

DOI:10.13288/j.11-2166/r.2026.08.014

基于多模态人工智能技术的中西医结合炎症性肠病大数据平台及专病队列的构建

沈照峰, 张梦媛, 朱磊[✉], 沈洪

江苏省中医院/南京中医药大学附属医院, 江苏省南京市秦淮区汉中路 155 号, 210029

[摘要] **目的** 构建基于多模态人工智能 (AI) 技术的中西医结合大数据平台与专病队列, 为开展高质量中医药循证研究提供方法学支撑与数据基础。**方法** 以炎症性肠病 (IBD) 为研究对象, 参考国内外专病队列数据库构建模式, 依托江苏省中医优势病种诊疗大数据平台, 制定中医优势病种研究核心数据集, 整合江苏省 11 家三级甲等中医院多源异构数据, 搭建中西医结合 IBD 医疗大数据平台; 应用自然语言处理 (NLP)、机器学习等 AI 技术, 对真实世界数据进行深度治理 (包括抽取、转换、脱敏、加密、清洗、映射、人工标注), 实现专病数据的标准化、结构化、归一化; 建立动态患者筛选入组、数据采集与随访管理体系, 形成 IBD 专病队列; 采用光学字符识别 (OCR) 结合霍夫变换直线提取法实现肠镜报告文本识别与提取, 利用卷积神经网络 (CNN) 模型对舌象、肠镜图片进行特征分析与量化。**结果** 初步制定了中西医结合 IBD 专病数据元标准, 完成了 36 890 条 IBD 全量诊疗记录数据治理, 实现了中西医数据的标准化映射、整合与质控, 构建了多模态 IBD 大数据平台 (14 424 人次); 建立了包含 1781 例 IBD 患者的专病队列; 开发了 AI 辅助的中医预测模型。**结论** 形成了一套可推广的中西医结合多模态 IBD 大数据平台及专病队列建设技术框架, 实现了中西医真实世界数据的深度融合, 可为中医药循证研究提供高质量的数据支撑。

[关键词] 炎症性肠病; 医疗大数据; 大数据平台; 专病队列; 人工智能

炎症性肠病 (inflammatory bowel disease, IBD) 是一种慢性、复发性肠道炎症性疾病, 主要包括溃疡性结肠炎 (ulcerative colitis, UC) 和克罗恩病 (Crohn's disease, CD)^[1]。IBD 的发病机制尚不明确, 给临床诊疗和疾病管理带来巨大挑战。中医药具有整体观和辨证论治的多维度诊疗模式优势^[2]。健康医疗大数据是国家基础性战略资源, 对提高临床医疗服务质量和专科专病诊疗效果具有重要意义^[3]。随着精准医学、医疗大数据和人工智能 (AI) 技术的快速发展, 基于真实世界数据构建高质量的专病队列数据库已成为开展循证医学研究的

重要基础^[4-5]。

目前国外已经有 IBD 专病队列的建立, 但这些队列以生物标志物和临床指标为主, 多为单一维度疾病模型, 缺乏对患者整体性和长期动态演变的评估^[6-8]。我国已在心血管、肿瘤等领域搭建了区域性的专病数据库^[9], 但 IBD 专病队列数据库建设仍处于雏形阶段, 无法满足中西医结合临床诊疗的需求^[10]。基于此, 我们依托江苏省中医优势病种诊疗大数据平台, 利用 AI 大模型整合全省多中心多模态 IBD 健康医疗大数据, 构建中西医结合 IBD 大数据平台与专病队列, 打造中西医协同创新解决方案, 为 IBD 早期诊断、疗效评价和长期管理提供高质量真实世界证据。本研究方案经江苏省中医院伦理委员会批准 (批件号: 2023NL-039-01), 并在国际传统医学临床试验注册平台注册 (注册号: ITMCTR2025002038)。

基金项目: 江苏省中医药科技发展计划 (2022WTZX04); 国家中医药管理局高水平中医药重点学科建设项目 (国中医药人教函 [2023]85 号); 江苏省中医消化病医学创新中心 (苏卫科教 [2022] 15 号); 江苏省中医药领军人才项目 (SLJ0306); 国家中医药管理局重大疑难疾病克罗恩病中西医临床协作项目 (ZDYN-2024-A-087)

