

基于人工智能的健康医学数据平台建设 与深度治理*

曾 可¹ 王 谢¹ 吴 迪¹ 厚子豪¹ 徐腾达² 马 瑞² 胡海洋³ 张 锋¹

(¹ 中国医学科学院北京协和医院信息中心 北京 100730

² 中国医学科学院北京协和医院健康医学部 北京 100730

³ 中国医学科学院北京协和医院院办公室 北京 100730)

〔摘要〕 目的/意义 构建健康医学数据平台, 实现数据快速筛选、高效管理和智能化提取, 服务临床科研与患者健康管理, 促进医疗健康大数据规范化应用。方法/过程 基于医院信息系统、临床数据中心等系统, 整合提取健康体检数据, 进行规范化治理和安全存储, 建设标准化健康管理数据库; 利用人工智能和大数据技术构建深度学习模型, 实现健康医学数据的多维度检索和科研指标自动提取。结果/结论 数据平台已汇集 31 万余名体检患者的健康数据, 完成肥胖、酒精代谢、医疗人员职业健康 3 大专科库构建, 可支撑多场景应用, 满足临床科研数据采集需求, 为个性化健康管理服务奠定基础。

〔关键词〕 健康医学数据平台; 人工智能; 专科库; 健康管理服务; 数据治理

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.05.011

Construction and In-depth Governance of a Health and Medical Data Platform Based on Artificial Intelligence

ZENG Ke¹, WANG Xie¹, WU Di¹, HOU Zihao¹, XU Tengda², MA Rui², HU Haiyang³, ZHANG Feng¹

¹Department of Information Management, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China; ²Department of Health Management, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China; ³Department of Administrative Office, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

〔Abstract〕 **Purpose/Significance** To construct a health and medical data platform to realize rapid data screening, efficient management and intelligent extraction, so as to serve clinical research and patient health management, and promote the standardized application of medical and health big data. **Method/Process** Based on hospital information system (HIS), clinical data repository (CDR) and other systems, the study integrates and extracts health examination data, conducts standardized management and safe storage, and constructs a standardized health management database. Artificial intelligence (AI) and big data technology are used to build deep learning models to realize multi-dimensional retrieval of health and medical data and automatic extraction of scientific research indicators. **Result/Conclusion** Through the data platform, health data from over 310 000 medical examination patients have been aggregated, and 3 specialized disease databases on obesity, alcohol metabolism, and occupational health of healthcare workers have been established, which can support multi-scenario applications, meet the needs of clinical scientific research data collection, and lay the foundation for personalized health management services.

〔Keywords〕 health and medical data platform; artificial intelligence (AI); specific disease database; health management service; data governance

〔修回日期〕 2025-01-13

〔作者简介〕 曾可, 助理研究员, 发表论文 8 篇; 通信作者: 张锋。

〔基金项目〕 中央高水平医院临床科研业务费资助项目 (项目编号: 2022-PUMCH-B-032)。

1 引言

健康医疗大数据对于改进健康医疗服务模式、加快经济社会发展具有重要作用,是国家重要的基础性战略资源^[1]。医疗大数据的广泛应用是实现传统医学向精准医学转变的必要前提和核心动力^[2-3]。近年来,国家高度重视对医疗大数据的分析和利用,出台多项政策鼓励医疗机构、企业、研究机构等进行相关应用和探索。2018 年 4 月《全国医院信息化建设标准与规范(试行)》^[4]出台,指出要利用人工智能、大数据平台等先进技术处理和研究临床医疗数据,实现医疗数据源于患者、服务于患者。2021 年 6 月《国务院办公厅关于推动公立医院高质量发展的意见》^[5]中指出要推动大数据等新一代信息技术与医疗服务深度融合。2024 年 11 月国家卫生健康委员会等三部委联合发布《卫生健康行业人工智能应用场景参考指引》^[6],明确人工智能在健康管理、慢性病管理、疾病筛查和预测等方面的应用场景,推进卫生健康行业“人工智能+”应用创新发展。随着生活水平不断提高,公众健康意识显著增强,越来越多的人选择定期进行系统、全面的健康检查,医院积累了大量健康体检数据。利用这些数据开展科研工作和临床转化,实现基于数据的智能感知和预警,对医疗健康行业发展具有战略意义^[7-11]。我国健康服务行业发展潜力巨大,但在数据方面存在范围局限、碎片化、不统一、质量不高、核心价值难提取等问题^[12-15],数据的统一治理、标准化建设及健康管理一体化顶层设计迫在眉睫。基于人工智能(artificial intelligence, AI)和大数据技术构建智能化健康管理数据平台,实现健康医学体检数据的快速筛选、高效管理,以及科研指标智能化提取,将有效提升医疗机构整体临床科研能力,加速学术研究成果转化,为开展高效智能的健康管理奠定坚实基础。

