

DOI:10.13288/j.11-2166/r.2026.07.010

基于大语言模型与潜在类别模型的混合痔 中医证型分布特征及危险因素分析 ——一项多中心真实世界研究

邓茹月, 丁康, 朱雨欣, 李猛, 朱慧婷, 杜蕾✉

南京中医药大学附属南京市中医院, 江苏省南京市秦淮区大明路 157 号, 210022

[摘要] **目的** 基于多中心真实世界诊疗数据, 构建混合痔中医证型标准化分类模型, 揭示其证型分布规律及核心危险因素, 为促进中医辨证规范化与实施精准辨证论治干预提供循证依据。**方法** 采用多中心横断面研究设计, 纳入 2023 年 9 月 1 日至 2024 年 12 月 31 日江苏省 8 所医院的 13 283 例混合痔患者。整合 DeepSeek-R1-Distill-Qwen-7B 及 LLaMA-3.3 大语言模型与潜在类别分析方法, 对中医证候信息进行无监督学习与潜在类别建模。通过单因素分析筛选潜在危险因素, 并进一步运用 Logistic 回归分析确定各证型的独立危险因素。**结果** 模型各项指标在不同病历段落类型中表现稳定可靠: 门诊病历中既往史 ($F1=99.7\%$)、现病史 ($F1=94.9\%$)、专科检查 ($F1=90.7\%$); 住院病历中既往史 ($F1=98.2\%$)、现病史 ($F1=91.2\%$)、专科检查 ($F1=90.3\%$)、出院情况 ($F1=90.6\%$)。潜在类别模型识别出混合痔四种核心中医证型: 脾虚气陷证 (915 例, 6.89%)、湿热下注证 (10 820 例, 81.46%)、气滞血瘀证 (1252 例, 9.43%) 及风伤肠络证 (296 例, 2.22%); 潜在类别概率分别为 0.069、0.815、0.094、0.022。Logistic 回归分析显示, 性别、年龄、病程、高血压病、糖尿病、高脂血症、便秘、吸烟史、饮酒史是不同证型分化的独立危险因素 ($P<0.05$)。疗效验证评估发现, 脾虚气陷证与气滞血瘀证患者的治愈率高于湿热下注证 (校正后 $P<0.05$), 其余证型间差异无统计学意义。**结论** 湿热下注证为混合痔的主要证型, 性别、年龄、病程、高血压病、糖尿病、高脂血症、便秘、吸烟史、饮酒史是不同证型分化的独立危险因素。

[关键词] 混合痔; 证候分布; 大语言模型; 潜在类别模型; 真实世界研究

痔病是最常见的肛肠疾病之一, 我国成年人群患病率高达 51.56% ^[1], 且在不良生活方式、环境及遗传因素协同作用下, 呈持续上升趋势^[2-3]。混合痔作为痔病最主要的临床亚型, 不仅构成疾病负担的主体, 更因反复便血、痔核脱垂嵌顿等并发症显著降低患者生活质量^[4]。《中国痔病诊疗指南 (2020)》将混合痔的中医辨证分为气滞血瘀证、湿热下注证、脾虚气陷证和风伤肠络证四型^[5], 但

对各证候分布规律与影响因素的研究仍有不足。

作为中医诊疗的核心环节, 辨证论治的准确性对临床疗效具有决定性影响。传统辨证方法因依赖医师主观经验, 存在可重复性不足与判别模糊性的局限^[6]。中医证型本质是多维症状的潜在聚类过程, 与潜在类别模型的降维简化原理深度契合。作为一种以个体为中心的探索性统计模型, 潜在类别分析通过无监督学习挖掘显变量间内在关联, 可观测症状间的潜在结构, 客观识别并量化不同证型的分类特征^[7-8], 从而为辨证标准化提供方法学支撑。本研究旨在运用潜在类别分析方法, 以数据驱动的方式探索混合痔患者群体中是否存在客观的、具有不同症状特征的潜在亚组, 并进一步检验这些亚组与中医证型之间的映射与补充关系, 为中医证候的客观化与规范化研究提供新的量化分析视角。大语

基金项目: 江苏省中医肛肠疾病临床医学创新中心重点项目 (苏中医科教函〔2023〕1 号); 南京大学中国医院改革发展研究院全国范围立项课题 (NDYG029/025); 国家中医药管理局监测统计中心中医药监测统计研究课题 (2025JCTJE2); 国家中医药管理局规划财务司项目 (GZY-GCS-2025-006); 中华人民共和国工业和信息化部人工智能医疗器械创新任务揭榜挂帅项目 (2025-AIFZZL-02009)

✉ 通讯作者: fsyy00237@njucm.edu.cn

言模型基于海量参数架构与深层语义解析能力,在医学文本挖掘领域具有独特优势^[9-10]。大语言模型的上下文感知与知识推理特性,能高效处理真实世界研究中的高维异构数据^[11],为中医证型文本的结构化处理与特征萃取提供关键技术范式。基于此,本研究联合大语言模型与潜在类别模型,依托 13 283 例多中心真实世界数据,系统构建混合痔中医证型标准化模型并揭示证型分布规律,深入解析不同证型的影响因素,以建立可验证的辨证标准化智能范式,为患者提供精准症状管理路径,并为人群防治策略制定提供循证决策支持。

