

DOI: 10. 13288/j. 11-2166/r. 2026. 08. 015

文献研究

基于大语言模型的中医黄疸病古籍语料库的构建

张艺然, 李海燕[✉], 聂莹

中国中医科学院, 北京市东城区东直门内南小街 16 号, 100700



[摘要] **目的** 基于大语言模型构建结构化中医黄疸病古籍语料库, 为现代黄疸病临床与科研提供数据支撑, 也为中医古籍知识的深度挖掘与利用提供思路。**方法** 系统整合互联网开源数字化中医古籍资源, 采用文献计量学方法构建黄疸病名术语库, 利用 Python 正则表达式进行关键词匹配和语料抽取, 设计“病-证-症-治-效”五元信息抽取模型, 制定标注指南并采用双人背对背标注结合第三人仲裁形成标准测试集, 为筛选适配实体信息抽取的最优模型, 采用 DeepSeek-R1 与 ChatGPT-o3 2 个大语言模型进行对比性能评估, 通过调用大模型应用程序编程接口 (API) 进行实体信息抽取。**结果** 构建了涵盖从先秦至清代 561 本古籍、10 243 条语料的中医黄疸病古籍语料库。模型对比评估结果显示, DeepSeek-R1 在 17 类实体识别的性能全面优于 ChatGPT-o3, 并基于该优选模型从语料库中抽取 41 407 个实体, 涵盖病名 (6238 个)、证型 (1622 个)、方剂 (2631 个)、中药 (2706 个) 等 17 个实体类型。**结论** 基于 DeepSeek-R1 大语言模型构建的语料库覆盖了黄疸病核心辨证论治要素, 可为后续中医黄疸病历史演变和辨证论治规律研究奠定基础, 并为 AI 驱动的中医古籍知识挖掘提供可复制的技术范式。

[关键词] 中医古籍; 黄疸病; 语料库; 大语言模型; 信息抽取

中医古籍是中医学术传承数千年的重要知识载体, 也是中医临床诊疗的宝贵资源。黄疸病是中医临床优势病种, 两千多年来中医学对黄疸病病因病机、证治方药的认识不断深化并逐渐形成了独特的理论体系。古代医籍对黄疸病有丰富记载, 如汉代张仲景《金匮要略》有“黄疸病脉证并治”专篇, 将黄疸分为五类, 提出“阳黄”“阴黄”分型, 首次对黄疸进行系统辨证论治, 确立了清热利湿、温化寒湿等治则, 为黄疸辨证施治奠定了基础^[1]。然而相关论述分散在古籍中, 存在缺少整理与归纳、数据标准化不足、研究方法局限以及知识整合困难等问题^[2-5]。

随着人工智能技术的蓬勃发展, 通过现代信息技术和数据挖掘等手段, 从海量的古籍文献资料中获取可理解的、可用的知识, 成为中医古籍文献研究的前沿方向^[6]。大语言模型 (large language

models, LLMs) 是参数规模达数十亿至万亿级的神经网络模型, 拥有强大的通用建模能力和泛化能力, 在自然语言处理任务中表现出色, 所展现出的智能化逻辑推理能力为中医古籍知识挖掘注入了新的动力^[7]。本研究基于不同大语言模型构建中医黄疸病古籍语料库, 实现对中医古籍文本内容的信息抽取, 将中医古籍中蕴含的关于黄疸病证的大量未系统化的隐性知识显性化, 形成结构化、数字化表达的黄疸病古籍知识库, 并比较不同大语言模型构建古籍语料知识库的效能, 有助于挖掘中医古籍蕴藏的治疗黄疸病的大量临证经验与用药规律, 也为中医古籍知识的深度挖掘与利用提供思路。

1 中医黄疸病古籍语料库的构建

1.1 语料来源

为确保语料的全面性与权威性, 本研究广泛整合了互联网开源且学术认可度较高的数字化中医古籍资源, 语料主要来源于以下三个数据库: 1) 古今医案云平台 (<https://www.yiankb.com>) 收录的专病医案中黄疸病相关医案; 2) 博览医书数据

基金项目: 中国中医科学院科技创新工程 (CI2021B002); 江苏省前沿技术研发计划 (BF2025076); 中国中医科学院基本科研业务费 (ZZ170307)

✉ 通讯作者: li_haiyan01@126.com

库 (<https://www.imedbooks.com/>) 专科古籍库中收录的黄疸病相关的古籍; 3) 殆知阁·中国古典文献全文检索平台 (<https://daizhige.org/>) 收录的医藏数据库的中医古籍。通过对上述三个来源的数据检索与下载, 初步获得了 974 本以 txt 格式存储的中医古籍原始文本, 构成了原始中医古籍文本库。