2 健康医学数据平台的建设要点

2.1 健康医学数据深度整合治理

健康医学数据包含丰富的临床科研信息,但存

在语言模糊、语法结构多样、语义表达复杂等问题,给临床科研目标变量提取带来困难。本研究针对增长的医疗数据及历史数据,对非结构化和非标准化数据进行深度治理解析,运用规则逻辑及并行计算、语义识别、知识挖掘等 AI 能力,提取关键科研信息,满足多场景临床和科研需求。

2.2 健康医学标准化数据库构建

基于医院信息系统(hospital information system, HIS)、临床数据中心(clinical data repository, CDR)、影像存储与传输系统(picture archiving and communication system, PACS)、检验信息系统(laboratory information system, LIS)、体检系统等,对临床数据整合、汇聚,对多源多维健康体检数据进行规范化采集、治理、安全存储及分析,形成健康体检数据集规范,建设健康医学标准化数据库。在此基础上搭载智能科研检索系统,实现病例筛选、队列构建、指标提取等功能,助力循证医学临床科学研究。

2.3 基于算法模型的健康医学数据挖掘

通过健康医学标准化数据库建设,实现技术标准规范及科研指标结合,利用深度学习模型确定关键技术,融合检测-评估-干预三位一体数据模型,实现重点人群健康大数据挖掘及智能化服务,从数据挖掘到真实世界数据应用。

3 健康医学数据平台整体设计

3.1 健康医学数据治理体系

数据治理是基于通用科研数据模型,对专病研究涉及的多源异构数据进行深度处理、分析、整合及提取的数据管理方式。按数据仓库分层设计的思路,将健康医学数据平台标准化治理体系划分为数据存储层、数据整合层、数据治理层、数据治理管理层,见图 1。首先将体检、HIS、CDR 等系统原始数据抽取至健康医学数据库,结合数据模型、分析模型、通用知识等进行集成存储;其次通过服务总线、服务引擎和数据抽取工具实现数据的外部交换和整合;再次通过标准化数据对照和规则转换进

行数据清洗，去重、处理缺失值、纠正错误数据后，依据体检数据集规范数据格式、编码和命名，完成体检数据治理；最后，构建数据治理管理体系，实现从数据汇集、治理、建模到应用的全过程

