值；③ 为医学专业人员和数据科学专家建立可高效合作的交流平台，形成肿瘤临床数据共享生态，促进对医疗数据的创新与探索。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: August 21, 2022; **Revised date:** September 27, 2022

Corresponding author: Li Pengfei is a senior engineer from the Advanced Institute of Information Technology, Peking University. His major research field is intelligent processing, analysis, and application of health and medical big data. E-mail: pfli@aiit.org.cn

Funding project: National Natural Science Fund project (72125009)

参考文献

[1] Ferlay J E M, Lam F, Colombet M, et al. Global cancer observatory: Cancer today [R]. Lyon: International Agency for Research on Cancer, 2020.

[2] Cao M M, Li H, Sun D Q, et al. Cancer burden of major cancers in China: a need for sustainable actions [J]. Cancer Communications, 2020, 40(5): 205–210.

[3] Wang Z Z, Zhou C M, Feng X S, et al. Comparison of cancer incidence and mortality between China and the United States [J]. Precis Cancer Med, 2021, 4: 31.

[4] Xia C F, Dong X S, Li H, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants [J]. Chinese Medical Journal, 2022, 135(5): 584–590.

[5] Li P F, Ma L, Liu J, et al. Surveillance of noncommunicable diseases: Opportunities in the era of big data [J]. Health Data Science, 2022: 1–10.

[6] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks [J]. Nature, 2017, 542(7639): 115–118.

[7] McKinney S M, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening [J]. Nature, 2020, 577(7788): 89–94.

[8] Lu M T, Raghu V K, Mayrhofer T, et al. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model [J]. Annals of Internal Medicine, 2020, 173(9): 704–713.

[9] Yala A, Lehman C, Schuster T, et al. A deep learning mammography-based model for improved breast cancer risk prediction [J]. Radiology, 2019, 292(1): 60–66.

[10] Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer [J]. NPJ Digital Medicine, 2019, 2(1): 1–10.

[11] Capper D, Jones D T, Sill M, et al. DNA methylation-based classification of central nervous system tumours [J]. Nature, 2018, 555(7697): 469–474.

[12] Wang S, Shi J Y, Ye Z X, et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning [J]. European Respiratory Journal, 2019, 53(3): 1–10.

[13] Chen M Y, Zhang B, Topatana W, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning [J]. NPJ Precision Oncology, 2020, 4(1): 1–7.

[14] Wójcikowski M, Siedlecki P, Ballester P J. Building machine-learning scoring functions for structure-based prediction of intermolecular binding affinity [J]. Methods in Molecular Biology, 2019: 1–12.

[15] Tong Z, Zhou Y, Wang J. Identifying potential drug targets in hepatocellular carcinoma based on network analysis and one-class support vector machine [J]. Scientific Reports, 2019, 9(1): 1–9.

[16] Chaudhary K, Poirion O B, Lu L, et al. Deep learning-based multi-omics integration robustly predicts survival in liver cancerusing deep learning to predict liver cancer prognosis [J]. Clinical Cancer Research, 2018, 24(6): 1248–1259.

[17] Litchfield K, Reading J L, Puttick C, et al. Meta-analysis of tumor-and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition [J]. Cell, 2021, 184(3): 596–614.

[18] Ginsburg G S, Phillips K A. Precision medicine: From science to value [J]. Health Affairs, 2018, 37(5): 694–701.

[19] 姜文华, 王菁蕊. 医疗大数据在肿瘤早期筛查标志物中的研究现状和前景 [J]. 生物医学工程与临床, 2018, 22(1): 116–121.
Jiang W H, Wang J R. Medical big data in early cancer screening: Current landscape and perspective [J]. Biomedical Engineering and Clinical Medicine, 2018, 22(1): 116–121.

[20] Hamet P, Tremblay J. Artificial intelligence in medicine [J]. Metabolism, 2017, 69: S36–S40.

[21] 凌红, 陈龙. 发达国家医院信息系统发展研究及启示 [J]. 中国医院管理, 2014 (6): 78–80.
Ling H, Chen L. Research and enlightenment of hospital information system development in developed Countries [J]. Chinese Hospital Management, 2014 (6): 78–80.

[22] Klaassen B, van Beijnum B J, Hermens H J. Usability in telemedicine systems—A literature survey [J]. International Journal of Medical Informatics, 2016, 93: 57–69.

[23] Yu P, Kibbe W. Cancer data science and computational medicine [J]. JCO Clinical Cancer Informatics, 2021, 5: 487–489.

[24] Henson K E, Elliss-Brookes L, Coupland V H, et al. Data resource profile: National cancer registration dataset in England [J]. International Journal of Epidemiology, 2020, 49(1): 16–26.

[25] Tuppin P, Rudant J, Constantinou P, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France [J]. Revue d'epidemiologie et de sante publique, 2017, 65: 149–167.

[26] Boffa D J, Rosen J E, Mallin K, et al. Using the National Cancer Database for outcomes research: A review [J]. JAMA Oncology, 2017, 3(12): 1722–1728.