1.2 古籍筛选标准

1) 纳入成书于 1912 年之前的中医典籍, 排除民国及以后出版的书籍文献; 2) 纳入中国本土医家著作, 排除国外典籍, 如日本丹波元坚《杂病广要》等汉方著作; 3) 纳入内容涵盖医案、医话、医论、医经、方书、本草、临证、内科、针灸推拿、诊法、养生、综合著作等类型著作; 4) 不同版本的中医古籍仅纳入一个版本, 在多个版本并存时优先选取学术界公认的权威版本或最早刊行版本作为核心语料来源; 5) 纳入与黄疸病相关论述。

1.3 语料抽取

中医古籍中黄疸病病名术语描述繁杂多样, 为确保获取语料的全面性, 本研究采用文献计量学与文本挖掘相结合的方法建立了黄疸病名术语库, 并基于术语库采用 Python 正则表达式完成了中医古籍文本库中黄疸病描述相关语料信息抽取。

1) 设计文献检索策略。以黄疸病名 (如黄疸、黄病、黄瘰等) + 医学史研究关键词 (源流、变迁、文献、古籍、理论、认识、思考、记载、沿革、发展、流变、嬗变、演变、考辨等) 为主题词进行检索, 对中国知网 (CNKI)、万方数据库、维普网、读秀学术等中文学术数据库进行系统检索, 检索时间截至 2025 年 3 月 4 日, 检索到 185 篇相关文献。

2) 设计文献筛选标准。纳入专门探讨黄疸病历史源流、病名演变、理论发展的研究性文献; 排除现代临床治疗、药理实验等未涉及历史考证的文献; 排除仅提及病名而无实质性考辨内容的文献; 排除会议摘要、科普文章等非正式学术文献。逐篇筛选后最终纳入符合条件的文献 151 篇 (其中期刊论文 107 篇, 博士、硕士学位论文 44 篇)。

3) 构建黄疸病名术语库。通过遍历检索获取的文献全文, 定位并提取文献中记载的所有黄疸病病名, 共收集到 98 个不同的黄疸病名, 详见表 1。该词库涵盖了从“黄疸”“瘰病”等总称, 到“谷疸”“酒疸”“女劳疸”等具体分类, 乃至“急黄”“瘟黄”“癖黄”等特殊或罕见称谓。

表 1 基于文献检索的黄疸病名术语库

Table 1 Terminology database of jaundice names based on literature retrieval

序号	病名	序号	病名	序号	病名	序号	病名
1	白疸	26	阳黄酒疸	51	湿疸候	76	癖黄
2	疸	27	胃疸	52	肝黄	77	火黄
3	目黄	28	风黄疸	53	心黄	78	水黄
4	发黄	29	黄病候	54	脾黄	79	人黄
5	胎疸	30	急黄候	55	肺黄	80	癰黄
6	风疸	31	黄汗候	56	肾黄	81	痢黄
7	赤疸	32	犯黄候	57	胆黄	82	白黄
8	马黄	33	劳黄候	58	脑黄	83	走精黄
9	五疸	34	脑黄候	59	肠黄	84	灰黄
10	三十六黄	35	阴黄候	60	体黄	85	土黄
11	黄病	36	内黄候	61	胃黄	86	房黄
12	脾疸	37	行黄候	62	髓黄	87	劳黄
13	久黄	38	癖黄候	63	爪黄	88	食黄
14	积黄	39	噤黄候	64	鬼黄	89	忧黄
15	热疸	40	五色黄候	65	奸黄	90	酒黄
16	虚劳疸	41	风黄候	66	立黄	91	酒疸
17	湿疸	42	因黄发血候	67	黑黄	92	风黄
18	谷疸	43	黄疸候	68	脊禁黄	93	女劳疸
19	劳食疸	44	酒疸候	69	气黄	94	走马黄
20	瘟黄	45	谷疸候	70	瘟黄	95	急黄
21	阴黄	46	女劳疸候	71	血黄	96	瘰
22	黄肿	47	黑疸候	72	惊黄	97	黄瘰
23	正黄疸	48	九疸候	73	花黄	98	瘰病
24	血疸	49	胞疸候	74	疰黄		
25	黄疸沙	50	风黄疸候	75	行黄		