规范管理。通过深度治理，解析和提取体检健康各类文本数据，保障数据指标准确性和完整性，为后续专病健康管理研究和应用提供支持。

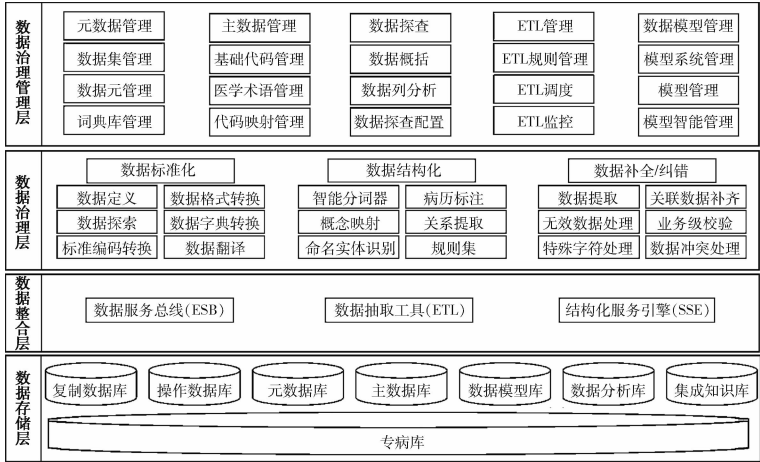


图 1 健康医学数据平台数据治理体系

3.2 智能科研检索

基于体检健康管理的业务过程，提取基本信息、检查检验信息、费用信息、体检报告及建议等数十张关键数据表，同时构建标准化体检模型数据集，定义体检数据的属性、信息和值域代码；基于科研和临床场景设计高效智能数据搜索工具，支持海量复杂健康

医学数据的多维度检索，包括关键词检索、高级检索等。高级检索具备多条件设置、限定条件和事件搜索等复杂功能，见图 2。利用语义分析和关联数据，提供近义词拓展查询、历史记录展示及收藏等功能，搜索结果支持导入/导出、分组展示、多维度展示及可视化统计分析等操作，助力用户高效利用检索结果，提升工作效率与准确性。

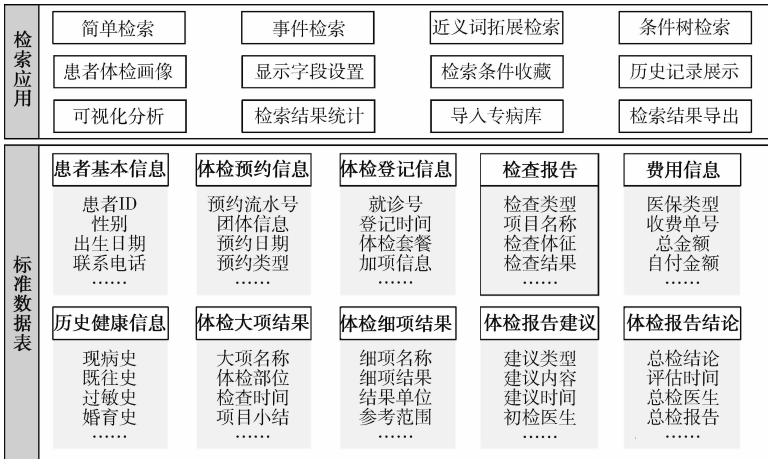


图 2 健康医学数据平台数据检索模型

3.3 科研专病库

科研专病库是一种高效的数据管理工具，通过

筛选特色专病数据，实现对指标、模型、模板等的管理，并支持自定义规则与统计分析。其构建流程包括：了解用户关于患者服务和科研管理等方面的

数据需求，基于通用科研数据模型构建全量科研指标库；由用户基于科研指标建立专病模型；平台根据用户设置的入组条件和数据提取规则自动筛选入库样本；经用户确认后，建立专病库，生成专病科

研数据，见图 3。为确保数据安全、可控，根据医院信息化管理要求设计权限管理、脱敏展示、任务审核和日志记录等功能，确保健康体检数据在存储、使用等环节“可防、可视、可控”。

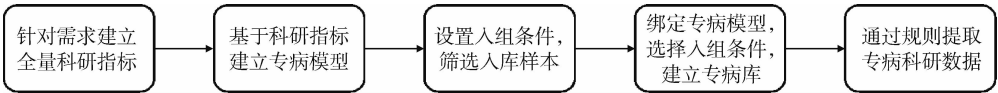


图 3 健康医学数据平台科研专病库构建流程

4 健康医学数据平台技术方案

4.1 自动化数据采集与处理

体检原始数据量大，业务实时性和稳定性要求高。因此将健康医学源数据分为人员、设备等基础数据，以及临床、体检等业务数据。基础数据定时完全抽取，保证完整性；业务数据增量抽取，减少系统干扰。数据抽取在业务闲时进行，利用多 CPU 资源，设置 14 个并行任务，确保采集效率。平台提供自动采集、计算任务管理、任务调度、资源管理、应用开发等功能，集成数据过滤、数据转换、随机取样、数据拆分、SQL 脚本等 13 种数据预处理组件，实现从数据采集到处理的自动化，见图 4。日志监测显示，每日数据采集任务成功率超过 99%。已完成 2005—2024 年 31 万余名体检患者 90 万份体检报告、180 万条体检建议的自动更新入库。

