◁信息管理▷

脑血管病临床科研大数据平台的设计与实现*

易行健，单思源，杨扬，殷海翔
（郑州大学第一附属医院，郑州 450001）

【摘要】 **目的** 设计并实现一个脑血管病临床科研大数据平台，以提升临床科研与成果转化实力，改善医疗服务质量。**方法** 采用B/S架构，构建包含数据湖、大数据中心和领域数据中心的多层数据架构，运用临床知识图谱、自然语言处理和深度学习技术进行数据治理和科研支持。**结果** 成功建立了一个能够集成海量医疗数据、支持临床科研和提高医疗服务效率的大数据平台，实现了数据的标准化处理和质量控制，提供了高效的数据检索和可视化工具。**结论** 该平台通过大数据技术实现了临床科研数据的有效管理和利用，有助于提升脑血管病的诊疗和科研水平，对医院的信息化建设和医疗服务改进具有重要意义。

【关键词】 脑血管病；大数据平台；临床科研；数据治理；医院信息化
【文献标志码】 A　【文章编号】 1672-4232(2025)05-0084-04
【DOI编码】 10.3969/j.issn.1672-4232.2025.05.023

Design and Implementation of a Big Data Platform for Clinical Research on Cerebrovascular Diseases/*YI Xing-jian, SHAN Si-yuan, YANG Yang, YIN Hai-xiang*(*The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450001, China*)

【**Abstract**】 **Objective:** To design and implement a clinical research big data platform for cerebrovascular diseases, in order to enhance the strength of clinical research and technology transfer, and improve the quality of medical services. **Methods:** A B/S architecture is adopted to build a multi-layer data architecture including a data lake, a big data center and a domain data center. Clinical knowledge graphs, natural language processing and deep learning technologies are used for data governance and research support. **Results:** A big data platform capable of integrating massive medical data, supporting clinical research and enhancing the efficiency of medical services has been successfully established. And it has achieved standardized data processing and quality control, and provided efficient data retrieval and visualization tools. **Conclusions:** This platform has achieved effective management and utilization of clinical research data through big data technology, which is conducive to improving the diagnosis, treatment and research level of cerebrovascular diseases, and is of great significance to the informatization construction of hospitals and the improvement of medical services.

【**Key words**】 cerebrovascular disease; big data platform; clinical research; data governance; hospital informatization

据原卫生部对我国第三次居民死因抽样调查的结果通报，脑血管疾病是我国城乡居民的主要死亡原因之一[1]，通常具有较高的发病率、死亡率、致残率和复发率，治疗难度大，费用高昂。无论对患者还是临床医生来说都是非常沉重的负担。尽管脑血管疾病诊疗新技术、新方法不断涌现，基于最新循证医学证据的指南和指导规范快速更新，临床实践与指南间仍存在着巨大的鸿沟，急需加强医院脑血管疑难病症诊治综合能力建设，保障医院能够持续不断地提供优质的医疗资源服务全省的脑血管病患。同时，发挥医院在脑血管疑难病症方面的研究优势及医疗资源优势，提升全省脑血管疾病医院的诊治水平。

随着科学技术的发展，医疗行业正经历从数字化到平台化再到智能化的转型升级，脑血管病专科领域也同样应该与时俱进，积极探索。为了提升医院临床科研与成果转化实力，亟须依托信息化手段，完善科研平台建设，以此推动脑血管病诊疗工作的规范化、标准化与同质化，进一步规范脑血管病的诊疗行为，改善医疗服务质量，不断提升卒中救治水平。

*基金项目：河南省医学科技攻关计划软科学重点项目（RKX202201007）

1　需求分析

作为危害我国人民的主要病种之一，如何提高针对该病种的预防和救治能力，成为广大临床工作者的主要研究方向。以某省级三甲医院为例，该院年门急诊量近千万人次，出院患者约80万人次，产生的医疗数据量非常庞大，但原本的临床数据平台多为分散的、结果性的，难以满足临床科研的需要。因此，亟须通过数据治理实现海量数据的集成融合，并针对不同疾病领域需求建设专病数据库，以满足科研对数据可用性与适配性的要求。