1 临床资料

1.1 研究对象

本研究由江苏省卫生健康委员会组织数据质控,选取南京市中医院、无锡市中医医院、徐州市中医院、张家港市中医医院、常熟市中医院、苏州市中医医院、江阴市中医院、泰州市中医院共 8 家医院参与,运用大语言模型标准化整合门诊、住院诊疗记录及检验、检查结果等全流程医疗数据,建立江苏省混合痔专病大模型数据库。选取该数据库中 2023 年 9 月 1 日至 2024 年 12 月 31 日记录的混合痔门诊及住院患者的临床资料。本研究方案经南京市中医院伦理委员会审批(批件号:KY2025129)。

1.2 诊断标准

诊断标准参照《中国痔病诊疗指南(2020)》^[5]及《痔中西医结合诊疗指南(2025 版)》^[12]。

1.3 纳入标准

1) 符合上述诊断标准;2) 年龄 ≥ 18 岁,性别不限;3) 拥有完整、可追溯的电子或纸质病历记录,能够满足研究所需的核心数据提取要求(至少包括基本信息、现病史、既往史、中医诊断、治疗方案、关键疗效指标)。

1.4 排除标准

1) 严重精神疾病及认知功能障碍;2) 合并心、脑、肝、肾、呼吸、神经、免疫、血液等系统严重疾病,感染性疾病及恶性肿瘤;3) 妊娠期、哺乳期或近期有妊娠计划者。

2 方法

2.1 基于大语言模型的数据库建立

针对现病史、既往史、专科检查等大段文本类非结构化病历,本研究采用 DeepSeek-R1-Distill-Qwen-7B 模型进行专病结构化开发,搭配 LLaMA-3.3 大语言模型用于协同实现病历后结构化的交叉

验证。经前期多轮上百份人工标注样本训练后,对非结构化病历进行解析,从中提取如病程、临床类型、疾病分期、吸烟史、饮酒史、合并疾病等数据项,实现病历后结构化处理并提取专项分析指标。训练过程中,采用 $2e^{-5}$ 的学习率,并基于低秩适应模型(LoRA)微调方法,设定秩维度 $r=64$,同时在 Q/K/V 矩阵及 MLP 层进行适配,以兼顾模型性能与计算效率。模型经过基于多轮人工标注样本的迭代训练与梯度下降优化,每轮标注均经临床专家复核与一致性检验,确保标注质量与模型收敛稳定,最终形成拥有标准化数据集的数据库平台为后续研究提供支撑。在模型性能评估方面,采用准确率(Accuracy)、召回率(Recall)与 F1 值(F1-score)三类指标进行系统验证^[13]。其中,准确率为完全匹配的实体占比,召回率为识别出的实体占标注实体的比例,F1 值为准确率与召回率的调和平均数,用于综合反映模型的整体识别效能。

2.2 资料提取

对于部分需要在自然文本语言中提取的字段,基于通用大模型非结构化数据解析处理,根据正则表达式进行提取变量。对于标准规范化字段则直接提取。包括:人口学资料(年龄、性别、婚姻状况)、病程、主要症状、合并症(高血压病、糖尿病、高脂血症、便秘)、吸烟史、饮酒史、治疗转归。参照《中国痔病诊疗指南(2020)》^[5],确定肿物脱出、便血、瘙痒、血栓、疼痛、大便性状及肛门坠胀为主要症状。治愈率依据文献中的疗效标准计算。治愈:临床症状及痔体完全消失,创面完全愈合;有效:痔体明显减小,临床症状明显好转,创面愈合;无效:临床症状、痔体大小均无明显改善。治愈率=治愈例数/总病例数 $\times 100\%$ 。

2.3 混合痔的中医辨证分型

参照诊疗指南^[5, 12]及既往研究^[14],将混合痔分为 4 种证型。具备主症 3 项加次症 1 项,或主症 2 项加次症 2 项即可确定证型。

风伤肠络证。主症:1) 肿物脱出肛外、水肿;2) 大便滴血、射血或带血,血色鲜红;3) 大便干结;4) 肛门瘙痒。次症:1) 口干咽燥;2) 舌红、苔黄,脉浮数。

湿热下注证。主症:1) 肛门肿物外脱、肿胀;2) 便血色鲜红,量较多;3) 大便时肛门灼热疼痛;4) 便干或溏。次症:1) 小便短赤;2) 舌质红、苔黄腻,脉弦数。

气滞血瘀证。主症:1) 肿物脱出肛外、水肿;

2) 有血栓形成; 3) 有嵌顿, 表面紫暗、糜烂、渗液; 4) 疼痛剧烈, 触痛明显, 肛管紧缩; 5) 大便秘结。次症: 1) 小便不利; 2) 舌质紫暗或有瘀斑, 脉弦或涩。

脾虚气陷证。主症: 1) 肿物脱出肛外, 不易复位; 2) 便血色淡; 3) 肛门坠胀; 4) 排便乏力; 5) 大便异常(溏、烂、先硬后溏、时溏时硬)。次症: 1) 面色少华, 头晕神疲, 食少乏力, 少气懒言; 2) 舌淡胖、苔薄白, 脉细弱。