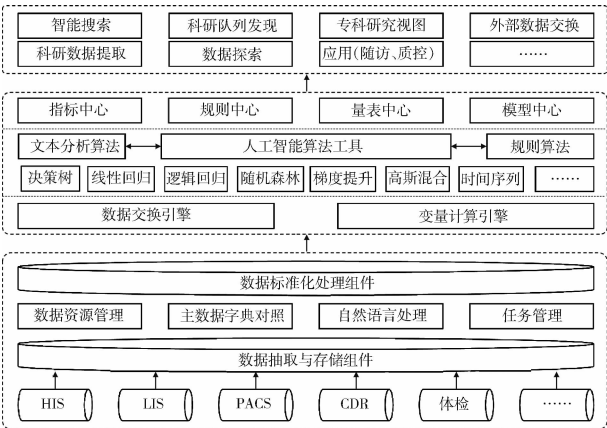


图 4 健康医学数据平台技术架构

4.2 基于人工智能的数据挖掘

针对体检数据标准化、结构化程度较低的特点，平台装载 23 种机器学习算法工具包，包括决策树、线性回归、逻辑回归、随机森林等，并提供相似度计算、文本关键词提取、文本摘要提取 3 种文本分析算法，以及评分卡、决策表两种规则算法，确保大数据开发和应用部署快速可行。在此基础上，建立健康医学大数据表达、管理和访问的标准化技术体系，利用自然语言处理、医学本体模型等技术将数据自动结构化、标准化，形成多维、动态、异构、多层次的健康医学大数据平台。目前已完成肥胖专病库十万级数据的挖掘及建模分析，通过平台预置的 XGBoost 算法学习各类体检指标与身体质量指数（body mass index，BMI）关系，构建回归模型，迭代训练捕捉非线性关系和复杂交互特征，实现 BMI 值高效预测。通过特征重要性评估，挖掘与肥胖群体 BMI 值强相关的预测指标，为健康干预和诊疗提供数据支持。

4.3 基于大数据的搜索和分析引擎

为满足海量健康医学数据的高级检索、事件检索、条件树检索等复杂场景，采用 ElasticSearch 结合人工智能技术构建分布式、高扩展、高可用的搜索和分析引擎。平台基于小模型提取疾病、诊断、症状等标签；基于大模型和医学知识库，结合自然语言理解技术提取推理指标；基于深度学习和知识图谱捕捉更深度的语义信息，此外还提供大数据处理和调度能力，为后续研究队列构建、专科专病研究视图、数据交换、患者随访和全生命周期健康管理等应用提供有效支撑。

5 健康医学数据平台价值分析与展望

5.1 对患者管理和临床研究的重要意义

健康医学数据平台能快速检索海量体检数据,精准提取关联病历中的科研信息。对患者而言,汇集个人医疗健康数据,深入挖掘健康信息,可以发现潜在风险或疾病倾向,提供个性化、定制化的健康管理方案。对老年人、慢性病患者等重点人群进行深度分析,掌握健康状况及变化趋势,协助其做好慢病监测和居家管理。对医生而言,平台提供全面高效的工作平台,促进医疗信息交流利用,帮助其快速识别目标患者,开展临床研究。通过系统性学习大量病例,医生可获得临床科研的新思路、新方法。通过数据平台建立以健康管理为视角的信息交互模式,促进不同领域医生开展协作研究,推动医疗机构从关注疾病向关注全生命周期健康管理转变。目前,平台已完成肥胖、酒精代谢、医疗人员职业健康 3 大专科病数据库。其中,肥胖、酒精代谢和医疗人员职业健康专科病数据库达到十万量级,已构建医院特色健康科研队列,为逐步建立健全健康医学亚专业研究奠定基础。基于平台的结构化数据提取和生成式语言模型,已实现对体检报告的智能化解读和自动化问答功能,为患者提供更加细致的异常指标解读和更加深入的医学知识,每日服务患者超 300 人。