4) 抽取黄疸病描述相关语料。基于构建的黄疸病名术语库, 采用 Python 正则表达式对获取的 974 本中医黄疸病古籍文本库进行关键词匹配与描述段落抽取, 在每本古籍文本中搜索并定位任何包含黄疸病名术语库中任意一个病名的文本段落, 为提高准确性, 将匹配窗口扩展至包含该病名关键词的上下文段落, 以确保抽取出的语料具有相对完整的语义单元。匹配后自动记录并结构化存储书名、卷名、朝代、作者、包含病名关键词的原文描述段落等元数据信息。所有抽取结果被组织并存储在结构化的表格中, 形成了中医黄疸病古籍基本数据集, 见图 1。

1.4 语料清洗

对生成的中医黄疸病古籍基本数据集语料数据进行人工校对与清洗处理, 见附图 1 (请扫描本文二维码获取)。1) 剔除多源数据整合导致的重复版本古籍 49 本, 最终用于语料抽取的古籍共 925 本, 基于黄疸病名术语库抽取到与黄疸病描述相关的语

料 13 980 条, 涉及古籍 683 本。2) 修正重复与异常数据。删除完全重复的黄疸病古籍描述语料共 816 条, 确保语料的唯一性; 剔除不符合古籍纳入标准的语料共 796 条 (包括 1912 年后语料 291 条、年代不详语料 100 条、域外典籍语料 405 条); 删除非黄疸病描述的语料共 2125 条。最终纳入语料 10 243 条, 古籍 561 本。3) 处理数据缺失值。人工查询与补充校对朝代、作者、卷名等信息缺失或采集错误的记录, 确保数据信息的准确性和完整性。4) 数据一致性维护。对朝代、古籍名等信息进行规范化处理, 如“清”“清朝”“清代”统一规范为“清”, 将“黄帝内经素问”“黄帝内经·素问”统一规范为“黄帝内经·素问”等。

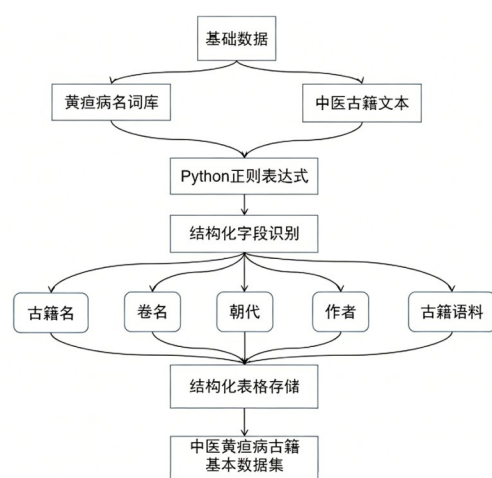


图1 基于大语言模型的中医黄疸病古籍语料库构建的语料抽取流程图

Figure 1 Flowchart of corpus extraction for constructing a traditional Chinese medicine jaundice-related ancient text corpus based on large language models

1.5 构建结果

经过数据清洗流程, 最终得到中医黄疸病历时古籍语料库包括从先秦两汉到清代 7 个朝代的语料数据, 涵盖古籍 561 本, 语料 10 243 条, 每条语料包含古籍名、卷名、朝代、作者和古籍语料。其中先秦两汉纳入古籍 9 本, 语料 172 条, 包括《伤寒杂病论》《金匮要略》《黄帝内经》等; 魏晋南北朝时期古籍 9 本, 语料 110 条, 包括《肘后备急方》《针灸甲乙经》等; 隋唐时期古籍 14 本, 语料 406 条, 包括《备急千金要方》《外台秘要》《诸病源候论》等; 宋代古籍 45 本, 语料 1135 条, 包括《圣济总录》《太平圣惠方》《经史证类备急本草》等; 金元时期古籍 43 本, 语料 353 条, 包括《世医得效方》《儒门事亲》《卫生宝鉴》等; 明代古籍 107

本, 语料 2978 条, 包括《普济方》《古今医统大全》《医学入门》等; 清代古籍 334 本, 语料 5089 条, 包括《古今图书集成医部全录》《医宗金鉴》《伤寒论纲目》等。

2 中医黄疸病古籍语料库信息抽取模型的构建

2.1 确定领域和范围

为实现从非结构化的古籍文本到结构化知识的转化, 需要构建一个覆盖黄疸病辨证论治核心要素的古籍知识本体模型, 以支撑后续的结构化信息抽取。本体的领域明确限定为中医古籍黄疸病辨证论治知识, 以构建的语料库 (561 本古籍, 10 243 条语料) 所涵盖的知识内容为本体模型的范围。