✉ 通讯作者: zhulei5100@njucm.edu.cn

1 研究设计与方法

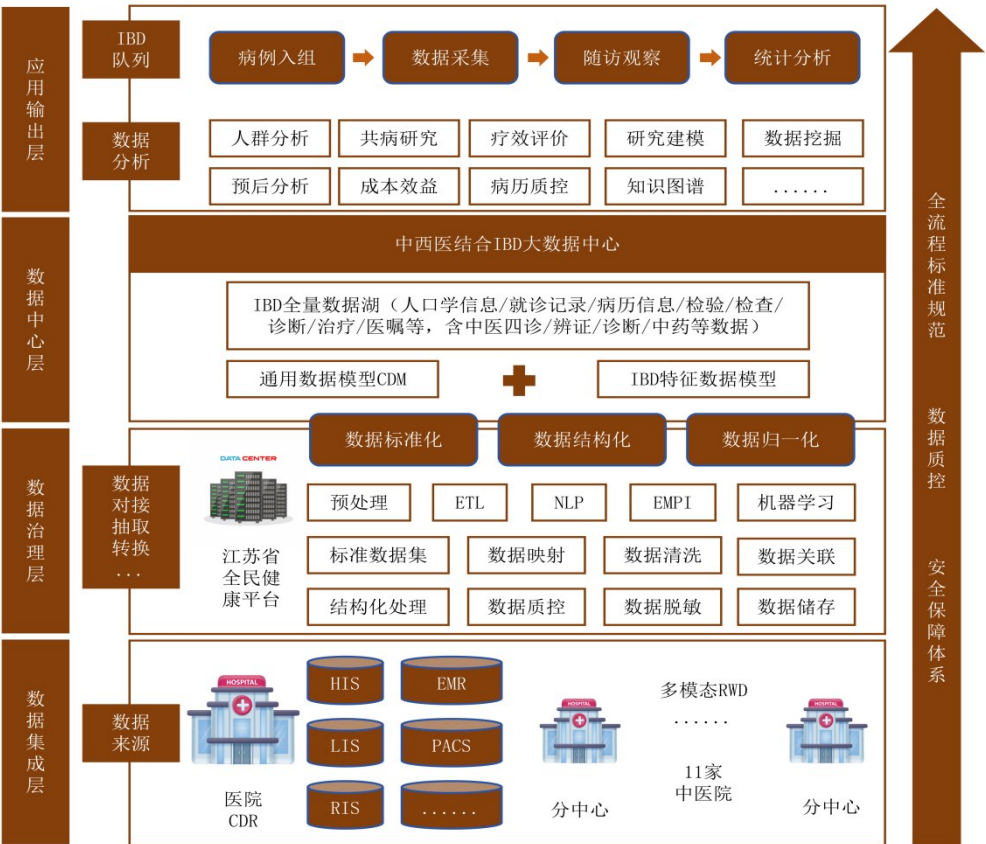
研究总体设计参考 BRIDGE (Best Practice for Developing Integrative Chinese-Western Medicine Databases using Electronic Health Records) 方法学框架^[11], 遵循 STROBE 声明 (Strengthening the Reporting of Observational Studies in Epidemiology) 在队列研究中的报告规范^[12]。

1.1 平台框架

成立 IBD 研究工作组, 成员由临床医学、流行病与卫生统计学、人工智能、软件开发、科研管理等领域具有高级职称的 11 名专家共同组成。参考《国际疾病分类》第 11 版 (ICD-11)^[13]、国家卫生健康委员会相关标准^[14]、《信息安全技术 健康医疗数据安全指南》^[15] 等, 遵循“数据标准化、治理智能化、功能模块化、管理信息化、结果可视化”原则, 设计平台框架总体包含四层核心架构, 即数据集成层、数据治理层、数据中心层、应用输出层

层, 整体架构如图 1 所示。

以 IBD 患者为研究对象, 诊断标准严格参照《中国溃疡性结肠炎诊治指南 (2023 年·西安)》^[16]、《中国克罗恩病诊治指南 (2023 年·广州)》^[17]、《溃疡性结肠炎中医诊疗专家共识 (2023)》^[18]、《克罗恩病中医诊疗专家共识意见 (2024 年)》^[19]。源数据来自全省 11 家中医医疗机构临床数据中心 (CDR), 包含医院信息系统 (HIS)、电子病历系统 (EMR)、实验室信息管理系统 (LIS)、影像归档和通信系统 (PACS) 等, 涉及所有医疗表单。借助数据引擎以及大数据技术 [如物化视图、提取-转化-加载 (ETL) 技术等], 稳定高效且定时自动地抽取 IBD 相关诊疗数据, 实现多源异构数据的同步采集与信息集成。以 IBD 为研究对象, 制定中医优势病种核心数据集。在临床数据管理规范和数据安全管理体系框架下, 采用观察性医疗结果合作计划通用数据模型 (OMOP



注: IBD, 炎症性肠病; CDM, 通用数据模型; ETL, 提取-转化-加载; NLP, 自然语言处理; EMPI, 患者主索引; CDR, 临床数据中心; HIS, 医院信息系统; LIS, 实验室信息管理系统; RIS, 放射科信息管理系统; EMR, 电子病历系统; PACS, 影像归档和通信系统; RWD, 真实世界数据。

图 1 中西医结合 IBD 大数据平台和专病队列架构图

Figure 1 Architecture diagram of an integrated traditional Chinese and western medicine ibd big data platform and special disease cohort