2.4 潜在类别分析

采用 Mplus 7.4 软件对混合痔患者的中医证型进行潜在类别分析。具体分析流程如下。1) 模型设定: 主要参数包括潜在类别概率及观测症状指标的条件概率。2) 参数估计与模型拟合评价: 采用最大似然法进行参数估计, 其迭代过程通过期望最大化算法实现。模型拟合优度评价指标包括: 对数函数似然值(LL)、赤池信息准则(AIC)、贝叶斯信息准则(BIC)、样本矫正的贝叶斯信息准则(aBIC)、熵值、似然比率检验(LMRT)和基于Bootstrap的似然比检验(BLRT)。较低的AIC、BIC、aBIC和较高的熵值表示模型拟合较好^[15-16]。LMRT与BLRT用于比较潜在类别模型的拟合差异, 如果两个值的P值达到显著水平, 则表明k个类别模型优于k-1个类别模型。3) 潜在类别划分: 依据贝叶斯定理计算个体所属的最优潜在类别的后验概率。4) 结果解释: 根据各潜在类别下症状指标的条件概率分布特征, 结合混合痔中医辨证分型理论, 对潜在类别的临床意义进行诠释。若某潜在类别中包含3个及以上主症, 则判定该类别符合对应中医证型的特征表现。

2.5 统计学方法

采用 STATA 17 软件进行统计分析。计量资料均服从正态分布, 以均值±标准差($\bar{x} \pm s$)表示, 计数资料以例(%)表示。应用 χ^2 检验筛选与中医证型相关的单因素指标, 运用二元 Logistic 回归分析探讨混合痔患者不同中医证型的危险因素; 不同证型疗效的组间比较采用 χ^2 检验, 多个独立样本两两比较的事后检验进行 Bonferroni 校正。 $P < 0.05$ 为差异有统计学意义。使用 R 4.4.3 软件制作森林图。

3 结果

3.1 大语言模型的建立及评估结果

运用DeepSeek-R1-Distill-Qwen-7B及LLaMA-3.3

大语言模型, 构建标准化专病数据库平台, 模型评估显示, 各项指标在不同病历段落类型中表现稳定可靠, 具体结果如下: 门诊病历中既往史(F1=99.7%)、现病史(F1=94.9%)、专科检查(F1=90.7%); 住院病历中既往史(F1=98.2%)、现病史(F1=91.2%)、专科检查(F1=90.3%)、出院情况(F1=90.6%)。

3.2 一般资料与主症分布

本研究共纳入 13 283 例混合痔患者, 其中男性 6092 例(45.86%), 女性 7191 例(54.14%); 年龄 18~94 岁, 平均年龄(45.54±14.26)岁; 病程范围 0~70 年, 平均病程(2.80±4.99)年。一般人口学、疾病相关特征详见表 1, 主症分布情况详见表 2。

表 1 13 283 例混合痔患者的一般情况
Table 1 General characteristics of 13 283 mixed hemorrhoid patients

项目	例 (%)
性别	
男性	6092 (45.86)
女性	7191 (54.14)
年龄	
18~29 岁	1619 (12.19)
30~39 岁	3696 (27.83)
40~49 岁	2845 (21.42)
50~59 岁	2676 (20.15)
≥60 岁	2447 (18.42)
婚姻状况	
已婚	10 885 (81.95)
其他	2398 (18.05)
病程	
<1 年	4814 (36.24)
1~<6 年	6544 (49.27)
6~10 年	1380 (10.39)
>10 年	545 (4.10)
合并症	
高血压病	1915 (14.42)
糖尿病	479 (3.61)
高脂血症	861 (6.48)
便秘	3125 (23.53)
吸烟史	1501 (11.30)
饮酒史	1681 (12.66)
治疗转归	
治愈	12 571 (94.64)
未治愈	712 (5.36)

3.3 潜在类别建模

3.3.1 潜在类别建模确定最佳潜在类别数 运用潜在类别模型对选取的 7 项主症指标进行建模。依

表2 13 283 例混合痔患者的主要症状分布
Table 2 Distribution of main symptoms in 13 283 mixed hemorrhoid patients

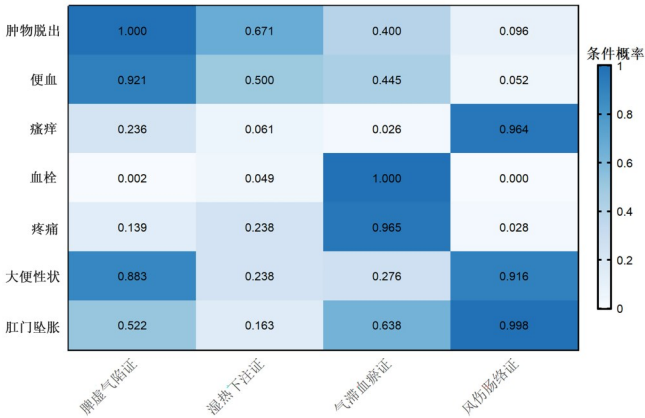
主要症状	例 (%)
肿物脱出	
助纳/无法回纳	8945 (67.34)
自行回纳	4338 (32.66)
便血	
大量 (肛门射血)	7120 (53.60)
少量 (肛门滴血)	6163 (46.40)
瘙痒	
有	1313 (9.88)
无	11 970 (90.12)
血栓	
有	1700 (12.80)
无	11 583 (87.20)
疼痛	
中度/剧烈疼痛 (VAS≥4分)	4439 (33.42)
轻度疼痛 (VAS<4分)	8844 (66.58)
大便性状	
异常 (便干或溏)	4439 (33.42)
正常	8844 (66.58)
肛门坠胀	
有	3558 (26.79)
无	9725 (73.21)