表1对脑血管病的主要特征及其对应的数据平台需求进行了系统性概括，从临床特征、诊疗过程、数据特性、科研需求和管理难点5个维度出发，明确了构建高效数据平台所需解决的关键问题。这种分类分析帮助识别了在数据采集、整合、存储、安全以及科研支持等方面的具体需求，为后续的数据治理和平台优化提供了清晰的方向。该院通过平台建设，在脑血管病疑难危重症的诊疗、临床科技创新与成果转化、医疗信息化建设以及区域内临床诊疗的辐射带动等方面获得显著提升，旨在建设国内一流、具有一定国际影响的脑血

管病诊疗中心,形成科学合理、运转高效、多学科协作的脑血管病疑难急危重症诊断及救治体系,全面提升脑血管病疑难急危重症的诊断及救治能力。

表1　脑血管病诊疗特征与数据平台需求分析

维度	病情特点	对数据平台要求
临床特征	症状多样,如偏瘫、失语、认知障碍等 诊断依赖多种检查手段 高发病率和复发率	多源数据采集与整合
诊疗过程	诊疗环节复杂,包括急救、手术、康复等多个阶段 需高效协作与实时数据更新	标准化诊疗活动模型 诊疗数据规范化存储
数据特征	数据类型多样,包括结构化、非结构化病历及影像数据 数据量庞大且敏感	构建数据湖与大数据中心 数据清洗 数据安全
科研需求	需要精准、海量的病例数据 涉及队列分析、模式识别及多中心协作	科研大数据平台RDR 病例快速检索 数据共享
管理难点	数据分散、质量参差 数据共享困难 需保障数据隐私与合规	建立数据治理体系 数据脱敏 可视化工具

2　平台设计

为了减轻对临床设备系统的负荷,降低系统维护与升级的成本和工作量,系统整体采用B/S架构,部署在应用服务器端,总体架构见图1。

脑血管病临床科研大数据平台不仅需要具备科室级的专病库,还有针对各个不同研究的课题库,可用于积累专病科研病例数据资产,科室级的专病库可以为不同的科研课题提供支撑[2]。另外,在医院现有临床数据中心或大数据中心与专病库之间,增加了数据湖,便于对原始数据进行管理和对历史数据的溯源,并在此基础上再构建一层临床科研大数据平台。该数据平台将医院原始数据进行清洗与转换。基于临床术语与知识图谱技术、自然语言处理与文本结构化技术进行数据清洗,将数据转换为科研数据模型,为科研人员提供科研需要的病例数据,在降低医院信息部门工作量的同时,通过科研能力反哺临床,为临床提供决策支持和知识库应用,更好地为临床提供服务。