CDM) 对多源数据进行标准化映射与重构, 以多模态 AI 技术 [如自然语言处理 (NLP)、机器学习算法和深度学习模型等] 为核心驱动, 结合医学人工标注, 实现 IBD 相关临床诊疗信息的结构化转换与自动化提取, 建立数据标准规范、汇集机制和深度治理体系^[20]。基于大规模可扩展分布式网络架构 (SOA), 确保数据完整性与安全性, 构建一套集数据采集、存储、治理、质控、分析与应用于一体的技术框架, 形成实时全息数据集, 即中西医结合 IBD 大数据平台。平台预留数据拓展接口, 可对接其他数据源综合管理, 实现结构化和非结构化数据的标准化存储。通过系统授权与安全认证, 数据通过应用程序编程接口 (API) 的方式供前端调用, 涵盖数据检索、随访管理、质量控制、数据可视、数据分析等多元化应用服务, 满足患者智能筛选入组、中医药知识图谱构建、中医病历文本解析、多中心联合建模、影像图片辅助诊断、疾病动态风险预测等不同需求。在此基础上, 对 IBD 患者进行全生命周期管理, 根据预先设定的研究目的, 形成 IBD 专病队列, 定期主动收集无法通过既有健康医疗数据系统获取的随访信息, 如中医体质 (王琦九种中医体质量表^[21])、中医证候积分、治疗效果 [如参照炎症性肠病治疗目标选择 II 共识 (STRIDE-II 共识)^[22]] 评价临床有效率、临床缓解率、症状消失率等]、并发症、患者生活质量、焦虑抑郁评分、舌面图谱、生物组学信息等, 描绘 IBD 患者特征画像及知识图谱, 建立中西医结合预测模型, 辅助临床诊疗。

1.2 关键技术

1.2.1 标准数据集制定 数据元标准化是搭建高质量、可交互的大数据平台的首要环节^[23]。为构建符合中国 IBD 人群特征的标准数据集, 我们借鉴了国际疾病分类第十一次修订本 (ICD-11)^[13]、HL7 临床文档架构 (CDA)^[24]、观测指标标识符逻辑命名与编码系统 (LOINC)^[25] 等多种行业标准规范, 参考了国家卫生健康委员会、国家市场监督管理总局等多部门技术标准^[14-15], 同时结合了相关 IBD 诊疗指南^[16-19]。

我们参考国际通用数据模型 (CDM), 融合中医特色诊疗优势, 创建中西医结合 IBD 标准数据集及其数据元, 并制定 IBD 值域字典。根据项目建设需求, 工作组在制定通用数据集时, 参考了观察性健康数据科学和信息学协作组开发的通用数据模型 (OMOP-CDM)^[26]、临床数据交换标准协会建立的

数据获取协调模型 (CDASH)^[27]、以患者为中心结局研究机构开发的通用数据模型 (PCORnet-CDM)^[28], 在此基础上针对 IBD 疾病特点, 再根据专病数据集个性化科研需求进行 IBD 特色诊疗数据维度拓展, 最终建立起通用数据集和特征数据集相结合、动态化、标准化的 IBD 专病数据集。数据集共分为 5 个维度, 包括 16 个一类域、49 个二类域、100 个三类域, 含 33 个模块和 620 个数据单元, 每个数据单元包含单元编码、域标签、模块标签、维度类别、中文名称、文本类型、值域、参考值、提取源、扩展字段、数据等级 11 项属性, 保证数据的标识度和溯源性。

IBD 专病数据集主要包括人口学数据集、门急诊数据集、住院数据集、诊断学数据集、医嘱与处方数据集、检验数据集、检查数据集、经济学数据集、随访数据集等, 涵盖门诊病历、住院病案首页、中医临床路径、病程记录、出院记录、检查报告、医嘱等所有医疗表单, 涉及人口信息学、主诉和现病史、既往史、家族史、生命体征、体格检查、肠道表现、内镜影像、西医诊断、中医辨证、中医诊断、中医四诊信息、实验室检查、西医药物治疗、中医治则治法、中医护理 (如中药灌肠、灸法、针灸、穴位按摩等)、用药依从性、不良事件记录等模块信息。

1.2.2 多源异构数据集成 在搭建中西医结合 IBD 大数据平台时, 首要步骤是整合 IBD 患者临床诊疗数据, 这些数据散落在各家医院内部的 CDR 系统中, 即 HIS、EMR、LIS、PACS 等业务系统。通过制定标准化的数据采集方案, 提前明确数据抽取和对接规则, 实现多源异构数据集成。研究采用 ETL 技术采集 IBD 相关诊疗数据, 配置 CDA 转换采集脚本^[29]。根据数据源的类型和结构选择不同的抽取算法, 通过 CDR 封装接口抽取数据, 统一采用 GB2312 编码, 以 XML、JSON 格式传输, 实现数据的传递、更新、同步和共享^[30]。为实现全省多中心 IBD 数据的标准化整合, 我们借助容器技术的分布式架构, 采用 Java 语言开发, 以 B/S 架构部署于江苏省健康医疗大数据中心。通过拓展项目, 结合读写分离、数据分片策略, 构建大规模并行处理 (MPP) 分布式数据库。针对图像形式的数据, 进行拷贝存储并分类管理, 建立 IBD 专病数据仓库, 数据映射后建立关联规则, 有效整合多源异构数据的对应关系^[31]。在临床数据管理规范体系下, 实现 IBD 数据集成, 独立存储于医疗健康大数据中心