注：VAS，视觉模拟评分法。

次建立 1-6 个类别的潜在类别模型，进行模型拟合度分析，混合痔中医证型 1-6 个类别模型的拟合指标，见表 3。AIC、BIC、aBIC 值均随类别数目的增多逐步减小，表明随着类别的增加，模型整体的拟合优度在逐步提高。增加至 4 个类别模型时，熵值达到 0.854，提示分类准确超过 90%^[17]。综合考虑模型的简洁性与准确性，本研究最终选取 4 个类别模型作为最佳潜在类别模型。

3.3.2 混合痔潜在类别模型的特征分析 基于各指标的条件概率与混合痔中医辨证分型理论^[5, 14],

对 4 种混合痔中医证型的特征进行描述与命名。类别 1 痔核脱出需助纳/无法回纳、大量便血、大便秘性状异常、肛门坠胀等症状的条件概率较高，符合气虚下陷的核心病机，因此归为“脾虚气陷证”；类别 2 痔核脱出需助纳/无法回纳、大量便血、肛门轻度疼痛等症状的条件概率较高，与湿热蕴结、下注肛肠的典型表现一致，因此归为“湿热下注证”；类别 3 痔核脱出可自行回纳、少量便血、血栓、中度/剧烈疼痛等症状的条件概率较高，体现了气机阻滞、血行瘀阻的病理特点，因此归为“气滞血瘀证”；类别 4 痔核脱出可自行回纳、少量便血、瘙痒、轻度疼痛、大便异常、肛门坠胀等症状的条件概率较高，符合风邪挟热湿伤及肠络的证候特点，因此归为“风伤肠络证”。13 283 例患者中脾虚气陷证 915 例 (6.89%)，湿热下注证 10 820 例 (81.46%)，气滞血瘀证 1252 例 (9.43%)，风伤肠络证 296 例 (2.22%)；潜在类别概率分别为 0.069、0.815、0.094、0.022。基于症状的潜在类别模型特征见图 1。4 个类别混合痔中医证型潜在



注：图中颜色深度用于表征脾虚气陷证、湿热下注证、气滞血瘀证、风伤肠络证对应症状的条件概率，颜色越深，概率越高。

图1 13 283 例混合痔患者潜在类别模型的条件概率热图
Figure 1 Conditional probability heatmap for the latent class model in 13 283 mixed hemorrhoid patients

表3 13 283 例混合痔患者潜在类别模型的拟合优度
Table 3 Goodness of fit for the latent class model in 13 283 mixed hemorrhoid patients

类别数	AIC	BIC	aBIC	熵值	LMRT	BLRT
1	102 158.807	102 211.267	102 189.021	N/A	N/A	N/A
2	99 184.082	99 296.495	99 248.827	0.962	<0.001	<0.001
3	97 780.431	97 952.799	97 879.707	0.831	<0.001	<0.001
4	96 891.819	97 124.140	97 025.625	0.854	<0.001	<0.001
5	96 603.640	96 895.916	96 771.977	0.673	<0.001	<0.001
6	96 349.650	96 701.879	96 552.518	0.699	<0.001	<0.001

注：AIC，赤池信息准则；BIC，贝叶斯信息准则；aBIC，样本校正的贝叶斯信息准则；LMRT，似然比率检验；BLRT，基于Bootstrap的似然比检验；N/A，不适用项。

表 4 13 283 例混合痔患者 4 个潜在类别模型的条件概率

Table 4 Conditional probabilities for the four latent class models in 13 283 mixed hemorrhoid patients

症状变量	脾虚气陷证	湿热下注证	气滞血瘀证	风伤肠络证
肿物脱出				
助纳/无法回纳	1. 000	0. 671	0. 400	0. 096
自行回纳	0. 000	0. 329	0. 600	0. 913
便血				
大量	0. 921	0. 500	0. 445	0. 052
少量	0. 079	0. 500	0. 555	0. 948
瘙痒				
有	0. 236	0. 061	0. 026	0. 964
无	0. 764	0. 939	0. 974	0. 036
血栓				
有	0. 002	0. 026	1. 000	0. 000
无	0. 998	0. 974	0. 000	1. 000
疼痛				
中度/剧烈疼痛(VAS≥4分)	0. 139	0. 238	0. 965	0. 028
轻度疼痛(VAS<4分)	0. 861	0. 762	0. 035	0. 972
大便性状				
异常(便干或溏)	0. 883	0. 238	0. 276	0. 916
正常	0. 117	0. 762	0. 724	0. 084
肛门坠胀				
有	0. 522	0. 163	0. 638	0. 998
无	0. 478	0. 837	0. 362	0. 002

注：VAS，视觉模拟评分法。

类别模型与其在 7 项主要症状指标上应答的条件概率详见表 4。

3.4 混合痔不同证型的危险因素分析

3.4.1 单因素分析 表 5 示，混合痔各中医证型患者间的性别、年龄、婚姻状况、病程、合并高血压病、合并高脂血症、合并便秘、吸烟史、饮酒史比较差异有统计学意义 ($P<0.05$)。

3.4.2 多因素分析 为进一步探究混合痔不同证型的危险因素，采用二元 Logistic 回归模型进行分析。为避免预测变量遗漏，将单因素分析中 $P<0.10$ 的变量作为自变量，4 种混合痔中医证型的虚拟变量 (0=否，1=是) 作为因变量进行回归分析。结果以比值比 (OR) 及其 95%CI 表示，其中 $OR>1$ 表示该因素与证型呈正相关， $OR<1$ 表示呈负相关。