2.2 复用领域本体

为保障术语的规范性与知识的可交互性, 本研究充分复用并参考已有的标准与领域本体。1) 参考《中医病证分类代码》^[8]、《中医临床诊疗术语》^[9] 对疾病、证型、症状等核心概念术语进行界定。2) 借鉴《中医内科学》^[10]、《中医诊断学》^[11]、《方剂学》^[12]、《中医基础理论》^[13] 等教材确保概念的分类符合中医临床共识。3) 复用中医药语言系统 (TCMLS) 和中医临床术语系统 (TCMCTS) 领域本体的“中药”“方剂”“病机”“治则”等实体及其属性。4) 参考《中医药文本数据命名实体信息分类与代码》^[14]、《中医药文本数据关系信息分类与代码》^[15] 中对“证症关系”“治效关系”等实体关系分类及定义, 指导本研究实体关系体系设计。

鉴于古今术语存在差异, 本研究在复用上述标准时, 坚持以古籍原文语义为核心, 现代标准主要作为参考框架和辅助映射工具, 以确保最终生成的结构化数据既忠实于古籍原意, 又具备良好的规范性。

2.3 确定层级关系

基于中医黄疸病古籍语料库, 构建“病-证-症-治-效”为核心的五元关系模型, 定义本体的核心类及其层级体系。实体类别包括 5 个一级实体分类和 17 个二级实体分类, 覆盖中医古籍中黄疸病的病名、证候、症状、治则治法及疗效等辨证论治要素, 其中“病”类实体是指黄疸病的病名, 包括“病名”1 个二级实体类别; “证”类实体是指黄疸病的证候, 包括“证型”“病因”“病机”“病位”“病性”“病势”6 个二级实体类别; “症”类实体是指黄疸病的症状, 包括“症状”“舌象”“脉象”3 个二级实体类别; “治”类实体是指黄疸病的治则治法和方剂中药等, 包括“治则”“治法”“方剂”“中药”“经络”“腧穴”6 个二级实体类

别；“效”类实体是指治疗黄疸病的疗效，包括“疗效”1个实体类别。

2.4 定义属性与关系

对象属性用于定义不同类之间的交互关系，本研究基于中医古籍文本中实体所蕴含的内在逻辑关联，定义了5个一级实体关系和15个二级实体关系，各级实体关系名称及定义详见表2。

表2 基于中医古籍语料库定义实体关系类别及定义

Table 2 Definition of entity relationship categories and definitions based on the TCM classics corpus		
一级实体关系	二级实体关系	定义
病证关系	病证关系	指每个中医疾病有其相关的证型
	病因关系	指引发疾病的病因
证症关系	证症关系	指每个证候的疾病对应的症状
	症因关系	指引起每个症状的病因病机
证治关系	证治关系	指每个证型的治则治法
	证方关系	指每个证型使用的方剂
	证药关系	指每个证型使用的中药
症治关系	症治关系	指不同症状、舌象、脉象采用的治则治法
	症方关系	指不同症状、舌象、脉象使用的方剂
	症药关系	指不同症状、舌象、脉象使用的中药
治效关系	方治关系	指不同治则治法对应的方剂
	药治关系	指不同治则治法使用的中药
	方药关系	指不同方剂对应的中药
	方效关系	指不同方剂对应的疗效
	药效关系	指不同中药对应的疗效

2.5 创建实例

基于上述本体模型，从语料库中抽取具体知识创建实例，详见表3。

3 基于大语言模型的中医黄疸病古籍信息的抽取研究

3.1 设计标注指南

为确保抽取实体信息的准确性和一致性，设计

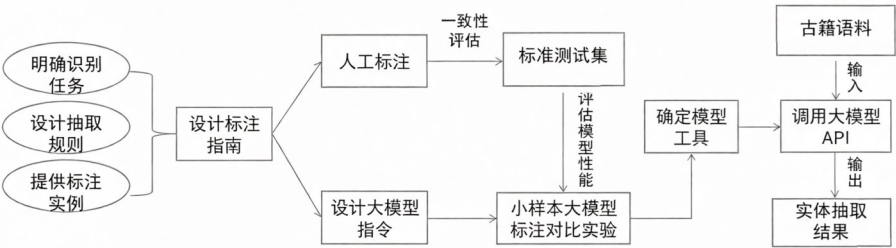
表3 基于中医古籍语料库创建实体实例及来源
Table 3 Entity instances and the sources based on the TCM classics corpus