内网专用服务器,通过中心交换与前置机方式实现各中心信息资源的采集、整合与共享。

1.2.3 数据治理体系建设 为实现多中心、多模式 IBD 真实世界数据的系统集成与证据生产,构建了一套规范的中西医结合数据治理体系。该体系遵循国际通用数据治理框架,并针对中西医整合数据的特殊性进行了增强设计。数据治理贯穿平台建设的整个周期,主要包括多源数据的清洗与关联、结构化数据的映射与转换、非结构化数据的后结构化 3 个维度,基于机器学习的 ETL、NLP、知识图谱等 AI 技术在数据治理过程中发挥了关键作用。

从 CDR 中抽取原始数据后,需要对其进行数据清洗,以确保符合平台建设的要求。数据清洗主要针对缺失数据、重复数据、无效数据等情况,通过特殊字符清除、时间格式清洗、整数清洗等方式,进行数据整理、填补、规约与集成。在数据清洗过程中,以 IBD 标准数据集为基准,对真实世界数据进行标准化预处理。同时基于 CDR 标准医学术语的映射集,实现人口学信息、生命体征、检验报告等结构化数据与数据元值域字典表的映射,对医学文字表述不同但含义一致的字段信息进行归一,转换成规范化语义。

对于非结构化数据和半结构化数据,制定了后结构化路径。采用基于 NLP 的机器学习模型,如监督学习、条件随机场、深度神经网络等算法,通过分析文本隐含语义和上下文结构关系构建多层次语义分析模型,探索挖掘结构化所需的知识模块,主要包括分词、同义词、主题和属性词关系图谱、上下文匹配逻辑、规则库、正则模板等核心信息,实现结构化转换^[32]。此外,基于知识图谱的医学分支逻辑,运用决策树、Logistic 回归等算法,构建分类合并的最优模型。通过信息增益的特征值对逻辑节点进行时间耦合度分析,使诊疗过程数据处于连续的时序排序状态。

针对肠镜报告、舌诊图像等以图片格式储存的信息,借助 OCR 技术结合霍夫变换直线提取法^[33]实现肠镜报告文本信息的识别与提取,采用 CNN 模型对舌象图片和肠镜影像进行特征分析与量化,根据特征对图像进行分类处理,实现图像数据的结构化采集与存储^[34]。机器学习算法提取的特征在进入数据库前,均由至少 2 名相关专业高级职称医师进行人工审核与校正;审核未通过的特征数据会返回模型重新迭代优化,同时数据库中对机器标注与人工校正的结果进行了分别标记,便于数据使用

人员区分。

以首次病程记录中中医症状文本信息抽取为例:首先,使用 R 语言导入 IBD 专病医学文本信息,进行预处理,然后随机选择 100 份样本,通过自主学习和标注构建提取规则,进行结构化抽取,统计有效性评估指标;其次,对于自动标注结果不理想即准确率较低的数据元,由工作组进行人工标注;再次,通过机器学习并建模,增加样本量迭代,进一步完善标注模型与提取规则;最后,人机结合对标注结果进行审核评估,通过不断迭代优化词典、规则、逻辑运算,直到准确率达到规定的值域范围。文本信息抽取流程示意图见附图 1 (请扫描本文二维码获取)。

1.2.4 数据质量及安全管理 在项目设计及实施过程中采取一系列控制措施来保证研究的可靠性。一是构建并推行统一的操作流程以及定期沟通机制,依据中医临床路径等规范指南对临床诊疗过程进行比对,达成关键质控指标的匹配,从源头上提升数据质量。二是开发云端实时质量控制工具,通过构建数据质量评估模型,从数据的准确性、完整性、一致性、合规性、及时性、重复性等多个维度对数据质量进行综合评估,提高数据资源的准确性与可靠性。三是组织专家对项目进行过程化监督,定期对数据内容进行抽查与质量评估。通过“制度机制-AI 初筛-专家复核”三级质控机制,确保数据完整性、一致性及术语标准化。为确保入库病例诊断的准确性,我们制定了多维度的病例质控措施:在病例入组环节,严格遵循 UC 和 CD 诊断标准^[16-19],要求入库病例必须具备规范的肠镜检查及病理组织学资料(肠镜检查由具备 5 年以上内镜操作经验且通过医院内镜诊疗技术资质认证的医师完成,操作过程规范记录病变部位、形态、范围等特征并同步采集高清影像;病理诊断采用国际公认分级标准^[35]由具备中级及以上职称的医师进行阅片),疑难病例实行 3 名及以上专家盲审。在数据整合阶段,借助大数据平台标准化质控模块,运用 NLP 技术提取肠镜、病理报告中的关键信息,与 EMR 中临床症状(如腹泻、黏液脓血便等)、实验室指标(如 C 反应蛋白、血沉等)进行交叉验证,若存在诊断与相关指标不匹配的情况,立即启动专家人工复核程序。此外,平台建立季度质控反馈机制,对各基地病例数据质量进行统计分析并生成报告,针对诊断质控问题开展专项培训与整改,持续保障入库病例诊断的准确性。三级质控机制见