图 2 示，脾虚气陷证高发于女性 ($OR=1.514$)、60 岁及以上人群 ($OR=1.535$)、病程 1~<6 年 ($OR=3.128$) 及 6~10 年 ($OR=2.719$) 者 (均 $P<0.05$)；合并高血压病 ($OR=1.695$ ， $P<0.001$)、糖尿病 ($OR=1.702$) 或高脂血症 ($OR=1.991$) 时风险显著增加，合并便秘 ($OR=0.149$)、吸烟 ($OR=0.389$) 或饮酒 ($OR=0.421$) 病史的患者风

险显著降低 (均 $P<0.05$)。湿热下注证多见于男性 ($OR=0.847$)、30~39 岁人群 ($OR=1.211$)、未婚者 ($OR=1.521$) 及病程 >10 年 ($OR=2.307$) 者 (均 $P<0.05$)；合并高脂血症 ($OR=1.204$)、便秘 ($OR=2.569$) 以及有吸烟史 ($OR=1.438$)、饮酒史 ($OR=2.841$) 是其主要风险因素，合并高血压病 ($OR=0.638$)、糖尿病 ($OR=0.757$) 与该证型呈显著负相关 (均 $P<0.05$)。气滞血瘀证多见于女性 ($OR=1.218$)、已婚者 ($OR=0.622$)，其风险随年龄增长而降低 (均 $P<0.05$)；与病程 <1 年者相比，病程 1~<6 年 ($OR=0.168$)、6~10 年 ($OR=0.596$) 及 >10 年 ($OR=0.336$) 均与该证型呈负相关 (P 均 <0.05)；合并高血压病 ($OR=1.278$)、便秘 ($OR=2.785$) 会显著增加其风险，合并高脂血症 ($OR=0.564$)、饮酒史 ($OR=0.292$) 则与其呈显著负相关 (均 $P<0.05$)。风伤肠络证集中于男性 ($OR=0.508$)、已婚 ($OR=0.638$)、病程 1~<6 年 ($OR=3.569$) 者 (均 $P<0.05$)；合并高脂血症 ($OR=3.581$)、吸烟史 ($OR=7.598$)、饮酒史 ($OR=5.628$) 是其显著的风险预测因素 (均 $P<0.05$)，合并便秘 ($OR=0.073$) 与该证型呈显著负

表 5 13 283 例混合痔患者不同证型一般人口学、疾病相关特征的单因素分析结果

Table 5 Univariate analysis of demographic and disease-related characteristics for different syndrome types in 13 283 mixed hemorrhoid patients					[例 (%)]	
项目	脾虚气陷证(915 例)	湿热下注证(10 820 例)	气滞血瘀证(1252 例)	风伤肠络证(296 例)	χ^2 值	P 值
性别					24. 55	<0. 001
男性	485 (53. 01)	4901 (45. 30)	587 (46. 88)	119 (40. 20)		
女性	430 (46. 99)	5919 (54. 70)	665 (53. 12)	177 (59. 80)		
年龄					47. 22	<0. 001
18~29 岁	87 (9. 51)	1337 (12. 36)	168 (13. 42)	27 (9. 12)		
30~39 岁	203 (22. 19)	3059 (28. 27)	367 (29. 31)	67 (22. 64)		
40~49 岁	216 (23. 61)	2292 (21. 18)	256 (20. 45)	81 (27. 36)		
50~59 岁	199 (21. 75)	2157 (19. 94)	254 (20. 29)	66 (22. 30)		
≥60 岁	210 (22. 95)	1975 (18. 25)	207 (16. 53)	55 (18. 58)		
婚姻状况					82. 61	<0. 001
已婚	817 (89. 29)	8712 (80. 52)	1091 (87. 14)	265 (89. 53)		
其他	98 (10. 71)	2108 (19. 48)	161 (12. 86)	31 (10. 47)		
病程					1156. 66	<0. 001
<1 年	86 (9. 40)	3831 (35. 41)	859 (68. 61)	38 (12. 84)		
1~<6 年	744 (81. 31)	5354 (49. 48)	207 (16. 53)	239 (80. 74)		
6~10 年	65 (7. 10)	1154 (10. 67)	150 (11. 98)	11 (3. 72)		
>10 年	20 (2. 19)	481 (4. 45)	36 (2. 88)	8 (2. 70)		
合并症						
高血压病	96 (10. 49)	1621 (14. 98)	175 (13. 98)	23 (7. 77)	25. 01	<0. 001
糖尿病	21 (2. 30)	406 (3. 75)	45 (3. 59)	7 (2. 36)	6. 50	0. 090
高脂血症	20 (2. 19)	674 (6. 23)	163 (13. 02)	4 (1. 35)	130. 12	<0. 001
便秘	702 (76. 72)	2084 (19. 26)	115 (9. 19)	224 (75. 68)	2139. 09	<0. 001
吸烟史	376 (41. 09)	941 (8. 70)	174 (13. 90)	10 (3. 38)	910. 39	<0. 001
饮酒史	405 (44. 26)	976 (9. 02)	289 (23. 08)	11 (3. 72)	1100. 85	<0. 001