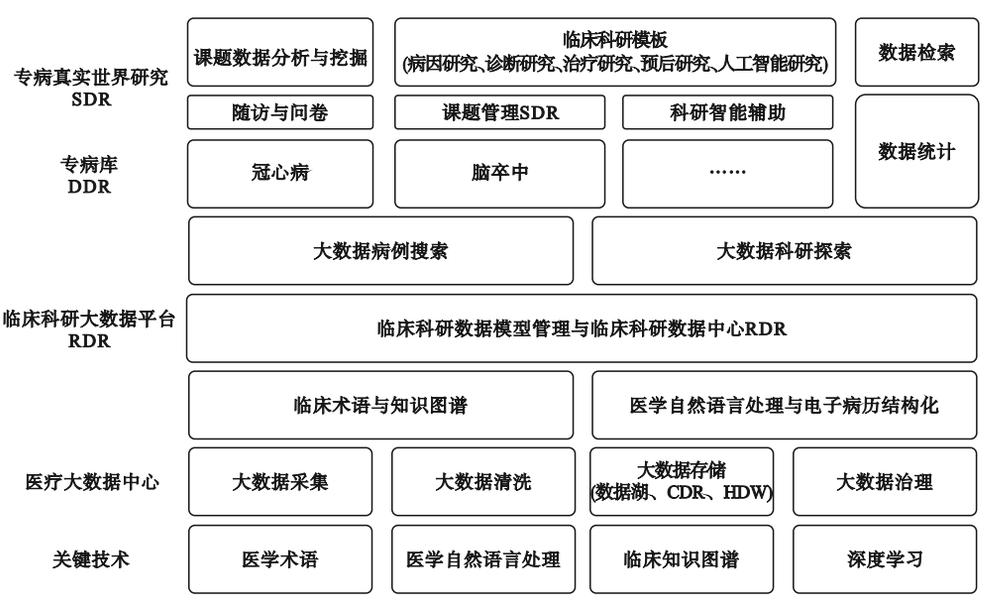
2.1　数据架构

脑血管病临床科研大数据平台架构由数据湖(全量原始数据)、大数据中心(全量数据仓库)和领域数据中心(包括临床数据中心、运营数据中心、科研数据中心等)组成[2]。

数据湖是一种低成本的数据存储环境,它支持将不同来源和类型的原始数据以未加工的形式进行存储。这种架构能够容纳各种数据类型,包括但不限于结构化数据、半结构化数据、文本文档、图像、文件以及消息流等[3]。引入数据湖的数据管理思路,解决数据集成问题,通过低成本技术例如分布式存储技术、低成本硬件、自动化运维等来存储海量的原始数据,将应用建模交由开发者。

全院大数据中心(全量数据仓库),是以大数据采集管理为基础,以大数据安全隐私为核心,以大数据治理体系确保数据质量,以大数据监管系统确保符合法律法规要求,整合大数据资产管理等组件来构建全院数据仓库,并可通过有效管控,对临床、管理、科研等不同方向提供数据服务[4]。

用于储存和管理全院科研数据的科研数据中心在对数据进行初步的自然语言、归一化等处理后形成相对规范的数据,以进行医学科学研究与探索。大数据病例检索可满足临床科研人员直接在临床科研大数据



注:SDR-原始数据审核,DDR-专病库,RDR-科研数据中心,CDR-临床数据中心,HDW-医院数据仓库。

图1　临床科研大数据平台架构

平台上找到科研所需病例[3]。全院科研大数据平台的建立需要跨部门合作，例如医院 IT 部门、临床部门、科研部门和行政部门等。在大数据科研的探索中，对应的病例需要经过预实验，对科研选题的可行性进行初步判断，通过数据搜集与预实验设计验证病例数据和科研选题的可行性，最后，将病例数据导入到科室的专病库中。

2.2　关键技术

2.2.1　临床知识图谱构建算法。门控循环单元（gated recurrent unit，GRU）是一种用于处理序列数据的神经网络结构。基于双向 GRU 的疾病名称规范化技术，利用上下位识别算法和同义词算法将疾病数据与 ICD-10 疾病编码相关联，构建疾病的上下位关系以及同义词关系。采用的上下位识别算法基于构成疾病的汉字序列或词语序列，使用双向 GRU 对中文疾病名称进行向量建模，同时利用注意力机制，以充分考虑疾病的内在结构和语义信息（见图 2）。

2.2.2　构建临床知识图谱的术语体系。基于临床知识图谱的应用，结合双向 GRU 的疾病名称规范化技术，构建实用、统一、受控的术语体系：受控的标准临床术语知识及其关系图谱；症状、诊断、药品、检查、检验、手术、科室等标准术语以及关系知识图谱；统一医院实用临床术语及其关系知识图谱；医院临床描述与标准术语的映射，增加实用同义词关系知识图谱。