附图 2 (请扫描本文二维码获取)。

数据安全始终是平台开发与利用的关键防线。我们遵循相关法律法规,明确制定了数据开放共享与安全管理责任制度,从平台对接、数据传输、数据应用等多个层面对数据资产进行安全管控。首先,IBD 大数据平台通过应用虚拟专用网络(VPN)、堡垒机、防火墙等网络安全技术和设备,利用对称加密算法(DES)保护数据的隐私及安全,将真实世界诊疗数据限制在服务器系统内,原始数据不出域,严格限定敏感数据的复制与导出功能。其次,IBD 大数据平台采用加密算法和安全传输协议进行多中心数据网络传输,定期更新动态密钥,通过区块链技术记录数据使用轨迹,记录所有加密和传输密钥版本号,最大限度防止患者隐私数据泄露。最后,IBD 大数据平台在数据应用层面采用混淆通用唯一识别码(UUID)等可逆性/不可逆机制,将姓名、家庭住址等可溯源的患者信息进行隐私化处理;通过角色与权限控制(RBAC)机制,采取商密 4 号分组密码(SM4)及李维斯特、萨莫尔、阿德曼公钥(RSA)等加密算法进行平台授权登录,实现数据集中管理和分级管理。通过数据脱敏、加密和授权管理,全方位确保数据安全和隐私保护。

1.2.5 IBD 专病队列建立 依托大数据平台,整合多中心、多维度、动态的 IBD 真实世界诊疗数据,构建 IBD 专病队列,为探索 IBD 发生发展规律、解析患者全生命周期的诊疗情况提供重要支撑。采用多中心、双向性的队列构建模式,基于 OMOP-CDM,根据严格的纳入、排除标准,通过 NLP 算法自动筛查 ICD-11 编码(K50.x、K51.x)及临床诊断关键词,自动生成 IBD 研究队列和数据集。通过宏观与微观相结合的研究方法,全面收集了 IBD 患者的流行病学调查、生命体征、体格检查、临床生化检测、肠镜报告数据、生物标本、中医四诊信息、辨证分型、中医体质、中药复方、疗效指标、药物不良反应、生存质量等纵向数据。以建立高质量中西医结合 IBD 专病队列为目标,拟通过长期随访研究,系统采集 IBD 生命健康大数据,为辨体-辨病-辨证诊疗模式提供参考;开展 IBD 中医药疗效评价真实世界研究,构建中西医结合关联规则与 IBD 疾病(诊断、治疗)预测模型,产出针对 IBD 的中医诊疗规范、技术标准、临床路径和防控策略。

2 结果

2.1 高质量多模态 IBD 数据集

针对中医优势病种 IBD 疾病特征,根据平台建设实际需求,完成了通用数据维度、专科数据维度、专病数据维度的数据字典梳理和定义,建立起通用数据集和特征数据集相结合、动态化、标准化的 IBD 专病数据集。IBD 专病数据集主要包括人口学数据集、门急诊数据集、住院数据集、诊断学数据集、医嘱与处方数据集、检验数据集、检查数据集、经济学数据集、随访数据集等,涉及所有医疗表单。借助 ETL 等技术,构建了一套符合中国 IBD 人群特征的、规范的、可互通、可推广的标准数据集,为规范中西医临床诊疗数据采集、推进多中心数据交换共享、挖掘真实世界数据价值提供重要支撑。详见附图 3 (请扫描本文二维码获取)。

2.2 中西医结合 IBD 大数据平台

搭建了中西医结合 IBD 大数据平台,整合了全省 11 家中医医疗机构 IBD 临床诊疗真实世界数据,实现了区域内数据集成与共享。截至 2025 年 8 月 31 日,平台以 IBD 患者为研究对象,完成了 IBD 全量诊疗数据(包括 33 个维度、620 个指标)的抽取和清洗,专病结构化抽取总体有效性评估。根据 IBD 在 CDR 就诊记录中的时间先后进行排序,累计汇集了 2023 年 1 月至今就诊的 36 890 人次 IBD 患者数据,研究病例共 14 424 例,其中门诊 13 008 例、住院 1416 例。此外,平台融合了中医特色模块,形成了标准化的 IBD 专病数据模型(包括通用数据模型和 IBD 特征数据模型),建立了常态化病例入库管理机制,相关信息在不断更新,必将成为区域性专病大数据平台建设的新范式。平台基本情况见附图 4 (请扫描本文二维码获取)。

2.3 IBD 专病队列

在 IBD 大数据平台建设基础上,通过对 IBD 患者全生命周期管理,形成了真实世界数据与主动采集数据相结合的 IBD 专病队列,包括 UC 专病队列和 CD 专病队列。根据预先设定的研究目的,设定纳入和排除标准,以 2020 年 1 月 1 日至 2024 年 12 月 31 日记录的全部 IBD 患者为研究对象,建立了回顾性队列,纳入 1781 例 IBD 患者(其中门诊患者 1496 例、住院患者 285 例);2025 年 1 月 20 日启动前瞻性队列建设,已入组 IBD 患者 324 例,其中 UC 患者 285 例,CD 患者 39 例。通过定期主动收集无法通过既有医疗数据系统获取的随访信息,如中