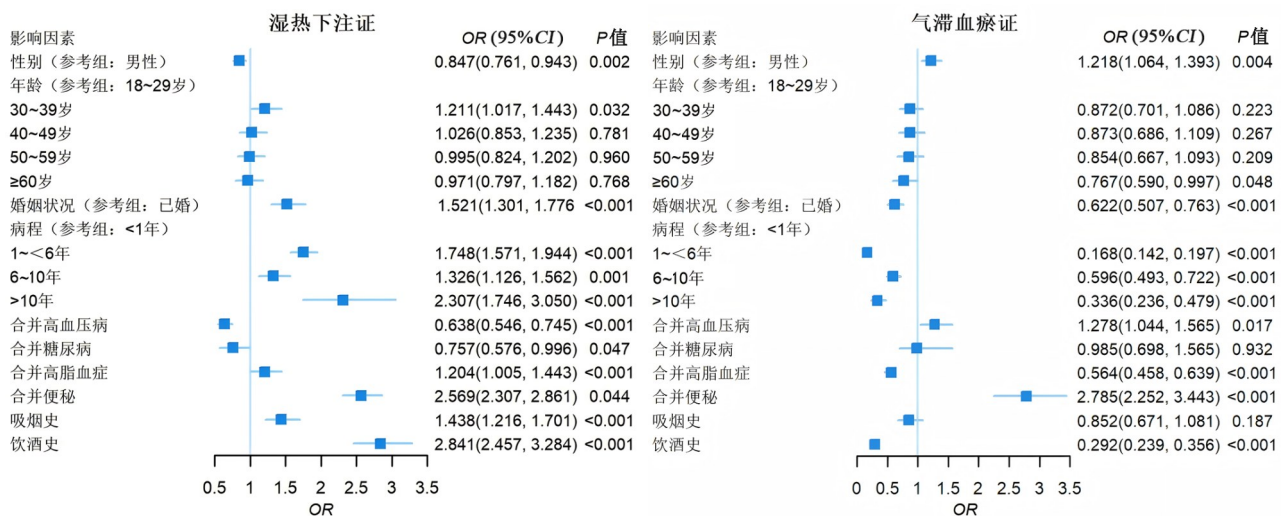


图 2 混合痔中医证型危险因素二元 Logistic 回归分析结果

Figure 2 Binary logistic regression analysis of risk factors for TCM syndrome types in mixed hemorrhoids

相关 (均 $P < 0.05$)。

3.5 混合痔不同证型的疗效分析验证

表 6 示, 混合痔各中医证型患者间的治愈率比较差异有统计学意义 ($\chi^2 = 44.16$, $P < 0.05$)。经

Bonferroni 校正的事后检验进一步表明, 脾虚气陷证、气滞血瘀证患者治愈率高于湿热下注证 (校正后 $P < 0.05$), 而其他证型之间的比较差异无统计学意义 ($P > 0.05$)。

表 6 13 283 例混合痔患者不同证型的疗效比较
Table 6 Comparison of treatment efficacy for different syndrome types in 13 283 mixed hemorrhoid patients [例(%)]

证型	例数	治愈	未治愈
脾虚气陷证	915	898 (98. 14)	17 (1. 86)
湿热下注证	10 820	10 175 (94. 04)	645 (5. 96)
气滞血瘀证	1252	1214 (96. 96)	38 (3. 04)
风伤肠络证	296	284 (95. 95)	12 (4. 05)

4 讨论

本研究基于 13 283 例多中心真实世界数据，分析混合痔中医证型的分布规律及其影响因素，为中医辨证标准化与精准干预提供证据。运用 DeepSeek-R1-Distill-Qwen-7B 及 LLaMA-3. 3 大语言模型。构建标准化专病数据库平台，并通过系统训练、严格质控与综合性能评估，保障了病历文本解析的准确性与鲁棒性，为后续分析奠定了高质量数据基础。通过潜在类别模型的无监督学习方法构建证型分类模型，识别出四大证型及其占比，各证型内部症状的条件概率分布客观印证了其与传统中医病机理论的高度一致性，这种源于真实世界数据的系统性对应，为中医辨证体系的合理构建与临床应用提供了重要支持。证型分布呈现以湿热下注证为主，气滞血瘀证与脾虚气陷证次之，风伤肠络证占比最低的特征，与既往流行病学调查结果一致^[18-19]。值得注意的是，湿热下注证占比显著高于既往文献报道^[20-22]，考虑原因可能与江苏地区湿热气候及高脂高糖饮食模式^[23]密切相关，凸显地域性真实世界数据的独特价值。

在影响因素层面，证型分布与多重因素存在深刻关联。在性别差异上，脾虚气陷证与气滞血瘀证多见于女性，可能与月经、妊娠、分娩等生理现象使女性腹腔及盆腔脏器受压有关，影响肛肠的血液循环和排便^[24-25]。湿热下注证与风伤肠络证则以男性为主，可归因于饮食偏嗜肥甘及久坐职业习惯^[26-27]。年龄分布显示，患者平均发病年龄为(45. 54±14. 26) 岁，30~50 岁为发病高峰。其中老年人群(≥60 岁)更易罹患脾虚气陷证，与既往报道^[14]相符；30~39 岁青壮年高发湿热下注证，或与青壮年饮食不节、饮酒过度等因素相关^[18]；18~29 岁青年则多见气滞血瘀证，可能源于久坐少动致肛垫血流速降低^[27]。