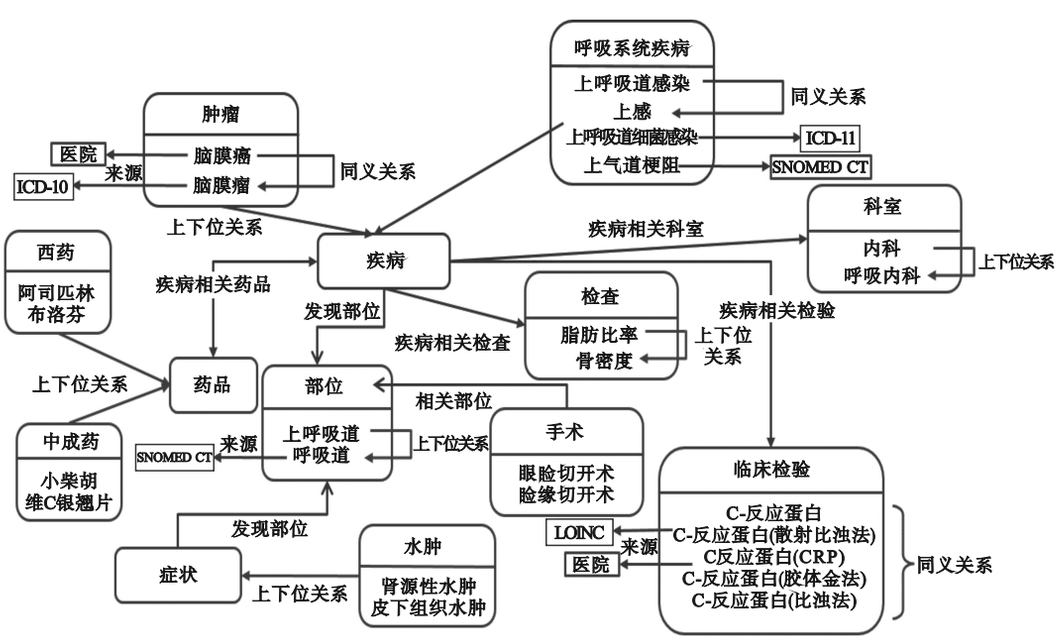
2.2.3　自然语言处理。平台采用自然语言处理技术来实现将非结构化的病历文书与检查报告结构化。在医学领域，结构化数据对于提升医疗服务的质量和效率至关重要，在医学自然语言处理中，需要更注重医学领域语义处理，而不是基于语法的自然语言处理，其中主要包括：首先，医学实体识别，如症状、检查检验等；其次，医学实体属性识别，如药物的剂量等，最后，医学实体与术语关系识别并链接[5]。

2.2.4　机器学习。机器学习是人工智能的一个分支，它使计算机系统能够从数据中学习并改进其性能，而无需进行显式的编程。机器学习算法能够通过分析大量数据来识别模式和趋势，然后利用这些模式和趋势来预测未来事件或进行决策[6]。机器学习可以分成如下几种类型，如有监督学习、无监督学习和半监督学习等，在平台中都有所体现[7]。

3　应用效果

脑血管病临床科研大数据平台从数据采集、管理、分析、协作等多个方面为脑血管病研究提供全方位支撑。

首先，该平台实现了多源数据的全面采集与整合。它能够接入医院近十年的住院和门诊临床数据，包括医院信息系统、电子病历、影像系统、检验系统等多个系统，构建起一个脑血管病专病数据湖。该数据湖采用统一标准的临床文档架构（clinical document architecture，CDA），优化了数据的存储、检索和分析流程。平台还提供数据清洗与治理功能，对原始数据进行清洗、结构化和术语归一化，形成可计算统计的数据集，并支持人工审核和修正机器清洗结果，确保数据质量。此外，平台还提供数据映射规则，将结构化数据映射到诊疗活动模型中，为后续的数据分析奠定基础。



注：ICD-国际疾病分类，SNOMED CT-临床医学术语体系，LOINC-观测指标标识符逻辑命名与编码系统。

图 2　构建临床知识图谱

其次,平台建立了完善的脑血管病数据管理体系。它定义了脑血管病病例应包含的各类信息,为数据模型的构建提供标准化基础。针对脑血管病特有的病历类型,为其定义符合CDA文档模型的结构,并将清洗后的病历资料存储到相应的活动模型中。同时,平台还维护脑血管病特有的诊疗活动模型,例如诊断、治疗、用药、手术等,为脑血管病研究提供全面、精准的数据支持。

在数据检索与分析方面,平台提供了强大的工具和方法。基于医学本体和知识图谱,平台构建了临床术语体系,支持同义词、上下位关系检索,解决了传统搜索"搜不全"和"搜不准"的问题。平台还提供更方便、更智能的语义化检索方式,例如输入疾病名称即可自动生成相关过滤条件,帮助临床科研人员快速筛选脑血管病病例。此外,平台还支持科研大数据探索,基于搜索到的脑血管病病例,提供快速横断面分析和快速队列分析,帮助研究人员验证科研选题的可行性,并支持构建回顾性队列和前瞻性队列,为脑血管病研究提供可靠的数据支持。

平台对促进科研协作与数据共享也有积极的促进作用。它支持多中心参与脑血管病研究,建立登记方案,进行队列构建和数据收集,并提供数据进度跟踪和可视化展示功能,方便各中心协同工作。平台支持数据导出,供第三方统计工具使用,进一步扩大了数据的使用范围。平台还支持脑血管病专家进行远程会诊和移动会诊,促进医疗资源共享,提高诊疗效率。