医体质、中医证候积分、治疗效果、并发症、患者生活质量、焦虑抑郁评分、舌面图谱、生物组学信息等，建立了长期持续、安全规范的数据资产积累模式，可为今后开展多中心中医药循证研究提供高质量数据支持。

2.4 中医药循证研究和成果转化

IBD 医疗大数据平台的建立，为疾病病因、诊断、治疗和预后研究提供了科学依据。基于 IBD 大数据平台和专病队列开展大规模多中心真实世界研究，分析不同治疗方案的临床价值和卫生经济学，可帮助医保决策和优化中医优势病种诊疗方案/临床路径。基于 IBD 大数据平台和专病队列设计了系列中医药循证医学研究方案，研究方向覆盖 IBD 的疾病特征、发病机制、危险因素、共病分析、中医诊疗方案、治疗效果评估、成本效益疾病诊断相关分组（DRG）分析、疾病预后预测模型等多个维度，“讲清楚”中医药临床疗效、“说明白”中医药基础机制。此外，采用全卷积神经网络（FCN）分析技术，对肠镜影像、舌象图片等医学图像进行分割、识别和建模，进而判断 IBD 进展，进行疾病的分类诊断；利用数据挖掘等技术，开展医疗机构制剂或名老中医经验方人用经验数据资料收集，生产高质量真实世界证据，推动 IBD 中药新药研发与成果转化。

3 讨论

随着对 IBD 认识的深入，IBD 的诊疗模式已从传统临床治疗向长期随访管理转变，需要更多基于中国 IBD 人群的流行病学和真实世界数据，以支持构建更精准的防治策略和更规范的诊疗体系。通过“互联网+医疗”以及 AI 技术的创新性交叉融合，基于多模态 IBD 真实世界诊疗数据，构建区域性、标准化中西医结合大数据平台和专病队列成为可能。

我们在遵循国际通用 CDM 基础上，融合中医特色元素，探索建立了通用数据集和特征数据集相结合、标准化的 IBD 专病数据模型，制定 IBD 标准数据集；借助 AI 技术完成全省 11 家中医院 IBD 诊疗数据的对接和抽取，实现多源异构真实世界数据系统集成；利用机器学习算法进行数据清洗和结构化处理，初步形成一套规范的中西医结合数据治理、质控与安全体系，成功开发中西医结合 IBD 大数据平台；通过对 IBD 患者进行长期动态随访、纵向观察，建立起中医特色 IBD 专病队列。

我们构建的专病队列与平台突破了传统单中

心、单维度队列的局限，实现了全省多中心、中西医多模态数据的整合与治理。一方面，这不仅为 IBD 临床研究提供了高质量的真实世界数据资源，也为探索疾病异质性、优化个体化治疗提供了可能。另一方面，平台将中医证候、舌象、脉诊、中药复方等特征数据纳入标准化数据模型，并建立“病-证-症-药”知识图谱，凸显中医药在 IBD 诊疗中的特色优势，实现了中西医数据的深度融合。此外，平台支持多中心隐私保护下的联邦学习，为多中心联合科研攻关提供了技术保障。

大数据平台和队列建设是一项系统工程，我们探索了“政府主导-医院联动-多学科交叉”的有组织科研模式，有效整合了临床医学、流行病学统计、信息科学等多学科团队；提出了“西医标准数据集+中医特色扩展”的 IBD 数据模型和数据治理框架，创新性地融入中医四诊信息、中医证候术语集（TCMIO）和舌象影像分析标准，构建了中西医结合全生命周期 IBD 管理队列，解决了中西医数据孤岛的难题；率先将多模态 AI 技术系统性应用于 IBD 专病队列建设，如利用 NLP [如基于预训练的医疗自然语言处理模型（BERT-MED）模型] 从非结构化文本中提取中医证候信息，应用 CNN 对舌象、内镜图像进行特征提取与量化分析。IBD 大数据平台与专病队列仍存在一定局限：1) 数据层面。本研究的核心数据来源于省内中医院，未覆盖全国大型中医院，可能存在选择偏移和信息偏倚。2) 技术层面。多模态 AI 模型在复杂中西医数据治理、关联推理、时空动态预测方面的性能有待提升，可能需要引入更先进的时序模型来提升对 IBD 疾病动态演变的预测能力。3) 应用层面。当前中西医结合 IBD 队列仍以住院患者为主，需加强前瞻性长期随访，纳入患者报告结局（PRO）等数据，并探索临床信息与多组学数据的整合，以“讲清楚”“说明白”IBD 病因病机及中医药作用机制。未来将持续深入优化，并致力于推动跨区域的数据共享与合作研究。通过中西医结合诊疗与 AI 技术的深度融合，必将为 IBD 的精准防治以及中药新药研发注入全新动力，并为其他慢性病中西医结合专病平台队列建设提供可借鉴的方案。

参考文献

- [1] DART R J, ZLATAREVA I, VANTOUROUT P, et al. Conserved $\gamma\delta$ T cell selection by BTNL proteins limits progression of human inflammatory bowel disease [J].