在病程特征上，脾虚气陷证与湿热下注证患者的病程普遍较长，提示此二证的形成可能与长期病损累积或病情迁延不愈相关；相较而言，气滞血瘀

证与风伤肠络证患者的病程较短，符合急性气血阻滞或风邪骤袭所致实证的发病特点。在合并症方面，既往研究已证实高血压病、糖尿病、高脂血症、便秘均为痔病的危险因素^[26, 28-29]。进一步分析可见，代谢性疾病(高血压病、糖尿病、高脂血症)与脾虚气陷证存在显著关联，高脂血症及便秘易诱发湿热下注证，高血压病合并便秘可促进气滞血瘀证的形成，同时高脂血症亦是风伤肠络证的危险因素。此外，吸烟、饮酒等不良生活习惯已被证实是湿热下注证与风伤肠络证的核心诱因^[30-31]，易致体内生湿生热，使湿热之气、浊气留注肛门而发为痔，这从病因学角度阐释了外邪侵袭的动力学机制。

在治愈率方面，脾虚气陷证与气滞血瘀证患者的治愈率高于湿热下注证，其他证型间比较差异无统计学意义，和本团队临床经验一致，原因可能是脾虚气陷证患者的主要诉求为解决脱出症状，气滞血瘀证患者则是解决疼痛诉求为主，手术的疗效明显优于其他手段，而湿热下注证患者存在多个明显症状，同时患者数量最多，因此有效率较低。

综上所述，本研究联合大语言模型与潜在类别模型，依托多中心真实世界数据集，建立了混合痔中医证型标准化模型，并分析了各证型的影响因素。本研究存在一定局限性：第一，样本仅来源于江苏省，限制了结论的普适性；第二，横断面数据难以推断因果关系；第三，疗效评估仅基于治愈率，缺乏多维度症状与远期结局指标；第四，未控制中心随机效应，可能影响模型的外推稳健性。未来将开展前瞻性、多中心队列研究，构建包含标准化症状量表与复发率的复合终点体系，并采用多水平模型控制中心效应，以提供更全面稳健的证据。

致谢

江苏省中医药管理局指导及江苏省中医优势病种诊疗大数据平台支持。

参考文献

[1]江维,张虹玺,隋楠,等.中国城市居民常见肛肠疾病流行病学调查[J].中国公共卫生,2016,32(10):1293-1296.

[2]XIONG YY, LIU H. Application intention and influencing factors of traditional Chinese medicine in patients with mixed hemorrhoids: a cross-sectional study in China [J]. Patient Prefer Adher, 2025, 19(1): 1673-1680.

[3]郭呈华,车晓玉,林芝,等.中国某健康体检人群痔的流行病学特征[J].北京大学学报(医学版),2024,56

- (5): 815-819.
- [4] SUN XW, XU JY, ZHU CZ, et al. Analysis of factors impacting postoperative pain and quality of life in patients with mixed hemorrhoids: a retrospective study [J]. World J Gastrointest Surg, 2024, 16(3): 731-739.
- [5] 中国中西医结合学会大肠肛门病专业委员会. 中国痔病诊疗指南(2020)[J]. 结直肠肛门外科, 2020, 26(5): 519-533.
- [6] 杨涛, 朱学芳. 中医辨证智能化研究现状及发展趋势[J]. 南京中医药大学学报, 2021, 37(4): 597-601.
- [7] 王琪, 胡良平, 高颖, 等. 潜在类别分析在缺血性中风病患者证候数据中的应用[J]. 中国卫生统计, 2013, 30(5): 638-640.
- [8] 李伟珂, 伊明洋, 倪园园, 等. 基于潜在类别结合隐结构模型的慢性阻塞性肺疾病肺气虚证分级量化诊断研究[J]. 中医杂志, 2025, 66(7): 710-716.
- [9] 陈子佳, 彭文茜, 张德政, 等. 大语言模型在中医药领域的应用、挑战与前景[J]. 协和医学杂志, 2025, 16(1): 83-89.
- [10] 李欣桐, 马素芬, 张丰聪, 等. 中医药领域大语言模型的研究进展与应用前景[J]. 南京中医药大学学报, 2024, 40(12): 1393-1403.
- [11] 陈子佳, 韩宇, 刘欣, 等. 基于大模型的肺癌表皮生长因子受体突变患者生存预测模型构建与验证[J]. 世界中医药, 2025, 20(8): 1379-1390.
- [12] 智从从, 李雪, 程一乘, 等. 痔中西医结合诊疗指南(2025版)[J]. 中国全科医学, 2025, 28(26): 3217-3228.
- [13] 赵雪, 张海, 王东波. 大语言模型评测研究现状、应用、问题与趋势分析[J]. 情报学报, 2025, 44(8): 1058-1074.
- [14] 谢珉宁, 徐璇, 张磊, 等. 混合痔中医辨证分型与流行病学特点的相关性研究[J]. 江苏中医药, 2016, 48(11): 38-40.
- [15] MCCUTCHEON AL. Basic concepts and procedures in single- and multiple-group latent class analysis [M]. Cambridge: Cambridge University Press, 2002: 56-86.
- [16] SCOTTO RN, BAER JC. Latent class analysis: a method for capturing heterogeneity [J]. Soc Work Res, 2012, 36(1): 61-69.
- [17] 曾宪华, 肖琳, 张岩波. 潜在类别分析原理及实例分析[J]. 中国卫生统计, 2013, 30(6): 815-817.
- [18] 乔敬华, 何佳伟, 周军惠. 基于流行病学调查的农村社区居民痔病中医药防治对策探讨[J]. 上海中医药杂志, 2019, 53(6): 14-19.
- [19] 蓝可彰, 潘海涛, 罗建琼, 等. 广西忻城县 546 例混合痔患者的临床分析[J]. 蛇志, 2017, 29(1): 17-18.
- [20] 刘建乔, 李宇栋, 李雪, 等. 139 例粪便潜血阳性混合痔患者结肠镜检查结果与中医证型的回顾性分析[J]. 北京中医药, 2022, 41(2): 207-209.
- [21] 蔡华花, 郑梅娟, 曾凤荣, 等. 120 例闽北地区混合痔中医证素规律研究[J]. 中国卫生标准管理, 2021, 12(15): 110-113.
- [22] 康健, 余乐来, 龚光明. 500 例混合痔患者的中医体质研究[J]. 湖北中医杂志, 2015, 37(9): 20-21.
- [23] 朱谦让, 戴月, 谢玮, 等. 江苏省居民膳食结构与营养素摄入状况评价[J]. 江苏预防医学, 2017, 28(3): 259-261, 265.
- [24] AVSAR AF, KESKIN HL. Haemorrhoids during pregnancy [J]. J Obstet Gynaecol, 2010, 30(3): 231-237.
- [25] 梁建民. 女性肛肠病的临床发病特点及治疗分析(附 600 例报告)[J]. 中国肛肠病杂志, 2007, 27(7): 2.
- [26] 徐红, 黄卫国, 郭淳. 1795 例常规体检者中痔病患病情况分析[J]. 中国普外基础与临床杂志, 2011, 18(8): 884-886.
- [27] HONG J, KIM I, SONG J, et al. Socio-demographic factors and lifestyle associated with symptomatic hemorrhoids: big data analysis using the National Health insurance Service-National Health screening cohort (NHIS-HEALS) database in Korea [J]. Asian J Surg, 2022, 45(1): 353-359.
- [28] DELCÒ F, SONNENBERG A. Associations between hemorrhoids and other diagnoses [J]. Dis Colon Rectum, 1998, 41(12): 1534-1541.
- [29] HONG YS, JUNG KU, RAMPAL S, et al. Risk factors for hemorrhoidal disease among healthy young and middle-aged Korean adults [J]. Sci Rep, 2022, 12(1): 129.
- [30] 杜勇军, 黄德铨, 杨洁, 等. 混合痔危险因素病例对照研究[J]. 重庆医学, 2017, 46(32): 4553-4555.
- [31] 张晶, 王剑. 赣南地区痔病的危险因素[J]. 中国老年学杂志, 2015, 35(22): 6554-6555.