实体类型	实体实例	来源
病名	[谷疸,女劳疸,酒疸]	《证治准绳》
证型	[胆虚气逆]	《黄帝内经·素问》
病因	[伏邪]	《温疫论》
病机	[肠中风火燥结]	《儒门事亲》
病位	[脾]	《圣济总录》
病性	[寒,热,虚]	《古今图书集成医部全录》
病势	[久久不愈]	《针灸甲乙经》
症状	[身重目黄,心下支满,口干燥渴]	《千金翼方》
舌象	[舌卷焦黄黑]	《伤寒总病论》
脉象	[脉浮,沉弦]	《诸病源候论》
治则	[补中益气]	《神农本草经》
治法	[外敷,点眼,搐鼻]	《证治准绳》
方剂	[茵陈五苓散]	《金匱要略》
中药	[白术,茯苓,泽泻,猪苓,桂枝,茵陈]	《医宗金鉴》
经络	[手阳明大肠经]	《黄帝内经·灵枢》
腧穴	[神门穴]	《子午流注针经》
疗效	[便痊,神效]	《女科要旨》

注：实体实例以语料库实例形式呈现。

详细的中医黄疸病古籍信息抽取的“标注指南”，用于指导人工标注形成标准测试集，评估不同大语言模型在本研究信息抽取性能，见图2。

1) 明确识别任务：明确病名、证型等17个实体类别的定义，区分易混淆概念。2) 设计抽取规则：(1) 唯一性规则，要求每个实体仅归属最匹配类别；(2) 语义聚焦规则，仅标注黄疸相关实体（如文本含“咳嗽”“风寒”等无关黄疸病则不标注），避免因古籍文本语料的其他疾病干扰识别结果；(3) 颗粒度控制规则，语料拆分为最小语义单元（如“脉弦数”标注为[“弦”，“数”]），同



注：API，应用程序编程接口。

图2 基于大语言模型的中医黄疸病古籍语料库构建的信息抽取流程图

Figure 2 Flowchart of information extraction for building a traditional Chinese medicine jaundice-related ancient text corpus based on large language models

时要求结合上下文进行语义理解以保留语料完整性；(4) 遵循古籍原文规则，在指令设计中强调抽取的实体信息须完全遵循并保留古籍原文原始语料表述，而非依赖现代术语标准进行反推。3) 提供标注实例：进行标注员培训或大模型小样本学习，减少主观理解偏差。

3.2 人工标注与一致性评估

为评估大模型性能提供标准测试集，本研究首先进行了小样本高质量的人工标注。

1) 标注过程。从语料库中随机抽取 100 条语料，选取两位中医学专业博士研究生作为标注员，在接受统一培训后，依据“标注指南”独立进行“背靠背”标注。2) 仲裁与修正。当两位标注员结果不一致时，邀请一位具有副主任医师职称的中医临床专家作为仲裁者，对标注员标注不一致情况进行仲裁，分析主要争议点并确定解决方式，及时修正标注偏差，最终形成标准测试集。3) 一致性评估。采用 Cohen’s Kappa 系数评估标注者间一致性^[16]，计算公式为： $\kappa = \frac{1 - P_e}{P_o - P_e}$ （ P_o 为观察一致率， P_e 为随机一致率）。如表 4 所示，17 类实体的 Kappa 值介于 0.87 到 1.00 之间，平均值为 0.94。根据 Landis & Koch 标准（ $\kappa > 0.81$ 表示“几乎完全一致”）^[17]，证明本研究的标注标准清晰、可靠，

表 4 基于大语言模型的中医黄疸病古籍信息抽取结果
测试集的一致性检验结果

Table 4 Consistency test of the information extraction results of TCM classics for jaundice based on the large language models

序号	实体类型	Kappa 值
1	病名	0.87
2	证型	0.93
3	病因	0.87
4	病机	0.90
5	病位	0.93
6	病性	0.91
7	病势	0.92
8	症状	1.00
9	舌象	1.00
10	脉象	0.96
11	治则	0.94
12	治法	0.97
13	方剂	0.96
14	中药	0.94
15	经络	1.00
16	腧穴	1.00
17	疗效	0.87

所形成的 100 条标注语料可作为高置信度的标准测试集，为后续大语言模型性能的对比评估提供了可信的基准。

3.3 大模型小样本标注对比实验

为选择最适合本研究任务的大语言模型，本研究选取了在国内开源并在复杂推理任务及中文语料理解上表现出独特优势的 DeepSeek-R1（2025-01-20 版）与国际上在通用自然语言处理任务上表现卓越并广泛应用的 ChatGPT-o3（2025-04-16 版）两个具有代表性的模型进行对比。

1) 实验设计。基于标注指南的核心规则和标注实例，设计大模型实体识别指令提示词，分别调用 DeepSeek-R1（<https://api.deepseek.com>）和 ChatGPT-o3（<https://cfwus02.opapi.win/v1>）大模型应用程序编程接口（API），使用相同的参数（temperature=1.0，最大 token 10 000），分别对上述 100 条标准测试语料进行实体识别。