- Science, 2023, 381(6663): eadh0301.
- [2] 叶放, 周学平, 程海波, 等. 基于整体观理解和创新中医病机辨证新理论[J]. 南京中医药大学学报, 2017, 33(5): 436-439.
 - [3] 陈赛娟. 发展重大疾病专病数据库意义重大[J]. 科技创新与品牌, 2022(3): 20.
 - [4] IACUCCI M, MAEDA Y, SANTACROCE G, et al. AI-enabled 'endo-histo-omics': breaking down intestinal barriers in IBD[J]. Nat Rev Gastroenterol Hepatol, 2025, 22(6): 362-363.
 - [5] 吴敏敏, 王鑫钰, 王伟炳. 医学人工智能的技术发展和场景应用[J]. 复旦学报(医学版), 2025, 52(3): 470-474.
 - [6] HOLTEN K I A, BERNKLEV T, OPHEIM R, et al. Fatigue in patients with newly diagnosed inflammatory bowel disease: results from a prospective inception cohort, the IBSEN III study[J]. J Crohns Colitis, 2023, 17(11): 1781-1790.
 - [7] RAFFALS L E, SAHA S, BEWTRA M, et al. The development and initial findings of a study of a prospective adult research cohort with inflammatory bowel disease (SPARC IBD) [J]. Inflamm Bowel Dis, 2022, 28(2): 192-199.
 - [8] DARBANI B, BRODERSEN T, LILJENSE A, et al. Cohort profile: Copenhagen hospital biobank-chronic inflammatory disease-inflammatory bowel disease (CHB-CID: IBD) genetic cohort[J]. Eur J Epidemiol, 2025, 40(6): 721-734.
 - [9] 冯雨萱, 郑锐, 于广军. 我国临床专病数据库建设情况概述[J]. 上海医药, 2024, 45(9): 10-13, 18.
 - [10] 吴晓萍, 郑涛, 房静远, 等. 炎症性肠病队列数据库的建设和应用[J]. 胃肠病学, 2023, 28(4): 237-242.
 - [11] REN Y, JIA Y L, LIANG W X, et al. Best practice for developing integrative Chinese-western medicine databases using electronic health records[J]. World J Integr Tradit West Med, 2025, 11(3): 174-186.
 - [12] VON ELM E, ALTMAN D G, EGGER M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies[J]. Ann Intern Med, 2007, 147(8): 573-577.
 - [13] HARRISON J E, WEBER S, JAKOB R, et al. ICD-11: an international classification of diseases for the twenty-first century[J]. BMC Med Inform Decis Mak, 2021, 21(Suppl 6): 206.
 - [14] 中华人民共和国国家卫生健康委员会. 卫生健康信息数据集元数据标准[EB/OL]. (2023-04-01)[2025-11-20]. <https://www.nhc.gov.cn/cms-search/downloadFiles/b4e61d83c6014b04a556810f04fce3a8.pdf>.
 - [15] 国家市场监督管理总局, 中国国家标准化管理委员会. 信息安全技术 健康医疗数据安全指南[EB/OL]. (2020-12-14)[2025-11-20]. <https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=239351905E7B62A7DF537856738247CE>.
 - [16] 中华医学会消化病学分会炎症性肠病学组, 中国炎症性肠病诊疗质量控制评估中心. 中国溃疡性结肠炎诊治指南(2023年·西安)[J]. 胃肠病学, 2024, 29(3): 145-173.
 - [17] 中华医学会消化病学分会炎症性肠病学组, 中国炎症性肠病诊疗质量控制评估中心. 中国克罗恩病诊治指南(2023年·广州)[J]. 胃肠病学, 2024, 29(4): 200-235.
 - [18] 中华中医药学会脾胃病分会. 溃疡性结肠炎中医诊疗专家共识(2023)[J]. 中华中医药杂志, 2024, 39(1): 288-296.
 - [19] 中华中医药学会脾胃病分会. 克罗恩病中医诊疗专家共识意见(2024)[J]. 中医杂志, 2024, 65(16): 1734-1740.
 - [20] 章萌, 沈鹏, 刘志科, 等. 观测性医疗结果合作组织通用数据模型在中国的应用研究现状[J]. 中华流行病学杂志, 2025, 46(5): 907-913.
 - [21] 朱燕波. 《中医体质量表》应用中的问题及其使用规范[J]. 中华中医药杂志, 2022, 37(9): 5066-5070.
 - [22] TURNER D, RICCIUTO A, LEWIS A, et al. STRIDE-II: an update on the selecting therapeutic targets in inflammatory bowel disease (STRIDE) initiative of the international organization for the study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD [J]. Gastroenterology, 2021, 160(5): 1570-1583.
 - [23] 董志威, 王钧岚, 谢韬, 等. 基于优劣解距离算法联合熵值法的针刺治疗前列腺癌术后尿失禁的循证决策研究[J]. 中医杂志, 2024, 65(23): 2434-2441.
 - [24] 国家市场监督管理总局, 国家标准化管理委员会. 健康信息学 数据交换标准 HL7 临床文档架构(版本 2) [EB/OL]. (2023-03-17)[2025-12-20]. <https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=A358014699D2D48295C9C529FBB70F3A>.
 - [25] 张林, 赵英杰, 陈兴. 观测指标标识符逻辑命名与编码系统(LOINC)数据库介绍[J]. 河北省科学院学报, 2004(4): 66-68, 71.
 - [26] AHMADI N, PENG Y, WOLFIEN M, et al. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review[J]. Int J Mol Sci, 2022, 23(19): 11834.