自该平台上线以来,通过汇集海量数据,并运用一系列复杂的查询条件和智能算法,对病历进行了深入的结构化处理,精准提取关键诊断信息。平台经过计算不同病历之间的相似性,并根据预设的查询规则,从庞大的患者档案库中筛选出符合特定条件的病历数据。这一流程显著提高了处理脱髓鞘疾病、帕金森病、运动障碍以及认知障碍等疾病的结构化病历时的数据处理效率,为临床和科研人员的研究工作提供了坚实的数据支撑。同时,平台利用多样化的算法对大规模数据集进行深入分析,构建并不断优化多种算法模型,从而打破信息孤岛和数据烟囱的现象,促进了医院内部信息的整合与协调。

4　总结

通过应用先进的大数据集成解决方案,可以有效地收集、整合、管理和存储临床研究数据,这一切都以患者的需求为核心,旨在为医疗领域的临床实践、科研探索和管理决策提供支持[8]。本平台通过创新的架构设计和先进的技术应用,为脑血管疾病的临床研究提供了强大的数据支持。未来,随着技术的发展和数据量的进一步扩展,该平台将在推动精准医疗和临床科研方面发挥更加重要的作用[9]。

鉴于医疗数据的庞大体量和高敏感性,以及数据泄露可能带来的严重后果,数据处理过程中必须包括数据的匿名化和加密措施。数据处理还应包括数据的结构化、规范化和标准化,且必须遵循国家的相关标准,以将原本难以直接利用的临床数据转化为有价值的信息资源。此外,基于大数据平台,可以开发出多种数据检索方案,如全文搜索、结构化查询和语义搜索等,以解决当前临床数据在准确性、性能和规模方面的挑战[10]。通过这些技术,可以提高数据检索的效率和准确性。同时,利用数据可视化技术,可以帮助管理人员更好地理解和利用临床数据资产,辅助临床和科研人员发现新的研究思路,并准确评估科研方向的可行性。

参　考　文　献

[1] 陈竺.全国第三次居民死因回顾抽样调查报告[M].北京:中国协和医科大学出版社,2008:33-85.

[2] 谭淼,祝圆,余夏晓波,等.基于大数据的心脑血管疾病防治的信息化设计及评价[J].中国心脏起搏与心电生理杂志,2023,37(4):331-334.

[3] 张琼瑶,黄基,林兰,等.基于全院统一大数据平台的急诊专病库建设的实践探索[J].中国数字医学,2022,17(4):19-25.

[4] 李鹏,聂刚,刘庆金,等.基于医院科研大数据中心的专病数据库建设实践[J].中国数字医学,2023,18(9):85-89,120.

[5] 台耀永,武胜勇,罗枭,等.机器学习在脑血管疾病预后预测中的应用现状及展望[J].中国数字医学,2023,18(10):83-91.

[6] 本刊新媒体部.大数据中心为临床决策、医院管理及科研赋能[J].中国数字医学,2022,17(6):119-120.

[7] 萧锘,曹磊,叶琪,等.医院临床科研大数据平台数据资源分层设计研究[J].中国数字医学,2022,17(9):90-94.

[8] 胡军军,谢晓军,石彦彬,等.电信运营商数据湖技术实施策略[J].电信科学,2019,35(2):84-94.

[9] 国家卫生健康委统计信息中心.医院数据治理框架、技术与实现[M].北京:人民卫生出版社,2019:20-40.

[10] 程楠,侯豪,牛亚军,等.基于NLP技术后结构化处理的电子病历应用[J].河南医学研究,2021,30(24):4510-4513.

[11] 史榴,梁鹏晨,常庆,等.机器学习在医用金属材料特性研究中的应用[J].中国组织工程研究,2024,28(17):2766-2773.

[12] 张国明,陈安琪.基于区域健康信息平台的医疗大数据利用探索[J].中国卫生信息管理杂志,2016,13(3):290-294.

[13] 汪鹏,吴昊,罗阳,等.医疗大数据应用需求分析与平台建设构想[J].中国医院管理,2015,35(6):40-42.

通信作者:杨扬(1978-)女,硕士研究生,高级工程师;研究方向:医学信息化。

收稿日期:2025-02-25
修回日期:2025-03-06　　(编辑　张瀚予)