2) 实验结果。以人工标注的标准测试集结果为参照，计算两模型的精确率（Precision）、召回率（Recall）和 F1 值（F1-Score）。结果显示，DeepSeek-R1 标注病名等 17 类实体平均精准率、召回率和 F1 值分别为 91.01%、92.87% 和 92.87%，ChatGPT-o3 分别为 75.21%、79.84% 和 73.96%，详见表 5。

3) 模型评估。经对比可见 DeepSeek-R1 在本研究任务中的优势主要体现在以下几方面：一是上下文语义理解深度更强，能精准捕捉中医术语的隐含关联。如病机类实体 DeepSeek-R1（F1=89.36%）显著优于 ChatGPT-o3（79.43%），差距近 10%，古籍文本中“湿热郁蒸，胆汁外溢”的描述，DeepSeek-R1 能通过“郁蒸”与“外溢”的因果关联，准确标注“湿热”为病因、“胆汁外溢”为病机；而 ChatGPT-o3 可能仅孤立识别“湿热”而忽略动态病机过程，将动态病机“胆汁外溢”误判为症状。二是中医领域专业推理能力突出，能破解古籍中的隐晦逻辑关系。如在治则实体上，DeepSeek-R1（F1=98.18%）对 ChatGPT-o3（72.73%）实现显著领先（差距 25.45%）。如“黄疸当利小便”，DeepSeek-R1 能结合“黄疸属湿”的病性知识，推理出“利小便”是针对湿邪的治则，并关联到具体治法“清热利湿”；而 ChatGPT-o3 易误判为普通症状描述。三是复杂语料适应性更优，能精准识别文言结构与罕见表述，如病势、病性等易于区分的实体。DeepSeek-R1 对生僻字词、省略句式的强解析力，使其在非标准化古籍语料中表现更稳健。例如古文“黄疸久羁，色晦如

表 5 DeepSeek-R1 和 ChatGPT-o3 大模型的实体识别精准率、召回率和 F1 值比较

Table 5 Comparison of precision rate, recall rate and F1 score for entity category annotation of the corpus by DeepSeek-R1 and ChatGPT-o3

实体类别	精准率/%		召回率/%		F1 值	
	DeepSeek-R1	ChatGPT-o3	DeepSeek-R1	ChatGPT-o3	DeepSeek-R1	ChatGPT-o3
病名	98.11	82.46	98.11	88.68	98.11	88.68
证候	92.41	77.50	89.02	75.61	89.02	75.61
病因	96.24	86.92	92.09	81.29	92.09	81.29
病机	93.33	83.58	89.36	79.43	89.36	79.43
病性	95.61	84.87	77.30	71.63	77.30	71.63
病位	93.33	83.58	89.36	79.43	89.36	79.43
病势	94.00	71.70	83.93	67.86	83.93	67.86
症状	93.33	83.58	89.36	79.43	89.36	79.43
舌象	100	92.86	83.33	72.22	83.33	72.22
脉象	95.59	78.48	97.01	92.54	97.01	92.54
治则	98.18	70.18	98.18	72.73	98.18	72.73
治法	97.10	66.67	91.78	71.23	91.78	71.23
经络	100	42.11	100	72.73	100	72.73
腧穴	100	100	100	66.67	100	66.67
中药	100	95.61	100	99.38	100	99.38
方剂	100	82.11	100	98.73	100	98.73
疗效	100	89.29	100	87.72	100	87.72
平均	91.01	75.21	92.87	79.84	92.87	73.96

烟熏”，能通过“久羁”判定病势（慢性迁延），而 ChatGPT-o3 因“羁”等古语词解析失败，易漏标或误归类为症状。四是关键实体识别准确率碾压，核心诊疗要素识别近乎完美。在直接影响诊疗决策的证候等实体上，DeepSeek-R1 证候实现精准率（92.41%）优于 ChatGPT-o3（77.50%），而 ChatGPT-o3 还存在漏检的情况（如病势召回率仅 67.86%）。

3.4 实体信息抽取结果

基于对比实验结果，本研究选用性能更优的 DeepSeek-R1 大模型对中医黄疸病古籍语料库进行实体信息抽取，通过调用 DeepSeek-R1 API，将每一条古籍语料与设计好的抽取指令相结合，提交给模型，模型返回结构化的 JSON 格式结果。

本研究通过对 10 243 条语料进行批量处理，最终从中医黄疸病古籍语料库中抽取病名等 17 类别共 41 407 个实体，涵盖了病名（6238 个）、证型（1622 个）、方剂（2631 个）、中药（2706 个）等黄疸病辨证论治核心诊疗要素，为下一步开展黄疸病辨证论治规律研究奠定了坚实的数据基础。详见表 6。