5.2 相较于传统数据平台的优势

与传统数据平台相比,基于人工智能的健康医学数据平台在处理海量多源异构数据方面更高效、实时、精准,满足科研数据标准化、结构化和高质量要求。针对复杂研究场景,利用自然语言处理和深度学习技术,平台可提供全面数据指标检索和灵活项目管理,提高科研效率,降低数据处理门槛。基于人工智能和大数据技术挖掘分析群体健康数据,发掘疾病或亚健康状态的特征指标和高危因素,对疾病风险预测模型构建、重大慢性病防控、个性化健康指导和医疗资源精准分配具有重要意义。

5.3 不足及未来展望

健康医学数据平台是健康大数据应用的有效尝试,满足了医疗机构在数据分析及利用等方面的需求,但其设计和建设存在不足。一是复杂医疗文本解读效果不理想。目前平台利用逻辑规则和人工智能技术可对大部分医疗文本进行解读、分析及科研指标提取,但面对部分结构多样、逻辑性强的复杂文本,仍存在循证推理难、推理精准性不足、大模型“幻觉”、溯源可解释性差等问题。未来将探索任务多步推理和证据检索循证等方法,提高医疗认知模型推理的准确性与可解释性。二是跨学科的健康管理科研与临床合作有待进一步加强。虽然平台建设促进了不同专科临床医生之间的科研交流合作,但基于平台的专病研究规模较小。未来将建立完善的跨学科协作机制,促进不同领域专家深度交流合作,加强培训和宣传,提高用户对平台的接受度和使用度,鼓励医院与高校或科研机构等建立合作,共同推进医疗大数据开发与应用。

6 结语

健康医学数据平台通过规范化获取和分析医学数据,建设标准化医疗大数据平台,全方位服务于患者健康管理和临床前沿科研工作,有力推动医疗健康大数据治理朝着规范化、高效化的方向发展。平台借助创新的实时数据治理体系,将健康体检数据高效转化为极具价值的研究数据,实现了科研数据资产的科学规划与有效管理,加速了健康医学科技成果转化,使前沿医学成果能够更快地惠及公众。

作者贡献: 曾可负责研究设计、系统构建、论文撰写;王谢、吴迪、厚子豪负责系统建设、应用数据分析;徐腾达、马瑞、胡海洋负责科研检索方案设计、专病库流程设计与科研应用;张锋负责论文审核。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 关于促进和规范健康医疗大数据应用发展的指导意见

[EB/OL]. [2024 - 10 - 27]. https://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.

2 胡碧莲, 胡海波. 医疗健康数据国内外研究主题演化路径与述评 [J]. 中国卫生事业管理, 2024, 41 (12): 1434 - 1440.

3 牛刚, 王建军, 乔小妮. 试论医疗健康大数据对医疗信息化发展的影响 [J]. 智慧健康, 2024, 10 (26): 17 - 19, 38.

4 全国医院信息化建设标准与规范 (试行) [EB/OL]. [2024 - 10 - 27]. <http://www.nhc.gov.cn/ewebeditor/uploadfile/2018/04/20180413162542120.pdf>.

5 关于推动公立医院高质量发展的意见 [EB/OL]. [2024 - 10 - 27]. https://www.gov.cn/zhengce/zhengceku/2021-06/04/content_5615473.htm.

6 卫生健康行业人工智能应用场景参考指引 [EB/OL]. [2024 - 10 - 27]. <http://www.nhc.gov.cn/guihuaxxs/gongwen12/202411/647062ee76764323b29a1f0124b64400.shtml>.