Syndrome Patterns Distribution and Risk Factors of Mixed Hemorrhoids in Traditional Chinese Medicine : A Multicenter Real-world Study Using Large Language Models and Latent Class Analysis

DENG Ruyue, DING Kang, ZHU Yuxin, LI Meng, ZHU Huiting, DU Lei

Nanjing Hospital of Chinese Medicine Affiliated to Nanjing University of Chinese Medicine, Nanjing, 210022

ABSTRACT **Objective** To develop a standardized classification model for traditional Chinese medicine (TCM) syndrome patterns of mixed hemorrhoids using multi-center real-world data, and unveil their distribution patterns and core risk factors, thereby providing evidence-based support for standardizing TCM syndrome differentiation and implementing precision interventions. **Methods** A multi-center cross-sectional study was conducted, enrolling 13 283 mixed hemorrhoid patients from eight hospitals in Jiangsu Province between September 1st, 2023 and December 31st, 2024. DeepSeek-R1-Distill-Qwen-7B and LLaMA-3. 3 large language models (LLM) were integrated with latent class analysis (LCA) to perform unsupervised learning and latent class modeling of TCM symptomatology. Potential risk factors were screened via univariate analysis, followed by logistic regression to identify independent risk factors for each syndrome pattern. **Results** The model's performance indicators were stable and reliable across different clinical data types, i. e. in the outpatient records, past medical history (F1=99. 7%), current medical history (F1=94. 9%), and specialist examination (F1=90. 7%); in inpatient records, past medical history (F1=98. 2%), current medical history (F1=91. 2%), specialist examination (F1=90. 3%), and discharge status (F1=90. 6%). Latent class modeling identified four core TCM syndrome patterns including spleen deficiency and *qi* sinking syndrome (915 cases, 6. 89%), damp-heat pouring downward syndrome (10 820 cases, 81. 46%), *qi* stagnation and blood stasis syndrome (1252 cases, 9. 43%), and wind injuring intestinal collaterals syndrome (296 cases, 2. 22%), with respective latent class probabilities of 0. 069, 0. 815, 0. 094, and 0. 022. Logistic regression demonstrated that gender, age, disease duration, hypertension, diabetes, hyperlipidemia, constipation, smoking history, and alcohol consumption were independent risk factors for pattern differentiation ($P<0. 05$). The efficacy validation evaluation revealed that the cure rates for patients with spleen deficiency and *qi* sinking syndrome and *qi* stagnation and blood stasis syndrome were higher than those for patients with damp-heat pouring downward syndrome (adjusted $P<0. 05$), with no statistically significant differences among other syndrome patterns. **Conclusion** Damp-heat pouring downward syndrome is the predominant syndrome in mixed hemorrhoids. Gender, age, disease duration, hypertension, diabetes, hyperlipidemia, constipation, smoking history, and alcohol consumption are independent risk factors for the differentiation of syndrome types.

Keywords mixed hemorrhoids; syndromes distribution; large language model; latent class analysis; real-world study

(收稿日期: 2025-09-17; 修回日期: 2026-01-16)

[编辑: 贾维娜]