抽取示例，例如输入语料：“唐童年面黄，能食目眩，发热不时，由湿甚生热，热蒸变黄，胃热谷消，此为谷疸。宜猪肚丸。入秋，食后胀眩便

表 6 DeepSeek-R1 大模型的不同类型实体抽取数量

Table 6 Entity information extraction results using the large language model DeepSeek-R1

序号	实体类型	实体数量
1	病名	6238
2	证型	1622
3	病因	4095
4	病机	3094
5	病位	3129
6	病性	2964
7	病势	514
8	症状	7222
9	舌象	266
10	脉象	1107
11	治则	1910
12	治法	1537
13	方剂	2631
14	中药	2706
15	经络	507
16	腧穴	404
17	疗效	1461
总计		41 407

溲，脉虚小，热与湿搏，由太阴不运，少阳化风。主理脾阳，佐以熄风。生白术、潞参、陈皮、薏苡仁、鸡内金、半夏曲（俱炒）、茵陈、赤苓、甘菊

(炒)、天麻(煨),服愈。谷疸之为病,寒热不食,食即头眩,心胸不安,久久发黄为谷疸,茵陈蒿汤主之。”模型能准确输出对应实体,详见附图2(请扫描本文二维码获取)。

4 小结

本研究通过系统整合互联网开源的数字化中医古籍资源,构建了跨越先秦至清代七个朝代、涵盖五百余部古籍、万余条条文的中医黄疸病专病古籍历时语料库。在此基础上,本研究创新性地设计了“病-证-症-治-效”五元信息抽取模型,并通过严谨的人工标注与一致性评估,验证了DeepSeek-R1大语言模型在中医古籍复杂语义理解与实体抽取任务中的卓越性能,最终自动化生成了包含四万余个实体的结构化知识库,将分散在中医古籍中的非结构化的黄疸病隐性知识转化为显性化、结构化与语义化的知识体系,为现代中医黄疸病临床和科研提供历史依据和基础数据支撑,也为AI驱动的中医古籍知识挖掘与传承创新提供了可复制的技术范式。

本研究构建的语料库的朝代分布呈现显著的不均衡性,这种“近多古少”的分布状况主要受现实因素影响,主要是古籍保存与流传的概率随时代推移而递减,年代越久远的典籍越难完整存世,清代距现代最近,文献存量相对丰富,且数字化整理程度较高。为降低此种不均衡性可能对黄疸病历时演变规律结论产生的潜在影响,在后续分析中需要采取针对性措施,一方面将按朝代对语料进行分层加权分析,在考查某一概念的历时趋势时,要关注其在各朝代内部语料中的相对占比与出现语境,另一方面要强化对《黄帝内经》《金匱要略》等关键早期典籍的重点深度剖析,确保其核心思想得以充分揭示。下一步我们将进一步结合知识挖掘与知识图谱等方法,探索在揭示黄疸病病名的动态演变、发现跨时代核心药组的作用机理、捕捉治则演变的关键节点等方面实现突破。

参考文献

- [1]荣三群.明清时期黄疸病中医外治法及用药规律研究[D].长沙:湖南中医药大学,2015.
- [2]刘子晴.中医门诊电子病历关键临床信息抽取方法研究[D].广州:广州中医药大学,2021.
- [3]林凤.中医临床症状信息自动抽取与编码方法研究[D].武汉:湖北中医药大学,2025.
- [4]刘彬.基于BERT预训练模型的中医医案命名实体识别研究[D].长沙:湖南中医药大学,2023.
- [5]蒋继彪.基于语料库的《伤寒论》英译本翻译风格比较研究[J].中国中医基础医学杂志,2023,29(9):1531-1534.
- [6]陈璟.基于聚类的中医临床术语语义关系的研究[D].镇江:江苏科技大学,2023.
- [7]王欣宇,杨涛,谢佳东,等.大语言模型在中医诊疗场景的应用思路探讨[J].中医药信息,2025,42(6):12-17.
- [8]国家市场监督管理总局,国家标准化管理委员会.中医病证分类与代码[S].北京:中国标准出版社,2021.
- [9]国家市场监督管理总局,国家标准化管理委员会.中医临床诊疗术语[S].北京:中国标准出版社,2021.
- [10]张伯礼,吴勉华.中医内科学[M].北京:中国中医药出版社,2017.
- [11]李灿东,吴承玉.中医诊断学[M].北京:中国中医药出版社,2021.
- [12]李冀,连建伟.方剂学[M].北京:中国中医药出版社,2021.
- [13]郑洪新,何清湖.中医基础理论[M].北京:中国中医药出版社,2021.
- [14]中国中医药信息学会.中医药文本数据命名实体信息分类与代码[EB/OL].(2023-07-15)[2025-11-10].<https://www.ttbz.org.cn/StandardManage/Detail/90184>.
- [15]中国中医药信息学会.中医药文本数据关系信息分类与代码[EB/OL].(2023-07-05)[2025-11-10].<https://www.ttbz.org.cn/StandardManage/Detail/90186>.
- [16]COHEN J. A coefficient of agreement for nominal scales[J]. Educ Psychol Meas, 1960,20(1): 37-46.
- [17]LANDIS J R, KOCH G G. The measurement of observer agreement for categorical data[J]. Biometrics, 1977,33(1): 159-174.