7 王璠, 贾良春. 医院体检中心数字化转型现状、挑战与前景 [J]. 中国当代医药, 2024, 31 (23): 129 - 133.

8 袁梦雪. 国内外健康医学科学数据管理平台对比分析 [J]. 数字图书馆论坛, 2020 (1): 11 - 19.

9 于琦, 王映辉, 李宗友, 等. 智能化中医健康管理云平台构建与服务 [J]. 医学信息学杂志, 2023, 44 (1): 54 - 58, 69.

10 梁婷, 李运明, 杨孝光, 等. 健康体检数据治理平台建设实践 [J]. 数字技术与应用, 2023, 41 (12): 173 - 175.

11 顾东兴, 曾斐, 王丹丹, 等. 数据驱动的精准健康管理平台设计与应用 [J]. 中国卫生资源, 2022, 25 (5): 613 - 617.

12 余玉刚, 王耀刚, 江志斌, 等. 智慧健康医疗管理研究热点分析 [J]. 管理科学学报, 2021, 24 (8): 58 - 66.

13 顾东兴, 曾斐, 王丹丹, 等. 数据驱动的精准健康管理平台设计与应用 [J]. 中国卫生资源, 2022, 25 (5): 613 - 617.

14 杨娅, 吕国卿, 马培元, 等. 基于大数据的医院健康管理平台的构建与实践 [J]. 济宁医学院学报, 2022, 45 (3): 189 - 191.

15 王维霞, 高山, 王华, 等. 我国健康医疗大数据相关标准及体系建设分析 [J]. 信息技术与标准化, 2024 (6): 70 - 73, 82.

(上接第 55 页)

8 VINCENT A C S R, SENGAN S. Edge computing - based ensemble learning model for health care decision systems [J]. Scientific reports, 2024, 14 (1): 26997.

9 孙清兰. 高频词与低频词的界分及词频估算法 [J]. 中国图书馆学报, 1992 (2): 78 - 81, 95 - 96.

10 VAN ECK N, WALTMAN L. Software survey: VOSviewer, a computer program for bibliometric map [J]. Scientometrics, 2010, 84 (2): 523 - 538.

11 端木飞雪. 我国档案数据治理研究趋势与研究热点分析 [J]. 江苏科技信息, 2024, 41 (16): 82 - 88, 104.

12 BECHER T, TROWLER P. Academic tribes and territories [M]. London: McGraw - Hill Education (UK), 2001.

13 崔雷, 刘伟, 闫雷, 等. 文献数据库中书目信息共现挖掘系统的开发 [J]. 现代图书情报技术, 2008 (8): 70 - 75.

14 单磊. 基于期刊主题词共现网络的生物医学领域单篇论文新颖性评价研究 [D]. 沈阳: 中国医科大学, 2022.

15 唐果媛, 张薇. 基于共词分析法的学科主题演化研究进展与分析 [J]. 图书情报工作, 2015 (5): 128 - 136.

16 邓君, 马晓君, 毕强. 社会网络分析工具 Ucinet 和 Gephi 的比较研究 [J]. 情报理论与实践, 2014, 37 (8): 133 - 138.

17 BRIN S, PAGE L. The anatomy of a large - scale hypertextual web search engine [J]. Computer networks and ISDN systems, 1998, 30 (4): 107 - 117.

18 张东红, 张冬芳, 马一凡. PageRank 算法在引文网络中的应用研究 [J]. 电脑知识与技术, 2021, 17 (27): 44 - 46.

19 薛钰千. 基于双加权引证网络主路径分析的学科主题演化研究 [D]. 南京: 南京理工大学, 2023.

20 董建成. 医学信息学的现状与未来 [J]. 中华医院管理杂志, 2004, 20 (4): 232 - 235.

21 代涛. 医学信息学的发展与思考 [J]. 医学信息学杂志, 2011, 32 (6): 2 - 16.