- [27]CHENG A C, FACILE R, OWEN J, et al. Making clinical data acquisition standards harmonization (CDASH) electronic case report forms available on the REDCap shared data instrument library [J]. J Soc Clin Data Manag, 2022,2(3):7.
- [28]GARZA M, DEL FIOLO G, TENENBAUM J, et al. Evaluating common data models for use with a longitudinal community registry[J]. J Biomed Inform, 2016,64(12): 333-341.
- [29]刘海,张瞩熹,任雯,等. 面向异构数据源的分布式集成工具研究与设计[J]. 计算机应用研究, 2020,37(S1):204-206.
- [30]王传立,李金才,吴移桃. Struts 与 Web Service 集成模式研究与应用[J]. 计算机工程与科学, 2011, 33(8): 181-184.
- [31]姜维宁,李晓明,栾兵,等. 基于提高临床科室管理能力的医院信息集成平台的建立及应用效果[J]. 中国卫生统计, 2023, 40(6):869-871,874.
- [32]STANLEY J, RABOT E, REDDY S, et al. Large language models deconstruct the clinical intuition behind diagnosing autism[J]. Cell, 2025,188(8):2235-2248.
- [33]张炫,杨艳. 基于改进随机霍夫变换的直线提取算法[J]. 信息技术,2017(12):6-9.
- [34]GAO C, ZHU Q, WANG P, et al. Structured multimodal attentions for TextVQA [J]. IEEE Trans Pattern Anal Mach Intell,2022,44(12):9603-9614.
- [35]MATSUMOTO T, HISAMATSU T, ESAKI M, et al. Guidelines for endoscopic diagnosis and treatment of inflammatory bowel diseases [J]. Dig Endosc, 2025, 37(4):319-351.

Construction of a Multimodal AI-Based Integrated Traditional Chinese and Western Medicine Big Data Platform for Inflammatory Bowel Disease with Specialized Disease Cohort

SHEN Zhaofeng, ZHANG Mengyuan, ZHU Lei, SHEN Hong

Jiangsu Province Hospital of Chinese Medicine/Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, 210029

ABSTRACT Objective To develop an integrated traditional Chinese and western medicine big data platform with specialized disease cohort based on multimodal real-world data (RWD), providing methodological support and data foundation for high-quality evidence-based traditional Chinese medicine (TCM) research. **Methods** Following the domestic and international construction models for specialized cohort database, a core dataset for TCM-dominant disease, that is inflammatory bowel disease (IBD), was developed relying on the Jiangsu TCM Dominant Diseases Diagnosis & Treatment Data Platform, which integrated multi-source heterogeneous data from 11 tertiary grade-A TCM hospitals in Jiangsu province, and an integrated traditional Chinese and western medicine big data platform for IBD was constructed thereafter. Leveraging artificial intelligence technologies including natural language processing (NLP) and machine learning, RWD is deeply governed through extraction, transformation, desensitization, encryption, cleaning, mapping, and manual annotation to achieve the standardization, structuring, and normalization of patient-specific data. A dynamic management system was developed for patient screening, data collection, and follow-up to form an IBD-specific cohort. Furthermore, optical character recognition (OCR) combined with the Hough transform line extraction method was used to recognize and extract texts from colonoscopy reports, and convolutional neural network (CNN) models were employed for feature analysis and quantification of tongue images and colonoscopy photos. **Results** A preliminary standard for the integrated traditional Chinese and western medicine IBD-specific disease metadata was developed, which processed 36,890 complete IBD diagnosis and treatment records, achieving standardized mapping, integration and quality control of TCM and western medicine data. A multi-modal IBD big data platform was established with 14,424 cases, and a specialized disease cohort was constructed including the full lifecycle management of 1781 IBD patients. Additionally, an AI-assisted TCM prediction model has been developed. **Conclusion** The research has established a replicable technical framework for building a multimodal integrated traditional Chinese and western medicine big data platform and disease cohort for IBD, which achieves the deep integration of RWD from both TCM and western medicine, highlighting TCM's unique advantages, and providing high-quality data support for future evidence-based studies in TCM.

Keywords inflammatory bowel disease; medical big data; big data platform; specialized disease cohort; artificial intelligence

(收稿日期: 2025-09-23; 录用日期: 2025-12-18; 修回日期: 2025-12-26)

[编辑: 叶晨]