Construction of a Traditional Chinese Medicine Classics Corpus for Jaundice Based on Large Language Models

ZHANG Yiran, LI Haiyan, NIE Ying

China Academy of Chinese Medical Sciences, Beijing, 100700

ABSTRACT Objective To construct a structured corpus of traditional Chinese medicine (TCM) classics on jaundice, providing data support for modern clinical practice and scientific research on jaundice, and facilitating the deep mining and utilization of knowledge from traditional Chinese medicine (TCM) classics. **Methods** Open-source digitized resources of TCM classics available on the internet were systematically integrated. Bibliometric methods were employed to construct a terminology database for jaundice. Python regular expressions were used for keyword matching and corpus extraction. A five-element information extraction model of "disease-syndrome-symptom-treatment-efficacy" was designed. Annotation guidelines were developed, and a standard test set was formed through back-to-back annotation by two annotators followed by arbitration through a third party. To select the optimal model for entity information extraction, two large language models, DeepSeek-R1 and ChatGPT-o3, were chosen for comparative performance evaluation by invoking their application programming interfaces (APIs) for entity extraction. **Results**

A TCM classics corpus for jaundice was constructed, encompassing 561 ancient books from the pre-Qin period to the Qing Dynasty, and 10,243 pieces of text. The comparative evaluation results showed that DeepSeek-R1 outperformed ChatGPT-o3 in the performance of entity recognition across all 17 entity types. Based on this optimized model DeepSeek-R1, 41,407 entities were extracted from the corpus, covering 17 entity types, including disease names (6238), syndrome types (1622), formulas (2631), and Chinese medicinal herbs (2706). **Conclusion** The corpus constructed using DeepSeek-R1 large language model covers the essential syndrome differentiation and treatment elements for jaundice, laying a solid foundation for subsequent research on the historical evolution and syndrome differentiation and treatment rules of jaundice in TCM. It also provides a replicable technical paradigm for AI-driven knowledge mining from TCM classics.

Keywords traditional Chinese medicine classics; jaundice; corpus; large language models; information extraction

(收稿日期: 2025-08-30; 录用日期: 2025-11-20; 修回日期: 2026-02-05)

[编辑: 焦 爽]

广告

《中医杂志(英文版)》欢迎投稿

Journal of Traditional Chinese Medicine [《中医杂志(英文版)》] (ISSN 0255-2922, CN 11-2167/R, 双月刊) 由中华中医药学会与中国中医科学院联合主办, 是我国中医药界具有权威的中医药英文科技期刊, 现发行和交换 90 多个国家和地区。

《中医杂志(英文版)》自 1981 年创刊以来一直坚持正确的办刊宗旨, 国际影响力不断提升。2010 年入选 SCI-E (《科学引文索引》扩刊版), 2011 年获得中国中医科学院自主选题项目资助, 2013 年获得“中国科技期刊国际影响力提升计划”资助, 自 2012 年起连续被评为“中国最具国际影响力学术期刊”, 2023 年再次入选“T1 级中医药英文科技期刊目录”。本刊先后被国际知名文献数据库 MEDLINE、《化学文摘》(CA)、《医学文摘》(EM) 等收录, 是 MEDLINE 收录的第一份中国出版的中医药期刊, 是 SCI 收录的 30 余种全科医学和替代医学期刊中的中国传统医学综合性学术期刊之一。

《中医杂志(英文版)》主要栏目包括循证研究、理论研究、基础研究、临床观察、中药研究、文献研究, 思路与方法、综述等。内容以报道中医药基础、临床研究成果和科研前沿进展为主, 覆盖中医药学科各个领域。欢迎广大专家学者积极投稿。作者可直接登录本刊网站 <http://www.journaltcm.cn> 投稿。

联系邮箱: jtcmen@126.com。联系电话: 010-64089190。

联系地址: 北京市东城区东直门内南小街 16 号, 邮政编码: 100700。