[27] Colicchio T K, Cimino J J, Del Fiol G. Unintended consequences of nationwide electronic health record adoption: Challenges and opportunities in the post-meaningful use era [J]. Journal of Medi-

cal Internet Research, 2019, 21(6): e13313.

[28] 杨丽, 王婷, 敖敏, 等. 肺结节与肺癌全程智能管理云平台的构建及临床应用 [J]. 中华肺部疾病杂志(电子版), 2022, 15(1): 11–14.
Yang L, Wang T, Ao M, et al. Construction and clinical application of cloud platform for intelligent management of lung nodules and lung cancer [J]. Chinese Journal of Lung Diseases (Electronic Edition), 2022, 15(1): 11–14.

[29] 刘景丰, 刘红枝, 陈振伟, 等. 肝病和肝癌大数据平台建设体系及其初步应用 [J]. 中华消化外科杂志, 2021, 20(1): 46–51.
Liu J F, Liu H Z, Chen Z W, et al. Construction system and preliminary application of big data platform for liver disease and liver cancer [J]. Advances in Digestive Surgery, 2021, 20(1): 46–51.

[30] 袁骏毅, 张琛, 潘常青, 等. 肺癌早筛管理平台设计与实现 [J]. 医学信息学杂志, 2020, 41(7): 75–79.
Yuan J Y, Zhang C, Pan C Q, et al. Design and realization of lung cancer early screening management platform [J]. Journal of Medical Informatics, 2020, 41(7): 75–79.

[31] Coroller T P, Agrawal V, Huynh E, et al. Radiomic-based pathological response prediction from primary tumors and lymph nodes in NSCLC [J]. Journal of Thoracic Oncology, 2017, 12(3): 467–476.

[32] Krafft S P, Rao A, Stingo F, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis [J]. Medical Physics, 2018, 45(11): 5317–5324.

[33] Coudray N, Ocampo P S, Sakellaropoulos T, et al. Classification and mutation prediction from non‐small cell lung cancer histopathology images using deep learning [J]. Nature Medicine, 2018, 24(10): 1559–1567.

[34] Christie J R, Lang P, Zelko L M, et al. Artificial intelligence in lung cancer: Bridging the gap between computational power and clinical decision-making [J]. Canadian Association of Radiologists Journal, 2021, 72(1): 86–97.

[35] 国家卫生健康委统计信息中心. 2021年11月底全国医疗卫生机构数 [R]. 北京: 国家卫生健康委统计信息中心, 2021.
Statistical Information Center. Number of national medical and health institutions at the end of November 2021 [R]. Beijing: Statistical Information Center, National Health Commission of the PRC, 2021.

[36] 中国医院协会信息专业委员会. 2019—2020年度中国医院信息化状况调查报告 [R]. 北京: 中国医院协会信息专业委员会, 2021.
China Hospital Information Management Association. Investigation report on informatization status of hospitals in China during 2019—2020 [R]. Beijing: China Hospital Information Management Association (CHIMA), 2021.

[37] Ross J S, Krumholz H M. Ushering in a new era of open science through data sharing: The wall must come down [J]. Jama, 2013, 309(13): 1355–1356.

[38] Taichman D B, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: A requirement of the International Committee of medical journal editors [J]. Annals of Internal medicine, 2017, 167(1): 63–65.

[39] 詹启敏, 董尔丹. 健康医疗人工智能指数报告2020 [M]. 北京: 科学出版社, 2021.
Zhan Q M, Dong E D. Health and medical artificial intelligence index report 2020 [M]. Beijing: Science Press, 2021.

[40] Naik A, Edla D R, Dharavath R. A deep feature concatenation approach for lung nodule classification [R]. Switzerland: Proceedings of the International Conference on Machine Learning and Big Data Analytics, 2021: 213–226.

[41] Overhage J M, Ryan P B, Reich C G, et al. Validation of a common data model for active safety surveillance research [J]. Journal of the American Medical Informatics Association, 2012, 19(1): 54–60.

[42] 王安然, 吴思竹, 钱庆. 面向标准化数据整合的医学通用数据模型探析 [J]. 中华医学图书情报杂志, 2019, 27(11): 4–15.
Wang A R, Wu S Z, Qian Q. Medical common data model for standardized data integration [J]. Chinese Journal of Medical Library and Information Science, 2019, 27(11): 4–15.

[43] Weiskopf N G, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research [J]. Journal of the American Medical Informatics Association, 2013, 20(1): 144–151.

[44] Estiri H, Stephens K A, Klann J G, et al. Exploring completeness in clinical data research networks with DQe-c [J]. Journal of the American Medical Informatics Association, 2018, 25(1): 17–24.

[45] Bonner S, McGough A S, Kureshi I, et al. Data quality assessment and anomaly detection via map/reduce and linked data: A case study in the medical domain [C]. Santa Clara: Proceedings of the 2015 IEEE International Conference on Big Data, 2015.

[46] Zheng R S, Zhang S W, Zeng H M, et al. Cancer incidence and mortality in China, 2016 [J]. Journal of the National Cancer Center, 2022, 2(1): 1–9.